UNIVERSIDAD DE SALAMANCA

Departamento de Estadística

CONTRIBUCIONES BASADAS EN EL ANÁLISIS BIPLOT AL DISEÑO Y GESTIÓN DE REDES DE TELECOMUNICACIÓN

Francisco Javier Delgado Álvarez

2015

CONTRIBUCIONES BASADAS EN EL ANÁLISIS BIPLOT AL DISEÑO Y GESTIÓN DE REDES DE TELECOMUNICACIÓN

VNiVERSiDAD D SALAMANCA

Departamento de Estadística
Universidad de Salamanca

Memoria que para optar al Grado de Doctor, por el Departamento de Estadística de la Universidad de Salamanca, presenta:

Francisco Javier Delgado Álvarez

Salamanca

2015

Departamento de Estadística
Universidad de Salamanca

DRA. Mª PURIFICACIÓN GALINDO VILLARDÓN

Profesora del Área de Estadística e Investigación Operativa de la Universidad de Salamanca

---

CERTIFICA:

Que **D. Francisco Javier Delgado Álvarez,** ingeniero de telecomunicación, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que para optar al Grado de Doctor, presenta con el título: "CONTRIBUCIONES BASADAS EN EL ANÁLISIS BIPLOT AL DISEÑO Y GESTIÓN DE REDES DE TELECOMUNICACIÓN". Y para que conste, firma el presente certificado en Salamanca, a 23 de noviembre de 2015.

"Tomó la levedad de la hoja y la fugacidad del cervatillo,
la alegría de los ratos de sol y las lágrimas de la neblina;
la inconsistencia del viento y la timidez de la liebre;
la vanidad del pavo real y la suavidad del plumón del cuello de la golondrina.
Añadió la dureza del diamante, la dulzura de la miel,
la crueldad del tigre, el calor del fuego y el frío de la nieve.
Lo mezcló todo y se formó a una mujer.
Luego se lo ofreció como presente al hombre."

[Creación de la mujer - Mitología hindú]

A mis padres y a mi hermana.

AGRADECIMIENTOS

Han sido muchos días los transcurridos desde que comencé a elaborar esta Tesis Doctoral, incluso muchos más de los inicialmente previstos. Confieso que cuando inicié los estudios de tercer ciclo, en 1996, no tenía la menor esperanza de encontrar un tema dentro de la Estadística Multivariante afín a las Telecomunicaciones. Como casi siempre en estos casos, la suerte me acompañó y, todo hay que decirlo, el último día de los cursos surgió la idea que subyace en todo el desarrollo de la Tesis, prácticamente tal cual será expuesta.

En primer lugar tengo que agradecer infinitamente a mi Directora, la Dra. Purificación "Puri" Galindo la calidez, cercanía, profesionalidad y cariño con el que me acogió en "su" departamento, en su vida y las muestras que de todo ello me sigue dando cada día desde entonces. No tengo palabras suficientes para agradecerle todo lo que ha hecho por mi en estos años, tanto profesionalmente, como personalmente. Si la formación encierra además de un aprendizaje de nuevas aptitudes, un desarrollo personal, desde luego esta Tesis es responsable de lo que soy ahora y de muchos de los cambios personales que he experimentado en todos estos años. Ella sabe las muchas alegrías y algunos sinsabores que esta tesis doctoral encierra entre sus líneas. Espero y confío no haberla decepcionado en el pasado, que se sienta orgullosa de hasta donde hemos llegado con esta investigación y que en el futuro sigamos trabajando juntos. Porque deseo que esta Tesis lejos de ser un punto y final a una productiva etapa de mi vida, sea solo un punto y aparte hacia otra etapa aún más productiva, en la que pueda desplegar todo lo aprendido hasta ahora y lo mucho que aún espero aprender en el futuro.

También tengo que agradecer a todo el Departamento de Estadística "Aplicada" [sic] el cariño con el que me han tratado todos estos años. Tanto a los compañeros que ya estaban en el departamento cuando empecé mis estudios de tercer ciclo, como a las posteriores incorporaciones. Muy especialmente recordaré SIEMPRE aquel primer congreso de Vigo y el último al que asistí en Tenerife, que serán para mi inolvidables.

Dedico esta Tesis a mi familia, a mis padres y a mi hermana. Sin su apoyo y aliento hubiese sido desde luego imposible, ya no terminar, sino simplemente iniciar un trabajo tan laborioso como este. Han sido siempre un modelo a seguir y a ellos les dedico este trabajo.

Salamanca, 23 de noviembre de 2015.

RESUMEN

La importancia de la redes de telecomunicación en nuestra sociedad es innegable. Desde la telefonía, tanto fija como móvil, hasta la red Internet están presentes en la mayoría de los hogares, empresas y administraciones públicas. Garantizar su correcto funcionamiento es de una importancia clave y la herramienta fundamental para este objetivo es un adecuado diseño y gestión de la red.

Los métodos biplot, formulados por Gabriel en 1971, permiten representar una matriz de datos en forma de un gráfico que utiliza marcadores individuales para cada una de las filas y las columnas de la matriz de partida, respetando determinadas propiedades de los datos originales.

En el diseño y gestión de redes se pueden utilizar múltiples tipos de matrices conteniendo diversos datos sobre su operación y configuración. Destacan entre ellas las matrices de tráfico, las matrices de topología y combinaciones de ambas. Por otro lado, las representaciones gráficas permiten a los diseñadores y gestores de la red identificar de manera eficiente y eficaz el estado de la red de comunicaciones.

Esta tesis doctoral propone la utilización de los métodos biplot, en general, y del HJ-Biplot, propuesto por Galindo en 1986, en particular, en los procesos de diseño y gestión de redes de comunicación, presentando aplicaciones sobre las redes de datos más habituales hoy en día.

Las propuestas se centran en tres casuísticas generales que cubren un amplio espectro de posibles aplicaciones: detección de anomalías, análisis de series temporales y análisis de la topología de redes.

La detección de anomalías se aplica en un primer ejemplo sobre datos de una red Ethernet real. Se demuestra que es posible utilizar la representación HJ-Biplot con dos objetivos: modelar la red con una representación adecuadamente robusta y detectar incidencias con la suficiente sensibilidad.

En un segundo supuesto se aplica a la detección de un ataque de negación de servicio, como caso especial de anomalía, para lo que se utiliza un juego de datos publicados para la verificación del funcionamiento de este tipo de sistemas. En este apartado se incluye la aplicación del método STATIS para la detección de la anomalía, y finalmente el HJ-Biplot para la diagnosis concreta de la incidencia ocurrida en la red.

El análisis de series temporales utilizando el HJ-Biplot mejora la propuesta realizada por Lakhina *et al* en 2005, que aplicaba el Análisis de Componentes Principales (ACP) a una matriz de tráfico Origen-Destino. El HJ-Biplot tiene en consideración la existencia simultánea de correlaciones temporales y espaciales en la matriz de tráfico y además permite localizar el punto de ocurrencia de la incidencia.

Finalmente, la combinación de la teoría espectral de grafos, aplicada a redes de comunicación, y la metodología biplot en general, y el HJ-Biplot en particular, permite obtener representaciones gráficas de las redes de comunicación con información sobre su topología, incluso incorporando información sobre tráfico cursado, simétrico o asimétrico, entre nodos.

La tesis doctoral presenta algunas contribuciones de los métodos biplot al análisis y gestión de las redes de comunicación más utilizadas en nuestros días. La herramienta propuesta permite mejorar los procedimientos de diseño y gestión de redes constituyendo una potente herramienta de visualización del estado de la red de comunicación.

PALABRAS CLAVE:

ESTADÍSTICA, ANÁLISIS MULTIVARIANTE, BIPLOT, TELECOMUNICACIONES.

TABLA DE CONTENIDO

1. INTRODUCCIÓN

"La unión de pueblos y ciudades puede ayudar a la gente a organizarse y trabajar conjuntamente para resolver problemas locales y regionales, desde mejoras en el abastecimiento del agua hasta la prevención de la deforestación. Para promover, proteger y preservar la libertad y la democracia, debemos hacer del desarrollo de las telecomunicaciones una parte integral del desarrollo de todas las naciones. Cada nexo de unión que establezcamos reforzará los vínculos de libertad y democracia alrededor del mundo."

Al Gore, Vicepresidente EEUU [1]

1. INTRODUCCIÓN

1.1. JUSTIFICACIÓN

El Diccionario de la Lengua Española define la "comunicación" con varias acepciones:

1. Acción y efecto de comunicar o comunicarse.
2. Trato, correspondencia entre dos o más personas.
3. Transmisión de señales mediante un código común al emisor y al receptor.
4. Unión que se establece entre ciertas cosas, tales como mares, pueblos, casas o habitaciones, mediante pasos, crujías, escaleras, vías, canales, cables y otros recursos.
5. Medio que permite que haya comunicación (unión) entre ciertas cosas.
6. Papel escrito en que se comunica algo oficialmente.
7. Escrito sobre un tema determinado que el autor presenta a un congreso o reunión de especialistas para su conocimiento y discusión.
8. Petición del parecer por parte de la persona que habla a aquella o aquellas a quienes se dirige, amigas o contrarias, manifestándose convencida de que no puede ser distinto del suyo propio.
9. Correos, telégrafos, teléfonos, etc.

El nacimiento de la comunicación como tal puede perfectamente datarse coetáneamente con la aparición de vida "social" en la Tierra: los animales, de hasta incluso los más bajos escalones de la pirámide ecológica, intercambian signos, sonidos, y símbolos para transmitir información.

El diccionario de la Lengua Española define la "telecomunicación" como "sistema de transmisión y recepción a distancia de señales de diversa naturaleza por medios

electromagnéticos". La importancia hoy en día de las redes de telecomunicación en nuestra sociedad es algo obvio.

John Mayo, presidente de los Laboratorios Bell entre 1991 y 1995, formuló el objetivo principal de las redes de comunicación [2]: proveer acceso a voz, datos e imágenes, o cualquier combinación de ellas, en cualquier lugar, en cualquier momento… de manera adecuada y económica. Este simple pronunciamiento se ha convertido en el *leit motiv* del diseño y gestión de redes desde aquellos días hasta hoy.

Las redes de telecomunicación, en general, abarcan aspectos tan tradicionales como las clásicas redes telefónicas analógicas, hasta las modernas redes de transmisión de datos, en las que convergen tanto las comunicaciones telefónicas clásicas, ahora digitalizadas, como el acceso a Internet o la televisión, entre otros servicios. En muchos casos todas estas formas de comunicación emplean la red Internet como mecanismo de transporte, por lo que, en realidad, es esta red Internet la que de manera efectiva sirve como verdadero canal de comunicación. Así pues, la convergencia ha culminado en una única red de comunicación, que sustenta todos los tipos de información posibles, y que por lo tanto debe ser adecuadamente diseñada y gestionada.

En tan solo los 20 años transcurridos desde la irrupción de la World Wide Web en la red Internet, como paradigma de democratización de las telecomunicaciones, ambas tecnologías se han vuelto omnipresentes. Algo que en un principio parecía más restringido a ámbitos científicos o empresariales ha llegado a nuestros hogares, para quedarse: según publica la Comisión del Mercado de las Telecomunicaciones (CMT) en su Informe sobre los consumos y gastos de los hogares españoles en los servicios de comunicaciones electrónicas, correspondiente al primer semestre de 2014 (publicado en marzo de 2015), el 70,0% de los hogares españoles disponían de conexión a Internet, con un crecimiento de más de un 1,7% en los 12 meses precedentes. Por lo que respecta al comercio electrónico, como uno de los servicios más relevantes de Internet[1], la CMT publica en su informe sobre el comercio electrónico en España a través de las entidades de medios de pago, correspondiente al primer trimestre de 2015, que el volumen de negocio del comercio electrónico alcanzó los 4.455,7 millones de euros, lo que supone un incremento del 24,5% interanual; con un total de 67,7 millones de operaciones. Estas operaciones abarcan

[1] Aunque es evidente que los términos Internet y World Wide Web no son equivalentes, en algunos apartados no técnicos se podrían utilizar de manera intercambiable. La interpretación del concepto concreto al que nos referimos en cada caso es considerada evidente en función del contexto en el que se utilice el respectivo término.

prácticamente cualquier rama de actividad: agencias de viajes y operadores turísticos, transporte aéreo, marketing directo, transporte terrestre de viajeros, espectáculos artísticos, deportivos y recreativos, vestido, juegos de azar, publicidad, alimentación y, por último pero no menos importante, administración electrónica.

Internet ha experimentado una tasa de crecimiento superior a otras tecnologías tales como la radio, la televisión o incluso la misma microinformática personal. Internet ha impulsado la colaboración a través de la red, la educación, el comercio y el entretenimiento. El futuro será aún más exigente con la convergencia de voz, video y datos (*triple play*) compartiendo redes [3].

En el ámbito empresarial se han desarrollado, al amparo de las nuevas tecnologías de la información y la comunicación, nuevos paradigmas de gestión empresarial, uno de los cuales es el de la Intranet. Una Intranet es una red que conecta un conjunto de ordenadores y dispositivos utilizando protocolos de Internet. En una Intranet los nodos de red se encuentran confinados en la entidad que la opera, y están protegidos de mundo exterior [4]. Otros autores, no obstante, afirman que lo que caracteriza una Intranet no es su "confinamiento", sino su uso por una empresa como canal de comunicación para el intercambio interno de información [5]. Así, las Intranets pueden ser desplegadas para su utilización exclusiva interna sobre redes locales pero también sobre redes de área extensa, o incluso a través de conexiones externas a través de Internet. Las Intranets están cambiando los entornos corporativos: nacieron como soporte a los procesos productivos y están cambiando esos mismos procesos y la relación de las corporaciones con el entorno. El funcionamiento de muchas organizaciones dependen de sus Intranets, así pues, su gestión es de gran importancia [5].

Las situaciones planteadas anteriormente sobre la extensión del uso de la red Internet en la sociedad y de las Intranet en ámbito empresarial justifican, en nuestra opinión, la elección de los escenarios sobre los que se aplicarán las propuestas metodológicas de esta tesis doctoral.

Pero hay otro aspecto interesante en la aplicación de los métodos que serán propuestos al diseño y principalmente a la gestión de redes de telecomunicación de datos. Sin entrar en profundidad en consideraciones técnicas sobre el propio concepto de gestión de redes, aceptaremos inicialmente la definición básica propuesta por Hegering, *et al* [6] para quien la gestión de redes comprende todas las acciones y precauciones tomadas para garantizar el uso efectivo y eficiente de los recursos

físicos y lógicos de los sistemas finales distribuidos y las redes de comunicación que los interconectan. La gestión de redes de computadores es esencial para mantener hoy en día la salud de las redes de muchas organizaciones. Las redes que sufren de una débil gestión, usualmente experimentan problemas de prestaciones y baja disponibilidad [7]. Algunos expertos [8]–[13] abogan por utilizar representaciones visuales de la información recogida a través de los sistemas de gestión de redes. También en este aspecto las técnicas estadísticas multivariantes con representación gráfica propuestas en esta tesis doctoral pueden aportar eficacia informativa.

Porque, como indica el experto en gestión de redes Bruce Boardman [8], “en las estadísticas [asociadas a la gestión de redes], lo malo son los detalles. Hay que terminar con la tortura de estadísticas difícilmente comprensibles y carentes de sentido, se precisan herramientas estadísticas claras y concisas que sean útiles y estén dotadas de autoridad, aportando una visión clara del caos que supone la gestión de una red de gran tamaño y heterogeneidad”.

1.2. ESTRUCTURA DE LA TESIS DOCTORAL

La presente tesis doctoral se desarrolla en siete grandes apartados:

En primer lugar se realizará una introducción a las diferentes tecnologías y paradigmas relacionados con las redes de telecomunicación, y su diseño y gestión. Se expondrán someramente los mecanismos de funcionamiento de las redes de comunicación. Se pondrá énfasis en especial en la red Internet y en las redes basadas en la norma popularmente conocida como Ethernet. Posteriormente se realizará un análisis a alto nivel de las implicaciones y algunos procedimientos que pueden utilizarse para el diseño y gestión de redes basadas en Ethernet y también sobre la red Internet y su infraestructura de soporte, que en algunos casos puede ser la propia norma Ethernet. En este punto estaremos ya en condiciones de exponer los objetivos que se establecieron para la presente investigación.

En segundo lugar se analizarán con detalle aspectos relacionados con la visualización de información relativa a redes de telecomunicación y su importancia en las fases de diseño y gestión de redes. Las técnicas de visualización, al incluir al observador humano en el análisis de las imágenes obtenidas, permiten alcanzar muy buenos resultados al facilitar al administrador información fidedigna y fácilmente interpretable para la toma de decisiones

El tercer gran apartado de esta tesis doctoral esta dedicado a una exposición formal de los métodos estadísticos multivariantes sobre los que se apoyan las contribuciones al diseño y gestión de redes que se proponen. Se detallarán los diferentes métodos Biplot propuestos en la literatura y también los métodos STATIS, ya que estos últimos serán utilizados en algún punto del desarrollo de la tesis. Como no podría ser de otra manera, se enumerarán algunas de las muchas aplicaciones publicadas de los métodos Biplot en general, y del HJ-Biplot en particular, en campos del conocimiento dispares.

El cuarto apartado está dedicado a la detección de anomalías en redes de comunicación. Como veremos, este aspecto es de especial relevancia en la gestión de redes de comunicación, ya que puede facilitarnos la detección de una incidencia en la red incluso antes de que los efectos más perniciosos de la misma aparezcan y sea percibida por los usuarios. Tras un extenso repaso a la bibliografía existente sobre detección de anomalías, que nos ha servido como guía para nuestro trabajo en este campo, se presentarán dos ejemplos de aplicación del HJ-Biplot a la detección de

anomalías, cada uno de ellos sobre un escenario diferente. En el segundo supuesto, además, incluiremos un método STATIS para separar las fases de detección de la anomalía y de diagnóstico de la misma, para lo que se aplicará el método HJ-Biplot.

El quinto apartado analiza la aplicabilidad del HJ-Biplot al estudio de las series numéricas, en particular a las series temporales, y más concretamente a aquellas obtenidas de lecturas de tráfico de redes de comunicación. Veremos cuales son las especiales características de estas series temporales a través de una revisión bibliográfica profunda y como los métodos Biplot en general, y el HJ-Biplot en particular, pueden ayudarnos a explorar múltiples series temporales simultáneamente, proporcionándonos información sobre su comportamiento que difícilmente podríamos obtener mediante una representación "más tradicional". Además, por su construcción matemática, los métodos Biplot nos permiten, de una manera sencilla y fácilmente comprensible, abordar algunos de los problemas que algunos autores han detectado en el análisis de este tipo de series temporales, en los que están presentes correlaciones entre las propias series temporales y entre los intervalos de tiempo en los que se han obtenido cada una de sus muestras.

El sexto apartado propone aplicar los métodos Biplot, y en particular el HJ-Biplot, al análisis de las topologías de redes, incluso incorporando datos de tráfico cursado en ellas. La posibilidad de representar las redes de comunicación en forma de grafos y estos a su vez en forma de matrices con diferentes definiciones, abrió la puerta a la aplicación de técnicas matemáticas establecidas para los grafos a las redes de comunicación. Pero, es más, la capacidad de representación que provee el HJ- Biplot sobre matrices de datos se coaliga perfectamente, como veremos, con las técnicas aplicables a los grafos, para ofrecer nuevas posibilidades de análisis de redes de comunicación.

Finalmente el séptimo y último apartado está dedicado en exclusiva a la exposición formal de las conclusiones derivadas de los estudios realizados en el marco de esta tesis doctoral.

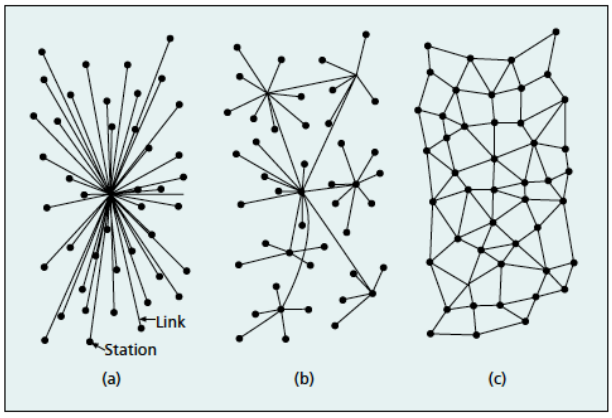
1.3. INTRODUCCIÓN A LAS REDES DE COMUNICACIÓN DE DATOS

1.3.1. LA TECNOLOGÍA DE CONMUTACIÓN DE PAQUETES

Quizá la mejor manera de introducir las redes de comunicación de datos sea repasar su historia de la mano de uno de sus creadores, Paul Baran, considerado como el padre de la tecnología de conmutación de paquetes con sus artículos de principios de los años 60 [14], [15]. Más recientemente, antes de su fallecimiento el 26 de marzo de 2011 [16], se publicó un artículo [17] en el que rememoraba la invención y que utilizaremos como hilo conductor de esta introducción.

En contra de lo que últimamente se suele afirmar [18], la tecnología de conmutación de paquetes que subyace en las redes de telecomunicación actuales es consecuencia directa de la Guerra Fría, que alcanzaba su punto álgido en los años sesenta. Estudios con simulaciones demostraron que en el caso de un ataque contra Estados Unidos, las fuerzas estratégicas podrían quedar inoperativas debido a los daños sufridos por la red de telefonía, todo ello a pesar de que gran parte del sistema de comunicaciones quedase en pie. Esta posibilidad, unida a la necesidad de que precisamente esa capacidad quedase intacta para poder dar respuesta al primer golpe de manera contundente (en lo que se llamó Destrucción Mutua Asegurada o MAD *Mutual Assured Destruction*), motivó por sí misma la investigación en redes robustas: una infraestructura de comunicaciones para comando y control (C&C) es obligatoria para mantener las armas enfundadas entre los adversarios.

Pero ¿por qué algunas redes de comunicaciones son tan vulnerables? Consideremos tres topologías simples de redes de comunicación:



Tres topologías de redes a) Centralizada b) Descentralizada c) Distribuida. [17]

a) **Red centralizada**: Todos sus nodos están conectados a un nodo central que se encarga de comunicar todos los nodos que forman la red, siendo la red altamente vulnerable a la fiabilidad de este único punto de fallo.

b) **Red descentralizada**: Es el esquema clásico de las redes de telefonía tradicionales (denominadas en la literatura "*POTS"* por *Plain Old Telephone System*). Presenta mejor capacidad de supervivencia que la anterior, ya que no existe un único punto de fallo. En lugar de un único punto de conmutación, la red tiene varios puntos que se encargan de las comunicaciones entre nodos a corta distancia, y sólo las comunicaciones a larga distancia se encaminan por enlaces a larga distancia.

c) **Red distribuida**: Es una red sin estructura jerárquica; no hay ningún punto de fallo que haga vulnerable a la red al completo. La robustez de la red puede además reforzarse mediante diferentes niveles de redundancia en los enlaces entre nodos. De hecho posibilita la construcción de redes extremadamente fiables aún con enlaces poco fiables, con la utilización de redundancia.

La estrategia planteada por Baran implicaba la digitalización y paquetización (troceado) de toda la información que viajaba a través de la red. La transmisión digital hacía además posible la regeneración de la señal en los nodos que atravesase, lo que evitaba la degeneración causada por recorrer la red. Por otro lado de alguna manera la propia información debería contener referencias suficientes sobre la ruta que debería ésta seguir para llegar a su destino.

La primera prueba de concepto que se llevó a cabo consistió en una red robusta de teletipos. El reto consistía en diseñar una estructura de conmutación escalable, capaz de dirigir dinámicamente tráfico de "alta velocidad" (de la de entonces) entre un elevado número de usuarios potenciales y con requerimientos no conocidos con antelación. El planteamiento era, como veremos más adelante, muy similar al de las redes de área local actualmente en servicio.

Un concepto importante que resolver en una red de comunicación como la planteada, era como descubrir fácil y rápidamente los caminos que debe de seguir la señal a lo largo de la red, especialmente cuando está siendo atacada y, por ejemplo, algunos enlaces o nodos dejan de estar operativos. Simulaciones demostraron que una red con esta nueva tecnología tenía muchas posibilidades de sobrevivir a la pérdida instantánea del 50% de los enlaces. El procedimiento que se utilizó para encaminar el tráfico (protocolo en enrutamiento) era muy simple: cada bloque de mensaje,

denominado paquete, disponía de un apartado de dirección “desde” y “para” y un contador que se incrementaba en una unidad cada vez que atravesaba un nodo. El valor de este contador era un estimador de la longitud de la ruta seguida por el paquete. Cada nodo aprendía (y olvidaba periódicamente) esta información, lo que le permitía responder a posibles cambios en la disponibilidad de los nodos.

El procedimiento se asimilaba a situar un “cartero” en cada nodo. El cartero podría estimar la mejor dirección para enviar en el futuro la correspondencia destinada a una dirección de la que proviene un carta a partir de la “fecha” más cercana del matasellos de la carta recibida. Dicho de otro modo, observando el trafico que atraviesa un nodo y registrando el contador de la estación de origen, junto con el enlace del que proviene, el cartero imaginario puede determinar el mejor enlace para dirigir un paquete a una dirección, e incluso ordenarlos por prioridad. Cuando el enlace mejor esta ocupado o no está disponible, se puede enviar a través del segundo mejor.

Una cuestión importante es que al dividirse la información en paquetes y poder seguir estos rutas diferentes, no está garantizado en modo alguno que lleguen a su destino en orden, así pues es también necesario colocar un número de secuencia, para que el nodo de destino pueda reordenar el mensaje correctamente. Por último, y para garantizar la integridad del mensaje, se añade un código de redundancia cíclica (CRC) que nos permita comprobar si el mensaje ha sufrido alguna alteración en su camino.

Volvamos a un aspecto importante, las direcciones que deben constar en los paquetes de información. Una diferencia con las redes de conmutación de circuitos tradicionales es la separación entre direcciones físicas direcciones lógicas. De nuevo esta separación, habitual aún actualmente, proviene en parte de los requerimientos de un sistema diseñado para comunicaciones de comando y control que debe evitar la existencia de puntos de fallo únicos. Eliminando el vínculo entre la dirección física y lógica, se introduce un nuevo grado de libertad que es muy útil en múltiples situaciones: donde quiera que se sitúe el emisor en la red (dirección física), puede mantener su dirección lógica.

A pesar del tiempo transcurrido desde la concepción de las ideas anteriores, todos los principios expuestos siguen vigentes hoy en día por ejemplo en la red Internet.

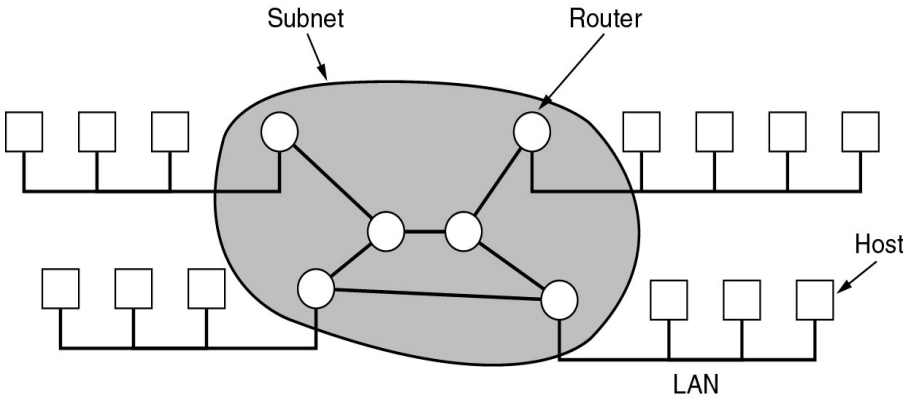
1.3.2. CLASIFICACIÓN DE LAS REDES POR SU ÁREA GEOGRÁFICA

No hay una única clasificación aceptada a la que se ajusten las redes de ordenadores [19], pero hay dos que destacan de manera importante: aquellas relativas a la tecnología de transmisión y a su escala. Nos centraremos en esta última clasificación.

Las redes de área local, generalmente conocidas como redes locales, LANs (*Local Area* Network) o incluso RAL, son redes de propiedad privada que se encuentran en un solo edificio o como mucho en un complejo de edificios de pocos kilómetros de longitud. Se utilizan ampliamente para conectar ordenadores en oficinas, fábricas y hogares, permitiendo compartir recursos tales como impresoras, dispositivos de almacenamiento o accesos a Internet, por ejemplo. Las redes locales están caracterizadas, por lo tanto, por su tamaño, lo que implica que el tiempo de transmisión de la información entre sus extremos, en el peor de los casos, es limitado y conocido de antemano. El hecho de conocer este límite permite utilizar ciertos tipos de diseño que de otro modo no serían posibles.

Las redes de área metropolitana o MAN (*Metropolitan Area Network*) suelen abarcar el área de una ciudad. Al principio eran sistemas diseñados con fines específicos (por ejemplo la televisión por cable) si bien en épocas más recientes comenzaron a utilizarse específicamente también como redes de comunicación de datos.

Finalmente las redes de área amplia o extendida o WAN (*Wide Area Network*) abarcan una gran área geográfica, con frecuencia un país, un continente, o todo el planeta (de hecho incluso más allá de nuestro propio planeta [20]). Contienen un conjunto de ordenadores, que se suelen denominar *hosts* y que habitualmente están conectados a redes de área local, que a su vez están conectadas entre sí mediante otras redes y/o líneas de comunicación, utilizando unos equipos denominados *routers* o enrutadores.



Relación entre *hosts*, redes locales (LAN) y redes WAN. [19]

Debido a las diferentes características de las estos tres tipos de redes, principalmente su tamaño y por consiguiente el tiempo de transmisión, las tecnologías utilizables en cada una de ellas suelen ser diferentes. No obstante una tecnología sobresale en este planteamiento, por su versatilidad y escalabilidad, y que presenta aplicaciones posibles en todas las tipologías de redes anteriores: Ethernet [21]–[23].

1.4. REDES "ETHERNET"

1.4.1. INTRODUCCIÓN

Ethernet es una tecnología de redes de área local (abreviadamente RAL o LAN por *Local Area Network*) que permite en la actualidad la transmisión de información a velocidades superiores al Gigabit por segundo entre equipos situados a distancias intermedias y a través de un medio de transmisión físico, que puede ser cobre (o en general "metálico") o fibra óptica. Hoy en día es habitual disponer en el puesto de trabajo en la empresa o institución, e incluso en el hogar, de velocidades de transmisión de 1 Gb/s, en equipos dotados con puertos Ethernet.

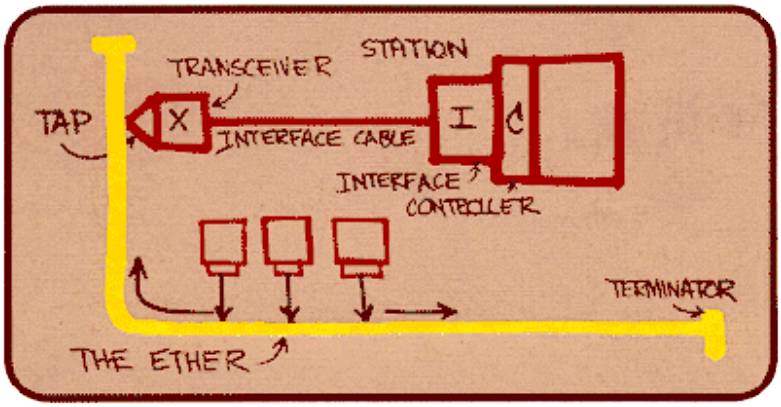
Desde hace algún tiempo Ethernet es prácticamente la tecnología de redes de área local hegemónica [24]. Si bien durante sus estadios iniciales compartió mercado con otras soluciones, ya en 1994 había instalados en el mundo 40 millones de nodos Ethernet [25]. En 2006 se fabricaron un total de 182.817.000 puertos Ethernet [21]. Solo en el año 2012 se vendieron 360 millones de puertos Ethernet 1GbE, 10GbE, 40GbE y 100GbE en equipos para empresa y operadores de telecomunicación [26].

La amplia popularidad de esta tecnología ha permitido que disponga de unos precios muy competitivos, así, la totalidad de fabricantes de ordenadores personales incluyen en sus equipos puertos Ethernet para facilitar su conexión a redes de área local. Se trata de una tecnología sencilla, económica, abierta, flexible, escalable y normalizada internacionalmente [23], [27]–[29]. No obstante, inicialmente existía una cierta ambigüedad entre tecnologías, y hablar de Ethernet no era equivalente al 100% a la norma internacional publicada [19]. El mismo "padre" de esta tecnología ha bromeado sobre su éxito, apuntando que la razón de su supervivencia durante tanto tiempo es la voluntad de la gente de redefinir una y otra vez lo que representa el término Ethernet. Dejando a un lado todos los aspectos lingüísticos, la verdad, en nuestra opinión, es seguramente la inversa: Ethernet ha triunfado por su capacidad de asimilar nuevos escenarios. A pesar de todos los cambios que ha sufrido durante sus 40 años de existencia, Ethernet sigue siendo la misma en su núcleo [30]. Y todo ello a pesar de

que a mediados de los 90 se pensaba que Ethernet alcanzaría su límite de prestaciones sobre cables de pares trenzados con velocidades de 100 Mb/s para distancias no superiores a 100 metros. Porque, se preguntaban ¿quién necesitaría más de 100 Mb/s en el puesto de trabajo? Con ello, la entonces naciente normativa de Gigabit Ethernet (GbE) se enfrentó a muchas burlas. Hoy en día la mayoría de los ordenadores personales incorporan puertos Ethernet a 1 Gb/s, ya comienzan a instalarse puertos a 10 Gb/s [31] y desde hace algún tiempo se planean despliegues a 100 Gb/s tanto en centros de datos corporativos [24] como en redes de área expandida [22]. Mientras en el mundo real sucede esto, en los laboratorios y grupos de estudio de la *Ethernet Alliance* comienza a hablarse de la norma Ethernet para velocidades de 400 Gb/s [32], [33].

1.4.2. HISTORIA DE LAS REDES ETHERNET

Ethernet fue concebida en los años 70 en el Xerox Palo Alto Research Center (PARC) por el Dr. Robert “Bob” M. Metcalfe. Concretamente el concepto de red Ethernet fue esbozado por primera vez por el Dr. Metcalfe el día 22 de mayo de 1973, habiendo cumplido ya 40 años [34].



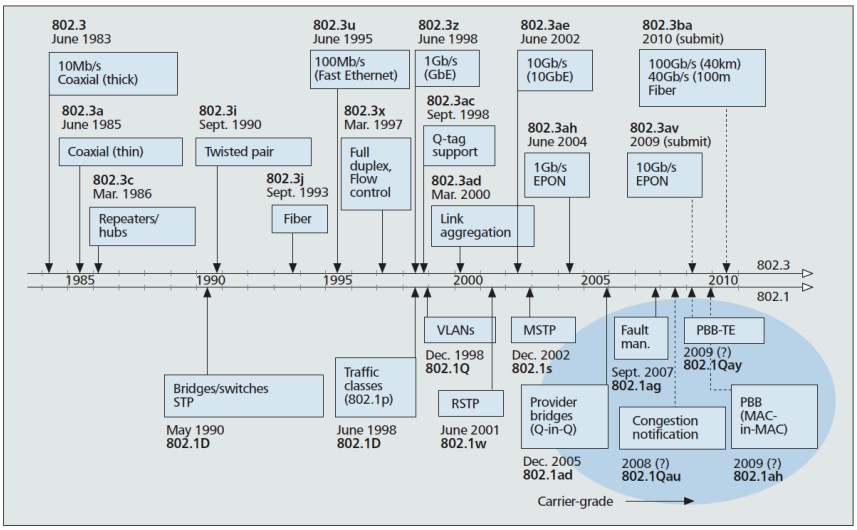
Primer boceto de la tecnología Ethernet (fuente: www.hp.com)

En la ya publicación histórica de 1976 en la que se proponía formalmente esta tecnología [35] los autores (Metcalfe y Boggs) definían Ethernet como un sistema de comunicaciones de difusión para transportar paquetes de datos digitales entre estaciones distribuidas localmente. En ese sistema el “*ether*” es un medio pasivo de transmisión sobre el que no existe un control centralizado. La coordinación en el acceso al medio de transmisión se encuentra distribuida entre las estaciones que integran la red, utilizando un sistema de arbitraje estadístico que expondremos brevemente más tarde.

Como simple curiosidad, es interesante remarcar el motivo por el que denominaron Ethernet a su propuesta [19]: El concepto de “Ether” (=éter) proviene de *luminiferous æther*, que era la “sustancia” a través de la que se pensaba antiguamente que se propagaba la radiación electromagnética. Efectivamente, cuando en el siglo XIX el físico inglés James Clerk Maxwell descubrió que la radiación electromagnética se podía describir mediante una ecuación de onda (en sus famosas ecuaciones de Maxwell), los científicos coetáneos supusieron que el espacio exterior debería estar lleno de algún medio “etéreo” en el que se propagasen dichas ondas, como sucede en los líquidos. El experimento de Michelson-Morley [36] en 1887 probó que la radiación electromagnética se podía propagar en el vacío y que por lo tanto no era precisa la intervención de medio (físico) “etéreo” alguno. La denominación de “*Ether*”net es, pues, un tributo a aquel medio físico inexistente.

En el caso de la red Ethernet el medio de transmisión no era el “ether”, sino un cable coaxial que posibilitaba en sus primeras versiones una velocidad de transmisión de hasta 2.94 Mb/s. La tecnología Ethernet propuesta por Xerox fue tan exitosa que DEC, Intel y Xerox formularon un estándar en 1978 para una Ethernet ya a 10 Mb/s que denominaron estándar DIX y que con escasos cambios se convirtió en el estándar IEEE 802.3.

Como a veces sucede, la empresa Xerox mostró poco interés en el desarrollo de Ethernet, así que Metcalfe fundó en 1979, conjuntamente con otros emprendedores, su propia empresa, 3Com Corporation (“*Computers, Communication and Compatibility”)*, para comercializar adaptadores Ethernet para ordenadores. Esta empresa fue adquirida en 2010 por Hewlett-Packard [37].



Hitos temporales en la normalización de la tecnología Ethernet. [29]

Desde 1978 Ethernet continuó su desarrollo y de hecho aún continúa en evolución [38]. En 1985 se estandarizó la norma para transmisión a 10Mb/s sobre cable coaxial. Evolucionó posteriormente con versiones sobre cable coaxial, cable de pares trenzados y fibra óptica, con velocidades [29] de 10 Mb/s, 100 Mb/s, en 1995 (IEEE 802.3u), 1 Gb/s, en 1998 (GbE, 802.3z), y 10 Gb/s, en 2002 (802.3ae) y para distancias de hasta 2 Km. En el año 2010 se aprobó el estándar 802.3ba para Ethernet a velocidades de 40 Gb/s y 100 Gb/s [39]. Ethernet incluso ha roto la barrera de las redes de área local, normalizándose versiones para redes de área metropolitana [23].

1.4.3. FUNCIONAMIENTO DE LAS REDES ETHERNET

Veamos a continuación y brevemente el funcionamiento de una red basada en la norma Ethernet [25].

Cada equipo dotado de un puerto Ethernet opera independientemente de los otros equipos conectados a la misma red de área local, esto es, no existe un "control central de la red", o similar, como sí sucede con otras tecnologías. En principio todos los equipos conectados a una misma red comparten un único medio de transmisión (el célebre "éter"). Si un equipo debe enviar una información primero escucha si en el medio de transmisión hay alguna transmisión en curso, y cuando el medio está libre, transmite la información en forma de una trama o paquete Ethernet. No obstante, y debido al hecho de que la velocidad de transmisión es finita, puede darse el caso de que dos estaciones hayan detectado el medio libre en el mismo momento, e inicien casi simultáneamente la transmisión, lo que transcurrido el tiempo necesario, hará que ambas señales se superpongan, en lo que se denomina una "colisión".

El protocolo someramente así descrito se denomina CSMA/CD (*Carrier Sense Multiple Access / Collision Detection*). En un símil CSMA/CD funciona como una cena en una habitación a oscuras. Todos los asistentes alrededor de la mesa deben escuchar hasta detectar un tiempo de silencio antes de hablar. Una vez que han detectado ese intervalo de silencio, todos los comensales tienen igual prioridad para comenzar a hablar. Si dos personas comienzan a hablar en el mismo instante, lo detectan y dejan de hablar.

Traducido este símil en términos técnicos, cada interface/tarjeta debe esperar hasta que no haya señales en el canal de comunicación, entonces comienza a transmitir. Cuando una interface está transmitiendo hay una señal específica en el canal que se

denomina portadora (*carrier*). Todas las estaciones deben esperar a que la portadora desaparezca antes de intentar realizar una transmisión, este proceso es el denominado *Carrier Sense*.

Dado que como se ha indicado todas las interfaces tienen igual prioridad para enviar tramas a la red, se habla de que estamos en presencia de un acceso múltiple (*Multiple Access*). Al emplear las señales un tiempo finito en viajar de un extremo a otro de la red Ethernet en cuestión, el inicio de la transmisión, y la propia portadora, no alcanzan todos los puntos de la red simultáneamente. Así, es posible que dos interfaces detecten el canal libre a la vez e inicien sus respectivas transmisiones casi simultáneamente. Cuando esto sucede se produce una colisión, las interfaces detectan la colisión como una mezcla de señales eléctricas sin sentido y cesan la transmisión. Se inicia en este punto el sistema de arbitraje estadístico para evitar (intentar evitar, para ser más precisos) una nueva colisión. El procedimiento se denomina algoritmo de *backoff* y consiste en la elección por todas las estaciones implicadas en la colisión de un intervalo temporal aleatorio en el que deberán mantener silencio antes de reintentar la transmisión. Aunque el termino “colisión” pueda aparentar la ocurrencia de un problema, avería, o incidencia técnica en la red, la verdad es que las colisiones son normales en redes Ethernet y de hecho es igualmente normal el incremento de las mismas conforme aumenta el número de estaciones en la red, la red tiene mayores dimensiones y/o el tráfico en la misma aumenta. En una red Ethernet con cargas elevadas de tráfico puede incluso suceder que ocurran colisiones múltiples, esto es, colisiones consecutivas dentro de un mismo intento de transmisión. De nuevo esto forma parte de la normalidad de operación de la red. El sistema prevé que si ocurren colisiones múltiples las estaciones implicadas comenzarán a ampliar los intervalos de espera, dentro de un patrón aleatorio, antes de intentar transmitir de nuevo su información para evitar con mayores probabilidades la posibilidad de una nueva colisión.

Ciertamente este es el modo de operación en un único dominio de colisión, todas las estaciones que constituyen una red comparten “interlocución”. Es posible segmentar una red Ethernet para establecer varios dominios de colisión con el objetivo de reducir el número de “interlocutores” por cada segmento y así disminuir la probabilidad de colisiones. En el límite, esta propuesta puede llevarse al extremo de definir un único dominio de colisión por equipo conectado a la red, en lo que se denominan redes conmutadas. Típicamente los conmutadores (o *switches*) tienen una dirección MAC en cada puerto que comparte dominio de colisión en la estación (o estaciones) a las que

sirven. Estos equipos construyen (¡y mantienen!) una tabla de direcciones origen (SAT – *Source Address Table*) asociando cada dirección de origen de las tramas entrantes con su correspondiente puerto en un proceso de aprendizaje continuo. Si la dirección de destino de un paquete a retransmitir no está en la lista de alguno de los puertos de salida, se envía el paquete a todos (difusión) [29]. Otra de las mejoras es la operación en modo *full dúplex*, permitiendo a los nodos transmitir y recibir simultáneamente aprovechando la existencia de enlaces físicos diferentes para cada sentido de comunicación.

1.4.4. ALGUNOS TIPOS DE ANOMALÍAS EN LAS REDES ETHERNET

En las redes Ethernet, y también análogamente en otras tecnologías distintas, pueden acaecer diversos tipos de incidencias que afectan al normal funcionamiento de la red local, algunas de la cuales se exponen a continuación por su relevancia en nuestro trabajo [2], [40]:

- **Tormenta de difusión (*broadcast storm*)**: Un mensaje de difusión es una trama o paquete especial que todos los equipos que forman la red deben recibir y procesar. Una tormenta de difusión es una situación en la que los mensajes de difusión son demasiado elevados, pudiendo potencialmente afectar al normal funcionamiento de la red. Esta incidencia ocurre por lo general debido a una avería en la red.
- **Nodo balbuceante (*Babbling node*):** Transmisión de paquetes aleatorios y sin sentido en la red, con frecuencia debido a una tarjeta de red averiada.
- ***Runts***: Paquetes que son más pequeños que los paquetes de menor longitud permitidos por el protocolo de red, en el caso de Ethernet 60 bytes.
- ***Jabbers***: Paquetes que son más grandes que el mayor tamaño permitido por el protocolo de red, en el caso de Ethernet 1518 bytes.
- **Desemparejamiento Ethernet dúplex**: En un entorno de red autoconfigurado, los dos puertos Ethernet de un mismo enlace pueden encontrarse en modos dúplex diferentes después de la fase de negociación (uno en modo *full* y otro en *half*). Este desemparejamiento produce problemas en la red, lo que requiere la reconfiguración manual.
- **Corrupciones en el enlace**: Se refiere a daños físicos en el medio de transmisión que no llegan a ocasionar un corte en el enlace, pero sí problemas con las transmisiones tales como tamaños de trama incorrectos, colisiones

excesivas para el nivel de tráfico existente, errores en los códigos de redundancia cíclica, prestaciones incongruentes, entre otras posibles incidencias.

- **Sobrecargas en la red**: el ancho de banda es ocupado por tráfico no deseado. Este puede ser causado, por ejemplo y entre otras posibilidades, por un gusano, o por una excesiva cantidad de paquetes ARP en lo que se denomina una tormenta de ARP (*ARP storm*), o incluso por paginación de memoria a través de la red, esto es, equipos que operan sin disco local ejecutan un trabajo para que el que no tienen suficiente memoria física y paginan virtualmente la que precisan en un disco ubicado en la red, causando un elevado tráfico en la misma.
- **Ruido eléctrico**: ocasionado por interferencias debido a la proximidad de fuentes de ruido tales como maquinaria de potencia (ascensores, por ejemplo).
- **Caídas de corriente**: interrupciones o fluctuaciones en el suministro de fluido eléctrico.

Won propone [40] diversas métricas para la detección de condiciones anómalas. Se enumeran a continuación aquellas de entre las propuestas que son más específicas de las redes Ethernet:

Métrica	**Condición de alarma**
Colisión de tramas	Primera aparición o basado en umbral
Trama Jumbo (≥1514 bytes)	Primera aparición
Trama Runt (≤64 bytes)	Primera aparición
Trama con error CRC (redundancia)	Primera aparición
Tiempo (ms) entre llegada de paquetes	Incremento sobre el valor previo
Prestaciones (b/s)	Decremento, caída a 0, o patrón analizado sobre el periodo monitorizado
Paquetes por segundo o pico de paquetes	Incremento, decremento, caída a 0, o patrón analizado sobre el periodo monitorizado
Paquetes de difusión	Basado en umbral
Paquetes de protocolo no soportado	Basado en umbral

La detección automática de anomalías en redes Ethernet comenzó a plantearse a principios de la década de los 90, con los trabajos de Maxion [41] y Feather *et al* [42] y, como veremos posteriormente, ha continuado hasta nuestros días.

1.5. LA RED INTERNET

1.5.1. INTRODUCCIÓN

De nuevo nadie mejor para explicar la historia del nacimiento de Internet, que uno de sus “padres”, Leonard Kleinrock, que la puede contar en primera persona [43], [44].

Uno de los eventos precursores de la red Internet, afirma Kleinrock, fue el lanzamiento del Spútnik, el 4 de octubre de 1957. Este evento preocupó notablemente a los Estados Unidos y en respuesta a él se creó la Agencia de Investigación de Proyectos Avanzados (ARPA, *Advanced Research Projects Agency*), para promover la investigación que asegurase que “los comunistas nunca batiesen a América en cualquier carrera tecnológica” (sic). Uno de los departamentos de ARPA era la Oficina de Técnicas de Procesado de Información (IPTO *Information Processing Techniques Office*), que inició la investigación en ordenadores, aportando muchos avances que hoy en día están generalizados, desde las propias redes (como Internet) hasta la inteligencia artificial. En 1965, auspiciado por ARPA, se estableció una conexión informática entre un ordenador en el *Massachusetts Institute of Technology* (MIT) *Lincoln Laboratory* y otro situado en la empresa *System Development Corporation* (SDC, Santa Mónica, California), a una velocidad de 1.200 bps y con conexión bajo demanda (esto es, mediante un “módem con marcación”). Esta prueba de concepto permitió conocer las dificultades técnicas que se encontraban bajo este tipo de conexiones y la necesidad de diseñar protocolos específicos para resolver las problemáticas detectadas.

Coetáneamente a este experimento, también en el MIT, Kleinrock completaba en 1962 su tesis doctoral, en la que formulaba la teoría matemática de la redes de paquetes, la tecnología subyacente en Internet (y realmente, añadiríamos, en todas las actuales redes de comunicación de datos). En su tesis se desarrollaban los principales aspectos de esta tecnología: escalabilidad, prestaciones, diseño, control, enrutamiento, compartición de recursos, demanda y paquetización de mensajes.

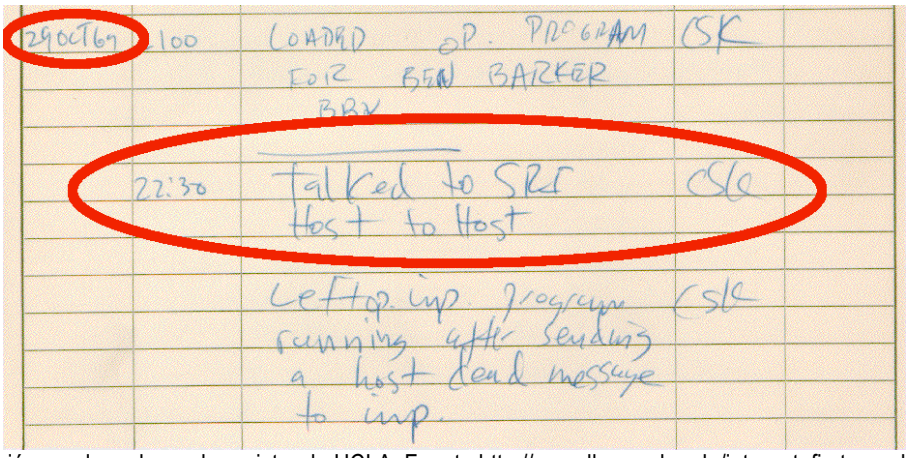
Por otro lado Paul Baran, en RAND Corporation (también en Santa Mónica, California), trabajaba en comunicaciones militares con el objetivo de diseñar redes de comunicaciones robustas, empleando redundancia y tecnología digital. En 1962 publicó un artículo con sus resultados e introducía también el aspecto de la paquetización de mensajes y el enrutamiento alternativo, aspecto este nuevo entonces [15].

En la parte europea, Donald Davies, del *National Physical Laboratory* (NPL) en el Reino Unido, comenzaba a analizar las redes de paquetes en 1965, y de hecho acuñaba el concepto de "paquete". Los trabajos de Baran y Davies estaban enfocados en aspectos de ingeniería y arquitectura de las redes, mientras que los trabajos de Kleinrock estaban dirigidos hacia los aspectos matemáticos de las redes.

Entre 1963 y 1964 las líneas de investigación del MIT y de ARPA se fusionan, cuando los investigadores de ambos proyectos son conscientes de la existencia del otro grupo. En abril de 1967 se organizó una reunión con el objetivo de diseñar ARPANET, que sería el embrión de la actual Internet. Basado en esa reunión se presentó un artículo al Simposium de la *Association for Computing Machinery* (ACM) de Principios de Sistemas Operativos [45].

El 3 de julio de 1969 la Universidad de California en Los Ángeles hacía público un comunicado de prensa anunciando la inminente puesta en marcha de ARPANET. Un paso importante en ese hito fue la conexión de los IMP (*Interface Message Processor*), algo así como las primeras tarjetas de red, a los ordenadores de UCLA y SRI (*Stanford Research Institute* –SRI- en Menlo Park, California), interconectados entre sí con el primer enlace a "alta velocidad" (sic) de Internet a 50 kb/s.

El 29 de octubre de 1969 a las 22:30 horas un programador llamado Charlie Kline y el propio Leonard Kleinrock se "logaron" en el equipo de SRI desde el de UCLA, en lo que se considera en primer mensaje de equipo a equipo enviado a través de Internet.



Anotación en el cuaderno de registro de UCLA. Fuente http://www.lk.cs.ucla.edu/internet_first_words.html

Desde ese 29 de octubre de 1969 hasta nuestros días, el crecimiento de Internet ha sido imparable [43]. Internet se ha vuelto omnipresente en nuestro mundo, es la tecnología dominante sobre la que comunicamos nuestras voces, nuestras palabras y nuestras imágenes [46]. A pesar de lo que muchos afirman [18], un tecnología con un

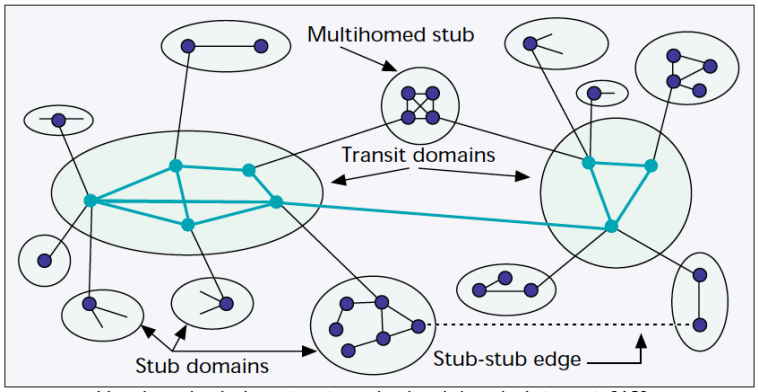
trasfondo militar, se “civiliza” y se populariza, en un fenómeno que tiene como secuela más cercana, la tecnología de posicionamiento global por satélite (GPS) [47].

1.5.2. FUNCIONAMIENTO DE LA RED INTERNET

Con objeto de comprender los escenarios que se expondrán posteriormente, es preciso esbozar brevemente cuál es el funcionamiento de la red Internet.

1.5.2.1. Arquitectura de la red Internet

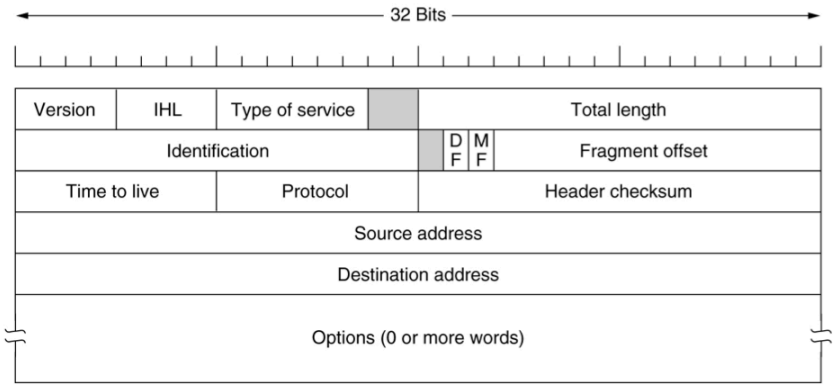
Como su propio nombre indica, la red Internet está compuesta por la interconexión de múltiples redes entre sí, facilitando la comunicación entre los equipos que las constituyen. Así [48], Internet puede ser vista como una colección de dominios interconectados. Cada dominio es un grupo de nodos (equipos), bajo una única administración técnica. Estos dominios pueden ser clasificados como dominios “extremo” (*stub*) o dominios de “tránsito” (*transit*). Un dominio “extremo” transporta solo tráfico que se origina o termina en el mismo dominio. Los dominios de “tránsito” no presentan esta restricción. El propósito de los dominios de tránsito es interconectar los dominios extremo de manera eficiente; sin ellos, cada par de dominios extremo necesitaría estar directamente conectados entre sí para poder intercambiar información. Los dominios de extremo corresponden a redes de campus u otras agrupaciones de redes de área local (LAN) interconectadas, mientras que los dominios de tránsito son casi siempre redes de área expandida (WAN) o metropolitanas (MAN). Un dominio de tránsito consiste en un conjunto de nodos troncales. En un dominio de tránsito cada nodo de troncal puede también conectar a varios dominios extremo, pero también pueden interconectar a otros dominios de tránsito. Los dominios extremo pueden clasificarse en multiconectados (*multihomed*), si están conectados a más de un dominio de tránsito, o simplemente conectados (*single-homed*) si solo están conectados a un dominio de tránsito. Por último, los dominios de tránsito puede estar a su vez organizados en jerarquías.



Un ejemplo de la estructura de dominios de Internet. [48]

1.5.2.2. El protocolo IP

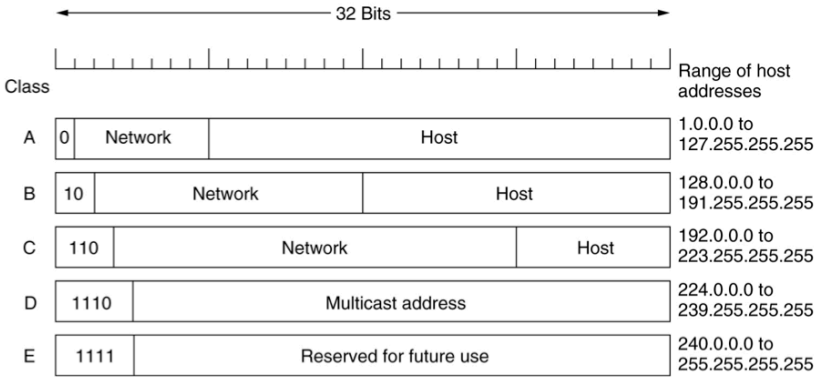
Un punto adecuado para comenzar la presentación del protocolo de Internet o protocolo IP (*Internet Protocol,* establecido en la RFC-791) es considerar el formato de los datagramas o paquetes (o tramas) IP. Un datagrama IP, como las cartas del correo, consiste en una parte de encabezado y una parte de texto. El encabezado tiene una parte fija de 20 bytes y una parte opcional de longitud variable. El formato del encabezado se muestra a continuación y contiene diversos campos con información sobre el propio datagrama. Especialmente importantes son los campos que contienen las direcciones de origen y de destino del datagrama.



Encabezado de IPv4. [19]

1.5.2.3. Direccionamiento en la red Internet

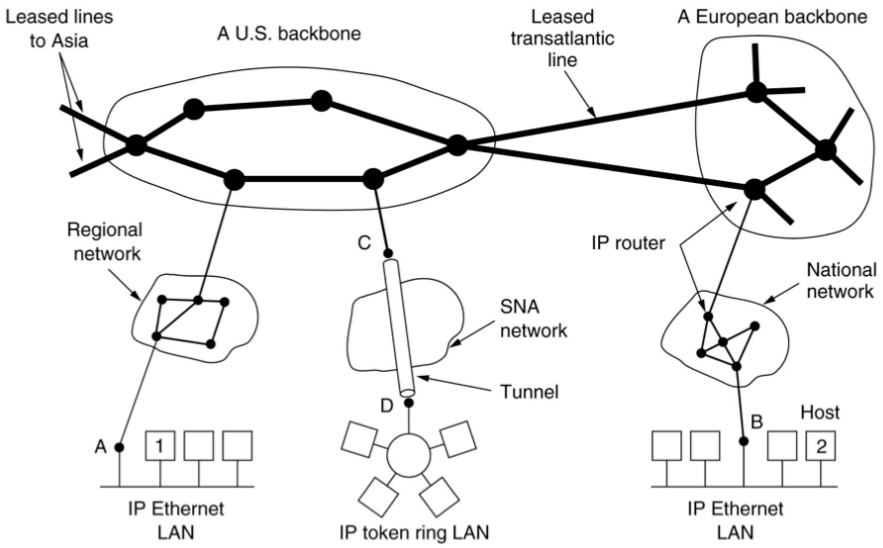
Cada nodo (*host*) y enrutador (*router*) en Internet tienen una dirección IP [19], que codifica su número de red y su número de nodo. Esta combinación es única: no hay dos equipos en la red Internet que tengan la misma dirección IP (excepto en el caso de redes con numeraciones “privadas” [49], aunque realmente esas direcciones no están propiamente en la red Internet). Todas las direcciones IP son de 32 bits (4 bytes) de longitud (en IPv4, versión actualmente en uso, en IPv6 la longitud es/será de 16 bytes) y se utilizan en los campos “Dirección Origen” y “Dirección Destino” de los paquetes IP. Las direcciones de red generalmente se escriben en notación decimal con puntos. En este formato, cada uno de 4 *bytes* que forman la dirección de 32 bits se escribe en decimal, de 0 a 255. Así, por ejemplo, una dirección IP válida sería la 192.168.1.1. La dirección IP menor sería la 0.0.0.0 y la mayor la 255.255.255.255. Estos dos valores extremos tienen significados especiales. La dirección 0.0.0.0 representa “esta red” o “este *host*” y solo se utiliza temporalmente durante la puesta en marcha de los equipos. La dirección 255.255.255.255 es la dirección de “difusión” y representa al conjunto de todos los host de la red indicada. Finalmente, para facilitar su manejo, el espacio completo de direcciones IP se dividió en 5 grupos, que contenían direcciones para otros tantos tipos de redes de diferentes tamaños.



Formatos de dirección IP y clases [19].

1.5.2.4. Enrutamiento de paquetes

La función principal de los enrutadores o *routers* en Internet consiste en reenviar paquetes con información a sus destinos finales [50]. Este es el procedimiento por el que se encaminan los paquetes con información a su destino a través de la red Internet, en los puntos de interconexión entre las diferentes redes y dominios. Para llevar a cabo esta función, un router debe decidir para cada paquete que recibe dónde debe enviarlo a continuación. Más exactamente, la decisión de envío consiste en encontrar la dirección del router del siguiente salto, así como el puerto de salida a través del cual se debe enviar el paquete. Esta información de reenvío se almacena en una tabla de reenvío que el router calcula sobre la base de la información recogida por los protocolos de enrutamiento. Para consultar la tabla de reenvío, el router utiliza la dirección de destino del paquete como una clave; esta operación se denomina búsqueda de direcciones. Una vez que se obtiene la dirección del nuevo destino, el router puede transferir el paquete desde el enlace entrante hacia el enlace saliente apropiado, en un proceso llamado conmutación.



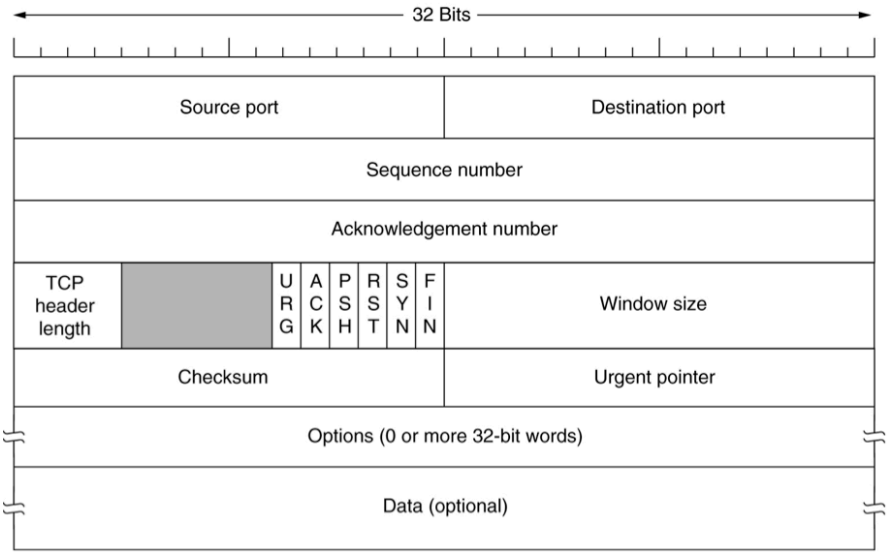
Internet es un conjunto interconectado de muchas redes. [19]

1.5.2.5. Los protocolos de transporte en Internet: TCP

Internet tiene dos protocolos principales en la capa de transporte, uno orientado a conexión y otro no. El protocolo no orientado a conexión es el denominado UDP, pero nos centraremos, por su relevancia en nuestro trabajo, en el protocolo TCP (establecido en la RFC-793) que sí es orientado a conexión. Los protocolos orientados a conexión precisan el establecimiento de una conexión "lógica" antes de poder enviar datagramas con información. Proporcionan una entrega de información en secuencia y confiable. De hecho TCP se diseñó específicamente para proporcionar un flujo de bytes confiable de extremo a extremo a través de una red no confiable.

El servicio TCP se obtiene al hacer que tanto el servidor como el cliente creen unos puntos de conexión terminales denominados *sockets*. Así, cada *socket* dispone de un número, que consiste en la dirección IP del *host* y un número que es local a ese *host* y que se denomina puerto (*port*). Un mismo *socket* puede utilizarse para múltiples conexiones simultáneas. Estas conexiones se identifican mediante los identificadores de *socket* de los dos extremos. Los números de puerto menores de 1024 se denominan "puertos bien conocidos" ("*well known ports"*) y son, efectivamente, ampliamente conocidos los servicios de red a que se refieren. Por ejemplo el puerto 23 corresponde al terminal (protocolo Telnet), el puerto 80 al protocolo http y el puerto 443 al SSL o https, estos dos últimos utilizados en la World Wide Web. Detrás de cada puerto hay una aplicación encargada de procesar las comunicaciones del correspondiente protocolo.

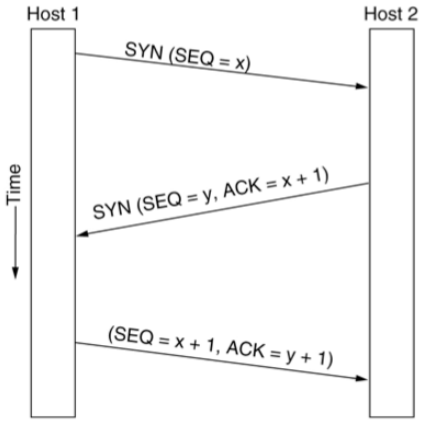
Como sucedía en IP, existe un encabezado que contiene la información específica del paquete TCP/IP.



Encabezado TCP. [19]

1.5.2.6. Establecimiento de una conexión TCP/IP

Veamos a continuación cómo se establecen las conexiones en TCP. Las conexiones TCP se establecen siguiendo un mecanismo de negociación en tres pasos denominado "*three way handshake*". Por un lado tenemos el servidor que está esperando pasivamente una conexión entrante (Host 2). En el otro extremo tenemos el cliente (Host 1) que desea establecer una conexión TCP con el servidor (Host 1). El cliente envía un segmento TCP con el bit SYN activado, un número de secuencia "x" y el bit ACK desactivado. Al llegar el segmento a su destino, el servidor revisa si hay algún proceso escuchando en el puerto indicado en el campo Puerto de destino. Si es así, devuelve al Host 1 un segmento confirmando de recepción con un nuevo número de secuencia "y" e incrementando en una la secuencia enviada por el servidor. Por ultimo el cliente confirma enviando un nuevo paquete incrementando la secuencia "y". Ambos nodos han establecido una conexión identificada por (x,y). Las aplicaciones situadas en ambos extremos deben reservar espacios de memoria para llevar a cabo las comunicaciones establecidas, por lo que si bien es posible que un puerto receptor mantenga varias comunicaciones simultáneas, su número no es ilimitado: un exceso de demanda puede saturar el servicio y bloquearlo. Esto, de ser una actuación premeditada, se conoce como un ataque de negación de servicio (denominado *DoS* por *Denial of Service*).



Caso normal de establecimiento de una conexión TCP. [19]

1.5.2.7. El Sistema de Nombres de Dominio (DNS)

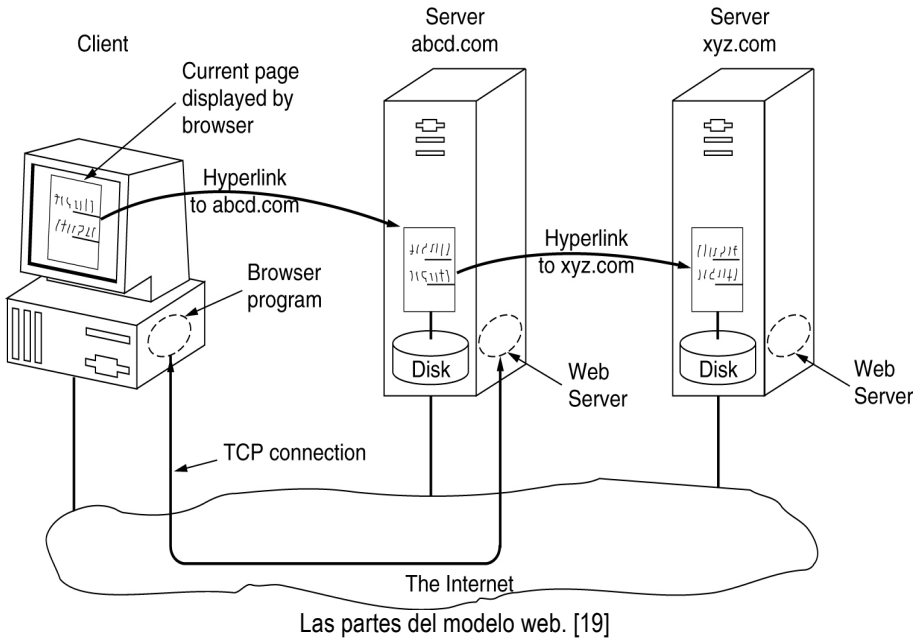
Cualquiera que haya utilizado alguna vez un navegador y lea el mecanismo de funcionamiento de Internet antes descrito, detectará de inmediato la existencia de un brecha: sólo hemos hablado de direcciones IP numéricas, y sin embargo lo más normal es utilizar direcciones (denominadas URL) que están en forma de texto. Esto

es así porque existe un servicio el Sistema de Nombres de Dominio [19] que se encarga de traducir las “direcciones de Internet” (www, correo, FTP, etc.) en sus correspondientes direcciones IP . No expondremos su funcionamiento ya que no es relevante para el desarrollo de la exposición objeto de este trabajo.

1.5.2.8. La World Wide Web (WWW)

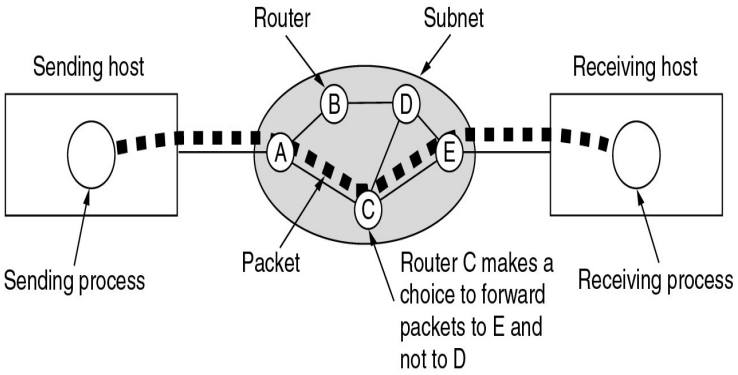
Aunque muchas veces se utiliza de manera análoga, Internet no es la World Wide Web. La World Wide Web, o abreviadamente la Web, es una aplicación de Internet que sirve para acceder a documentos vinculados entre sí y distribuidos por millones de servidores diferentes. La web comenzó en el Centro Europeo de Investigación Nuclear (CERN) como una necesidad para la colaboración entre grandes grupos de investigadores dispersos internacionalmente y que precisaban trabajar sobre un conjunto cambiante de informes, planos, dibujos, fotos y otros tipos de documentos [19].

La propuesta inicial [51] surgió del físico Tim Berners-Lee en marzo de 1989. El primer prototipo estaba operativo 18 meses después y en diciembre de 1991 se hizo una demostración pública en la conferencia Hypertext ’91 que tuvo lugar en San Antonio (Texas). El primer navegador se liberó en febrero de 1993 y en 1994 se fundo la World Wide Web Consortium, como organización dedicada al desarrollo de la web, su estandarización y el fomento de la interoperabilidad entre sitios.



Las partes del modelo web. [19]

El funcionamiento de la Web, utilizando los elementos sistémicos antes expuestos, sería el siguiente: Cuando en un navegador se introduce la dirección o URL de un servidor web, el equipo utiliza el servicio de DNS para convertir la dirección textual en una dirección IP (omitimos los pasos de más bajo nivel). Una vez obtenida la dirección IP el cliente solicita la apertura de una conexión al puerto 80 del servidor identificado por dicha IP, para ello utiliza el protocolo "*three way handshake*" antes descrito. Antes de confirmar la apertura de la conexión, el servidor reserva memoria con la que gestionar el intercambio de paquetes, y una vez responde a esta petición de apertura, con la tercera pata del procedimiento, comienza el intercambio de información. La descarga de una única página web (la principal o *home*, por ejemplo) puede llegar a efectuar la apertura de cientos de conexiones TCP cuasi-simultáneas.



Comunicaciones en la red Internet. [19]

1.5.3. ALGUNOS TIPOS DE ANOMALÍAS EN LA RED INTERNET

El funcionamiento "normal" de la red Internet puede verse alterado por múltiples tipos de incidentes. Lakhina en su artículo de 2004 [52] cita los siguientes:

Anomalía	Definición	Características	Ejemplos
Alpha	Inusualmente elevada tasa de transferencia punto a punto	Pico en el tráfico en B, P o BP, atribuible a un único par origen-destino. Corta duración (menos de 10 minutos) y limitado a un único flujo O-D	Experimentos de medida de ancho de banda.
DOS y DDOS	Ataque de negación de servicio (distribuido o no) contra una única víctima	Pico en el tráfico en P, F o FP, la fracción dominante del mismo dirigida a una única IP, con dirección IP de origen no dominante. Puede implicar múltiples fijos y típicamente de duración inferior a 20 minutos	Múltiples instancias donde un gran número de paquetes se envían a una única IP de destino que son blanco frecuentes de ataques DoS (por ejemplo el puerto 0)
FLASH CROWD	Inusual incremento de la demanda para un recurso/servicio	Pico en el tráfico en F o FP hacia una IP de destino predominante y hacia un puerto predominante. Típicamente de corta duración y limitado a un flujo único.	Múltiples instancias de un gran número de peticiones web a una única IP al puerto 80.

Anomalía	Definición	Características	Ejemplos
SCAN	Rastreo de un nodo buscando un puerto vulnerable (*port scan)* o rastreo de una red buscando un equipo vulnerable (*network scan*)	Pico en el tráfico F, con similar número de paquetes como flujos desde un origen predominante; no hay combinación predominante de IP destino y puerto. Pueden implicar múltiples flujos OD y típicamente dura menos de 10 minutos.	Rastreo de red buscando el puerto 139 (NetBIOS)
WORM	Código autopropagado que se difunde a través de la red aprovechando brechas de seguridad.	Punta en el tráfico F sin destino dominante y sólo un puerto dominante.	Flujos con puerto dominante 1433 asociado al gusano MS SQL-Snake
POINT TO MULTI-POINT	Distribución de un contenido desde un servidor a muchos usuarios.	Punta en tráfico P, B o BP desde una fuente predominante a numerosos destino, todos al un único puerto (y conocido)	Un único servidor difundiendo al puerto 119 (news nntp service) a un gran número de destinos.
OUTAGE	Incidentes que causan una disminución en el tráfico cursado en un par OD	Decremento en el tráfico BFP, usualmente a cero. Puede ser de larga duración (horas) y en todas las instancias, afectando a múltiples flujos OD	Mantenimientos programados.
INGRESS-SHIFT	Cliente cambia el tráfico de un punto de entrada a otro.	Disminución en el tráfico F para un flujo OD y una punta en el tráfico F en otro. Sin atributos dominantes. Implica múltiples flujos OD.	Cambios de ruta masivos por avería en la red de un operador.

Anomalías detectables en (B) Bytes, (P) Paquetes, (F) Flujos y combinaciones

En nuestra primera propuesta se pretende la detección de aquellas anomalías que impliquen una variación en el volumen de información transmitida, por lo que la gran mayoría de las enumeradas podrían estar incluidas en dicho grupo.

1.6. DISEÑO Y GESTIÓN DE REDES

1.6.1. UNA APROXIMACIÓN AL DISEÑO DE REDES DE COMUNICACIÓN

El diseño de redes de comunicación y de cualquier otro servicio o producto puede y debe estar centrado en la experiencia de usuario. Una posible aproximación es el diseño centrado en el usuario que asegura que este es considerado en el proceso de diseño de productos, sistemas y servicios [53]. El mayor beneficio de este planteamiento es que las tecnologías no son desarrolladas para su propio motivo, sino para satisfacer una necesidad de los usuarios o deseo, así se maximiza la probabilidad de que el nuevo desarrollo sea utilizado y las personas lo utilicen, para de esta manera aportar valor en el mercado.

El diseño centrado en el usuario es proceso de tres etapas incorporando las siguientes tareas:

1. Derivar los requerimientos a partir del análisis del usuario y el contexto de uso.
2. Seguir una aproximación estructurada de diseño incluyendo el prototipado testeado con los requerimientos del usuario.
3. Evaluar el diseño con los requerimientos.

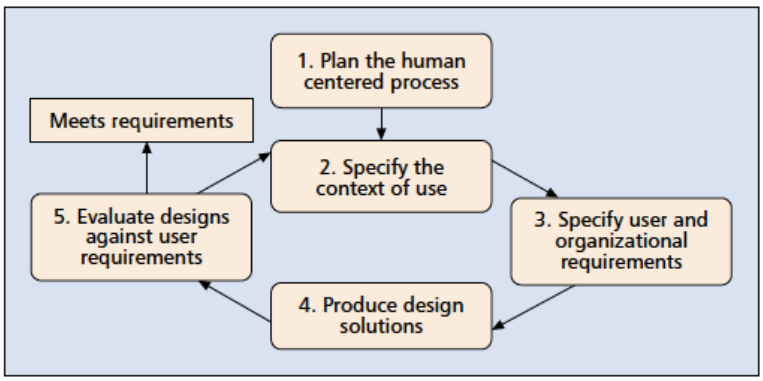
Obsérvese que todo el proceso de diseño es altamente interactivo e interrelacionado, por lo que tanto el análisis, diseño y evaluación puede modificarse conforme el proceso avanza, incardinándose incluso en fases posteriores de explotación de la red.

Durante la fase de análisis las características de los usuarios y tareas que realizan, y el entorno organizativo y físico definen el contexto en el que el sistema será diseñado. Las actividades de análisis estructuradas incluyen los siguientes pasos: análisis de los usuarios, análisis de las tareas, análisis del entorno y análisis comparativo.

Durante la fase de diseño un proceso de diseño estructurado soporta la orientación hacia el usuario del desarrollo del sistema. Existe un amplio conocimiento de información sobre la utilización del usuario, aquí se pueden utilizar requerimientos de los usuarios y datos de tráfico pasados así como estimaciones futuras.

Finalmente en la fase de evaluación el diseño de la solución se verifica para comprobar su correcto funcionamiento.

Existe una normativa, la UNE-ISO-13407, que materializa el diseño orientado al usuario [54]. Esta norma facilita un procedimiento para el diseño centrado en el usuario a través del ciclo de vida de sistemas informáticos interactivos. Es una herramienta para aquellos que gestionan procesos de diseño, y provee una guía sobre fuentes de información y normativa relevante a la aproximación de diseño centrado en el usuario. Describe el diseño centrado en el usuario como una actividad multidisciplinar, que incorpora factores humanos y ergonómicos y técnicas con el objetivo de mejorar la efectividad y eficacia, mejorando las condiciones de trabajo y contrarrestando posibles efectos adversos del uso sobre la salud, seguridad y desempeño.



Diseño centrado en el usuario de ISO-13407 [53].

Toda red de comunicación, considerada como un sistema en sí mismo, deberá estar caracterizada por los siguientes principios [55] que pueden entenderse como las condiciones de contorno cualitativas en el diseño y gestión de una red de datos:

- **Seguridad**: Bajo este concepto se incluyen aspectos relativos tanto a la seguridad física de los diferentes elementos que forman la red como a la información almacenada o circulante en la red.

- **Flexibilidad**: La red deberá ofrecer respuestas evolutivas rápidas ante las nuevas necesidades que demanden los usuarios con el paso del tiempo. En este aspecto la escalabilidad de las tecnologías utilizadas es fundamental.

- **Disponibilidad / Supervivencia / Fiabilidad**: Aspectos definidos en términos de capacidad para responder a las necesidades operativas en cualquier punto, en cualquier momento y en cualquier entorno de trabajo. Deberá responder a las necesidades operativas en cualquier ubicación física y en cualquier momento. Una gestión incorrecta de la red o poco atenta puede derivarse en frecuentes “caídas” de la red con las consiguientes pérdidas económicas. Será preciso un mantenimiento predictivo, esto es, analizar el funcionamiento de la red día a día para que sea posible adelantarse a la aparición de posibles incidencias en la red.

- **Interoperabilidad**: Entendida como la capacidad de los sistemas instalados para atender y solicitar peticiones desde y hacia otros sistemas cooperando entre ellos para alcanzar los objetivos propuestos.

- **Abordabilidad**: Deberá requerir inversiones económicas asumibles para su instalación y explotación. Esta restricción económica marcará la condición de contorno última para la selección de tecnologías, servicios y aplicaciones.

Las redes constituyen hoy en día una infraestructura crucial, interconectando empresas con el mundo y clientes, proveedores y colaboradores a través de una miríada de medios, dispositivos, aplicaciones, protocolos y herramientas [56]. Para conseguir que una red de comunicaciones posea las características anteriores se deberán superar obstáculos tanto económicos como técnicos y organizativos. Los problemas con las prestaciones de las redes y cortes eventuales ocasionan impactos muy costosos en empresas, por lo que la gestión de redes para evitar este tipo de problemas es una cuestión estratégica hoy en día [56].

El rápido incremento del tamaño de las redes de ordenadores ha conllevado la insuficiencia de obtener un diseño de la red que simplemente funcione, es necesario obtener un diseño óptimo que satisfaga las necesidad de ancho de banda con el mínimo coste posible. El diseño y gestión de redes se apoya en la siguientes áreas [57]:

- Gestión de dispositivos, detección de fallos, aislamiento de fallos, balanceo de carga y desplazamiento de tráfico de recursos sobrecargados.
- Predicción de degradaciones de servicio y sobrecarga.
- Planificación y localización de recursos en despliegue de redes.
- Planificación de mantenimiento y reparaciones.

No obstante la puesta en marcha de iniciativas encaminadas al establecimiento de marcos de operación de red a nivel corporativo encuentra problemas, en algunos casos, de difícil solución [57]:

- Las reglas teóricas relativas al diseño y gestión de redes son sólo mínimamente aplicables en la práctica dada la diversidad de escenarios posibles.

- No existe un procedimiento normalizado para el diseño y gestión de redes, el administrador de la red tiene (debe tener) la última palabra

El diseño de una red de telecomunicación implica numerosos aspectos [58]. Se deben de considerar aspectos tales como protocolos de red que se utilizarán (orientados a conexión u orientados a sesión, prioridad, enrutamiento de protocolos, arquitectura de aplicaciones, direccionamiento…), consideraciones de tiempo y retardo (tiempo de acceso, tiempo de respuesta, velocidad-vs.-ancho de banda, bloqueo-vs.-almacenamiento-vs.-encolado), conectividad (usuario-red y red-red, requerimiento

geográficos, estructura, infraestructura actual), disponibilidad, fiabilidad, soportabilidad, control del usuario, escalabilidad… Uno de los aspectos principales es la determinación del patrón que sigue el tráfico en la red, que puede extrapolarse a partir del comportamiento estadístico del usuario. Esto permite modelar tanto el tráfico de voz, como el de datos. La obtención de las prestaciones de que debe disponer la red para satisfacer los requerimientos de los usuarios es también una tarea necesaria. Este paso estará en función tanto de las prestaciones de la red, como de la conectividad. En general el diseño de una red se efectúa sobre los valores extremos más desfavorables a su funcionamiento, a través del calculo sobre los momentos temporales de mayor tráfico o de retardos más desfavorables, por ejemplo.

La naturaleza dinámica de las infraestructuras en términos de topología, carga y disponibilidad requiere que el modelo utilizado sea constantemente actualizado de acuerdo a la situación operativa y configuración actual de la red. Es también importante que el modelo refleje más que el simple estado y configuración de los dispositivos individuales y componentes, incluyendo relaciones y dependencias así como comportamientos de grupos de elementos [59]. La utilización de simuladores de redes permiten analizar las prestaciones de un diseño siendo posible obtener estimaciones sobre los principales indicadores [57]. Principalmente se deben analizar:

- **Dimensionamiento de la red** [60] : El dimensionamiento de la red proporciona la configuración a medio y largo plazo de los recursos de la red. El dimensionamiento no proporciona valores absolutos, sino en la forma de rangos a largo plazo. El objetivo es que los requerimientos futuros se encuentren en esos rangos. Los objetivos del diseño incluyen además evitar que partes de la red se encuentre sobrecargadas mientras que otras estén infrautilizadas y proveer el servicio al menor coste.

- **Gestión de rutas y recursos** [60]: La gestión de rutas suele encontrarse en la redes distribuida y operan los enrutadores. Es responsable de la gestión del proceso de enrutamiento de acuerdo con las reglas facilitadas por el dimensionamiento de la red. En las fases iniciales, el propio dimensionamiento de la red facilita la información de enrutamiento a partir de la configuración de los enlaces. Cuando el sistema está en operación, la información recogida permite realizar los cambios necesarios para mantener la situación operativa de la red en niveles óptimos.

Con el objetivo de soportar los procesos de negocio, los gestores de TI (tecnologías de la información) necesitan mantener y gestionar del conjunto de la infraestructura de TI basándose en el profundo conocimiento del despliegue actual y el estado operacional. Una mala comprensión de la situación de la red puede acarrear rápidamente en costes significativos en el nivel de negocio. Un método esencial para capturar el despliegue y su estado operativo es crear una descripción o representación de la infraestructura y su estatus. Esta descripción es típicamente denominada modelo [59].

Los procedimientos habitualmente utilizados para construir modelos de redes e infraestructuras de servicios han alcanzado hoy en día sus límites ante los constantes cambios del entorno. En particular, los despliegues flexibles de recursos virtuales en una única plataforma física han complicado esta tarea del modelado. También, el incremento de la movilidad y la heterogeneidad de recursos, tales como teléfonos inteligentes, ordenadores portátiles y tabletas, crean dificultades en mantener un modelo actualizado para nuestra red [59].

Además de un adecuado diseño, otra de las herramientas de posible utilización para conseguir que una red de comunicaciones disponga de los atributos de seguridad, flexibilidad, disponibilidad, interoperabilidad y abordabilidad, antes mencionados [55], sin caer en los errores que ocasionarían unos costes de operación innecesariamente altos, es la correcta administración o gestión de la misma.

1.6.2. INTRODUCCIÓN A LA GESTIÓN DE REDES DE COMUNICACIÓN

Decíamos anteriormente que la gestión de redes, sistemas y aplicaciones comprende todas las medidas preventivas y correctoras tomadas para garantizar un uso efectivo y eficiente de los recursos físicos y lógicos de los sistemas distribuidos y elementos de comunicación subyacentes que constituyen un sistema de información [6]. La gestión de redes es la suma de todas las actividades relacionadas con la configuración, control y monitorización de redes y sistemas, con el objetivo de asegurar su efectiva operación [61].

En todo tipo de redes debe coordinarse su crecimiento y organización impidiendo la proliferación caótica de elementos, arquitecturas y protocolos con el objetivo de garantizar un crecimiento sostenido y sostenible. La gestión de la red es un aspecto

fundamental para el buen funcionamiento de la misma, mediante la planificación, administración y mantenimiento de los elementos integrantes [6].

La importancia de la gestión y administración de redes es tan importante o más que el diseño en sí mismo de la red. La selección de protocolos de comunicaciones normalizados en la fase de diseño sólo garantiza algún grado de interoperabilidad e interconexión entre los sistemas que la integran, sin embargo el problema de la gestión de estos dispositivos y servicios no es resuelto en estas fases del proyecto. Es preciso que, con posterioridad, se redacte y asuma una política general que sirva de marco para controlar el comportamiento de los agentes que comprende una red corporativa, incluyendo usuarios, administradores, personal de soporte, aplicaciones ejecutándose en la red y los equipos físicos que la integran [6].

La complejidad de esta tarea reside en el hecho de que los componentes gestionados han evolucionado desde un conjunto de sistemas aislados, homogéneos y controlables dentro de un dominio de gestión, a un entorno de comunicaciones enorme, heterogéneo y distribuido a lo largo de múltiples dominios de gestión. Tenemos que trabajar con una variedad de componentes de red, múltiples dominios de gestión, entornos de servicios integrados y sistemas altamente heterogéneos. Teniendo que afrontar estos retos han emergido diferentes soluciones de gestión de redes a lo largo de los años [61].

Tan importante es el concepto de gestión que existen diversas iniciativas de normalización de las arquitecturas de gestión, entre ellas destacamos [62]:

- **TINA**: (Telecommunication Information Networking Architecture). Se trata de una arquitectura abierta para servicios de telecomunicación basados en computación distribuida. Intenta responder a preguntas tales como qué tipologías de servicios ofrecer y cómo y cuándo hacerlo.

- **OSI**: Dentro de la arquitectura para la interconexión de sistemas abiertos se ha desarrollado un esquema de gestión basada en seis pilares fundamentales: facturación de servicios, conexión, fallo, prestaciones, configuración y seguridad.

Un adecuado marco de administración de una red comprenderá una descripción general de los procedimientos de gestión de los elementos de las tecnologías de la información y las comunicaciones desde un punto de vista de proveedor de servicios

[63]. Esta política de administración incluirá los siguientes conceptos como áreas de gestión funcionales [64]–[67]:

- **Gestión de fallos** : La gestión de averías (o mantenimiento) es un conjunto de funciones que permite detectar, aislar y corregir un funcionamiento anormal de la red de telecomunicaciones y de su entorno. Proporciona facilidades para la realización de las fases de mantenimiento. Las mediciones de la protección de la calidad del servicio para la gestión de averías involucran mediciones de los componentes de fiabilidad, disponibilidad y supervivencia. Incluye tareas como el escalado de eventos y alarmas de la red, transmisión de las alertas de red a los administradores, procedimientos a emplear por el personal de soporte para subsanar las incidencias acaecidas.
- **Gestión de seguridad**: Las funciones correspondientes a este grupo permiten la gestión de la seguridad. Además, la gestión de la seguridad se necesita en todas las áreas funcionales de gestión y todas las transmisiones. La gestión de la seguridad figura como parte de la función de seguridad. La funcionalidad gestión de la seguridad comprende los servicios de seguridad de las comunicaciones y la detección y notificación de eventos de seguridad:
    - Los servicios de seguridad de las comunicaciones son los servicios de autenticación, control de acceso, confidencialidad de los datos, integridad de los datos y no rechazo que pueden prestarse en el curso de cualquier comunicación entre sistemas, entre clientes y sistemas y entre usuarios internos y sistemas. Se define además, un conjunto de mecanismos penetrantes de seguridad que son aplicables a cualquier comunicación, por ejemplo, la detección de eventos, la gestión de pistas de verificación de seguridad y la recuperación de la seguridad.
    - Mediante la detección y notificación de eventos de seguridad se comunica a las capas superiores de seguridad cualquier actividad que pudiera ser interpretada como una violación de la seguridad (por ejemplo, la actuación de un usuario no autorizado, la manipulación indebida del equipo, etc.).

  Se encargará, entre otras cuestiones de los derechos de acceso a la red y de los procedimientos de investigación en caso de penetración en el sistema.
- **Gestión de configuración**: La gestión de la configuración proporciona las funciones con las que ejercer el control sobre, identificar, recoger datos de, y suministrar datos a, los elementos de la red. La gestión de la configuración

comprende los siguientes grupos de conjuntos de funciones: planificación e ingeniería de la red; instalación; planificación y negociación de servicios; provisión; situación y control. Se encargará entre otras cuestiones de los procedimientos para provisión de servicios o configuración de dispositivos, cambios en procedimientos operativos.

- **Gestión de prestaciones**: La gestión de calidad de funcionamiento o de prestaciones proporciona funciones destinadas a evaluar el comportamiento de equipos de telecomunicación e informar al respecto, así como en relación con la efectividad de la red o del elemento de red. Su cometido consiste en reunir y analizar datos estadísticos para supervisar y corregir el comportamiento y la efectividad de la red, del elemento de red o del equipo de red, y facilitar la planificación, la provisión, el mantenimiento y la medición de la calidad. En este sentido, realiza la fase de medición de calidad de funcionamiento. Velará por lo tanto por el estado de "salud" de la red, la calidad de los servicios ofrecidos, los procedimientos para mejorar y/o garantizar una adecuada calidad en los servicios.
- **Gestión de facturación**: La gestión de la contabilidad permite la medición del uso de los servicios de red y la determinación del coste que representa para el proveedor de servicios, así como la cantidad que se ha de cobrar al cliente por el mencionado uso. Permite también la determinación de los precios de los servicios. Tendrá como tareas la valoración del coste de los distintos servicios de red y facturación a los usuarios de los recursos utilizados, entre otras.

Existen fundamentalmente dos protocolos normalizados para el acceso a la información de configuración de los dispositivos [19], [68]:

- **CMIP**: (Common Management Information Protocol). Es un estándar OSI (Open System Interconnection, modelo de interconexión de sistemas abiertos) que, al igual que ocurre con los protocolos OSI, es poco utilizado dada la popularidad de su competidor.

- **SNMP**: (Simple Network Management Protocol). Pertenece al conjunto de protocolos IP y es actualmente el más extendido.

Hoy en día la mejor elección como protocolo para gestionar una red es el SNMP. Este protocolo se basa en un arquitectura cliente/servidor en la que el cliente que accede a la información contenida en los dispositivos se denomina Gestor y los servidores se

denominan Agentes. Existen tres tipos de bases de datos cuya consulta es posible (dependiendo de los Agentes integrantes de la arquitectura). Estas bases de datos se denominan MIB (Management Information Base) y almacenan la información en forma estructurada [19], [68]:

- **MIB I**: Contiene información general del dispositivo. Todos los fabricantes comparten la definición de las tablas integrantes de esta base de datos.

- **MIB II**: Específica del dispositivo y/o fabricante. Es preciso que el programa de gestión conozca su definición para poder acceder a ella.

- **MIB RMON**: Contiene datos de monitorización remota de tráfico. Viene a solventar deficiencias de las MIBs anteriores para realizar estadísticas de tráfico.

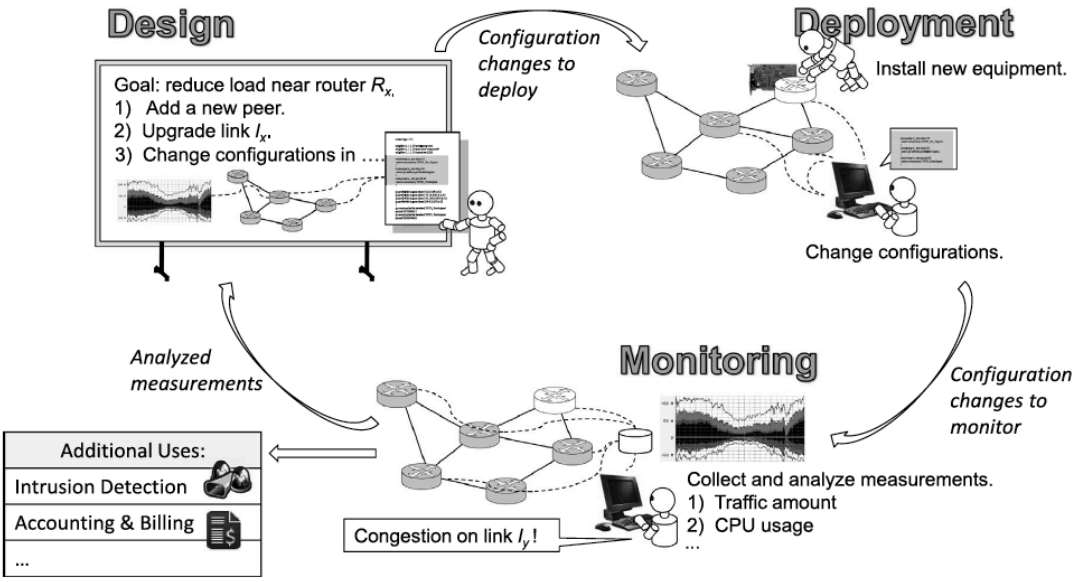
El último reto de la gestión de redes es la gestión de una manera segura del ancho de banda disponible, con un coste abordable, mientras gestiona elevadas cargas de trafico, esperando que la red este siempre operativa funciones de manera optimizada [56]. Puede ser necesario un conocimiento más profundo de la estructura física subyacente (incluyendo, por ejemplo, no solo los anchos de banda sino también las capacidades de los nodos) para evaluar la vulnerabilidad de la red a ataques, para la planificación de futuras expansiones de su capacidad y para reconsiderar la eficiencia de una infraestructura existente a la luz de los nuevos avances tecnológicos [69].

1.6.3. MONITORIZACIÓN DE REDES DE COMUNICACIÓN

El incremento del tamaño, complejidad y el carácter dinámico de las redes infraestructuras de servicios requieren funciones de gestión autónomas y adaptativas, así como superiores niveles de gestión de redes. Los sistemas tradicionales de monitorización no son ya capaces de ayudar adecuadamente a los administradores con tareas críticas en la gestión de fallos, de seguridad, de configuración y de prestaciones [59]. Como reacción, la tendencia es el desplazamiento desde la gestión individual de recursos hacia la gestión distribuida y autónoma que pueda ser enlazada a objetivos de nivel de negocio. [59].

La monitorización de redes es crucial para las operaciones de gestión de red y es usada para múltiples tareas críticas [70]. La función más importante de la monitorización de redes es la identificación temprana de tendencias y patrones en el tráfico de red y dispositivos. De acuerdo a estas medidas, los operadores de la red

comprenden la situación actual de la red y pueden reconfigurarla de manera que el estado de la red observado por los usuarios pueda mejorar. Debido a la significativa importancia de la monitorización de redes se ha dedicado un esfuerzo significativo para avanzar en su contenido. No obstante, los operadores de redes aún dedican la mayor parte de su trabajo a monitorizar y resolver problemas en sus redes. Los cortes de red continúan ocurriendo y evitan el acceso a las redes durante horas.



Clasificación de las operaciones de gestión de red en tres grupos: (i) monitorización del comportamiento de la red, (ii) diseño de los cambios de configuración de conformidad con los requerimiento, y (iii) despliegue de los cambios en la configuración y la infraestructura [70].

En el caso de Internet una comprensión detallada de las muchas facetas que componen su estructura a múltiples escalas es fundamental para muchos problemas de investigación. Entre ellos se incluyen la evaluación de prestaciones de los protocolos de red utilizados, la evaluación de la eficiencia y eficacia de las técnicas propuestas para proteger la red de ataques perniciosos e intrusiones, y el desarrollo de nuevos diseños para la provisión de recursos [69].

El análisis del tráfico es crucial para redes operativas. Tradicionalmente facilita información sobre posibles problemas y utilización real de los enlaces, protocolos, servidores, aplicaciones,… Además de los beneficios a nivel de los elementos que componen la infraestructura, los datos recopilados puede ser útil como base para el análisis de aplicaciones superiores y sobre los flujos de trabajo: la ejecución de casi cualquier aplicación de negocio deja una huella en el tráfico de red al implicar el acceso e intercomunicación entre recursos interconectados. Esta tendencia será aún

más fuerte conforme se incremente la utilización de aplicaciones compuestas a través de arquitecturas SOA (*Service-Oriented Architectures*) y la Web 2.0 [59].

1.6.4. INGENIERÍA DE TRÁFICO

La ingeniería de tráfico requiere de la observación del estado de la red a través de sistemas de monitorización para permitir la aplicación de actuaciones de control para mantener la red en la situación deseada [60]. La ingeniería de tráfico es un proceso iterativo y continuo de mejora de las prestaciones de la red. Los objetivos de optimización pueden cambiar a lo largo del tiempo conforme se impongan nuevas políticas de gestión y requerimientos a la red, por lo que los sistemas de monitorización deben ser suficientemente genéricos para poder enfrentarse a estos cambios [60]. Un sistema de monitorización de red debe facilitar información útil para las tres categorías siguientes de tareas [60]:

- Asistir a los procesos de ingeniería de tráfico para la toma de decisiones sobre provisión que optimicen la utilización de recursos de manera acorde con los cambios estimados a corto y medio plazo.

- Asistir a los procesos de ingeniería de tráfico proporcionando información de prestaciones y del tráfico analizado para la planificación a largo plazo de manera que optimice la utilización de la red y evite situaciones no deseadas.

- Verifique el cumplimento de los acuerdos de nivel de servicio (SLA: Service Level Agreements) establecidos para la operación de la red.

La ingeniería de tráfico dentro de un mismo dominio ha ganado en popularidad en los últimos años: buenas herramientas de ingeniería de tráfico permiten contribuir significativamente a la gestión y mantenimiento de prestaciones de grandes redes. Hay dos componentes importantes de la ingeniería de tráfico, la comprensión de los flujos de tráfico y la configuración (y diseño) de los protocolos de enrutamiento. Ambos componentes están relacionados, y es ampliamente aceptado que una buena comprensión de la matriz de tráfico y la dinámica de los flujos de tráfico pueden conducir a una mejor utilización de las capacidades de los enlaces a través de un mejor enrutamiento del tráfico [71]. No obstante, la medida y predicción de las demandas de tráfico son problemas engañosos. Las medidas de flujo están raramente disponibles en todos los enlaces y puntos de entrada/salida de la red, y es aún más complicado estimar los volúmenes de tráfico agregados para los flujos entre Orígenes y Destinos. Además, las demandas son variables en el tiempo, al menos en ciclos

diarios y con menor predictibilidad como resultado de eventos especiales o fallos internos o externos a la red [71]. Estos problemas están siendo abordados con modelos y herramientas de medida que permiten efectuar extrapolaciones y estimar demandas de tráfico [71]. No obstante, está ampliamente aceptado que es necesaria la comprensión de las demandas de tráfico para obtener una buena utilización de la red, pero esta creencia nunca ha sido adecuadamente cuantificada: ¿cuánto puede mejorar las prestaciones de una red un enrutamiento diseñado sin un conocimiento (o un conocimiento parcial) de la matriz de tráfico? ¿Con cuánta precisión necesitamos estimar la demanda de tráfico para garantizar una buena utilización de la red? Cuando las demandas de tráfico varían, ¿qué rango de cambio es tolerable para garantizar las prestaciones de la red? ¿Cómo opera un enrutamiento que ha sido diseñado óptimamente para una matriz de tráfico concreta, cuando esta matriz cambia? [71].

Varios tipos de análisis toman los datos medidos como entrada. Los campos más importantes para análisis son los de mapas topológicos, detección de aplicaciones, descubrimiento de relaciones, estimación de prestaciones y análisis de características [59]. La extensa cantidad de datos capturados en redes corporativas o de proveedores requieren capacidades de agregación y almacenamiento que, por ejemplo, reducen la fiabilidad de los datos a lo largo del tiempo. Se precisan algoritmos eficientes para obtener diversos estadísticos. Tales algoritmos tratan habitualmente medidas de tráfico en línea como flujos de datos y procesan la información conforme les llegan. Sin embargo, las grandes redes presentan distintos flujos de datos que requieren el almacenamiento de grandes cantidades de información que consumen significativos recursos de memoria. [59].

La información sobre las relaciones de tráfico directas, en el sentido de que una fuente envía paquetes a un destino, sobre un determinado protocolo, se analiza mediante inspección directa de los paquetes y mediciones sobre flujos. En muchos casos la información sobre tráfico cursado revela información sobre el rol de los dos sistemas implicados. Así un sistema final puede ser identificado como un servidor a partir del patrón de tráfico cursado [59].

Otra importante función en la gestión de redes es la monitorización y determinación de un mapa de red. La topología de interés en esencia un grafo en el cual los nodos representan los enrutadores y las aristas representan los enlaces entre dos enrutadores [59].

La caracterización del tráfico de Internet se ha convertido en los pasados años en uno de los mayores retos en las redes de telecomunicación [72]. El profundo conocimiento de la dinámica y la composición del tráfico es esencial en la gestión y supervisión de las redes.

Dado que Internet es una colección de miles de redes más pequeñas, cada una bajo su propio control administrativo, no hay un único punto en el que se pueda obtener una visión completa de su topología. Además, el temor de los operadores a perder ventajas competitivas frente a otros operadores les ha llevado a no compartir información sobre la topología de sus redes. Así, dado que la inspección directa de la red es generalmente inviable, la tarea de "descubrir" la topología de Internet ha sido campo de estudio de investigadores que han intentado inferir su estructura utilizando métodos empíricos y teóricos [69].

Algunos autores [59] abogan por la evolución de las aproximaciones tradicionales de monitorización hacia otros que consistan en mediciones de tráfico distribuidas y técnicas de monitorización; capacidades dinámicas para el almacenamiento, recolección y medida de tráfico; así como análisis de tráfico para inferencia y deducción.

Nuevos avances en monitorización de tráfico, medida y análisis juegan un papel importante en la construcción y mantenimiento automático de un modelo que capture la infraestructura de los recursos y servicios, así como sus restricciones y utilización dinámica. Estos nuevos métodos varían desde descubrimientos activos y pasivos, medidas de tráfico, detección de aplicaciones y servicios a sofisticadas técnicas de agregación, análisis y almacenamiento de información de tráfico medido [59].

1.6.5. RETOS EN LA GESTIÓN DE REDES DE COMUNICACIÓN

Los sistemas de gestión de redes reúnen grandes cantidades de datos que deben ser agregados, filtrados y visualizados con el objetivo de hacer que la información comprensible sea fácilmente accesible a los operadores humanos [73].

Aunque el análisis de datos y su visualización sea un concepto temporalmente maduro de la gestión de redes, parece que las técnicas e interfaces disponibles no satisfacen del todo a los operadores, por las siguientes razones [73]:

- Las visualizaciones tradicionales, especialmente aquellas basadas en mapas geográficos, no permiten un buen escalado conforme se incrementa la

información a representar. Es más, si hay implicadas una multitud de capas diferentes e intentamos una visualización de la topología en muchas o en todas las capas de forma simultánea, el problema de escalabilidad solo empeorará.

- La información recogida y sus estadísticas son visualizadas habitualmente en una forma casi estática. Habitualmente no hay herramientas, y si las hay están muy limitadas, que permitan explorar los datos aplicando filtros, ampliando zonas o correlando información, de una manera interactiva.

- La visualización de tráfico típicamente se enfoca en la visualización de elevados volúmenes de componentes de tráfico o flujos. Aunque esto es ciertamente útil para planificación y quizá para objetivos de facturación, hay también una creciente necesidad de extraer y remarcar tráfico inusual y patrones extraños. Especialmente para propósitos de seguridad es a veces mucho más deseable descubrir y localizar pequeños volúmenes de tráfico altamente inusuales.

- Muchas de las herramientas disponibles están diseñadas para análisis y visualización fuera de línea. Sin embargo hay una creciente necesidad de análisis y visualización en línea y casi en tiempo rea, para reducir los tiempo de detección y reacción. Con velocidades en las redes del orden de decenas de gigabits por segundo, este reto es cualquier cosa menos trivial. La tendencia hacia la captura estadística de datos para redes de alta velocidad también requiere la visualización de la precisión de los datos.

Se han realizado investigaciones básicas de técnicas de visualización como parte del proyecto CAIDA (*Cooperative Association for Internet Data Analysis*). El proyecto CAIDA ha desarrollado varias técnicas de visualización de topología, algunas basadas en mapas geográficos y otras en representaciones más abstractas [73]. La investigación en técnicas de visualización tridimensionales ha sido también objeto de algunos proyectos. Muchas de estas visualizaciones son espectaculares, si bien su usabilidad nunca ha sido bien valorada, y la creación de dichas imágenes requieren herramientas de hardware y software muy sofisticadas. Así pues estas técnicas no son ampliamente accesibles y utilizables. Dados los recientes avances en capacidades gráficas de los equipos informáticos y los nuevos formatos interactivos, existen oportunidades para que herramientas de visualización multidimensionales sean desarrolladas en el próximo futuro. Para los mapas geográficos, es previsible que

Google Earth actúe como un catalizador para el desarrollo de nuevas técnicas en la que las capacidades de visualización y exploración al vuelo sean más ampliamente accesibles [73].

Kind *et al* plantean también algunos retos en la gestión de redes [59]. Uno de los problemas que mencionan es el volumen de información a procesar y el ancho de banda que consume el envío de la información sobre tráfico. El segundo aspecto pendiente de investigación, más interesante para nosotros, es la extracción de información de interés de la medidas de tráfico. Por ejemplo, los administradores de las redes desean conocer cuánto ancho de banda consumen las diferentes aplicaciones. La respuesta a esta pregunta requiere modelar el comportamiento de las aplicaciones, analizar los patrones de las aplicaciones e identificar tales patrones en los datos de tráfico. Se necesita inferencia y algoritmos de minería de datos para algunos problemas similares, tales como, detección de ataques, mapeado de dependencias de servidores e identificación de problemas de configuración de la red. Un problema importante aquí es que los patrones de tráfico cursado en la red cambian con el tiempo, especialmente cuanto entran en juego nuevas aplicaciones o servicios.

Lee *et al* [70] insisten en que uno de los principales problemas de la monitorización de redes es la gestión de los grandes volúmenes de información almacenada que debe ser analizada y visualizada. Dado que las mediciones son continuamente recogidas, el tamaño de la información recopilada aumenta rápidamente. Este volumen de información necesita ser almacenada, analizada y presentada al operador de la red. Así pues, los operadores de la red precisan una metodología de monitorización que reduzca el tamaño de los datos acumulados sin reducir la precisión del análisis. Además de los requerimiento de mediciones en tiempo real, los operadores retienen el histórico de anteriores mediciones para análisis futuros. Estos análisis posteriores implican tendencias a largo plazo para planificación de la red y análisis detallados de utilización para facturación y propósitos legales.

El modo más habitual de reducir la cantidad de información a almacenar y analizar es la consolidación [70]. La consolidación gradualmente reduce la resolución de los datos de medidas pasadas consolidando los datos con el uso de diferentes funciones (por ejemplo la media, el máximo, o el total). Esto también descarta medidas que sean anteriores a un determinado momento temporal. Para poder utilizar la consolidación los operadores deben definir una función de consolidación y un intervalo temporal en función de la precisión que necesiten para análisis. La agregación es otro modo habitual de reducir la cantidad de datos a ser analizados y presentados. Por ejemplo,

estadísticas de múltiples flujos pueden ser agregados cuando estos flujos tengan como destino la misma red. Nótese que las mediciones de flujos pueden ser agregadas conforme a múltiples dimensiones, tales como, dirección IP de origen, dirección IP de destino, número de puerto, protocolo y tiempo. Los analistas serán los responsables de elegir la dimensión más apropiada para sus fines.

Algunos estudios recientes han adoptado una aproximación diferente [70]. En lugar de reducir el tamaño de los datos, han optado por incrementar la velocidad del análisis de estos grandes volúmenes de datos utilizando sistemas en paralelo y plataformas de computación distribuida. Estos análisis son llevados a cabo utilizando tecnología de Big Data tales como, MapReduce y Hadoop, además de Apache S4 o Storm [70].

2. OBJETIVOS

2. OBJETIVOS

Se plantean los siguientes objetivos:

1. Analizar el papel de las modernas técnicas de “Visual Analytics” en la inspección del tráfico de redes de comunicación.

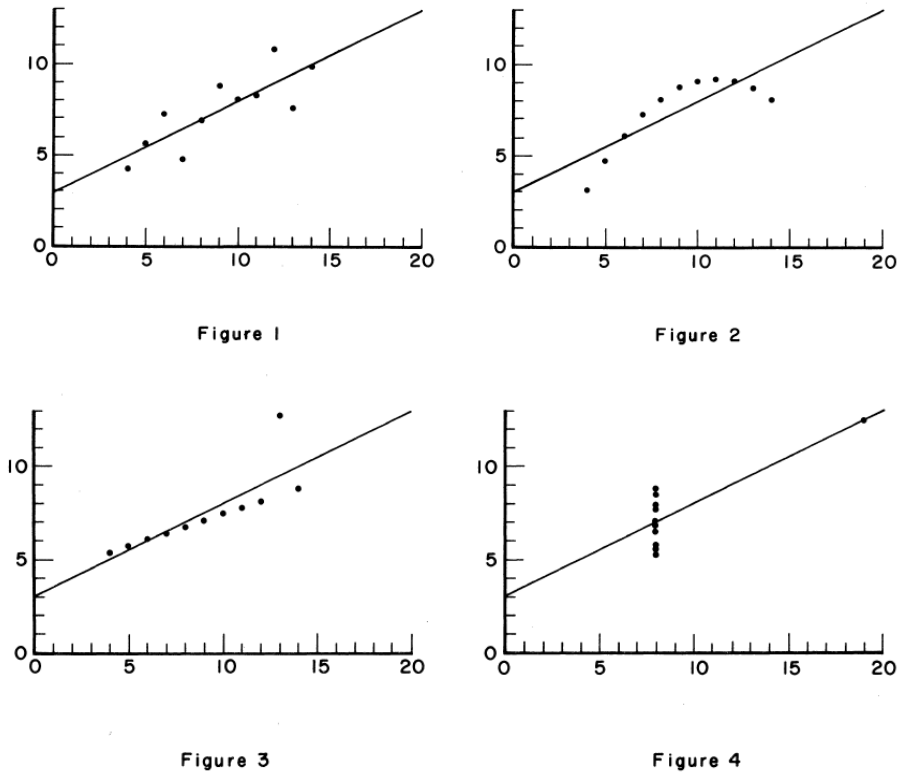
2. Proponer una metodología multivariante que nos permita detectar anomalías de volumen y diagnosis específica de las anomalías detectadas.

3. Investigar las posibles relaciones existentes entre la topología de una red de comunicación y la HJ-bigeometría.

3. TÉCNICAS DE VISUALIZACIÓN DE DATOS DE REDES DE TELECOMUNICACIÓN

3. TÉCNICAS DE VISUALIZACIÓN DE DATOS DE REDES DE TELECOMUNICACIÓN

Sería inconcebible escribir una sola línea sobre bibliografía de visualización de datos en cualquier escenario imaginable sin referenciar el que podría considerarse el primer libro sobre el tema. Si bien se puede situar un abordaje a las técnicas de visualización en el trabajo de Tukey de 1977 [74], en el año 1983 Tufte publica la primera edición de su libro "The visual display of quantitative information" que vino seguida de una segunda edición en el año 2001 [75], [76]. El libro es un compendio de lo que posteriormente se ha denominado Analítica Visual o *Visual Analytics*. Incluye pormenorizadamente los aspectos que deben ser tenidos en cuenta a la hora de representar visualmente datos cuantitativos. Afirma el autor que los gráficos delatan los datos, que verdaderamente las representaciones visuales pueden ser más precisas y reveladoras que los cálculos estadísticos convencionales. Pone como ejemplo ilustrativo de su afirmación el conocido como cuarteto de Anscombe [77], que presenta cuatro conjuntos de datos con idéntico modelo lineal y iguales estadísticos descriptivos básicos, pero cuyas representaciones visuales son totalmente diferentes.



Representaciones visuales y modelo lineal asociado al cuarteto de Anscombe [77]

Expone Tufte que la excelencia en los gráficos estadísticos consiste en la comunicación de ideas complejas con claridad, precisión y eficiencia. Las representaciones gráficas deber:

- Mostrar los datos
- Inducir al observador a pensar sobre la sustancia, en lugar de sobre la metodología, el diseño gráfico, la tecnología de producción de los gráficos, o cualquier otra cuestión.
- Evitar distorsionar lo que los datos tienen que decir
- Presentar muchos números en un pequeño espacio
- Hacer coherentes grandes conjuntos de datos
- Facilitar al ojo la comparación de diferentes subconjuntos de datos
- Revelar los datos a diferentes niveles de detalles, desde una amplia visión de conjunto a la estructura más fina
- Servir a un razonable propósito claro: descripción, exploración, tabulación o decoración
- Estar integrado con la descripción estadística y verbal del conjunto de datos.

Tufte también expone que los gráficos pueden distorsionar los datos que representan, haciendo difícil para el observador descubrir la verdad. Pero las representaciones gráficas, continua, no son diferentes de las palabras a este respecto, cualquier método de comunicación puede ser utilizado para engañar: no hay razones para creer que los gráficos sean especialmente vulnerables a la explotación por mentirosos. Referencia Tufte los trabajos de Tukey en los años 60, que convirtieron las representaciones gráficas estadísticas en algo respetable, poniendo fin a la idea de que los gráficos estaban solo para decorar unos pocos números.

Enumera Tufte lo que a su juicio son los principios de la excelencia de los gráficos:

- La excelencia en los gráficos es una presentación bien diseñada de datos interesantes, una cuestión de fondo, de estadística y de diseño.

- La excelencia en los gráficos consiste en idea complejas comunicadas con claridad, precisión, y eficiencia.

- La excelencia gráfica es lo que da al observador el mayor número de ideas en el menor tiempo y con la menor tinta en el espacio más pequeño.

- La excelencia gráfica es casi siempre multivariante.

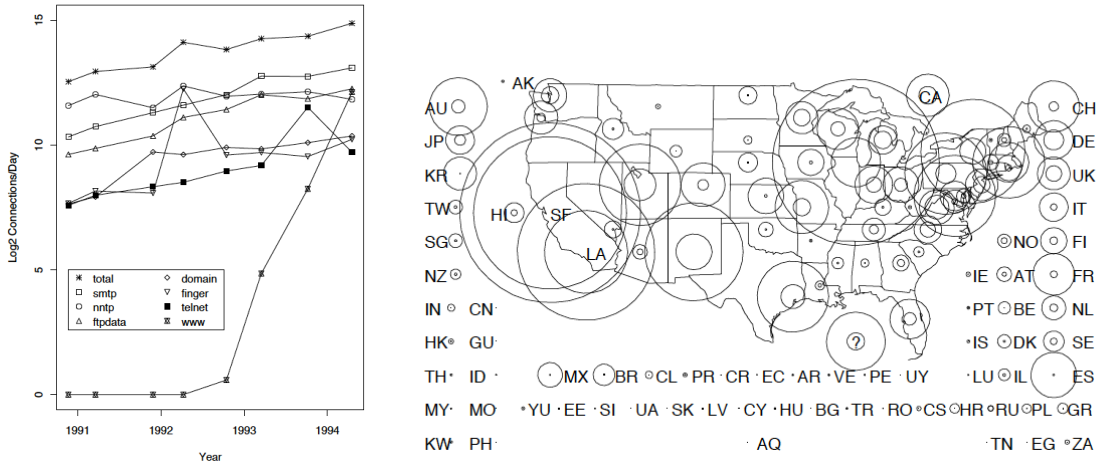
- Y la excelencia gráfica requiere decir la verdad sobre los datos.

En el año 1990 encontramos una de las primeras referencias, sino la primera, que relaciona los conceptos "visualización", "redes" y "gráficos". Becker *et al* [78] presentan en la primera conferencia del IEEE sobre visualización su propuesta "*Dynamic Graphics for Network Visualization*". Los autores ya afirmaban hace más de dos décadas que, junto con el explosivo crecimiento de las redes de ordenadores en aquellos años, se había producido un crecimiento análogo en el volumen de los datos medidos en las redes, tales como capacidad, flujo, bloqueos y retados. El análisis de todos esos datos, afirmaban, era complicado debido a la necesidad de incorporar en el análisis la frecuentemente enmarañada estructura de la red. Se había prestado poca atención a la representación gráfica de las redes, más allá de la visualización de la información topológica. Atendiendo a la creciente importancia de la gestión de redes y su análisis, se precisaban nuevas herramientas de visualización. Es necesario indicar, por nuestra parte y por colocar en contexto las afirmaciones de los autores, que en 1990 Internet era prácticamente embrionaria y las redes de área local comenzaban a extenderse por el mundo. Así pues, si en 1990, con un escenario muy limitado, se realizan esas afirmaciones, hoy en día la justificación de análisis y métodos de representación más avanzados esta fuera de toda duda.

Becker *et al* ya establecen en su artículo algunos precedentes importantes. Especificaban que los datos en las redes pueden dividirse en dos grupos, los asociados a los enlaces y los asociados a los nodos. Igualmente afirmaban que a veces es deseable estudiar la evolución de un nodo o un enlace a través del tiempo, por lo que planteaban la utilización de gráficos dinámicos. Estos gráficos permitían comprender patrones de utilización en grandes volúmenes de datos y también la rápida detección de patrones inusuales o eventos anómalos.

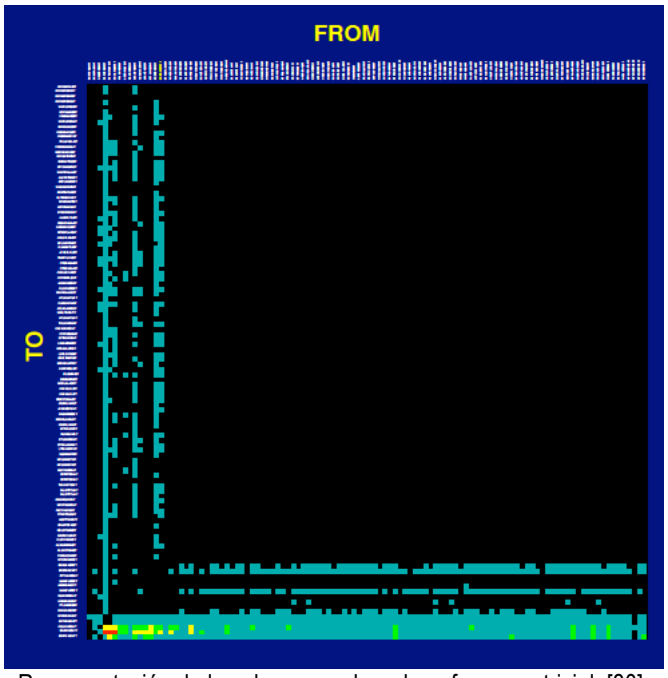
En 1994 Paxson publica un artículo [79] sobre la tendencia en el crecimiento de la utilización del protocolo TCP en redes WAN. El artículo analiza la evolución del tráfico entre el Lawrence Berkeley Laboratory – Universidad de California (LBL-UCBerkeley) y el resto del mundo. Es un estudio que comprende varias anualidades y expone en realidad el crecimiento del tráfico de Internet. Presenta numerosos gráficos mostrando

diferentes aspectos evolutivos. Se trata de gráficos de dispersión o de líneas, que ya permiten intuir algunas posibles aplicaciones de la visualización de datos de redes.



Gráficos de tasas de conexiones diarias por protocolo y distribución geográfica [79].

En 1995, de nuevo Becker *et al* [80] publican en el primer volumen del *IEEE Transactions on Visualization and Computer Graphics* un artículo sobre la visualización de datos de redes. Este artículo tuvo un antecedente en una presentación en el año 1993 [81]. Si analizamos las referencias del mismo se podría de nuevo afirmar que se trata de una de las primeras publicaciones sobre el tema. Una vez más en el artículo dividen los datos generados por las redes en aquellos asociados a los nodos y los asociados a los enlaces. Además estos datos pueden ser categóricos, como el tipo de enlace o nodo, o cuantitativos, como por ejemplo la capacidad del los enlaces. Y por último, estos datos pueden ser estáticos o dinámicos. El objetivo que se plantean los autores para la utilización de técnicas gráficas es comprender los datos multivariantes asociados a las redes, y no las redes en si mismas. Así, la estructura y conectividad presenta una importancia secundaria, lo que se quiere analizar son los datos asociados a los enlaces y nodos. En algunas de las propuestas que presentan se detectan problemáticas que serán arrastradas en el tiempo, tales como representaciones desordenadas y visualmente confusas. Es especialmente interesante en el entorno de esta investigación la representación de la carga de una red en forma de matriz, ya que esta visualización funciona mucho mejor que la tradicional representación de nodos y enlaces cuando hay muchos elementos en el gráfico.



Representación de la sobrecarga de red, en forma matricial. [80]

Los autores identifican igualmente diversos parámetros de interés en las representaciones, tales como:

- Estadístico de la red elegido para ser representado.
- Nivel de detalle (capilaridad) de los elementos representados.
- Representación geográfica / topológica de la red.
- Referencias de tiempo consideradas en la representación.
- Posible agregación del estadístico sobre los diferentes niveles de detalle.
- Tamaño del gráfico.
- Colores utilizados.

En 1997 aparece en un documento cuyo objetivo es planificar la puesta en marcha de la nueva generación de Internet [82] y en el que se formula la obligación de disponer de herramientas para recopilar y analizar información relativa a ingeniería y gestión de Internet. Se concreta esta carencia aún más, definiendo la necesidad de herramientas de monitorización y análisis para todos los niveles de protocolos y para todas las velocidades de transmisión, además de la exigencia de una visualización a gran tamaño de la información, mostrando la desviación sobre los modelos planificados y la eficiencia de las comunicaciones.

En 1999 la organización CAIDA (*The Cooperative Association for Internet Data Analysis*) albergó una reunión del grupo de trabajo ISMA (*Internet Statistics and*

Metrics Analysis) sobre Visualización en redes [83]. En la reunión participaron proveedores de acceso a Internet, vendedores de equipamiento y, por supuesto, investigadores. Esta mezcolanza permitió obtener importantes conclusiones sobre un tema que les interesaba a todos ellos, con sus diferentes puntos de vista. Así, los proveedores de acceso a Internet reclamaban herramientas de visualización que incluyesen:

- Representaciones simples de la actividad de tráfico, en forma de diagramas de dispersión en función del tiempo.
- Enlaces a otros tipos de datos, informes o análisis, combinando, por ejemplo, diversas fuentes y que proporcionasen visualizaciones que permitiesen mejorar significativamente la calidad y celeridad para resolver problemas, para ello era necesario:
    - Proporcionar correlaciones entre múltiples tipos de datos.
    - Agregar información sobre prestaciones para los clientes.

De hecho, cada grupo de participantes identificaron varios tipos de datos que les ayudarían en la realización de sus diferentes funciones:

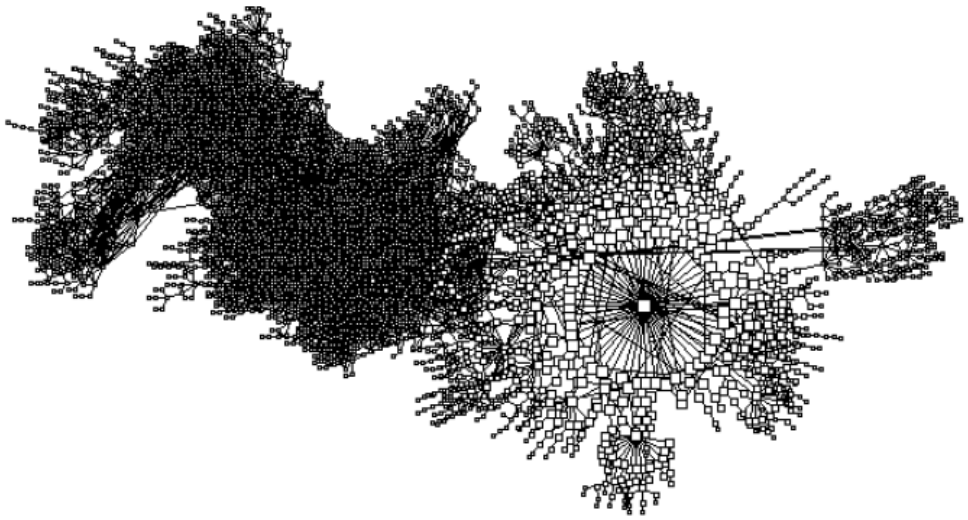
- Centros de operación de red: Visualizaciones en tiempo real de la red que les ayuden en la identificación de fallos y problemas.
- Ingenieros: Acceso a visualizaciones de ingeniería que incluyesen cuestiones tales como prestaciones, enrutamiento o topologías lógicas.
- Gestores: Situaciones actuales y tendencias futuras con un nivel de agregación suficiente para operar como resúmenes ejecutivos.
- Planificadores de redes: Tráficos cursados con datos de origen y destino, que incluyesen naturaleza del tráfico, tipos de protocolos, herramientas de modelado y simulación, especialmente si son capaces de incluir datos de tráfico reales.
- Investigadores: requerimientos similares a los ingenieros y planificadores.
- Clientes: Prestaciones de la red, especialmente latencias y pérdidas de paquetes, que les permitan disponer de información sobre la calidad de servicio de la red.
- Empresas de servicios de soporte: localizaciones geográficas de enlaces.
- Administradores de sistemas y redes: infraestructuras físicas y lógicas, localizaciones y relaciones entre los diferentes elementos que las forman.

- Reguladores: los organismos reguladores pueden beneficiarse de las herramientas visuales para detectar interacciones entre intereses competitivos en el sector.

En concreto todos ellos detectan, entre otros, los siguientes aspectos como merecedores de atención en el campo de la visualización:

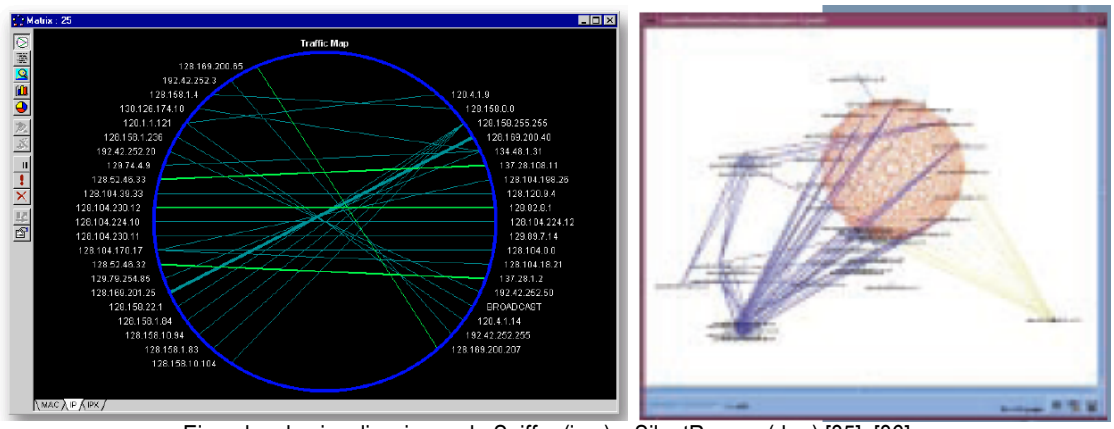
- Mejoras en técnicas automáticas de clustering y agregación, en función de los elementos que forman la red, lo que es especialmente importante conforme los tamaños de la redes se incrementan para hacer las representaciones más eficaces.
- Estudios sobre cómo la visualización de datos se aplica en otros sectores y su posible aplicación a Internet.
- Desarrollo de nuevas herramientas de visualización, para representar información sobre prestaciones de tráfico (particularmente latencia y pérdidas de paquetes), localización y priorización de elementos críticos de la red y conectividad de la red.
- Mejores métodos de monitorización, caracterización y visualización de patrones de trafico reales.

Ese mismo año 1999 Huffaker [84], también miembro de CAIDA, presentó la herramienta Otter que permite la visualización de diferentes tipos de datos asociados a nodos, enlaces o rutas. Por sus funciones la herramienta presentaba usos en una amplia variedad de aplicaciones: topología, carga, prestaciones, enrutamiento. Permitía trabajar con “grandes volúmenes de datos” y obtenía representaciones eficaces.



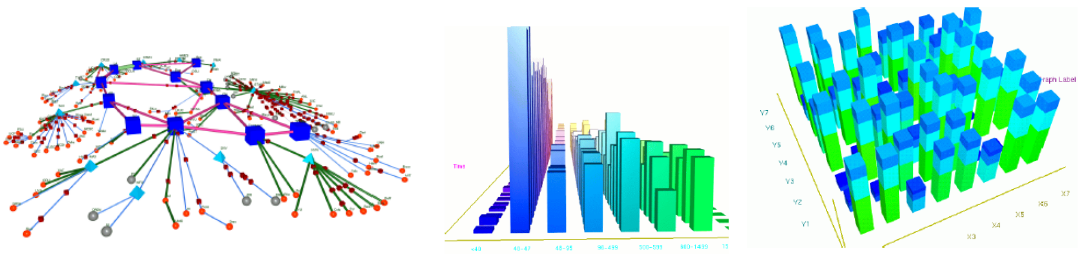
Representación de Otter para 30.000 nodos. [84]

También en 1999 comienzan a aparecer aplicaciones informáticas de gestión y monitorización de redes, como Sniffer [85] (de Computer Associates) y SilentRunner [86] (de Raytheon) que disponen ya de capacidades muy útiles de representación de información del tráfico de la red.



Ejemplos de visualizaciones de Sniffer (izq.) y SilentRunner (der.) [85], [86].

Un año después, en el 2000, Brown *et al* [87] presentan en el *Passive and Active Measurement Workshop* (PAM2000) unos atractivos gráficos obtenidos con la herramienta *Cichlid,* gráficos que, entre otras cuestiones, aprovechaban la mejora de los ordenadores en la obtención de representaciones gráficas de calidad. Los autores justifican la herramienta, como en otros casos anteriores, en la necesidad de utilizar herramientas de visualización para analizar los grandes volúmenes de datos obtenidos de las redes de comunicación. *Cichlid* es una herramienta de visualización que genera representaciones animadas tridimensionales de diversos tipos de datos posibilitando incluso interactuar con las representaciones en tiempo real. Las funciones de *Cichlid* incluyen la posibilidad de interactuar con el gráfico para obtener información del mismo, etiquetado y coloreado condicional de los gráficos.



Diferentes tipos de gráficos generados con Cichlid. [87]

Cichlid soportaba dos tipos de gráficos: gráficos de barras tridimensionales, útiles para representar cantidades numéricas que sean función de dos variables independientes, y gráficos de vértices/bordes, para representar topologías. Los gráficos de barras pueden utilizarse para representar, en una forma parecida a la expuesta por Becker en

1995 [80], retardos entre nodos, distribuciones de tráfico cursado entre orígenes y destinos, por ejemplo. Y, como decíamos, estas representación podían ser animadas para visualizar la evolución en el tiempo de los diferentes datos recopilados.

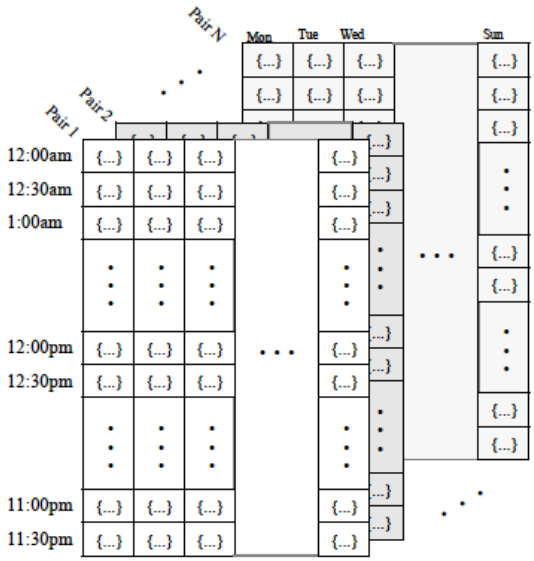
En ese mismo año 2000 otro grupo de trabajo entre empresas denominado XIWT (*Cross-Industry Working Team*) publica el documento "*Internet Service Performance: Data Analysis and Visualization*" [88]. En él identifican tres objetivos prioritarios del análisis y representación de datos en redes:

1. Establecimiento de perfiles basales, para determinar la elasticidad de las infraestructuras para absorber nuevas aplicaciones y servicios. Sin comprender las prestaciones de la red en el pasado, los riesgos de invertir en nuevos desarrollos y despliegues no pueden ser adecuadamente gestionados.

2. Detección de anomalías, para conocer si la infraestructura está alcanzado los requerimientos de prestaciones y fiabilidad requeridas por los servicios y aplicaciones.

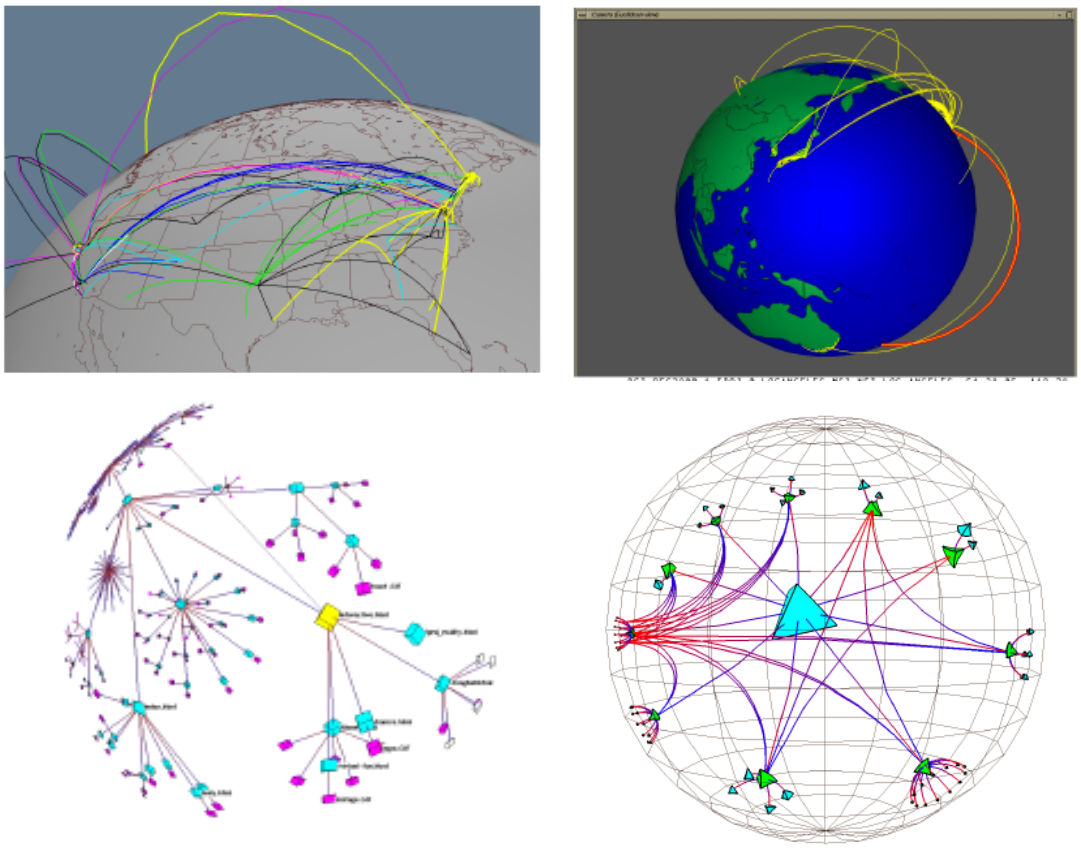
3. Identificación de tendencias, que ayuden a predecir las prestaciones futuras de las infraestructuras y a planificar proactivamente las capacidades de las redes.

En los aspecto más concretos de visualización de datos, la representación de datos de prestaciones y sus perfiles basales asociados con esas métricas permiten extraer informaciones útiles de los grandes juegos de datos que suelen considerarse para estos cometidos. Estas métricas pueden ser, de manera general, dividas en unidimensionales, tales como retardo o pérdidas, y bidimensionales, tales como la disponibilidad, que es una función de dos métricas unidimensionales (retardo y pérdidas). Evidentemente la dimensión de la métrica condiciona el abordaje de su visualización. Para datos unidimensionales proponen la utilización box-plots, mientras que para bidimensionales utilizan clásicos gráficos de dispersión (métrica1-vs.-métrica2, o métrica-vs.-tiempo). No obstante, quizá lo más interesante del documento sea su apéndice B en el que se formulan diversas alternativas conceptuales para el tratamiento de los datos multidimensionales de las redes en forma de matricial.



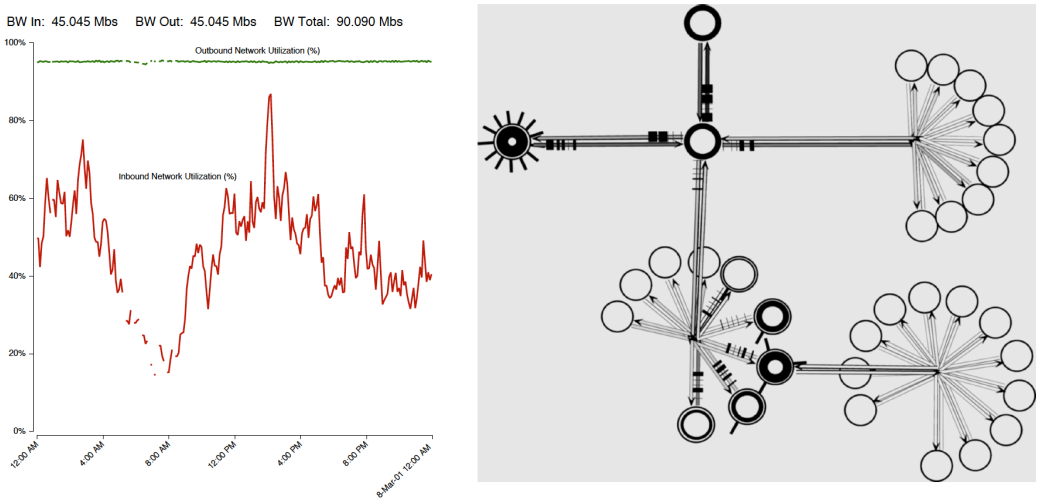
Conceptualización de conjuntos de datos de redes. [88]

También en el año 2000 Munzner [89] presenta su tesis doctoral sobre visualización de grandes gráficos y redes. Munzner fue, además, una de las participantes en el workshop organizado por CAIDA en 1999 [83] comentado anteriormente. Presenta en su tesis unos espectaculares gráficos que, debido a la técnica de proyección que utiliza, permiten representar con un aceptable nivel de eficacia gráficos con un gran número de elementos y con resultados altamente expresivos.



Algunas de las representaciones gráficas propuestas por Munzner en 2000. [89]

En 2001 Erbacher presenta una ponencia sobre monitorización visual del tráfico [90], si bien los modelos gráficos obtenidos son relativamente básicos, ya que se trata de gráficos de utilización de red en función del tiempo y grafos de red con información de tráfico.



Gráficos para la monitorización visual de redes propuestos por Erbacher en 2001. [90]

En 2002 Nyarko [91] *et al* presentan un analizador visual para detección de intrusos que incluye integración sensorial (*haptic integration*). Tal y como afirman los autores, la detección de intrusos requiere el análisis de grandes cantidades de datos a través del tiempo. Además, tras esa recopilación masiva, los datos deben ser clasificados, analizados y representados de forma que se separen los eventos sospechosos o interesantes de aquellos que constituyan la actividad normal de la red. Esta ardua tarea, afirman los autores, solo utiliza técnicas de visualización rudimentarias. La visualización de información consiste en la transformación de datos de entrada en representaciones gráficas como salida y es un potente nexo entre los dos sistemas de procesado de información dominantes: la mente humana y los ordenadores. La visualización en la detección de intrusos se presta muy bien a la utilización de técnicas de animación. Las capacidades de representación bidimensional o tridimensional próximas a tiempo real que posibilitan los modernos ordenadores personales y estaciones de trabajo abren el camino a nuevos y excitantes métodos de visualización para las actividades en las redes. Por otro lado la integración sensorial (*haptic*) se refiere al proceso por el que se perciben objetos virtuales a través de interfaces entre ordenadores y humanos que permiten tocar y manipular objetos conceptuales generados por el ordenador.

En 2002 Keim publica un artículo genérico sobre la minería de datos visual y la visualización de información [92]. En él defiende el análisis de datos visual como una ayuda a la integración del "factor humano" en el proceso de exploración de datos, aplicando las habilidades sensoriales humanas al análisis de grandes conjuntos de datos disponibles en los sistemas informáticos actuales. Las técnicas de minería de datos visuales han demostrado, afirma, un elevado valor en el análisis exploratorio de datos y tienen un potencial para explorar grandes conjuntos de datos. Expone como principales ventajas del análisis exploratorio visual frente a las técnicas de minería de datos automáticas las siguientes:

- Las técnicas visuales pueden fácilmente tratar con datos no homogéneos y con presencia de ruido.
- Las técnicas visuales son intuitivas y no requieren la comprensión de complejos procedimientos o conceptos estadísticos o matemáticos.

El análisis exploratorio visual de datos usualmente sigue un proceso en tres etapas: en primer lugar obtener una visión del conjunto, luego ampliar y filtrar, finalmente obtener detalles, en lo que se ha denominado el "Mantra de la búsqueda de información".

Keim establece diversas clasificaciones de técnicas de minería de datos visual.

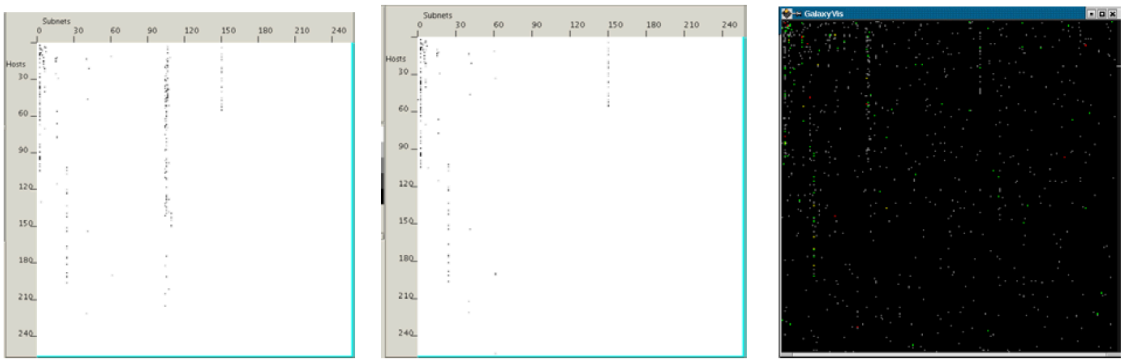
1. En función del tipo de datos a ser visualizado:
    a. Datos unidimensionales, tales como series temporales.
    b. Datos bidimensionales, tales como mapas geolocalizados.
    c. Datos multidimensionales, tales como tablas relacionales.
    d. Información textual, tales como artículos o páginas web.
    e. Grafos y jerarquías, tales como tráfico o estructuras de webs.
    f. Algoritmos y programas, tales como operaciones de depuración.
2. En función de la técnica de visualización utilizada:
    a. Representaciones bidimensionales y tridimensionales, tales como gráficos de barras y de dispersión.
    b. Representaciones transformadas, tales como coordenadas paralelas.
    c. Representaciones basadas en iconos, tales como las caras de Chernoff.
    d. Representaciones de alta densidad, tales como las técnicas de patrones recursivos y de segmentos circulares.

e. Representaciones apiladas, tales como los dendogramas.

3. En función de la interacción y la técnica de distorsión utilizada:
    a. Proyecciones interactivas.
    b. Filtrado interactivo.
    c. Ampliación interactiva.
    d. Distorsión interactiva.
    e. Enlazado y ajuste interactivo.

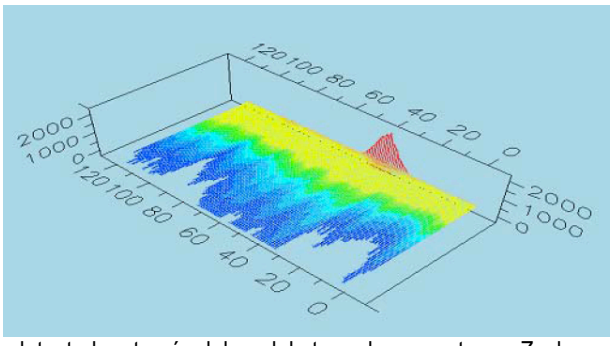
En 2003 Wood *et al* publican un extenso informe [93] que forma parte de la documentación para la certificación como analista de intrusiones dentro del grupo de certificaciones globales de aseguramiento de la información promovidas por el SANS Institute (www.sans.org). Reafirman los autores que la idea principal que reside tras la visualización de información en el análisis de tráfico, es que los datos puedan ser presentados al usuario en un formato que esté optimizado para una fácil comprensión, y que permita identificar de manera más fácil el tráfico y posibles patrones anómalos. Afirmación ésta que extrae de otro documento de SANS Institute elaborado por Sheffler [94].

También en 2003 Yurcik *et al* presentan un prototipo de herramienta para la detección visual de intrusos en el tráfico de redes [95]. La primera acción que plantean para "asegurar" una red es conocerla (*Know thy network!*, ordenan). La ignorancia, afirman, es felicidad, pero también es muy arriesgada. El conocimiento de la situación en seguridad informática ha evolucionado desde ¿hay algún problema? hasta ¿dónde está el problema? y por último ¿cuál es el problema?. No obstante las herramientas no facilitan la visión de redes de cierto tamaño en su conjunto. Los requerimientos para estas técnicas, propugnan, serán dos: monitorizar la red al completo y monitorizarla continuamente. Plantean una herramienta para visualizar el tráfico en nodos, subredes y a través de un subespacio completo de direcciones de red.



Tráfico entrante, saliente y diferencia, representaciones propuestas por Yurcik *et al* [95]

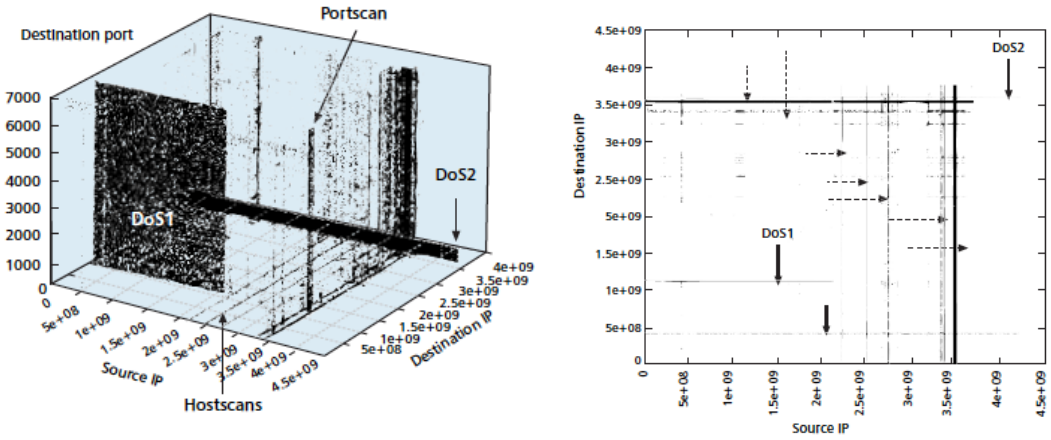
En 2004 Zachary *et al* presentan una metodología para la monitorización de redes y detección de anomalías en tiempo real [96]. Afirman los autores que el desarrollo de una infraestructura informática en red que sea confiable y segura depende de la monitorización en tiempo real y de la detección de eventos anómalos en ella. En el núcleo de su propuesta se encuentra la idea de modelar la red apoyándose en la idea básica en física estadística de caracterizar un sistema dinámico complejo a través de su Hamiltonianio. Así, la analogía del Hamiltoniano en el espacio de estados de la red (para ser más precisos se trataría de la inversa del Hamiltoniano) sería el conjunto de reglas que gobernarían la dinámica de intercambio de conversaciones en la red. Es más, descriptores habituales termoinformativos, tales como la entropía, energía y temperatura pueden ser directamente derivados desde ese modelo termodinámico. El modelo obtenido es similar al concepto termodinámico de las curvas de probabilidad de Boltzmann. Bajo esta analogía, el área bajo la curva de estado es una representación proporcional a la temperatura. La pendiente de la curva de estado es una representación proporcional a la entropía. Fluctuaciones macroscópicas de la energía son representadas como cambios entre intervalos temporales, que corresponderían con reacciones endotérmicas (absorciones de calor) o exotérmicas (emisiones de calor). Los histogramas obtenidos a partir de intervalos temporales contiguos se combinan para formar una variedad ("manifold") termal. Bajo condiciones normales de la red, esta variedad termal se caracteriza por el equilibrio, y las fluctuaciones alrededor del equilibrio se establecen por las condiciones de contorno. Las anomalías se representan como perturbaciones abruptas en la fluctuación alrededor del equilibrio. Los sistemas más actuales de detección de anomalías proveen mecanismos para trazar un conjunto eventos desencadenantes de una alerta. Esta capacidad es importante para análisis forenses posteriores al evento y para la valoración de mejoras en el proceso de detección.



Anomalía detectada a través del modelo termal propuesto por Zachary *et al*. [96]

En 2004 Kim *et al* [97] publican un artículo sobre la visualización en tiempo real de ataque en enlaces de alta velocidad. La visualización de ataques en curso es, afirman

los autores, algo de lo que los administradores de redes pueden hacer uso antes de proceder a análisis más complicados del tráfico, ya que proporciona un percepción rápida de pistas indiciarias de que algo va mal en la red. Esta rápida detección de actividades sospechosas que permitan disponer de una alerta temprana es una actividad tan esencial como difícil de llevar a cabo. Los autores obtienen representaciones tridimensionales similares a las de Becker [80] y Brown [87].



Representación tridimensional y proyección en el plano O-D. [97]

En 2005 Goodall presenta un artículo especificando los que en su opinión son los requerimientos de una herramienta de detección de intrusos basada en visualización [98]. Afirma el autor que un sistema de detección de intrusos puede disparar miles de alarmas al día, de las cuales el 99% serán falsos positivos. Discriminar entre el elevado número de falsos positivos aquellos debidos a actividad maliciosa real es una tarea que recaerá en el conocimiento y experiencia del personal técnico. Aunque se busca un sistema totalmente automático de detección de intrusos, una aproximación más realista es implicar al personal técnico en el diagnóstico. Si bien es cierto que los ordenadores son capaces de analizar elevados volúmenes de datos de bajo nivel, no pueden siquiera igualar las capacidades de análisis humanas. El situar al personal técnico en el "bucle de diagnóstico" se puede llevar a cabo a través de sistemas de visualización que utilicen gráficos informáticos para mejorar la comprensión aprovechando las capacidades de percepción humanas. Las herramientas de visualización constituyen sistemas para ofrecer a los analistas un conocimiento de la situación con una visión de conjunto de una red al completo en una única pantalla. Las representaciones de enlaces y nodos son útiles para visualizar pequeñas cantidades de datos sobre el tráfico en la red, mientras que las coordenadas paralelas, por ejemplo, permiten analizar relaciones entre datos multidimensionales de redes. Las tareas técnicas a llevar a cabo para la detección de intrusos se desarrollan en tres

fases: monitorización, análisis y respuesta. La visualización puede jugar un importante papel en las dos primera fases:

- En la etapa de monitorización, la visualización permite a los técnicos procesar eficazmente los datos de manera semi-atendida, esto es, sin necesidad de prestar sobre ellos una especial atención.

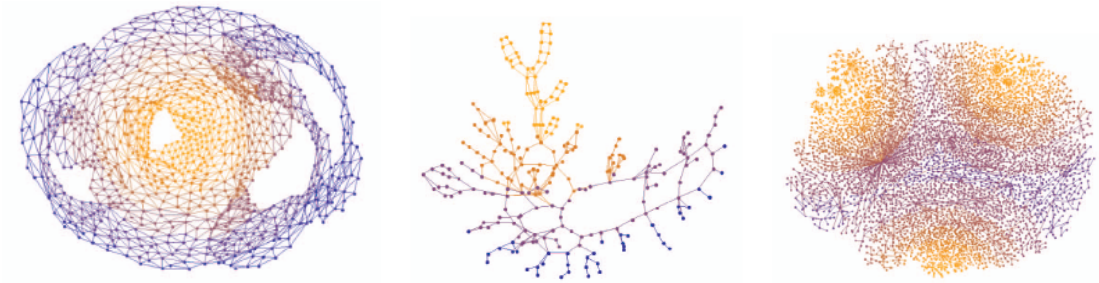
- En la etapa de análisis, sin embargo, se debe enfatizar la completitud y precisión frente a la rapidez del procesado humano. Las técnicas de visualización aquí deben soportar análisis multidimensionales y presentar la información desde varios puntos de vista. Igualmente deben facilitar la correlación de múltiples fuentes de datos en una única representación.

Incide el autor en la necesidad de que la herramienta de visualización sea flexible y capaz de adaptarse a cada red, la que deberá, coincide en esto con otros autores, ser perfectamente conocida por el analista. La visualización igualmente deberá ser capaz de facilitar la diagnosis y no solo la detección: no es suficiente con proveer al técnico de las habilidades para fácilmente reconocer una anomalía, debe ser capaz de identificar la razón de dicha anomalía con objeto de construir un diagnóstico preciso.

También en 2005 Van Wijk presenta un artículo [99] en el que defiende el valor de las técnicas de visualización. La visualización de datos hace posible para investigadores, analistas, ingenieros y audiencia general obtener información de esos datos, de una manera eficiente y efectiva, gracias a las capacidades de sistema visual humano, lo que permite detectar rápidamente interesantes características y patrones. El concepto de visualización, afirma, es un término ambiguo. Puede referirse a una disciplina de investigación, a una tecnología, a un procedimiento especifico, o propiamente al resultado visual. Si la visualización se considerada como una tecnología, esto es, como una colección de métodos, técnicas y herramientas desarrolladas y aplicadas para satisfacer una necesidad, entonces es posible aplicar medidas habituales: la visualización tiene que ser eficaz y eficiente. En otras palabras, la visualización debe hacer lo que se supone tiene que hacer, y tiene que hacerlo utilizando una mínima cantidad de recursos. Una de las inmediatas y obvias implicaciones de esto es que no podemos juzgar la visualización aisladamente, sino que tenemos que considerar el contexto en el que es usada: deberemos considerar su coste y los beneficios de la misma.

En 2005 Abdullah *et al* [100] presentan otra propuesta para la aplicación de herramientas de visualización en la detección de intrusos. Si bien las herramientas que utilizan son simples gráficos de barras, introducen en su artículo un importante elemento nuevo a considerar: aunque podría desprenderse típicamente un mayor interés en las visualizaciones de datos de tráfico en tiempo real, hay un campo en el que la rapidez en pasar de la captura a la visualización no es tan apremiante, los análisis forenses. Los análisis forenses de un ataque suelen ser estudios más dilatados en el tiempo que los propios ataques (segundos frente a días), y aunque permiten métodos de análisis menos exigentes en cuanto a prestaciones, el objetivo de los mismos suele ser igualmente la reducción de los tiempos necesarios para la obtención de resultados.

En 2005 Gansner *et al* publican un artículo [101] sobre visualización de redes, en este caso centrado en topologías grandes, con unas representaciones visualmente próximas a las obtenidas por Munzner [89].



Visualizaciones de ojo de pez para topologías grandes propuestas por Gansner *et al*. [101]

Erbacher *et al* presentan también en 2005 un artículo en el que exponen otros métodos gráficos para detección de intrusos en redes. En su introducción exponen las capacidades que las técnicas de visualización deben tener, en su opinión, para operar con las especificaciones propias de sistemas de detección de intrusos:

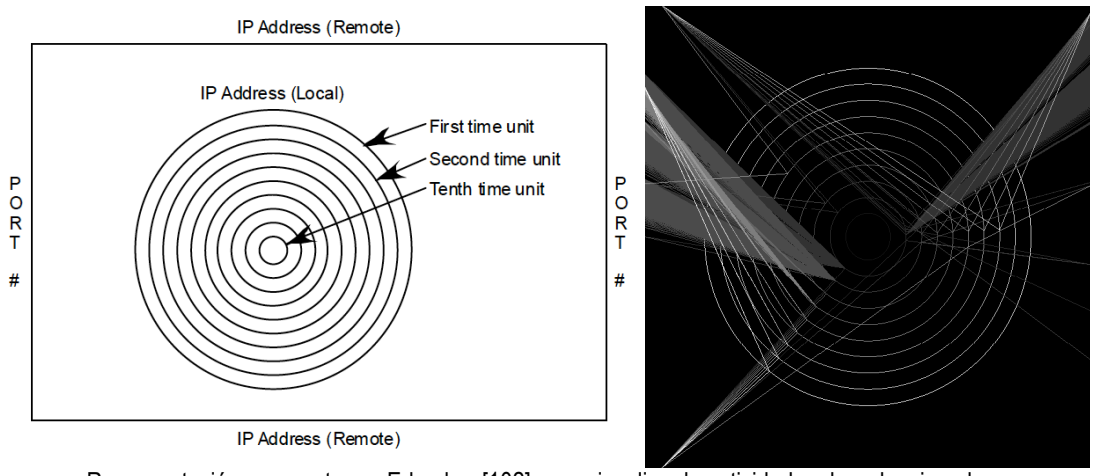
- Escalabilidad: El análisis de datos para funciones de IDS (*Intrusion Detection System*, sistema de deteccción de intrusos) requiere la intervención de enormes volúmenes de datos. Esto es particularmente cierto con ataques sofisticados en los que deben ser analizados y correlados datos recopilados durante días o semanas.

- Elevada dimensionalidad: Los datos relativos a IDS presentan una elevada dimensionalidad, pero no solo eso, datos únicos concretos presentan *per se* ya una elevada dimensionalidad.

- Complejidad y correlaciones: Los datos recopilados no pueden ni deben ser considerados de manera aislada. Es la correlación entre eventos a lo largo del tiempo lo que permite detectar ataques.

- Temporalidad: Los datos recopilados son temporales por naturaleza y deben ser consideradas largos periodos para posibilitar la detección de todos los tipos de ataques posibles. Esto es especialmente cierto con sofisticados ataques que intentan evitar la detección utilizando técnicas de sigilo (ataques de baja intensidad y lentitud en ejecución).

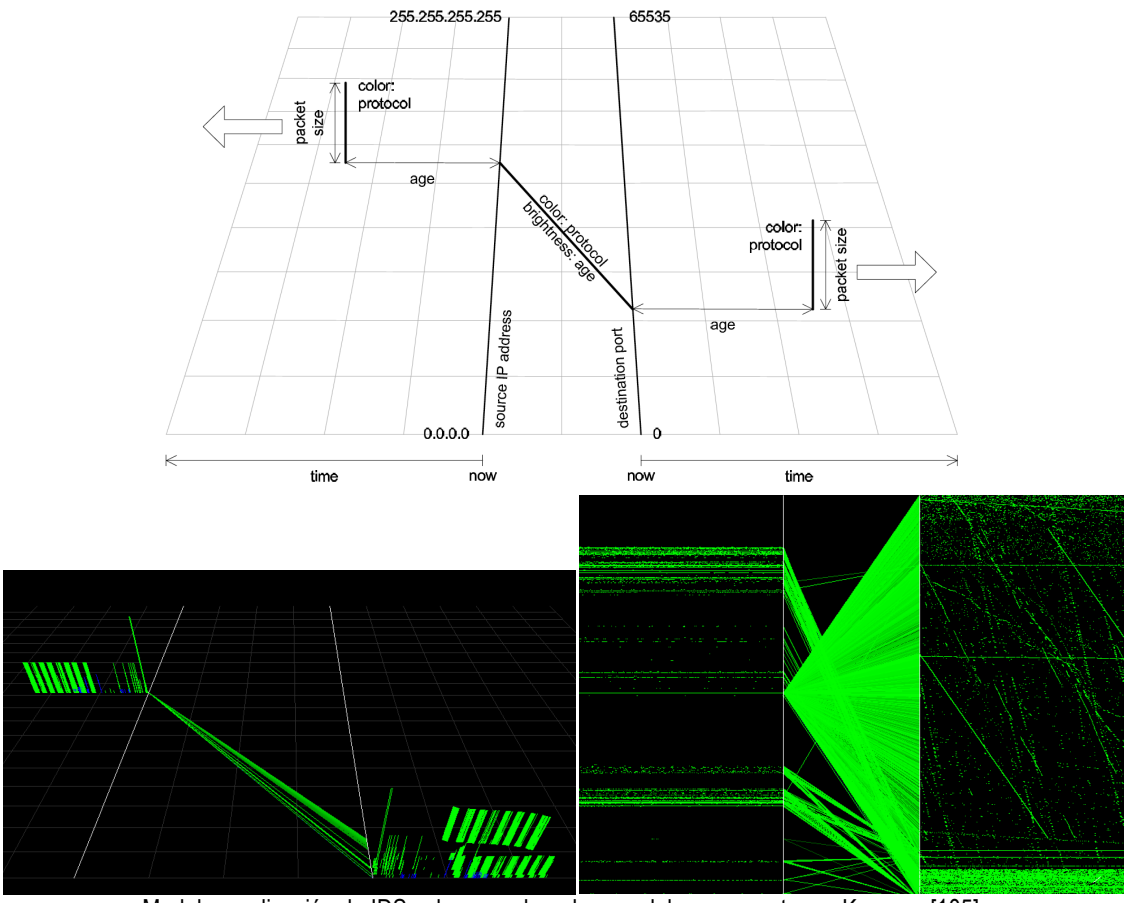
La visualización propuesta por los autores se basa en la representación de la actividad de la red sobre un gráfico circular en el que cada circulo concéntrico representa una unidad temporal, a lo largo de los círculos se sitúan las direcciones IP locales y en el rectángulo que rodea esta figura, se ubican las direcciones IP remotas y sus puertos.



Representación propuesta por Erbacher [102] para visualizar la actividad en la red y ejemplo.

El gráfico propuesto es muy versátil, ya que permite representar tanto el tráfico cursado en paquetes, como las conexiones realizadas. Por otro lado los intervalos temporales transcurridos entre círculos concéntricos puede ser igualmente escalados según la necesidad del investigador. Se trata de una especie de coordenadas concéntricas, en analogía a las coordenadas paralelas propuestas por Inselberg en 1985 [103]. En la segunda representación mostrada anteriormente puede detectarse cómo se está produciendo una actividad compatible con un ataque de rastreo de puertos (*port scan*), visualizado en el grupo de líneas que parten de puertos incrementales hacia direcciones IP internas en círculos concéntricos sucesivos (intervalos temporales).

En la conferencia del IEEE Integrated Network Management celebrada en Niza en 2005 (*IEEE IM-2005*) Delic y Dayal presentaron un análisis sobre la aplicación de métodos analíticos a la gestión de redes [104]. En el mismo hacían referencia a la necesidad de herramientas capaces de correlar datos de diversa naturaleza. Por ejemplo, correlaciones entre patrones de tráfico y amenazas de seguridad, correlación entre comportamientos de usuarios y problemas en la red. Concluyen la necesidad de utilizar herramientas de visualización para aumentar la percepción humana y comprensión de eventos en la red, mejorando las toma de decisión en la gestión de redes mediante simulación, pronóstico, predicción, optimización y diagnosis. Coinciden con la aplicabilidad estas técnicas en seguridad, gestión de prestaciones y planificación.



Modelo y aplicación de IDS sobre coordenadas paralelas propuesta por Krasser. [105]

En 2005 Krasser *et al* presentan en un workshop sobre seguridad una aplicación de coordenadas paralelas en detección de intrusos [105]. La aplicación de coordenadas paralelas propuesta difiere de la versión "clásica" en dos aspectos: incorpora una tercera dimensión para representar el tamaño de los paquetes y es animada/dinámica, para representar la evolución temporal. La propuesta de Krasser comprende tanto análisis en tiempo real como forense. El analista puede seleccionar diferentes escalas

temporales y enfocar su área de interés a zonas específicas de la representación para captar más detalles. Los autores proponen la metodología visual, que afirman es poco utilizada en áreas de seguridad de redes, para obtener resultados más rápidos que, por ejemplo, utilizando analizadores de protocolos más clásicos (Ethereal), si bien posteriormente se podría recurrir a ellos para el análisis en detalle de los eventos evidenciados en las representaciones gráficas.

Pero el trabajo más relevante, a nuestra manera de ver, sobre representación visual de información en su sentido más general y no solo de redes, es el libro de Thomas y Cook que fue publicado en 2005 [106]. Se trata de un informe respaldado por el Departamento de Seguridad Nacional y el Centro Nacional de Visualización y Analítica, ambos de Estados Unidos. Proponen una detalladísima agenda de acción para aplicar la analítica visual en la seguridad nacional de Estados Unidos. Afirman los autores que es necesario el análisis de una cantidad abrumadora de información dispar, contradictoria y dinámica para identificar y prevenir nuevas amenazas, proteger nuestras fronteras [las suyas, se entiende] y responder en caso de un ataque o cualquier otro desastre. Este proceso de análisis requiere el juicio humano para realizar lo mejor posible la evaluación de la información incompleta, inconsistente y potencialmente engañosa de cara a situaciones que cambien rápidamente. Se requieren nuevos métodos que permitan analizar esta riada de información masiva, multidimensional, multifuente, variable en el tiempo para adoptar decisiones en un escenario en el que el factor tiempo es crítico. La mente humana puede comprender información compleja recibida a través de canales visuales. La analítica visual se construye sobre esta habilidad para facilitar el proceso de razonamiento analítico.

En este informe se define la Analítica Visual como la ciencia del razonamiento analítico ayudado por interfaces visuales interactivas. La gente, reflexionan, usa técnicas y herramientas analíticas visuales para sintetizar la información y obtener conocimiento a partir de datos masivos, dinámicos, ambiguos, y con frecuencia contradictorios; detectar lo esperado y descubrir lo inesperado; proporcionar evaluaciones oportunas, defendibles y comprensibles, y comunicar de manera eficaz la evaluación para la toma de acciones.

La analítica visual es un campo multidisciplinario que incluye las siguientes áreas de interés :

- Técnicas de razonamiento analítico que permiten a los usuarios obtener una visión profunda que apoyan directamente la evaluación, la planificación y la toma de decisiones

- Las representaciones visuales y técnicas de interacción que se aprovechan del amplio ancho de banda del ojo humano hacia la mente para permitir a los usuarios ver, explorar y comprender grandes cantidades de información a la vez

- Representaciones de datos y las transformaciones que convierten todos los tipos de datos contradictorios y dinámicos en formas que apoyen la visualización y el análisis

- Técnicas de apoyo a la producción, presentación y difusión de los resultados de un análisis para comunicar la información en el contexto adecuado para una variedad de audiencias.

La analítica visual busca maridar técnicas de visualización de información con técnicas de transformación de datos y análisis de datos. Son también necesarias visualizaciones que combinen varios tipos de datos para apoyar el razonamiento analítico amplio en ciertas situaciones. Desarrollar teorías y prácticas para transformar los datos en nuevas representaciones escalables que representen fielmente el contenido de los datos subyacentes.

Se deben desarrollar transformaciones matemáticas y representaciones que sea escalables para hacer frente a grandes cantidades de datos de manera oportuna. Estos enfoques deben proporcionar una representación de alta fidelidad del verdadero contenido de la información de los datos subyacentes. Deben apoyar la necesidad de analizar un problema a diferentes niveles de abstracción y considerar los mismos datos desde múltiples puntos de vista. Los métodos de transformación deben incluir técnicas para detectar cambios, anomalías y aparición de tendencias.

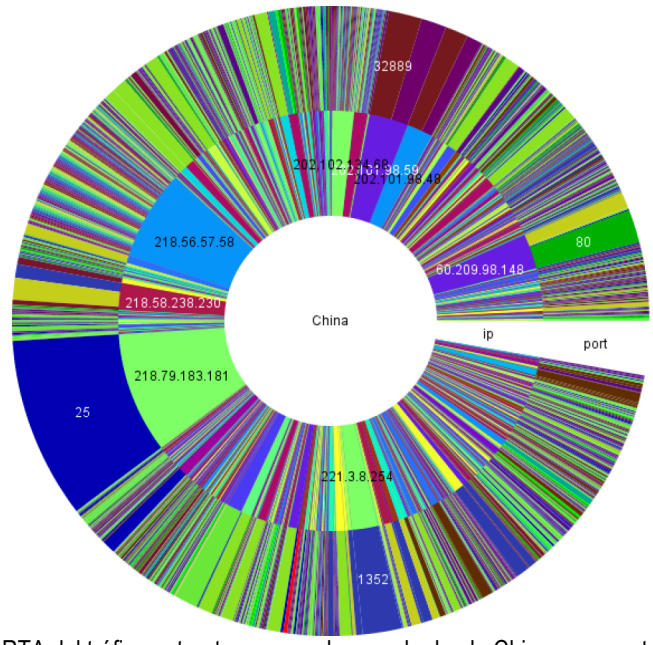
En 2005 Keim *et al* publican un capítulo [107] en el libro de Hansen y Johnson “The visualization Handbook” [108] sobre técnicas de minería de datos visuales. Además de la implicación directa de la visión humana en el proceso de la exploración de datos visual, mencionan los autores las siguientes ventajas sobre las técnicas de minería de datos más automáticas:

- La exploración de datos visual puede fácilmente tratar con datos con ruido y altamente no homogéneos.

- La exploración de datos visual es intuitiva y no requiere la comprensión de matemática compleja o algoritmos estadísticos o parámetros.

- La visualización puede proveer una visión de conjunto cualitativa de los datos, permitiendo aislar los datos interesantes para un análisis cuantitativo posterior.

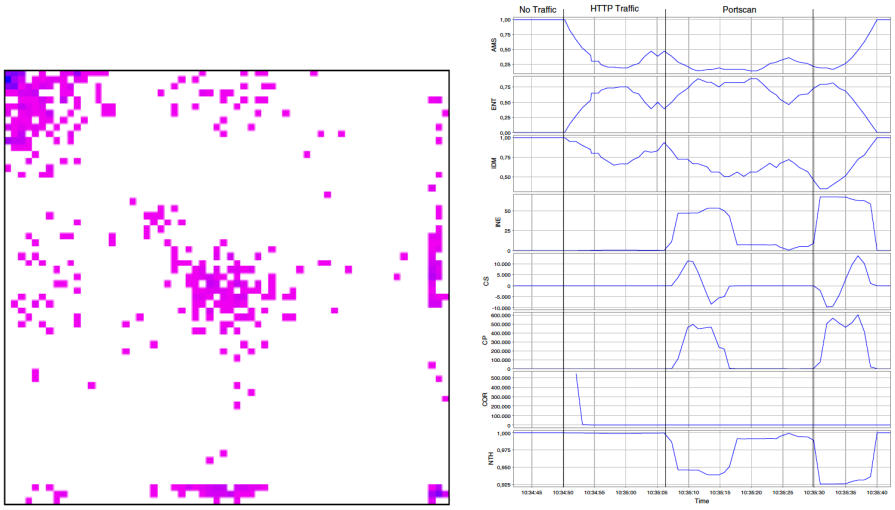
En 2006, de nuevo Keim *et al* [109] presentan una propuesta de representación para monitorizar el tráfico de redes basada en representaciones radiales (*Radial Traffic Analyzer – RTA*). En el gráfico siguiente, por ejemplo, se muestra el tráfico cursado desde un país, China, mostrando las direcciones IP de origen y los puertos de destino en la red interna. En esta representación pueden detectarse varias anomalías: Se observa un ataque de rastreo de puertos desde la IP 218.56.57.58, así como un elevado número de intentos de efectuar conexiones con el servidor de correo (puerto 25) desde la IP 218.79.183.181.



Representación RTA del tráfico entrante a una red, cursado desde China, propuesta por Keim. [109]

En 2007 Kisner presenta [110] una aplicación de visualización de tráfico utilizando matrices dinámicas de “co-ocurrencia” ya propuestas por Oka en 2004 [111] (por cierto basadas en ACP) pero en cuyo trabajo no se consideraban aspectos de visualización. Como en casos precedentes los autores afirman que la visualización de tráfico es importante en áreas tales como planificación de redes y monitorización, análisis de tráfico y detección de intrusos. Aunque los humanos no somos capaces de entender a

la vez grandes cantidades de datos en modo texto, tenemos capacidades excepcionales en procesado de imagen, por lo que los métodos de representación de datos han sido objeto de investigación desde tiempo atrás. Dado que, en la mayoría de los casos, se trabaja con datos multidimensionales se ha puesto mucho esfuerzo en técnicas de reducción de la dimensionalidad de datos multivariantes para ser capaces de su representación en el plano. Así, se representa el tráfico de redes como series temporales y se traspasarán al dominio del procesado digital de imagen mapeando las muestras de tráfico en una representación. Con este método se hace uso de técnicas de análisis estadístico para analizar y visualizar el conjunto de los datos. En este campo se utilizan el método de matrices de auto co-ocurrencia ya reseñado. En la representación de la serie temporal del tráfico se observa la naturaleza autosimilar y fractal de las observaciones.



Visualización de la matriz de concurrencias y parámetros a partir de los que se obtiene. [110]

En 2007 Pras *et al* publican, como vimos someramente en la introducción a la gestión de redes, los retos claves en la investigación sobre gestión de redes [73]. En él aparece un extenso apartado dedicado al análisis de datos y visualización. Dado que estos sistemas de gestión recopilan grandes cantidades de datos de monitorización y medición estos deben ser agregados, filtrados y visualizados, con el objetivo de sacar a la luz información significativa fácilmente comprensible por los operadores. Aunque el análisis de datos y su visualización es un aspecto considerado tradicional en la gestión de redes, aparentemente las herramientas disponibles no satisfacen realmente a los gestores por diversas razones:

- Las representaciones topológica clásicas, especialmente las basadas en mapas georeferenciados, no disponen de la escalabilidad necesaria con el

incremento de elementos en las redes. Además las diferentes capas que constituyen las redes solo agravan el problema de la escalabilidad.

- Los conjuntos de datos recopilados y sus estadísticas son frecuentemente representadas de una manera más bien estática. Las capacidades para realizar un análisis exploratorio de los datos de manera interactiva o son muy limitadas o simplemente no existen.

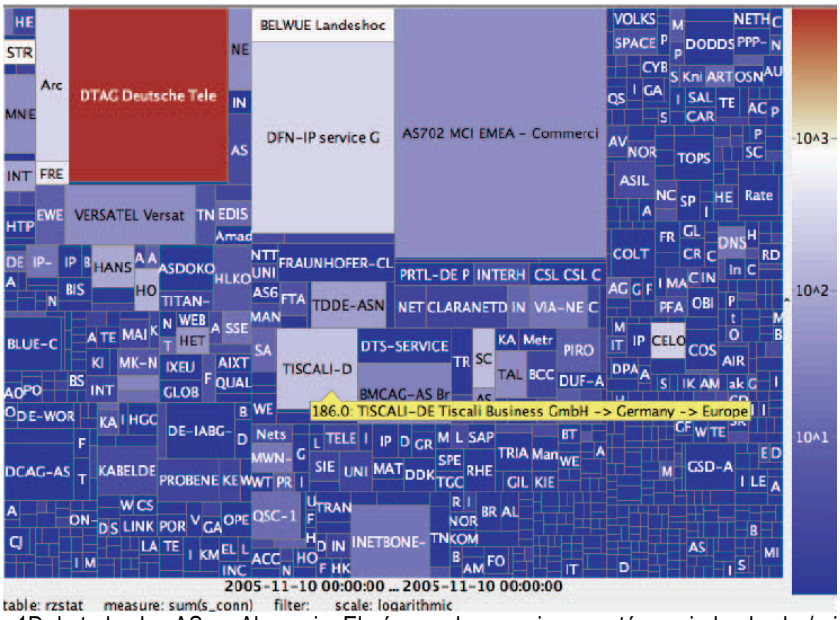
- Las representaciones de tráfico típicamente se centran en la visualización de grandes volúmenes de componentes de tráfico o flujos. Aunque esto es útil para planificación, y quizá para contabilización, hay una creciente necesidad de extraer y destacar el tráfico inusual y los patrones anómalos. Especialmente para propósitos de auditoría de seguridad, con frecuencia es mucho más deseable descubrir e identificar pequeños volúmenes de tráfico anómalo o patrones extraños.

- La mayoría de herramientas están diseñadas para el análisis y visualización fuera de línea. No obstante hay una creciente necesidad de análisis y visualización en tiempo real, o próxima a tiempo real, para reducir los tiempos de detección y reacción. Con velocidades de transmisión en las redes del orden de múltiplos de decenas de gigabits por segundo, este objetivo es cualquier cosa, menos trivial. Las iniciativas que avanzan hacia la captura estadística de datos en redes de alta velocidad también precisan de una representación precisa de los datos capturados.

Algunas investigaciones básicas en técnicas de representación de datos han sido llevadas a cabo dentro del proyecto CAIDA. La mayoría de las técnicas se basan en representaciones bidimensionales, aunque también se ha realizado alguna investigación en técnicas tridimensionales.

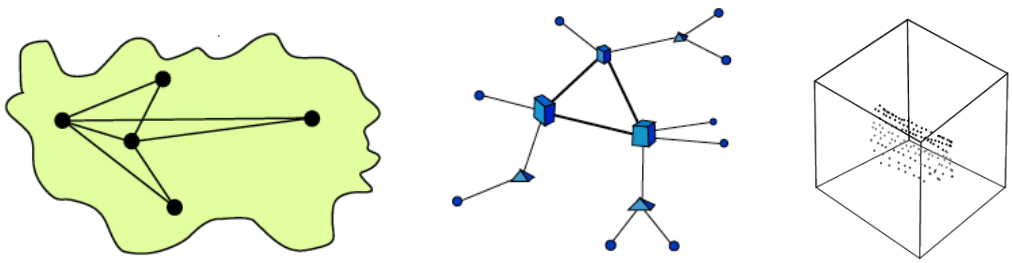
En 2007 Mansmann *et al* publican una propuesta de mapas de red jerárquicos [112]. Plantean inicialmente los exigentes requisitos que la visualización interactiva de datos de seguridad y de redes grandes demandan de las bases de datos que almacenan la información. Una buena elección para este almacenamiento son las arquitecturas OLAP (*On-Line Analytical Processing*). El modelo de datos multidimensional subyacente posibilita mapear los datos en un espacio multidimensional o hipercubo. Los valores numéricos bajo análisis estarán caracterizados por sus valores descriptivos desde su número de dimensiones. Así,

los autores proponen un mapa de red jerárquico que permite representar en el plano un espacio de dimensiones y el tráfico cursado.



HistoMap 1D de todos los AS en Alemania. El número de conexiones está asociado al color/gris. [112]

En 2007 Withall *et al* publican una completa revisión sobre la "visualización de redes" [113]. Establecen que la visualización de redes es una técnica para la presentación del enorme volumen de información producido a través de la monitorización de dichas redes. Establecen una clasificación de tres tipos de visualización: visualizaciones geográficas, en las que los datos están presentados con relación a la localización física de los nodos en la red; visualizaciones topológicas abstractas, en las que las relaciones entre los nodos están presentadas independientemente de las localizaciones físicas; y visualizaciones basadas en gráficos, en las que el foco es un punto único de la red, y con frecuencia las variables capturadas son representadas respecto al tiempo. Inciden en el "mantra" de la búsqueda visual de información: vistazo de conjunto, primero; ampliación y filtrado, luego detalles bajo demanda. Reafirman la existencia de 7 tipos de datos diferentes: uni-, bi-, tri-dimensionales, temporales y multidimensionales, y árboles y datos de la red.



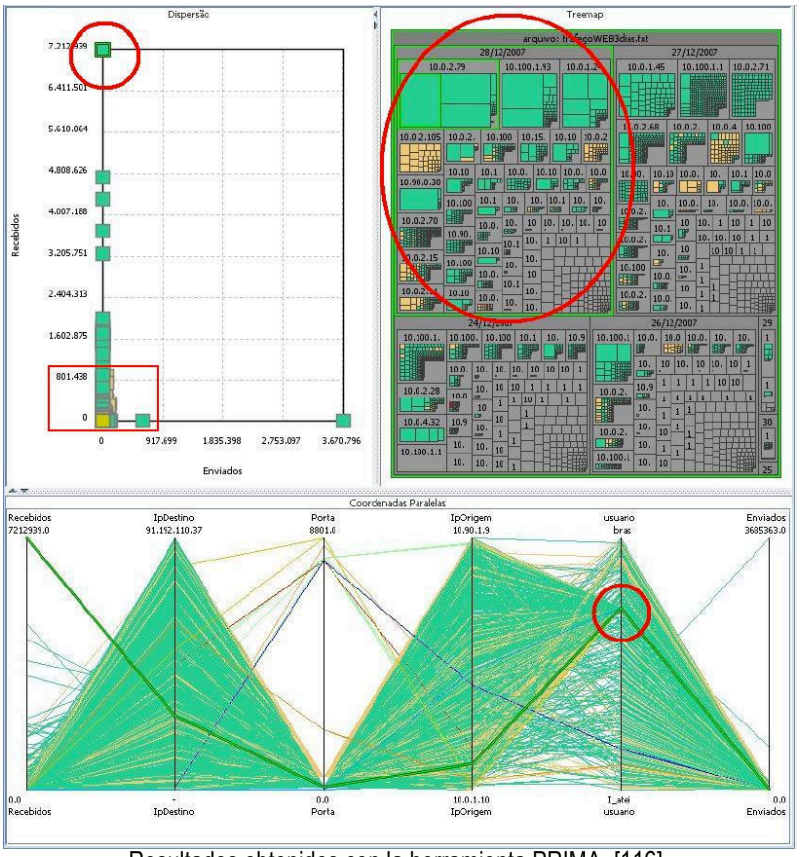
Ejemplos de representaciones geográficas, topológica abstracta y gráfica, recopiladas por Withall.[113]

En 2008 de nuevo Keim *et al* publican dos artículos [114], [115] sobre el concepto "*Visual Analytics*", que podría traducirse bien directamente como "Analíticas Visuales" o de manera libre como "Estadística gráfica". La idea básica de la analítica visual es la representación gráfica de la información de forma que permita al humano interactuar directamente con la información, para una mayor comprensión de los datos, extraer de ellos conclusiones y por último adoptar las mejores decisiones al respecto. La analítica visual combina técnicas automatizadas de análisis con visualizaciones interactivas sobre grandes y complejos conjuntos de datos. Plantean los autores la diferencia entre analítica visual y visualización de información, que definen como la comunicación de datos abstractos relevantes a través de la utilización de interfaces interactivas. En la visualización, afirman, hay tres retos importantes: presentación, análisis confirmatorio y análisis exploratorio. Coinciden con Van Wijk [99] en que la visualización no es "buena" por definición, frente a otras alternativas de análisis. Así, los desarrolladores de nuevos métodos deberán dejar claro porqué la información buscada a través de la visualización no puede ser extraída automáticamente. Reiteran el conocido mantra del análisis visual:

"Primero analizar – Mostrar lo importante – Ampliar, filtrar y analizar de nuevo – Detalles bajo demanda"

Enumeran, así mismo, diversos casos prácticos en los que el análisis visual es importante. En el caso de aplicaciones de seguridad exponen la necesidad de que el análisis responda a las tres preguntas importantes que deben afrontarse ante cada incidente: ¿Qué?, ¿Cuándo? y ¿Dónde?. En muchas aplicaciones, el éxito del análisis dependerá, además, de que la información correcta esté disponible en el momento oportuno. Plantean dos conjuntos de datos básicos en los datos de redes: datos espacio-temporales y gráficos de redes.

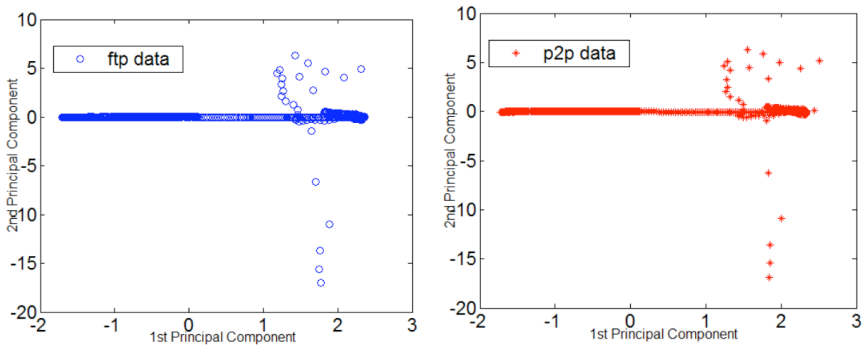
En 2008 Kauer *et al* [116] presentan una nueva aplicación de representación de tráfico que ofrece tres técnicas de visualización diferentes: coordenadas paralelas, árboles y diagramas de dispersión. El análisis del tráfico de redes permite conocer el comportamiento de los usuarios y mejora la concienciación de los usuarios para una mejor seguridad de la información en la entidad. Esto es importante, ya que el 50% de los incidentes de seguridad están originados en el interior de la redes. Las herramientas para analizar el tráfico están mayoritariamente compuestas por técnicas estadísticas, de minería de datos y de visualización de información.



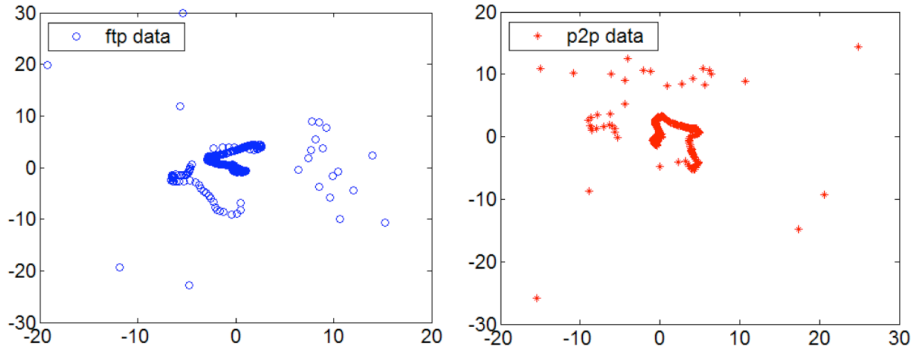
Resultados obtenidos con la herramienta PRIMA. [116]

También en 2008 Han y Hemert presentan [117] un sistema de discriminación de patrones de tráfico basado en *wavelets*. Las aplicaciones de las *wavelets* han sido utilizadas en otras propuestas debido a su capacidad de extraer simultáneamente información temporal y frecuencial de los perfiles de tráfico. No obstante, esta aplicación presenta algunas dificultades: las representaciones obtenidas presentan una alta dimensionalidad, por lo que son difíciles de analizar y las estructuras que las componen no puede ser fácilmente extraíbles con técnicas de representación convencionales; en segundo lugar los datos multidimensionales, con muchos parámetros intervinientes en el modelo, incrementan la complejidad de los modelos y los requerimientos de cálculo; por último es complicado optimizar el procedimiento debido al volumen de datos generados. Proponen una técnica de proyección cuyo objetivo es representar los datos de entrada en un espacio de baja dimensión, de tal manera que la estructura de los datos se preserve tan fielmente como sea posible. Utilizan para ello un ACP, pero no obstante el ACP no considera fielmente las estructuras no lineales ni las estructuras que consisten en agrupaciones superpuestas o subespacios curvos, dado que el ACP describe los datos en un subespacio lineal. Proponen los autores, además, la utilización del mapeado de Sammon, que es una técnica de escalado multidimensional. El mapeado de Sammon es una técnica no lineal de proyección que mapea un

espacio n-dimensional en uno bidimensional. Se trata de un procedimiento iterativo que utiliza el gradiente del error para minimizar el error: los puntos se posicionan en el espacio de baja dimensión de manera que la distancia entre los puntos sea lo más próxima posible a la distancia en el espacio de dimensión elevada. La carga de cálculo para esta operación es considerable, ya que cada evaluación del error requiere la obtención de n(n-1)/2 distancias entre puntos. El procedimiento propuesto primero realiza una descomposición de *wavelets* de Daubechies en base 4 de la serie de datos de tráfico, para construir la matriz que posteriormente será sometida al ACP. Pero dada la dificultad del ACP para captar el comportamiento no lineal, se propone utilizar el mapeado de Sammon.



Comparativa del resultado del análisis *wavelet*+ACP sobre tráfico FTP y P2P. [117]

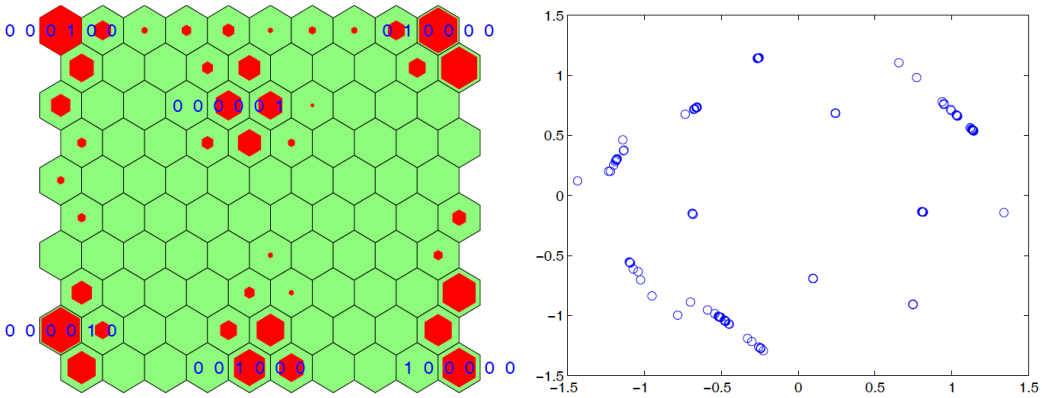


Comparativa del resultado del análisis *wavelet*+Sammon sobre tráfico FTP y P2P. [117]

Como puede apreciarse a primera vista en las representaciones precedentes, la opción de Sammon permite captar más variabilidad que el ACP.

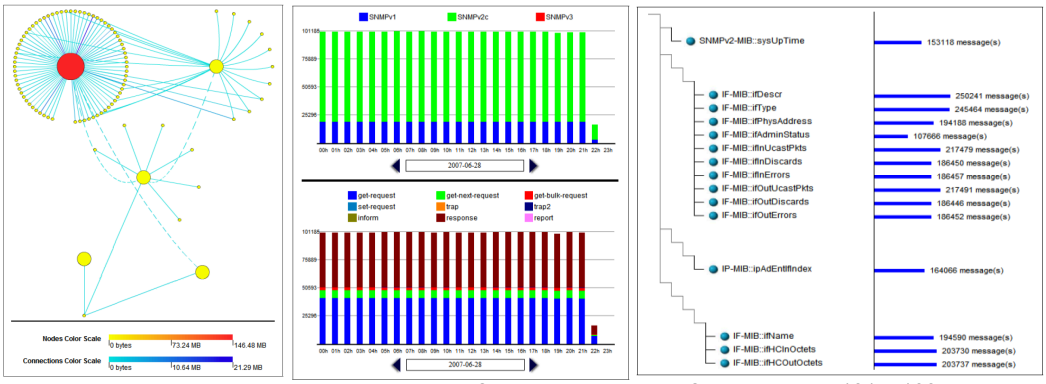
Nayak *et al* publican en 2008 [118] una nueva propuesta para la visualización y análisis de los datos recopilados por un sistema de monitorización. Como se viene afirmando por todos los autores, para la comprensión eficaz y eficiente por parte del personal técnico de los datos multidimensionales provenientes de una infraestructura de TI, es precisa su proyección en un espacio de dimensión reducida. A lo largo de los años se han propuesto para esa función multitud de técnicas de reducción de la dimensionalidad. La aplicación directa de esas técnicas para la visualización de datos de monitorización se enfrenta dos retos. El

primero es que los datos de monitorización no se apoyan en un espacio métrico. El segundo problema es que esos datos son intrínsecamente multidimensionales, en el sentido que un único evento puede provocar un cascada de ellos. Las representaciones en baja dimensión que no tengan en cuenta la naturaleza multidimensional de la información de monitorización presentarán por lo tanto importantes limitaciones. Nayak *et al* proponen la utilización de los denominados Mapas de Características Autoorganizados (*Self-Organizing Feature Maps – SOFM*) propuestos originalmente por Kohonen [119], [120]. La figura siguiente muestra un ejemplo de proyección de un espacio 6D en un espacio 2D utilizando dos técnicas de mapeado diferentes, la propuesta de Kohonen y la ya analizada anteriormente de Sammon. Se asume que los eventos visualizados son mutuamente excluyentes en el tiempo y que cada vector de seis dimensiones de un evento está compuesto por todos los elementos "cero" excepto un "uno". El mapa de SOFM es capaz de visualizar los seis clusters diferenciados, la propuesta de Sammon no lo es. La técnica propuesta por Nayak *et al* utiliza redes neuronales para efectuar la reducción de la dimensionalidad y precisa por lo tanto de entrenamiento previo.



Comparativa entre resultados obtenidos con el mapeado de Kohonen (izq.) y de Sammon (der.). [118]

En 2008 Salvador *et al* presentan en dos conferencias sendos artículos [121], [122] proponiendo técnicas de visualización para analizar datos de trazas SNMP. El enorme volumen de datos generados por los sistemas de gestión de redes hacen precisa la utilización de técnicas visuales que faciliten la interpretación de las situaciones. Proponen tres tipos de gráficos diferentes, uno topológico que incluyen información sobre volumen de información cursada, histogramas más clásicos e histogramas "jerárquicos" apoyados en la estructura de la MIB.



Propuesta de representación de datos SNMP realizadas por Salvador *et al*. [121], [122]

En 2008 Tong *et al* presentan una propuesta que denominan COLIBRI [123] para el análisis de gráficos dinámicos y estáticos de gran tamaño. No es propiamente una artículo sobre visualización, pero aporta varias ideas interesantes apoyadas en la descomposición de matrices. Afirman los autores que las descomposiciones "estándar" no respetan la escasez de datos. Esta situación dio lugar al desarrollo de otros métodos de descomposición, tales como CUR [124] y CMD [125]. Reiteran los autores que una matriz es una representación muy común de un gráfico. Así la matriz de proximidad unipartita, donde cada fila/columna corresponde a un nodo del gráfico y cada elemento no nulo es una arista del gráfico, y las matrices de interacción para gráficos bipartitos en los que filas y columnas corresponde a dos tipos diferentes de nodos y los elementos no nulos indican aristas entre ellos.

Como afirman los autores, las aproximaciones de bajo rango proporcionan herramientas para descubrir patrones en los gráficos, tanto estáticos como dinámicos. Formalmente las **A'** aproximaciones de bajo rango c de una matriz **A** se representan usualmente como factorizaciones de una matriz

A' = LMR con **L**, **M**, **R**, matrices de rango r

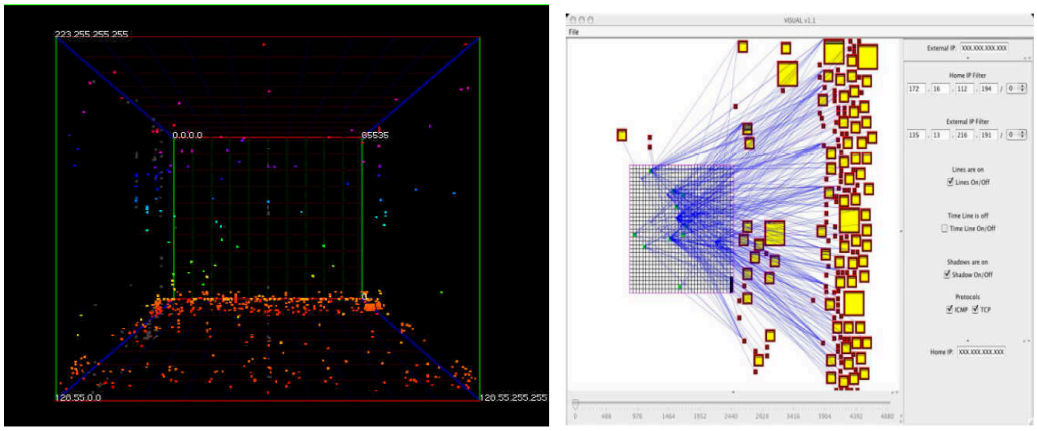
En el caso de la Descomposición en Valores Singulares (DVS), **L** y **R** son matrices ortogonales cuyas filas/columnas con vectores singulares y **M** es una matriz diagonal de valores singulares. Entre todas las aproximaciones de bajo rango c, la DVS proporciona la mejor aproximación en términos del error cuadrático. No obstante, la DVS es usualmente "densa", esto es, la mayoría de sus elementos son no nulos, incluso si la matriz original está escasamente poblada. Además, los vectores singulares con elementos abstractos de la mejor base ortonormal, lo que no les otorga significados intuitivos para la interpretación de los resultados. Con esa situación aparecieron las descomposiciones CUR y CMD, que utilizan las filas y columnas de las matriz original para construir **L** y **R**. A estas descomposiciones las denominan

“aproximaciones de bajo rango basadas en ejemplos”. El mayor beneficio es que proveen una representación intuitiva y también escasamente poblada, ya que **L** y **R** están directamente muestreadas de la matriz original. Por el contrario, la aproximación es frecuentemente subóptima comparada con la DVS y la matriz **M** no es diagonal, lo que significa una interacción más complicada. Los autores proponen otro método, que denominan Colibri-S, para la representación de gráficos estáticos que mejora la eficiencia de los existentes aprovechando la correlación lineal existente entre las diferentes columnas muestreadas de la matriz original, y con resultados similares en cuanto a precisión que las propuestas CUR/CMD. Proponen también otro procedimiento, que denominan Colibri-D, para el análisis de gráficos dinámicos (que evolucionan en el tiempo), con características similares al Colibri-S cuando lo comparan con las propuestas existentes en la literatura para análisis de gráficos dinámicos apoyados en DVS, series dinámicas, análisis dinámico de tensores y clustering espectral incremental, entre otros. No obstante no proporcionan ni una sola representación visual de sus resultados.

En 2009 Goodall presenta un estudio comparativo evaluando los resultados que se obtienen utilizando una herramienta visual para analizar el tráfico en una red y una tradicional basada en texto [126]. Se pretende evaluar una serie de hipótesis que permitan concluir si la representación visual de la información proveniente de la gestión de redes presenta mejoras frente la tradicional. Concluyen que los participantes obtienen una mayor precisión en el desarrollo de una tareas concretas utilizando la herramienta visual. Además estas tareas se realizan en menos tiempo utilizando la herramienta visual. En lo que respecta a la detección exploratoria de detalles, de nuevo la herramienta visual presenta un mejor comportamiento que la tradicional. Por último, los participantes prefieren la herramienta visual frente a la textual. Con todo ello, el autor concluye que la representación visual es mejor que la textual.

Ten presenta en 2009 un estudio sobre las herramientas de visualización para la monitorización de redes [127]. Los resultados de la monitorización pueden ser presentados en representaciones multivariantes, bidimensionales y tridimensionales. La utilización de sistemas de monitorización de redes integrados con innovadoras herramientas de visualización pueden mejorar notablemente los tiempo de respuesta de los administradores de redes y acelerar la resolución de problemas. El mayor problema del análisis del tráfico de redes reside en el constante incremento del volumen del tráfico y la incapacidad de las herramientas tradicionales de

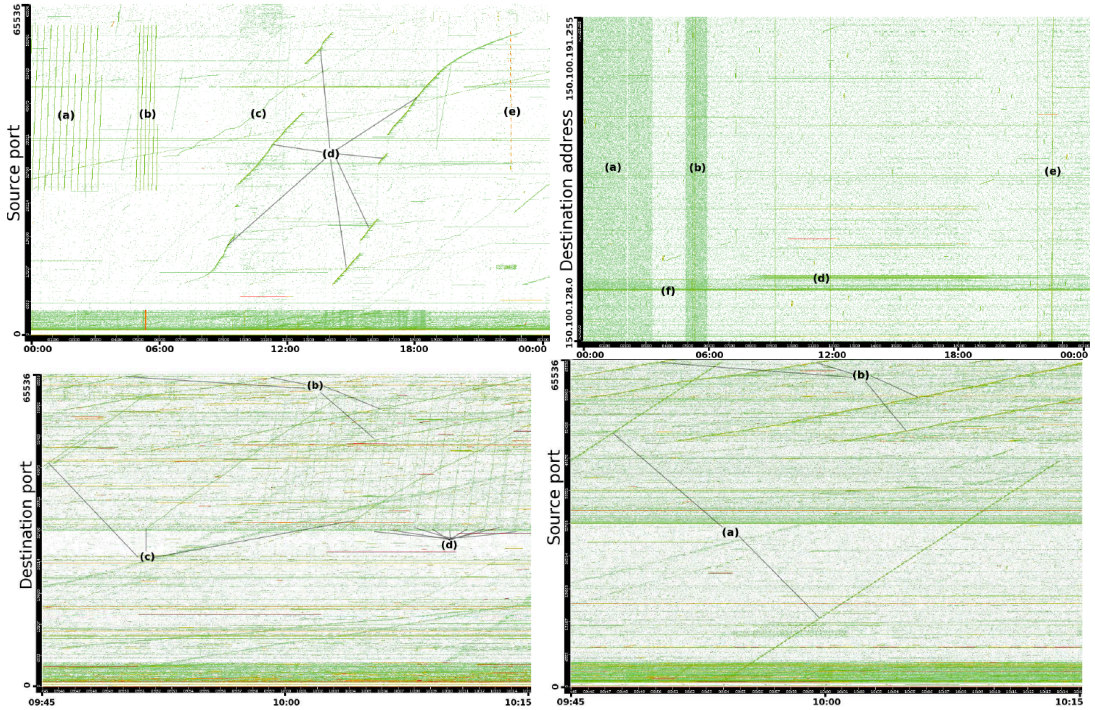
monitorización para ofrecer una buena visión de conjunto de los patrones de tráfico. Es difícil mejorar el conocimiento de la naturaleza del tráfico utilizando las herramientas tradicionales ya que el elevado volumen de información se torna rápidamente inmanejable.



Representaciones de las herramientas SCPD y VISUAL extraídas del estudio de Ten. [127]

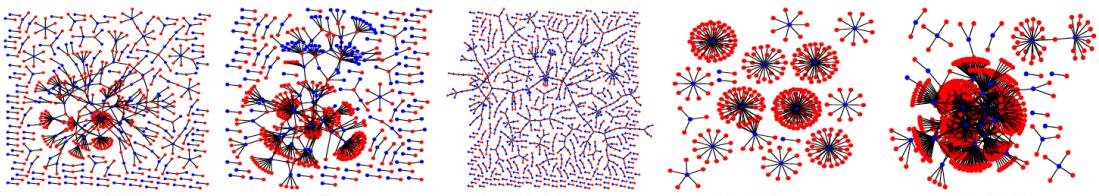
En 2009 Fontugne *et al* presentan una herramienta para explorar anomalías en el tráfico de las redes [128]. Partiendo de la base de que la detección rápida y precisa de anomalías en el tráfico de redes es importante y presenta complicaciones debido al número de dimensiones implicadas y el elevado volumen de datos a interpretar, las herramientas de visualización proporcionan una herramienta muy eficaz para resolver el problema existente. El método propuesto utiliza gráficos de dispersión bidimensionales, partiendo de datos originales de 5 dimensiones (puerto origen, puerto destino, dirección origen, dirección destino y tiempo), que son proyectados en diversos planos, en los que el eje horizontal siempre representa el tiempo, pero el eje vertical varia: puerto origen, puerto destino, dirección origen, dirección destino, número de paquetes, número de bytes y tamaño medio del paquete. No obstante, a nuestro modo de ver, los resultados obtenidos distan mucho de ser intuitivos.

Así, por ejemplo, en las representaciones mostradas a continuación pueden observarse dos regiones, (a) y (b), que se corresponden con un intento de acceso a múltiples equipos en un intervalo muy corto de tiempo, al igual que la región (e). Las líneas (c) y (d) también corresponden también se corresponden con ataques, en este caso dirigidos a puertos 80 y 161.



Representaciones obtenidas por Fontugne *et al* [128].

En 2009 Jin *et al* presentan una propuesta de descomposición de gráficos de actividad de red (*TAG – Traffic Activity Graphs*) basada en la factorización ortogonal de matrices no-negativas (*tNMF – orthogonal nonnegative matrix tri-factorization*) [129]. En los TAG los nodos son direcciones IP (*hosts*) y los vértices representan comunicaciones de interés entre nodos. Dependiendo del objeto del estudio son posibles diferentes tipos de TAGs en función del criterio de utilizado para seleccionar el tráfico de interés. En general los TAGs suelen ser grandes, dispersos, aparentemente complejos y altamente conectados. Los TAGs fueron inicialmente propuestos y estudiados por Iliofotou en 2007 [130] y Jin *et al* proponen su análisis mediante su descomposición utilizando tNMF para capturar las interacciones más significativas entre los grupos dominantes de hosts.



(a) HTTP (b) Email (c) AOL Messenger (d) BitTorrent (e) DNS

Ejemplos de TAGs de diferentes aplicaciones de red [129].

Un TAG se define formalmente como un gráfico bipartito en el que direcciones internas de la red y direcciones externas (dos tipologías) aparecen “conectadas” por un vértice si al menos existe una comunicación o flujo entre ellas dos. También puede

considerarse una versión con “pesos”, en la que el peso de ese vértice venga atribuido por el volumen de tráfico cursado, por ejemplo.

Por lo que respecta al método de análisis propuesto por Jin *et al* fue presentado por Ding *et al* en 2006 [131] éste se formula de la siguiente manera:

Dada una matriz no-negativa $\mathbf{A}_{mxn}$ podemos descomponerla en tres matrices no-negativas de menor rango $\mathbf{R}_{mk}$, $\mathbf{H}_{kl}$, $\mathbf{C}_{nl}$ tales que minimicen la siguiente función J bajo la condición de ortogonalidad en **R** y **C**.

$$\min_{R\geq 0, C\geq 0, H\geq 0} \left[J(\mathbf{R},\mathbf{H},\mathbf{C}) = \left\| \mathbf{A} - \mathbf{RHC}^T \right\|_F^2 \right] \text{ con las restricciones } \mathbf{R}^T\mathbf{R}=\mathbf{I} \text{ y } \mathbf{C}^T\mathbf{C}=\mathbf{I}$$

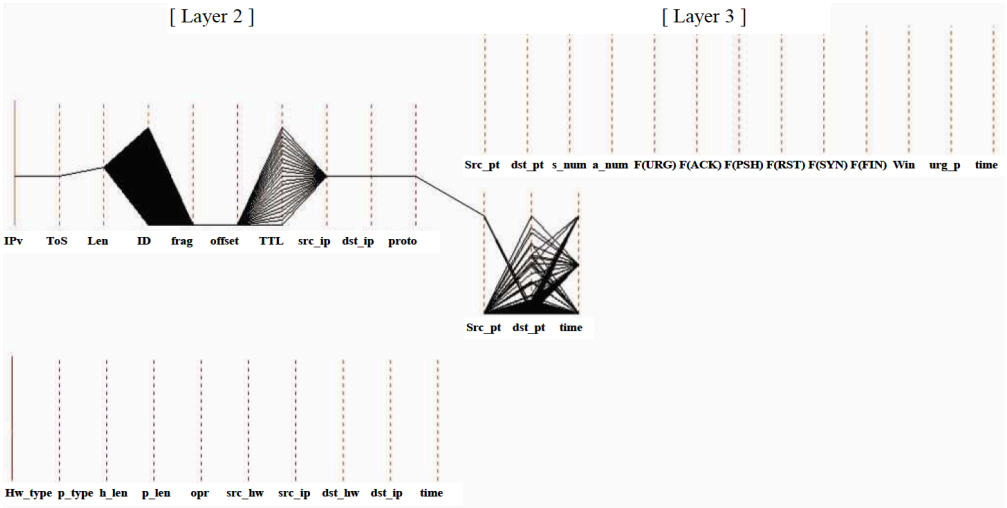
Y donde $||.||_F$ es la norma de Frobenius y k,l << min(m,n). La restricción de ortogonalidad diferencia el método propuesto de otros existentes en la literatura que permiten agrupar simultáneamente filas y columnas de una matriz. Para resolver el problema planteado se utiliza un criterio de optimización iterativo hasta que el error cuadrático baja de un determinado umbral. De nuevo los autores no proponen representaciones gráficas en su propuesta.

En 2009 Kim *et al* proponen [132] un avance sobre las coordenadas paralelas clásicas consistente en la visualización de los protocolos de red superiores en coordenadas “multiparalelas” y así mejorar el análisis de ataques. Parten, entre otras, de la propuesta de Krasser [105], sobre la que los autores reflejan que utiliza pocas características de entre las posibles. Utilizando más características del tráfico es posible descubrir más información sobre los ataques, representándola en un gráfico bidimensional.

La visualización de datos multivariantes en coordenadas paralelas muestra datos N dimensionales en un plano bidimensional sin utilizar ningún algoritmo de proyección, esto es, sin emplear ninguna transformación, lo que reduce la posibilidad de errores. Las ventajas de la utilización de coordenadas paralelas son, según los autores, la ausencia de preponderancia/peso otorgado a las características, todas se representan bajo las mismas condiciones, no presenta limitaciones en el número de características a representar, el usuario puede seleccionar tantas como desee, y por último el método permite visualizar correlaciones y tendencias entre los datos bajo estudio. La propuesta de Kim *et al* lleva más lejos las coordenadas paralelas, visualizando todas las propiedades contenidas en las cabeceras de los paquetes, en las que, para

mejorar la comprensión de los gráficos, estos se separan en dos representaciones de coordenadas paralelas para cada capa de análisis de la red. De facto se utilizan 4 representaciones en coordenadas paralelas diferentes, para otros tantos protocolos:

- Características del protocolo IP
- Características del protocolo ARP
- Características del protocolo TCP
- Características del protocolo UDP



Representación de un ataque por rastreo de UDP según el método propuesto por Kim. [132]

En 2009 Read *et al* presentan un enfoque unificado para la visualización del tráfico de redes y aplicaciones de seguridad [133]. Afirman los autores que la visualización juega un importante papel en la seguridad de las redes y en la gestión del tráfico en las redes para aportar una mayor comprensión de la vasta cantidad de datos tradicionalmente provenientes de diferentes fuentes que son utilizadas en ambos campos, con un elevado solapamiento en los mismos datos.

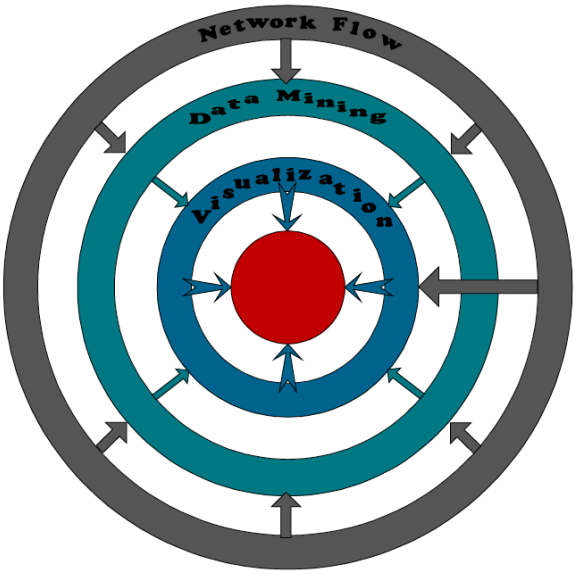
Los tres grandes requerimientos para las visualizaciones según Read *et al* son:

1. Monitorización: herramientas que proporcionan representaciones de alto nivel de los datos con escasas opciones para interacción.
2. Análisis: con vistas más en profundidad de los datos, con mayor énfasis en la interacción.
3. Respuesta: la capacidad para registrar eventos y almacenar históricos.

Read *et al* presentan de hecho una arquitectura de intercambio de datos que demuestra que la línea entre seguridad de redes y tráfico de redes es cada vez más

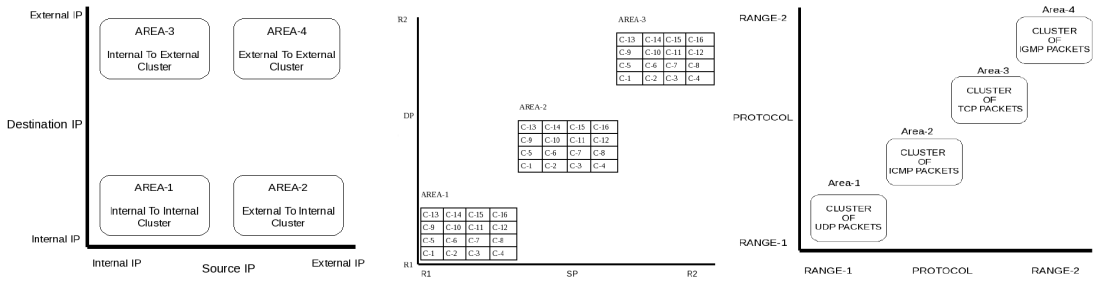
difusa. Como ejemplo de la intersección clara entre ambos campos exponen la detección de un ataque de negación de servicio distribuido DDoS por una herramienta de gestión de tráfico, mostrando el incremento de ancho de banda en uso.

Tal y como indican Shahrestani *et al* en 2009 [134], la minería de datos permite reconocer desde patrones de interés a identificar irregularidades en grandes volúmenes de datos. La minería de datos permite extraer de esos grandes volúmenes de datos un subconjunto suficiente para su análisis. Como consecuencia se pueden reducir los volúmenes de datos que almacenar y sobre los que trabajar. Un amplio rango de técnicas de minera de datos pueden ser utilizadas para extraer información de utilidad a partir de los datos de tráfico, como por ejemplo, correlación, clasificación, clustering, agregación y otras técnicas de análisis estadístico. En ellas, la visualización juega un papel importante posibilitando la detección de tráfico malicioso por parte del personal técnico aprovechando la habilidad conceptual de la visión humana para extraer información del entorno.



Arquitectura propuesta por Shahrestani para la monitorización visual de amenazas. [134]

Singh y Subramanian presentan en 2009 otra aplicación de visualización [135] para la identificación de anomalías en redes. Proponen la aplicación del análisis de cluster de K-medias simple para visualizar el tráfico. Utilizan para ello las direcciones IP internas y externas de las comunicaciones, los respectivos puertos, duraciones, tamaños, marcas temporales y protocolos utilizados. Las representaciones obtenidas no son únicas, en el sentido que presentan diferentes planos (direcciones IP de origen vs. direcciones IP de destino, puertos origen vs. puertos destino, etc.). No obstante, los resultados obtenidos distan mucho de ser fácilmente interpretables.

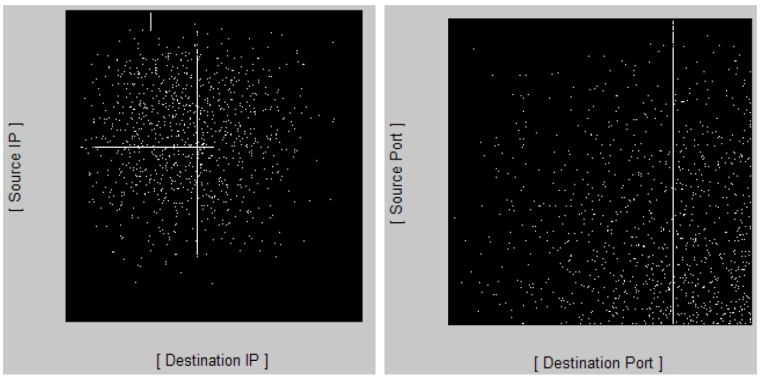


Diferentes planos analizados en la propuesta de Singh y Subramanian [135]

En 2010 Jin *et al* presentan un método para la detección de patrones anormales basado en visualización [136]. La propuesta aplica la metodología a ataques de negación de servicio distribuido (DDoS) y rastreo de puertos (*port scan*). Parte de una representación bidimensional en un gráfico de dispersión en el que presenta en el eje x las direcciones IP de destino y en el eje y las direcciones de origen. Si hay comunicación entre una dirección de origen y otra de destino, el pixel se activa, de otro modo queda a apagado. Una aportación interesante es como se transforman las direcciones IP para su traslado a los ejes correspondientes:

$$A.B.C.D \Rightarrow x = \sqrt{A \times B \times C \times D}$$

Lo que permite que el rango de valores de las direcciones se sitúe en entre 0 y 65.535.

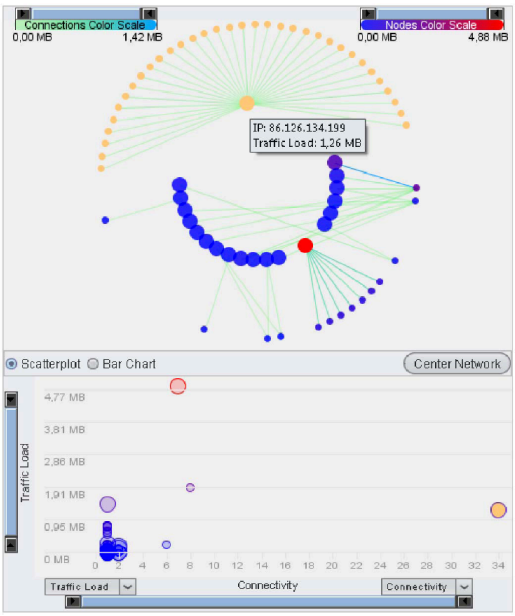


Representaciones obtenidas por Jin *et al.* [136]

En las representaciones obtenidas, los patrones de ataque aparecen como líneas rectas. Los autores exponen además un método para retirar el ruido de las imágenes e incluso la aplicación de redes neuronales para la identificación automática del evento. Los autores exponen que una desventaja de su propuesta es la transformación planteada para reducir es espacio de direcciones, ya que direcciones IP distintas aparecen en el mismo punto del eje.

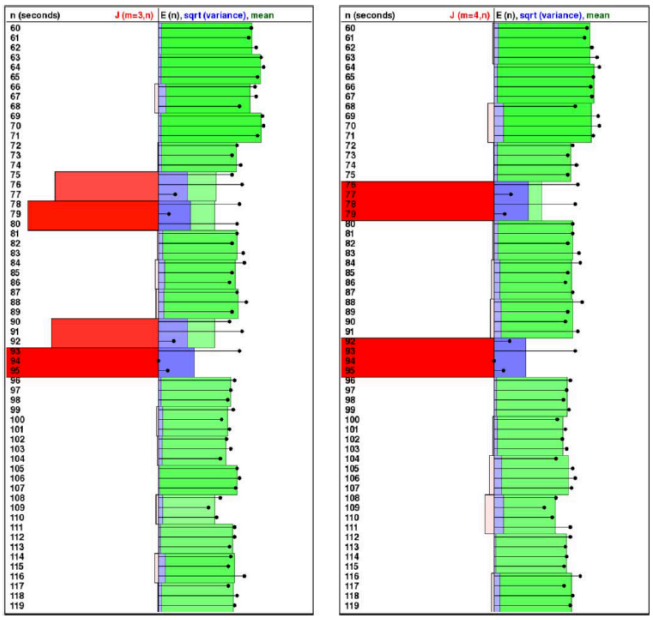
También en 2010 Barbosa y Granville [137], además de exponer una serie de preguntas a realizarse en el ámbito de la representación de datos en redes para

evaluar las representaciones obtenidas, obtienen unas representaciones combinadas de gráficos de dispersión para visualizar el tráfico cursado en las redes.



Representación extraída de Barbosa y Granville.[137]

En 2010 Celenk *et al* presentan una herramienta de detección y visualización predictiva de anomalías en redes basada en la entropía [138]. Combinan en su propuesta la entropía con el filtrado de Wiener (para eliminar el ruido de la "señal") y un predictor ARMA. Para la identificación de patrones anómalos definen un índice J que es una variación del Discriminante Lineal de Fisher modificado. Con esto elaboran una representación dinámica de la entropía (similar al box plot) y este índice J.

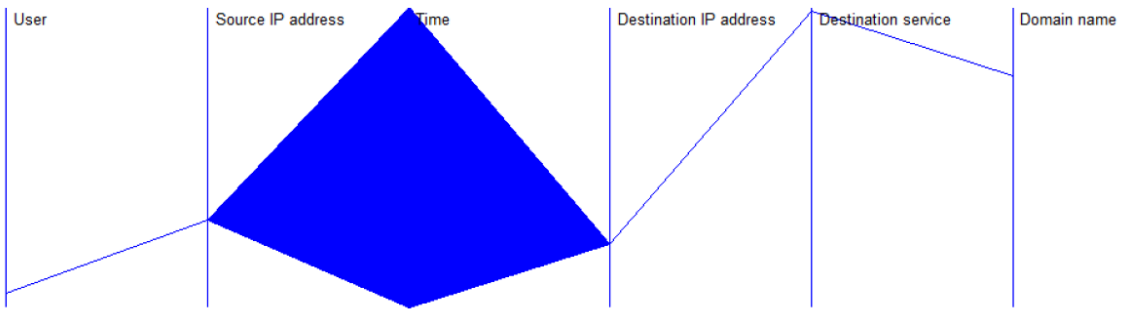


Visualización para identificación de anomalías propuesta por Celenk *et al* [138]

Las columnas de valores más elevados del índice J permiten identificar patrones anómalos, en los que además simultáneamente se observan comportamientos de la entropía igualmente diferentes.

En 2010 Lee y Kim [13] publican un artículo en el que exponen la necesidad de comprender y analizar las configuraciones de las redes de forma rápida y precisa para las operaciones y gestión de las mismas. Durante el proceso de configuración, verificación y búsqueda de problemas, las fuentes de datos con las que los gestores de redes deben operar se pueden clasificar en dos tipos: configuraciones y medidas. Dentro de las configuraciones se incluyen comandos y parámetros establecidos en los dispositivos de red. Las medidas incluyen paquetes, rutas y otras situaciones de la red, recopiladas en distintos momentos. Por lo general estos dos grupos de datos se analizan conjuntamente, dado que se encuentran fuertemente relacionados. La visualización de las medidas obtenidas ha recibido mucha atención dado que permite el análisis y comprensión de los datos obtenidos con rapidez.

En 2011 Promrit *et al* proponen la utilización de coordenadas paralelas para el análisis forense de redes [139]. El principal problema de análisis forense de redes es que los datos son abrumadores y difíciles de interpretar, ya que la mayoría están constituidos por texto o son numéricos o estadísticos. Frecuentemente las investigaciones forenses son presentadas en forma gráfica, lo que posibilita identificar rápidamente las anomalías identificadas por la forma y el color de los elementos. El objetivo del trabajo de Promrit *et al* es crear un sistema que permita elaborar relaciones multidimensionales de los registros y el tráfico de las redes para ayudar en la investigación criminal. Los autores prueban su propuesta con diversos ataques: rastreo de equipos, rastreo de puertos, ataques de negación de servicio desde una fuente única, rastreo de equipos y puertos simultáneos.



Representación de un ataque de negación de servicio desde una única dirección de origen. [139]

En 2011 Von Landesberger *et al* publican una revisión sobre el análisis visual de gráficos grandes [140]. Remarcan los autores que la representación de gráficos

“grandes” es el objetivo principal de la “analítica visual”. Sitúan el inicio del crecimiento de estas técnicas en 2005, con el libro de Thomas and Cook [106]. Von Landesberger *et al* en su revisión retoman la teoría de grafos, en la que un gráfico dirigido con vértices ponderados es también denominado “red”. En visualización de información, el término “red” es con frecuencia utilizado en un sentido más amplio denotando un grafo con atributos asociados a vértices y bordes. Estos grafos pueden también evolucionar con el tiempo, constituyendo así grafos dinámicos, en contraposición con grafos estáticos. Además los grafos pueden ser identificados por sus propiedades topológicas. En la visualización de grafos, el preprocesados de los mismos incluye frecuentemente la simplificación del grafo para reducir su tamaño, mientras se mantiene la estructura principal subyacente. Hacen los autores un detalladísimo repaso a las diferentes técnicas y métodos de representación utilizados.

En noviembre de 2012 la revista IEEE Network publica una entrega especial dedicada a la visualización en redes de ordenadores. Como exponen los editores [141], las características de las redes de ordenadores hacen difícil comunicar los resultados obtenidos de las monitorización de la red, clasificación de tráfico e identificación de tráfico malicioso o anormal. Así, los administradores de redes y analistas de seguridad requieren herramientas que les ayuden a entender, razonar y adoptar decisiones a partir de la información que los sistemas les facilitan. Para estos objetivos la visualización de información y los métodos estadísticos gráficos prometen hacer la información accesible, usable y manejable aprovechando las ventajas de la percepción visual humana. Las técnicas de visualización ayudan a los administradores de redes y analistas de seguridad a reconocer rápidamente patrones y anomalías; integrar visualmente datos de fuentes heterogéneas; ubicar en contexto eventos críticos. Y aún queda mucho por hacer. Mientras los datos continúan creciendo exponencialmente, las capacidades de visualización de los sistemas crecen de manera aparentemente lineal, lo que convierte en un reto generar análisis más fácilmente interpretables junto con sumarios más sucintos. Además de métodos para manejar mayores volúmenes de datos en aplicaciones de visualización, se precisan herramientas para procesar los datos en línea más rápidamente, permitiendo a los analistas tomar medidas sobre los eventos ocurridos en momentos más próximos a cuando estos tienen lugar. Para propósitos de seguridad, la mayoría de los datos utilizados hoy en día fueron inicialmente concebidos para otros propósitos, lo que en algunos casos hace necesario desarrollar nuevos descriptores específicamente enfocados a cuestiones de seguridad. Se precisan métodos para identificar y relacionar distintas fuentes, sean físicas o políticas de gestión. La mayor parte del potencial de sistema visual humano

permanece aún sin explotar, pendiente de más investigaciones en este campo, afirman.

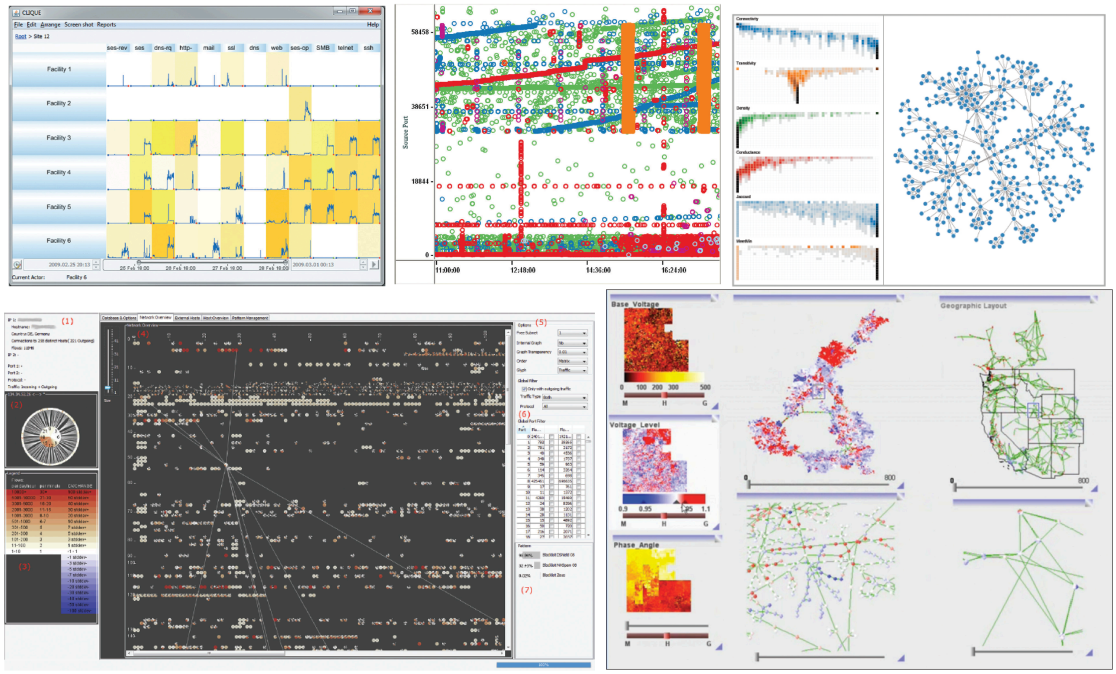
En el primer artículo de ese número especial, Harrison y Lu [142] hacen un repaso al estado del arte de la visualización en el campo de la seguridad en redes, analizando la situación actual y como avances en otros campos pueden ser aplicados en este. De hecho, afirman, la investigación en técnicas de visualización se ha prodigado en los últimos años en múltiples dominios. Esta situación parece lógica si consideramos el incremento que ha habido tanto en volumen de datos, como de su complejidad. Este incremento del volumen de información constituye un reto especialmente preocupante para los analistas de seguridad, ya que deben enfrentarse simultáneamente a este incremento de volumen y complejidad tanto en los datos a analizar como en los propios ataques. A pesar de que la seguridad es una componente fundamental de la gestión de redes, y de que la visualización se tornado en una componente también fundamental del análisis de redes, sin embargo se han producido muchos avances en la aplicación de técnicas de visualización en otros campos diferentes de la seguridad, afirman los autores. Han sido identificadas, prosiguen, limitaciones a las aplicaciones de visualización en seguridad, que están típicamente relacionadas con la imposibilidad de operar con datos en diferentes escalas e integrar diversos orígenes de datos. Por ejemplo, exponen, mientras algunas técnicas de visualización funcionan bien en términos cualitativos, como los histogramas, esas mismas técnicas presentan limitaciones en cuanto dimensionalidad. Por otro lado, visualizaciones que presentan ventajas en dimensionalidad, como las coordenadas paralelas, son casi inutilizables con grandes volúmenes de datos. Finalmente, algunas veces los propios algoritmos que se encuentran tras la representación visual pueden presentar problemas de escalabilidad. Así, aunque los histogramas presenten un buen comportamiento visual, una serie de datos de volumen elevado puede hacer la técnica prácticamente inútil.

La defensa en las redes, indican Harrison y Lu [142] reside típicamente en una variedad de sistemas que generan datos, incluyendo en ellos sistemas de detección de intrusos, registros de cortafuegos y registros de eventos del sistema, entre otros. La visualización de esta información siempre juega un papel importante. Algunos sistemas comerciales de gestión de eventos e información utilizan visualizaciones simples para facilitar a los analistas la monitorización de actividad de red. No obstante estas aproximaciones son limitadas, ya que proveen una visión de conjunto en la que solo facilitan al observador qué tipo de evento se ha producido y cuándo ha sucedido. Para responder a las preguntas de dónde, porqué y cómo el evento ha tenido lugar se

requiere más información sobre la red. En otras palabras, los sistemas comerciales actuales hacen un uso razonablemente efectivo de la visualización para la monitorización de eventos, pero no aprovechan la efectividad de la visualización para ayudar al analista a escalar, correlar y responder a los eventos. Del repaso que efectúan los autores a algunas herramientas disponibles concluyen que, efectivamente las representaciones visuales proporcionadas presentan limitaciones, por ejemplo, la mayoría hacen uso exclusivo del tráfico de red, cuando en realidad la seguridad reside en más cuestiones que el tráfico de red. Hay una clara necesidad, afirman los autores, de metáforas visuales escalables, técnicas de interacción y algoritmos que admitan datos de red heterogéneos. Exponen los autores claros parecidos entre la red eléctrica inteligente y la seguridad de las redes; por ejemplo, ambas deben ser analizadas a varias escalas, y ambas son interdependientes en el sentido de que cambios en un nodo o enlace de la red pueden afectar significativamente a otros elementos.

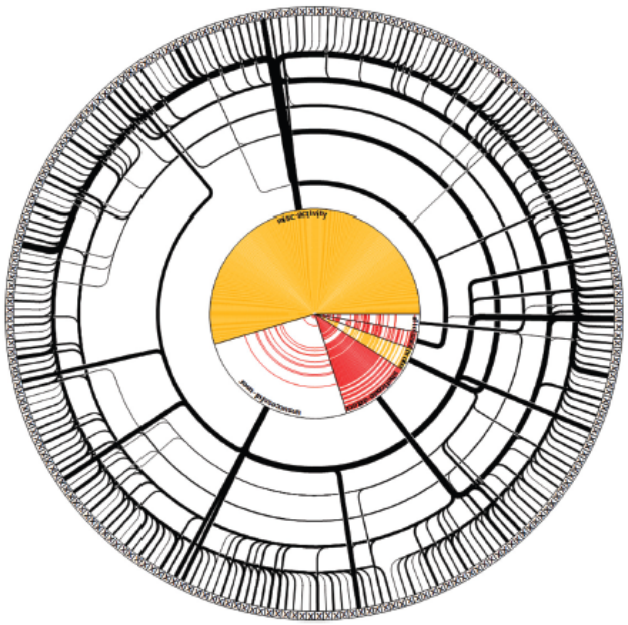
Presentan los autores varias representaciones obtenidas con herramientas disponibles en la actualidad, algunas basadas en coordenadas paralelas, representaciones matriciales, o gráficos de dispersión, entre otras, utilizando diversas métricas estadísticas.

Concluyen los autores [142] que la visualización de redes está volviéndose más difundida, no solo como herramienta para la representación de datos, sino también como importante componente en el análisis de su comportamiento.



Ejemplos de representaciones recopiladas por Harrison y Lu. [142]

En el siguiente artículo Dumas *et al* [143] proponen una herramienta, que denominan *AlertWheel*, enfocada al análisis de alertas generadas por sistemas de detección de intrusos, aplicando métodos para evitar el *clutter* visual en la representación.

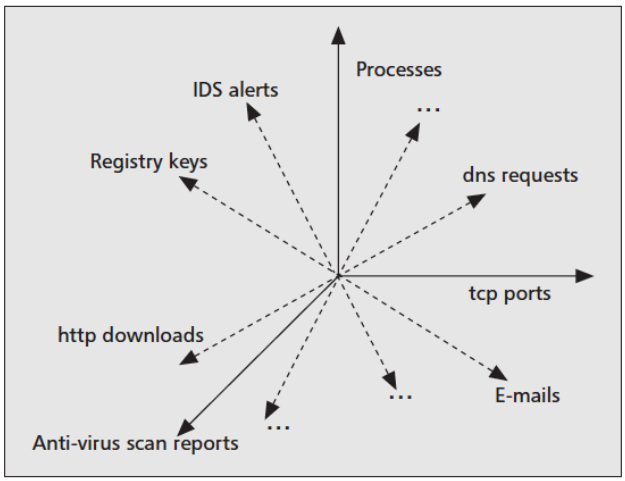


Representación obtenida con AlertWheel. [143]

AlertWheel permite la representación de gráficos bipartitos, esto es, un gráfico constituido por dos conjuntos de nodos (direcciones de origen y categorías) donde todos los vértices conectan un nodo de un conjunto y otro nodo en el otro conjunto. *AlertWheel* utiliza una representación radial y muestra la localización, tiempo y naturaleza de múltiples eventos de forma simultánea. Para evitar el desorden en la representación utiliza enlaces curvos cuidadosamente diseñados y tres técnicas diferentes de agrupación para clarificarlos. Los nodos más exteriores de la representación corresponden con rangos de direcciones IP origen, o agrupaciones de ellas. En el interior están agrupadas las categorías de alertas producidas, en las que el color representa la severidad de la alerta. Cada enlace desde un nodo exterior a la categoría se corresponde con, al menos, una alerta; el grosor del enlace es proporcional al logaritmo del número de alertas producidas. En la transición entre los nodos exteriores e interiores los enlaces agrupan a través de tantos elementos concéntricos como categorías hay en el elemento interior, de manera que a ese punto solo llega un único enlace.

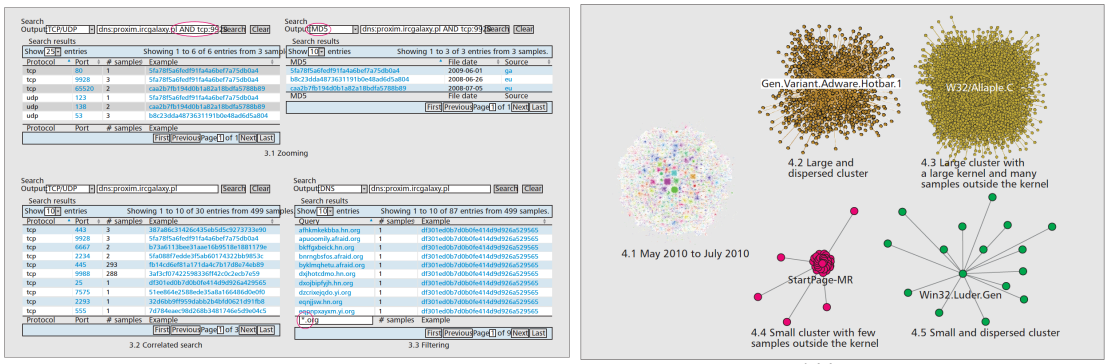
El tercer artículo elaborado por Couture *et al* [144] efectúan una interesante, a nuestro modo de ver, aportación algo diferente de las restantes, en cuanto a su objetivo. Este objetivo no es tanto la visualización de un evento o grupo de eventos en si mismo, sino la representación de lo que denominan “espacio de código malicioso”: a través de

representaciones gráficas identifican de manera explícita las relaciones entre diferentes ejemplos de código malicioso, facilitando la investigación de incidentes de seguridad y revelando las tendencias a lo largo del tiempo. Especialmente interesante, por su "familiaridad" en cuanto a su forma, es la representación que denomina "espacio del código malicioso".



Espacio multidimensional del código malicioso. [144]

Para analizar este espacio de código malicioso Couture *et al* proponen dos herramientas con objetivos diferentes: *BeAVER*, diseñada para navegar por dicho espacio y facilitar la tarea a los analistas de extraer y correlar atributos del código malicioso, y *MTR* (*Malware Threat Radar*), para representar las relaciones entre ejemplos de malware, así como su evolución a lo largo del tiempo. *BeAVER* es una herramienta basada en texto, mientras que MTR es una representación gráfica sobre el espacio multidimensional del código malicioso.



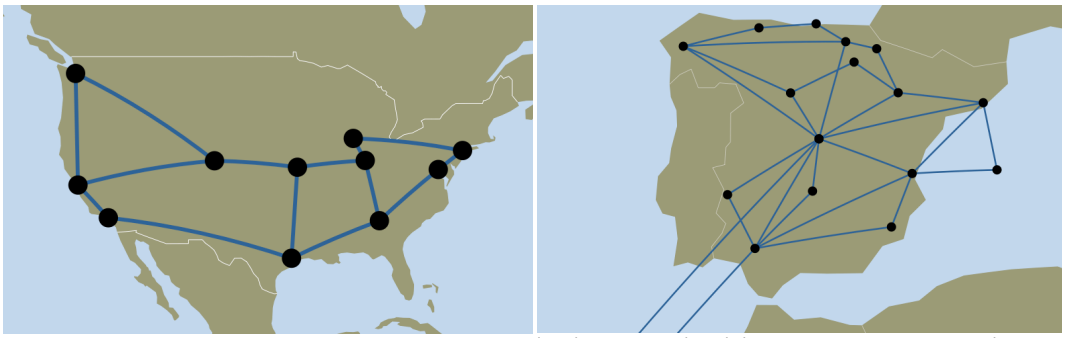
Ejemplos de resultado obtenidos con BeAVER y MTR. [144]

El siguiente artículo de este número especial es el elaborado por Knight *et al* [145] que aporta una interesante herramienta abierta para la recolección de mapas de red actuales e históricos, en un formato estandarizado, que denominan *Internet Topology Zoo*. Tras una interesante aproximación a la presentación de los mapas de red a

través de la historia de los mapas de metro, los autores se plantean cuáles son las características importantes de una red que deben representarse en una visualización y cómo deben ser presentadas. Afirman los autores que no hay para esas preguntas una única respuesta "correcta", y que esa respuesta dependerá de las prioridades de la representación que formulen los operadores de la red y el objetivo de la visualización:

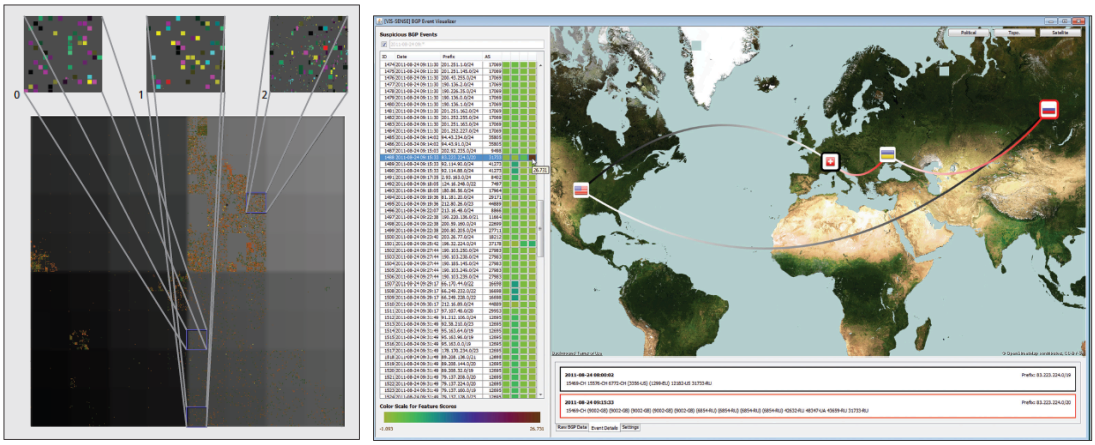
- La gestión de red, por ejemplo, superponiendo cortes y volumen de tráfico en el mapa de la red.
- La planificación de actualizaciones en la red.
- La comercialización de la red a sus clientes.

Lo interesante de este proceso es comprender las implicaciones y los resultados que se derivan de las decisiones adoptadas en esos aspectos. Como se ha indicado, el artículo se centra en la representación topológica de redes, si bien combinan información en algunos casos sobre situación de las redes, dando lugar a lo que se denominan "mapas del tiempo en la red" (*network weather maps*).



Dos ejemplos de *Internet Topology Zoo*: Abilene (izq.) y RedIris (der.) (de www.topology-zoo.org)

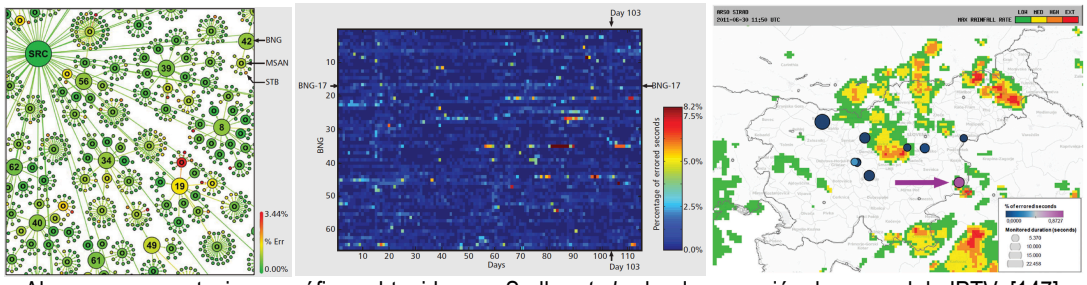
El penúltimo artículo de la entrega corresponde a Biersack *et al* [146], en el que proponen una visualización del protocolo de enrutamiento en Internet BGP. Tras analizar las propuestas existentes para la visualización de BGP, los autores presentan un ejemplo de visualización sobre un incidente de secuestro de prefijo, demostrando como las herramientas de representación pueden ayudar a los analistas. Los autores analizan un total de nueve herramientas de visualización relacionadas con labores de análisis del protocolo BGP, exponiendo las técnicas de visualización utilizadas en cada caso, sus características principales y casos de uso de cada propuesta.



Ejemplos de visualización de eventos relacionados al protocolo en enrutamiento BGP. [146]

Concluyen los autores que la disponibilidad de herramientas eficientes de monitorización de las redes es de máxima importancia. Los administradores de redes son desafiados actualmente por el gran volumen de datos que deben analizar, especialmente cuando se trata de datos BGP y su seguimiento. El artículo describe los métodos y herramientas que se están desarrollando para el seguimiento de protocolo BGP en el marco de VIS-SENSE, un proyecto europeo de investigación que se centra en la explotación de la analítica visual para mejorar las técnicas forenses en Internet.

El último artículo de esta entrega especial es el elaborado por Sedlar *et al* [147] y en el que se expone una metodología para la monitorización de la calidad de servicio en una red de televisión a través de Internet (IPTV). La información en tiempo real es recopilada a través de agentes distribuidos en la red y agregada en un nodo central. Los análisis que proponen no se realizan en tiempo real, y utilizan herramientas de análisis genéricas tales como Matlab y Tableau. La representación visual de esta información y su enriquecimiento a través de fuentes externas permiten una comprensión más profunda de la red analizada para posibilitar un análisis de causa final de las anomalías en la misma, a través del descubrimiento de correlaciones entre eventos producidos. Así, por ejemplo, combinan datos sobre la situación de la red de comunicaciones con la lluvia. Las representaciones son de tipo topológico, con información geoposicional o no, y también de tipo matricial.

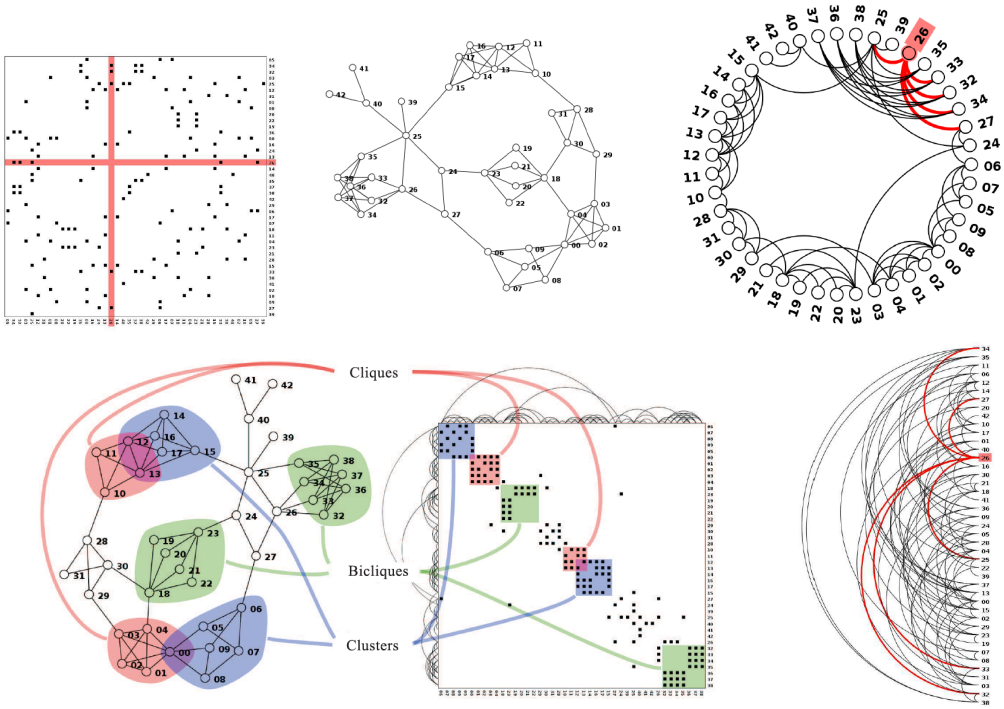


Algunas representaciones gráficas obtenidas por Sedlar *et al* sobre la operación de una red de IPTV. [147]

En 2012 Davey *et al* publican un capítulo sobre la aportación de la analítica visual en Internet y su seguridad [148]. Con referencia al concepto de *Big Data*, afirman los autores que los datos en sí mismos no suelen ser importantes, que lo verdaderamente relevante es la información contenida en los datos. Así, la sobrecarga de datos puede acarrear consecuencias negativas, ya es posible perderse en los datos debido a que los datos con los que se trabaja sean irrelevantes, se procesen de manera inapropiada, o se presenten de manera inapropiada. Los autores plantean aplicaciones de analítica visual en diversos campos de redes de telecomunicación:

- Planificación y verificación de infraestructuras.
- Seguridad en redes
- Monitorización en tiempo real.

En 2012 McGuffin publica un tutorial sobre algoritmos para visualización de redes [149]. En su artículo McGuffin repasa diversas visualizaciones de redes, principalmente limitados a sus aspectos topológicos. Especialmente interesante es la representación de matrices de proximidad: dados dos nodos *i* y *j*, las posición (i,j) y (j,i) de la matriz contiene información sobre el vértice o vértices entre ambos nodos. El caso más sencillo es que el valor sea una variable booleana que indique si existe o no vértice entre los nodos, pero podría contener otro tipo de información. Si el gráfico es "no dirigido", la matriz es simétrica. Si el gráfico tiene dirección, la matriz será asimétrica. Veremos más adelante aproximaciones más concretas a este aspecto.



Algunos ejemplos de visualización de redes propuestos por McGuffin. [149]

Finalmente el autor expone algunas consideraciones sobre aplicaciones más avanzadas de las propuestas inicialmente. Así comenta la posibilidad de crear gráficos a partir de datos multidimensionales, utilizando diagramas de dispersión y coordenadas paralelas, y gráficos dinámicos, que muestren la dimensión temporal de las relaciones entre nodos.

En 2012, en último lugar por su importancia para nuestro trabajo, Liu *et al* presentan una ponencia [150] sobre la monitorización de gráficos de actividad de red (los TAG ya analizados por Jin en 2009 [129]) mediante aproximaciones matriciales de bajo rango. Afirman los autores que el análisis de tráfico juega un papel esencial en la seguridad y gestión de las redes, lo que cubre un amplio rango de temas de investigación, tales como mediciones, clasificación y detección de anomalías, entre otras. Para comprender, analizar y modelar sistemas en red complejos y de rápida evolución, así como el comportamiento de sus usuarios, se precisan metodologías efectivas y nuevas y apropiadas a las características del tráfico, para capturar patrones de comunicación a lo largo de toda la red y las complejas y a veces sutiles estructuras y relaciones que ocultan. Jin en 2009 [129] ya aplicó, como hemos visto anteriormente, métodos de "trifactorización ortogonal de matrices no negativas" (tNMF) para descubrir las estructuras de comunidades en varios gráficos de actividad de red en aplicaciones conocidas. Monitorizar TAGs en redes de gran tamaño y alta velocidad, como Internet, presenta numerosos retos derivados de los elevados volúmenes de tráfico a manejar. Si bien para soslayar esa situación es común emplear un muestreo de paquetes sobre el tráfico, algunos autores, dicen en su ponencia Liu *et al,* prefieren analizar estos gráficos con aproximaciones de bajo rango. Como se ha indicado, los métodos tNMF, que aproximan los TAGs por tres matrices de bajo rango no negativas, permiten descubrir patrones de comunicación. Sin embargo los algoritmos existentes para esa descomposición presentan elevadas cargas computacionales, por lo que son difícilmente aplicables para monitorizar TAG en redes reales.

Un gráfico de actividad de red, o TAG (*Traffic Activity Graph*) representa los patrones de comunicación entre nodos a lo largo de una red. A veces también se les denomina Gráficos de Dispersión de Tráfico (*Traffic Dispersion Graph*) [130]. En Internet, continúan los autores, un equipo envía una secuencia de paquetes con información para comunicarse con otro equipo, y este conjunto de paquetes constituye un flujo (o *flow*) desde el origen al destino. Cuando la comunicación concluye, el flujo se termina. Un equipo puede crear simultáneamente múltiples flujos para comunicarse con

diferentes equipos. En un TAG, un nodo representa un equipo distinto, y un vértice dirigido indica que hay al menos un flujo desde el origen al destino.

Así un TAG se define como un gráfico bipartito $G=\{V_o \times V_d, E\}$ donde V_o / V_d son los conjuntos de nodos que representan los orígenes / destinos en la red, y E es el conjunto de vértices correspondientes a los flujos entre nodos. Un vértice dirigido (i,j) de E indica que existe al menos un flujo desde el origen i en V_o al destino j en V_d. Dado un TAG así definido se puede definir su matriz de proximidad A_{mxn} como $a_{ij}=1$ si hay al menos un flujo entre el origen i y el destino j. El problema de extraer información sobre estructuras de agrupaciones puede ser formulado como un problema de *clustering* sobre dicha matriz de proximidad.

Las estructuras en TAGs pueden ser descubiertas mediante métodos de descomposición matricial, como por ejemplo los mencionados tNMF que descomponen aproximadamente la matriz A en tres matrices no negativas de bajo rango [131]

$$A \approx FSG$$

Siendo F , S , G matrices positivas m x k, k x l, l x n, respectivamente. El criterio de optimización de la aproximación se define como

$$\min_{\mathbf{F}\geq 0,\mathbf{S}\geq 0,\mathbf{G}\geq 0} \left\{ J\left(\mathbf{F},\mathbf{S},\mathbf{G}\right) = \left\|\mathbf{A}-\mathbf{FSG}\right\|_{\mathrm{F}} \right\}$$

Con las restricciones $F^TF=I$ y $G^TG=I$, siendo $||.||_F$ la norma de Frobenius e I la matriz identidad y k,l <<min(m,n). Como ya vimos cuando comentamos el artículo de Ding *et al* [131].

El procedimiento para calcular la aproximación anterior es iterativo, las matrices F,G,S se inicializan con números aleatorios positivos y son actualizadas con arreglo a lo indicado en [131] hasta que el error cuadrático relativo sea menor que un umbral preestablecido.

$$\mathrm{ECR} = \frac{\left\|\mathbf{A}-\mathbf{FSG}\right\|_{\mathrm{F}}^2}{\left\|\mathbf{A}\right\|_{\mathrm{F}}^2}$$

Exponen también los autores la posibilidad de utilizar para descomponer la matriz ***A*** la descomposición en valores singulares $\mathbf{A} = \mathbf{UDV}^T$ que puede utilizarse, como bien sabemos, para obtener una aproximación de **A** de menor rango.

$$\mathbf{A}_r = \mathbf{U}_r\mathbf{D}_r\mathbf{V}_r^T$$

donde la matriz $\boldsymbol{A}_r$ minimiza la norma de Frobenius de los errores

$$\left\|\mathbf{A} - \mathbf{A}_r\right\|_F = \min_{\text{rank}(\mathbf{X})\le r} \left\{\left\|\mathbf{A} - \mathbf{X}\right\|_F\right\}$$

Otro método para descomponer la matriz es la descomposición CUR [124], ya comentada anteriormente en el trabajo de Tong [123], en la que las matrices **C** y **R** que forman la descomposición $\mathbf{A} \approx \mathbf{CUR}$ se obtienen de un muestreo de las columnas y filas de **A**, respectivamente, y **U** es una matriz que permite reconstruir aproximadamente **A**. Esta aproximación preserva los vectores singulares en las filas y columnas muestreadas.

El método que proponen Liu *et al* está basado en ambos métodos previos. Así proponen los autores una descomposición, que denominan CMR, utilizando columnas y filas muestreadas uniformemente de la matriz ***A***. Si la j-ésima columna de ***A*** es muestreada, entonces $\mathbf{c}_{*j} = \rho^{-1}\,\mathbf{a}_{*j}$. En otro caso, $\mathbf{c}_{*j} = 0$. La relación entre **C** y **A** puede ser escrita como **C** = **AQ**, con **Q** es una matriz diagonal tal que:

$q_{ii} = \rho^{-1}$ con una probabilidad ρ y 0 en otro caso

$q_{ij} = 0 \quad \forall\, i \neq j$

De esta relación se deduce que $E(\mathbf{Q}) = \mathbf{I}$ ya que $E(q_{ii}) = \rho^{-1}\,\rho = 1$, y $E(q_{ij}) = 0$

Similarmente se puede argumentar para la relación entre **R** y **A**, **R** = **PA** y E(**P**) = 1.

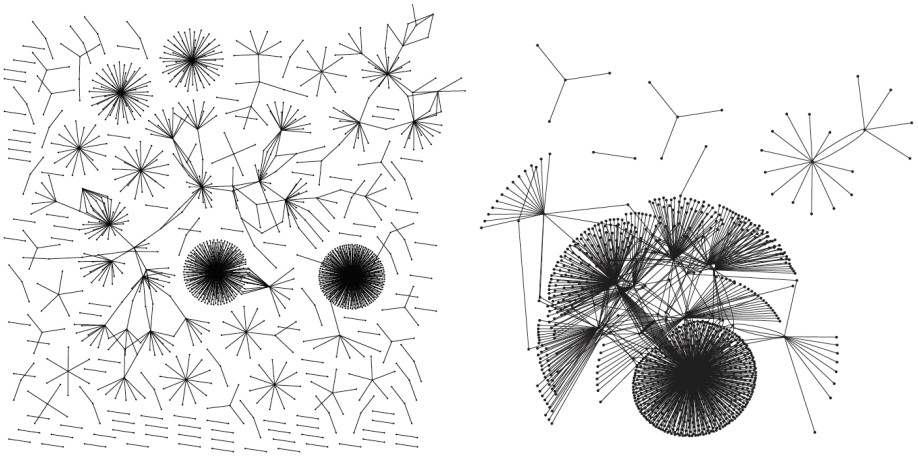
Definimos la matriz **M** como $\mathbf{M} = (\mathbf{PAQ})^-$ donde la matriz **PAQ** incluye los elementos comunes en las matrices **C** y **R** que provienen del flujo con un origen y destino específico con probabilidad ρ. El operador $(\,)^-$ indica la pseudo-inversa de Moore-Penrose de la matriz, esto es

$$\mathbf{X}^- = \mathbf{V}_X\Sigma_X^{-1}\mathbf{U}_X^T$$

Donde $\mathbf{U}_x\mathbf{\Sigma}\mathbf{V}^T$ es la DVS de **X**.

Por último, se calcula la aproximación de bajo rango de **A** como $\mathbf{A}_k=\mathbf{C}\mathbf{M}_k\mathbf{R}$ donde $\mathbf{M}_k$ es la aproximación de bajo rango de **M**.

Los autores demuestran que, dada una frecuencia de muestreo ρ la matriz $\mathbf{A}_k$ es próxima a la matriz original **A**, con la norma de Frobenius, con una elevada probabilidad y se puede por lo tanto utilizar como una aproximación de bajo rango de **A**.



Ejemplos de TAGs obtenidos con la propuesta de Liu *et al*. Izq. http TAG. Der. TAG trafico malicioso. [150]

Así la propuesta de Liu *et al* combina el muestreo de paquetes con las aproximaciones de bajo rango para monitorizar patrones de comunicación y agrupaciones en TAG, que proponen aplicar en el futuro a la detección de anomalías y ataques a redes.

Ya en el año 2013, en un número especial de la revista China Communications sobre el tema “Gestión y Visualización de datos de usuarios y redes”, Pfeffer publica un artículo sobre los fundamentos de la visualización de redes de comunicaciones [151]. Los autores exponen la proliferación actual de representaciones visuales de redes de comunicación, no exclusivamente de telecomunicaciones. Citan como ejemplo las visualizaciones de relaciones personales en redes sociales. (En este contexto, redes “sociales”, sitúan la aparición de las primeras representaciones de “sociogramas” en el año 1934 [152].) Los datos recogidos de las redes de comunicación son multivariantes, afirman los autores. Además de la propia información sobre “relaciones”, pueden existir datos adicionales disponibles, u otros pueden ser calculados. Representar esta información en una imagen puede ser un traducción efectiva de información a un sistema de elementos visuales, tales como, posición, tamaño, color, saturación, orientación, forma y textura. Nuestros cerebros están entrenados para procesar esta información incluso de manera no intencionada. El reto lo constituye el utilizar “correctamente” estos elementos para transmitir información.

Los autores afirman que la tarea más importante cuando se representa una red es la determinación del lugar a ubicar los nodos de la red. Una buena representación es aquella en la que la posición de los nodos revele la estructura de la red de una manera intuitiva. Por otro lado cualquier representación (salvo aquellas triviales) presentará cierta distorsión ocasionada por la "proyección" de datos multidimensionales en representaciones de dos o tres dimensiones. En general las redes de comunicación son redes de "dos modos" ya que se representan dos diferentes tipos de nodos.

También en 2013, Hu *et al* publican un artículo que combina la visualización de redes para el estudio de topologías, con el análisis espectral [153]. Afirman los autores que si bien la representación directa de la topología de una red puede ser muy eficaz, muchas de las características topológicas de una red pueden expresarse como una función explícita de sus autovalores (espectro) y autovectores.

Un grafo o una red G(V,E) está constituido por un conjunto de n nodos V conectados por un conjunto de m enlaces E. Se puede representar G como una matriz de proximidad "simétrica" (sic) $\mathbf{A}=(a_{ij})_{nxn}$. Los autores se centran en analizar matrices binarias, donde $a_{ij}=1$ si los nodos i y j están interconectados, o $a_{ij}=0$ si no lo están.

El análisis espectral del gráfico trata de los espectros de los nodos, autovalores y autovectores. Existe una relación intrínseca entre las características combinatorias de un grafo y las propiedades algebraicas de su matriz de proximidad: para un grafo con k agrupaciones, las coordenadas del nodo u en el espacio k-dimensional, indica la probabilidad de la vinculación del nodo u a dichas k agrupaciones. Los puntos correspondientes a un nodo con una comunidad forman una línea que atraviesa el origen en el espacio k-dimensional. Los nodos en las k agrupaciones forman k líneas cuasiortogonales en el espacio espectral. Veremos esta casuística más adelante con detalle.

Sun *et al* publican en 2103 una revisión de aplicaciones y técnicas de analítica visual [154]. Los pasos clave en la analítica visual incluyen transformación de los datos, mapeado visual, análisis basado en modelos, e interacciones del usuario. Identifican así mismo cinco categorías de aplicaciones: espacio-temporales; multivariantes; texto; gráficos y redes; y otras aplicaciones. Dentro de las aplicaciones con datos multivariantes, establecen dos categorías, basadas en métodos de proyección y basadas en métodos visuales. En las técnicas de proyección sitúan el escalado multidimensional y el análisis de componentes principales como técnicas más importantes, pero citando otras más novedosas como el LAMP (*Local Affine*

Multidimensional Projection) propuesta por Joia *et al* [155] o el propuesto por Turkay *et al* [156] también en 2011, que consiste en un análisis interactivo realizado iterativamente sobre dos espacio, el de los ítems y el de las dimensiones del espacio, posibilitando un análisis conjunto de ítems y dimensiones, utilizando el Análisis de Componentes Principales. Dentro de los métodos visuales sitúan, por ejemplo, las coordenadas paralelas y las representaciones gráficas más tradicionales, entre las que colocan las representaciones matriciales.

4. MÉTODOS MULTIVARIANTES APLICABLES AL ANÁLISIS DEL TRÁFICO DE REDES

4. MÉTODOS MULTIVARIANTES APLICABLES AL ANÁLISIS DEL TRÁFICO DE REDES

4.1. LOS MÉTODOS BIPLOT

4.1.1. Introducción a los métodos Biplot

Un **Biplot** es una representación gráfica de datos multivariantes. Formulados por primera vez por Gabriel en 1971 [157] permiten representar una matriz **X** mediante un vector para cada fila de **X** y otro vector para columna de **X**, de manera que los elementos de la matriz **X** sean los productos interiores (escalares) de los vectores que representan las correspondientes filas y columnas.

Un **Biplot** para una matriz de datos **X,** de 'f' filas y 'c' columnas, es una representación gráfica mediante unos vectores, denominados marcadores $\mathbf{a_1}$, $\mathbf{a_2}$,…,$\mathbf{a_f}$ para las filas de **X** y $\mathbf{b_1}$, $\mathbf{b_2}$, … $\mathbf{b_c}$ para las columnas de **X**, de forma que el producto interno/escalar $\mathbf{a_i}^T\mathbf{b_j}$ corresponda con el elemento x_{ij} de la matriz original.

$$x_{ij} = \mathbf{a}_i^T\mathbf{b}_j$$

Por su relevancia en la interpretación gráfica posterior es oportuno traer a colación que aplicando la definición del producto escalar de dos vectores se tiene que:

$$x_{ij} = \mathbf{a_i^T b_j} = \text{proyeccion}\left(\mathbf{a}_i / \mathbf{b}_j\right)\left|\mathbf{b}_j\right|$$

Si consideramos los marcadores $\mathbf{a_1}$, $\mathbf{a_2}$,…,$\mathbf{a_f}$ como las filas de una matriz **A** y los marcadores $\mathbf{b_1}$, $\mathbf{b_2}$, … $\mathbf{b_c}$ como filas de una matriz **B** la relación anterior se puede escribir como:

$$\mathbf{X} = \mathbf{A}\mathbf{B}^T$$

La factorización anterior no es única, y puede ser reemplazada por cualquier otra a través de la expresión:

$$\mathbf{X} = \left(\mathbf{A}\mathbf{R}^{\mathbf{T}}\right)\left(\mathbf{B}\mathbf{R}^{-1}\right)^T$$

Siendo **R** cualquier matriz no singular. Las transformaciones $\mathbf{A} \rightarrow \mathbf{AR}^T$ y $\mathbf{B} \rightarrow \mathbf{BR}^{-1}$ pueden ser interpretadas como una rotación, escalado y posiblemente reflexión, de la representación **Biplot**.

La Descomposición en Valores Singulares de la matriz **X** puede ser expresada como [158]

$$\mathbf{X} = \mathbf{U}\,\Sigma\,\mathbf{V}^T$$

Donde **U** es la matriz cuyas columnas son los vectores propios de $\mathbf{XX}^T$.

$\mathbf{\Sigma}$ es la matriz diagonal de valores singulares λ_i de **X.**

V es la matriz cuyas columnas son los vectores propios de $\mathbf{X}^T\mathbf{X}$.

Para que esta Descomposición en Valores Singulares sea única y por lo tanto útil para la inspección de las relaciones entre las filas y entre las columnas de la matriz **X** es preciso imponer una métrica que establezca la unicidad de la factorización resultante. Para asegurar esa unicidad se establece que debe cumplirse que **U** y **V** sean ortonormales, esto es $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$, siendo **I** la matriz identidad.

Con lo que obtendríamos:

$$\mathbf{X} = \mathbf{AB}^T = \mathbf{U}\,\Sigma\,\mathbf{V}^T$$

Además, esta Descomposición en Valores Singulares [158] de la matriz **X** puede también expresarse como

$$\mathbf{X} = \sum_{n=1}^{r} \lambda_n \mathbf{u}_n \mathbf{v}_n^T$$

Lo que nos permitiría obtener, si fuese necesario, una aproximación de menor rango (s<r) de la matriz **X** [159]:

$$\mathbf{X}_{(s)} = \sum_{n=1}^{s} \lambda_n \mathbf{u}_n \mathbf{v}_n^T$$

Que en forma matricial se expresaría como

$$\mathbf{X}_{(s)} = \mathbf{A}_{(s)}\mathbf{B}_{(s)}^T = \mathbf{U}_{(s)}\,\Sigma_{(s)}\,\mathbf{V}_{(s)}^T$$

Si la matriz **X** fuese de rango r, esta no podría ser representada de manera exacta por un **Biplot** de dimensión menor que su rango r. No obstante, si la matriz **X** pudiese ser aproximada con una cierta "calidad" por una matriz de rango s<r, el Biplot de esta última matriz sí que podría ser útil para analizar la matriz original **X**. Como hemos indicado, en este supuesto los productos escalares de los marcadores filas y de los marcadores columnas nos permitirían obtener aproximaciones de los elementos de la matriz original **X,** pero no su valor exacto.

La bondad de ajuste global de esta aproximación puede ser calculada mediante la expresión

$$\rho^2_{(s)} = \left(\sum_{m=1}^{s} \lambda_m^2\right) \Big/ \left(\sum_{n=1}^{r} \lambda_n^2\right)$$

Con λ_i los valores singulares de **X.** Si esta bondad de ajuste es próxima a 1, la aproximación de la matriz **X** obtenida puede considerarse una buena representación de la matriz original **X**.

$$\mathbf{X}_{(s)} \approx \mathbf{X}$$

En una matriz de rango s=2, los marcadores $\mathbf{a_1}$, $\mathbf{a_2}$,…,$\mathbf{a_f}$ y $\mathbf{b_1}$, $\mathbf{b_2}$, … $\mathbf{b_c}$ son vectores de dimensión 2. Estos f+c vectores pueden ser representados en un plano, permitiendo obtener una representación de los f x c elementos de la matriz **X** mediante los productos escalares de los correspondientes marcadores. Esta representación es denominada **Biplot**.

Como ya se ha expuesto, los marcadores fila y columna de la representación **Biplot** pueden ser obtenidos a partir de la Descomposición en Valores Singulares (DVS) de la matriz de datos original:

$$\mathbf{X} = \mathbf{A}\mathbf{B}^T = \mathbf{U}\,\Sigma\,\mathbf{V}^T$$

De esta fórmula podemos obtener las diferentes expresiones para los marcadores fila y columna de los diferentes **Biplot**.

	Coordenadas fila **A**	Coordenadas columna **B**
GH-Biplot	$\mathbf{U = G}$	$\mathbf{V\Sigma = H}$
JK-Biplot	$\mathbf{U\Sigma = J}$	$\mathbf{V = K}$
HJ-Biplot	$\mathbf{U\Sigma = J}$	$\mathbf{V\Sigma = H}$
Caso general	$\mathbf{U\Sigma^{1-\alpha}}$	$\mathbf{V\Sigma^{\alpha}}$
SQRT-Biplot	$\mathbf{U\Sigma^{1/2}}$	$\mathbf{V\Sigma^{1/2}}$

Los GH-Biplot y JK-Biplot fueron formulados por Gabriel en 1971 [157] mientras que el HJ-Biplot lo fue por Galindo en 1986 [160], el cuarto Biplot se corresponde con una generalización de los dos primeros, en la que el caso particular α= ½ se denomina SQRT-Biplot o Biplot simétrico [161].

4.1.2. GH-BIPLOT

Cuando en la factorización de la matriz **X** optamos por la asociación

$$\mathbf{X=U\Sigma V^T} \qquad \mathbf{G=U} \quad \mathbf{H=V\Sigma} \qquad \mathbf{H^T=\Sigma V^T} \qquad \mathbf{X=GH^T}$$

para los marcadores fila **G** y columna **H**, respectivamente, estaríamos obteniendo en denominado GH-Biplot o CMP-Biplot (*Column Metric Preserving*).

Recordemos la restricción impuesta para garantizar la unicidad de la solución

$$\mathbf{G^TG = U^TU = I}$$

lo que nos lleva a que

$$\mathbf{X^TX = HH^T}$$

Esto es, si **X** es una matriz de F filas y C columnas

$$\mathbf{X} = \left(x_{ij}\right) = \left(\begin{array}{c|c|c} x_{11} & \cdots & x_{1C} \\ \hline x_{21} & \cdots & x_{2C} \\ \hline \vdots & & \vdots \\ \hline x_{F1} & \cdots & x_{FC} \end{array}\right)$$

Tenemos que:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \left(\begin{array}{c|c|c} x_{11} & \cdots & x_{F1} \\ \hline x_{12} & \cdots & x_{F2} \\ \hline \vdots & & \vdots \\ \hline x_{1C} & \cdots & x_{FC} \end{array}\right)\left(\begin{array}{c|c|c} x_{11} & \cdots & x_{1C} \\ \hline x_{21} & \cdots & x_{2C} \\ \hline \vdots & & \vdots \\ \hline x_{F1} & \cdots & x_{FC} \end{array}\right)$$

Pero ¿qué son cada uno de los elementos de la matriz $\mathbf{X}^{\mathrm{T}}\mathbf{X}$?

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)_{ij} = \mathbf{X}_{\bullet i}\mathbf{X}_{\bullet j}$$

Esto es, no son más que los productos escalares entre las columnas de la matriz de los datos originales:

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)_{ij} = \mathbf{X}_{\bullet i}\mathbf{X}_{\bullet j} = \sum_{k=1}^{F} X_{ki}X_{kj}$$

Por lo tanto los productos escalares entre los marcadores columna reproducen los productos escalares entre las columnas de la matriz original, esto es preservan la métrica (euclídea) entre las columnas (*Column Metric Preserving*).

Consideremos la matriz **X** de 'f' filas y 'c' columnas, esto es 'f' unidades taxonómicas y 'c' variables, supongamos que las columnas han sido centradas (media nula), entonces

$$\mathbf{S} = \frac{1}{f-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$$

en donde **S** es la matriz de varianzas y covarianzas de las 'c' variables estudiadas.

Volviendo a retomar la descomposición en valores singulares de la matriz **X** tenemos que:

$$\mathbf{X} = \left(\mathbf{p_1},\ldots,\mathbf{p_r}\right)\left(\lambda_1\mathbf{q_1},\ldots,\lambda_r\mathbf{q_r}\right)^{\mathrm{T}}$$

que con la factorización anterior nos permite obtener los marcadores

$$\mathbf{G} = \left(\mathbf{p_1},\ldots,\mathbf{p_r}\right)\sqrt{f-1}$$

$$\mathbf{H} = \left(\lambda_1\mathbf{q_1},\ldots,\lambda_r\mathbf{q_r}\right)\Big/\sqrt{f-1}$$

que se corresponden con el denominado **Biplot de componentes principales**, en el que los productos escalares entre los marcadores columna **H** reproducen la estructura de las varianzas y covarianzas entre variables, que son los productos escalares de las columnas de **X**.

Podemos considerar la aproximación de rango 2 $\mathbf{X}_{(2)}$ a efectos de obtener el GH-Biplot, eligiendo:

$$\mathbf{G}_{(2)} = \left(\mathbf{p}_1, \mathbf{p}_2\right)\sqrt{f-1}$$

$$\mathbf{H}_{(2)} = \left(\lambda_1\mathbf{q}_1, \lambda_2\mathbf{q}_2\right)\Big/\sqrt{f-1}$$

Luego

$$\mathbf{X} \approx \mathbf{G}\mathbf{H}^{\mathbf{T}}$$

$$\mathbf{Y}\mathbf{S}^{-1}\mathbf{Y}^{\mathbf{T}} \approx \mathbf{G}\mathbf{G}^{\mathbf{T}}$$

$$\mathbf{S} \approx \mathbf{H}\mathbf{H}^{\mathbf{T}}$$

Esto implica a su vez que

$$y_{ij} \approx \mathbf{g}_i^{\mathbf{T}}\mathbf{h}_j$$

4.1.2.1. Propiedades del GH-BIPLOT

El GH-Biplot presenta diversas propiedades [157], [160], [162]–[164]:

Propiedades de los marcadores columna (H)

a) Los productos escalares de las columnas de **X** coinciden con los productos escalares de los marcadores **H**.

$$\mathbf{X}^{\mathbf{T}}\mathbf{X}=(1)=[\mathbf{G}\mathbf{H}^{\mathbf{T}}]^{\mathbf{T}}[\mathbf{G}\mathbf{H}^{\mathbf{T}}]=(2)=\mathbf{H}\mathbf{G}^{\mathbf{T}}\mathbf{G}\mathbf{H}^{\mathbf{T}}=(3)=\mathbf{H}\mathbf{U}^{\mathbf{T}}\mathbf{U}\mathbf{H}^{\mathbf{T}}=(4)=\mathbf{H}\mathbf{H}^{\mathbf{T}}$$

(1) $\mathbf{X}=\mathbf{G}\mathbf{H}^{\mathbf{T}}$
(2) $(\mathbf{A}\mathbf{B})^{\mathbf{T}}=\mathbf{B}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}$
(3) $\mathbf{G}=\mathbf{U}$
(4) $\mathbf{U}^{\mathbf{T}}\mathbf{U}=\mathbf{I}$

Si consideramos los vectores columna de **X** como $\mathbf{x_j}$ tendremos que $\mathbf{x_j}=\mathbf{Gh_j}$

$\mathbf{x_j^T x_k}=(1)=(\mathbf{Gh_j})^T(\mathbf{Gh_k})=(2)=\mathbf{h_j^T G^T G h_k}=(3)=\mathbf{h_j^T U^T U h_k}=(4)=\mathbf{h_j^T h_k}$

(1) $\mathbf{x_j}=\mathbf{Gh_j}$
(2) $(\mathbf{AB})^T=\mathbf{B^T A^T}$
(3) Por $\mathbf{G=U}$
(4) Por $\mathbf{U^T U=I}$

b) Si hacemos que (5) $\mathbf{G}=(n-1)^{0.5}\mathbf{U}$ y $\mathbf{H}=(n-1)^{-0.5}\mathbf{V\Sigma}$ (Biplot de componentes principales) se sigue cumpliendo que $\mathbf{X=U\Sigma V^T=GH^T}$. Si consideramos que **X** es una matriz centrada por columnas (=variables), esto es que la media de las columnas es cero, y tenemos que **S** es la matriz de varianzas y covarianzas muestrales de **X**, sería $\mathbf{S=X^T X}/(n-1)$, con n = nº uds. taxonómicas = filas de **X**

$\mathbf{S=X^T X}/(n-1)=(1)=[\mathbf{GH^T}]^T[\mathbf{GH^T}]/(n-1)=(2)=\mathbf{HG^T GH^T}/(n-1)=(5)=$

$=\mathbf{H}(n-1)^{0.5}\mathbf{U}^T(n-1)^{0.5}\mathbf{U}\,\mathbf{H^T}/(n-1)=\mathbf{H}(n-1)^{0.5}(n-1)^{0.5}\mathbf{U^T U H^T}/(n-1)=(4)=\mathbf{HH^T}$

Esto es, $s_{jk}=\mathbf{h_j^T h_k}$ (recordemos que $\mathbf{x_j}$= vector columna j-ésima de **X**)

Si el factor de escala introducido en (5) no se aplica y por lo tanto estamos en presencia de un GH-Biplot habitual, y con **X** centrada por columnas, se sigue cumpliendo lo anterior, excepto por un factor de escala:

$\mathbf{S=X^T X}/(n-1)=(1)=[\mathbf{GH^T}]^T[\mathbf{GH^T}]/(n-1)=(2)=\mathbf{HG^T GH^T}/(n-1)=(3)=$

$=\mathbf{HU}^T\mathbf{UH^T}/(n-1)=(4)=\mathbf{HH^T}/(n-1)$

Esto es $\mathbf{HH^T}=(n-1)\mathbf{S}\propto\mathbf{S}$ (con n = nº uds. Taxonómicas = nº filas de **X**)

c) La longitud al cuadrado de los marcadores columna $\mathbf{h_j}$ (=fila j-ésima de **H**) aproxima la varianza de la variable $\mathbf{X^T_j}$ (=fila j-ésima de $\mathbf{X^T}$), por lo que la longitud aproxima la desviación estándar.

$$\mathbf{S} \propto \mathbf{H}\mathbf{H}^{\mathsf{T}} \Rightarrow \begin{bmatrix} s_{11} & \cdots & s_{1j} \\ \vdots & \ddots & \vdots \\ s_{i1} & \cdots & s_{ij} \end{bmatrix} \propto \begin{bmatrix} h_{11} & \cdots & h_{1j} \\ \vdots & \ddots & \vdots \\ h_{i1} & \cdots & h_{ij} \end{bmatrix} \begin{bmatrix} h_{11} & \cdots & h_{1j} \\ \vdots & \ddots & \vdots \\ h_{i1} & \cdots & h_{ij} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} h_{11} & \cdots & h_{1j} \\ \vdots & \ddots & \vdots \\ h_{i1} & \cdots & h_{ij} \end{bmatrix} \begin{bmatrix} h_{11} & \cdots & h_{i1} \\ \vdots & \ddots & \vdots \\ h_{1j} & \cdots & h_{ij} \end{bmatrix} =$$

$$= \begin{bmatrix} (h_{11} \cdots h_{1j})(h_{11} \cdots h_{1j}) & \cdots & (h_{11} \cdots h_{1j})(h_{i1} \cdots h_{ij}) \\ \vdots & \ddots & \vdots \\ (h_{i1} \cdots h_{ij})(h_{11} \cdots h_{1j}) & \cdots & (h_{i1} \cdots h_{ij})(h_{i1} \cdots h_{ij}) \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1\mathbf{h}_1 & \cdots & \mathbf{h}_1\mathbf{h}_i \\ \vdots & \ddots & \vdots \\ \mathbf{h}_i\mathbf{h}_1 & \cdots & \mathbf{h}_i\mathbf{h}_i \end{bmatrix}$$

luego

$$s_{11}=s_1^2=\sigma_1^2 \propto \mathbf{h}_1\,\mathbf{h}_1=||\mathbf{h}_1||^2=(1)=||\mathbf{x}_1||^2$$

(1) Por $\mathbf{x}_j^{\mathsf{T}}\mathbf{x}_k=\mathbf{h}_j^{\mathsf{T}}\mathbf{h}_k$

d) El coseno del ángulo que forman dos marcadores columna aproxima la correlación entre las variables asociadas a estas columnas:

$$cos(\mathbf{h}_j\mathbf{h}_k) = (1) = \frac{\mathbf{h}_j^{\mathsf{T}}\mathbf{h}_k}{\|\mathbf{h}_j\|\|\mathbf{h}_k\|} = (2) \propto \frac{s_{jk}}{s_j s_k} = (3) = r_{jk}$$

(1) Definición del producto escalar de dos vectores y sus propiedades.

(2) De la propiedad anterior (c).

(3) Definición coeficiente correlación lineal de Pearson.

e) La distancia euclídea entre dos vectores columna de **X** coincide con la distancia entre los respectivos marcadores columna **H**.

$$d^2(\mathbf{x}_j,\mathbf{x}_s) = \sum_i (x_{ij}-x_{is})^2 = (\mathbf{x}_j-\mathbf{x}_s)^{\mathsf{T}}(\mathbf{x}_j-\mathbf{x}_s) = (1) = (\mathbf{x}_j^{\mathsf{T}}-\mathbf{x}_s^{\mathsf{T}})(\mathbf{x}_j-\mathbf{x}_s) =$$

$$= (2) = \mathbf{x}_j^{\mathsf{T}}\mathbf{x}_j - \mathbf{x}_j^{\mathsf{T}}\mathbf{x}_s - \mathbf{x}_s^{\mathsf{T}}\mathbf{x}_j + \mathbf{x}_s^{\mathsf{T}}\mathbf{x}_s = (3) = \mathbf{x}_j^{\mathsf{T}}\mathbf{x}_j - (\mathbf{x}_s^{\mathsf{T}}\mathbf{x}_j)^{\mathsf{T}} - \mathbf{x}_s^{\mathsf{T}}\mathbf{x}_j + \mathbf{x}_s^{\mathsf{T}}\mathbf{x}_s =$$

$$= (4) = \mathbf{x}_j^{\mathsf{T}}\mathbf{x}_j - (\mathbf{x}_s^{\mathsf{T}}\mathbf{x}_j)^{\mathsf{T}} - \mathbf{x}_s^{\mathsf{T}}\mathbf{x}_j + \mathbf{x}_s^{\mathsf{T}}\mathbf{x}_s = \|\mathbf{x}_j\|^2 + \|\mathbf{x}_s\|^2 - (\mathbf{x}_s^{\mathsf{T}}\mathbf{x}_j)^{\mathsf{T}} - \mathbf{x}_s^{\mathsf{T}}\mathbf{x}_j = (6) =$$

$$= \|\mathbf{h}_j\|^2 + \|\mathbf{h}_s\|^2 - (\mathbf{x}_s^{\mathsf{T}}\mathbf{x}_j)^{\mathsf{T}} - \mathbf{x}_s^{\mathsf{T}}\mathbf{x}_j = d^2(\mathbf{h}_j,\mathbf{h}_s)$$

(1) Propiedad de la traspuesta de suma (resta) de matrices.

(2) Desarrollamos los paréntesis.

(3) Propiedad de las matrices $(\mathbf{AB})^{\mathsf{T}}=\mathbf{B}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}$.

(4) Reordenamos a expresión

(5) Definición de norma y producto escalar

(6) Por $\mathbf{x}_j^T\mathbf{x}_k=\mathbf{h}_j^T\mathbf{h}_k$

f) Las coordenadas de la matriz de marcadores columna **H** en el Biplot de componentes principales son equivalentes a la importancia de las variables a lo largo de los ejes principales, esto es:

$$\varphi = \mathbf{X}^T\mathbf{U}=(1)=(\mathbf{U\Sigma V}^T)^T\mathbf{U}=(2)= \mathbf{V}(\mathbf{U\Sigma})^T\mathbf{U}=(2)=\mathbf{V\Sigma}^T\mathbf{U}^T\mathbf{U}=(3,4)=\mathbf{V\Sigma}=\mathbf{H}$$

(1) $\mathbf{X}=\mathbf{U\Sigma V}^T$.

(2) Propiedad de las matrices $(\mathbf{AB})^T=\mathbf{B}^T\mathbf{A}^T$.

(3) $\mathbf{U}^T\mathbf{U}=\mathbf{I}$.

(4) $\mathbf{\Sigma}=\mathbf{\Sigma}^T$ (es una matriz diagonal).

Propiedades de los marcadores fila (G)

g) En un GH-Biplot la distancia de Mahalanobis entre dos filas i y s de **X** coincide con la distancia euclídea entre dos marcadores fila $\mathbf{g}_i$ y $\mathbf{g}_s$.

La distancia de Mahalanobis se define como $\delta^2_{is}=(\mathbf{x}_i-\mathbf{x}_s)^T\mathbf{S}^{-1}(\mathbf{x}_i-\mathbf{x}_s)$

La i-ésima fila de la matriz **X** se puede escribir como $\mathbf{x}_i=\mathbf{Hg}_i$ sustituimos en δ^2_{is}

$$\delta^2_{is}=(\mathbf{x}_i-\mathbf{x}_s)^T\mathbf{S}^{-1}(\mathbf{x}_i-\mathbf{x}_s)=(1)=(\mathbf{Hg}_i-\mathbf{Hg}_s)^T\mathbf{S}^{-1}(\mathbf{Hg}_i-\mathbf{Hg}_s)=(\mathbf{g}^T_i\mathbf{H}^T-\mathbf{g}^T_s\mathbf{H}^T)\mathbf{S}^{-1}(\mathbf{Hg}_i-\mathbf{Hg}_s)=$$
$$=(2)=(\mathbf{g}^T_i-\mathbf{g}^T_s)\mathbf{H}^T\mathbf{S}^{-1}\mathbf{H}(\mathbf{g}_i-\mathbf{g}_s)=(3)=(\mathbf{g}^T_i-\mathbf{g}^T_s)(\mathbf{\Sigma V}^T)\mathbf{S}^{-1}(\mathbf{V\Sigma})(\mathbf{g}_i-\mathbf{g}_s)=(4)=$$
$$=(\mathbf{g}^T_i-\mathbf{g}^T_s)(\mathbf{\Sigma V}^T)(\mathbf{HH}^T)^{-1}(\mathbf{V\Sigma})(\mathbf{g}_i-\mathbf{g}_s)=(3)=(\mathbf{g}^T_i-\mathbf{g}^T_s)(\mathbf{\Sigma V}^T)(\mathbf{V\Sigma\Sigma V}^T)^{-1}(\mathbf{V\Sigma})(\mathbf{g}_i-\mathbf{g}_s)=$$
$$=(5)=(\mathbf{g}^T_i-\mathbf{g}^T_s)(\mathbf{\Sigma V}^T)(\mathbf{\Sigma V}^T)^{-1}(\mathbf{V\Sigma})^{-1}(\mathbf{V\Sigma})(\mathbf{g}_i-\mathbf{g}_s)=(6)=(\mathbf{g}^T_i-\mathbf{g}^T_s)(\mathbf{g}_i-\mathbf{g}_s)=(\mathbf{g}_i-\mathbf{g}_s)^T(\mathbf{g}_i-\mathbf{g}_s)$$

(1) Sustituimos $\mathbf{x}_i=\mathbf{Hg}_i$.

(2) Sacamos factor común de **H** y $\mathbf{H}^T$

(3) $\mathbf{H}=\mathbf{V\Sigma}$ y $\mathbf{H}^T=\mathbf{\Sigma V}^T$

(4) $\mathbf{S}=\mathbf{HH}^T$

(5) $(\mathbf{AB})^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$

(6) Simplificando $(\mathbf{\Sigma V}^T)(\mathbf{\Sigma V}^T)^{-1}(\mathbf{V\Sigma})^{-1}(\mathbf{V\Sigma})=\mathbf{I}$

h) En el caso del Biplot de componentes principales $\mathbf{G}=(n-1)^{0.5}\mathbf{U}$ y $\mathbf{H}=(n-1)^{-0.5}\mathbf{V\Sigma}$, el i-ésimo marcador fila $\mathbf{g}_i$ coincide con las coordenadas el i-ésimo individuo en

las componentes principales estandarizadas, por lo que **G** se puede interpretar como el score de las componentes principales estandarizadas.

$$\mathbf{g}_i=(1)=(n-1)^{0.5}\mathbf{u}_i=(2)=(n-1)^{0.5}\mathbf{x}_i\mathbf{V}\mathbf{\Sigma}^{-1}$$

(1) $\mathbf{G}=(n-1)^{0.5}\mathbf{U}$

(2) $\mathbf{X}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ luego $\mathbf{U}=\mathbf{X}\mathbf{V}\mathbf{\Sigma}^{-1}$ y $\mathbf{V}=\mathbf{X}^T\mathbf{U}\mathbf{\Sigma}^{-1}$.

i) El producto interno entre filas de **X** es aproximadamente igual al producto interno entre marcadores fila **G**, introduciendo la métrica $(\mathbf{X}^T\mathbf{X})^{-1}$, esto es

$$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T=(1)=\mathbf{G}\mathbf{H}^T((\mathbf{G}\mathbf{H}^T)^T\mathbf{G}\mathbf{H}^T)^{-1}(\mathbf{G}\mathbf{H}^T)^T=(2)=\mathbf{G}\mathbf{H}^T(\mathbf{H}\mathbf{G}^T\mathbf{G}\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{G}^T=(3)=$$
$$=\mathbf{G}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{G}^T=(4)=\mathbf{G}\mathbf{H}^T(\mathbf{H}^T)^{-1}\mathbf{H}^{-1}\mathbf{H}\mathbf{G}^T=\mathbf{G}\mathbf{G}^T$$

(1) $\mathbf{X}=\mathbf{G}\mathbf{H}^T$

(2) $(\mathbf{A}\mathbf{B})^T=\mathbf{B}^T\mathbf{A}^T$

(3) $\mathbf{G}=\mathbf{U}$ y $\mathbf{U}^T\mathbf{U}=\mathbf{G}^T\mathbf{G}=\mathbf{I}$

(4) $(\mathbf{A}\mathbf{B})^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$

En un GH-Biplot las columnas de la matriz **X** (variables) aparecen con alta calidad de representación, pero la calidad de representación de las filas es baja.

$$\text{Bondad ajuste columnas} = \left(\sum_{k=1}^{2}\lambda_k^2 \Big/ \sum_{k=1}^{r}\lambda_k^2 \right) \times 100$$

$$\text{Bondad ajuste filas} = \left(\sum_{k=1}^{2}\lambda_k^0 \Big/ \sum_{k=1}^{r}\lambda_k^0 \right) \times 100 = \frac{2}{r} \times 100$$

Suponiendo que la representación GH-Biplot se efectúa en dimensión 2 y siendo r el rango de la matriz **X** y siendo λ_i el i-ésimo valor singular de **X**.

Se pueden obtener no tan sólo la bondad "global" de la representación, sino incluso la "calidad" de la proyección de cada elemento fila y columna de la matriz **X**. La

Contribución Relativa del Factor α al Elemento β es la variabilidad relativa de una variable explicada por un factor o dimensión.

$$CRF_{\alpha}E_{\beta} = g_{\beta\alpha}^2 \Big/ \sum_{\forall\alpha} g_{\beta\alpha}^2 \qquad CRF_{\alpha}E_{\beta} = h_{\beta\alpha}^2 \Big/ \sum_{\forall\alpha} h_{\beta\alpha}^2$$

Para un marcador la CRFE en un espacio n-dimensional se define como la suma de las CRFE para cada uno de las dimensiones. Solo marcadores con una calidad de representación “alta” pueden ser adecuadamente interpretados en la representación.

En resumen, como conclusiones más relevantes sobre el GH-Biplot, tenemos que por simple inspección del gráfico de dispersión obtenido representando los marcadores fila y marcadores columna de la matriz podemos estimar:

- La variabilidad de las variables, observando la longitud del correspondiente vector: mayor longitud del vector, mayor variabilidad, y viceversa.
- La covariación entre las variables, observando el ángulo que forman: el coseno del ángulo está en relación directa con la covariación entre esas variables.

4.1.3. JK-BIPLOT

Cuando en la factorización de la matriz **X** optamos por la asociación

$$\mathbf{X=U\Sigma V^T} \qquad \mathbf{J=U\Sigma} \qquad \mathbf{K=V} \qquad \mathbf{X=JK^T}$$

para los marcadores fila **J** y columna **K**, respectivamente, estaríamos obteniendo en denominado JK-Biplot o RMP-Biplot (*Row Metric Preserving*).

Recordemos la restricción impuesta para garantizar la unicidad de la solución

$$\mathbf{G^TG = U^TU = I}$$

lo que nos lleva en este caso a que

$$\mathbf{XX^T = JJ^T}$$

Si **X** es una matriz de F filas y C columnas

$$\mathbf{X} = \left(x_{ij}\right) = \begin{pmatrix} x_{11} & \cdots & x_{1C} \\ x_{21} & \cdots & x_{2C} \\ \vdots & & \vdots \\ x_{F1} & \cdots & x_{FC} \end{pmatrix}$$

Tenemos que $\mathbf{XX}^T$:

$$\mathbf{XX}^T = \begin{pmatrix} x_{11} & \cdots & x_{1C} \\ x_{21} & \cdots & x_{2C} \\ \vdots & & \vdots \\ x_{F1} & \cdots & x_{FC} \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{F1} \\ x_{12} & \cdots & x_{F2} \\ \vdots & & \vdots \\ x_{1C} & \cdots & x_{FC} \end{pmatrix}$$

Pero ¿qué son cada uno de los elementos de la matriz $\mathbf{XX}^T$?

$$\left(\mathbf{XX}^T\right)_{ij} = \mathbf{X_i.X_j.}$$

Esto es, no son más que los productos escalares entre las filas de la matriz de los datos originales:

$$\left(\mathbf{XX}^T\right)_{ij} = \mathbf{X_i.X_j.} = \sum_{k=1}^{C} X_{ik} X_{jk}$$

Por lo tanto los productos escalares entre los marcadores fila reproducen los productos escalares entre las filas de la matriz original, esto es preservan la métrica (euclídea) entre las filas (*Row Metric Preserving*).

4.1.3.1. Propiedades del JK-BIPLOT

El JK-Biplot presenta diversas propiedades [157], [160], [162]–[164]:

Propiedades de los marcadores fila (J)

a) Los productos escalares de las filas de **X** coinciden con los productos escalares de los marcadores fila **J**, con la métrica identidad.

$$\mathbf{XX}^T=[\mathbf{JK}^T]\,[\mathbf{JK}^T]^T=(1)=\mathbf{JK}^T\mathbf{KJ}^T=(2)=\mathbf{JV}^T\mathbf{VJ}^T=(3)=\mathbf{JJ}^T$$

(1) $\mathbf{X=JK}^T$

(2) $(\mathbf{AB})^T=\mathbf{B}^T\mathbf{A}^T$

(3) $\mathbf{K=V}$

(4) $\mathbf{V}^T\mathbf{V}=\mathbf{I}$

b) La distancia entre marcadores fila **J** coincide con la distancia euclídea entre las filas de la matriz **X**

$$\left(\mathbf{x}_i-\mathbf{x}_s\right)^T\left(\mathbf{x}_i-\mathbf{x}_s\right)=(1)=\left(\mathbf{K}\mathbf{j}_i-\mathbf{K}\mathbf{j}_s\right)^T\left(\mathbf{K}\mathbf{j}_i-\mathbf{K}\mathbf{j}_s\right)=(2)=\left(\mathbf{j}_i^T\mathbf{K}^T-\mathbf{j}_s^T\mathbf{K}^T\right)\left(\mathbf{K}\mathbf{j}_i-\mathbf{K}\mathbf{j}_s\right)=$$
$$=(3)=\left(\mathbf{j}_i^T-\mathbf{j}_s^T\right)\mathbf{K}^T\mathbf{K}\left(\mathbf{j}_i-\mathbf{j}_s\right)=(4)=\left(\mathbf{j}_i^T-\mathbf{j}_s^T\right)\left(\mathbf{j}_i-\mathbf{j}_s\right)=(5)=\left(\mathbf{j}_i-\mathbf{j}_s\right)^T\left(\mathbf{j}_i-\mathbf{j}_s\right)$$

(1) $\mathbf{X}_i=\mathbf{K}\mathbf{j}_i$.

(2) $(\mathbf{A}+\mathbf{B})^T=\mathbf{A}^T+\mathbf{B}^T$ y $(\mathbf{AB})^T=\mathbf{B}^T\mathbf{A}^T$.

(3) Factor común de $\mathbf{K}^T$ y $\mathbf{K}$

(4) $\mathbf{K}^T\mathbf{K}=\mathbf{V}^T\mathbf{V}=\mathbf{I}$

(5) $\mathbf{A}^T+\mathbf{B}^T=(\mathbf{A}+\mathbf{B})^T$

c) Los marcadores fila **J** coinciden con las coordenadas de las filas en el espacio de las componentes principales.

$$\varphi=\mathbf{XV}=(1)=(\mathbf{U\Sigma V}^T)\mathbf{V}=(2)=\mathbf{U\Sigma}=\mathbf{J}$$

(1) $\mathbf{X}=\mathbf{U\Sigma V}^T$

(2) $\mathbf{V}^T\mathbf{V}=\mathbf{I}$

Esta propiedad implica que es posible estudiar similitudes entre individuos con mínima pérdida de información.

Propiedades de los marcadores columna (K)

a) Las coordenadas para las columnas **K** coinciden con las proyecciones de los ejes originales (base canónica = matriz identidad **I**) en el espacio de las componentes principales.

$$\varphi=\mathbf{V}=\mathbf{IV}=\mathbf{IK}=\mathbf{K}$$

b) El producto interno entre marcadores columna **K** aproxima el producto interno entre columnas de **X** con la métrica $(\mathbf{XX}^T)^{-1}$.

$$\mathbf{x}_j^T\left(\mathbf{XX}^T\right)^{-1}\mathbf{x}_s=(1)=\left(\mathbf{J}\mathbf{k}_j\right)^T\left(\mathbf{XX}^T\right)^{-1}\mathbf{J}\mathbf{k}_s=(2)=\left(\mathbf{k}_j^T\mathbf{J}^T\right)\left(\mathbf{XX}^T\right)^{-1}\mathbf{J}\mathbf{k}_s=(3)=$$
$$\left(\mathbf{k}_j^T\mathbf{J}^T\right)\left(\mathbf{JJ}^T\right)^{-1}\mathbf{J}\mathbf{k}_s=(4)=\mathbf{k}_j^T\mathbf{J}^T\left(\mathbf{J}^T\right)^{-1}\mathbf{J}^{-1}\mathbf{J}\mathbf{k}_s=(5)=\mathbf{k}_j^T\mathbf{k}_s$$

(1) $\mathbf{x}_j=\mathbf{J}\mathbf{k}_j$, $\mathbf{x}_s=\mathbf{J}\mathbf{k}_s$

(2) $(\mathbf{AB})^T=\mathbf{B}^T\mathbf{A}^T$

(3) $\mathbf{X}\mathbf{X}^T=\mathbf{J}\mathbf{J}^T$

(4) Propiedad matriz inversa $(\mathbf{AB})^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$

(5) Propiedad matriz inversa $\mathbf{A}\mathbf{A}^{-1}=\mathbf{A}^{-1}\mathbf{A}=\mathbf{I}$

c) La similitud entre columnas de **X** se aproxima utilizando la inversa de la matriz de dispersión entre individuos, no siendo posible interpretar los ángulos en términos de correlación, ya que:

$$(\mathbf{x}_j-\mathbf{x}_s)^T(\mathbf{X}\mathbf{X}^T)^{-1}(\mathbf{x}_j-\mathbf{x}_s) \cong (\mathbf{k}_j-\mathbf{k}_s)^T(\mathbf{X}\mathbf{X}^T)^{-1}(\mathbf{k}_j-\mathbf{k}_s)$$

En un JK-Biplot las filas de la matriz **X** aparecen con alta calidad de representación, pero la calidad de representación de las columnas es baja.

$$\text{Bondad ajuste filas} = \left(\sum_{k=1}^{2} \lambda_k^2 \Big/ \sum_{k=1}^{r} \lambda_k^2 \right) \times 100$$

$$\text{Bondad ajuste columnas} = \left(\sum_{k=1}^{2} \lambda_k^0 \Big/ \sum_{k=1}^{r} \lambda_k^0 \right) \times 100 = \frac{2}{r} \times 100$$

Suponiendo que la representación JK-Biplot se efectúa en dimensión 2 y siendo r el rango de la matriz **X** y siendo λ_i el i-ésimo valor singular de **X**.

Al igual que en el caso del GH-Biplot, en el JK-Biplot también se pueden obtener no tan sólo la bondad "global" de la representación, sino incluso la "calidad" de la proyección de cada elemento fila y columna de la matriz **X**.

$$CRF_\alpha E_\beta = j_{\beta\alpha}^2 \Big/ \sum_{\forall\alpha} j_{\beta\alpha}^2 \qquad CRF_\alpha E_\beta = k_{\beta\alpha}^2 \Big/ \sum_{\forall\alpha} k_{\beta\alpha}^2$$

En resumen, como conclusiones más relevantes sobre el JK-Biplot, tenemos que por simple inspección del gráfico de dispersión obtenido representando los marcadores fila y marcadores columna de la matriz podemos estimar:

- Los marcadores para las filas coinciden con las coordenadas de los individuos para las componentes principales.
- La similitud entre las columnas se aproxima utilizando como métrica la inversa de la matriz de dispersión entre los individuos.

4.1.4. HJ-BIPLOT

El HJ-Biplot [160] es una representación gráfica multivariante de una matriz **X** de f-filas y c-columnas, mediante los marcadores $\mathbf{j}_1,\ldots,\mathbf{j}_f$ para sus filas y $\mathbf{h}_1,\ldots,\mathbf{h}_c$ para sus columnas, elegidos de forma que ambos marcadores puedan ser superpuestos en un mismo sistema de referencia con máxima calidad de representación, por ello se lo denomina también RCMP-Biplot (*Row Column Metric Preserving*)

Partiendo, como en los casos anteriores, de la Descomposición en Valores Singulares de **X** obtenemos los marcadores fila **J** y columna **H** deseados:

$$\mathbf{X}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \qquad \mathbf{H}=\mathbf{V}\mathbf{\Sigma} \quad \mathbf{J}=\mathbf{U}\mathbf{\Sigma}$$

Es muy importante tener en cuenta que en este caso $\mathbf{X}\neq\mathbf{J}\mathbf{H}^T$, esto es, el producto interior de los marcadores fila **J** y columna **H** en el HJ-Biplot NO REPRODUCE la matriz original **X**.

4.1.4.1. Propiedades del HJ-Biplot

El HJ-Biplot presenta diversas propiedades [157], [160], [162]–[164]. Las demostraciones de las propiedades siguientes se encuentran en los respectivos apartados anteriores del GH-Biplot y JK-Biplot:

Propiedades de los marcadores columna (H)

a) Los productos escalares de las columnas de **X** coinciden con los productos escalares de los marcadores **H**.

$$\mathbf{X}^T\mathbf{X}=\mathbf{H}\mathbf{H}^T \qquad \mathbf{x}_j^T\mathbf{x}_k=\mathbf{h}_j^T\mathbf{h}_k$$

b) Si hacemos que $\mathbf{G}=(n-1)^{0.5}\mathbf{U}$ y $\mathbf{H}=(n-1)^{-0.5}\mathbf{V}\mathbf{\Sigma}$ (Biplot de componentes principales) y si consideramos que **X** es una matriz centrada por columnas

(=variables), esto es que la media de las columnas es cero, y tenemos que **S** es la matriz de varianzas y covarianzas muestral de **X**, sería $\mathbf{S}=\mathbf{X}^T\mathbf{X}/(n-1)$, luego

$$\mathbf{S}=\mathbf{H}\mathbf{H}^T$$

Si el factor de escala introducido anteriormente no se aplica y con **X** centrada por columnas, se sigue cumpliendo lo anterior, excepto por un factor de escala:

$\mathbf{S}= \mathbf{H}\mathbf{H}^T/(n-1)$ Esto es $\mathbf{H}\mathbf{H}^T= (n-1)\mathbf{S} \propto \mathbf{S}$

c) La longitud al cuadrado de los marcadores columna $\mathbf{h}_j$ (=fila j-ésima de **H**) aproxima la varianza de la variable $\mathbf{X}^T_j$ (=fila j-ésima de $\mathbf{X}^T$), por lo que la longitud aproxima la desviación estándar.

$$s_{11}=\sigma_{11}^2 \propto \mathbf{h}_1\,\mathbf{h}_1=||\mathbf{h}_1||^2 = ||\mathbf{x}_1||^2$$

d) El coseno del ángulo que forman dos marcadores columna aproxima la correlación entre las variables asociadas a estas columnas:

$$cos\left(\mathbf{h}_j,\mathbf{h}_k\right) = (1) = \frac{\mathbf{h}_j^T\mathbf{h}_k}{\left\|\mathbf{h}_j\right\|\left\|\mathbf{h}_k\right\|} = (2) \propto \frac{s_{jk}}{s_j s_k} = (3) = r_{jk}$$

e) La distancia euclídea entre dos vectores columna de **X** coincide con la distancia entre los respectivos marcadores columna **H**.

f) Las coordenadas de la matriz de marcadores columna **H** en el Biplot de componentes principales son equivalentes a la importancia de las variables a lo largo de los ejes principales, esto es:

$$\varphi = \mathbf{X}^T\mathbf{U} = \mathbf{H}$$

Propiedades de los marcadores fila (J)

a) Los productos escalares de las filas de **X** coinciden con los productos escalares de los marcadores fila **J**.

$$\mathbf{X}\mathbf{X}^T = \mathbf{J}\mathbf{J}^T$$

b) La distancia entre marcadores fila **J** coincide con la distancia euclídea entre las filas de la matriz **X**, o de otro modo, la distancia euclídea entre dos vectores fila de **X** coincide con la distancia entre los respectivos marcadores fila **J.**

c) Los marcadores fila **J** coinciden con las coordenadas de las filas en el espacio de las componentes principales.

$$\varphi = \mathbf{XV} = \mathbf{J}$$

Esta propiedad implica que es posible estudiar similitudes entre individuos con mínima pérdida de información.

Propiedades conjuntas de los marcadores fila (J) y columna (H)

a) Los marcadores fila **J** y columna **H** se pueden representar en el mismo sistema de referencia.

Esta afirmación se basa en el hecho de que ambas nubes de puntos **H** y **J** están referidas a los mismos valores propios y por lo tanto están relacionadas. Esto es evidente ya que los valores propios de $\mathbf{X}^T\mathbf{X}$ y $\mathbf{X}\mathbf{X}^T$ son los mismos, ya que los valores propios de **A** y $\mathbf{A}^T$ son los mismos (siendo **A** una matriz cuadrada) y $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.

Las relaciones que ligan los vectores propios **U** y **V** son:

$$\mathbf{U}=\mathbf{X}\mathbf{V}\mathbf{\Sigma}^{-1} \qquad \text{y} \qquad \mathbf{V}=\mathbf{X}^T\mathbf{U}\mathbf{\Sigma}^{-1}$$

Sabemos que los marcadores en el HJ-Biplot son $\mathbf{H}=\mathbf{V\Sigma}$ y $\mathbf{J}=\mathbf{U\Sigma}$, luego

$$\mathbf{H}=\mathbf{V\Sigma}= \mathbf{X}^T\mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{\Sigma} = \mathbf{X}^T\mathbf{U}= \mathbf{X}^T\mathbf{X}\mathbf{V}\mathbf{\Sigma}^{-1}= \mathbf{X}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^{-1}=\mathbf{X}^T\mathbf{U}=\mathbf{X}^T\mathbf{J}\mathbf{\Sigma}^{-1}$$

Las coordenadas para las filas **J** son medias ponderadas de las columnas, siendo las ponderaciones los valores originales de la matriz **X**.

$$\mathbf{J}=\mathbf{U\Sigma}= \mathbf{X}\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{\Sigma} = \mathbf{X}\mathbf{V} =\mathbf{X}\ \mathbf{X}^T\mathbf{U}\mathbf{\Sigma}^{-1}= \mathbf{X}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\ \mathbf{U}\mathbf{\Sigma}^{-1}=\mathbf{X}\mathbf{V}= \mathbf{X}\mathbf{H}\mathbf{\Sigma}^{-1}$$

Las propiedades específicas del HJ-Biplot son las siguientes:

1. Las filas y columnas pueden ser representadas en el mismo sistema de referencia.
    a. Puntos fila y columna pueden relacionarse mediante combinaciones lineales simétricas.
    b. Ambas nubes de puntos (filas y columnas) presentan la misma dispersión.
2. **V** son los vectores propios de $\mathbf{X}^T\mathbf{X}$, luego podemos afirmar que los marcadores para las filas **J** en un HJ-Biplot coinciden con las coordenadas de las filas respecto a los ejes factoriales.
3. **U** son los vectores propios de $\mathbf{X}\mathbf{X}^T$, luego podemos afirmar que los marcadores para las columnas **H** en un HJ-Biplot coinciden con las coordenadas de las columnas respecto a los ejes factoriales

Al igual que sucedía con el GH-Biplot y el JK-Biplot podemos obtener las diferentes bondades de ajuste para filas y columnas:

$$\text{Bondad ajuste filas} = \left(\sum_{k=1}^{2} \lambda_k^2 \Big/ \sum_{k=1}^{r} \lambda_k^2 \right) \times 100$$

$$\text{Bondad ajuste columnas} = \left(\sum_{k=1}^{2} \lambda_k^2 \Big/ \sum_{k=1}^{r} \lambda_k^2 \right) \times 100$$

4. Al igual que en los casos del GH-Biplot y del JK-Biplot también se pueden obtener no tan sólo la bondad "global" de la representación, sino incluso la "calidad" de la proyección de cada elemento fila y columna de la matriz **X**.

$$CRF_\alpha E_\beta = h_{\beta\alpha}^2 \Big/ \sum_{\forall\alpha} h_{\beta\alpha}^2 \qquad CRF_\alpha E_\beta = j_{\beta\alpha}^2 \Big/ \sum_{\forall\alpha} j_{\beta\alpha}^2$$

Estas propiedades se añaden a las referidas a los marcadores **H** y **J** antes enumeradas para los GH-Biplot y JK-Biplot.

En resumen, como conclusiones más relevante para el HJ-Biplot tenemos que por simple inspección del gráfico de dispersión obtenido representando los marcadores fila y marcadores columna de la matriz podemos estimar:

- La variabilidad de las variables, observando la longitud del correspondiente vector: mayor longitud del vector, mayor variabilidad, y viceversa.
- La covariación entre las variables, observando el ángulo que forman: el coseno del ángulo está en relación directa con la covariación entre esas variables.
- La similitud entre entidades taxonómicas puede estimarse a partir de la distancia euclídea entre los puntos que los representan.

En este caso el producto interno (escalar) de los marcadores NO nos permite obtener la matriz **X** original.

4.1.5. SQRT-BIPLOT

Si en la expresión del Biplot genérico optamos por α=1/2 obtenemos el SQRT-Biplot o Biplot Simétrico, en el que

$$\mathbf{X}=\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{T} \qquad \mathbf{A}=\mathbf{U}\boldsymbol{\Sigma}^{1/2} \qquad \mathbf{B}=\mathbf{V}\boldsymbol{\Sigma}^{1/2}$$

Este Biplot asigna papeles simétricos tanto a las filas como a las columnas de **X**.

Las factorizaciones en este caso no son únicas, razón por la que este tipo de Biplot solo se utiliza cuando el objetivo del análisis consiste fundamentalmente en la aproximación de los elementos de la matriz **X**, por ejemplo, en el caso de la diagnosis de modelos en tablas de contingencia donde el papel de las filas y columnas es simétrico . En este tipo de Biplot no es de fundamental importancia la interpretación de los marcadores, ya que el objetivo que se persigue es la diagnosis de modelos

4.1.5.1. Propiedades del SQRT-Biplot

El SQRT-Biplot presenta algunas de las propiedades de otros Biplot [157], [160], [162]–[164]:

a) El producto interior de los marcadores fila **A** y columna **B** reproduce la matriz original **X**.

$$\mathbf{AB}^T = \mathbf{U\Sigma}^{1/2}(\mathbf{V\Sigma}^{1/2})^T = (1) = \mathbf{U\Sigma}^{1/2}(\mathbf{\Sigma}^{1/2})^T\mathbf{V}^T = (2) = \mathbf{U\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{V}^T = \mathbf{U\Sigma V}^T = (3) = \mathbf{X}$$

(1) $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$

(2) Al tratarse de una matriz diagonal $\mathbf{\Sigma}^{1/2} = (\mathbf{\Sigma}^{1/2})^T$

(3) $\mathbf{X} = \mathbf{U\Sigma V}^T$.

Recíprocamente, $\mathbf{BA}^T = \mathbf{X}^T$

b) El producto interno entre columnas de **X** es igual al producto interno entre marcadores columna **B**, introduciendo la métrica **Σ**.

$$\mathbf{X}^T\mathbf{X} = (1) = (\mathbf{U\Sigma V}^T)^T(\mathbf{U\Sigma V}^T) = (2) = \mathbf{V}(\mathbf{U\Sigma})^T(\mathbf{U\Sigma V}^T) = (2) = \mathbf{V\Sigma}^T\mathbf{U}^T(\mathbf{U\Sigma V}^T) = (3) = \mathbf{V\Sigma U}^T\mathbf{U\Sigma V}^T = (4) =$$
$$\mathbf{V\Sigma\Sigma V}^T = (5) = \mathbf{V\Sigma}^{1/2}\mathbf{\Sigma\Sigma}^{1/2}\mathbf{V}^T = (2) = \mathbf{V\Sigma}^{1/2}\mathbf{\Sigma}(\mathbf{V\Sigma}^{1/2})^T = (6) = \mathbf{B\Sigma B}^T$$

(1) $\mathbf{X} = \mathbf{U\Sigma V}^T$

(2) $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$

(3) Al tratarse de una matriz diagonal $\mathbf{\Sigma} = \mathbf{\Sigma}^T$

(4) $\mathbf{U}^T\mathbf{U} = \mathbf{I}$

(5) $\mathbf{\Sigma} = \mathbf{\Sigma}^{1/2}\,\mathbf{\Sigma}^{1/2}$

(6) $\mathbf{B} = \mathbf{V\Sigma}^{1/2}$

c) El producto interno entre filas de **X** es igual al producto interno entre marcadores fila **A**, introduciendo la métrica **Σ**.

$$\mathbf{XX}^T = (1) = \mathbf{U\Sigma V}^T(\mathbf{U\Sigma V}^T)^T = (2) = \mathbf{U\Sigma V}^T\mathbf{V}(\mathbf{U\Sigma})^T = (2) = \mathbf{U\Sigma V}^T\mathbf{V\Sigma}^T\mathbf{U}^T = (3) = \mathbf{U\Sigma V}^T\mathbf{V\Sigma U}^T = (4) =$$
$$= \mathbf{U\Sigma\Sigma U}^T = (5) = \mathbf{U\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{U}^T = \mathbf{U\Sigma}^{1/2}\mathbf{\Sigma\Sigma}^{1/2}\mathbf{U}^T = (2) = \mathbf{U\Sigma}^{1/2}\mathbf{\Sigma}(\mathbf{U\Sigma}^{1/2})^T = (6) = \mathbf{A\Sigma A}^T$$

(1) $\mathbf{X} = \mathbf{U\Sigma V}^T$

(2) $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$

(3) Al tratarse de una matriz diagonal $\mathbf{\Sigma} = \mathbf{\Sigma}^T$

(4) $\mathbf{V}^T\mathbf{V} = \mathbf{I}$

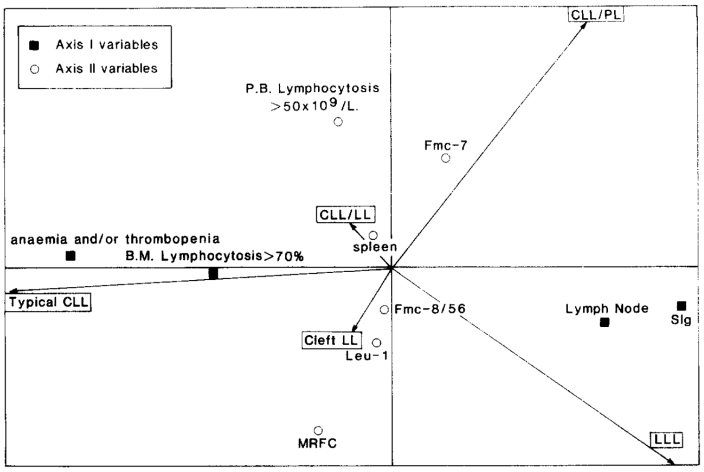
(5) $\mathbf{\Sigma} = \mathbf{\Sigma}^{1/2}\,\mathbf{\Sigma}^{1/2}$

(6) $\mathbf{A} = \mathbf{U\Sigma}^{1/2}$

4.1.6. ALGUNAS APLICACIONES DE LOS MÉTODOS BIPLOT

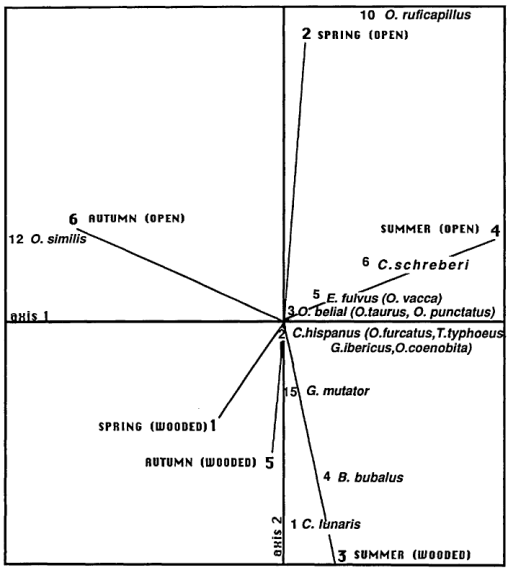
Desde que en el año 1971 Gabriel formulase los primeros métodos Biplot [157] han aparecido en la literatura aplicaciones de los mismos en los campos más diversos. Una revisión más exhaustiva de aplicaciones de los métodos Biplot puede encontrarse en el trabajo de Frutos *et al* de 2013 [165].

En **inmunología**, por ejemplo, en 1988 Orfao *et al* publican un estudio sobre la leucemia [166] en el que utilizan un HJ-Biplot para establecer los diferentes grupos de clasificación de linfocitos. El HJ-Biplot permite analizar la separación entre los grupos, así como identificar las variables que más contribuyen a dicha diferenciación.



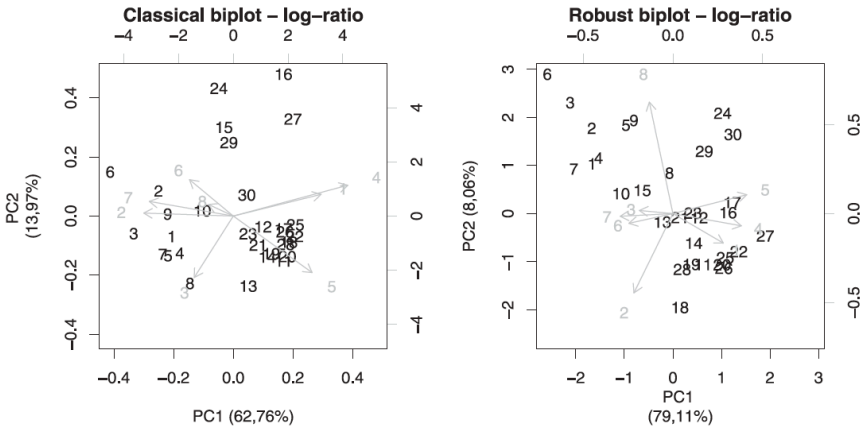
HJ Biplot para la caracterización clínica y de superficie de grupos morfológicos de linfocitos [166].

En **entomología**, en 1991 Galante *et al* publican un estudio sobre el patrón de distribución espacial de los escarabajos en la dehesa mediterránea [167] utilizando también un HJ-Biplot.



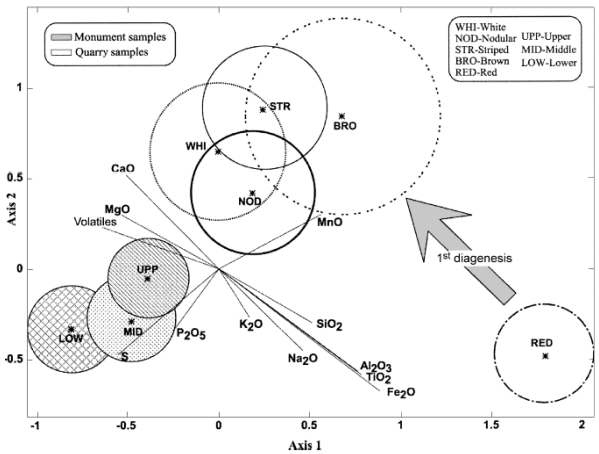
HJ-Biplot conjunto de escarabajos y su entorno [167].

En **enología** se han publicado varios artículos. Así, en 1991 Santos *et al* publicaron un análisis que aplicaba el HJ-Biplot en la caracterización de vinos tintos jóvenes [168], y posibilitaba la clasificación de los vinos de las denominaciones de origen "Ribera de Duero" y "Toro". En 1993 Rivas-Gonzalo *et al* aplican de nuevo el HJ-Biplot a los vinos de "Ribera de Duero" y "Toro" para identificar las variables más importantes que intervienen en la diferenciación de ambas denominaciones de origen [169]. En 2012 Hron *et al* aplican de nuevo biplots composicionales (tradicional y robusto) en el estudio estadístico de vinos [170].



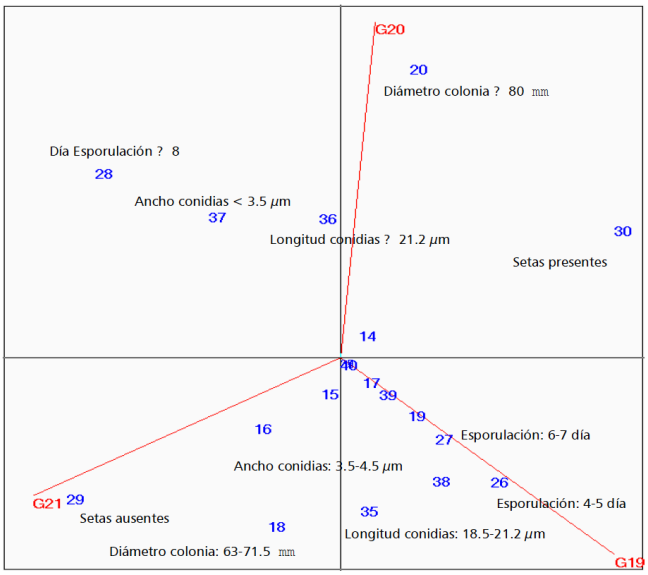
Biplot clásico y robusto obtenido por Hron en enología [170].

En **geología** también se han publicado varias aplicaciones de los métodos Biplot. En 1999 García-Talegón *et al* publican una aplicación del HJ-Biplot en la determinación del origen del granito utilizado en la construcción de la catedral de Ávila [171]. En 2005 Iñigo *et al* vuelven a aplicar el HJ-Biplot en la identificación el origen de las piedras utilizadas en la construcción del puente romano de Salamanca [172]. También en 2005 Varas *et al* vuelven a aplicar un Biplot, en este caso el canónico, para estudiar los diferentes materiales utilizados en la construcción de la catedral de Ciudad Rodrigo (Salamanca) [173].



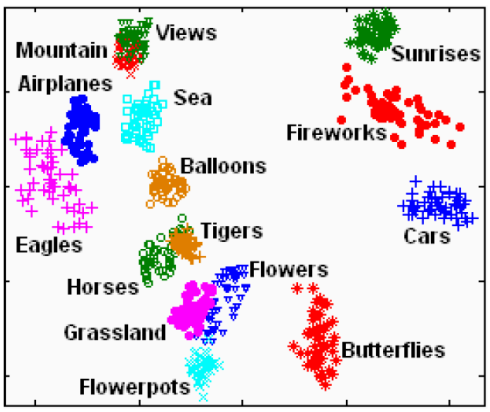
Biplot canónico de los elementos constructivos de la catedral de Ciudad Rodrigo [173].

En **micología**, Correa *et al* publican en 2007 una nueva aplicación del Biplot sobre tablas de contingencia para la detección de las características con mayor capacidad de discriminación entre grupos de aislamientos de *Colletotrichum* [174].



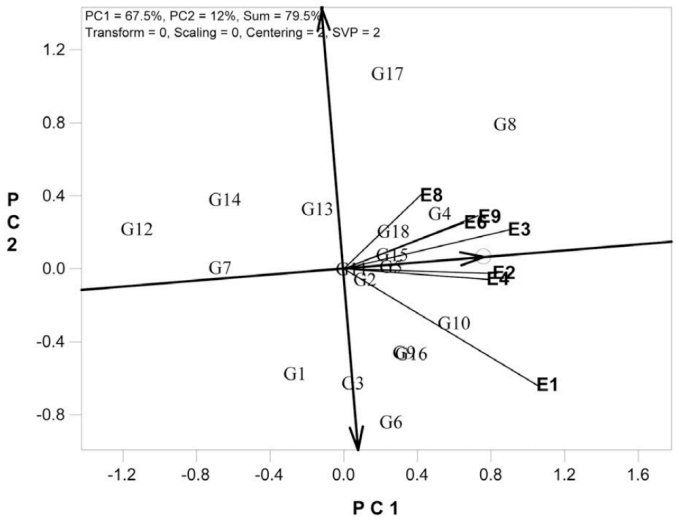
Representación Biplot en estudio micológico [174].

En **análisis de imágenes**, Theoharatos *et al* presentan en 2007 una aplicación de los métodos biplots en la clasificación de imágenes en una base de datos [175] que ofrece una representación de la organización intrínseca de la base de datos que mejor refleja la percepción del usuario. El poder de discriminación que introduce el Biplot, afirman los autores, constituye una atractiva herramienta para la recuperación de imágenes y es una interfaz flexible para tareas de minería de datos visuales.



Resultado de la aplicación de biplots en la organización de bases de datos de imágenes [175].

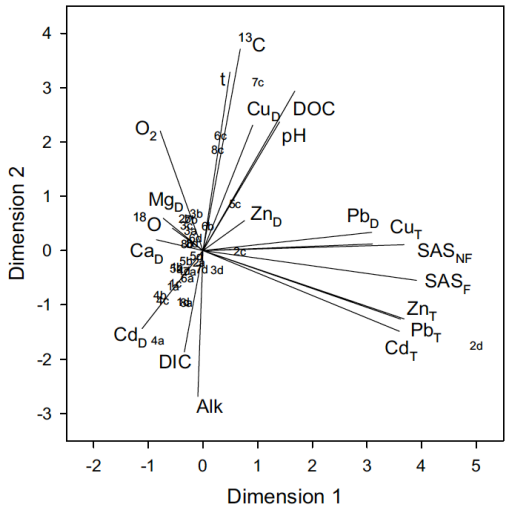
En **genética** Yan *et al* publican en 2007 una aplicación de los métodos Biplot que aprovecha su poder de discriminación y representatividad para la evaluación del entorno y mejora otras herramientas disponibles (análisis de interacción multiplicativa con efecto principal aditivo –AMMI-) [176].



Aplicación de biplots en genética [176].

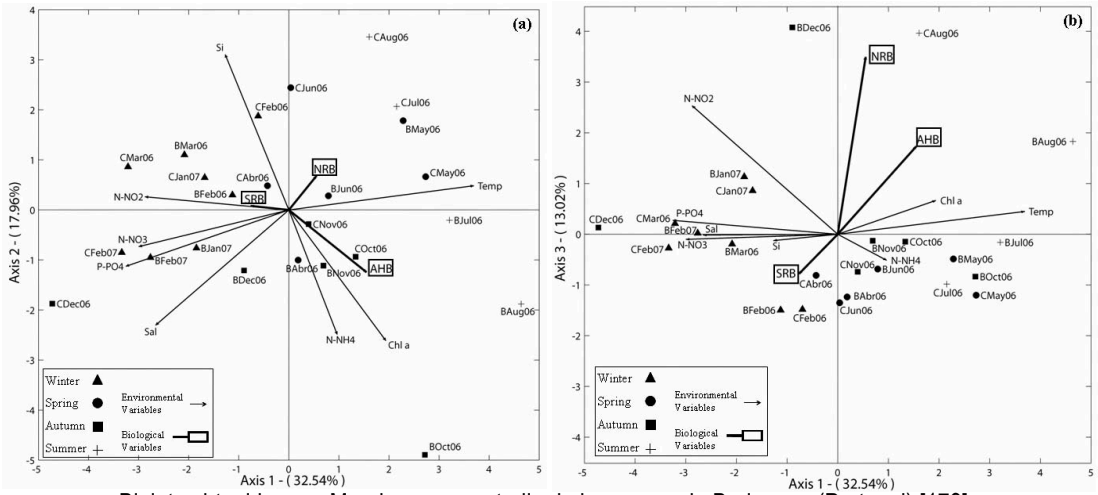
Theoharatos *et al* presentan en 2008 una aplicación de "biplots" en **imaginería médica** para la identificación de osteoartritis en radiografías [177], si bien los autores utilizan el término "Biplot" más para referirse a una representación de baja dimensionalidad sobre un escalado multidimensional clásico, que en el sentido de los biplots de Gabriel.

Cukrov *et al* presentan en 2009 en **hidrología** un análisis estadístico multivariante sobre parámetros físico-químicos del río Krka (República de Croacia) para estudiar la influencia del hombre en dichos parámetros [178]. No obstante, una vez más, el término "Biplot" es utilizado más por los autores para referirse a una representación bidimensional que al método introducido por Gabriel. De hecho los autores utilizan una técnica de desplegamiento multidimensional (MDPREF, *multidimensional preference analysis*) para obtener las coordenadas de los diferentes ítems, si bien se refieren directamente a los biplots de Gabriel.



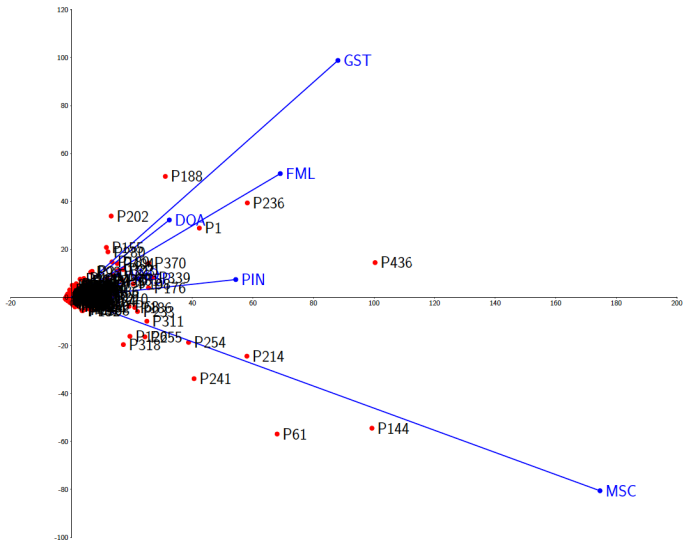
Aplicación de Biplot en estudio de características del río Krka [178].

En 2009 Mendes *et al* publican una aplicación del HJ-Biplot en **biología marina**, estudiando la dinámica del bacterioplankton en la reserva natural de Berlengas (Portugal) [179]. El HJ-Biplot permite representar simultáneamente en baja dimensión los tres grupos principales de bacterias implicadas en el ciclo del carbono y los parámetros del entorno.



Biplots obtenidos por Mendes en su estudio de la reserva de Berlengas (Portugal) [179].

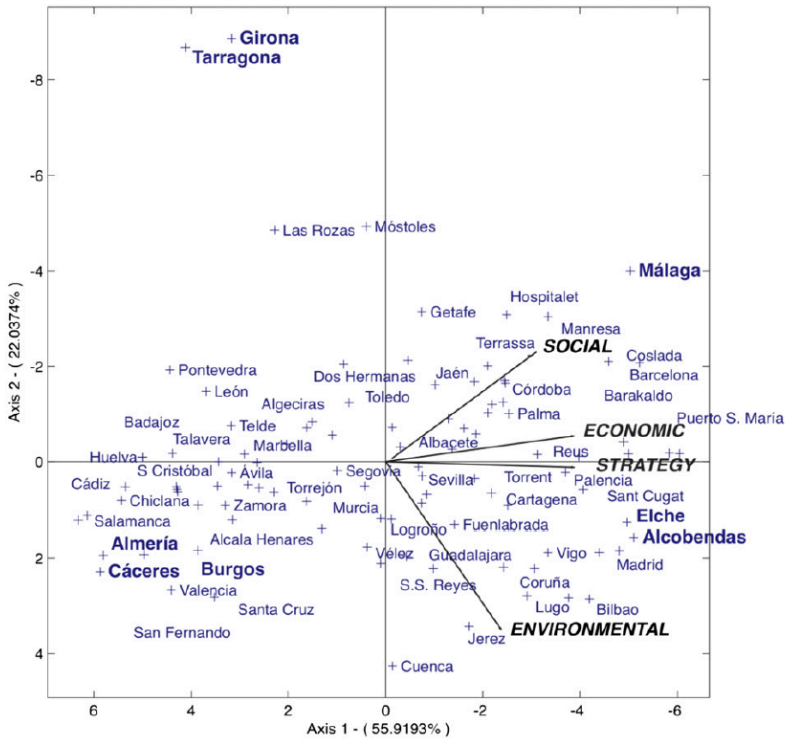
En botánica, Afendi *et al* presentan en 2010 un estudio **botánico** utilizando biplots sobre la relación entre las plantas indonesias y la eficacia del *jamu*, una medicina natural compuesta de una mezcla de plantas medicinales [180]. El método Biplot utilizado permite identificar 190 plantas como responsables de la eficacia del *jamu*, entre 465 plantas totales utilizadas en 3138 combinaciones diferentes de *jamu*.



Uno de los Biplot obtenidos por Afendi en su estudio del *jamu* [180].

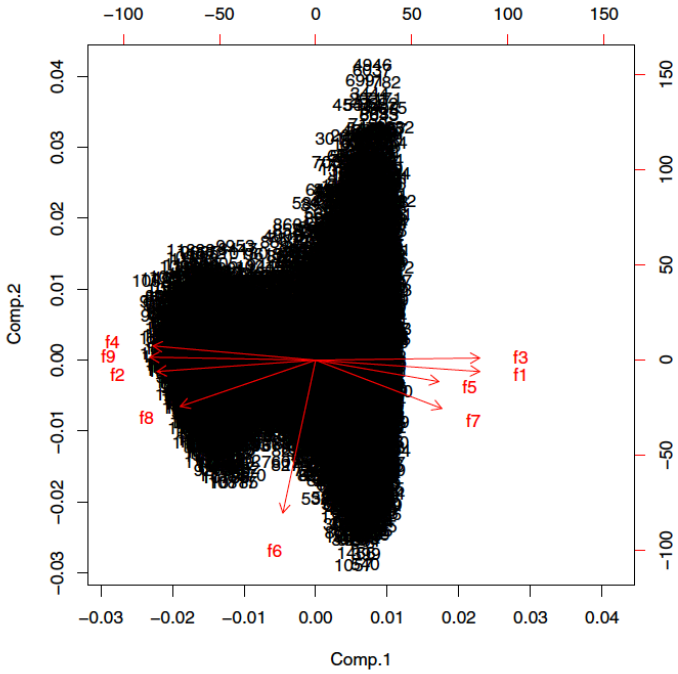
En **sociología** se han publicado varias aplicaciones de los métodos Biplot. Así, por ejemplo, en 2010 Castela y Galindo aplican el método HJ-Biplot como herramienta de

análisis para la caracterización de votantes abstencionistas en elecciones portuguesas [181]. El HJ-Biplot permite identificar en el primer plano factorial seis grupos diferentes de votantes caracterizando su actitud electoral, y su evolución a lo largo de varios procesos electorales, a partir de su localización geográfica. En 2011 Vicente-Galindo utiliza un Biplot logístico para el estudio de la capacidad de innovación en Portugal, identificando las principales características, de entre las 10 consideradas, de las 623 instituciones más innovadoras [182]. Como último ejemplo en este campo, en 2013 García-Sánchez *et al* publica un estudio sobre la transparencia y sostenibilidad de 102 municipios españoles utilizando el método HJ-Biplot y los resultados del Índice de Transparencia de Ayuntamientos (ITA) publicados en 2010 por Transparencia Internacional España, así como otras variables sociodemográficas geolocalizadas disponibles [183].



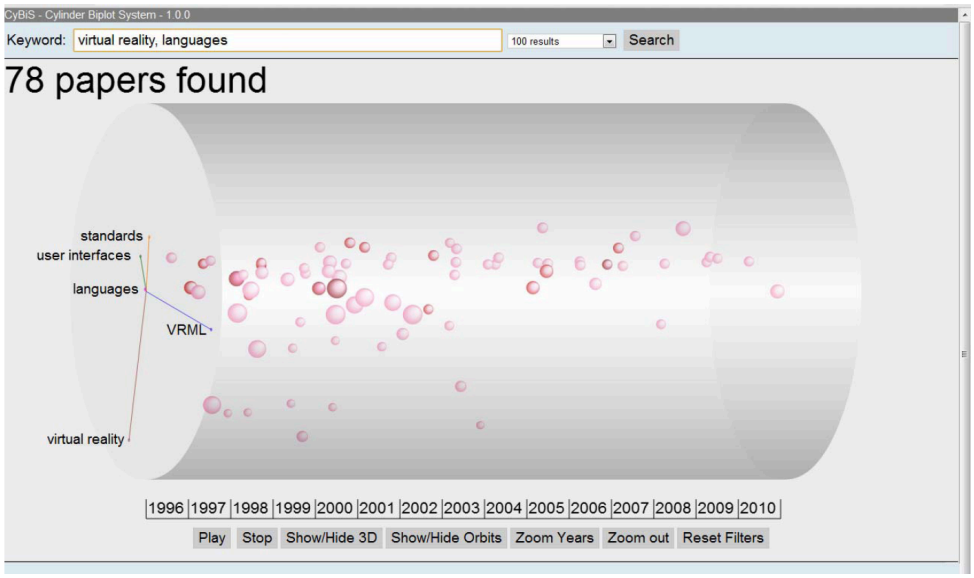
Aplicación del Biplot al estudio de la transparencia en ayuntamientos españoles [183].

En un área del conocimiento más general como es la **teoría de la decisión**, en 2010 Costa y Oliveira presentan una aplicación del GH-Biplot para la detección gráfica de la relación entre objetivos y soluciones incluso en presencia de conjuntos de soluciones que contienen valores dominantes posibilitando la reducción de objetivos en la toma de decisiones [184].



Aplicación de los métodos Biplot en teoría de la decisión para la conformación de una señal de radar [184].

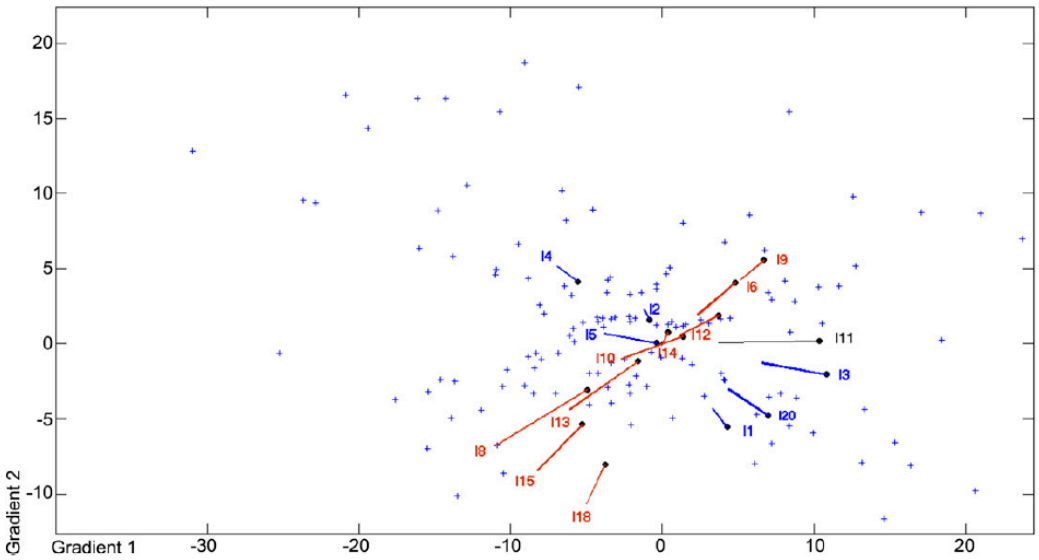
En el área de **biblioteconomía y documentación**, en 2011 Costagliola y Fuccella aplican un Biplot cilíndrico para la representación tridimensional de los resultados de búsquedas bibliográficas presentando en un vista única los datos más relevantes de los artículos recuperados [185].



Interface de recuperación de búsquedas basada en Biplot cilíndrico propuesta por Costagliola y Fuccella [185].

Como último ejemplo de aplicación de la metodología Biplot, comentaremos la aplicación de biplots logísticos en **ecología** y más concretamente al análisis de indicadores ambientales de empresas internacionales, publicado por Gallego-Álvarez y Vicente-Villardón en 2012 [186]. En este estudio se analizan 162 firmas internacionales de diferentes sectores productivos y 20 características ambientales

binarias. Los resultados obtenidos permiten concluir la existencia de dos gradientes principales que resumen la información proporcionada por las 20 variables analizadas.



Biplot logístico obtenido del estudio de 20 indicadores ambientales de 162 empresas [186].

4.2. LOS MÉTODOS STATIS

4.2.1. INTRODUCCIÓN A LOS MÉTODOS STATIS

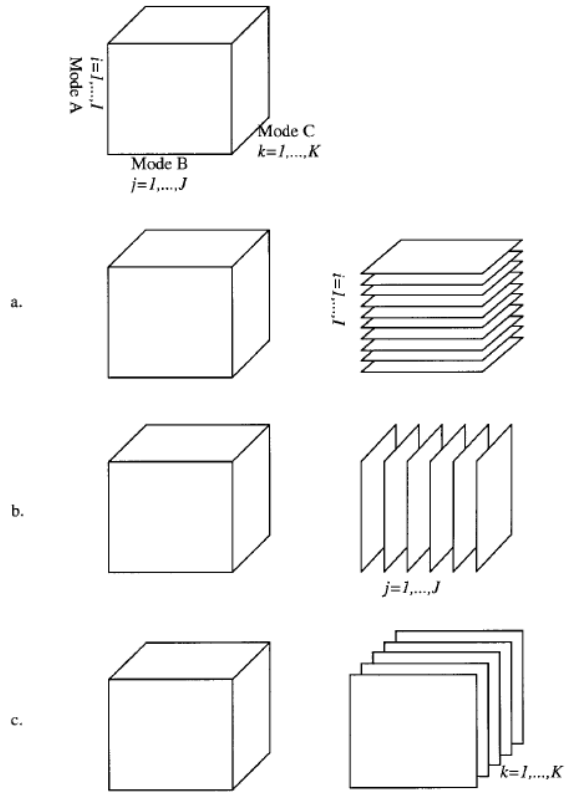
STATIS es un acrónimo de la expresión francesa "*Structuration des Tableaux à Trois Indices de la Statistique*". Esta técnica es también conocida bajo el acrónimo de "ACT", en este caso proveniente de la también expresión francesa "*Analyse Conjointe de Tableaux*". En el fondo no es más que una generalización del Análisis de Componentes Principales (ACP) cuyo objetivo es analizar varios conjuntos de variables medidas (no necesariamente las mismas) en el mismo juego de observaciones (individuos u objetos), o varios juegos de observaciones (no necesariamente las mismas) medidas sobre mismo juego de variables, en este caso se denomina STATIS Dual [187], [188].

El método STATIS fue presentado en 1976 en la Tesis Doctoral de L'Hermier des Plantes [189] y posteriormente, en 1978, en la Tesis Doctoral de Jaffrenou [190] se presentó otro método, denominado Análisis Triádico (aunque a veces se le denomina X-STATIS, o Análisis Triádico Parcial [188]), que no es más que una variación del STATIS para ser utilizado cuando en todos los conjuntos de datos son medidas las mismas variables y las mismas observaciones. Son también relevantes las aportaciones de Lavit en 1988 [191] sobre el método STATIS y Thioulouse y Chessel en 1987 [192] sobre el Triádico.

STATIS es una herramienta exploratoria para análisis de datos de 3 vías. La idea principal es comparar diferentes tablas de datos (matrices) obtenidas bajo varias condiciones experimentales, pero conteniendo el mismo número de filas y/o columnas [191], [193]. En realidad, la mejor manera de comprender el planteamiento del problema es acudir a Kiers que en su artículo de 2000 [194] propone una normalización de la notación y terminología de los conceptos utilizados en el análisis multivía.

Un conjunto de variables medidas sobre un grupo de elementos (u observaciones) pueden agruparse en una matriz bidimensional **X**, con x[i,j] correspondiendo al valor de la variable j-ésima en el elemento i-ésimo (u observación). Pero si tenemos ese mismo conjunto de variables y elementos, que medimos en diferentes circunstancias (tiempos, por ejemplo), la matriz **X** sería ahora "tridimensional" con los diferentes valores x[i,j,k] correspondiendo ahora al valor de la variable j-ésima en el elemento i-ésimo en el instante (por ejemplo), k-ésimo.

En la notación propuesta por Kiers, a esta matriz tridimensional se la denominaría de 3 vías, y cada una de las dimensiones, se denominaría "modo".



Matriz de tres modos, con cortes (a) horizontales, (b) laterales y (c) frontales. [194]

El objetivo del método STATIS es [187]

1. Comparar y analizar la relación entre los diferentes juegos de datos.
2. Combinar los juegos de datos en una estructura común denominada "compromiso" o "consenso" que pueda ser analizada con un ACP para revelar la estructura común entre las observaciones (individuos u objetos, por ejemplo.)
3. Proyectar cada uno de los juegos de datos iniciales en el compromiso para analizar aspectos comunes y discrepantes.

El número y/o naturaleza de las variables utilizadas para describir las observaciones pueden variar de un conjunto de datos a otro, pero las observaciones (individuos u objetos) deben ser las mismas en todos los conjuntos de datos.

4.2.2. EL MÉTODO STATIS

El algoritmo del método STATIS puede descomponerse en los siguientes pasos principales: El primero consiste en analizar la similaridad en las estructuras de los diferentes juegos de datos (matrices) y que se denomina análisis de la "interestructura". Este primer análisis, como veremos, proporciona un juego de pesos que serán utilizados en el segundo paso, que consiste en analizar la intraestructura o estructura interna de las tablas.

Veamos los pasos ahora con más detalle [187]:

Análisis la interestructura

0. Partimos de T conjuntos de datos. Cada uno de esos conjuntos de datos es una matriz rectangular $\mathbf{Y}_{[t]}$ de dimensiones I x $J_{[t]}$ con I siendo el número de observaciones (objetos o individuos, por ejemplo) y $J_{[t]}$ el número de variables recogidas sobre las observaciones en el conjunto de datos t-ésimo (como hemos indicado, en STATIS los diferentes conjuntos de datos pueden contener diferentes variables). Estas matrices de datos son normalizadas (centradas por columnas y/o varianza unidad, por ejemplo) y a estas matrices normalizadas las denotaremos por $\mathbf{X}_{[t]}$.

1. Cada matriz $\mathbf{X}_{[t]}$ se transforma en otra matriz I x I (recordemos que I es el número de observaciones) denominada $\mathbf{S}_{[t]}$ y calculada

$$\mathbf{S}_{[t]} = \mathbf{X}_{[t]} (\mathbf{X}_{[t]})^T$$

2. Para comparar las estructuras entre los T conjuntos de datos obtendremos una nueva matriz **C** de dimensiones T x T cuyo elemento c[t,t'] es el coseno entre estudios t y t'. Este coseno se le conoce también como el coeficiente RV formulado por Escoufier [195] y que se define como:

$$c\left[t,t'\right] = \mathrm{RV} = \frac{\mathrm{traza}\left\{\mathbf{S}_{[t]}^T \mathbf{S}_{[t']}\right\}}{\sqrt{\mathrm{traza}\left\{\mathbf{S}_{[t]}^T \mathbf{S}_{[t]}\right\}\mathrm{traza}\left\{\mathbf{S}_{[t']}^T \mathbf{S}_{[t']}\right\}}}$$

3. La descomposición de la matriz coseno **C** permite obtener la estructura entre los diferentes juegos de datos. Esta descomposición se lleva a cabo sometiendo a la matriz **C** a un ACP SIN centrado previo.

$$\mathbf{C}=\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}^T \qquad \text{con } \mathbf{P}^T\mathbf{P}=\mathbf{I}$$

Donde **P** es la matriz de autovectores de **C** y $\boldsymbol{\Sigma}$ es la matriz diagonal de autovalores. Un elemento de un autovector dado representa la proyección de una matriz / juegos de datos en ese autovector.

Los diferentes juegos de datos pueden ser representados como puntos en el espacio de los vectores propios para estudiar sus similitudes, estas proyecciones pueden ser calculadas del modo habitual

$$\mathbf{G}=\mathbf{P}\boldsymbol{\Sigma}^{1/2}$$

Tomando las dos primeras columnas de **G** (proyecciones de los juegos de datos en la primera y segunda componente principal) podemos representar en el plano los diferentes juegos de datos.

4. Los pesos utilizados para obtener el compromiso se obtienen del ACP de la matriz coseno **C**, que al no estar centrada, su primer autovector o vector propio representa la parte "común" de los diferentes juegos de datos. Así, estudios con valores grandes en la correspondiente componente del primer autovector son más parecidos a los restantes estudios y tendrán un mayor peso en la obtención del compromiso. En la práctica los pesos se obtienen reescalando los elementos del primer autovector de **C** para que su suma sea la unidad. Denominaremos a este vector de pesos $\boldsymbol{\alpha}$. Así pues, α_t será el peso del juego

de datos t-ésimo (componente t-ésima del vector α) y la matriz compromiso será:

$$\mathbf{S}_{[+]} = \sum_{t=1}^{T} \alpha_t \mathbf{S}_{[t]}$$

5. Algunos autores indican que no existe indicador de la calidad de representación en STATIS, aunque al menos podemos obtener un índice de la calidad de presentación del compromiso a través de la relación del primer autovalor con la suma de todos los autovalores:

$$\text{Calidad} = \frac{\lambda}{\sum_{\forall \ell} \lambda_\ell} = \frac{\lambda}{\text{traza}\{\Sigma\}}$$

6. La matriz compromiso puede ser sometida a un ACP para explorar la estructura del conjunto de observaciones.

$$\mathbf{S}_{[+]} = \mathbf{Q}\Lambda\mathbf{Q}^T \quad \text{con } \mathbf{Q}^T\mathbf{Q} = \mathbf{I}$$

Donde **Q** es la matriz de autovectores de $\mathbf{S}_{[+]}$ y Λ es la matriz diagonal de autovalores.

Las diferentes observaciones (individuos u objetos) pueden ser representados como puntos en el espacio de estos vectores propios para estudiar sus similitudes, estas proyecciones pueden ser calculadas del modo habitual

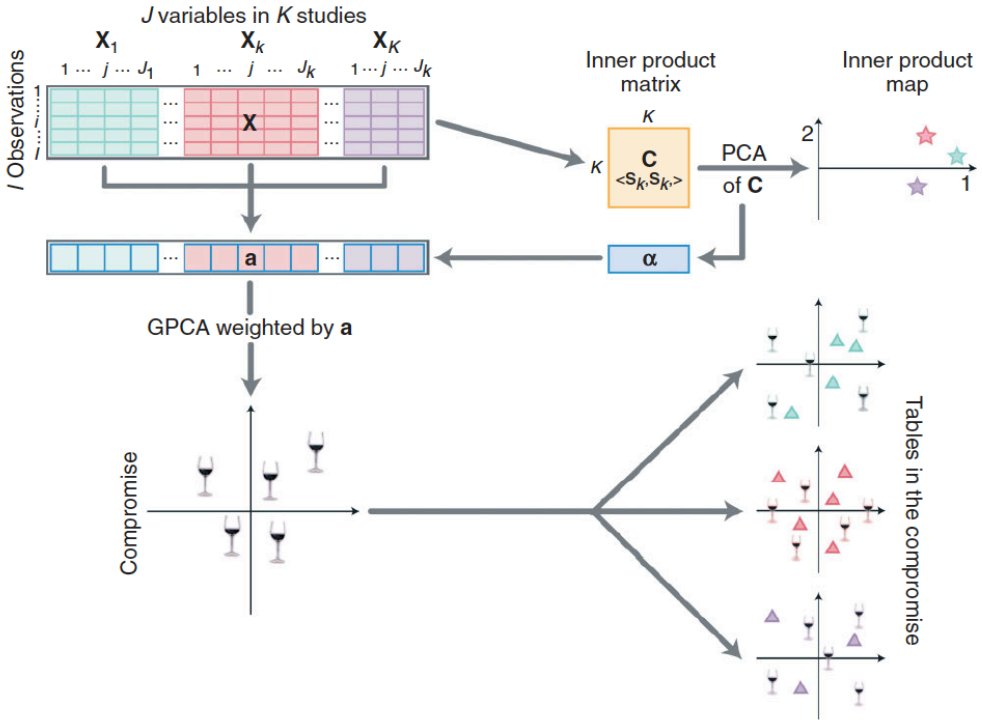
$$\mathbf{F} = \mathbf{Q}\Lambda^{1/2}$$

Como siempre, podemos tomar las proyecciones en las dos primeras componentes principales (dos primera columnas de **F**) para representar las observaciones en el plano. De nuevo la suma de los dos primeros autovalores indican la absorción parcial de inercia en el plano.

7. El análisis del compromiso revela la estructura del espacio común de las observaciones con relación a los diferentes conjuntos de datos, pero además podemos querer estudiar como cada conjunto de datos queda en este espacio. Esto se consigue calculando la matriz de proyección que transforma el producto escalar en cargas (*loadings*)

$$F = S_{[+]}Q\Lambda^{-1/2}$$

Esto se utiliza para representar en el mismo plano factorial las observaciones y los diferentes conjuntos de datos.



Pasos del método STATIS. [188]

Análisis la intraestructura

8. Por último podemos integrar las variables originales en el análisis según el procedimiento habitual del ACP, calculando las cargas (*loadings*), esto es, la correlación entre las variables originales y las puntuaciones de los factores (*factor scores*). Esto permite comparar las diferentes variables utilizadas en cada juego de datos, o intraestructura.

4.2.3. EL MÉTODO STATIS DUAL

En el método STATIS Dual, los datos consisten en K conjuntos de observaciones medidas sobre el mismo conjunto de variables. Aquí, en lugar de calcular los productos cruzados entre las observaciones, calcularemos las matrices de covarianzas entre las variables (una para cada conjunto de observaciones). El STATIS Dual continua desde ese punto con los mismos pasos que el STATIS, y permitirá obtener un espacio compromiso para las variables (en lugar de para las observaciones como en el STATIS) y las cargas parciales para cada tabla.

Para efectuar un análisis STATIS-DUAL es posible aplicar el STATIS "normal" a la matrices originales de datos traspuestas.

Este será el método que aplicaremos en nuestra propuesta, como expondremos.

4.2.4. EL MÉTODO X-STATIS O ANÁLISIS TRIADICO (PARCIAL)

El Análisis Triádico Parcial (*Partial Triadic Analysis – PTA*) también denominado X-STATIS [188], como dijimos fue descrito inicialmente por Jaffrenou [190] es una variación sobre el STATIS que puede ser utilizada para aquellos casos en los que todos los juegos de datos miden las mismas variables para las mismas observaciones. Su nombre Análisis Triádico "Parcial" proviene de que puede ser también visto como un caso simplificado del análisis Tucker-3 [188].

El algoritmo sigue los pasos del STATIS excepto dos diferencias:

1. El producto de matrices utilizado para el cálculo de los pesos α se obtiene desde las matrices $\mathbf{X}_{[k]}$ en lugar de las $\mathbf{S}_{[k]}$

2. El compromiso se obtiene también a partir de las matrices $\mathbf{X}_{[k]}$ en lugar de las $\mathbf{S}_{[k]}$

5. DETECCIÓN DE ANOMALÍAS EN REDES

5. DETECCIÓN DE ANOMALÍAS EN REDES

5.1. INTRODUCCIÓN A LA DETECCIÓN DE ANOMALÍAS

Con un servicio sobre el que circulan literalmente decenas de miles de euros por minuto solo en España, la seguridad en las transacciones y la disponibilidad de los servicios en las redes son cuestiones innegociables. Los requerimientos de las redes de comunicación en entornos que pueden interpretarse como de máxima criticidad como los militares no difieren tanto de los existentes actualmente en entornos "civiles", como negocios, bancos, administración, etc. [196]. Y no digamos si introducimos la utilización de este tipo de servicios en otro entorno también extremadamente crítico como es el médico [197], las medidas de seguridad se incrementan en este caso de manera aún más notable.

Con el crecimiento explosivo de Internet y las infraestructuras de comercio electrónico, la detección de suaves anomalías rápida y proactivamente en redes de área expandida es un prerrequisito para una recuperación de fallos inmediata que evite caídas en los servicios. La degradación de prestaciones y fallos en las redes es el preludio de fallos en los servicios [198]. Estas eventualidades implican siempre pérdidas económicas y/o de imagen importantes para las entidades.

Para mantener la disponibilidad y fiabilidad de la red se debe disponer de mecanismos que permitan detectar y diagnosticar problemas potenciales en la red e iniciar las acciones de recuperación apropiadas. La detección proactiva de anormalidades en la red implica la predicción de anomalías antes de que estas causen fallos catastróficos. Sólo aquellos incidentes que originen cambios en el comportamiento estadístico de variables podrán ser detectados proactivamente. Entre estos se incluyen, por ejemplo, la degradación de prestaciones, el malfuncionamiento de algunos dispositivos, fallos en servidores, degradación de los medios físicos de transmisión, tormentas de difusión…[199]

Como veremos en apartados posteriores han sido numerosos los autores que han dedicado grandes esfuerzos a la elaboración de propuestas con el objetivo de detectar situaciones anómalas no sólo en redes de telecomunicación. Chandola *et al* publican en 2009 un estudio [200] en el que, tras definir la detección de anomalías como el problema de descubrir patrones en datos que no se ajusten a un comportamiento esperado, repasan este concepto en múltiples escenarios: desde médicos (ECG y EEG, TAC/RM,…), industriales (prestaciones de motores, defectos estructurales,...),

procesado de imagen (videovigilancia, satélites,…) y por supuesto detección de anomalías en redes (detección y diagnosis de averías, detección de intrusos,…). En este último caso se han seguido varios esquemas generales para aproximarse al problema: por las propias características de la información objeto de estudio, las series temporales han motivado no pocas propuestas. La introducción de métodos estadísticos tales como el Análisis de Componentes Principales y el Análisis de Componentes Independientes, ha venido a veces acompañada de la utilización de funciones de preprocesado de la serie de datos más complejas. Así, por ejemplo, se ha utilizado la entropía, y variaciones de la misma, para acentuar las anormalidades existentes en la serie temporal. La aplicación del Análisis de Componentes Principales ha permitido extender el estudio desde un único enlace a múltiples enlaces de manera simultánea. Otros autores han optado por utilizar diferentes métodos de representación gráfica de las señales que permitiesen poner de manifiesto la ocurrencia de una anomalía dentro de un patrón de normalidad. La combinación de ambos métodos en uno único, aplicando métodos estadísticos multivariantes gráficos aprovechando ambos aspectos, es algo novedoso en este campo.

En nuestro estudio propondremos la aplicación de las metodologías estadísticas multivariantes sobre las que son las dos tipologías de redes de telecomunicación más prevalentes en la actualidad: las redes de área local soportadas sobre la norma Ethernet y la red de área expandida por antonomasia, Internet.

Ciertamente ambas elecciones tecnológicas presentan, además, una sinergia no desdeñable. Una red de área local que utiliza los protocolos de Internet como portadores de información en una empresa, como definición más general, se la conoce como una Intranet [201].

5.2. DETECCIÓN DE ANOMALÍAS EN REDES: ESTADO DEL ARTE.

En 1990 Maxion publica [41] el que consideramos el primer artículo sobre detección de anomalías en redes. Con la red Internet en forma embrionaria (hasta el 30 de abril de 1993 no se liberó el primer conjunto de aplicaciones para navegar por la web [202]) el artículo de Maxion se circunscribe a la red Ethernet instalada en la *School of Computer Science* de la Universidad de Carnegie Mellon. El artículo, en su parte analítica, es muy elemental comparado con los planteamientos actuales. Establece un simple modelo lineal que opera como un filtro de señales para detectar las variaciones anómalas en el patrón de volumen de tráfico cursado, que es representado como el porcentaje de uso del ancho de banda de la red Ethernet. No obstante en su parte

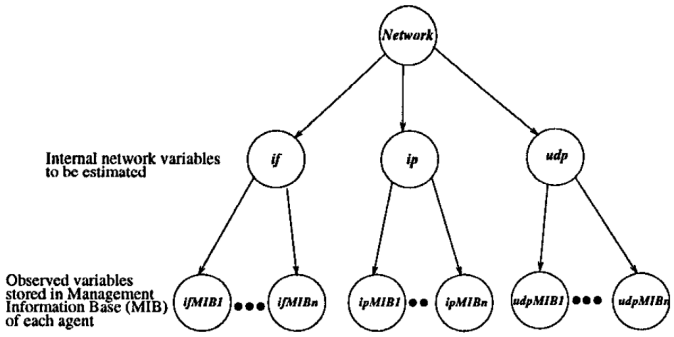
menos analítica establece muchas definiciones que permanecen vigentes hasta nuestros días: detección, diagnosis, anomalía, etc. Identifica igualmente lo que constituyen los principal retos de la detección de anomalías en redes: el entorno altamente dinámico y con fuerte presencia de “ruido”, las características de las series temporales utilizadas, así como la determinación del concepto de “situación normal” (o “nominal”). Estos aspectos siguen actualmente constituyendo el caballo de batalla de la detección de anomalías en redes de comunicación.

En 1993 Feather *et al* [42] publican un nuevo artículo sobre detección de anomalías en redes Ethernet. En su trabajo proponen un método basado en la identificación de patrones de anomalías que utiliza técnicas generales de minería de datos basadas en distancias. Utiliza dos mecanismos de detección de anomalías. El primer detector es un simple umbral que permite establecer un comportamiento anómalo si es sobrepasado. El segundo mecanismo utiliza una combinación de elementos heurísticos y métodos estadísticos. En este último caso define un vector de características para establecer la presencia de una situación anómala. Las variables que utiliza para ello son: la carga de la red, paquetes transmitidos, colisiones, tamaño de paquetes, paquetes de difusión, direcciones de origen y destino, principalmente. Posteriormente monitoriza la red en esas determinadas variables y las compara, mediante el cálculo de una distancia, con los vectores establecidos para cada condición de error. Los autores identifican para ello los errores más comunes en redes Ethernet, que se expusieron en el apartado anterior de esta tesis relativo a la norma Ethernet.

En 1995 Katzela y Schwartz [203] proponen un modelo para la correlación entre alarmas y posterior localización de puntos de fallo. Este modelo tiene en consideración las dependencias existentes entre los objetos que forman la red de comunicaciones. Basándose en dicho modelo proponen un algoritmo, también heurístico, capaz de colegir una explicación plausible para las alarmas recibidas. La situación que plantean es el habitual disparo de múltiples alarmas debido a un fallo único. Las múltiples alarmas dificultan la diagnosis del problema. El algoritmo funciona especialmente bien cuando todos los elementos de la red tienen la misma probabilidad independiente de fallo. En su artículo realizan un repaso de la bibliografía disponible hasta esa fecha, más centrados propiamente en la diagnosis del problema que en la misma detección. Su objetivo es el diseño de un algoritmo que, a partir de un modelo para el sistema de comunicación y un conjunto de alarmas ya detectadas, elabore distintas hipótesis de

posibles puntos de fallo y les asigne una figura de mérito a cada una, según una métrica de confianza establecida.

Hood y Ji [204] presentan en 1997 un sistema estadístico adaptativo para la monitorización de redes. Demuestran en su trabajo que es posible detectar automáticamente un fallo sin especificación concreta del modelo de fallo. El sistema aprende el comportamiento "normal" de cada variable bajo estudio conforme los datos con capturados. Las desviaciones de ese comportamiento asumido como normal son detectadas como "anomalías" y dicha información se combina en una red bayesiana para su análisis. Dado que las relaciones entre los nodos de la red bayesiana y las variables de la MIB (base de datos información de gestión *Management Information Base,* vinculada al protocolo de gestión de red SNMP) son complejas y probablemente no del todo conocidas, proponen un modelo simple que no presupone a priori relaciones entre las diferentes funciones de red. Los beneficios de este planteamiento incluyen la habilidad para detectar fallos de tipología desconocida, la capacidad para correlar información en espacio y tiempo así como la detección de fallos proactivamente.

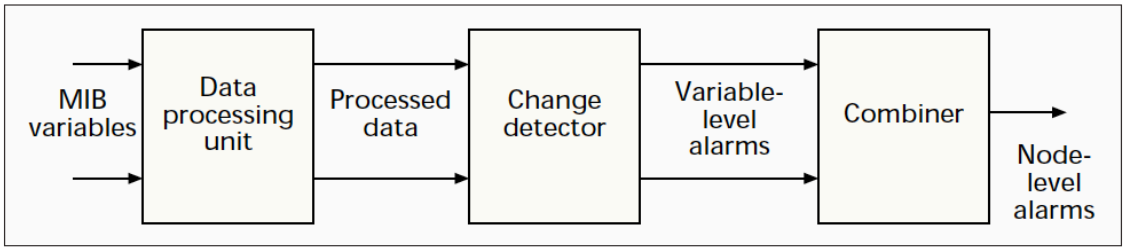


Red Bayesiana para la detección de fallos. [204]

Posteriormente en 1998 y 1999 Thottan y Ji [199], [205] avanzan en este esquema y proponen de nuevo un método proactivo para la detección de anomalías. Recordemos que la proactividad en la detección de anomalías implica la predicción de problemas en la red antes de que éstos causen fallos "severos" en su funcionamiento. La proactividad, en contraposición a la reactividad, que implica acciones después del fallo, posibilita mantener índices de disponibilidad más alto al tratar de "prevenir" averías siendo capaces de detectar precozmente sus "síntomas" cuando estos están aún surgiendo. La detección proactiva es llevada a cabo, de nuevo, mediante el análisis del comportamiento estadístico de algunas de las variables incluidas en las MIB. Las anomalías susceptibles de ser detectadas proactivamente pueden ser, entre otras, el descenso de prestaciones, las averías en dispositivos de red, los fallos en

servidores, la degradación de los medios de transmisión y las tormentas de difusión. En ambos trabajos Thottan y Ji caracterizan los cambios estadísticos en las variables de la MIB que preceden a la ocurrencia del fallo. Para ello utilizan modelos estacionarios autoregresivos (AR) que ya habían sido utilizados para describir series temporales. Posteriormente realizan un test *Generalized Likelihood Ratio (GRL)* secuencialmente sobre las distintas hipótesis propuestas (H_0: No existe cambio entre los segmentos temporales en la variable considerada, H_1: Existe cambio). El algoritmo de detección opera a nivel de cada variable objeto de estudio por separado y consiste en dos pasos sucesivos:

- Detección de cambios dentro de la serie temporal obtenida mediante mediciones de tráfico.
- Determinación de si el cambio detectado corresponde a una situación anormal.



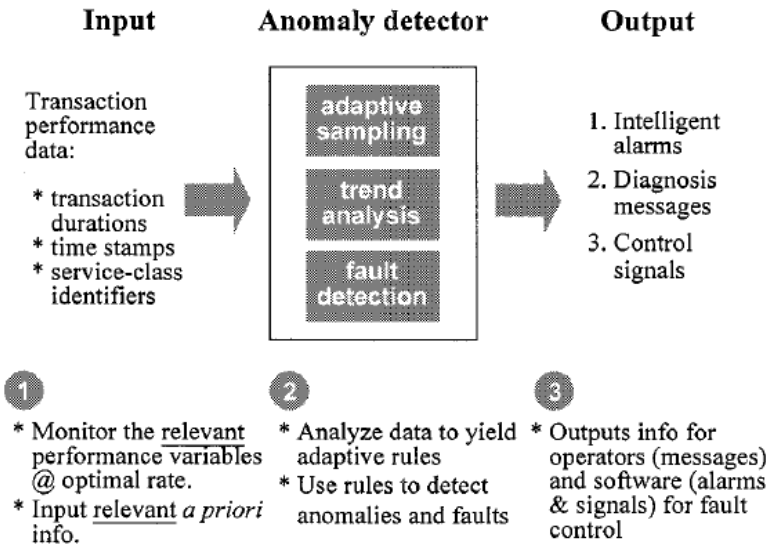
Modelo de detección de anomalías. [199]

A partir de la información extraída anteriormente se combinan las alarmas declaradas como tales mediante una red de inferencia bayesiana, de hecho la misma que aparece en [204], que posibilita correlar la información entre las diferentes variables objeto de análisis. La propuesta implica un modelado preciso del comportamiento considerado "normal" lo que conlleva disponer de un gran número de muestras de prueba. El otro problema es la escala de tiempos en la que se efectúa la detección, esto es, la "ventana de observación", que está a su vez condicionada por la frecuencia de muestreo de las diferentes variables consideradas en el estudio.

En el año 2000 Ho *et al* [198] proponen un detector de anomalías para redes transaccionales. Este tipo de redes están caracterizadas por transmisiones simples (un único paquete de información que contiene una consulta u orden y otro de respuesta con el resultado de la operación) y de corta duración (del orden de segundos) entre un conjunto de terminales y servidores. El ejemplo típico de redes transaccionales es el servicio de tarjetas de crédito que emplea mensajes de longitud reducida entre los terminales de lectura de tarjeta situados en los comercios

(datáfonos) y los centros de compensación correspondientes. El método de detección propuesto está basado en tres etapas:

- Muestreo de la red para resaltar anomalías en el servicio o en la misma red.
- Establecimiento de perfiles basales extraídos a partir de datos históricos.
- Comparación entre los datos muestreados en la red y los perfiles basales para detectar anomalías y posteriormente estimar el diagnóstico del fallo en la red.

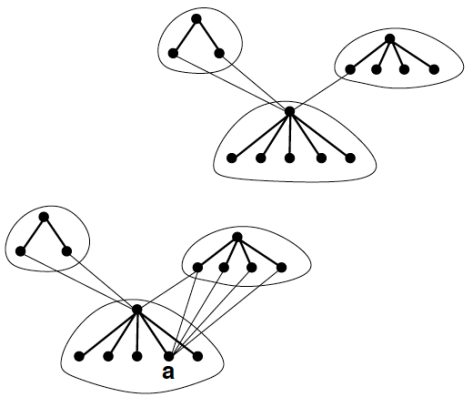


Entradas y salidas de un detector de anomalías orientado a redes transaccionales [198]

El factor diferencial con estudios previos reside, fundamentalmente, en que los perfiles y el muestreo de la red no son generales, sino que son específicos para cada tipo de servicio que ofrece la red transaccional y por lo tanto no son intercambiables. Además, establece varios umbrales superiores e inferiores a partir de los perfiles basales obtenidos anteriormente para declarar el fallo en función de determinadas condiciones periódicas y ajenas a la red. Así pues, existen diferentes umbrales de alarma en función de si el día es hábil, o se trata de un sábado o domingo o un período vacacional. Estos umbrales se establecen a partir del valor de la mediana de los datos históricos recopilados a lo largo del tiempo. Las prestaciones a priori de los diferentes servicios que se establecen a partir de los patrones anteriores juntamente con las desviaciones de los mismos, tanto en magnitud como en duración, conforman un conjunto de criterios de fallo. La variable principal que se emplea para la detección de fallos es la intensidad de tráfico cursado por la red para cada tipo de servicio disponible. La propuesta deja pendiente para futuras investigaciones el método utilizado para el establecimiento de los umbrales de fallo y la aplicación de esta metodología a otros entornos de redes, tales como redes inalámbricas y redes IP en las que las degradaciones de las prestaciones y fallos pueden afectar severamente la

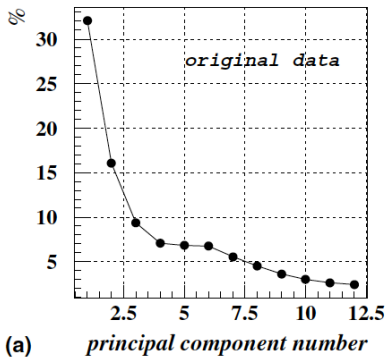
calidad del servicio y la disponibilidad de la red, al tratarse de redes con tiempos de respuesta, por lo general, más exigentes que en caso de las transaccionales.

En 2001 Niggemann *et al* [9] introducen en la detección de anomalías en redes un nuevo elemento, en nuestro criterio muy interesante y que se situaba hasta este momento en un segundo plano: la representación visual del trafico. Si bien la aplicación que proponen se circunscribe a la detección de intrusos (IDS), este aspecto puede ser considerado, como sabemos, un caso particular del problema general de detección de anomalías. La visualización propuesta permite obtener conclusiones a alto nivel sobre el tráfico analizado, incluyendo patrones de tráfico cursado, análisis de la topología de la red o, como ya ha sido mencionado, la detección de ataques de negación de servicio. Es especialmente interesante el modelado que efectúa de la red de comunicación: Los nodos de la red forman un conjunto de puntos y las líneas físicas un conjunto de segmentos que los interconectan. Así pues, la información contenida en la matriz de tráfico puede ser interpretada como un conjunto de "pesos" asociados a dichos bordes del gráfico. En particular, la matriz de tráfico define una asignación para cada línea existente del volumen de tráfico entre los nodos que relaciona. Este gráfico con pesos es denominado "gráfico de tráfico" y constituye el punto de partida de la propuesta y al que se le aplica un análisis de *clusters* que posteriormente es representado visualmente. La aplicación del análisis de *clusters* constituye un hito destacable en la aplicación de métodos estadísticos multivariantes en la bibliografía consultada. La estructura visual obtenida mediante la propuesta realizada es, por supuesto, variable en el tiempo y será desde la comparación entre las representaciones correspondientes a diferentes intervalos de tiempo desde donde se extraerán las conclusiones sobre el tráfico cursado y la ocurrencia de posibles incidencias en la red objeto de estudio.



Estructura de comunicación típica y con un rastreo desde nodo "a" [9]

En el año 2002, Akritas *et al* [206] efectúan una interesantísima y muy relevante aportación al tema: introducen por vez primera (en contra de lo que se considera, como veremos posteriormente) el Análisis de Componentes Principales (ACP) en el estudio del tráfico en redes de comunicación. Si bien los autores no dirigen su propuesta específicamente a la detección de anomalías, sino al análisis de tráfico, el hilo conductor de su estudio es similar al de otros artículos posteriores que lo aplican a la detección de anomalías. Utilizan el ACP para determinar la dimensionalidad del modelo del tráfico en la red, y así poder reducir la dimensionalidad del espacio original. En su estudio proponen, además, un modelo que permite generar una serie temporal, mediante una red neuronal, que simula el tráfico real en la red, demostrando así que el modelo que proponen funciona correctamente. Aunque no directamente, ya apuntan en algunas partes de su trabajo a la posibilidad de utilizar su propuesta para la detección de anomalías (de hecho serían errores en la predicción de su modelo). Utilizan el ACP como verificación del buen comportamiento del modelo de tráfico que proponen, confirmando que las dimensiones tanto del tráfico real como el generado por al red neuronal coinciden.



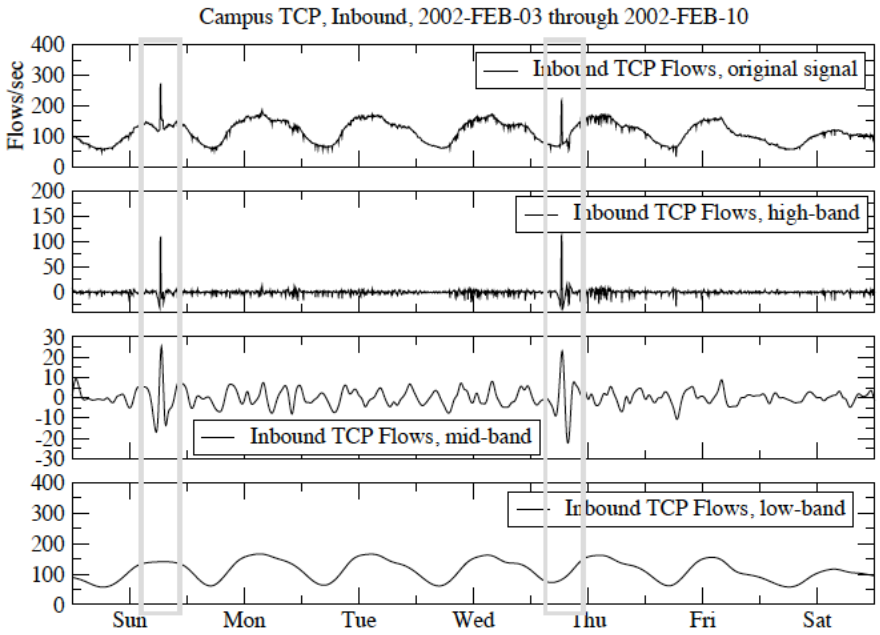
Scree-plot de los datos originales [206]

Barford *et al* [207] en 2002 inciden una vez más en la necesidad de la identificación rápida y precisa de posibles anomalías en las redes de comunicación. Colocan dentro de la categoría de anomalías tanto las pérdidas de prestaciones, los fallos en los enlaces, como los ataques a la red. Los autores aplican su propuesta a cuatro tipos diferentes de incidencias:

- Errores en la red debidos a fallos físicos o lógicos, tales como, caídas de enlaces de comunicación o defectos del software interno de los equipos electrónicos.
- Ataques de negación de servicio (*DoS*) llevados a cabo mediante el envío masivo de paquetes a un equipo víctima.

- Incidencias temporales debidas a procesos no habituales en la red, por ejemplo la distribución de un paquete de actualización.
- Problemas con la propia infraestructura de captura de información que ocasionen la pérdida de información.

Barford propone la captura de información de tráfico desde agentes SNMP y su análisis posterior. El análisis propuesto se basa en la forma de las representaciones en función del tiempo de las variables objeto de estudio. Se consideran las mismas como "señales" y se utiliza un método de análisis de forma de ondas (*wavelet analysis*) de aplicación habitual en teoría de la señal. Como sucede en el caso de las señales, es posible combinar representaciones temporales y frecuenciales de dichas formas de onda. Para ello consideran en su análisis como información de partida la cadena de mediciones de tráfico de la red que es procesada como si de una señal genérica se tratase. El análisis de *wavelets* que proponen como herramienta principal organiza la información sujeta a estudio en estratos jerárquicos cada uno de los cuales mantiene el tiempo como su variable independiente (o la frecuencia si se trata de un análisis en el dominio de la frecuencia). El proceso en sí mismo se realiza en dos pasos complementarios: el primero consiste en el análisis/descomposición de la forma de la señal y el segundo en su inverso, el proceso de reconstrucción/síntesis. El objetivo del subproceso de análisis es extraer de la señal original la anteriormente mencionada jerarquía de subseñales componentes de la misma. Este proceso de descomposición puede ser llevado a cabo tanto en el dominio del tiempo como en el de la frecuencia. El posterior proceso de reconstrucción tiene como objetivo hacer posible la supresión de información no relevante que haya sido aislada durante el proceso de análisis y cuyo efecto se desee eliminar. La señal resultante de la reconstrucción puede ser de nuevo objeto de análisis sin que el efecto eliminado sea considerado. Lógicamente los análisis efectuados sobre las formas reconstruidas pueden contener efectos no deseados (artefactos) relacionados con el proceso de reconstrucción al que han sido sometidas y que nada tienen que ver con la situación real de la red. A vista de pájaro el proceso propuesto se resume en la identificación en las mediciones de "elementos constructivos" a partir de los cuales podamos extraer conclusiones sobre el conjunto agregado de todos ellos. Los autores concluyen que su propuesta es capaz de detectar diversos tipos de anomalías, si bien es sensible a la configuración concreta de los diferentes parámetros de ajuste disponibles en el algoritmo.

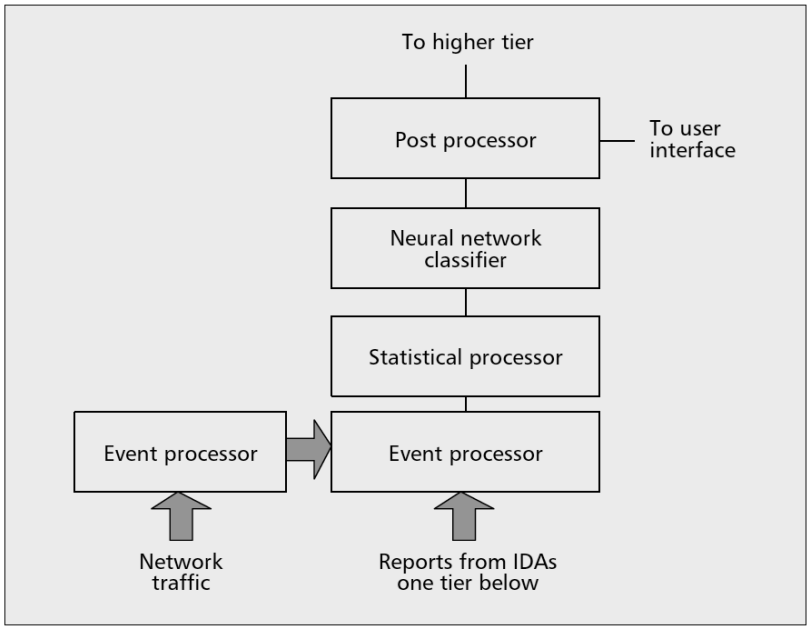


Señal original de una semana, y descomposición alta/media/baja con marca de 2 ataques DoS [207]

En ese mismo año 2002 Manikopoulos y Papavassiliou [208] proponen otro método de detección de anomalías basado en un análisis estadístico. Como en otros casos, la detección de intrusos forma parte del grupo de "anomalías" candidatas a ser detectadas. La propuesta utiliza una red neuronal para la clasificación de patrones e identificación de posibles anomalías. Al tratarse de una red neuronal se precisa un entrenamiento previo de la misma para que funcione correctamente. Las métricas de similaridad que se investigan están basadas en el test de Kolmogorov-Smirnov (*K-S test*) y los modelos de referencia y observados se caracterizan por sus funciones de densidad acumuladas. La principal ventaja que los autores exponen de la utilización del test K-S es la independencia explícita o implícita de hipótesis previas sobre la normalidad de la función de distribución. El algoritmo propuesto es el siguiente:

- La primera fase consiste en un "procesador estadístico" (*sic*) cuya salida es un vector ordenado k-dimensional cuyas componentes son los valores de similaridad entre las funciones de densidad medidas y las funciones de densidad de referencia de los k parámetros bajo estudio recogidos durante una ventana temporal de observación del sistema.
- La segunda fase consiste en el "clasificador multivariante" (*sic*), formado por una red neuronal a la que se le presenta como entrada el vector anterior.

Los autores destacan que la primera fase de procesado estadístico reduce los tiempos de entrenamiento requeridos por la red neuronal y la complejidad de cálculo de la misma.



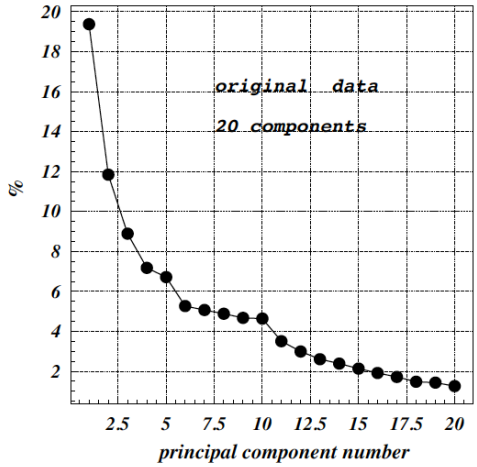
Estructura de un agente para detección de intrusos [208]

Los autores concluyen que el método propuesto puede constituir un eficaz procedimiento de detección de anomalías ya que detecta de manera fiable incidencias que impliquen intensidades anormales sobre el nivel de tráfico basal de tan sólo el 3% ó 5%, lo que, según ellos, constituye un resultado prometedor como alerta temprana de incidencias. Al requerir dos etapas de aprendizaje, por un lado es preciso establecer los patrones correspondientes a las funciones de densidad de referencia para su utilización en la primera fase del procesador estadístico, además es necesario, una vez concluida la determinación de los patrones basales anteriores, entrenar la red neuronal correspondiente al clasificador multivariante.

Los años 2003 y 2004 constituyen una época importante en el análisis de anomalías en el tráfico de redes mediante la aplicación del Análisis de Componentes Principales.

En 2003 Antoniou *et al* [209] vuelve a plantear la utilización del ACP para el modelado del tráfico en redes. El origen ruso de varios de los coautores introduce además la aplicación del muy interesante método SSA-Caterpillar (SSA, *Singular-Spectrum Analysis* [210]). Los autores demuestran que su propuesta es muy eficiente para la comprensión de las principales características de las componentes que forman el tráfico de la red. El ACP permite estimar las contribución de cada una de las componentes principales a la serie analizada. Su análisis demuestra que aproximadamente la mitad de las componentes principales constituyen la parte más relevante de la información de tráfico. Las restantes componentes (que consideran residuales) juegan un papel accesorio como variaciones irregulares que pueden ser interpretadas como ruido estocástico. Los autores concluyen que la simplificación

posibilita el estudio de la complicada estructura original de las serie temporal que representa el tráfico de red y abre múltiples posibilidades a modelos de tráfico dinámicos más realistas y mejor ajustados.



Contribuciones de las componentes principales (en %) al tráfico original. [209]

También en 2003, Shyu *et al* [211] vuelven a plantear la utilización del ACP, pero esta vez ya sí para la detección de anomalías. En su artículo Shyu trata, como hemos visto también anteriormente, las intrusiones como un tipo más de anomalía. Clasifica los detectores de intrusiones en dos tipos: reconocimiento de firmas o de patrones y detección de anomalías. En el primer caso, reconocimiento de firmas, el algoritmo compara los eventos registrados contra una base de datos de firmas. El método identifica un ataque cuando el evento coincide con un tipo registrado. Este es de hecho también el planteamiento que propuso Feather [42]. Por otro lado se encuentran los detectores de anomalías "puros" que se basan en la construcción de un modelo basal a partir del funcionamiento "normal" de la red, con el que contrastan la nueva información recogida de la red, generando una alerta ante desviaciones del modelo inicial. La ventaja de este último método frente la primero es que es capaz de detectar incidentes nuevos, sobre los que no se dispone de información previa, como la que precisan los detectores basados en firmas. En su artículo Shyu propone un esquema de detección de anomalías basado en el Análisis de Componentes Principales y detección de *outliers*. La asunción bajo el modelo propuesto es que los ataques aparecen como *outliers* de los datos basales. El uso ACP presenta varias ventajas destacando que no precisa ninguna hipótesis sobre la distribución estadística de los datos. Por otro lado, como se ha visto en artículos previos, es típico de este tipo de datos el que presenten una elevada dimensionalidad. En el esquema que propone Shyu utiliza ACP robusto para reducir la dimensionalidad y llegar a un modelo de dimensionalidad reducida, lo que reduce los tiempos de cálculo del clasificador. Los resultados experimentales muestran que el método presenta buenas características de

detección con una tasa de falsas alarmas baja, mejorando otros métodos propuestos hasta esa fecha. La autora afirma que el ACP ha sido aplicado hasta esa fecha como una técnica de reducción de la dimensionalidad de los datos, pero no como herramienta de detección de *outliers*. Si bien los métodos gráficos, afirma, son habituales para la detección de *outliers*, no son prácticos para aplicaciones en tiempo real. Como los test de detección de *outliers* precisan que los datos presenten ciertas características estadísticas para que los test de detección sean válidos, proponen un método de detección basado en componentes principales que pueda ser de aplicación en tiempo real y no imponga fuertes restricciones al tipo de datos a utilizar. El funcionamiento del detector propuesto es el siguiente: se efectúa un ACP sobre la matriz de correlación del grupo normal de datos. Se utiliza la matriz de correlación ya que cada característica/variable es medida en diferente escalas. Estos datos de entrenamiento deben encontrarse libres de *outliers* ya que serán posteriormente utilizados para determinar el criterio de alarma. Es también muy importante que esta matriz de correlaciones sea robusta, para lo que la autora utiliza la distancia de Mahalanobis y elimina iterativamente las observaciones que aparenten ser *outliers* hasta que la matriz se estabilice. El clasificador de componentes principales consiste en dos funciones generadas a partir de los factores de las componentes principales consideradas como esenciales y de los factores de las restantes componentes principales, normalizadas por los respectivos valores propios. Para la evaluación del método utiliza la matriz de confusión habitual en estos casos y la función característica de operación de un receptor (ROC - *Receiver Operating Characteristic*).

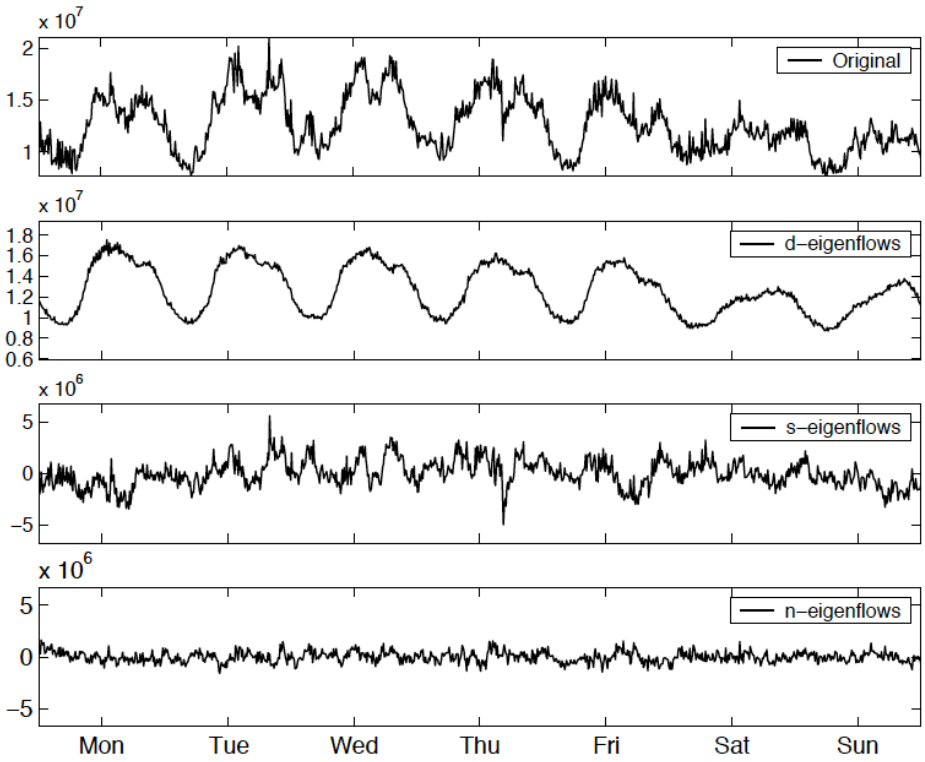
		Predicción de conexión	
		Ataque	Normal
Conexión Real	Ataque	Detección Correcta	Falso Negativo
	Normal	Falsa Alarma	Verdadero Negativo

Métricas de confusión para evaluación de los ataques. [211]

Y por fin, entre 2004 y 2005 tiene lugar la que se considera la "presentación en sociedad" del Análisis de Componentes Principales para el análisis de tráfico en redes y detección de anomalías e intrusos. Aunque, como hemos visto, otros autores ya habían propuesto antes su aplicación en este contexto [206], [209], [211]. Lakhina, Crovella y Diot (en el primero de los artículos acompañados por otros autores) publican una serie de artículos todos ellos basados en el ACP para el estudio del tráfico en redes y que a continuación reseñaremos brevemente.

En junio de 2004 Lakhina, Papagiannaki, Crovella, Diot, Kolaczyk y Taft presentan, en la prestigiosa conferencia SIGMETRICS que tuvo lugar ese año en Nueva York, su

artículo "*Structural analysis of network traffic flows*" [212]. Los autores proponen el análisis de un juego de datos consistente en series temporales de matrices Origen-Destino (O-D) capturados de la red Abilene y de la red europea Sprint. Un análisis preliminar del juego de datos de la red Abilene fueron ya publicados por los mismos autores en 2003 en un documento técnico de la Universidad de Boston [213]. Una de las primeras diferencias del trabajo de Lakhina *et al* es el alcance de los juegos de datos utilizados: como ellos mismos indican, los trabajos existentes hasta esa fecha estaban enfocados en el estudio del tráfico en un único enlace de manera aislada. Sin embargo en un amplio rango de problemas de ingeniería de redes los investigadores requieren modelar y analizar el tráfico en varios enlaces simultáneamente: en ingeniería de tráfico, estimación de la matriz de tráfico OD, detección de anomalías, detección de intrusos/ataques, predicción de tráfico y planificación de capacidad. En todos ellos la matriz de Origen-Destino (O-D) juega un papel fundamental. El principal reto que presenta estudiar las matrices O-D es la estructura multivariante que presentan con una elevada dimensionalidad. Por ejemplo, una red de tamaño medio, afirman, puede transportar cientos de pares O-D, cuyas series temporales tienen centenares de dimensiones. Como los autores indican el ACP, también conocido como descomposición de Karhunen-Loève y descomposición en valores singulares (sic), es la técnica más común para analizar estructuras de alta dimensionalidad: dado un objeto con una elevada dimensionalidad y su espacio de coordenadas asociado, el ACP descubre un nuevo espacio de coordenadas que es el mejor de posible utilización para reducir la dimensionalidad del objeto en cuestión. Una observación muy importante es que los autores utilizan un ACP un poco diferente, en comparación con el tradicional utilizado en Análisis Multivariante: se realiza un ACP sobre series temporales, esto es, se descompone una serie temporal en sus series temporales "principales" (ver Bouhaddou *et al* [214]). Salvo esta "peculiaridad" el ACP utilizado no difiere del "tradicional": se utiliza el *scree-plot* para determinar la dimensionalidad del espacio destino, descompone el espacio original en sus componentes principales (que en este caso se corresponden con series temporales), se analiza la fracción de la varianza capturada por cada componente principal (aquí tiene una interpretación de "energía", al tratarse de series temporales), etc. Los autores comprueban, además de la bondad del espacio de dimensión reducida para capturar la información esencial de la estructura del tráfico, la estabilidad en el tiempo de la estructura extraída de las componentes principales del conjunto de matrices O-D. Concluyen que la estructura obtenida por el ACP presenta tres características, una tendencia determinista, una parte correspondiente a impulsos y otra restante de ruido, y que corresponden con la descomposición en sus componentes principales.



Descomposición de una serie temporal O-D en sus tres componentes principales. [212]

En agosto de ese mismo año Lakhina *et al* presentan un nuevo artículo en la conferencia SIGCOMM en Portland [215] centrado en el diagnóstico de anomalías en redes de área expandida. Remarcan de nuevo que las anomalías pueden deberse a incidentes no intencionados o maliciosos, este último supuesto se correspondería con un ataque o intrusión. La diagnosis, que es lo que proponen en su artículo, constituye un problema importante, debido a que las causas de una anomalía pueden variar considerablemente. Indican igualmente que si bien en la literatura la caracterización del tráfico es un tópico muy abordado, no ha sido así el estudio de las posibles anomalías. Su propuesta se centra en las anomalías "de volumen" en flujos Origen-Destino (*O-D flows*): Como sabemos, una red troncal típica está compuesta por nodos interconectados por enlaces; definimos un flujo Origen-Destino como el trafico que entra en la red a través de un nodo Origen y sale de la red por otro nodo Destino. Los autores se refieren a "anomalías de volumen" como a súbitos incrementos o decrementos del tráfico en un flujo determinado. Utilizan de nuevo los mismos juegos de datos que en [212], [213] y un planteamiento similar a partir de la descomposición de las series temporales de matrices O-D en sus componentes principales vía ACP.

La diagnosis de anomalías de volumen se descompone en tres pasos diferenciados:

1. Detección: consistente en la identificación de los puntos de la serie temporal en los cuales la red esta experimentando una anomalía. Evidentemente el

algoritmo encargado de esta fase debe tener un buen desempeño, esto es, una probabilidad de detección elevada, manteniendo la probabilidad de falsa alarma reducida.

2. Identificación: consistente en la selección del tipo de anomalía que realmente está sucediendo, a partir de un conjunto de posibles anomalías tipificadas. En su artículo, más que el tipo de anomalía, lo que identifican es el par O-D concreto responsable de la anomalía.
3. Cuantificación: como la estimación del volumen de tráfico por encima o por debajo de lo que sería normal, lo que ofrece un indicador de la importancia de la anomalía sucedida.

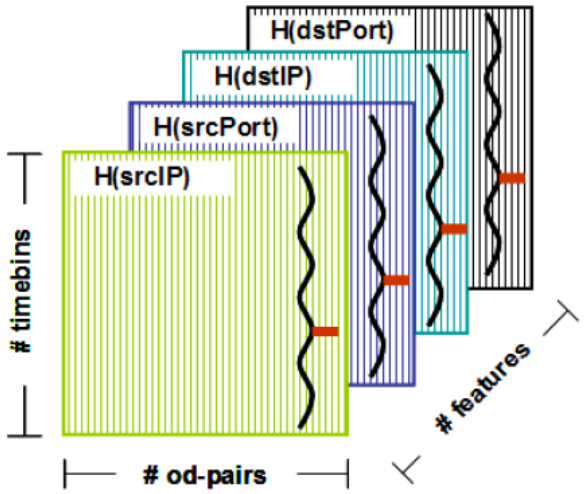
Esto es, el algoritmo debe ser capaz de detectar el tiempo en el que sucede la anomalía, identificar el flujo O-D responsable del evento y cuantificar el tamaño de la anomalía con relación a ese flujo O-D concreto.

En este artículo los autores utilizan el ACP sobre las series temporales para separar el tráfico considerado normal del considerado anómalo, permitiendo responder satisfactoriamente a todas las cuestiones planteadas: detección, identificación y cuantificación de la anomalía. Proponen igualmente que la propuesta que realizan, y que denominan “método del subespacio”, funcionaría igualmente con otro tipo de anomalías, y no solo de volumen.

En el mes de octubre de ese mismo año 2004, Lakhina *et al* presentan otro artículo en la conferencia IMC (*Internet Measurement Conference*) que tuvo lugar en Taormina (Sicilia) [52]. La principal novedad de este tercer artículo de la serie es la identificación no del flujo O-D responsable de la anomalías, sino del tipo concreto de anomalía ocurrida. Plantea 8 tipos distintos de anomalías identificables por el método del subespacio (ACP) así como las características observadas que permiten su clasificación (ver apartado previo dedicado a Ethernet, para detalles).

Utilizan los autores dos series de datos de la red Abilene entre el día 7 al 13 de abril de 2003 y entre el 8 de diciembre y el 28 también del 2003, en intervalos de 5 minutos, para una matriz O-D de 11x11 nodos. Es muy interesante la aportación que realizan indicando el número de anomalías descubiertas en el análisis, así como el tipo de las mismas. Dado que el juego de datos está disponible, esta información puede ser de gran utilidad para posibles comprobaciones de otros algoritmos. De hecho, este juego de datos será, como veremos, muy utilizado a partir de ese momento.

Por último Lakhina *et al* en 2005 [216] introducen el concepto de "minería de datos" dentro de la búsqueda de anomalías en los patrones de tráfico en redes. La propuesta se basa, una vez más, en el análisis de la información de origen y destino de los paquetes transmitidos. Para cada una de las cuatro características posibles (dirección IP origen, dirección IP destino, puerto origen, puerto destino) se calcula su entropía y de esa manera se construye una tabla de tres vías en la que cada elemento denota el valor de la entropía de una característica para un determinado intervalo de tiempo y para un flujo de datos (*flow*) entre un origen y un destino concreto.

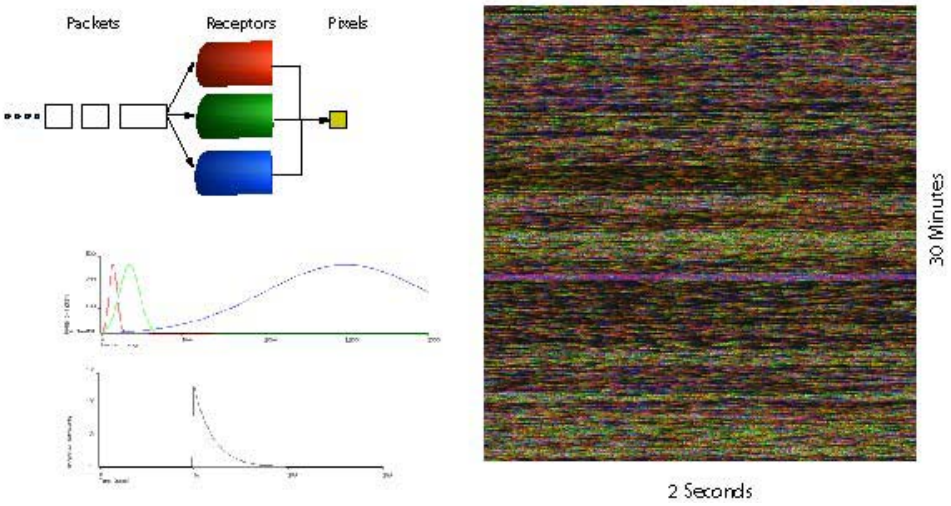


Datos multivariantes, multivía para analizar. [216]

A esta tabla multivariante los autores aplican un método de "desplegamiento" (*unfolding*) o de colapsabilidad para transformar la matriz de tres vías en una clásica de dos vías sobre la que aplicar su propuesta. Los autores se basan en el artículo de Kiers sobre análisis multivía [194]. La propuesta conceptual de desplegamiento que exponen consiste básicamente en una reordenación de la tabla de tres vías inicial para obtener una tabla de dos vías, en una técnica que Kiers denomina "matrización" [194] para evitar confusión con el método "unfolding". A dicha matriz posteriormente le aplican el método del subespacio, esto es, se aplica un Análisis de Componentes Principales a esta matriz para obtener un subespacio de dimensión reducida sobre el que proyectar los datos objeto de estudio. Se supone que los datos provenientes de situaciones derivadas de anomalías presentarán proyecciones diferentes que aquellos que se deban a situaciones normales. Este conjunto de puntos es sometido posteriormente a un análisis de *clusters* como algoritmo de clasificación no supervisada. Los autores concluyen que la entropía es una métrica efectiva para la captura de cambios de comportamiento inusuales inducidos por anomalías en las distribuciones de las características de tráfico, demostrando que métodos

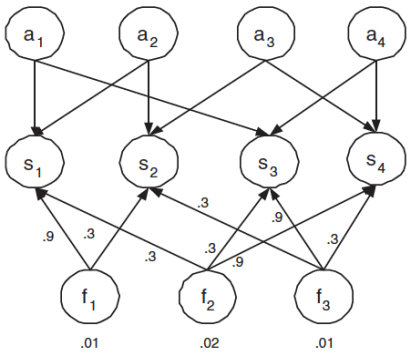
multivariantes posibilitan la detección de cambios anormales en la operación de la red a través del análisis simultáneo de múltiples variables características. En este artículo, en comparación con los precedentes, las aportaciones más importantes se corresponderían con el incipiente intento de aplicar métodos de tres vías (incipiente por el hecho de que se limita a transformar el problema en uno de dos vías y no a abordarlo con técnicas multivía propiamente dichas) y la utilización de la entropía, además, de la aplicación de métodos de cluster para establecer asociaciones.

En 2005 Rosenbluth y Pucci [10] presentan otra propuesta para representar visualmente el tráfico cursado en la red y detectar anomalías en su funcionamiento. El planteamiento es simple y a la vez eficaz: una imagen se compone de puntos o *pixels* a los que se les puede asignar un color. En el ejemplo que ilustra su póster asignan dicho color en función de la longitud de los paquetes cursados en la red. La imagen bidimensional se compone de puntos, cada línea horizontal corresponde a un intervalo temporal en sentido de izquierda a derecha de, por ejemplo, dos segundos de duración. Cuando la línea horizontal termina se pasa a la siguiente en sentido vertical descendente. El número de líneas verticales disponibles determinará el intervalo temporal al que corresponderá la imagen obtenida (30 minutos para 900 líneas de 2 segundos de duración). La asignación de colores de cada *pixel* se realiza mediante inferencia bayesiana y puede corresponder a diferentes atributos del tráfico, la longitud de los paquetes, a la fragmentación de los mismos, entre otros posibles. La representación obtenida por este sistema contiene patrones estructurales de la actividad de la red, que pueden ayudar a reconocer eventos en el entorno bajo estudio. Los autores indican que si bien la mayoría de herramientas gráficas para representar visualmente mediciones en las redes tienen como objetivo principal ayudar a expertos humanos en la realización de tareas tales como la detección de comportamientos anómalos, formulación de nuevos patrones de utilización de la red y/o aplicaciones, diagnóstico de fallos y control de tráfico, su propuesta genera representaciones que pretenden ser objeto posteriormente de un análisis automatizado. Indican que están realizando estudios sobre la aplicación a estas representaciones de algoritmos de reducción de dimensionalidad.



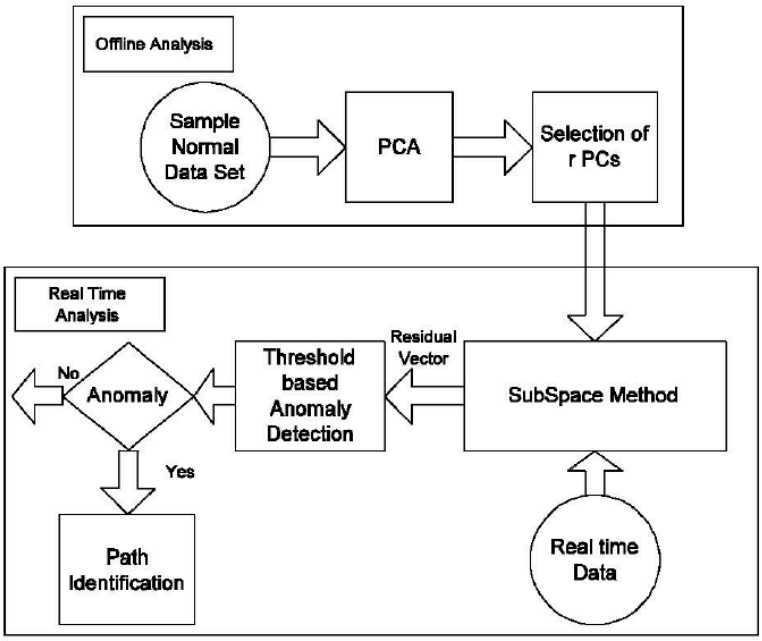
La imagen de la derecha representan 30 minutos de tráfico TCP/IP. Cada color representa una categoría de tamaño de paquetes. Las curvas representan las respuestas de los receptores. [10]

También en 2005 Tang *et al* presentan en un congreso [217] y posteriormente publican en 2008 [218] un método para la localización de fallos en redes de comunicaciones, aunque su trabajo está más asociado a la parte de "identificación" que a la de "detección", propiamente dicha. El sistema diagnostica el fallo relacionando las alarmas que se producen con los fallos en la red. Aquí el problema no es detectar el error, sino diagnosticar el problema: un único problema puede originar múltiples errores. Por lo general, atajar individualmente las incidencias que se producen no suele ser productivo, lo "interesante" es resolver el problema que los produjo. La propuesta que realizan mejora otras técnicas "activas" existentes en la literatura reduciendo las pruebas que requiere para un diagnóstico. Su modelo de Síntoma-Fallo-Acción es una interesante propuesta para asociar (s)íntomas, que pueden ser comprobados mediante una o varias (a)cciones, así como la asociación de (f)allos con (s)íntomas. En la versión publicada en 2008 [218] los autores incluyen la posibilidad de aplicar el método que proponen a redes "superpuestas" (*overlay*) sobre las que no se dispone de un control directo sobre la red de comunicaciones en la que se apoya.



Modelo Acción – Síntoma – Fallo. [217]

En 2006 Androulidakis *et al* [219] publican un estudio sobre el efecto del muestreo en la técnicas de detección de anomalías. Aplican su estudio a dos técnicas ampliamente utilizadas para la detección de anomalías: el denominado método de detección del punto de cambio, cuyo objetivo es determinar si la serie temporal observada es estadísticamente homogénea y si no lo es, detectar el punto en el tiempo en el que el cambio sucede, y el Análisis de Componentes Principales. El primer método se utiliza cuando estamos trabajando con una sola serie temporal / un solo enlace, mientras que el ACP se utiliza para el caso de múltiples series / múltiples enlaces.

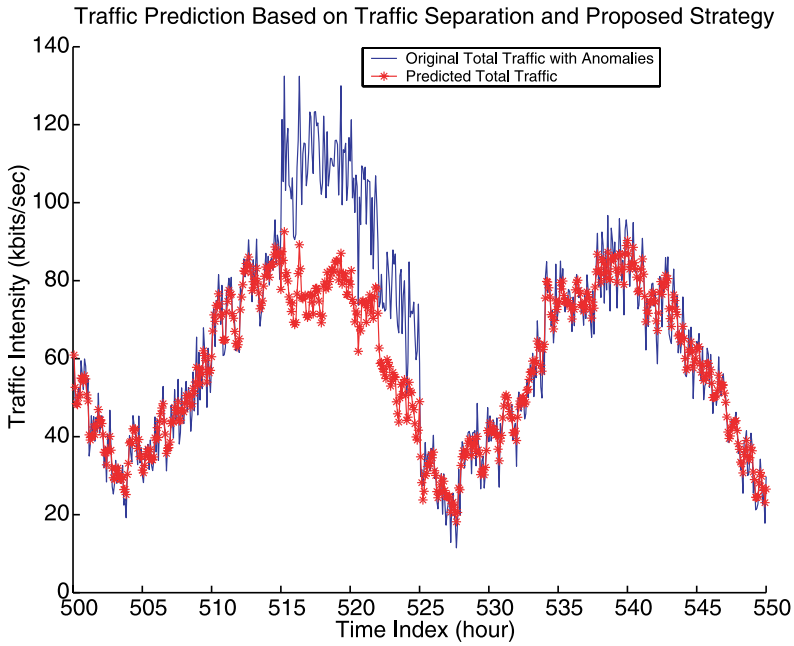


Representación a alto nivel de la metodología propuesta. [219]

Los autores analizan el efecto de varios tipos de muestreos, concluyendo que el muestreo sistemático, el más común, deteriora significativamente las prestaciones de los algoritmos cuando la frecuencia de muestreo es baja y la detección depende de ciertas características de los paquetes (por ejemplo, los indicadores TCP). El muestreo sistemático es aquel en el que los puntos de muestreo son seleccionados de acuerdo a una función determinista, los autores consideran en su estudio solo puntos equiespaciados. Aunque el muestreo sistemático es sencillo y proporciona buenos resultados para las funciones de gestión básicas, aparentemente es inadecuado para la detección de anomalías y se precisan técnicas de muestreo más sofisticadas. Es más, cuando se utilizan métricas basadas en flujo, como direcciones IP o número de flujos, las prestaciones de los algoritmos de detección residen fundamentalmente en la frecuencia de muestreo utilizada durante la detección. Con relación a los métodos basados en ACP, que utilizan métricas basadas en paquetes o en flujos, se observa que la efectividad reside en la frecuencia de muestreo y no en el método de muestreo.

En 2006 Farraposo *et al* [220] proponen un método de detección de anomalías basado en el análisis de las siguientes variables: número de paquetes, número de bytes, número de flujos (*flows*), dirección IP de origen y destino y puertos IP de origen y destino. Su hipótesis de trabajo es que cualquier anomalía de tráfico incidirá en la variabilidad de alguno o varios de dichos parámetros. Su objetivo secundario es detectar qué flujo (*flow*) es responsable de dichas variaciones. El punto de partida del método es la detección del intervalo o intervalos de tiempo en los que tiene lugar una variación "significativa" de alguno de los parámetros anteriores, que son analizados independientemente. No se detalla en el artículo como se declara que una variación de alguno de los parámetros es "significativa", esto es, no se especifica que tipo de contraste se realiza, aunque la explicación se asemeja al método de detección del punto de cambio [219]. Una vez localizados los intervalos (o intervalo) de interés aplica diferentes niveles de agregación en las direcciones IP tras lo que realiza un nuevo test para identificar el flujo (*flow*) responsable de la incidencia. El trabajo se presenta con un ejemplo de detección de una ataque de negación de servicio (DoS), como en algunos casos anteriores, caracterizado por un incremento en el número de flujos establecidos por unidad de tiempo en comparación con otros períodos.

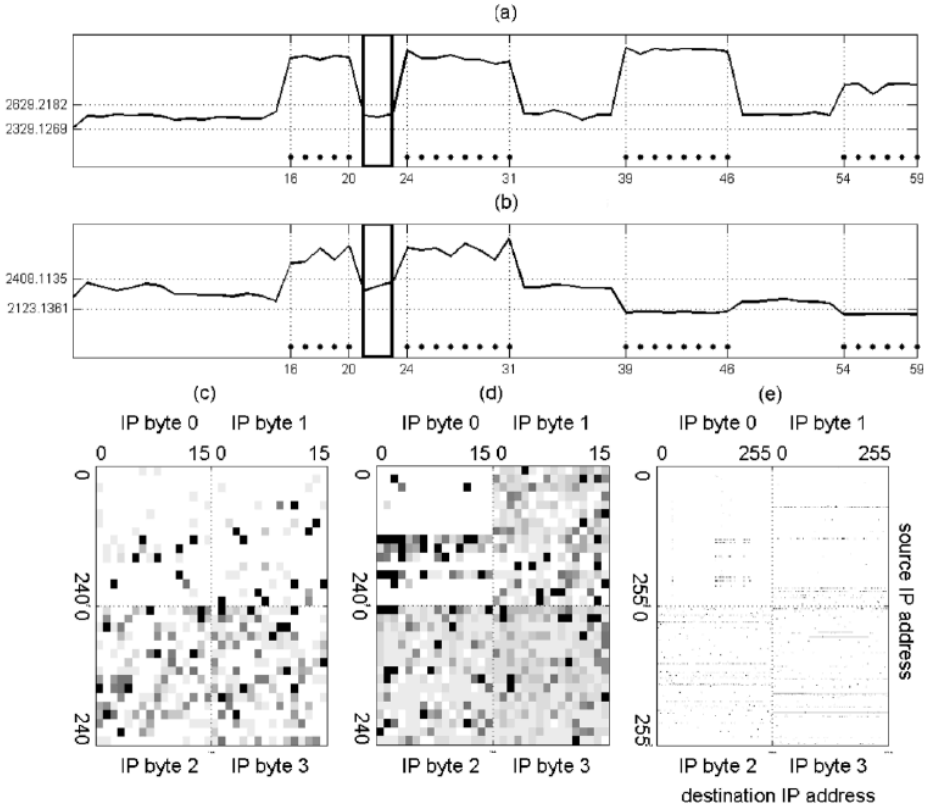
En 2006 Jiang y Papavassiliou [221] proponen un método, que será utilizado en otras propuestas posteriores, basado en la descomposición de la intensidad de tráfico cursado en componentes de alta y baja variabilidad. Esta descomposición de la intensidad de tráfico se efectúa en el dominio frecuencial, con una componente basal, más a largo plazo, y otra a corto plazo, más impulsiva. La separación entre ambas componentes la realizan mediante un filtro del Kolmogorov-Zurbenko (K-Z) que es simple y preserva la información de la señal aún cuando se aplica a muestreo no uniforme y/o con datos faltantes. Los autores utilizan el algoritmo ERAN (*Extended Resource-Allocating Network*) para predecir la componente basal del tráfico. Con respecto a la componente de tráfico a más corto plazo, y como hemos dicho más dinámica, proponen un modelo ARMA (*AutoRegressive Moving Average*) para estimar esta componente del tráfico. Así, con ambas componentes, estiman el tráfico total y a partir de esa predicción, y las desviaciones observadas, detectan las posibles anomalías.



Predicción de tráfico basada en la separación del tráfico y estrategia propuesta bajo estado anómalo.[221]

En ese mismo año 2006 Kim y Reddy [11] realizan una propuesta también basada en la construcción de imágenes a partir del tráfico de la red, si bien estas imágenes son procesadas posteriormente de manera automática para la detección de anomalías. Las imágenes bidimensionales obtenidas son aparentemente similares a las de Rosenbluth [10] aunque la diferencia principal con aquellas reside en el procedimiento utilizado para formar la imagen: si Rosenbluth [10] utilizaba un esquema temporal relativamente simple, Kim y Reddy utilizan información contenida en las cabeceras IP (dirección de origen/destino, puertos, protocolo…) para posicionar los valores de la métrica considerada en un plano, siendo la evolución de los "fotogramas" obtenidos la dimensión temporal. La información o métrica objeto de estudio puede ser el volumen de tráfico cursado en caracteres (*bytes*), el número de paquetes, el número de flujos (*flows*) o cualquier otra información que se considere de interés. Esta información es representada visualmente de manera que cada punto de la imagen (*píxel*) es asociado con, por ejemplo, el volumen de tráfico originado en una dirección IP concreta, los paquetes enviados a un destino particular, el tráfico entre un par de direcciones origen-destino o los puertos IP participantes en la comunicación. La representación permite la visualización de la información de tráfico como si de un fotograma de una película se tratase. Así pues, es posible aplicar a esta "película" técnicas específicas de análisis de video para descifrar patrones de tráfico, cambios en "escenas" o incluso algoritmos de compresión para facilitar su almacenamiento. El algoritmo de detección de cambios en las imágenes que se propone es diferente si se está realizando un análisis en tiempo real o no. En el primer caso utiliza la varianza de la intensidad de

los puntos que forman la imagen, en el segundo proponen la transformada coseno discreta (*DCT*). Esta dualidad es justificada por la elevada carga computacional que requiere el análisis en tiempo real utilizando la transformada coseno discreta. El ejemplo sobre el que verifican el correcto funcionamiento de su propuesta, es una vez más, un ataque de negación de servicio (*DoS*).



Visualización del tráfico. Rectángulos en (a) y (b) marcan los puntos actuales de muestreo. Las líneas puntos en (a) y (b) indican las anomalías. Las imágenes (c) y (d) muestran la intensidad de tráfico en las direcciones IP de origen y destino, respectivamente. El nivel de gris representa la intensidad del tráfico. El gráfico (e) muestra simultáneamente el nivel de tráfico en origen y destino por IP. [11]

En 2006 Kwitt y Hofmann [222] proponen un método de detección de anomalías basado en el Análisis de Componentes Principales (ACP) derivado de la propuesta de Shyu *et al* [211] antes revisada. Los autores exponen que el mayor problema que presenta el Análisis de Componentes Principales para su aplicación en la detección de anomalías, es su robustez. Argumentan que los resultados en sí mismos del ACP están contaminados al ser éste realizado en la presencia de *outliers* en los datos. Su propuesta se basa en la localización de los *outliers* y su eliminación previamente a la realización del Análisis de Componentes Principales. Para la identificación de los *outliers* postulan la utilización del Elipsoide de Volumen Mínimo (*MVE*) o el Determinante de Covarianza Mínima (*MCD*) como sustitutos de la distancia de Mahalanobis utilizada por Mei-Ling Shyu *et al* [211] para mejorar la robustez del ACP en la detección de anomalías. Los autores se inclinan abiertamente por la utilización

del MCD debido a la existencia de un algoritmo más rápido de cálculo, que en el caso del MVE, concluyendo la bondad de su propuesta.

El último trabajo localizado del año 2006 corresponde al presentado por Kim y Kim [223] que proponen una extracción de características basada en la combinación del ACP con el análisis de clusters de k-medias: en primer lugar aplican el ACP para la extracción de características y luego el algoritmo de *clustering* en cada juego de datos de entrenamiento. Basado en este último análisis se extraen las características para las que los factores son mayores que un determinado umbral. Es posible disminuir la redundancia de las características con factores elevados si están en el mismo cluster. A partir de este juego de patrones proponen dos métodos de clasificación no paramétricos, el de Parzen-Windows, para estimar la función de densidad y aplicar una regla de decisión de Bayes, y el k-vecino más cercano (*k-Nearest Neighbor*). Sus resultados muestran que las prestaciones del clasificador dependen del juego de entrenamiento utilizado. El principal inconveniente que remarcan es la elevada carga computacional de los métodos no paramétricos, con la ventaja, eso sí, de poder utilizarse con distribuciones arbitrarias.

En 2007 Samak *et al* [12] proponen un nuevo método para la visualización de tráfico que, al igual que Rosenbluth [10] y Kim [11], utiliza un método para formar las imágenes bidimensionales basado en el mapeado de un espacio multidimensional en uno bidimensional. Este método asigna los valores correspondientes al espacio multidimensional sobre el espacio bidimensional a través de una "ruta". Este proceso se realiza en dos fases diferenciadas, primero se "serializan" unidimensionalmente los valores objeto de análisis. Estos son posteriormente situados en el plano mediante el empleo de una curva de rellenado de espacio (*Space-Filling Curve, SFC*). Como sucedía en casos anteriores, la serie de imágenes bidimensionales obtenida constituye la evolución temporal del estado de la red y la secuencia de las mismas permite extraer conclusiones sobre dicho estado. Concretemos en un ejemplo el método de mapeado propuesto: supongamos que sólo deseamos analizar una característica, la intensidad del tráfico recibido en cada equipo identificado por su dirección IP. Dado que cada dirección IP se compone de un vector de cuatro números que pueden (en general) variar entre 0 y 256 (=16x16) el espacio completo de direcciones IP puede ser visto como un subplano compuesto de cuatro cuadrantes de 16x16 *pixels.* La intensidad de cada *pixel* vendrá determinada, en este ejemplo, por la intensidad de su tráfico. Si consideramos, por ejemplo, que el estudio está limitado a una red de clase C, sólo el último *byte* de la dirección IP será representativo. De esta

manera tendríamos representada la evolución de la intensidad de tráfico en la red como una película de fotogramas de 16x16 puntos. La *SFC* seleccionada entre las disponibles será la que determine como recorrer el espacio de la imagen para asignar los valores de intensidad, o diferentes colores, a cada *pixel*.

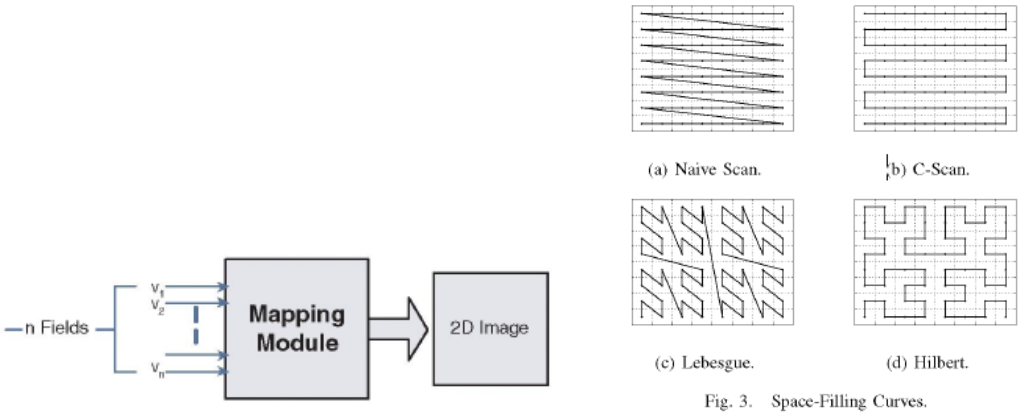


Diagrama de bloques del sensor y curvas de rellenado de espacio (SFC). [12]

Los autores concluyen que el método de representación propuesto presenta dos importantes propiedades para esta aplicación concreta, la actividad es fácilmente percibida, al tender los *pixels* intervinientes a agruparse en *clusters*, y por otro lado el método no genera con facilidad líneas rectas, lo que, según los autores, es ventajoso a la hora de aplicar posteriormente técnicas genéricas de tratamiento de imagen. Finalmente proponen, entre otras cuestiones, analizar la posibilidad de utilizar el dominio frecuencial para mejorar la efectividad del método propuesto.

En 2007 Agrawal *et al* [224] proponen diferentes técnicas para construir una infraestructura de monitorización pasiva para el diagnóstico de anomalías a nivel de enlace entre nodos.

El objetivo de los autores es la detección de anomalías tales como un excesivo número de paquetes perdidos o incrementos del retardo en la transmisión. La estrategia contempla la ubicación de un número reducido de dispositivos de monitorización pasiva en posiciones estratégicas del a red que serán los encargados de examinar el tráfico que atraviesa dichos enlaces.

Su propuesta engloba tres cuestiones diferentes:

1. Ubicación del nodo de prueba: La ubicación del nodo encargado de monitorizar el tráfico no es una elección simple y es dependiente de la topología específica de la red, bien sea esta mallada o en árbol.

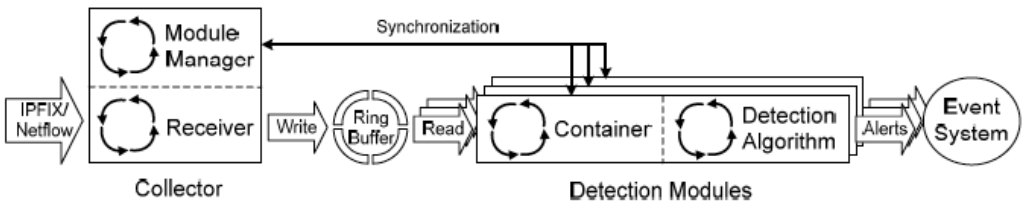
2. Elección de la ruta para detectar la anomalía: El problema de elegir el mínimo número de rutas que deben ser monitorizadas vuelve a ser un problema importante y de nuevo dependiente de la topología concreta de la red.
3. Elección de las rutas para diagnosticar la anomalía: Una vez detectada la existencia de la anomalía, el siguiente paso es diagnosticar la condición de anomalía.

Utilizan el concepto de "tomografía de red" [225] que implica la estimación de parámetros de los enlaces interiores de la red a partir de medidas de extremo a extremo. Las técnicas tomográficas de red persiguen un objetivo "similar" a las técnicas de imaginería médica: inferir situaciones en el "interior" de un sistema cerrado a partir de medidas "externas".

En 2007 Münz y Carle [226] estudian los pasos necesarios para detectar en el tráfico de un enlace la existencia de una anomalía general, un ataque de negación de servicio o la propagación de un gusano. En primer lugar describen un procedimiento genérico para analizar el flujo de datos en tiempo real, tanto para la detección de anomalías genéricas, como de ataques. Los hitos propuestos representan un modelo general válido para cualquier mecanismo de detección:

1. El flujo de datos es recibido de los dispositivos de monitorización y decodificado, para poder ser tratado.
2. La información es preprocesada con objeto de proporcionar una entrada adecuada al algoritmo de detección. Por lo general solo un subconjunto de los datos capturados pueden ser elegibles para su análisis.
3. Finalmente se aplica el algoritmo de detección para descubrir la anomalía o el ataque.

Si el análisis se debe efectuar en tiempo real, las tres tareas enumeradas son críticas en tiempo.



Arquitectura de detección propuesta por Münz y Carle. [226]

Los autores clasifican los algoritmos de detección presentes en la literatura en cuatro categorías:

1. Algoritmos basados en umbrales, basados en el conocimiento previo del tráfico. Los umbrales pueden ser predefinidos o adaptativos.
2. Clasificadores de componentes principales, capaces de detectar anomalías en series temporales multivariantes.
3. Algoritmos de detección de *outliers* que comparan los datos capturados con el comportamiento "normal" de la red.
4. Algoritmos de reglas capaces de "aprender" a clasificar el tráfico a partir de datos de entrenamiento conteniendo tráfico "normal" y "anómalo".

Un aspecto muy importante cuando la pretensión es realizar el análisis en tiempo real es la complejidad de cálculo asociada a los diferentes algoritmos de detección, que varía de decisiones muy simples (como en el caso de algoritmos basados en umbrales o reglas) a cálculos de distancias más complejos (detección de *outliers*) y transformaciones lineales (componentes principales).

En 2007 Ringberg *et al* [227] efectúan un análisis muy interesante sobre la sensibilidad del Análisis de Componentes Principales (ACP) para la detección de anomalías de tráfico. Los autores muestran como el ajuste del ACP para operar en este escenario es difícil y requiere la incorporación de otras formulaciones que le doten de mayor robustez, lo que coincide en gran medida con lo expuesto por Kwitt en 2006 [222]. El estudio identifica y evalúa cuatro retos importantes para poder utilizar ACP en la detección de anomalías:

- La tasa de falsos positivos es altamente sensible a pequeñas diferencias en el número de componentes principales retenidas en el subespacio de comportamiento normal.
- La efectividad del ACP es sensible al nivel de agregación de las medidas de tráfico.
- Una anomalía de duración prolongada puede contaminar el espacio de comportamiento normal.
- La identificación del flujo (*flow*) responsable de generar la alarma es un problema inherente al método.

Los autores concluyen que la capacidad del ACP para la detección de anomalías ha sido sobreestimada y que los métodos para optimizar la aplicación del ACP con este

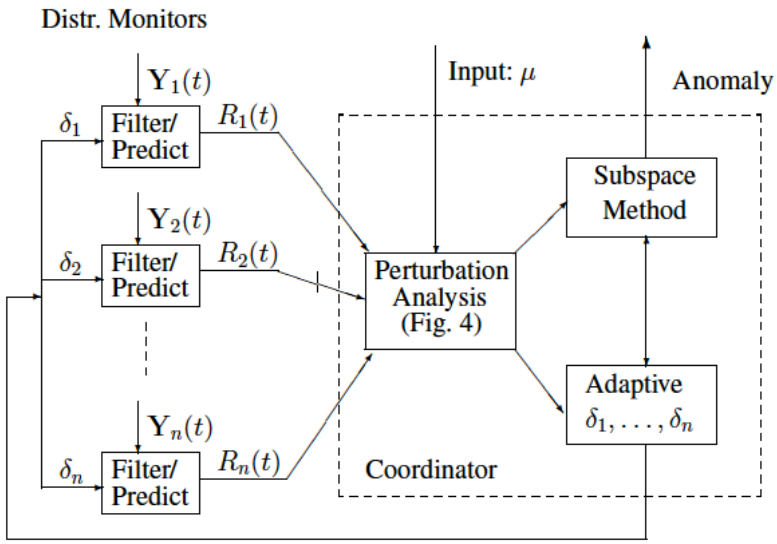
objetivo, son inadecuados. Así pues, la utilización del Análisis de Componentes Principales para la detección de anomalías de tráfico es mucho más difícil de lo aparentaba. El motivo que manifiestan es la dificultad en determinar la dimensión del espacio considerado "normal", y la imposibilidad de identificar qué esta sucediendo realmente para que se afirme que existe una "irregularidad" en el patrón de tráfico observado. De hecho, afirman que "no hay relación directa entre el subespacio de dimensión reducida obtenido por el ACP y la localización espacial original de la anomalía". Antes de que sea posible la utilización del ACP de manera automática y no supervisada se precisan técnicas más efectivas para determinar la dimensionalidad del subespacio de comportamiento normal, previniendo su contaminación e incorporando algún mecanismo para la identificación de flujos (*flows*) responsables de la alarma. En su trabajo en curso se encuentra la investigación de otras técnicas estadísticas que sean capaces de detectar e identificar tráfico anómalo de una manera más robusta que el ACP.

En 2007 Ahmed, Coates y Lakhina presentan una propuesta multivariante de detección de anomalías en línea [228]. Los autores, partiendo de los trabajos de Lakhina [212], [215], [52], [216] que concluían la baja dimensionalidad de los flujos de red, y la elevada covarianza espacial y temporal entre los flujos, proponen un nuevo método que denominan KOAD (*Kernel-based Online Anomaly Detection*). Consideremos un conjunto de medidas de tráfico de red multivariantes obtenidas en situación "normal". En un espacio de características apropiadas, las medidas correspondientes a la situación "normal" deben estar agrupadas en un *cluster*. Dicho de otra manera, dada la inherente baja dimensionalidad del tráfico de red, debe ser posible describir la región ocupada por las características del tráfico utilizando un diccionario relativamente reducido de elementos linealmente independientes. Una aproximación a este diccionario teórico puede ser construido a partir de las medidas capturadas. El hecho es que estas medidas capturadas pueden contener en realidad no solo tráfico "normal", sino también vectores medidos que se correspondan a situaciones anómalas. Por ello se precisa un procedimiento inicial para determinar cuando un vector medido es anómalo y así poder excluirlo del diccionario. Este procedimiento se basa en la hipótesis de que una anomalía debe estar distante del cluster formado por los vectores correspondientes al tráfico "normal". Una vez obtenido el diccionario, determinar las situaciones anómalas se propone que se realice a través del denominado *Kernel Recursive Least Squares* (KRLS). Este método provee, afirman los autores, una eficiente aproximación no paramétrica para efectuar una minera de datos en línea y se basa en un nivel umbral a partir del cálculo de una

distancia específica. La bondad del método se comprueba sobre el conocido juego de datos de Abilene y comparándolo con método el propuesto por Lakhina (ACP). La comparación se realiza mediante las curvas características del receptor (ROC-*Receiver Operating Characteristics*) que no son más que la relación entre la probabilidad de detección y la probabilidad de falsa alarma. El método propuesto presenta prestaciones similares al de Lakhina, aunque con menores cargas computacionales, afirman los autores. Posteriormente, en 2010, los autores realizan una nueva propuesta de utilización del algoritmo KOAD, esta vez para la detección de intrusos en imágenes de vigilancia por circuito cerrado de televisión [229].

En 2007 Huang *et al* [230] proponen un método para hacer escalable la detección de anomalías utilizando el ACP. El método propuesto por Lakhina *et al* [215], recordemos que basado en el ACP, precisa que los monitores de tráfico remitan toda los datos recopilados a un punto centralizado en el que se realiza el análisis de los mismos. Específicamente los monitores deben medir continuamente el volumen de tráfico cursado en cada enlace de la red, y periódicamente remitir todas las medidas al nodo que efectuará el ACP sobre la matriz construida a partir de todas las mediciones recibidas de los monitores. Recordemos que esta técnica permite revelar anomalías de tráfico que no serían detectables a nivel de estudios sobre un único enlace y que la detección es posible en parte debido a la dimensionalidad baja de los datos de tráfico. No obstante de las buenas prestaciones del método propuesto por Lakhina, Huang *et al* exponen dos claras limitaciones al mismo: el primero de ellos es la escala de tiempo. Los trabajo de Lakhina *et al* operan en intervalos de 5 y 10 minutos, sin embargo muchas anomalías suceden a escalas temporales mucho menores. Reducir la escala de tiempo conlleva incrementar el volumen datos a remitir al nodo central y la complejidad del cálculo. La segunda limitación de la propuesta de Lakhina tiene que ver con el efecto de incrementar el número de monitores de tráfico. En la propuesta realizada consistente en que TODOS los monitores remitan al nodo central TODOS los datos recopilados se crean dos problemas: el volumen de información a recibir puede saturar el nodo central y el envío de tal cantidad de información puede sobrecargar la red de comunicación, sobre todo si pensamos en la posibilidad de enlaces de baja velocidad. Así pues, Huang *et al* proponen un sistema de análisis distribuido, en el que los monitores procesan los datos de intensidad de tráfico obtenidos con aplicación de un filtrado local para suprimir el envío de información innecesaria, y un nodo central o coordinador adopta la decisión global que corresponda, enviando a los monitores información de actualización para sus filtros locales basado en las mediciones que observaron. El análisis global se sigue

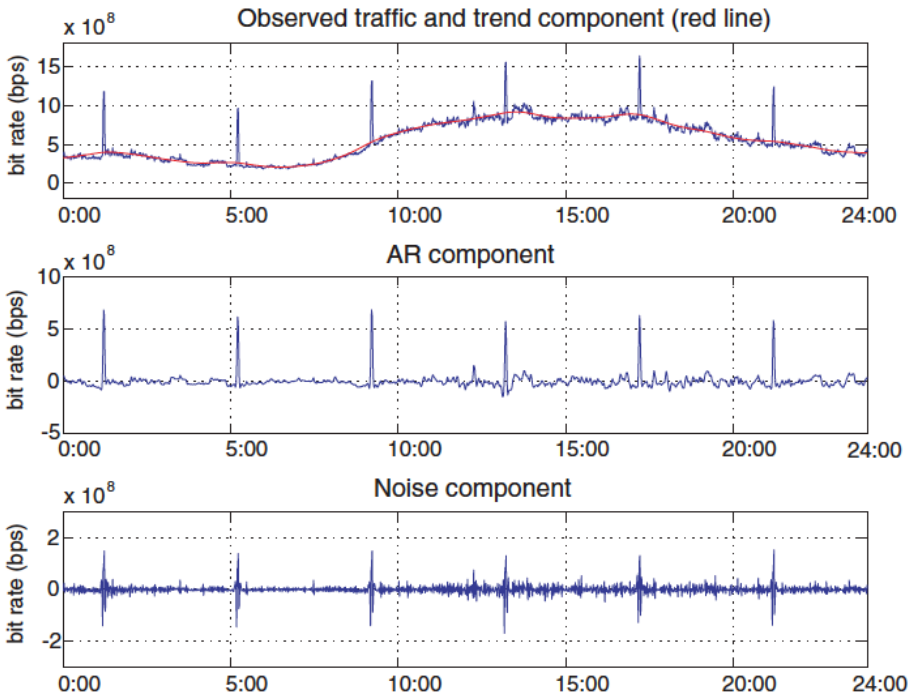
realizando mediante ACP, o método del subespacio, cuyo desempeño se verá afectado por el sesgo introducido por el filtrado en los monitores. Por su lado, los monitores efectúan el filtrado a partir de la predicción que el nodo central les envía para su consideración y la desviación de la medida en curso sobre dicha predicción. Los autores realizan un análisis sobre la perturbación que ocasiona en la decisión final el preanálisis o filtrado efectuado en los nodos de monitorización, concluyendo que el detector de anomalías opera con precisión incluso cuando entre el 80% y el 90% de los datos nunca son enviados al nodo coordinador. Es posible, por lo tanto, reducir drásticamente la necesidad de envío de información, a un coste relativamente bajo en cuanto a prestaciones del detector de anomalías de volumen.



Sistema de detección distribuida propuesto por Huang *et al*. [230]

Du *et al* en 2008 [231] proponen un método para detectar y trazar anomalías de volumen en un enlace troncal de una red concreta (SINET3). Basan su propuesta en la descomposición del volumen de tráfico cursado en tres componentes:

- Componente de tendencia, que captura los cambios graduales de la serie temporal del volumen de tráfico cursado.
- Componente autoregresiva (AR) que consiste en fluctuaciones del tráfico estocásticas y anomalías predecibles.
- Componente de ruido, que se asume blanco con media nula.

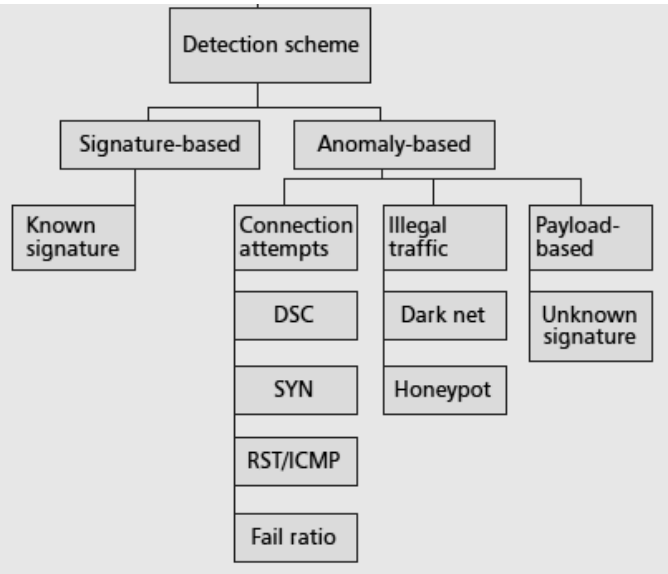


Ejemplo de descomposición del tráfico en un enlace. [231]

La detección de una anomalía tiene lugar cuando la componente autoregresiva toma valores "imprevistos".

El principio en el que basan su propuesta es comparable a la realizada por Lakhina *et al* [215] , en cuanto a la separación del volumen de tráfico en diferentes "subespacios", si bien, la propuesta de Lakhina, como vimos, contempla la aplicación del Análisis de Componentes Principales (ACP) y requiere una carga computacional elevada debido al análisis matricial implicado en el mismo. La propuesta de los autores efectúa la descomposición requiriendo menores potencias de cálculo utilizando un principio expuesto en el artículo de Maxion [41].

Li *et al* publican en 2008 [232] un estudio sobre las diferentes técnicas de detección de gusanos, cuya propagación no deja de ser un tipo específico de anomalía más en el tráfico de las redes. En el aspecto que más nos interesa en este punto (algoritmos de detección de anomalías) efectúan un repaso a los diferentes métodos existentes en la literatura. Si bien algunos de ellos son muy específicos del problema concreto, otros son más genéricos y podrían utilizarse para detectar otro tipo de anomalías, y no tan solo la propagación de gusanos



Clasificación procedimientos de detección de gusanos. [232]

Entre las técnicas genéricas destacaríamos:

- Detectores basados en firmas, son tradicionales en Sistemas de Detección de Intrusos (IDS) y se emplean normalmente para la detección de ataques conocidos. No precisan la determinación de modelos de tráfico “normales”.
- Detectores de anomalías, tienen la ventaja de no precisar un conocimiento previo sobre el ataque, pero precisan definir el comportamiento “normal” de la red.

Ambos métodos presentan ventajas e inconvenientes y, ciertamente, en muchos casos suelen combinarse detectores de ambos tipos para mejorar sus prestaciones. Los autores enumeran una serie de métricas a monitorizar para la detección de anomalías provocadas por gusanos.

En 2008 Naidu *et al* [233] proponen un conjunto de técnicas de monitorización de red basadas en medidas activas de extremo a extremo. Se trata de nuevo en un planteamiento de tomografía de red que, como ya se expuso, implica la estimación de parámetros a nivel de enlace a partir de medidas de extremo a extremo. El trabajo se basa específicamente en la detección de violaciones de la Calidad de Servicio (QoS), lo que se aleja notablemente de nuestro objetivo. No obstante se cita el artículo ya que provee otra interesante aplicación a la detección de anomalías en un escenario diferentes, ya que hasta ahora se circunscribía a anomalías de volumen, prácticamente. La detección de violaciones de QoS tiene especial interés en redes que cursen tráficos sensibles a estas violaciones, como por ejemplo las de telefonía sobre IP.

Este mismo 2008 de nuevo Kim y Reddy proponen [234] un detector de anomalías en tres fases consecutivas:

- La primera fase está constituida por un analizador de tráfico, en el que se genera una señal de correlación a partir de las cabeceras de los paquetes. Se puede utilizar información cualquier información contenida en la cabecera, como por ejemplo la dirección de destino y el puerto.
- La segunda fase implica la transformación de los datos para el análisis estadístico. La propuesta utiliza aquí *wavelets* para estudiar la correlación de las direcciones y puertos a lo largo del tiempo, estrategia ya planteada por Barford en 2002 [207].
- La etapa final es la detección, en esta fase se utilizan umbrales para la declaración de anomalías o ataques.

La correlación de las direcciones utilizada en el primer paso se formula del siguiente modo:

$$\rho(n) = \frac{\sum_m \left(p_{mn-1} - \overline{p}_{n-1}\right)\left(p_{mn} - \overline{p}_n\right)}{\sqrt{\sum_m \left(p_{mn-1} - \overline{p}_{n-1}\right)^2}\sqrt{\sum_m \left(p_{mn} - \overline{p}_n\right)^2}}$$

donde: p_{mn} es el número de paquetes de la dirección a_m enviados en el instante s_n

Los autores verifican la fuerte correlación positiva entre muestras adyacentes en un amplio intervalo de tiempo. Plantean igualmente una simplificación de la formula anterior, para mejorar la eficiencia computacional:

$$C(n) = \frac{\sum_m p_{mn-1}p_{mn}}{\sum_m p_{mn}}$$

Que verifican que presenta un comportamiento similar a la anterior propuesta.

Los autores concluyen que el método constituye un mecanismo efectivo para la detección de anomalías en un campus o en un borde de la red, a través de la correlación de las dirección IP de destino y los puertos en el tráfico de salida en el router.

También en 2008 Won *et al* publican un artículo [40] sobre la detección y diagnosis de fallos en sistemas industriales de misión crítica. Especialmente interesantes para nuestro trabajo es la enumeración de las métricas más comunes en redes Ethernet (y también TCP/IP) que identifican en su estudio (ver apartados previos dedicados a Ethernet e Internet para detalles). El método de detección de anomalías que se emplea es un simple nivel umbral obtenido a partir de la distribución biparamétrica de Weibull. Los autores no justifican en su artículo el motivo de porqué utilizan la mencionada distribución estadística en su estudio.

Samaan y Karmouch en 2008 [235] investigan la eficiencia en el diagnóstico de anomalías utilizando conceptos de análisis estadístico, en concreto modelos autoregresivos, y razonamiento basado en la evidencia, aplicando la teoría de Dempster-Shafer. En el primer paso se analizan los incrementos en las medidas obtenidas en cada una de las diferentes variables monitorizadas en una doble aplicación de un modelado autoregresivo. Así se obtiene una medida normalizada para detectar posibles cambios abruptos en un instante concreto. Las medidas de estas desviaciones se combinan en el motor de decisión para obtener una conclusión sobre la causa de fallo. Es este motor de decisión el que aplica la teoría de la evidencia de Dempster-Shafer como nueva aproximación para la clasificación de anomalías. Los resultados que obtienen los autores confirman la efectividad del esquema de clasificación. El objetivo del estudio es la detección de anomalías en el reenvío de paquetes IP, si bien planean investigar en el futuro la aplicación del esquema propuesto en el diagnóstico de otros tipos de anomalías tales como ataques de negación de servicio.

También en 2008 Chhabra *et al* [236] proponen un método para la detección de anomalías de manera distribuida. Con similares motivaciones que Huang *et al* [230] evalúan igualmente un método en dos fases. En la primer fase cada router identifica una serie de “anomalías candidatas” comparando las medidas de tráfico con sólo las mediciones locales que él ha podido realizar en los enlaces a los que está conectado. De hecho, cada router identifica *outliers* en el espacio que tiene visible 2N dimensional explotando las correlaciones en el tráfico local. En la segunda fase, se establece un conjunto de “anomalías consenso”, a partir de la comunicación de los routers a sus vecinos de sus “anomalías candidatas”, que pueden ser declaradas “anomalías consenso” simplemente si dos routers declaran un *outlier* en el mismo intervalo de tiempo. La principal desventaja del método que citan los propios autores es que no se trata de un método “global”, aunque sí que se trata de un procedimiento que considera

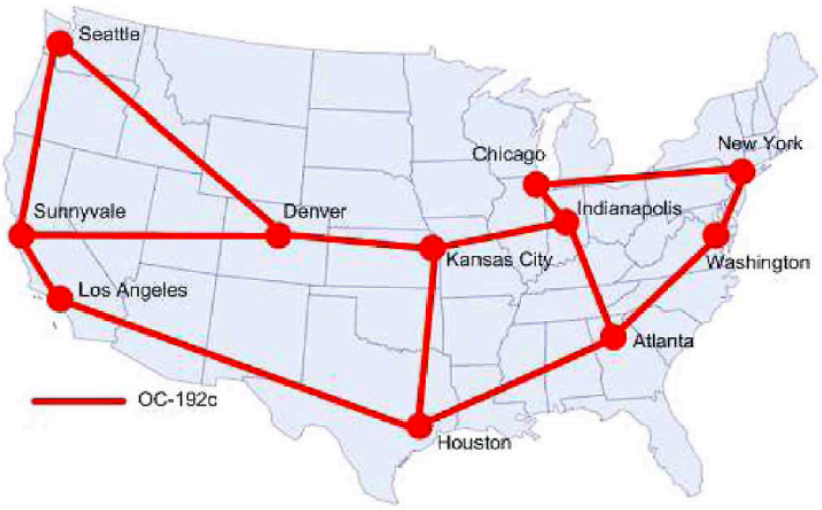
el efecto “espacial” más que el temporal. Por eso los autores lo denominan detección espacial de anomalías de volumen. El procedimiento que se propone para la detección de anomalías en la primera fase es el GQS (*Generalized Quantile Sets*). Formalmente el GQS encapsula regiones del espacio donde la masa de una función de densidad de probabilidad P está más concentrada, estableciendo un conjunto de niveles β definidos como el mínimo volumen euclídeo con probabilidad al menos β**.** El método se basa en la adopción de un marco probabilístico en el que las medidas de tráfico locales típicas (no anómalas) se consideran realizaciones independientes (ignorando aquí las dependencias temporales) de una distribución de probabilidad. La hipótesis principal es que las anomalías son *outliers* respecto a esta distribución y clasificamos los puntos según cuánto extremos sean respecto a la distribución de probabilidad nominal. Dado que el conjunto de datos es multidimensional, el criterio de “extremo” será un “volumen” de probabilidad. Los autores aplican su método sobre el conocido juego de datos de la red Abilene, del 7-13 de abril de 2003 y del 8 al 28 de diciembre también de 2003. Los resultados se comparan con la aplicación del ACP, concluyendo que la nueva estrategia se comporta de una manera similar al ACP centralizado, quedando aún margen, afirman, para mejoras en el método propuesto que sean más robustas, tolerantes a fallos y resistentes a ataques al método de detección.

Para terminar con la revisión del año 2008 Sun *et al* [237] proponen un método de defensa ante ataques de negación de servicio distribuidos (DDoS) basado en el ACP. El ACP extrae las características nominales del tráfico analizando las dependencias intrínsecas entre los valores de los atributos de cada paquete y diferenciando el tráfico sospechoso a través de la verificación de múltiples atributos. El procedimiento propuesto captura el tráfico “normal” y extrae sus características nominales, esto es, mientras no existe un ataque en curso. Cuando se detecta una congestión en el enlace de la víctima y los paquetes entrantes comienzan a ser descartados por ese efecto, el método basado en ACP se activa para descartar solo los paquetes declarados sospechosos de acuerdo a la probabilidad del paquete de ser legítimo: solo los paquetes con mayores probabilidad de ser legítimos obtendrán acceso a la red protegida. Aquí el ACP se aplica directamente sobre múltiples variables del tráfico: direcciones IP, valor de TTL (*Time-to-live*), puerto del servidor, tipo de protocolo, tamaño del paquete, longitudes de las cabeceras, combinaciones de etiquetas TCP (SYN, SIN, RST, ACK). La matriz formada a partir de las mediciones de todas estas variables en diferentes intervalos consecutivos es sobre la que se aplica el ACP. Es interesante que este artículo va un paso más allá de los vistos hasta ahora, al no plantear exclusivamente la “detección” de la anomalía (en este caso un ataque), sino

también la “mitigación” de la misma mediante una reacción de defensa automatizada. Los autores concluyen que el método propuesto diferencia de manera efectiva paquetes integrantes del ataque de aquellos que constituyen tráfico legítimo, para diversos tipos de ataques analizados.

En 2009 Paschalidis y Smaragdakis [238] presentan dos diferentes métodos para caracterizar el tráfico. El primero utiliza una aproximación libre basada en el método de tipos y el teorema de Sanov. El segundo se trata de una propuesta paramétrica utilizando procesos modulados de Markov (MMP). Con estas dos caracterizaciones del tráfico monitorizan el tráfico cursado y comparan el tráfico medido en realidad con la referencia obtenida del modelo. Si bien inicialmente la propuesta solo contempla información temporal, los autores proponen la incorporación de información espacial considerando diferentes vectores de actividad de tráfico en diferentes localizaciones de interés en la red bajo estudio. Consideran que las series $\mathbf{X}_1,\ldots,\mathbf{X}_n$ (donde $\mathbf{X}_i \in \mathbf{R}^d$) representan las “velocidades” de la características de interés de la red bajo estudio (por ejemplo, el número de bytes/paquetes/flujos durante el intervalo temporal i-ésimo) en todas las d localizaciones que deseamos monitorizar. El autor aplica en primer lugar su propuesta al archiconocido juego de datos de la red Abilene recopilado entre el 7 al 13 de abril de 2003, en el que constan tres diferentes tipos de anomalías:

- 133 ataques de negación de servicio distribuido contra una víctima única.
- 81 rastreos de puertos para localizar sistemas vulnerables
- 32 tasas inusualmente elevadas de transferencia de datos

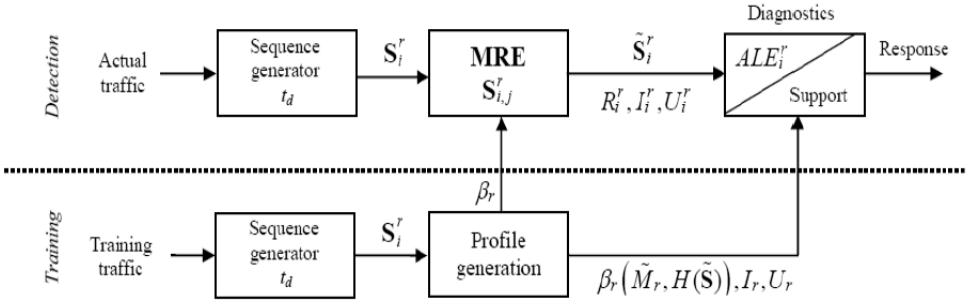


Red troncal Abilene y puntos de presencia. [238]

En total 270 anomalías diferentes. En segundo lugar utilizan el juego de datos de MIT Lincoln Laboratory de 1999 utilizado por la Agencia de Proyectos Avanzados (DARPA) para la evaluación de detectores de intrusos. Este juego de datos contiene múltiples y diversos tipos de ataques, en concreto 56 tipos de ataques distribuidos. Los resultados experimentales muestran que el método libre funciona "algo" mejor que el modelo paramétrico y postulan seguir investigando la robustez del método ante diferentes parámetros del modelo. En el apéndice los autores realizan un interesante análisis de la correlación en los datos de tráfico.

En 2009 Velarde *et al* [239] proponen el denominado Método de los Elementos Restantes para la detección anomalías basado en la caracterización del tráfico a través de la medida proporcional de la incertidumbre. Esta relación de incertidumbre les permite a los autores determinar los valores restantes de una secuencia de variables características. Afirman que esta relación de incertidumbre provee mejor sensibilidad para definir el corte entre elementos remanentes y significativos que la incertidumbre relativa. Los autores indican que la utilización de la entropía de Shannon para la detección de anomalías [216], [240] presenta inconvenientes, ya que los ataques distribuidos o a corto plazo no son claramente detectados porque la incertidumbre es despreciable o está distribuida. Para corregir el problema los autores proponen una modificación de la entropía. Posteriormente utilizan el cociente entre esa propuesta de nueva entropía y el conocido máximo de la entropía de Shannon, ese cociente lo denominan incertidumbre proporcional. El método de los elementos remanentes finalmente se basa detectar la ocurrencia de un ataque a partir de los elementos restantes que quedan por aparecer en la secuencia posible de elementos controlados, esto es, cuando el número de elementos observados dentro de las posibles ocurrencias de direcciones IP o puertos es muy elevado, la elevada diversidad en la actividad observada nos indicaría que estamos en presencia de un ataque. El método se apoya en un parámetro seleccionable β_τ denominado umbral de exposición, que controla la sensibilidad a la hora de calcular los elementos restantes en un conjunto de datos. Los autores verifican el funcionamiento del método propuesto en varios escenarios reales, dejando para futuros estudios el análisis de la influencia del umbral de exposición en la aparición de falsos positivos. En 2010 publican un nuevo artículo [241] en el que proponen una mejora de su método, ya que del análisis en profundidad del primer método concluyen que diversos patrones de ataque no pueden ser detectados. Es el caso de ataques generados con elevados volúmenes de paquetes que no presenten suficiente diversidad en sus características. La propuesta se apoya en un nuevo parámetro denominado ALE (*Anomaly Level Exposure*), resume

el comportamiento de las diferentes secuencias en términos de la exposición a anomalías, y presenta una mejor sensibilidad ante intrusiones que la alternativa previa.

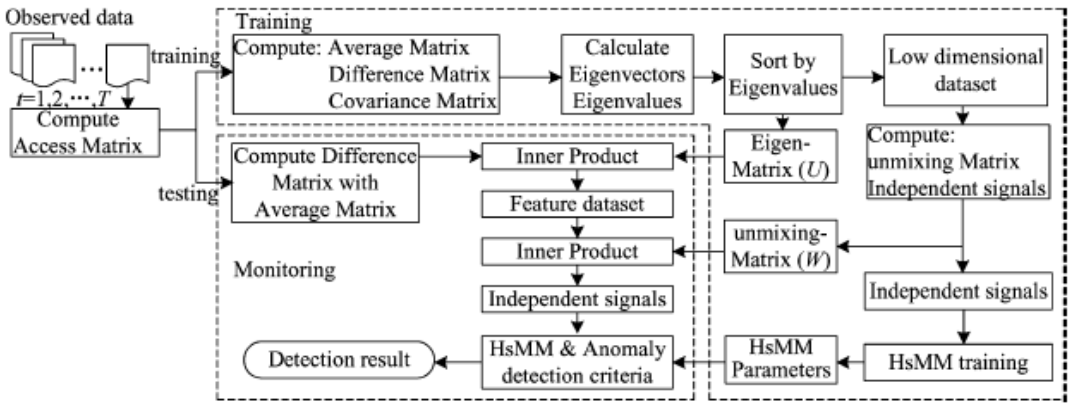


Arquitectura para detector de intrusos basado en detección de anomalías y MRE. [241]

En 2009 Xie y Yu publican una propuesta [242] para la monitorización de ataques DDoS en servidores web. En este caso la detección del ataque de negación de servicio se realiza sobre el tráfico de la aplicación, y no directamente sobre el tráfico de la red. En su artículo los autores definen una Matriz de Acceso para capturar el patrón espacio-temporal del ataque y monitorizarlo durante su ocurrencia. Sobre esta matriz aplicación modelos de Markov para lograr una detección automática y utilizan ACP y ACI (Análisis de Componentes Independientes [243]) sobre la información multidimensional del modelo de Markov.

Especialmente interesante es la definición de la matriz de acceso, en la que los elementos de la matriz son relaciones entre valores medios de solicitudes de páginas:

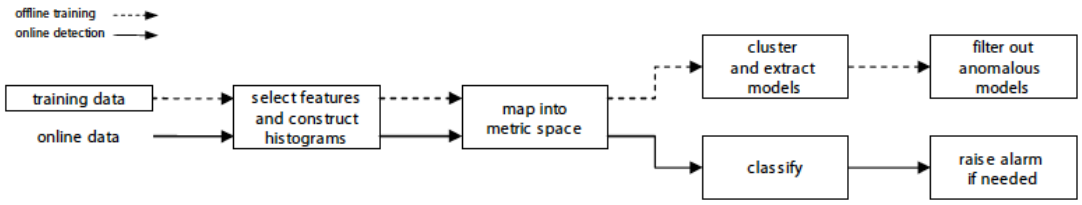
> r_{it} se define como la relación entre el número de medio de solicitudes por usuario del documento i-ésimo en el intervalo t-ésimo y el número de peticiones medias por usuario en el mismo intervalo t-ésimo.



Arquitectura de monitorización. [242]

En la fase final de la detección los autores proponen la utilización de la entropía y un nivel umbral, obtenido durante una fase de entrenamiento, para la declaración final. Los autores concluyen que su propuesta es capaz de capturar el desplazamiento del tráfico del servidor web y que la entropía puede ser utilizada para la objetivación de la anormalidad.

Ese mismo año 2009 Kind *et al* [244] presentan una propuesta para la detección de anomalías basada en histogramas. Los autores proponen la construcción de histogramas para describir los patrones de diferentes características del tráfico. Por ejemplo, un histograma puede reflejar el número de flujos asociados con diferentes puertos de destino a lo largo de un periodo de tiempo. Cada histograma es posicionado en un espacio métrico de manera que los histogramas se agrupen por su parecido. Se aplican técnicas de minería de datos para identificar los patrones "normales" que son finalmente utilizados para detectar desviaciones.



Itinerario de la detección de anomalías. La línea punteada representa la fase de entrenamiento. La continua el proceso de detección en la fase de explotación. [244]

La propuesta se descompone en cuatro pasos:

1. Seleccionar las características a analizar y construir los histogramas. Proponen 8 características que posibilitan la detección de varios tipos de anomalías: direcciones IP de origen/destino, puertos origen/destino, etiquetas TCP, número de protocolo, tamaño del paquete y duración de los flujos.
2. Mapear los histogramas en un espacio métrico en función de su similaridad. Los autores repasan diferentes métricas: Manhattan, Mahalanobis, Euclídea, Hamming. Además proponen el ACP para reducir la dimensión del modelo. Los autores especifican que esta utilización del ACP es distinta de la presentada por Lakhina [216], ya que aquí solo persiguen la reducción de la dimensión y no la clasificación del tráfico explícitamente.
3. Construir los *clusters* y extraer los modelos, para el primer paso estudian el clustering jerárquico y el k-medias. Este método ya ha sido propuesto anteriormente en 2006 por Kim y Kim [223].
4. Clasificar las anomalías, a partir de la comparación de los patrones extraídos con los modelos obtenidos.

Los autores concluyen que la dimensión de los histogramas puede ser reducida para obtener un modelo más manejable, los clusters obtenidos para el tráfico de entrenamiento ("normal") no presentan diferencias significativas con el incremento de la cantidad de tráfico utilizado para entrenamiento y por último, y muy interesante, las diferentes métricas dan como resultado clusters similares.

En la Conferencia Internacional de Comunicaciones celebrada en 2009 (ICC-2009) se presentaron dos artículos interesantes sobre detección de anomalías. En primer lugar Qin *et al* [245] propusieron la monitorización de trafico utilizando Análisis de Componentes Independientes (ICA), método que ya hemos visto aplicar antes en esta misma anualidad a Xie y Yu [242]. Los autores proponen un procedimiento para monitorizar el tráfico que no requiere fase de aprendizaje, lo que es importante, y que utiliza cuatro parámetros del tráfico: número de paquetes, tamaño del paquete, número de flujos y grado de conexión. El patrón de tráfico contiene dos tipologías: el tráfico normal y el tráfico anómalo, que son generalmente independientes, debido a la naturaleza intrínseca de cada tipo. Cada una de las series temporales correspondientes a los parámetros de tráfico se descompone en una series diferente para a cada tipo de tráfico considerado, normal y anómalo:

$$x_i(t)= a_{i1}s_1(t) + a_{i2}s_2(t) \qquad i=1,..,4$$

Los autores proponen la utilización del Análisis de Componentes Independientes (ICA) para la obtención de los diferentes coeficientes a_{ij} que posibilitan la descomposición anterior. Posteriormente a la descomposición se propone la aplicación de filtrado para suavizar las series obtenidas. Por último la detección de la anomalía propiamente dicha se realiza analizando las características de la serie anómala: su amplitud, la duración y el área. Los autores comparan su método con el ACP, probando que la nueva alternativa presenta mejores prestaciones, ya que la variabilidad intrínseca del tráfico no afecta los resultados de la detección y la tasa de falsa alarma es menor, todo ello con la ventaja de no requerir fase de entrenamiento previa.

El segundo trabajo presentado en ICC-2009 es el de Himura *et al* [246]. Al igual que el trabajo de Kind *et al* [244], proponen utilizar histogramas para la detección de anomalías. Los autores definen una anomalía en la red como una desviación de un comportamiento estadístico de referencia. Como ya hemos visto anteriormente el ajuste de los parámetros del método suele afectar a las prestaciones del detector. El artículo estudia cómo establecer los parámetros de un método concreto, y que pueden

ser extrapolable a otros métodos. El algoritmo de identificación se proponen se divide en tres pasos:

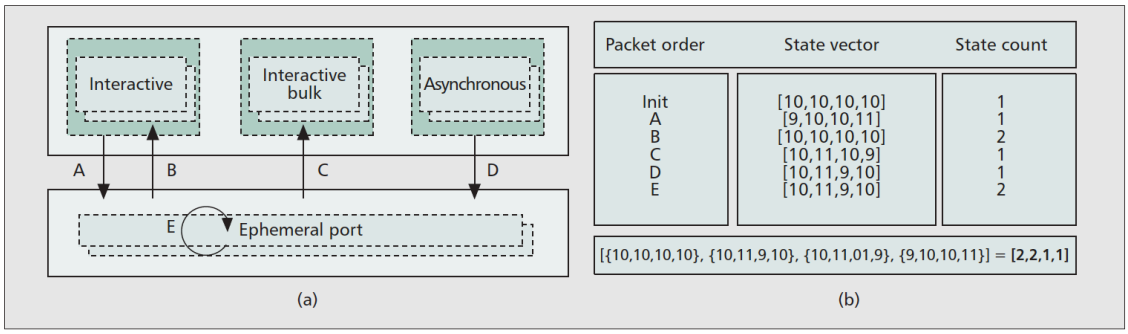
1. Bosquejo (*Sketch*): el tráfico se divide en conjuntos de paquetes con la misma dirección IP de origen.
2. Modelo de función gamma: cada conjunto de paquetes con la misma dirección IP de origen recibidos en un cierto intervalo de tiempo se aproxima como una distribución gamma, con dos parámetros, α que determina la forma del histograma y β la escala.
3. Identificación de anomalías: Los parámetros α de cada evento se comparan entre sí, y los eventos cuyos parámetros son *outliers* se declaran anomalías. Se procede de igual modo con los parámetros β.

Los autores concluyen que hay una elevada variabilidad del parámetro que se considera óptimo, y que se puede requerir un ajuste diario para mantener las prestaciones del método, situación esta que no es específica del método de detección utilizado por los autores en su estudio.

También en 2009 Han y Choi publican un artículo [247] en el que proponen la detección de ataques (anomalías) utilizando, una vez más, la entropía de los paquetes. Los autores comparan cuatro métodos diferentes:

1. EWMA (*Exponential Weighted Moving Average)*
2. Predicción Holt-Winters
3. ACP
4. Entropía

Los autores comienzan su trabajo con la observación de que la entropía varía abruptamente cuando una anomalía perturba el sistema, como ya indicaba Lakhina en su artículo [216]. Por ejemplo, como resultado de un rastreo de puertos (*port scanning*) se incrementa la entropía del puerto de destino, y el host infectado observará un decremento de la entropía de la dirección IP de origen. La entropía es una métrica particularmente efectiva para determinar el comportamiento normal o anormal de un sistema, la cuestión central radica en cómo medir de manera efectiva la entropía observando el intercambio de paquetes en una red. Veamos como proponen calcularlo:



Representación para el cálculo de la entropía propuesta por Han y Choi. [247]

Supongamos 4 sistemas, que denominaremos “interactivo” (*interactive*), “interactivo masivo” (*interactive bulk*), “asíncrono” (*asynchronous*), “alrededores” (surroundings, *ephemeral port*). Asignamos a cada sistema una puntuación inicial, por ejemplo 10, lo que nos permite crear un vector de estado con 4 componentes (10,10,10,10) y un contador de las observaciones de los diferentes vectores de estado, 1. Analicemos el movimiento de 5 paquetes diferentes:

- Paquete A: se mueve del sistema “interactivo” a los “alrededores”, la primera componente del vector disminuye una unidad, mientras que la última, se incrementa en 1: (9,10,10,11) el contador es 1 ya que es la primera vez que vemos este vector de estado.

- Paquete B: en este caso el paquete se mueve de los “alrededores” al grupo “interactivo”, el vector de estado vuelve a (10,10,10,10,10) y su contador pasa a 2.

Y así sucesivamente para los restantes paquetes. Ahora sólo resta calcular la entropía:

$$e_t = -\sum_{i=1}^{m} p_i \log p_i \qquad p_i = d_i \Big/ \sum_{i=1}^{m_t} d_i$$

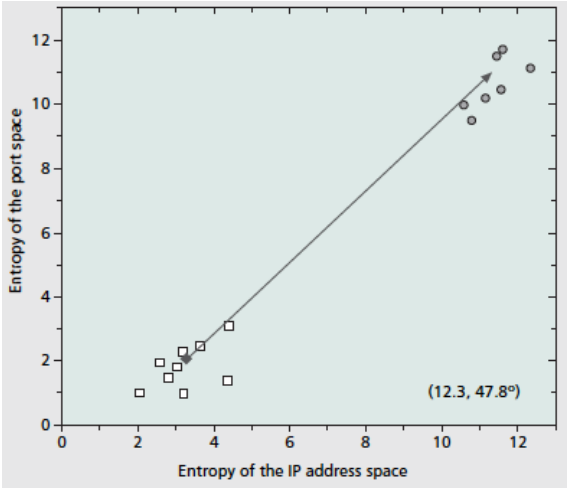
en donde e_t es la entropía en un tiempo dado t

d_i es el contador de estado para el vector de estado i-ésimo

m_t es el número de vectores de estado distintos para un tiempo t

Según esto, la entropía se incrementa bien cuando el número de vectores de estado distintos m_t aumenta o bien la varianza del contador de estado $Var(d_i)$ disminuye.

Especialmente interesante es la representación gráfica de la entropía. Uno de los ejemplos que presentan los autores se trata de la representación gráfica de la variación de la entropía para un ataque de negación de servicio DDoS.



Representación gráfica de la entropía vs. espacio de direcciones. [247]

El gráfico muestra 40 puntos, 20 corresponden con la situación “normal”, y otros 20 con la del “ataque”. Se muestra como la entropía se incrementa conforme se introduce el tráfico anómalo, debido a la aparición de un elevado número de paquetes provenientes de múltiples fuentes y con destino los puertos vulnerables, incrementando la entropía en ambos espacios.

La declaración de anomalía se produce al sobrepasar la variación un determinado nivel umbral. Los principales parámetros del método serán el número de valores de entropía considerados y el umbral de detección.

Los autores proponen la curva de detección para comparar la efectividad de los diferentes algoritmos de detección y concretamente el área bajo la curva: cuanto más grande sea el área bajo la curva de características del receptor (ROC), mejor será el algoritmo de detección. La propuesta que realizan los autores mejora las prestaciones de detección de los tres métodos con los que se compara. No obstante es preciso realizar una puntualización muy importante: si bien citan el artículo de Lakhina [216] para presentar la entropía, para la comparación del método de ACP utilizan otro artículo de Lakhina previo [215], ambos artículos utilizan ACP, pero el primero utiliza la combinación de PCA y entropía, planteamiento este último que sorprendentemente Han y Choi no evalúan.

También en 2009 Androulidakis *et al* publican un artículo [248] con énfasis en la técnica de muestreo utilizada para la obtención de la serie temporal de datos. La

técnica de detección de anomalías utiliza se basa en la entropía, cuya metodología de aplicación exponen con detalle, así como una enumeración de las diferentes características sobre la que se ha propuesto aplicarla en la literatura: direcciones IP origen/destino, puerto origen/destino y tamaño del flujo. Especialmente interesante es la clasificación que proponen de anomalías basada en los cambios de la entropía. Es similar a la propuesta por Lakhina en 2004 [52] aunque aquella contempla algún caso más y ejemplos. Androlulidakis *et al* concluyen el método de muestreo utilizado puede ser incluso adaptado al tipo de anomalía que deseemos detectar, demostrando que incluso con pequeñas tasas de tráfico anómalo, la utilización de un muestreo "inteligente" puede mejorar significativamente la efectividad de la detección de anomalías y en algunos casos revelar anomalías que de otro modo permanecían ocultas.

Ese mismo año 2009 Duffield *et al* [249] proponen un detector de anomalías basado en reglas aplicado a patrones de flujos. Los patrones iniciales son aprendidos a partir de un conjunto de patrones de referencia de paquetes y los datos conjuntos de paquetes/datos. Como criterio de prestaciones de la propuesta se vuelve a utilizar la curva de características operativas del receptor (ROC) que presenta la relación entre verdaderos positivos y falsos positivos. Como curva, ROC ofrece una información completa, pero para efectuar comparaciones es compleja de aplicar y se utiliza el "área bajo la curva ROC" (AUC) aunque si el método es muy bueno, incluso los valores que ofrece el AUC suelen ser problemáticos para interpretar/comparar ya que pueden ser superiores a 0,9999. En estos casos se propone utilizar la Precisión Media que provee la parte pesimista la optimista AUC. El método lo prueban sobre SNORT con trazas de cabeceras de paquetes de 4 semanas de duración, concluyendo:

- La clasificación con reglas de nivel de flujo de acuerdo a si actúan sobre la cabecera de los paquetes, información transportada o meta-información es un buen predictor cualitativo de la Precisión Media.

- La propuesta realizada es efectiva para descubrir asociaciones entre características de anomalías a nivel de paquetes y flujos y aprovechas estas asociaciones para disparar alarmas a nivel de flujo.

- A lo largo de dos semanas de pruebas no aparece fenómeno de deriva, entendido como el efecto sobre las prestaciones de la distancia temporal entre el juego datos de entrenamiento y los reales.

Pero a nuestro modo de ver, el artículo más interesante para nuestra investigación de entre los publicados ese año 2009 es el que publica Brauckhoff [250]. En él, y a partir de las propuestas de Ringberg *et al* [227], analizan el ACP para la detección de anomalías de volumen, concluyendo que el principal problema del ACP para ese objetivo reside en que NO considera la correlación temporal existente en los datos. En el desarrollo de su artículo Brauckhoff expone una extensión del ACP a procesos estocásticos (aplicando la expansión de Karhunen-Loève y el método de Galerkin, pero diferente de la aproximación "clásica" de Bouhaddou [214]) que, aún sin indicarlo de manera expresa, es aparentemente una aplicación del algoritmo MSSA (Multivariate Singular Spectrum Analysis) [243], [210].

Brauckhoff parte de un vector de dimensión K de procesos estocásticos discretos de media nula:

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_K)^T$$

Desde ahí construye una nueva matriz de observaciones de dimensión KN x (n-K), siendo N el número a partir del cual los valores de las covarianzas de los procesos son despreciables. La matriz que obtiene es la siguiente:

$$\mathbf{X} = \begin{pmatrix} x_1(1) & \cdots & x_1(n-N) \\ x_1(2) & \cdots & x_1(n-N+1) \\ \vdots & \ddots & \vdots \\ x_1(N) & \cdots & x_1(n) \\ x_2(1) & \cdots & x_2(n-N) \\ \vdots & \ddots & \vdots \\ x_2(N) & \cdots & x_2(n) \\ \vdots & \ddots & \vdots \\ x_K(1) & \cdots & x_K(n-N) \\ \vdots & \ddots & \vdots \\ x_K(N) & \cdots & x_k(n) \end{pmatrix}$$

De dicha matriz obtiene una nueva matriz de covarianzas espacio-temporales:

$$\Sigma = \frac{1}{n-N-1}\mathbf{X}\mathbf{X}^T$$

a la que la autora aplica el ACP, obteniendo los correspondientes autovectores y autovalores. Los autores exponen, en repetidas ocasiones, que la utilización de esta nueva matriz de datos, más compleja que la habitual, es inevitable ya que hay que tratar con observaciones que están correladas temporalmente. Es muy interesante también la aproximación que realizan del modelo obtenido con un banco de filtros de Respuesta Finita al Impulso (FIR-*Finite Impulse Response*). De hecho se utiliza esta última aproximación para obtener un filtro predictivo con el que comparar los valores reales de tráfico para declarar la situación de anormalidad.

Ya en 2010 Abdelkefi *et al* [251] incidiendo de nuevo en la sensibilidad del ACP para la detección de anomalías manifestada por Ringberg y Brauckhoff [227], [250] realizan una nueva propuesta basada en una variación "robusta" del ACP denominada PCP (*Principal Component Pursuit*), propuesta por Candes inicialmente en 2009 y revisada en 2011 [252], [253]. La elevada sensibilidad del ACP a la presencia de *outliers*, conocida como "envenenamiento del subespacio de bajo rango" o "envenenamiento del ACP" (*PCA-poisoning*), está fundamentada en la utilización de la norma L_2 (euclídea) en el ajuste, como indican Ke y Kanade en 2005 [254]. Este fenómeno representa la desviación de las componentes principales de la verdadera distribución de los datos, que sucede incluso en presencia de solo una pequeña fracción de *outliers*. Como consecuencia, el desplazamiento de las componentes principales conducen a una perturbación del espacio de bajo rango, que a su vez ocasiona detecciones inexactas y altas tasas de falsos positivos. Así los autores proponen la utilización de la norma L_1 (Manhattan) en lugar de la L_2 (euclídea), ya que concluyen que la primera, L_1, resulta más robusta frente al envenenamiento por *outliers*, según las curvas ROC obtenidas, en comparación con el método de PCA+KLE propuesto por Brauckhoff [250].

También en 2010 Barbosa y Zambenedetti efectúan una propuesta [137] que incide de nuevo [9]–[12] en la visualización de la información como método para el análisis de los grandes conjuntos de datos provenientes del tráfico en las redes. Los autores proponen tres prototipos de visualización de la información y postulan cuatro preguntas a realizarse ante propuestas de esta tipología:

- ¿posibilita la herramienta de visualización una precisa imagen de los conjuntos de datos de gestión de la red?

- ¿puede el administrador de la red ajustar el rango de la visualización a sus necesidades?

- ¿En qué medida la técnica de visualización facilita el proceso de detección de patrones en el tráfico de gestión de la red?

- ¿Estamos usando representaciones abstractas eficaces para mostrar los datos?

Ciertamente el análisis que los autores proponen se efectúa sobre un tipo de tráfico muy concreto (el de gestión de la red) pero las cuestiones que plantean para evaluar el desempeño de una propuesta de visualización de tráfico son muy relevantes, en nuestra opinión.

En 2010 Holanda y Bessa proponen un método para la predicción del tráfico de redes utilizando ACP y análisis de cluster k-medias [255]. Los autores inciden en las propuestas ya realizadas en 2006 por Kim y Kim [223] y en 2009 por Kind *et al* [244]. La metodología propuesta se basa en la predicción del tráfico origen-destino. El uso del ACP permite aplicar la predicción a un pequeño número de componentes principales, las cuales retienen una parte significativa de las propiedades originales de los datos, posibilitando la identificación del perfil de comportamiento común a varios flujos. Las tendencias de comportamiento y la predicción se aplican a la serie temporal de los factores calculados para cada componente principal utilizando k-medias y un modelo de regresión lineal aplicado localmente. Una de las hipótesis asumida para la aplicación de este modelo es la homocedasticidad de la serie temporal del tráfico, también conocida como varianza constante finita. Los autores estiman la varianza del tráfico, los valores medios (como tendencia) y el parámetro de Hurst. Este último mide el nivel de autosimilaridad presente en la serie y se ha demostrado por Leland en 1994 [256] y Paxson en 1995 [257] que el tráfico en la red Internet presenta propiedades de dependencia a largo plazo y no cumple la tradicional distribución de Poisson. Los modelos más simples de este tipo son procesos autosimilares, caracterizados por funciones de autocorrelación descendentes hiperbólicamente. Estos procesos son particularmente atractivos porque la dependencia a largo plazo puede ser caracterizada solo con el parámetro de Hurst.

En ese mismo año 2010 Kanda *et al* [258] proponen otro método para la utilización del ACP en la detección de anomalías. El método se descompone en los siguientes pasos:

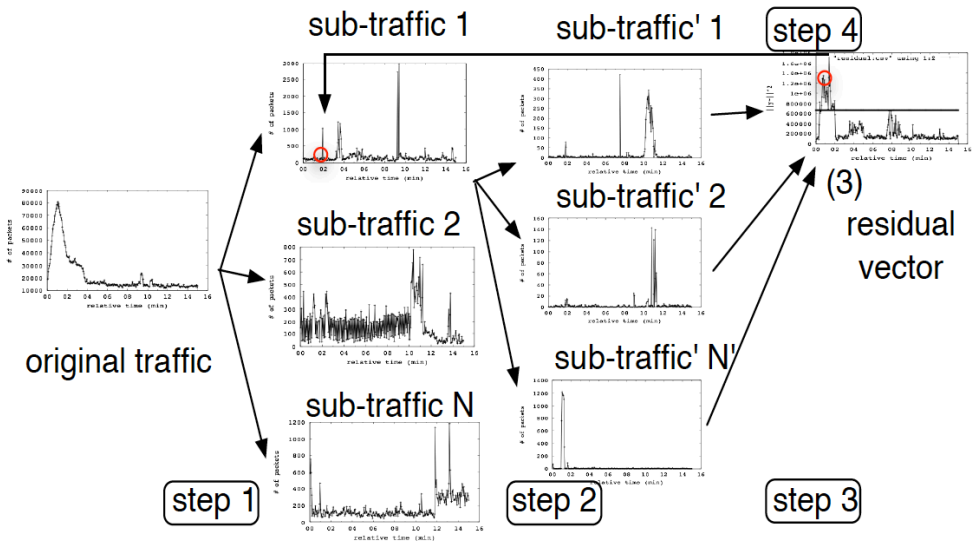
1. Proyecciones aleatorias (*Sketches*-Bosquejos): Se divide el tráfico original en N sub-tráficos, colocando cada paquete 1,…, I del trafico original de manera aleatoria en cada subtráfico 1,..,N utilizando M funciones *hash* distintas.

2. Nuevas proyecciones con diferentes *hashes*: Cada sub-tráfico anterior, se divide otra vez en N' sub-tráficos con otra función hash diferente de la utilizada en el primer paso.

3. Detector de anomalías basado en ACP: Los autores proyectan los subtráficos obtenidos en el paso anterior en el subespacio residual (anómalo), que es obtenido a partir de las series temporales. El número de componentes principales utilizado como umbral para establecer el subespacio normal –vs.- subespacio anómalo es aquel que retiene en el espacio normal el 70% de la varianza total.

4. Retener las direcciones IP de origen del subespacio: Las direcciones IP correspondientes a los puntos de tiempo anómalo de cada subtráfico 1,…,N son almacenados en una lista. Como empleamos M funciones *hash*, obtendremos M listas de direcciones IP.

5. Calcular la intersecciones de las funciones *hash*: La intersección de las M listas de direcciones IP de origen para identificar las direcciones IP de origen.



Esquema del método de detección de anomalías propuesto por Kanda *et al.* [258]

Los autores comparan su propuesta con otra similar a la presentada por Himura *et al* [246] en 2009 y que recordemos utilizaba también bosquejos y la función gamma.

Concluyen que la propuesta podría presentar problemas similares a los ya detectados derivados del uso del ACP y debidos a la contaminación del subespacio normal debido a la presencia de anomalías en las series de datos y a que el número de componentes principales utilizadas no retengan la mayoría de la varianza del tráfico.

En 2011 Callegari *et al* [259] presentan un nuevo método de detección de anomalías también basado en el ACP. Tras un preprocesado de los datos que utiliza una estructura de bosquejos similares a [258] y [246] con d funciones *hash,* el autor propone la construcción de un histograma para cada partición:

$$X(t)= [n_1(t), n_2(t),\ldots,n_N(t)]$$

Donde $n_i(t)$ es el número de bytes transmitidos por la direcciones IP i-ésima en el intervalo temporal t. Es evidente la complejidad del histograma obtenido, por lo que se concentrará toda la información en un único valor que contenga la mayoría de la información. Para ello los autores recurren, de nuevo, a la entropía:

$$H(t) = -\sum_{i=1}^{N} \frac{n_i(t)}{S} log_2 \frac{n_i(t)}{S} \qquad S = \sum_{i=1}^{N} n_i(t)$$

Desafortunadamente la entropía solo es capaz de capturar la información relativa a un único intervalo temporal, siendo mucho más importante capturar la diferencia entre las distribuciones en dos intervalos temporal adyacentes. Para ello los autores presentan una métrica adicional, la divergencia de Kullback-Leibler (K-L): dados dos histogramas X(t) y X(t-1) capturados en los intervalos t y t-1 respectivamente la divergencia K-L se define como:

$$D_{KL}(t) = \sum_{i=1}^{N} n_i(t-1) log \frac{n_i(t-1)}{n_i(t)}$$

Una vez que las series temporales son construidas, se aplica el ACP, lo que permite, como es habitual, separar las componentes principales en dos grupos: dominantes y despreciables, que serán las utilizadas para distinguir entre tráfico normal y anómalo. El número de componentes principales se elige mediante el *scree-plot*. El grupo de las r primeras componentes principales que capturen la mayoría de la varianza del tráfico permite obtener la base del subespacio “normal”

$$P=(v_1,v_2,\ldots,v_r)$$

Estos vectores nos permiten proyectar el tráfico total Y(t) en ambos subespacios obteniendo la descomposición del tráfico total como

$$Y(t)=Y^{*}(t)+Y^{**}(t)$$

Donde $Y^{*}(t)=PP^{T}Y(t)$ y $Y^{**}(t)=(I-PP^{T})Y(t)$.

La detección del tráfico anómalo puede realizarse mediante la utilización de la norma L_2 (euclídea) con un umbral preestablecido. Dado que disponíamos de d juegos de datos distintos provenientes de la separación con los *hash,* los autores proponen un simple método de "votación" para la determinar la presencia/ausencia de una anomalía en el intervalo temporal correspondiente.

El método propuesto permite también la identificación del flujo responsable de la anomalía, simplemente retirando del cálculo cada flujo de uno en uno cada vez, hasta que no se sobrepase el umbral establecido de alarma.

También en 2011 Nyalkalkar *et at* [260] efectúan una comparativa entre dos métodos de detección de anomalías: el primero basado en la combinación de entropía y ACP (ya conocido) y el segundo denominado HHH (*Hierarchical Heavy Hitter*) basado en el análisis de series temporales y *wavelets*. La utilización de la entropía y el ACP que aplica Nyalkalkar difiere de la propuesta tradicional de Lakhina [216]: no solo obtienen la entropía en términos de los contadores de flujos (en lugar de sobre paquetes que proponía Lakhina) sino que utilizan el ACP sin agregación (Lakhina agregaba a nivel de flujos de origen-destino). Veamos ahora en que consiste el método HHH. La idea básica consiste en dividir el tráfico en "cubos" de una determinada capacidad (HHH) y monitorizar las series temporales de estos cubos utilizando *wavelets*. La principal novedad de la propuesta es la combinación de estos dos procedimientos para la detección, ya que el método HHH aislado ya había sido propuesto con anterioridad por Zhang y Cormode en 2004 [261], [262]. Un *Heavy Hitter* (que podemos traducir como "peso pesado") es una entidad que acumula al menos una proporción especificada del total de actividad. Dado que las direcciones pueden ser entendidas como un ente jerárquico (a partir del prefijo de su dirección) se pueden construir *Hierarchical Heavy Hitters* o "pesos pesados jerárquicos", que se corresponderían con una entidad jerárquica que, como antes, acumula un porcentaje especificado de la actividad total. El método propuesto pasa por identificar los HHH y posteriormente analizar cada HHH individualmente. Los resultados muestran que para ataques de negación de servicio (DDoS) la combinación de HHH-*wavelets* obtiene mejores prestaciones que la

entropía+ACP, al menos para el método este último utilizado que, recordemos, no es exactamente el mismo que el propuesto por Lakhina [216]. Para otros tipos de ataques analizados no presenta mejoras significativas. Los autores vuelven a confirmar, de paso, la elevada sensibilidad del ACP para la detección de anomalías, ya mostrada, entre otros, por Ringberg [227].

En 2011 Thatte *et al* [263] desarrollan métodos paramétricos para detectar anomalías en redes utilizando exclusivamente estadísticas agregadas de tráfico, sin necesidad de separación de flujos o de una inspección en profundidad de los paquetes, cuestiones ambas complicadas de llevar a cabo en tiempo real por las prestaciones necesarias en los sistemas de detección y considerando las velocidades de transmisión actualmente en juego. El método que proponen dispone de autoaprendizaje y es capaz de evolucionar sin intervención externa. Emplean la velocidad de transmisión de los paquetes y la entropía, esta última para mejorar la robustez del método frente a falsos positivos. El método es denominado bPDM (*bivariate Parametric Detection Mechanism*) y es completamente pasivo, no introduce cargas adicionales a la red y opera sobre el tráfico agregado. Emplea una técnica de detección adaptativa para las dos características del tráfico agregado que controla: la velocidad y el tamaño de los paquetes. Los autores evalúan su propuesta utilizando simulaciones que posteriormente validan con ataques reales, concluyendo que el método permite detectar ataques en pocos segundos y con una sensibilidad elevada manteniendo reducida la tasa de falsos positivos.

En 2012 Pascoal *et al* [264] proponen un método de detección de anomalías que combina un algoritmo de selección de características y un método de detección de *outliers*, utilizan estadísticas robustas y también hacen uso de un procedimiento automático para determinar el número de características relevantes. La detección de *outliers* está basada en ACP robusto que, en contraposición al ACP clásico, no es sensible a *outliers*.

La selección de características se apoya en la métrica denominada Información Mutua IM (*MI - Mutual Information*), que captura dependencias lineales y no lineales y que afirman tiene gran aceptación. Han sido propuestos en la literatura múltiples estimadores para MI, pero los autores concluyen en el estimador empírico con discretización de intervalos iguales obtiene el mejor compromiso entre prestaciones y complejidad y tiene la expresión siguiente:

$$MI^{*} = -\sum_{i=1}^{m_x} p_i^{*} \log p_i^{*} - \sum_{j=1}^{m_y} q_j^{*} \log q_j^{*} - \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p_{ij}^{*} \log p_{ij}^{*}$$

Donde m_x y m_y son el número de intervalos considerados para la muestra x e y, p^{*}_i y q^{*}_j son las proporciones de observaciones de la muestra x e y que caen en el i-ésimo y j-ésimo intervalo, y finalmente p^{*}_{ij} es la porción de las observaciones (x_k,y_k) tales que x_k pertenece al i-ésimo intervalo e y_k pertenece al intervalo j-ésimo, asociados con la muestra x e y respectivamente. Los autores proponen un algoritmo para robustecer el estimador anterior. Es significativo el "parecido" de la expresión de la IM con la entropía ya aplicada anteriormente por otros autores.

La detección de *outliers*, esto es, de las anomalías, se realiza mediante ACP robusto. Como en otros casos precedentes [211], [244] se propone la utilización de la distancia de Mahalanobis, si bien para robustecer esta distancia es necesario obtener en primer lugar la matriz de covarianzas robusta, lo que puede obtenerse reduciendo el número de características mediante un ACP robusto. Los autores referencian el método de búsqueda de proyecciones (*Projection Pursuit*-PP) propuesto también por otros autores [251]. Para la detección de anomalías utilizan dos distancias diferentes:

- *Score Distance* (SD) que se corresponde con la distancia de Mahalanobis en el espacio de las componentes principales.

$$SD(\mathbf{x}_i) = \left(\sum_{j=1}^{k} z_{ij}^2 / \lambda_j^{*} \right)^{1/2}$$

Donde $\mathbf{z}_i=(z_{i1},\ldots,z_{ik})$ representa al sujeto i-ésimo en el subespacio de las componentes principales, en las k primera componentes principales.

- Distancia Ortogonal (OD) que es la distancia de una observación al espacio de las componentes principales.

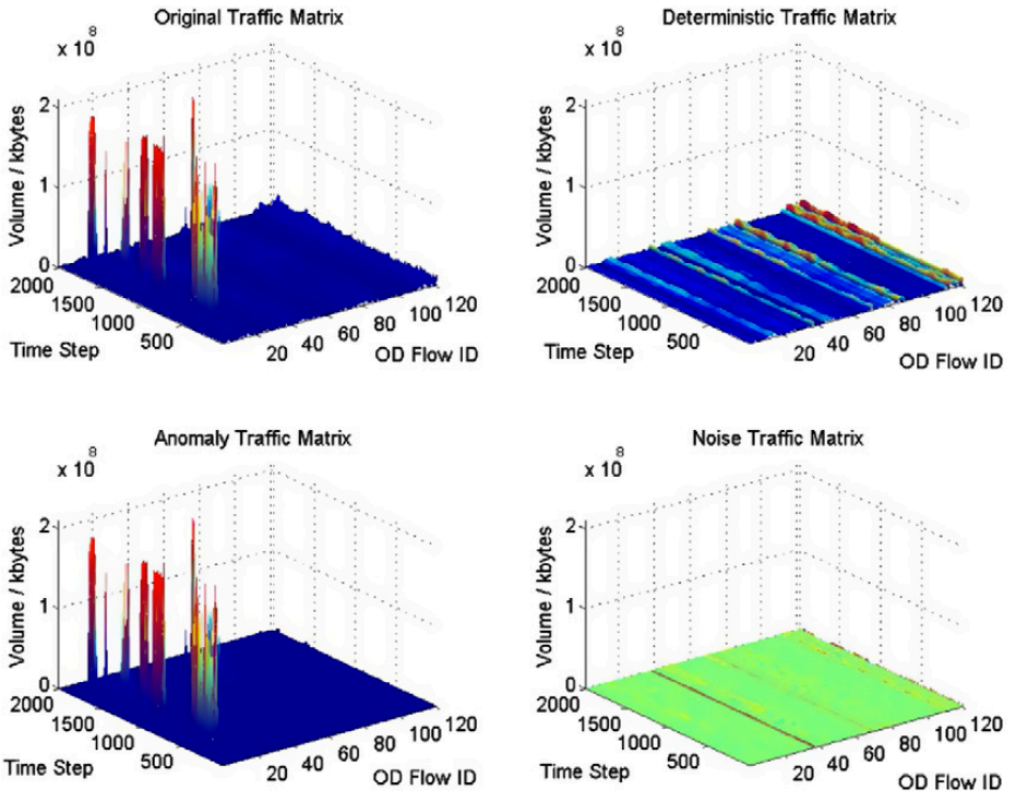
$$OD(\mathbf{x}_i) = \left\| \left(\mathbf{x}_i - \mu^{*} \right) - \mathbf{P}_{p,k}\mathbf{z}_i \right\|$$

Donde μ^{*} es la estimación del vector de medias y $\mathbf{P}_{p,k}$ es la matriz p x k que contiene las *loadings* de las primeras k componentes principales en columnas.

El mecanismo ha de ser entrenado para obtener los umbrales de ambas distancias, con las que se comparará el tráfico real para identificar las anomalías.

También en 2012 Wang *et al* [265] publican un estudio sobre la análisis estructural de la matriz de tráfico de red mediante el algoritmo de Búsqueda de Componentes Principales Relajada (*Relaxed Principal Component Pursuit*). Partiendo de los trabajos de Lakhina *et al* [212], [215] aplicando el ACP a las matrices de tráfico O-D y el de Ringberg [227] sobre la sensibilidad del método a la corrupción por las propias anomalías los autores analizan varios métodos de ACP modificados para solventarlas. Así tratan con el ACP Robusto, el Búsqueda de Componentes Principales Relajado (*Relaxed PCP*) utilizado a través del método del Gradiente Próximo Acelerado (APG – *Accelerated Proximal Gradient*) que es más eficiente que el RPCP. Básicamente el resultado es la descomposición de la matriz de tráfico original en tres matrices: la matriz determinística, la matriz de anomalías y la matriz de ruido.

Los autores concluyen que el ACP no descompone bien la matriz de tráfico en presencia de anomalías de volumen importantes, proponiendo el mencionado nuevo modelo de descomposición en tres matrices a través del ACP Robusto conocido, pero aplicando el método *Relaxed PCP*.

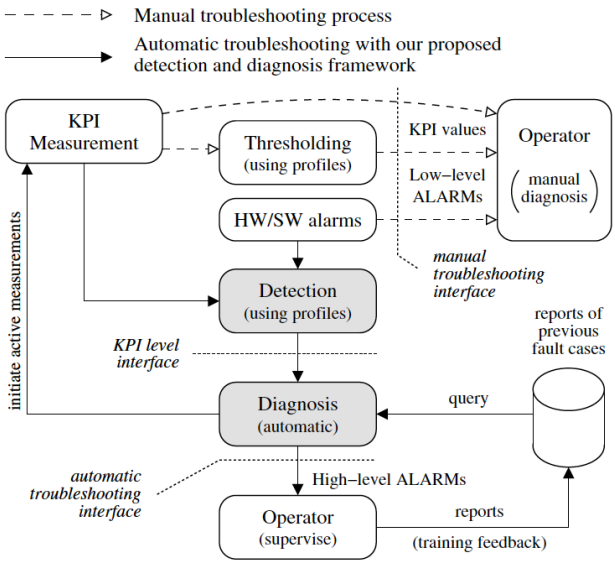


Resultado de la descomposición mediante Búsqueda de Componentes Principales Relajada. [265]

En 2012 Szilágyi y Nováczki [266] proponen un esquema de detección y diagnosis de anomalías en sistema de comunicación móviles. Si bien las redes móviles no constituyen, por ahora, un objetivo para nuestra propuesta, el trabajo nos ofrece un esquema de trabajo perfectamente aplicable a nuestro caso y desarrollado con un, a nuestro modo de ver, elegante paralelismo con la medicina.

El autor expone que en el lenguaje coloquial, la detección y el diagnóstico no están claramente separados. Con la frase "detectar un problema" a menudo se agrupan realmente dos cuestiones: en primer lugar la confirmación de que efectivamente existe un problema, y en segundo lugar la determinación de la naturaleza o tipo de problema presente. Así, si se detecta un comportamiento inusual, tiene que se llevarse a cabo un más profundo análisis para descubrir la causa de dicho comportamiento. Los autores definen la detección como la identificación de algo inusual en la red, mientras que la diagnosis implica la investigación de la raíz del problema que ha causado los síntomas detectados. Efectivamente, como se puede aventurar, por lo general a continuación de la detección y diagnosis se llevarán a cabo las medidas correctoras oportunas para resolver el problema.

Los síntomas en la analogía propuesta por los autores se corresponden con los Indicadores de prestaciones clave (KPI: *Key Performance Indicators*) que son monitorizados por el sistema. Más adelante de su artículo especifican que el conjunto de valores "normales" de los KPI se denomina perfil del KPI. Efectivamente, los síntomas se corresponderían con desviaciones de dichos valores normales. El autor propone, como hemos indicado, un flujo de trabajo automático para la detección y diagnosis de problemas, que compara con el tradicional método manual.



Esquema de resolución de problemas propuesto por Szilágyi y Nováczki. [266]

La propuesta permite utilizar tanto medidas activas de los KPI como pasivas. Utiliza una suerte de lógica difusa en la que el comportamiento de un KPI es comparado con su perfil previamente obtenido, resultando un indicador que denominan nivel del KPI, que es un número entero en el rango continuo [0,1], con nivel 0 significado perfecta adaptación al perfil y aproximándose asintóticamente a 1 conforme se desvía de dicho perfil. Como puede desprenderse del método, la obtención del perfil es una fase crítica del método. El modelo propuesto está basado en la obtención de una función de distribución normal a partir de la medidas en los KPI en ventanas temporales deslizantes.

Por último, pero no menos importante, más bien todo lo contrario, Roughan *et al* publican en junio de 2012 un artículo [267] que postula la aplicación de la descomposición de la matriz de tráfico **X** en la forma:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T = \mathbf{L}\mathbf{R}^T$$

donde $\mathbf{L}=\mathbf{U}\Sigma^{1/2}$ y $\mathbf{R}=\mathbf{V}\Sigma^{1/2}$

Esto es, como veremos, un **BIPLOT** del tipo SQRT. Los autores afirman que ellos mismos fueron los primeros que propusieron dicha descomposición en el año 2009: el artículo publicado en 2012 se corresponde directamente con una versión extendida de aquel previo. No obstante, existen antecedentes en la proposición de métodos BIPLOT (concretamente el HJ-Biplot) para la detección de anomalías en el tráfico de redes en el año 2000 [268] y 2007 [269] (autocitación). Por la relevancia del artículo de Roughan, analizaremos con detalle su línea expositiva, metodología y conclusiones.

Los autores comienzan su artículo exponiendo la evidente multidimensionalidad intrínseca de los datos relativos al tráfico en redes: la matriz Origen-Destino (O-D) bidimensional, se transforma en tridimensional al incorporar la dimensión tiempo. A este último tipo de objeto lo denominan Matriz de Tráfico que, como indican, es el dato de entrada de múltiples tareas, tales como, ingeniería de tráfico, planificación de capacidades de la redes y, por supuesto, la detección de anomalías. El artículo expone las dificultades para la obtención de la matriz O-D en redes reales, y cómo puede ser solo estimada a partir de las cargas de tráfico en los distintos enlaces que forman la red, al tratarse de un problema indeterminado. Por otro lado, las matrices O-D suelen presentar, a menudo, un número significativo de elementos faltantes, que dificultan los análisis que se pretendan aplicar. Una de las cuestiones a afrontar en

estos estudios es “completar” la matriz de tráfico, lo que se suele abordar bien mediante interpolación, o utilizando la “detección de compresión” (*compressive sensing*) una metodología genérica para hacer frente a la existencia de valores perdidos en la matriz, que aprovecha la presencia de cierto tipo de estructura y la redundancia de los datos recogidos. Los autores afirman que este último método, que goza de cierta popularidad, no funciona bien en el tipo de datos objeto del estudio. Así, proponen el análisis de la matriz de tráfico, a través de la descomposición SQRT-Biplot (aunque ellos la denominan *Sparsity Regularized Matrix Factorization* – SRMF, no es más que un SQRT-Biplot). Afirman igualmente que SRMF (o SQRT-Biplot) representa el primer modelo espacio-temporal genuino de la matriz de tráfico y que sus prestaciones con excelentes. La aproximación de bajo rango combinada con el modelo espacio-temporal opera bien en escenarios con un elevado número de elementos faltantes, mientras que interpolaciones locales permiten resultados igualmente buenos para escenarios con bajo número de elementos faltantes.

La matriz de tráfico es una matriz no negativa Z(i,j) que describe volúmenes de tráfico (en bytes, paquetes o flujos) entre un origen i y una destino j. Para una red con N nodos, la matriz de tráfico es una matriz cuadrada de N x N. No obstante esta definición precisa alguna puntualización, indican los autores:

- La matriz es medida sobre un intervalo de tiempo, y el valor de la matriz es un promedio. Así, se denota Z(i,j;t) el tráfico desde i a j promediado entre $[t,t+\Delta t)$. Por otro lado denominaremos una instantánea (*snapshot)* de la matriz de tráfico a Z(*,*,t) siendo conscientes de que en realidad se trata de un intervalo.
- Aunque es común hablar de “origen-destino” es a menudo difícil asignar las direcciones IP presentes en el tráfico a las direcciones IP de origen/destino reales, por lo que típicamente Z(i,j;t) representa el tráfico entrante en la red en el router i y saliendo por el router j.

Como ya se ha indicado la matriz de tráfico así definida es intrínsecamente tridimensional, siendo habitual tomar una instantánea de la matriz y apilar las columnas en forma de un vector columna que los autores denominan $\mathbf{x}_t$ (ver propuestas de Kiers [194]). Estos vectores pueden a su vez agruparse en una matriz **X** de n x m donde $n = N^2$ y m son los intervalos temporales bajo estudio, sobre la que es más fácil operar algebraicamente que en el caso de una matriz tridimensional.

Para construir una aproximación de bajo rango de una matriz **X** real (como es la de tráfico) los autores exponen la "popularidad" de la Descomposición en Valores Singulares (DVS)

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$$

En análisis de trafico la DVS aparece vinculada, generalmente, al Análisis de Componentes Principales (ACP), relacionándose ambos, como sabemos, por el hecho de que las columnas de **U** constituyen los ejes principales del ACP. De aquí continúan con la factorización ya expuesta anteriormente

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T = \mathbf{L}\mathbf{R}^T$$

donde $\mathbf{L}=\mathbf{U}\Sigma^{1/2}$ y $\mathbf{R}=\mathbf{V}\Sigma^{1/2}$

Los algoritmos típicos para calcular la DVS asumen en **X** es perfectamente conocida, dado que en el caso bajo estudio, por lo general, existirán datos faltantes será preciso incorporar información adicional para "rellenar" esos vacíos, para lo que proponen utilizar las restricciones o condiciones que se conozcan sobre los datos disponibles.

Los autores aplican su modelo a tomografía de red, predicción y detección de anomalías. En este último caso proponen varias estrategias:

1. Diferenciación: Se trata de una técnica habitual en series temporales consistente en eliminar tendencias lineales, de forma que se remarquen cambios bruscos. Implícitamente utiliza datos del intervalo temporal anterior como modelo. Aunque se trata de un método sencillo al que hemos de añadir algo de intuición, no ha sido muy utilizado. Analíticamente la diferenciación es una postmultiplicación por una matriz de Toeplitz (0,1,-1), una operación puramente temporal que hace uso de la correlación espacial entre los elementos de la matriz de tráfico.
2. PCA/SVD: Como viene siendo habitual su aplicación eligiendo el rango r del subespacio "normal" y proyectando los datos X en el espacio "anómalo", donde se localizan los eventos. Este método es puramente espacial y la reordenación de los datos en el tiempo (columnas de **X**) no tiene ningún efecto en los resultados.
3. SRMF (*Sparsity Regularized Matrix Factorization*) método que partiendo de la factorización ya expuesta $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T = \mathbf{L}\mathbf{R}^T$ unido al conocimiento de la

información del tráfico para completar la matriz, utilizando una combinación lineal de los k-elementos más cercanos al faltante, utiliza la estructura espacial y temporal de los datos, mediante la utilización en el proceso de factorización de unas matrices **S** y **T** que incorporan esta información.

Los autores recurren a la simulación para verificar su propuesta argumentando la simulación es válida para la comparación entre técnicas de detección de anomalías porque:

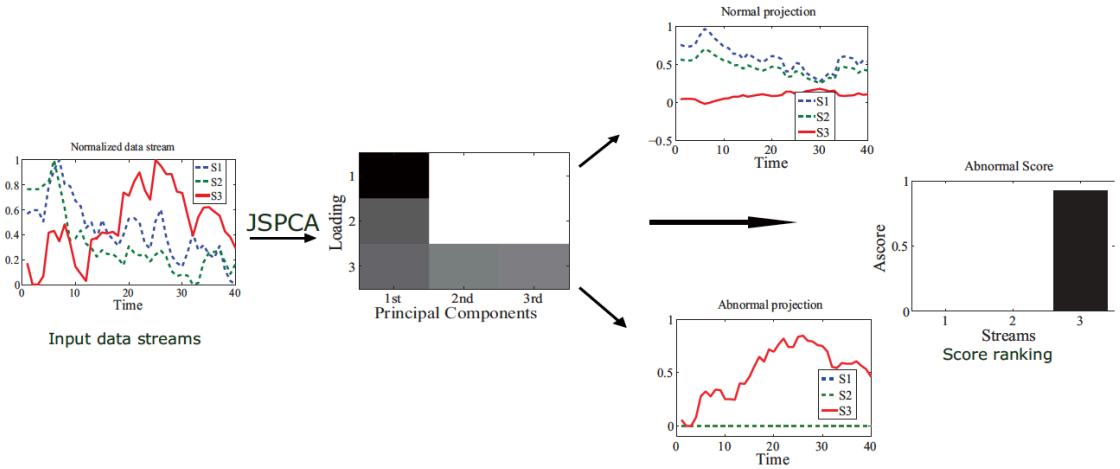
1. Es preciso conocer con total precisión el contenido del juego de datos utilizado para poder estimar la probabilidad de detección y falsa alarmas, ya que ambas probabilidades son necesarias. En juegos de datos reales, por definición, no conocemos con certeza su contenido.
2. Son precisos muchos resultados distintos para estimar correctamente las probabilidades.
3. Las simulaciones permiten variar parámetros tales como la duración o intensidad de la anomalía para estudiar sus efectos.

Si bien afirman que la simulación es necesaria para validar el método, pero no es suficiente. Como curiosidad, los autores otorgan a Ringberg *et al* [227] la argumentación precedente, aunque no hemos podido localizarla en dicho artículo.

Roughan *et al* concluyen que la ventaja obtenida por el método que proponen, y verificada por sus simulaciones, reside en la utilización de modelos espacio-temporales. Comienzan, así mismo, a plantear la posible utilización de métodos de análisis que permitan el estudio directo de matrices tridimensionales sin necesidad de colapsarlas a matrices bidimensionales, que será lo que se llevará a cabo en este trabajo con la aplicación del método STATIS Dual.

Jiang *et al*, presentan en 2013 [270] un conjunto de algoritmos basado en el ACP adaptado a las especificidades de los datos de tráfico en redes: correlación espacial y temporal de los datos, dificultad del ACP para la identificación de la fuente de la anomalía, entre otras. Los autores introducen también [212], [250] la extensión de Karhunen-Loève (KLE) para extender el ACP para considerar la correlación temporal de los datos, además de la espacial. Si bien algunos autores tratan la KLE como un método equivalente al ACP [212], otros la identifican como un caso general del ACP [250]. Los autores [270] incluyen un anexo en su artículo sobre la conexión existente entre el ACP tradicional y la KLE en el que apuntan a la KLE como una expansión del

ACP tradicional. Jiang *et al* proponen una descomposición trifactorial sobre este planteamiento de la matriz de datos en los habituales subespacios “normal” y de “anomalías”, si bien la descomposición presenta las peculiaridades antes indicadas para robustecerla y hacer que considere todas las correlaciones existentes en los datos de partida.



Arquitectura propuesta por Jiang *et al* aplicada a tres series de datos, dos “normales” y una anómala. [270]

Jiang *et al* concluyen que su propuesta mejora significativamente la aplicabilidad de las técnicas de detección y localización de anomalías basadas en el ACP. La incorporación de la extensión de Karhunen-Loéve posibilita la estabilización del método, al incorporar información espacial y temporal en el procedimiento. Finalmente los resultados experimentales sobre varios juegos de datos demuestran, en su opinión, la efectividad de la propuesta presentada.

5.3. DETECCIÓN DE ANOMALIAS EN LA RED DIPSANET

5.3.1. DESCRIPCIÓN DEL JUEGO DE DATOS DIPSANET-96

Se dispone de un juego de datos real, que hemos denominado DIPSANET-96, en el que disponemos de 6 lecturas de 4 variables de tráfico cursado medidas en la red Ethernet corporativa de la Diputación Provincial de Salamanca (DIPSANET)[2]. Las medidas fueron capturadas a través del protocolo SNMP entre los días 14 de noviembre de 1996 y 7 de enero de 1997.

Vista aérea de los edificios de la Diputación de Salamanca en los que estaba instalada la red Ethernet bajo estudio

Estas medidas se tomaron en 8 segmentos diferentes de la red DIPSANET que se encuentran distribuidos por 3 edificios administrativos anejos de la Diputación Provincial de Salamanca en la calle Felipe Espino 1 de Salamanca. Estos segmentos tenían en aquellas fechas conectados cada uno un número diferente e indeterminado en este momento de dispositivos de red y usuarios:

- **EXP**: Segmento al que estaba conectado exclusivamente el servidor principal.

[2] La utilización y difusión de esos datos ha sido autorizada por la Diputación Provincial de Salamanca en fecha 29 de marzo de 2000. El autor agradece a la Corporación Provincial el permiso de uso concedido.

- **LR-P2, LR-P1, LR-PB, LR-PS**: Cuatro segmentos correspondientes a cada una de las 4 plantas del edificio situado en la calle La Rua.
- **SP-A, SP-B**: Dos segmentos que abarcan equipos ubicados en otro edificio diferente, en este caso el situado en la calle San Pablo.
- **FEsp**: Segmento único englobando todas las plantas del edificio situado en la calle Felipe Espino.

Las variables bajo estudio para cada uno de los 8 segmentos mencionados son:

- **SCol**: Colisiones simples
- **MCol**: Colisiones múltiples
- **RxFrms**: Tramas recibidas
- **TxFrms**: Tramas transmitidas

Serie	Inicio	Final	Long.
0	14-nov-1996	28-nov-1996	14
1	28-nov-1996	02-dic-1996	4
2	02-dic-1996	10-dic-1996	8
3	10-dic-1996	17-dic-1996	7
4	17-dic-1996	02-ene-1007	16
5	02-ene-1007	07-ene-1997	5

Tabla resumen de las series de datos utilizadas medidas en la red DIPSANET

Los datos son brutos, esto es, no están promediados por número de usuarios o de equipos (por desconocerse su número en este momento), por lo que, en principio, este hecho podría constituir un condicionante al resultado de los análisis. De igual manera el número de tramas transmitidas está relacionado con el nivel de tráfico cursado, pero no directamente ya que, como se ha comentado, el tamaño de la tramas en las redes Ethernet no es fijo. No obstante es un indicador aceptado de intensidad de tráfico.

Serie 0	Single-Col	Multi-Col	Rcv-Frms	Xmit-Frms
EXP	30	26	5.090.532	6.570.283
LR-P2	831	225	109.236	1.486.173
LR-P1	601	211	55.005	1.409.410
LR-PB	2.190	796	243.129	1.589.567
LR-PS	725	209	313.348	922.803
SP-A	1.918	637	3.321.429	4.622.701
SP-B	0	2	117.770	1.326.499
FEsp	8.092	4.839	7.239.137	3.137.964

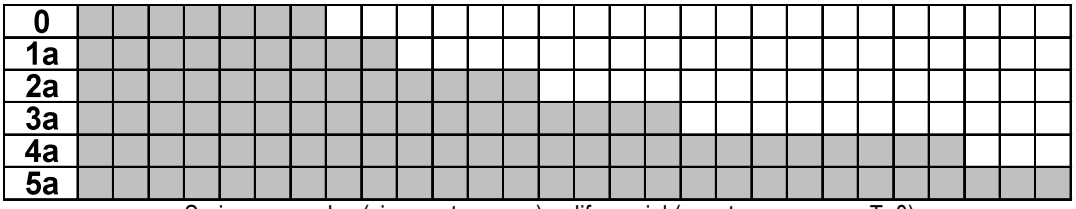
Ejemplo de Matriz de datos (T=0)

Las lecturas se realizaron de manera agregada, esto es, parten de la lectura efectuada en T=0 (14 de noviembre) y las variables bajo estudio aumentan conforme el tiempo transcurre. No obstante es trivial pasar de estas matrices "agregadas" a matrices "diferenciales" o "incrementales" simplemente restando los valores de las matrices agregadas.

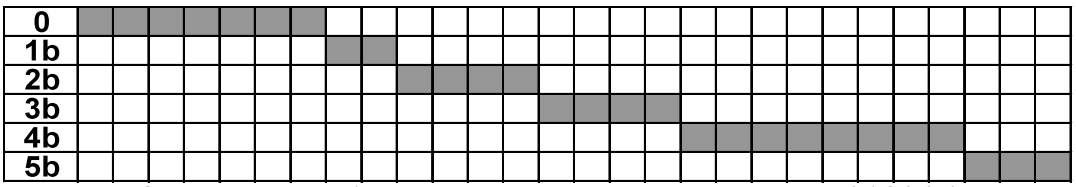
Así, sea **Xa**[0] la matriz de datos inicial (que no parte de valores nulos)

Matriz **diferencial** para el intervalo T=n: **Xd**[n] = **Xa**[n]-**Xa**[0]

Matriz **incremental** para el intervalo T=n: **Xi**[n] = **Xa**[n]-**Xa**[n-1]



Series agregadas (sin puesta a cero) o diferencial (puesta a cero para T=0)



Series incrementales (con puesta a cero al principio de cada intervalo T=0,1,2,3,4,5)

El juego de datos DIPSANET-96 tiene claramente tres vías, por lo que para realizar el análisis HJ-Biplot (2 vías) es posible trabajar con, al menos, dos proyecciones diferentes de nuestra matriz de datos tridimensional:

$$\mathbf{X}_{a,d,i} = x(s,v,t)$$

En donde s = segmento (EXP, LR-P2/P1/PB/PS, SP-A/B, FEsp)

v = variable (SCol, MCol, RxFrms, TxFrms)

t = tiempo N, (con N=0,…,5)

a,d,i = datos agregados, diferenciales, incrementales

En nuestro caso analizaremos las 6 diferentes matrices correspondientes a los intervalos de tiempo en los que se obtuvieron las muestras

$$\mathbf{X}_{a,d,i} = x(s,v,0:5)$$

5.3.2. NORMALIZACIÓN DE LA MATRIZ DE DATOS

Las matrices de datos se normalizan individualmente para que las variables (columnas) tengan media nula y presenten varianza unitaria [271].

Esto es, se aplica la siguiente fórmula a la matriz de datos original (para cada las variables correspondientes a cada intervalo por separado):

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{S_i} \qquad \bar{x}_i = \frac{1}{J}\sum_{j=1}^{J} x_{ij} \qquad S_i = \left(\frac{\sum_{j=1}^{J} \left(x_{ij} - \bar{x}_i \right)^2}{J-1} \right)^{1/2}$$

La aplicación de la normalización se justifica considerando las diferentes características de las variables intervinientes en el estudio [271].

5.3.3. ANÁLISIS HJ-BIPLOT PARA CADA INTERVALO TEMPORAL

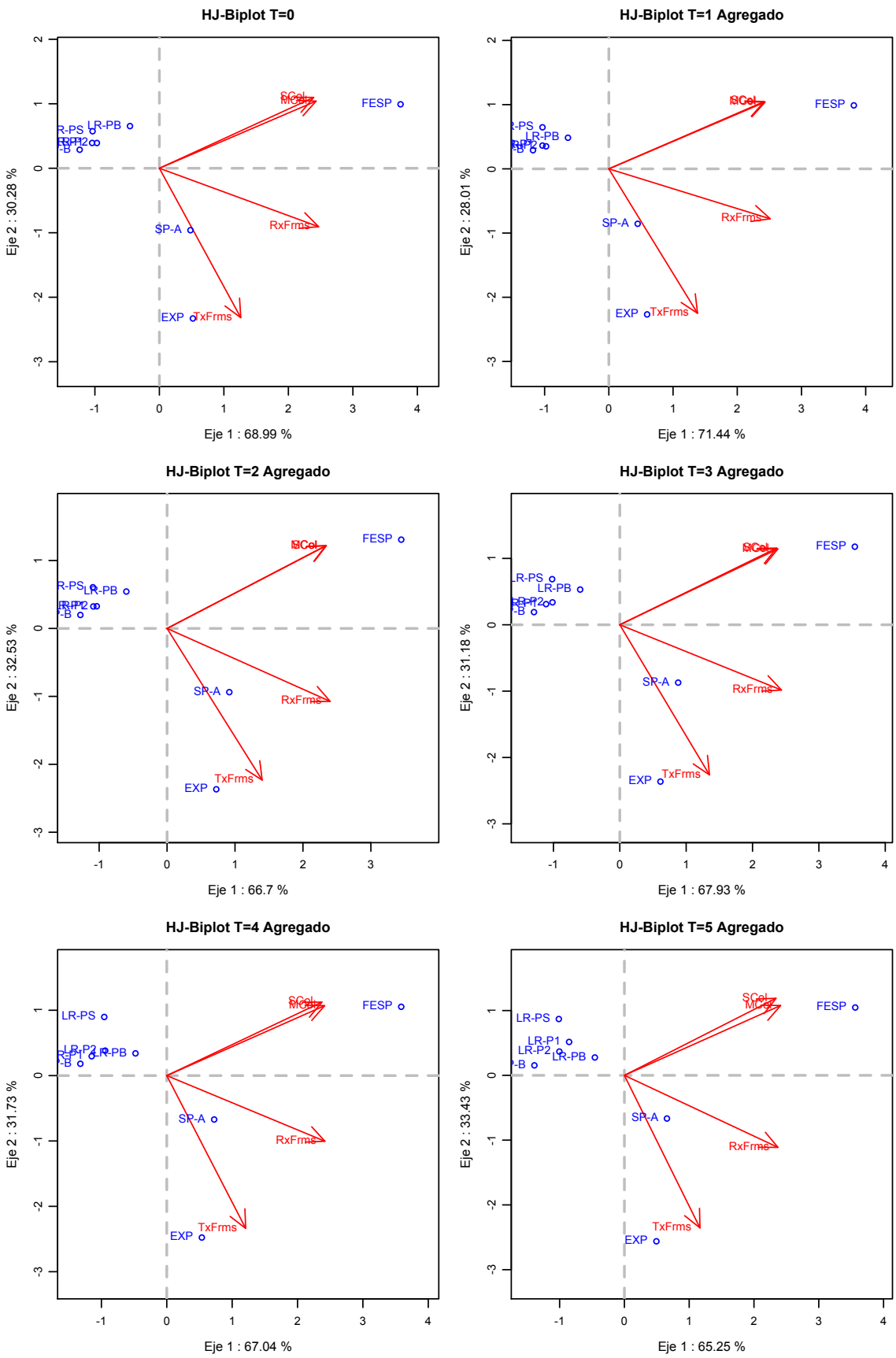
A continuación se procede a aplicar el análisis HJ-Biplot [160] a las diferentes matrices de datos. Utilizaremos para el estudio el paquete de software libre R [272].

5.3.3.1. Matrices agregadas para cada intervalo temporal

Se aplica un análisis HJ-Biplot sobre las 6 matrices de datos agregados que han sido previamente normalizadas por columnas para que presenten media nula y varianza unidad [271], obteniendo los resultados gráficos mostrados en la página siguiente.

Lo primero que se observa por simple inspección de los gráficos obtenidos es la **estabilidad** de la configuraciones obtenidas, todo ello a pesar de:

- Las diferentes longitudes de los intervalos temporales bajo estudio (entre 4 y 16 días)
- El diferente número de equipos/usuarios en cada segmento considerado (pero que permanece estable a lo largo de la toma de datos)



Secuencia temporal HJ-Biplot datos DIPSANET-96 Agregados y normalizados.

En los seis análisis realizados la varianza explicada en el plano (dos primeros ejes) es superior al 98%, lo cual no es de extrañar dado que la matriz de datos solo tiene 4 variables y la reducción de la dimensionalidad aplicada es, cuando menos, modesta (50%). La calidad de representación para las variables es buena y, en general, para los segmentos de red es aceptable. Algunos marcadores, puede comprobarse en los gráficos, se encuentran muy próximos al origen de coordenadas. Veamos como ejemplos las CRFE para el primer intervalo (T=0):

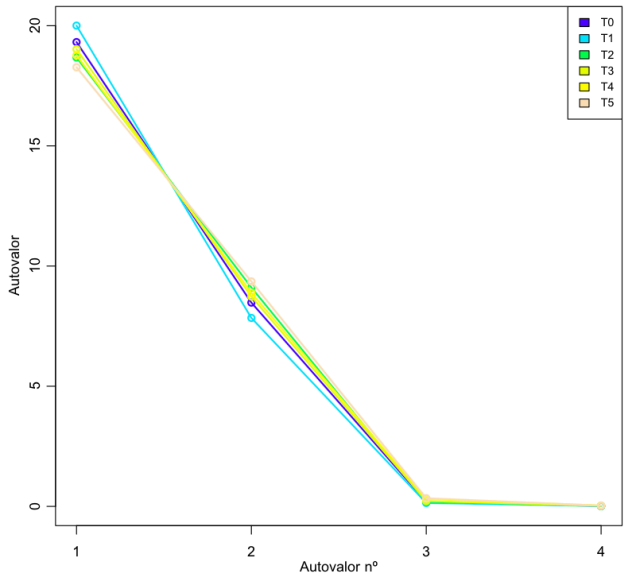
CRFE (T=0)

Fila	Eje 1	Eje 2	Acumulada
EXP	47.3	950.9	998.1
LR-P2	858.6	140.1	998.8
LR-P1	874.6	123.8	998.4
LR-PB	293.0	602.0	895.0
LR-PS	747.5	229.1	976.6
SP-A	192.8	767.4	960.2
SP-B	932.6	50.3	983.0
FEsp	933.9	65.7	999.6

Columna	Eje 1	Eje 2	Acumulada
Single-Col	818.0	172.5	990.5
Multi-Col	843.5	154.9	998.5
Rcv-Frms	870.7	117.3	988.0
Xmit-Frms	227.4	766.5	994.0

La CRFE acumulada en el plano no baja en ningún caso de los intervalos temporales, para todos los segmentos y variables, del 87,2% (para el segmento LR-P2 en el intervalo T=5) y solo en 5 casos del total (el total de casos bajo estudio sería de 6 intervalos, 4 variables, 8 segmentos = 192) bajó del 90%.

El gráfico de autovalores refleja las similitudes entre las 6 estructuras:



Screeplots superpuestos de las matrices DIPSANET-96 agregadas y normalizadas.

De estos resultados se puede concluir que el HJ-Biplot de los datos de tráfico agregados puede ser un buen método de modelado de la red y permitiría establecer un perfil "basal" o "típico" de su comportamiento, con el que poder comparar las sucesivas representaciones para detectar anomalías o comportamientos espurios.

5.3.3.2. Matrices diferenciales para cada intervalo temporal

En el caso anterior se podría pensar que la estabilidad de los marcadores entre intervalos temporales de deba a la elevada aportación de los contadores iniciales. Si comparamos porcentualmente los valores de las variables iniciales (T=0) con los valores al final del intervalo bajo estudio (T=5) mediante la expresión (**X**[t=0] / **X**[t=5]) obtenemos las siguientes relaciones:

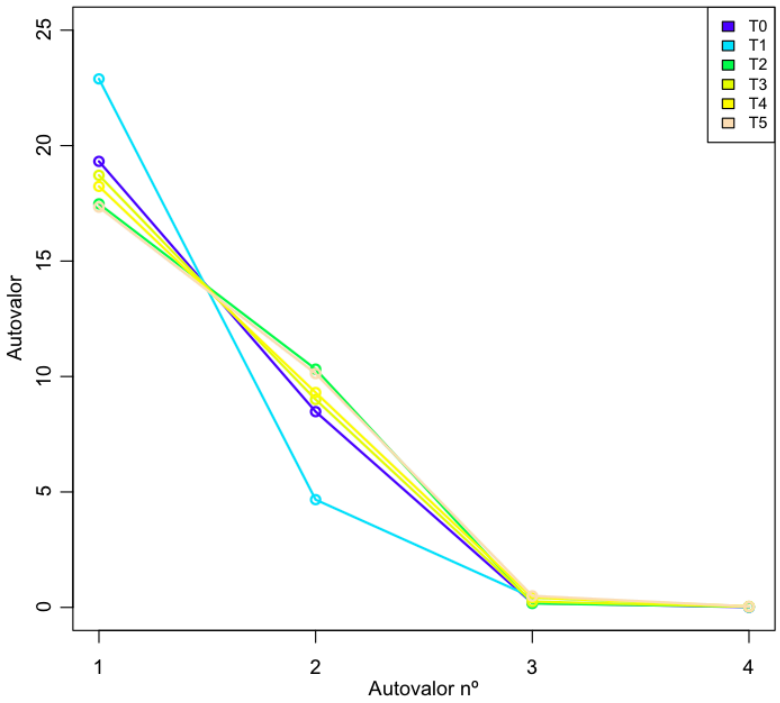
	Single-Col	Multi-Col	Rcv-Frms	Xmit-Frms
EXP	17%	16%	33%	31%
LR-P2	30%	29%	41%	26%
LR-P1	15%	20%	49%	26%
LR-PB	45%	50%	11%	21%
LR-PS	21%	20%	36%	45%
SP-A	27%	29%	38%	33%
SP-B	0%	50%	27%	25%
FEsp	42%	48%	40%	33%

Contribuciones de los valores para T=0 en los valores para T=5

No obstante, para analizar el efecto que sobre la representación HJ-Biplot tienen los valores iniciales de los marcadores, a continuación abordamos el análisis de las matrices que hemos denominado "diferenciales", que no son más que las matrices originales, considerando que los contadores se hubiesen puesto a cero en el inicio del experimento, aunque, por su relevancia, se mantendrán los valores para T=0.

Al igual que en análisis precedente, y por los motivos y expuestos, matrices se normalizan (media nula y varianza unidad por columnas/variables) antes de someterlas al análisis HJ-Biplot [271].

De nuevo el porcentaje de varianza explicada por los dos primeros ejes es superior al 98% en todos los casos, y las calidades de representación para las variables son buenas, y aceptables en general para los diferentes segmentos de red.



Screeplots superpuestos de las matrices DIPSANET-96 (datos diferenciales y normalizados).

Con idéntico comentario, la representación de los autovalores muestran comportamientos muy similares, si bien la serie T=1 es notablemente diferente.

Los gráficos continúan mostrando una significativa **estabilidad** a lo largo del tiempo, a pesar de las diferentes longitudes temporales consideradas y el diferente número de usuarios/equipos conectados a cada segmento. No obstante se observa una variabilidad mayor que en la situación precedente (series agregadas).

Veamos las CRFE de los marcadores en el último intervalo temporal, esto es, el que acumula todas las mediciones a lo largo del periodo bajo estudio.

Fila	Eje 1	Eje 2	Acumulada
EXP	35.7	959.8	995.5
LR-P2	877.8	117.5	995.3
LR-P1	714.3	260.6	974.9
LR-PB	644.8	235.7	880.5
LR-PS	550.2	405.8	956.0
SP-A	430.8	444.1	874.9
SP-B	963.8	11.8	975.6
FEsp	919.6	79.4	999.0

Columna	Eje 1	Eje 2	Acumulada
Single-Col	782.5	202.0	984.5
Multi-Col	830.6	166.0	996.6
Rcv-Frms	802.5	176.6	979.1
Xmit-Frms	194.4	792.4	986.8

CRFE para marcadores de acumulados (T=5 agregada)

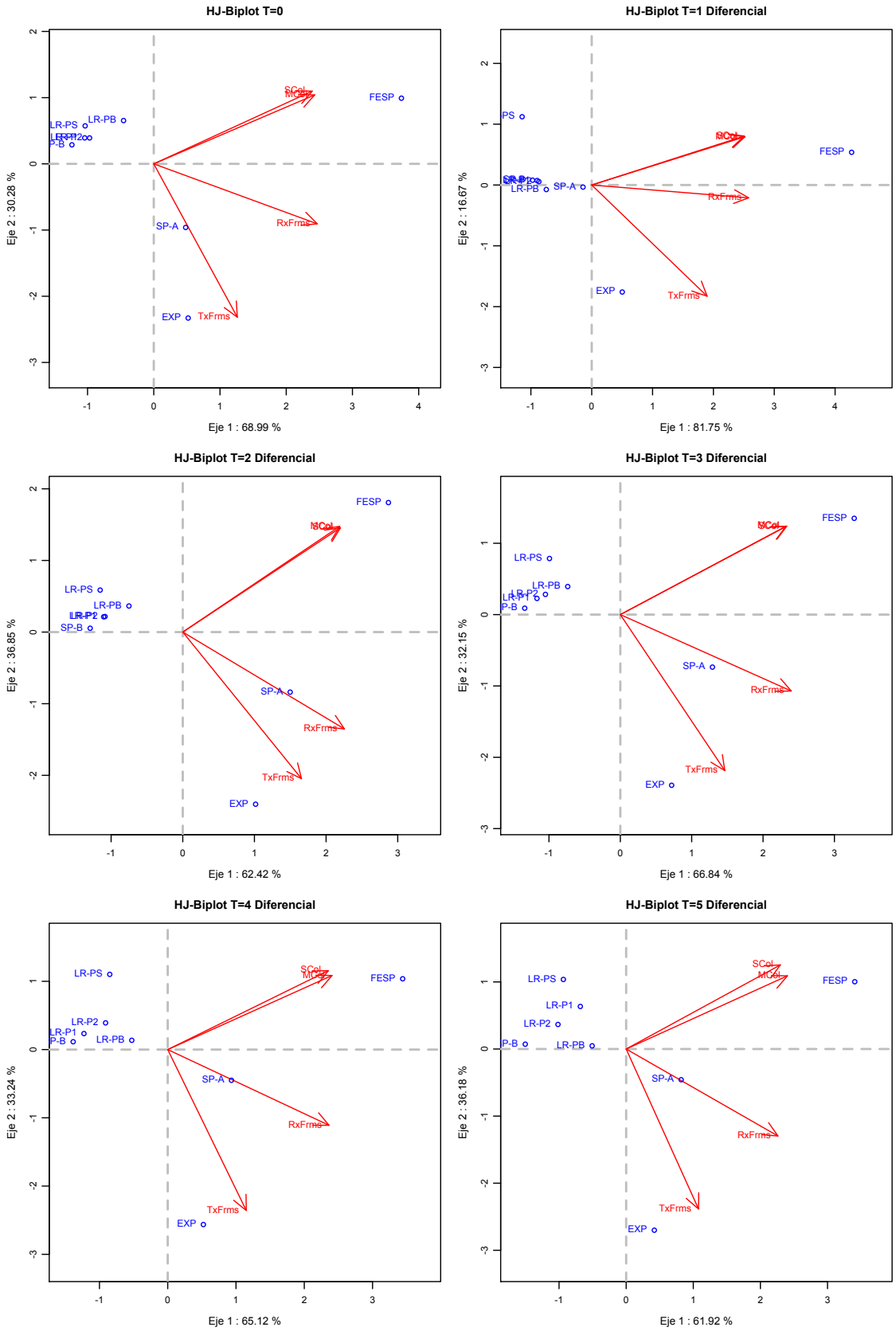
Observamos elevadas contribuciones para el plano, lo que nos indica una buena calidad de representación para todos los marcadores bajo estudio. Ciertamente el estudio de los autovalores nos indica ya un resultado en este sentido, ya que la inercia absorbida en el primer plano es superior al 95% en todos los intervalos.

Este caso, en el fondo, ofrece un resultado intermedio entre el agregado, más estable y con posible aplicación para el modelado, y el incremental que veremos en último lugar.

Los resultados gráficos obtenidos gráficos se muestran a continuación. En ellos podemos ver como se aprecia un grupo de segmentos que engloba a todos excepto FEsp, SP-A y EXP, mientras que estos últimos se encuentran aislados. Recordemos que estos tres segmentos corresponden a los dos segmentos con mayor número de usuarios y el correspondiente al servidor de red, encargado de suministrar los diferentes servicios y que por lo tanto tiene una significativa carga de tráfico.

Por lo que respecta a las variables, las correspondientes a las colisiones están fuertemente correlacionadas (ángulo muy bajo) lo que es lógico.

El segmento de FEsp presenta mayor preponderancia en colisiones, lo que también se deduce de un mayor número de equipos conectados y previsiblemente de tráfico cursado. En cuanto al tráfico la relación entre tramas transmitidas y recibidas se mantiene aproximadamente estable a lo largo del intervalo temporal bajo estudio y los marcadores de los segmentos oscilan alrededor de posiciones comparables.



Secuencia temporal HJ-Biplot datos DIPSANET-96 Diferenciales y normalizados.

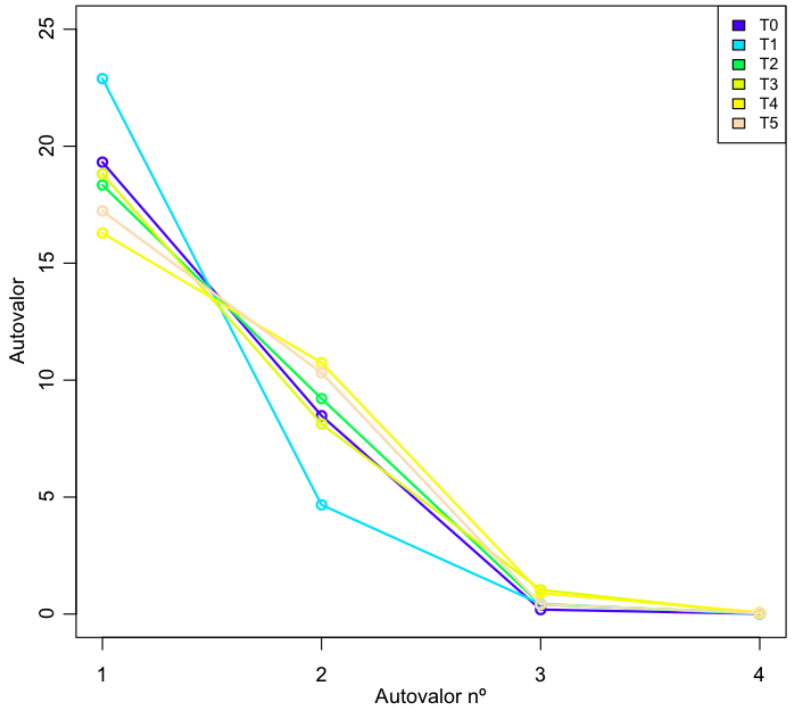
5.3.3.3. Matrices incrementales para cada intervalo temporal

El último análisis que nos resta por abordar es el relativo a las serie de matrices incrementales, esto es, en las que los contadores se "reinician" a cero al principio de cada intervalo temporal considerado.

En la página siguiente se muestran los resultados gráficos obtenidos.

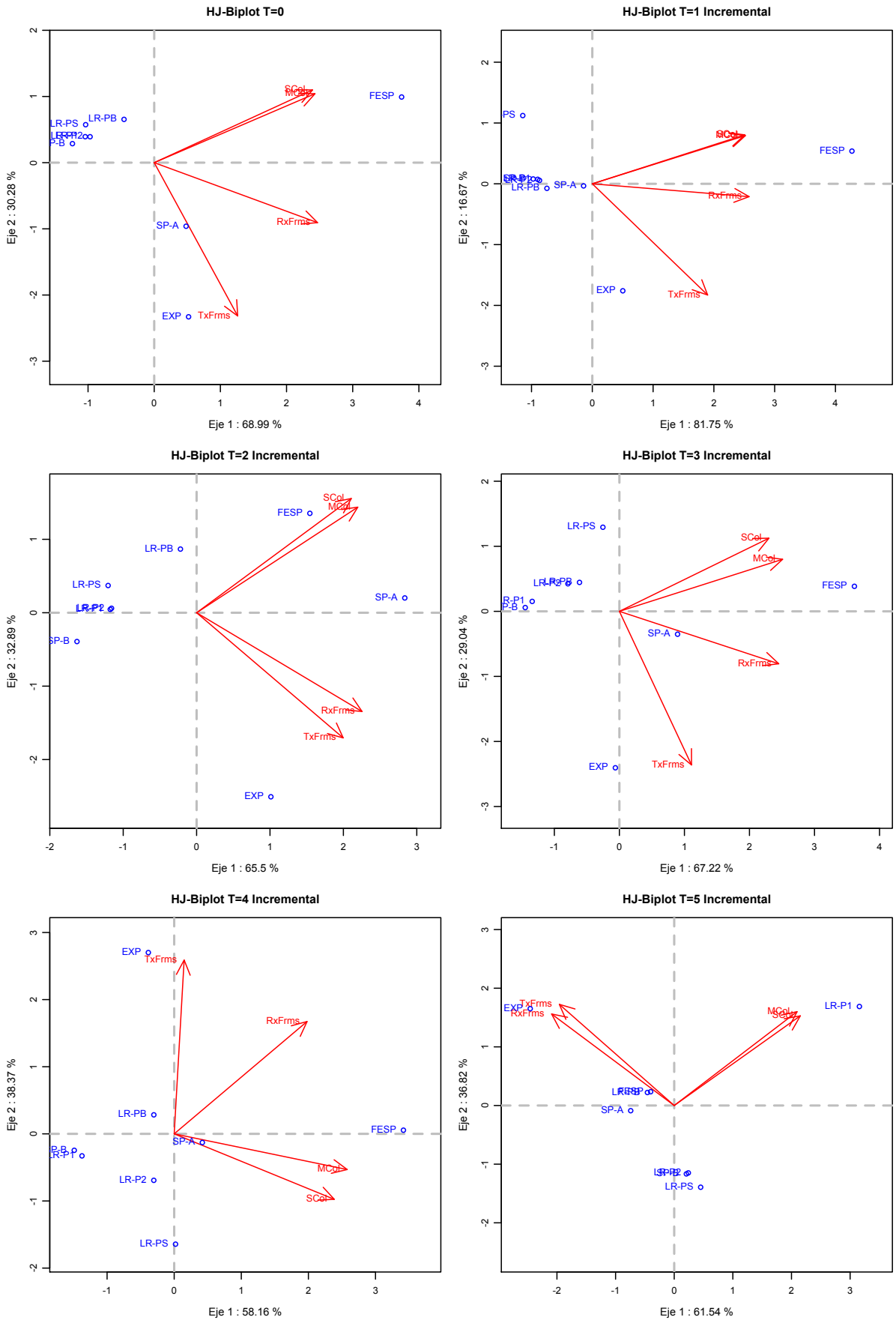
Previamente revisaremos los resultados de estos análisis. La varianza absorbida por el plano es en este caso superior al 96% en todos los casos (96,5% para el intervalo T=4), aproximadamente un 2% inferior a los estudios precedentes.

Es evidente que los autovalores presentan en este caso, como era de esperar, una mayor variabilidad que en los dos casos anteriores.



Screeplots superpuestos de las matrices DIPSANET-96 incremental y normalizadas

Por simple inspección de las representaciones HJ-Biplot obtenidas se concluye que en los cuatro primeros intervalos temporales existe una **moderada estabilidad** en las posiciones obtenidas por los marcadores, pero sin duda menor que en los anteriores análisis. Para el intervalo T=4 se observa un cambio del patrón que, a primera vista, parece relevante. No obstante un análisis más detallado permite concluir que el cambio solo se debe a que en el eje 2 se ha cambiado el signo de los marcadores, con lo que más adelante podemos girarlo para deshacer el artificio del análisis.



Secuencia temporal HJ-Biplot datos DIPSANET-96 Incrementales y normalizados.

Por lo que respecta a las CRFE, el caso T=4 es el que presenta mayor dispersión

Fila	Eje 1	Eje 2	Acumulada
EXP	19.9	978.0	997.9
LR-P2	132.0	685.5	817.5
LR-P1	916.5	52.3	968.8
LR-PB	395.1	344.5	739.6
LR-PS	0.1	970.7	970.8
SP-A	310.8	29.4	340.2
SP-B	910.6	24.7	935.3
FEsp	992.0	0.3	992.2

Columna	Eje 1	Eje 2	Acumulada
Single-Col	813.3	135.9	949.2
Multi-Col	949.8	40.4	990.2
Rcv-Frms	560.2	400.2	960.4
Xmit-Frms	3.1	958.1	961.3

No obstante, de nuevo, sólo un número reducido de marcadores (8) presentan calidades de representación por debajo del 90%. Eso sí, en este caso hay algunos marcadores con calidades bajas (SP-A, T=1, CRFE 20% ; SP-A, T=4, CRFE=34%).

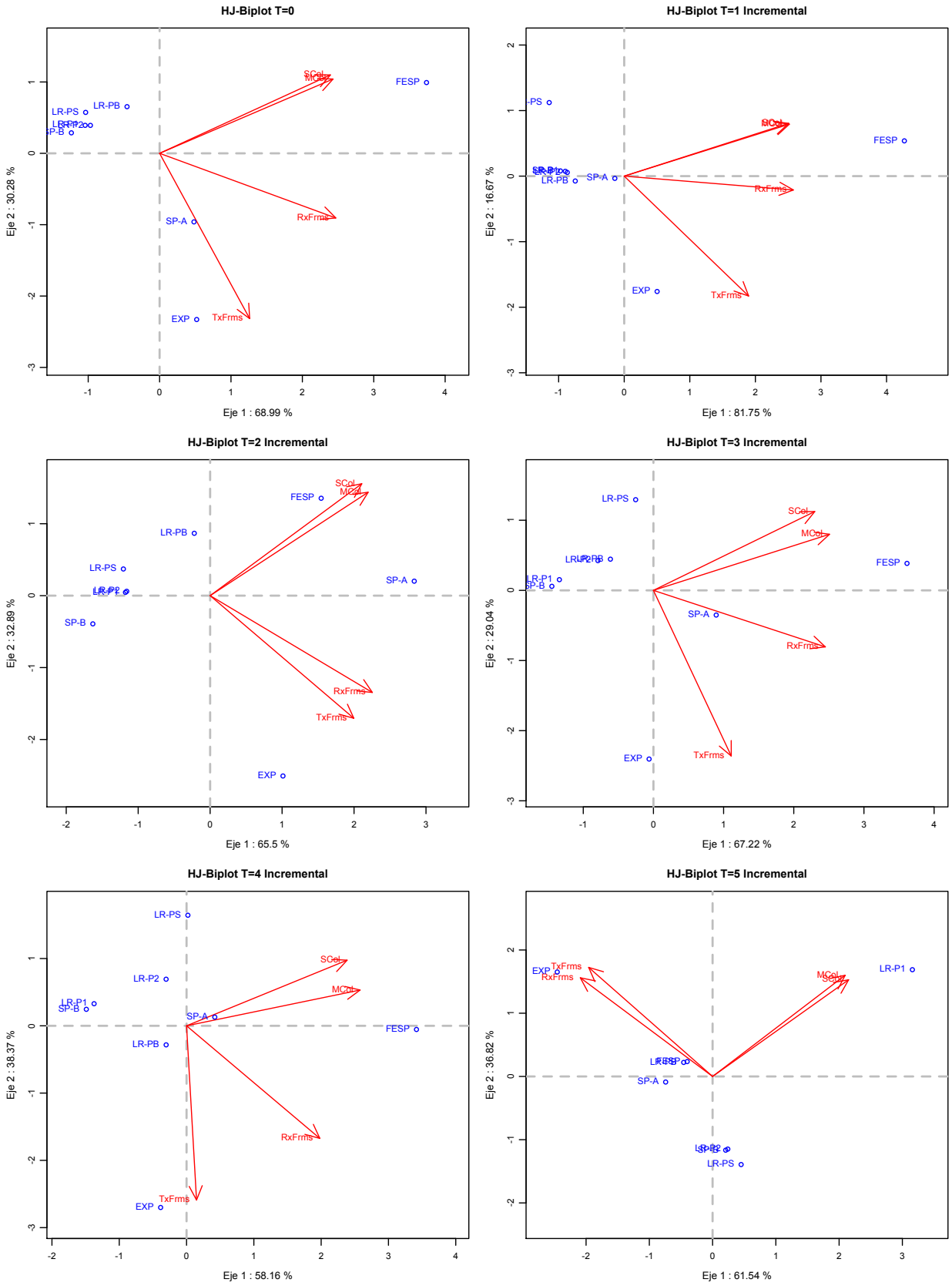
Los resultados se presentan ya de forma comparable en los gráficos de la siguiente página, habiendo modificando la representación para el intervalo T=4 “reflejando” los marcadores obtenidos sobre eje 2.

Se observa ahora que para los cinco primeros intervalos temporales continúa existiendo una **moderada estabilidad** de los marcadores fila y columna, pero que para el quinto intervalo temporal considerado, la representación obtenida es **claramente distinta**.

Esto nos permite concluir visualmente que en dicho intervalo temporal algo **anómalo** ha sucedido.

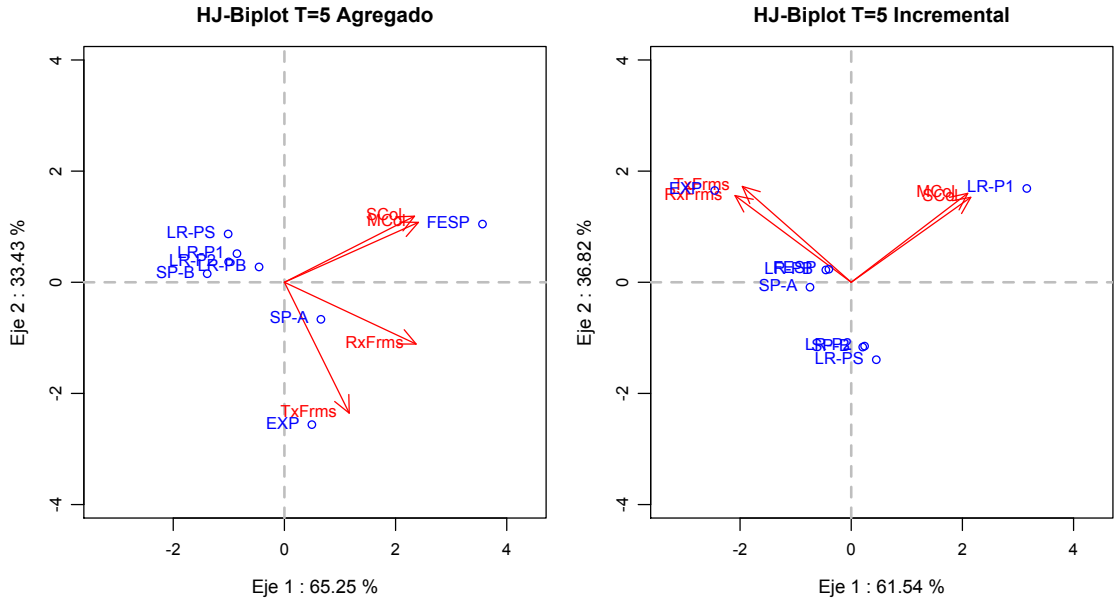
Así pues, la aplicación del HJ-Biplot nos permite **la detección de una anomalía** en la red que tuvo lugar entre los días 2 y 7 de enero.

Pero además, el HJ-Biplot nos permitirá la **diagnosis de la anomalía** que tuvo lugar.



Secuencia temporal HJ-Biplot datos DIPSANET-96 Incrementales y normalizados (T=4 eje y invertido)

Hasta ahora solo hemos utilizado los resultados del HJ-Biplot para compararlos entre sí, en una suerte de análisis interestructura, sin haber entrado a su estudio individual. Retomemos los gráficos obtenidos de la aplicación del HJ-Biplot para el intervalo T=5 en sus versiones agregadas e incremental:



HJ-Biplot datos DIPSANET-96 Agregado y Diferencial, para el intervalo temporal T=5

Utilizaremos el HJ-Biplot agregado como perfil basal con el que comparar la situación anómala.

Las absorciones de inercia en el primer plano factorial son en ambos casos muy elevadas, como ya se ha indicado previamente. Las calidades de representación de los distintos elementos individuales (CRFE) son, son como ya hemos dicho, en general buenas para los segmentos de red y muy buenas para las variables.

Si comparamos ambas representaciones Biplot y analizamos la situación de cada marcador individualmente observamos que, en realidad, no hay tanto cambio cualitativo: la práctica totalidad de los marcadores mantienen las posiciones relativas entre ambos casos. El evidente el cambio brusco experimentado por el marcador correspondiente al segmento LR-P1 que se sitúa como prevalente en las variables de Colisiones simples y Colisiones múltiples. Esto podría haber sido debido a un aumento busco del tráfico cursado en el segmento (ya de por sí, un tipo de anomalía, como vimos), pero no obstante el segmento no aparece como prevalente en las variables correspondientes a tráfico cursado (TxFrms y RxFrms), con lo que concluimos que la anomalía se debe a un funcionamiento anormal que previsiblemente ha ocasionado una tormenta de difusión en el segmento, el equipo transmite tramas no relacionadas a tráfico efectivo, colapsando el segmento por saturación de tráfico. Esto es también

deducible a partir del comportamiento de los marcadores de colisiones (simples y múltiples) y tramas (transmitidas y recibida), cuando se produce el evento, los ángulos que forman ambos grupos (que aparecen entre sí antes correlacionadas) forman un ángulo recto, indicando la desaparición de la relación intrínseca existente entre ambos.

5.3.4. APLICACIÓN DEL ANÁLISIS STATIS

Hemos visto hasta ahora como el método HJ-Biplot nos ha permitido la detección y diagnosis de una anomalía en una red Ethernet real. El procedimiento de detección de la anomalía ha requerido la comparación "*de visu*" de varios representaciones gráficas entre sí.

En este nuevo caso vamos a aplicar el método STATIS Dual para intentar que la comparación entre las matrices correspondientes a los distintos intervalos temporales se realice de un modo más sencillo.

En primer lugar justificaremos el motivo de la utilización del STATIS "Dual": El método STATIS permite el análisis de tablas en la que las variables puedan ser distintas en cada tabla temporal. En cambio el STATIS Dual permite el análisis de tablas en las que las unidades taxonómicas (individuos, en nuestro caso segmentos) puedan variar entre tablas. En el método Triádico, variables y unidades taxonómicas son estables en las diferentes situaciones. Realmente podríamos utilizar este último método, ya que en nuestro "ejemplo" se cumple dicha condición. No obstante, en el caso más general, lo más probable es que, una vez establecidas las variables que serán analizadas, lo que sí puedan variar entre intervalos temporales, sean los segmentos bajo estudio. Esto podría ocasionarse por la conexión de nuevos segmentos (o la eliminación de obsoletos), o simplemente por situaciones transitorias (averías, vacaciones del personal, mudanzas, etc.). Así pues, en nuestra opinión el análisis más apropiado al caso bajo estudio sería el STATIS Dual.

Para realizar el análisis utilizaremos el software libre R [272] y la librería ADE4 [273]. Los pasos a ejecutar son:

Paso 1: Cargar la matriz de tres vías con los datos, para ello se yuxtaponen las siete tablas original trasponiéndolas individualmente, manteniendo comunes las 4 filas de las variables. Previamente a la trasposición/yuxtaposición normalizamos individualmente las matrices por columnas (variables) para que tengan media nula y desviación típica unidad. Este paso es necesario dado que las variables analizadas

son conceptualmente muy distintas (tráfico cursado y colisiones) aunque exista una relación entre ambos grupos [271].

Paso 2: Construir un *data frame* como contenedor de la matriz (primer paso para obtener el tipo de objeto requerido como entrada de la función STATIS).

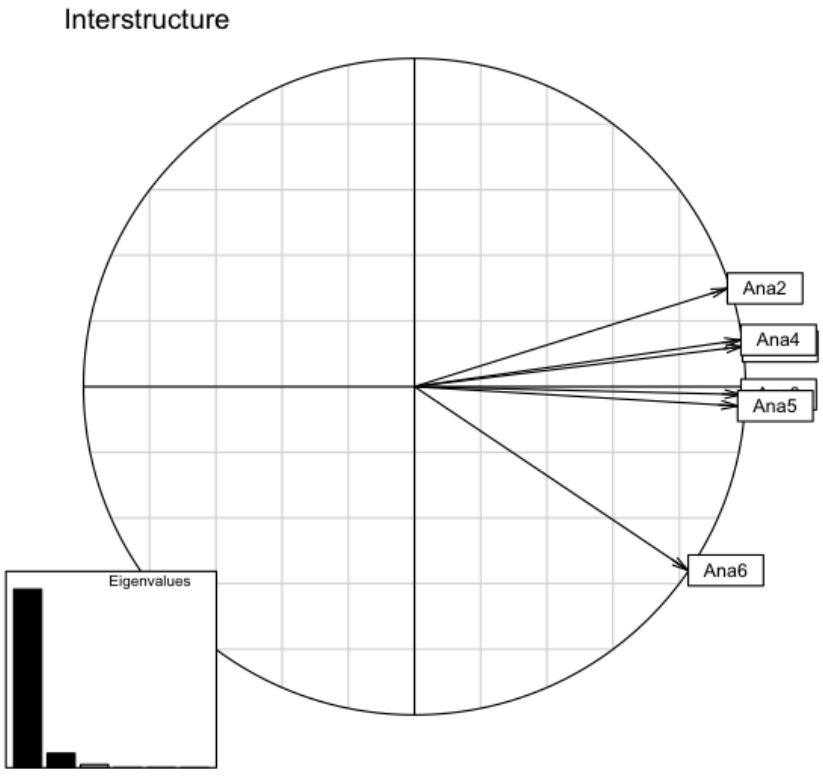
Paso 3: Construir una *K-Table* (*ktab*), que es el objeto que precisa la función STATIS sobre el que operar.

Paso 4: Aplicación del método STATIS (en nuestro caso Dual por la trasposición).

Analicemos los resultados obtenidos tras la realización de estos pasos.

En primer lugar la interestructura obtenida es claramente de primer eje (99% de inercia absorbida) lo que valida la utilización del método STATIS Dual.

Veamos la representación de la interestructura



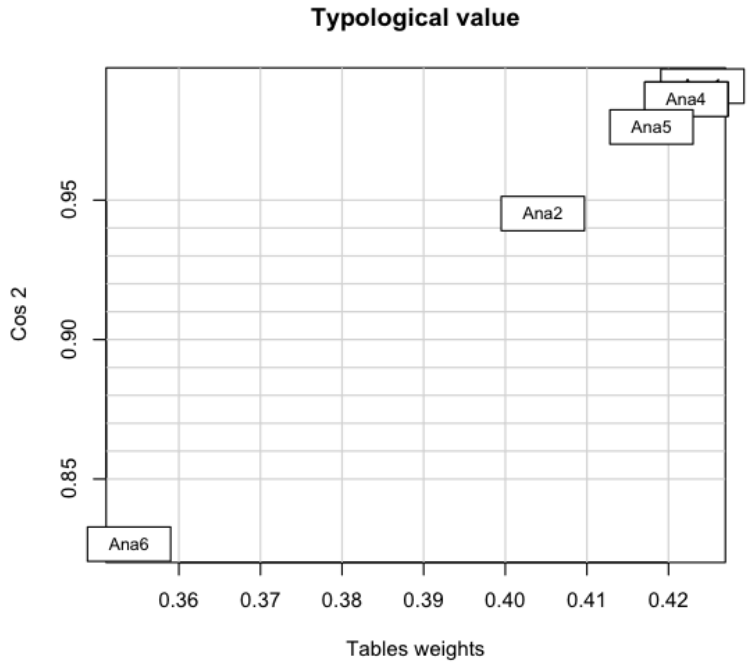
Resultado de la interestructura.

Se observa que la tabla sexta (Ana6) presenta un comportamiento diferente a las restantes (ángulo muy alejado del resto). Si vemos la tabla de los coeficientes vectoriales RV ya es detectable un comportamiento “diferente” de la matriz T=5.

	Ana1	Ana2	Ana3	Ana4	Ana5	Ana6
Ana1	1.0000000					
Ana2	0.9676974	1.0000000				
Ana3	0.9674825	0.9447765	1.0000000			
Ana4	0.9968629	0.9643256	0.9563599	1.0000000		
Ana5	0.9683350	0.8781745	0.9313228	0.9714739	1.0000000	
Ana6	0.7485914	0.6212458	0.8370112	0.7313704	0.8286632	1.0000000

Tabla de coeficientes RV

Este fenómeno es más evidente aún, si cabe, si representamos los pesos asignados a las diferentes tablas para elaborar el "compromiso" y el coeficiente Cos^2.



Resultado de las ponderaciones de las diferentes tablas.

5.3.5. RESUMEN

Hemos visto en este supuesto práctico, en el que se han aplicado los métodos HJ-Biplot y STATIS Dual a mediciones reales de tráfico de dato en una red Ethernet, como ambos métodos permiten la detección *de visu* de una situación anómala en la red, una tormenta de difusión en uno de los segmentos bajo estudio. El método HJ-Biplot parece posibilitar su utilización para el modelado de la red, dada la estabilidad de los marcadores agregados de la representación, que además tiene una calidad de representación notable. Igualmente, la representación visual que se obtiene con el HJ-Biplot es informativa del comportamiento de la red, y de ella se pueden extraer conclusiones compatibles con el modo de funcionamiento del protocolo Ethernet.

5.4. DETECCIÓN DE ANOMALÍAS EN LA RED INTERNET

5.4.1. INTRODUCCIÓN

El grupo de Tecnología y Cibersistemas del Massachusetts Institute of Technology (MIT) Lincoln Laboratory, con el patrocinio de la Agencia de Proyectos Avanzados de Investigación de Defensa (DARPA ITO) y del Laboratorio de Investigación de la Fuerza Aérea (AFRL/SNHS) recopiló y distribuyó en 1998 el primer corpus normalizado para la evaluación de sistemas de detección de intrusos en redes informáticas. Como resultado de ello se coordinaron las primeras evaluaciones formales, repetibles y significativas de sistemas de detección de intrusos. Estas evaluaciones incluían medidas de la probabilidad de detección y de falsa alarma para cada sistema sometido a evaluación [274]. Posteriormente se distribuyeron dos juegos de datos más, en los años 1999 y 2000 [275].

El objetivo del programa era contribuir al campo de la investigación en detección de intrusiones, proporcionando una guía para los investigadores y un procedimiento de calibración objetiva del estado de las diferentes técnicas en aquellas fechas. Ambas cuestiones eran y son de indudable interés para los investigadores. La evaluación fue diseñada para ser simple, poniendo el foco en los aspectos tecnológicos y promoviendo la mayor participación posible, eliminando para ello todos los problemas de seguridad y privacidad y proveyendo los tipos de datos utilizados habitualmente por la mayoría de los sistemas de detección de intrusos [274].

5.4.2. DESCRIPCIÓN DE LOS DATOS

El kit de evaluación proporcionado por DARPA consiste en dos juegos diferentes, uno para evaluación fuera de línea y otro en línea. Los sistemas de detección son probados utilizando la evaluación fuera de línea con tráfico de red y registros de auditoría recogidos en una red simulada. Los sistemas de detección de intrusos procesan esta información e intentan identificar los diferentes tipos de ataques dentro de la maraña de tráfico legítimo. Posteriormente los sistemas serían remitidos al AFRL para su conexión a la red de pruebas y verificación de su funcionamiento en tiempo real [274]. En esta demostración operaremos con el juego de datos fuera de línea.

Concretamente en nuestro ejemplo se utilizará el juego de datos denominado “LLDOS 2.0.2 – Escenario dos”. Este juego de datos incluye un ataque de negación de servicio distribuido llevado a cabo por un atacante muy silencioso. El ataque se lleva a cabo en 5 fases típicas:

1. Se prueba el equipo que será comprometido.
2. Se penetra en el equipo aprovechando una vulnerabilidad encontrada.
3. Se instala en el equipo vulnerable el troyano del agente *mstream* DDoS.
4. Se comprueban las posibles vulnerabilidades de otros equipos de la red local para intentar instalar el agente y así convertirlos en atacantes.
5. Finalmente, se lanza la orden de ataque contra un equipo exterior desde el equipo comprometido que es llevada cabo por los atacantes.

El ataque incluido en el juego de datos utiliza la herramienta mstream DDoS [276], [277], que está ya obsoleta y se encarga de generar y enviar los paquetes ACK que forman parte del ataque de negación de servicio distribuido (DDoS).

Nuestro objetivo será detectar el ataque en la fase 5, esto es, intentaremos detectar la anomalía de volumen producida por el ataque. El ataque de mstream DDoS consiste en el envío masivo de peticiones de conexión a múltiples puertos de la víctima. Todos los paquetes tienen falseada la dirección de origen (*IP spoofing* [278], [279]), para ocultar el origen y así dificultar la aplicación de contramedidas (por ejemplo, bloquear/”oscurecer” un rango de direcciones IP es imposible si los paquete aparentan provenir de miles de direcciones diferentes).

Los registros abarcan aproximadamente 1h 45 minutos en total, desde las 14:45 horas a las 16:28 del 16 de abril de 2000 (19:45 a 21:28) comenzando el ataque a las 16:06:15 (21:06:15) con una duración de 5 segundos [280].

5.4.3. ANÁLISIS DE LOS DATOS

Como es habitual en este tipo de análisis, el primer paso a realizar es un preprocesado de los datos capturados para adecuarlos al algoritmo de análisis [226], ya que por regla general solo una parte de esa información aportada suele ser elegible para el análisis, bien por su tipología (inadecuada para el tipo de análisis a realizar) bien porque no sea representativa (ineficaz para el objetivo planteado).

A continuación describiremos someramente las operaciones de preprocesado realizadas sobre los datos, antes de que sean sometidos al análisis propiamente dicho.

0. Obtención de los datos

En nuestro caso, la captura de los datos consiste en algo tan simple como la descarga del juego de datos utilizado y que está contenido en el fichero LLS_DDOS_2.0.2-inside.dump disponible en la dirección

http://129.55.10.6/IST/ideval/data/2000/LLS_DDOS_2.0.2/data_and_labeling/tcpdump_inside/LLS_DDOS_2.0.2-inside.dump.gz

En un caso real, la captura de la información es, como vimos en la literatura seleccionada, un aspecto de gran importancia. En nuestro ejemplo la información ya está capturada y volcada en un archivo binario.

1. Generación del fichero traza para interior de la red:

El archivo de información capturada de la red simulada contiene información binaria, que no es, en nuestro caso, directamente tratable. Hemos de generar un fichero de traza, en modo texto, que podamos procesar por nuestros algoritmos sin problemas. Para ello utilizaremos la aplicación tcpdump [281]:

```
tcpdump –n –r LLS_DDOS_2.0.2-inside.dump ip > inside.out
```

La opción –n evita que el programa efectúe una búsqueda inversa de las direcciones IP en el DNS para sustituirlas por el nombre del equipo. Esto, evidentemente, proporciona una mayor legibilidad a la información obtenida, pero ralentiza el proceso. El cambio entre IP->nombres se realizará posteriormente operando directamente sobre la información. La opción –r indica al programa tcpdump que la información se encuentra en el archivo binario indicado (por defecto captura directamente la información de la red de área local a la que el equipo está conectado). Finalmente la opción IP indica que solo nos interesan los paquetes del protocolo IP.

El resultado es un archivo de texto ya “comprensible” de la siguiente forma:

```
19:45:04.705889 IP 172.16.112.50.23 > 172.16.113.168.1266: P 6:7(1) ack 6 win 8760 (DF)
19:45:04.725383 IP 172.16.113.168.1266 > 172.16.112.50.23: . ack 7 win 32120 (DF)
19:45:04.833935 IP 194.7.248.153.63281 > 172.16.112.194.23: P 11:12(1) ack 12 win 33580 (DF)
19:45:04.834671 IP 172.16.112.194.23 > 194.7.248.153.63281: P 12:13(1) ack 12 win 32736 (DF)
19:45:04.853791 IP 194.7.248.153.63281 > 172.16.112.194.23: . ack 13 win 33580 (DF)
19:45:04.991179 IP 172.16.112.149.1472 > 172.16.115.20.53:  45060+ A? lambda.orange.com. (35)
```

Ejemplo de registro obtenido a través de la aplicación tcpdump.

2. Carga del archivo en la base de datos

Dado que el fichero es muy extenso (279.890 líneas/registros) es conveniente utilizar una base de datos relacional para preprocesar los datos, ya que es más eficiente que operar directamente sobre el fichero de texto, en primer lugar por la posibilidad de utilizar un lenguaje de alto nivel para efectuar consultas. Creamos una base de datos con los siguientes campos:

Nombre del campo	Observaciones	Formato
Time	Marca temporal del paquete	Texto
IP_From	Dirección IP de origen	Texto
Port_From	Puerto de origen	Número
IP_To	Dirección IP de destino	Texto
Port_To	Puerto de destino	Número
Flag	Flags activados en el paquete	Texto
Ack	Ack activado en el paquete	Texto

En dicha base de datos importamos los registros del archivo de texto y los mapeamos en los campos correspondientes.

La marca temporal de inicio del registro (primera entrada de la base de datos) corresponde con la hora 19:45:01.903961 y la marca final (última entrada de la base de datos) es la 21:27:48.357079. Según la información facilitada el ataque DDoS comienza a las 21:06:15.757496 (registro número 168.148) y dura 5 segundos.

3. Construcción del "cubo" de datos

Utilizando el lenguaje de interrogación de bases de datos relacionales SQL recopilamos la información necesaria para construir las diferentes matrices para cada uno de los siete intervalos de 15 minutos en que hemos divido el tiempo entre las 19:45 y 21:30. Para cada intervalo "contamos" mediante consultas SQL sucesivas los paquetes con el flag ACK, PUSH, RST, FIN, SYN activado (como en Sun *et al* [237]) que son dirigidos a cada uno de los equipos objeto de estudio (solo se han considerado los 15 equipos con más tráfico dentro del archivo, además de la propia "víctima", de un total de 34 equipos incluidos en el fichero de traza):

IP	name		IP	name
131.84.1.31	www.af.mil		172.16.113.168	finch.eyrie.af.mil
172.16.112.100	hume.eyrie.af.mil		172.16.113.169	swan.eyrie.af.mil
172.16.112.149	eagle.eyrie.af.mil		172.16.113.204	goose.eyrie.af.mil
172.16.112.194	falcon.eyrie.af.mil		172.16.113.207	pigeon.eyrie.af.mil
172.16.112.207	robin.eyrie.af.mil		172.16.113.50	zeno.eyrie.af.mil
172.16.112.50	pascal.eyrie.af.mil		172.16.113.84	duck.eyrie.af.mil
172.16.113.105	swallow.eyrie.af.mil		172.16.114.50	marx.eyrie.af.mil
172.16.113.148	crow.eyrie.af.mil		172.16.115.20	mill.eyrie.af.mil

Lista de equipos objeto de estudio

Son necesarias 35 consultas a la base de datos para obtener las cinco columnas de variables correspondientes a las siete tablas que serán objeto del posterior análisis estadístico.

A continuación se muestra un ejemplo de una de las consultas SQL utilizadas y una tabla de datos resultado.

```
SELECT DistinctIPTo.IP, IP2Name.name, Count(Datos.IP_To) AS ACK FROM (Datos INNER JOIN DistinctIPTo ON
Datos.IP_To = DistinctIPTo.IP) INNER JOIN IP2Name ON DistinctIPTo.IP = IP2Name.IP WHERE
(((Datos.Ack)="ack")) GROUP BY DistinctIPTo.IP, IP2Name.name ORDER BY DistinctIPTo.IP;
```

Consulta SQL para "contar" los ACK dirigidos a los equipos seleccionados y mostrar sus nombres

IP	name	ACK	PSH	RST	FIN	SYN
131.84.1.31	www.af.mil	44	7	0	7	7
172.16.112.100	hume.eyrie.af.mil	329	143	0	54	55
172.16.112.149	eagle.eyrie.af.mil	56	35	0	6	6
172.16.112.194	falcon.eyrie.af.mil	4458	2172	0	7	6
172.16.112.207	robin.eyrie.af.mil	400	208	0	6	7
172.16.112.50	pascal.eyrie.af.mil	3871	1931	0	4	4
172.16.113.105	swallow.eyrie.af.mil	350	303	0	10	10
172.16.113.148	crow.eyrie.af.mil	199	169	0	5	6
172.16.113.168	finch.eyrie.af.mil	1691	1622	0	12	11
172.16.113.169	swan.eyrie.af.mil	126	86	0	11	11
172.16.113.204	goose.eyrie.af.mil	193	159	0	6	6
172.16.113.207	pigeon.eyrie.af.mil	1367	524	0	187	187
172.16.113.50	zeno.eyrie.af.mil	440	218	0	1	1
172.16.113.84	duck.eyrie.af.mil	12	9	0	1	1
172.16.114.50	marx.eyrie.af.mil	444	223	0	3	3
172.16.115.20	mill.eyrie.af.mil	367	182	0	1	2

Tabla de datos obtenida para el intervalo de las 19:45 a las 20:00 horas

Con este paso se habría concluido el preproceso de la información "capturada" de la red de datos, con el resultado de una tabla de tres vías conteniendo información sobre el número de paquetes con cada tipo de flag activado dirigido a cada uno de los equipos bajo estudio y para cada subintervalo temporal en el que ha sido dividido el intervalo de análisis. Con ello podremos pasar a realizar el análisis propiamente dicho.

4. Análisis STATIS

Para realizar el análisis utilizaremos el software libre R [272] y la librería ADE4 [273]. Los pasos a ejecutar son:

Paso 1: Cargar la matriz de tres vías con los datos, para ello se yuxtaponen las siete tablas, manteniendo comunes las 4 filas de los servidores.

```
X=matrix(c(44, 7, 0, 7, 7, 0, 0, 0, … , 614, 0, 1, 2, 892, 435, 0, 2, 2 ),ncol=35,nrow=16, byrow=TRUE)
```

Observamos que la variable RST no aporta información, al menos en nuestro experimento, y puede constituir un problema la presentar un número muy elevado de ceros, así que la eliminaremos del análisis.

```
X= cbind(X[,1:2],X[,4:7],X[,9:12],X[,14:17],X[,19:22],X[,24:27],X[,29:32],X[,34:35])
```

Colocamos los nombres de filas y columnas a la matriz.

```
dimnames(X)<-
list(c('www','hume','eagle','falcon','robin','pascal','swallow','crow','finch','swan','goose','pigeon','zeno','duck','marx','mill')
,c('ack','psh','fin','syn','ack','psh','fin','syn','ack','psh','fin','syn','ack','psh','fin','syn','ack','psh','fin','syn','ack','psh','fin','syn','ack','psh','fin','syn'))
```

Vamos a trasponer cada matriz temporal para realizar un STATIS Dual. Dado que los servidores bajo estudio pueden variar de matriz a matriz, se considera conveniente que el análisis a realizar sea un STATIS Dual.

```
XT=cbind(t(X[,1:4]),t(X[,5:8]),t(X[,9:12]),t(X[,13:16]),t(X[,17:20]),t(X[,21:24]),t(X[,25:28]))
```

Paso 2: Construir un *data frame* como contenedor de la matriz (primer paso para obtener el tipo de objeto requerido como entrada de la función STATIS).

```
Xdf=data.frame(XT)
```

Paso 3: Construir una *K-Table* (*ktab*), que es el objeto que precisa la función STATIS sobre el que operar.

```
Xtb=ktab.data.frame(Xdf,c(16,16,16,16,16,16,16))
```

Paso 4: Aplicamos del método STATIS

```
statis1 <- statis(Xtb,scann=FALSE)
```

5. Resultados obtenidos

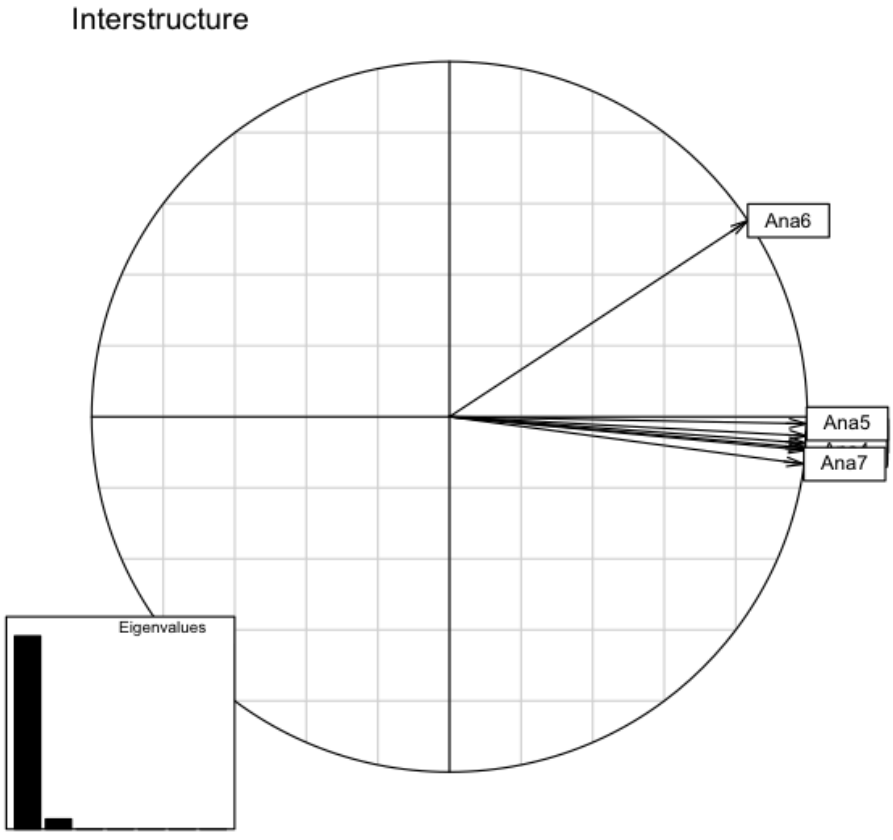
El primer resultado a analizar es lógicamente la interestructura detectada entre las tablas objeto de estudio. El *scree-plot* (ver página siguiente) nos muestra una clara estructura prácticamente unidimensional, con un primer eje absorbiendo más del 99% de la inercia.

El coeficiente de correlación vectorial, como se ha indicado, nos proporciona información sobre la similaridad entre las matrices, cuanto más próximo a 1, las matrices serán más parecidas.

RV Coeff.	Ana1	Ana2	Ana3	Ana4	Ana5	*Ana6*	Ana7
Ana1	1.0000000						
Ana2	0.9991486	1.0000000					
Ana3	0.9998161	0.9991167	1.0000000				
Ana4	0.9997563	0.9984228	0.9998758	1.0000000			
Ana5	0.9956224	0.9985951	0.9956680	0.9942119	1.0000000		
Ana6	*0.7901372*	*0.8022367*	*0.7821257*	*0.7780079*	*0.8206246*	*1.0000000*	
Ana7	0.9978353	0.9953171	0.9984682	0.9990348	0.9894652	*0.7524351*	1

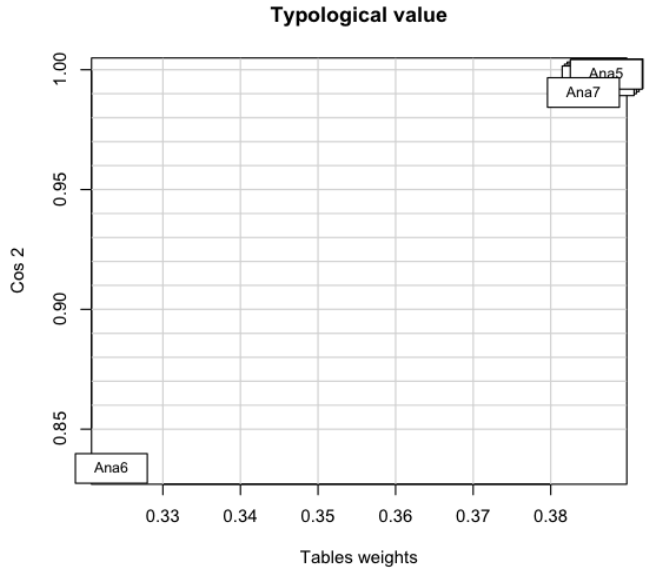
La simple inspección de la matriz de coeficientes RV ya nos pone de manifiesto que la matriz correspondiente al intervalo temporal número 6 (etiquetado como Ana6) es la más "diferente" a las restantes.

Este diferente comportamiento queda también en evidencia si representamos la interestructura en la que se observa de nuevo claramente que el mencionado intervalo número 6 es "anómalo". Dicho intervalo corresponde al tráfico capturado entre las 21:00 y 21:15 horas.



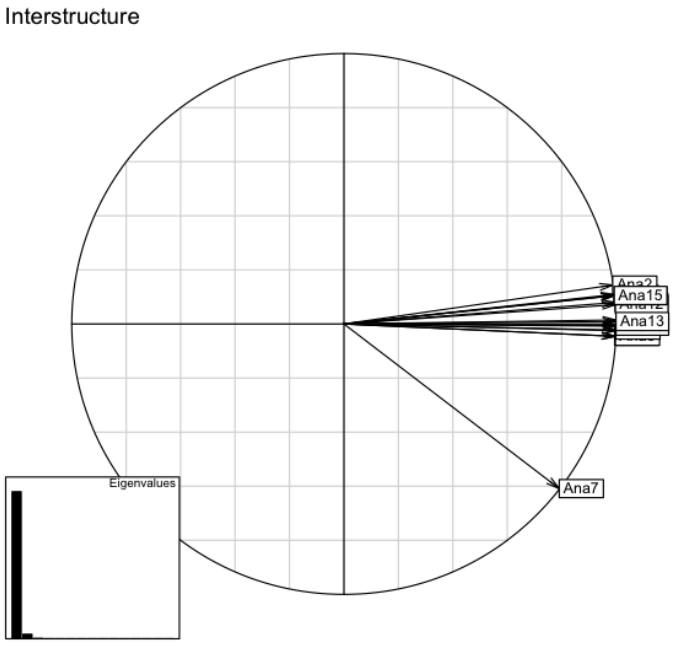
Representación de la interestructura entre las 19:45 y 21:30 en intervalos de 15 minutos.

Como en el ejemplo anterior sobre Ethernet, si representamos Cos^2-vs.-Ponderaciones de las tablas, esta separación queda también de manifiesto:



Representación de las ponderaciones de la tablas.

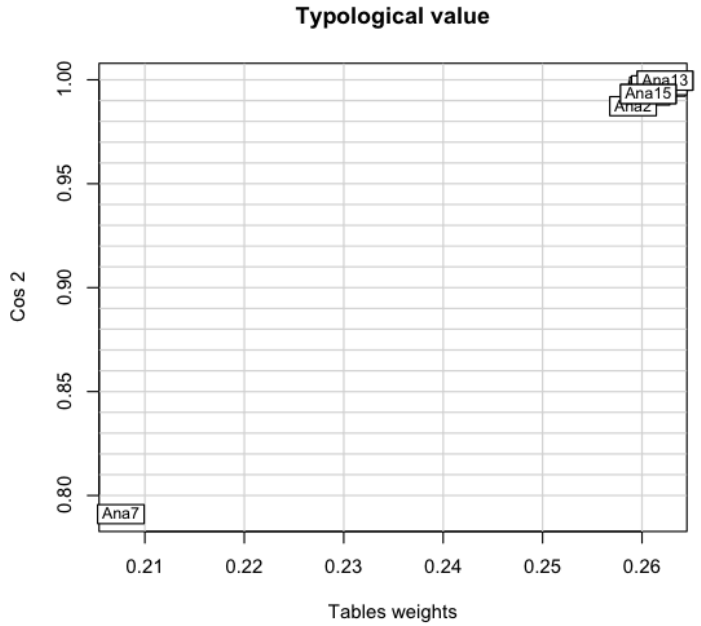
Con objeto de "afinar" más el momento en el que se produce la anomalía (en este caso el ataque de negación de servicio) repetimos todo el proceso de análisis exclusivamente para las captura entre las 21:00 y 21:15, en este caso con intervalos de 1 minuto.



Representación de la interestructura entre las 21:00 y 21:15 en intervalos de 1 minuto

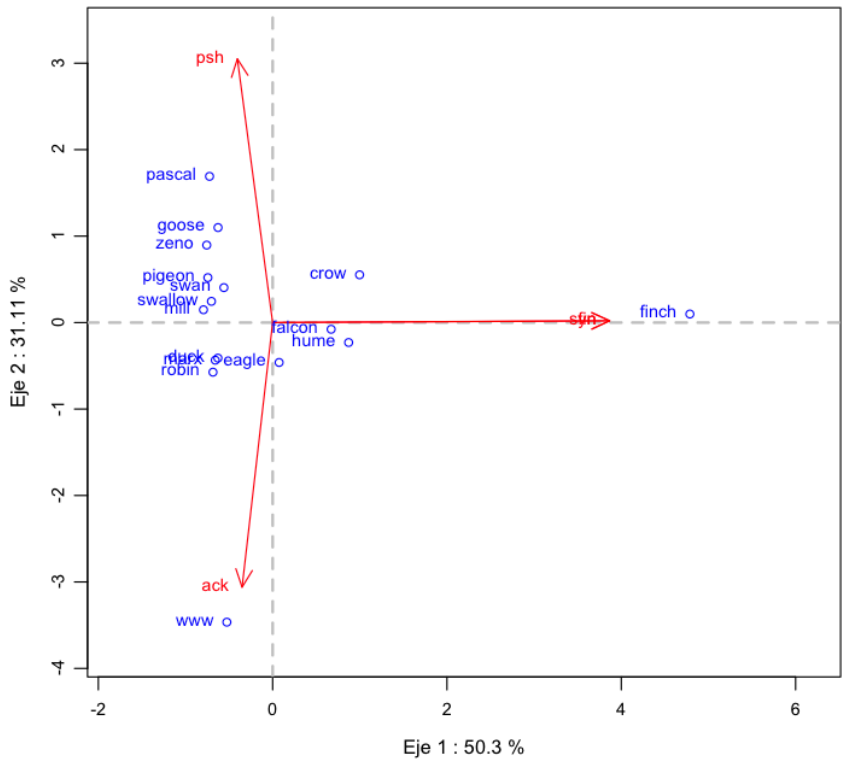
En este caso, el comportamiento "diferente" tiene lugar en el intervalo séptimo (Ana7) que corresponde a las capturas con etiqueta temporal entre las 21:06 y 21:07 horas. Recordemos que, según la información facilitada con el juego de datos, el ataque comienza a las 21:06:15 horas.

Igualmente en la representación cos^2 –vs.- pesos de las tablas es evidente la situación:



Representación de las ponderaciones de las tablas.

Según las denominaciones habituales ([41],[266] entre otros) se habría conseguido la "detección" de la anomalía, pero aún estaría pendiente la "diagnosis", el averiguar lo sucedido en el intervalo en el que se ha detectado el ataque. Para ello recurriremos al análisis intraestructura, que en como en el caso anterior, realizaremos mediante un HJ-Biplot de la matriz correspondiente al intervalo de ataque:



Representación HJ-Biplot del intervalo temporal entre las 21:00 y 21:15

En la que destaca el “emparejamiento” entre la variable “ack” y el equipo “www”, correspondiente con el tipo de paquete enviado por el agente para provocar la “negación de servicio” y el equipo “víctima del ataque”. Las inercias absorbidas por los dos primeros ejes suman más del 80%, lo que nos permite afirmar que estamos ante una estructura de plano, con una buena calidad de representación.

Analicemos brevemente la CRFE de las variables y hosts en esta representación:

CRFE	1	2	3	4
ack	8.2	624.3	367.4	0
psh	10.9	620.1	368.9	0
fin	996.5	0.0	3.3	0
syn	996.1	0.0	3.7	0

CRFE	1	2	3	4
www	17.9	776.9	205.2	0
hume	677.79	47.3	274.9	0
eagle	5.3	203.6	791.1	0
falcon	735.4	9.8	254.3	0
robin	229.7	161.4	608.6	0
pascal	89.5	488.2	422.3	0
swallow	822.5	100.5	76.5	0
crow	709.3	217.3	73.2	0
finch	996.8	0.4	2.9	0
swan	655.9	342.5	1.1	0
goose	184.6	567.5	247.9	0
pigeon	666.8	326.1	7.0	0
zeno	323.6	454.0	222.3	0
duck	276.7	118.6	604.7	0
marx	284.9	124.5	590.4	0
mill	869.4	30.3	100.0	0

CRFE HJ-Biplot del intervalo temporal entre las 21:00 y 21:15

Se observa que la estructura resultante es tridimensional, si bien puede representarse en dos dimensiones para todos los marcadores con una calidad de representación en el plano por encima de 600, en 15 marcadores de los 22.

5.4.4. RESUMEN

Con este ejemplo hemos demostrado la capacidad del método STATIS para la detección de anomalías de volumen, mediante el análisis de la interestructura, y posteriormente del HJ-Biplot para la diagnosis especifica de la anomalía detectada, mediante el análisis detallado de la intraestructura de la matriz de datos correspondiente al intervalo anómalo. De nuevo la representación HJ-Biplot nos ha permitido, además, extraer conclusiones visualmente de las posiciones de los marcadores que están relacionadas con el comportamiento real de la red.

6. EL HJ-BIPLOT COMO ALTERNATIVA A LA DESCOMPOSICIÓN DE KARHUNEN-LÒEVE (EKL)

6. EL HJ-BIPLOT COMO ALTERNATIVA A LA DESCOMPOSICIÓN DE KARHUNEN-LÒEVE (EKL)

6.1. INTRODUCCIÓN

El estudio de series numéricas, como resultado de la discretización o muestreo en el tiempo de variables, funciones o señales, ha constituido un campo fundamental de la matemática. Se han desarrollado técnicas para el estudio de series temporales tanto en el dominio del tiempo propiamente dicho, como en el dominio frecuencial. Destacamos entre ellas las transformadas de Laplace y las series y transformadas de Fourier, para el caso continuo, y la transformada Z y transformada discreta de Fourier, para el caso discreto [282]. Todo ello se engloba dentro de un aparato matemático denominado estadística aplicada, en este caso al campo de las telecomunicaciones [283].

Para el caso concreto de la descomposición de series numéricas en forma de una combinación lineal de elementos más “simples”, admitiendo en ello un posible error si se tratase solo de una aproximación, se han efectuado muchas propuestas diferentes, como por ejemplo las enumeradas anteriormente. Algunas otras se han visto ya aplicadas en capítulos anteriores, como por ejemplo la descomposición en *wavelets* [207], [234], y en la literatura se proponen otras muchas, como los modelos autoregresivos (AR *AutoRegressive model*, ARMA *AutoRegressive Moving Average,* ARIMA *AutoRegressive Integrated Moving Average*), entre otras [284].

En nuestra propuesta vamos a valorar la aplicación a la caracterización de series temporales de métodos relacionados con una técnica tan habitual en el análisis multivariante de datos, como es el Análisis de Componentes Principales [243] y en particular los métodos Biplot [157], [160]. La formulación inicial del Análisis de Componentes Principales se sitúa [285] en los estudios de Pearson de 1901 [286] y posteriormente de Hotelling en 1933 [287].

Es evidente que una serie temporal x(t) puede ser contemplada como una realización ξ fija de un proceso estocástico x(t,ξ) [288].

Bouhaddou *et al* propusieron en 1987 [214] la aplicación de una extensión del Análisis de Componentes Principales tradicional (ACP) al estudio de procesos estocásticos. Esta propuesta, indican los autores, ya fue descrita en trabajos previos de otros

autores franceses [289]. El Análisis de Componentes Principales de Procesos estocásticos (PCAP / ACPP) extiende de manera natural el ACP de n variables al análisis de un número infinito de variables. En la práctica los datos recopilados del proceso estocástico están constituidos por *p* observaciones en el tiempo de *n* variables distintas, ordenándose en forma de una matriz $p \times n$. Mientras el ACP clásico trata la matriz como si se estuviese en presencia de variables no relacionadas entre sí, aplicando la misma ponderación a todas ellas, el PCAP, por el contrario, tiene en consideración la estructura completa del fenómeno analizado, tratando las observaciones como si se estuviese en presencia del muestreo de funciones discretizadas en el espacio y, por lo tanto, la matriz a diagonalizar se modifica mediante un factor que tiene en consideración tanto la secuencia temporal entre observaciones como su posición espacial relativa. Ese factor está relacionado con el operador de covarianza (temporal y espacial) del proceso estocástico subyacente. Este método, afirman los autores, es al menos igual y en algunos casos superior al método de Kriging formulado por Matheron en 1971 [290] y que como hemos visto anteriormente ya se ha utilizado en el estudio del tráfico de redes [291]. El ACPP que proponen Bouhaddou *et al* [214] constituye una aproximación no paramétrica, lo que evita las asunciones de estacionalidad e isotropía, siendo necesarias solo unas condiciones básicas en la función de covarianza del proceso estocástico para hacer posible su aplicación.

Hay en la literatura diferentes planteamientos al respecto de la formulación del Análisis de Componentes Principales (ACP) y la conocida como Expansión o Descomposición de Karhunen-Loéve (EKL). Así, para Kirby y Sirovich [285] el Análisis de Componentes Principales es equivalente al conocido como Expansión (o Descomposición) de Karhunen-Loève o Transformada de Hotelling.

Sin embargo, para Dong *et al* [292] el ACP y la EKL se diferencian en que la primera utiliza la matriz de covarianzas y la segunda la matriz de correlaciones para obtener la base vectorial ortonormal. Así, exponen que si se escalan las series originales por la raíz cuadrada del recíproco de sus varianzas, el EKL es equivalente al ACP de la serie así reescalada, ya que cuando la matriz de covarianzas se normaliza por el vector de varianzas, la matriz resultante es la matriz de correlaciones [292].Fernando y Nicholson en 1980 [293] y Gerbrands en 1981 [294] exponen igualmente en sendos artículos las relaciones existentes entre la Expansión de Karhunen-Loève (EKL), la Descomposición en Valores Singulares (DVS) y el Análisis de Componentes Principales (ACP).

Gao *et al* en 2007 [295] también establecen diferencias entre el ACP y la EKL, señalando que el ACP es la descomposición en autovalores y autovectores de la matriz de correlación "o de covarianza" (sic), mientras que la EKL es la descomposición de la serie en un conjunto de funciones ortonormales conocidas previamente. En ese mismo sentido se pronunciaban Biglieri y Yao en 1989 [296], para quienes la DVS de una imagen (entendida ésta como matriz bidimensional) es "conceptualmente similar" a su descomposición de Karhunen-Loève. Sin embargo, indican, hay importantes diferencias en el fondo de ambos conceptos. Si la DVS se define sobre los "datos en bruto" y las imágenes de rango 1 que forman parte de la descomposición están únicamente determinadas por la imagen de partida que es sometida al análisis, en la EKL las imágenes de rango 1 en las que se basa la descomposición están predeterminadas por la matriz de covarianza de los procesos estocásticos que hipotéticamente generan cada posible imagen. Esto es, las imágenes ortogonales de la EKL son conocidas a priori, mientras que aquellas que forman la DVS son desconocidas hasta que la imagen concreta está disponible.

No obstante, de una u otra manera, tal y como ya indicaba Ozeki en 1979 [297], los métodos de EKL y ACP se reducen al mismo problema matemático de obtención de los autovectores-autovalores de una matriz. En algunos casos será importante la conocida limitación que presenta la aplicación de la DVS si en los datos hay elementos faltantes o presencia de *outliers* [254].

Stone y Cutler afirman en 1996 [298] que la aplicación del método de Análisis de Componentes Principales "o Descomposición de Karkhunen-Loève" (KLE) a problemas de modelado de sistemas dinámicos ha tenido mucha atención desde 1991. Ese mismo año 1996, por ejemplo, Graham y Kevrekidis [299] proponen la diagonalización de la matriz de covarianza de los datos yuxtapuestos, tras restarles la media para analizar solo las fluctuaciones. Esta matriz compuesta es utilizada para analizar diversos fenómenos espacio-temporales, como por ejemplo fenómenos meteorológicos. Newman también en 1996 [300] propone igualmente la utilización de la Descomposición de K-L para el modelado como procesos estocásticos de flujos, esto es, de una función del espacio y tiempo u(t,x) que describe la evolución de una particular entidad.

El método expuesto por Stone y Cutler [298] parte de un conjunto de datos $\{\mathbf{x}_i, i=1,\ldots,n\}$ donde $\mathbf{x}_i$ es un vector de m-componentes $\mathbf{x}_i=\{x_{1i},\ldots,x_{mi}\}^T$. Cada vector puede corresponder con un muestreo en el tiempo de una variable que está además

discretizada en una dirección espacial única. Se supone, sin pérdida de generalidad, que $E[\mathbf{x}_i]=0$. Recuerda Newman en su artículo [300] que la media, correlación y covarianza en función del tiempo de un proceso $\mathbf{X}_t$ se definen como:

Media: $\mu(t) = E[\mathbf{X}_t]$

Correlación: $\Re_v(t,s)=E[\mathbf{X}_t\mathbf{X}_s]$

Covarianza: $R_v(t,s)=)=E[(\mathbf{X}_t - \mu(t)\ )\ (\mathbf{X}_s - \mu(s))]$

Por lo que para procesos con media nula se tiene que $\Re_v(t,s) = R_v(t,s)$, esto es, las funciones de covarianza y de correlación son intercambiables.

La EKL / PCA consiste entonces [298] en la determinación de los autovectores/autovalores de la matriz de covarianzas $\mathbf{C}=(\mathbf{x}_i,\mathbf{x}_i^T)$ de dimensión m×m, simétrica y no-negativa que determina el conjunto de autovectores ortogonales Φ's y autovalores reales y no-negativos $\lambda_1\geq\lambda_2\geq\lambda_3\ldots\geq\lambda_m$.

$$\mathbf{x_i} = \sum_{j=1}^{m} a_{ij}\Theta_j \approx \sum_{j=1}^{p} a_{ij}\Theta_j \quad _{p<m}$$

En donde los coeficientes a_{ij} se obtienen proyectando los vectores de datos sobre cada autovector, esto es $a_{ij}=(\mathbf{x}_i, \Phi_i)$

Jiang *et al* en el anexo de su artículo publicado en 2013 [270], y que ya fue analizado en el apartado dedicado a detección de anomalías, exponen también la formulación del ACP y de la EKL. Según Jiang *et al* [270], la EKL fue primeramente considerada para representar un proceso estocástico como una combinación lineal infinita de funciones ortogonales continuas [301], [302] y posteriormente se consideró la formulación de su versión discreta. La versión unidimensional (que indican se corresponde con el ACP) ha sido aplicada en una amplia variedad de campos. La generalización del ACP a la EKL pasa por expandir la matriz original de datos $\mathbf{X} \in \Re^{n\ x\ p}$ a $\mathbf{X'} \in \Re^{(n-N+1\ x\ pN)}$ en el dominio temporal y espacial como sigue:

$$\mathbf{X'^T} = \begin{bmatrix} x_1(1) & \cdots & x_1(t) & \cdots & x_1(n-N+1) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_1(N) & \cdots & x_1(t+N-1) & \cdots & x_1(n) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_p(1) & \cdots & x_p(t) & \cdots & x_p(n-N+1) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_p(N) & \cdots & x_p(t+N-1) & \cdots & x_p(n) \end{bmatrix}$$

Donde N es un desplazamiento hacia delante en el dominio temporal.

El punto de partida es un proceso unidimensional x(t) con media cero sobre el intervalo temporal $t \in [a,b]$. Según la EKL, el proceso x(t) admite una descomposición de la forma:

$$x(t) = \sum_{i=1}^{\infty} \alpha_i \Psi_i(t)$$

Donde α_i son variables aleatorias incorreladas y la función $\Psi_i(t)$ son funciones determinísticas continuas y ortogonales tales que

$$\int_D \Psi_i(t)\Psi_j(t)dt = \delta_{ij} \qquad \text{con} \qquad \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

Supongamos que $K_x(t,s)$ es la función continua de covariación de x(t), esto es $K_x(t,s)=E[x(t)x(s)]$ (recordemos que la media es nula y que por lo tanto la correlación es igual a la covarianza), entonces Ψ_i son las autofunciones de $K_x(.,.)$ y se obtienen resolviendo la ecuación integral de Fredholm

$$\int_a^b K_s(t,s)\Psi_j(s)ds = \lambda_i \Psi_j(t)$$

Así, los coeficientes α_i serían calculados como

$$\alpha_i = \int_a^b x(t)\Psi_i(t)dt$$

Solo se conocen soluciones analíticas de la ecuación integral de Fredholm para geometrías simples y formas específicas de la función de autocovarianza, constituyendo esta la principal dificultad para la utilización de la EKL [284]. En el caso de problemas reales se requieren tratamientos numéricos para la obtención de soluciones. Estos métodos numéricos, como por ejemplo Galerkin [300], usualmente dan lugar a matrices densas que con computacionalmente costosas de procesar [284].

Pero como habitualmente para el análisis de datos se trabaja en un “dominio discreto” en el tiempo y con procesos finitos, ese planteamiento “continuo” debe ser convenientemente adaptado [270]. Así, el proceso estocástico continuo x(t) se considera muestreado de manera uniforme con periodicidad Δt en un intervalo temporal (a,b) con lo que obtendremos un vector n dimensional

$$\mathbf{x} = [x(1), x(2),\ldots, x(n)]^T \qquad \text{con } n=(b-a)/\Delta t.$$

Con este muestreo la función de covariación $K_s(t,s)$ se transforma en la matriz de covarianza $\Gamma_{xx}=E[\mathbf{x}\mathbf{x}^T]$. Para su estimación se suele utilizar el algoritmo de promediado de ventana desplazada, que esencialmente implica el promediado de los productos externos de una ventana desplazada sobre x, esto es

$$\Gamma_{xx} = \sum_{i=1}^{n-N+1} \mathbf{x}_i \mathbf{x}_i^T$$

donde $\mathbf{x}_i=[x_i,\ x_{i+1},\ldots,x_{i+N-1}]^T$ es el subvector de longitud N del vector **x**. El factor de normalización es generalmente ignorado ya que es irrelevante para la determinación de los autovectores de Γ_{xx}, no así de sus autovalores.

El sumatorio anterior puede ser expresado en forma matricial $\Gamma_{xx}=\mathbf{X}^T\mathbf{X}$ con la siguiente matriz de datos **X** "expandida" del vector **x** :

$$\mathbf{X}^T = \begin{bmatrix} x(1) & x(2) & \cdots & x(n-N+1) \\ x(2) & x(3) & \cdots & x(n-N+2) \\ \vdots & \vdots & \ddots & \vdots \\ x(N) & x(N+1) & \cdots & x(n) \end{bmatrix}$$

Con todo ello la anterior integral de Fredholm se transforma en un problema matricial de obtención de autovectores, para obtener el vector de KLE (o componente principal) asociado con **X**, esto es $\Gamma_{xx}\Psi_i=\lambda_i\psi_i$.

Los autovectores ψ_i capturan la correlación temporal de un proceso estocástico discreto, un único flujo, mientras que el ACP tradicional considera la correlación espacial entre diferentes flujos. Para considerar ambas correlaciones espacial y temporal se extiende la EKL unidimensional a múltiples dimensiones para tratar con múltiples procesos estocásticos p que definimos como

$$\mathbf{X}=[\mathbf{x}_1^T, \mathbf{x}_2^T,\ldots \mathbf{x}_p^T]^T$$

La i-ésima componente $\mathbf{x}_i$ de la i-ésima fuente tiene la forma $\mathbf{x_i} = [x_i(1), x_i(2),\ldots, x_i(n)]^T$. Según eso la nueva matriz de covarianza se define como $\Gamma_{xx}=E(\mathbf{X}\mathbf{X}^T)$ con la siguiente estructura:

$$\Gamma_{XX} = \begin{bmatrix} \Gamma_{x_1x_1} & \cdots & \Gamma_{x_1x_p} \\ \vdots & \ddots & \vdots \\ \Gamma_{x_px_1} & \cdots & \Gamma_{x_px_p} \end{bmatrix}$$

Si consideramos el estimador de la matriz de covarianza unidimensional y su correspondiente formato matricial, obtendremos la matriz $\mathbf{X'}^T$ definida anteriormente:

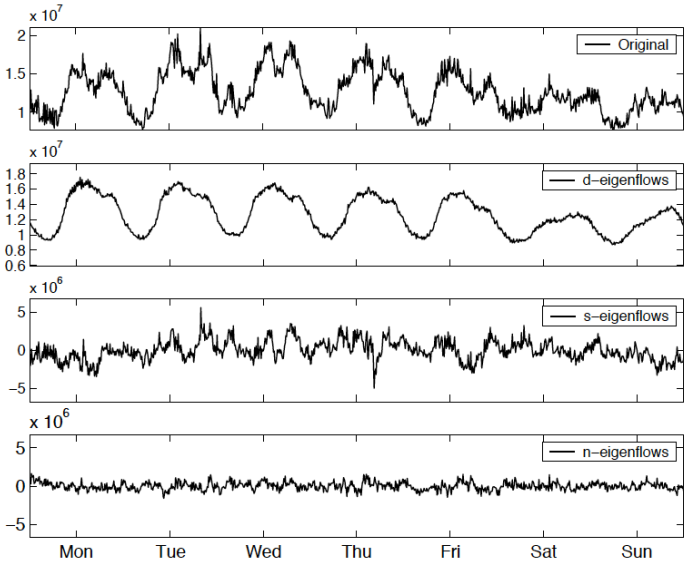
$$\mathbf{X'}^T = \begin{bmatrix} x_1(1) & \cdots & x_1(t) & \cdots & x_1(n-N+1) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_1(N) & \cdots & x_1(t+N-1) & \cdots & x_1(n) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_p(1) & \cdots & x_p(t) & \cdots & x_p(n-N+1) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_p(N) & \cdots & x_p(t+N-1) & \cdots & x_p(n) \end{bmatrix}$$

Sus autovectores pueden obtenerse resolviendo $\Gamma_{XX}\Psi_i=\lambda_i\psi_i$ y el resultado que se obtiene considera, como se desea, ambas correlaciones espacial y temporal.

6.2. Análisis de series temporales de matrices Origen-Destino basado en ACP

Como ya vimos en el capítulo dedicado específicamente a detección de anomalías, Lakhina *et al* en 2004 [212] publicaron el análisis de un juego de datos consistente en series temporales de matrices Origen-Destino (O-D) capturados de la red Abilene y de la red europea Sprint. Un análisis preliminar de dicho juego de datos de la red Abilene fue ya publicado por los mismos autores en 2003 en un documento técnico de la Universidad de Boston [213]. Como ellos mismos indican una de las primeras diferencias del trabajo de Lakhina *et al* [212] en comparación con los publicados hasta esa fecha sobre análisis de series de datos correspondientes a tráfico en redes es el alcance de los juegos de datos utilizados. Los trabajos existentes hasta entonces estaban enfocados en el estudio del tráfico en un <u>único enlace</u> de manera aislada. Sin embargo, en un amplio rango de problemas de ingeniería de redes los investigadores requieren modelar y analizar el tráfico de <u>varios enlaces simultáneamente</u>, por ejemplo ingeniería de tráfico, estimación de la matriz de tráfico O-D, detección de anomalías, detección de intrusos/ataques, predicción de tráfico y planificación de capacidad. En todos ellos la matriz de Origen-Destino (O-D) juega un papel fundamental [58]. El principal reto al que se enfrenta el investigador al estudiar las matrices O-D es la estructura multivariante que presentan, con una elevada dimensionalidad. Por ejemplo, afirman los autores que una red de tamaño medio puede transportar cientos de pares O-D, cuyas correspondientes series temporales presentan centenares de dimensiones. Como los autores indican, el Análisis de Componentes Principales (ACP), "también conocido" como descomposición de Karhunen-Loève (EKL) y Descomposición en Valores Singulares (DVS) [sic], es la técnica más común para

analizar estructuras de alta dimensionalidad: dado un objeto con una elevada dimensionalidad y su espacio de coordenadas asociado, el ACP "descubre" un nuevo espacio de coordenadas que es el mejor de posible utilización para reducir la dimensionalidad de la estructura en cuestión. Los autores analizan, además de la bondad del espacio de dimensión reducida para capturar la información esencial de la estructura del tráfico, la estabilidad en el tiempo de la estructura extraída de las componentes principales del conjunto de matrices O-D. Concluyen que la estructura obtenida por el ACP presenta tres componentes, una tendencia determinista, una parte correspondiente a impulsos de muy corta duración y otra restante de ruido, y que se corresponden con la descomposición de la serie temporal en sus componentes principales.



Descomposición de una serie temporal O-D en sus tres componentes principales. [212]

En este capítulo vamos a detenernos a analizar más en profundidad en lo que Lakhina *et al* propusieron en su artículo [212]. Como ya hemos indicado su propuesta se basa en la aplicación del Análisis de Componentes Principales a la matriz de series temporales O-D. Como exponen los autores, calcular las componentes principales es equivalente a la obtención de los autovalores y autovectores de la matriz $\mathbf{X}^T\mathbf{X}$, siendo la matriz **X** la matriz que contiene las sucesivas mediciones de tráfico entre un origen y un destino determinado a lo largo del tiempo, y tiene dimensiones de t filas y p columnas.

		p			
t	t	Origen A – Destino B	Origen A – Destino C	…	Origen N – Destino M
	1				
	2				
	3				
	…				

En su artículo Lakhina *et al* [212] realizan una interpretación geométrica para finalmente obtener una proyección de los datos originales en el espacio de las componentes principales, ponderado por el valor propio de la componente principal correspondiente:

$$\mathbf{u}_i = \frac{\mathbf{X}\mathbf{v}_i}{\sigma_i} \qquad i = 1,...,p$$

Siendo $\sigma_i = \sqrt{\lambda_i}$ con λ_i el autovalor correspondiente a $\mathbf{v}_i$ y cada componente principal $\mathbf{v}_i$ el i-ésimo autovector calculado a partir de la descomposición espectral de $\mathbf{X}^T\mathbf{X}$ esto es:

$$\mathbf{X}^T\mathbf{X}\mathbf{v}_i = \lambda_i\mathbf{v}_i \qquad i = 1,\ldots,p$$

Siendo p el número de pares O-D considerados.

Los autores exponen que los vectores $\mathbf{u}_i$ son de dimensión t y ortogonales por construcción. Estos vectores $\mathbf{u}_i$ capturan la variación temporal común a todos los flujos de tráfico a lo largo del eje principal i-ésimo. Dado que los ejes principales están ordenados en función de su contribución a la energía (o varianza), $\mathbf{u}_1$ captura la tendencia temporal más fuerte común a todos los flujos O-D, $\mathbf{u}_2$ la siguiente tendencia temporal más fuerte, y así sucesivamente. Ya que el conjunto de $\{\mathbf{u}_i\}_{i=1,...,p}$ captura las tendencias temporales comunes a todos los flujos Origen-Destino, los autores proponen denominarlos autoflujos (o "*eigenflows"*) de **X**.

El conjunto de componentes principales $\{\mathbf{v}_i\}_{i=1,...,p}$ puede ser ordenado como columnas de una matriz principal **V**, de dimensión p×p. Así, podemos construir una matriz **U** de orden t×p cuya columna i-ésima es el vector $\mathbf{u}_i$. Con este formato cada flujo O-D $\mathbf{X}_i$ se puede expresar como:

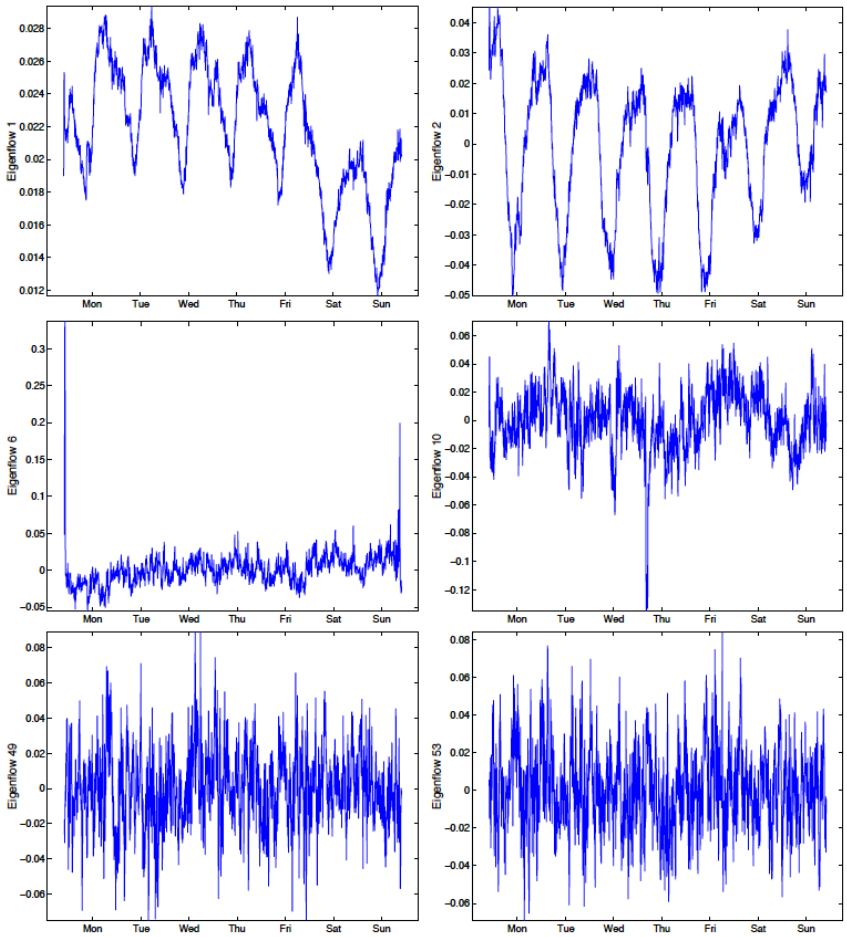
$$\frac{\mathbf{X}_i}{\sigma_i} = \mathbf{U}\left(\mathbf{V}^T\right)_i \qquad i = 1,...,p$$

Donde $\mathbf{X}_i$ es la serie temporal del i-ésimo flujo O-D,
$(\mathbf{V}^T)_i$ es la i-ésima fila de **V**
$\sigma_i = \sqrt{\lambda_i}$ con λ_i el autovalor correspondiente a $\mathbf{v}_i$.

Esto es, cada flujo O-D $\mathbf{X}_i$ es una combinación lineal de los autoflujos, con $(\mathbf{V}^T)_i$ como peso asociado al mismo.

Los autores realizan en este punto un estudio "tradicional" de Análisis de Componentes Principales de la matriz **X**, incluyendo los *scree plots* para analizar la

dimensionalidad de la matriz y la posibilidad de reducir la misma. Por último se concreta una taxonomía de los autoflujos clasificándolos, como hemos indicado anteriormente, entre “deterministas”, “picos” y “ruido”. Los primeros presentan una tendencia periódica, los segundos muestran picos abruptos y por último los terceros aparentan ser ruido aleatorio.



Ejemplos de autoflujos extraídos del conjunto de datos medidos en la red Abilene. [212].

Se ha intentado obtener el mismo conjunto de datos utilizado por Lakhina *et al* [212] ya que parecía muy interesante disponer de ellos para replicar inicialmente su propuesta, dado que en su trabajo precedente del año 2003 [213] se ofrecía la descomposición detallada de todos los flujos O-D de la matriz **X,** resultando la gestión infructuosa. Como veremos ha sido posible soslayar este inconveniente a través de un artículo posterior de otros autores [265].

En el apartado dedicado a detección de anomalías se comentaron diversos artículos publicados desde el año 2004 en los que se utilizaban el ACP para la detección de anomalías en redes, o se estudiaba la problemática asociada a su aplicación. Por la repercusión que tiene para el trabajo de Lakhina *et al* [212], y para otros que aplican el ACP, traemos a colación de nuevo aquí el trabajo que Ringberg *et al* publicaban en el

año 2007 [227]. En él ponían de manifiesto la sensibilidad del ACP para la detección de anomalías. Concluían los autores que la utilización del ACP para el análisis de datos de tráfico es más difícil de lo que el trabajo de Lakhina *et al* [212] sugiere. Esto es principalmente debido a la problemática existente con la determinación de la dimensión del espacio considerado "normal" y la imposibilidad de identificar qué está sucediendo realmente para que sea posible afirmar que existe una "irregularidad" en el patrón de tráfico observado. De hecho, se indica que "no hay relación directa entre el subespacio de dimensión reducida obtenido por el ACP y la localización espacial original de la anomalía". Como veremos más adelante, esta situación puede ser resuelta a través de nuestra propuesta.

Como también se vio en aquella sección, partiendo de la sensibilidad del ACP para su utilización en la detección de anomalías expuesta por Ringberg *et al* [227], Brauckhoff *et al* en 2009 [250] efectuaron un diagnóstico más detallado de la situación, concluyendo que el principal problema de la aplicación del ACP a este tipo de datos reside en que el ACP tradicional NO considera la correlación temporal de los datos. En el desarrollo de su artículo, Brauckhoff *et al* [250] exponen una extensión del ACP a procesos estocásticos discretos en la que aplicando la expansión de Karhunen-Loève y el método de Galerkin los autores obtienen una nueva matriz de correlación espacio-temporal sobre la que finalmente aplican el ACP tradicional. Esta matriz consiste realmente en una reordenación de los datos observados, planteamiento en el que se coincide con lo expuesto por Jiang *et al* [270].

Expongamos, con algo más detalle el desarrollo expuesto en Brauckhoff *et al* [250]. Consideremos un vector de dimensión K de procesos estocásticos discretos de media nula:

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_K)^T$$

Brauckhoff *et al* [250] proponen la construcción de una nueva matriz de observaciones de dimensión KN x (n-K) siendo N el número a partir del cual los valores de la covarianza de los procesos son despreciables.

La matriz que obtiene es la siguiente:

$$\mathbf{X} = \begin{bmatrix} x_1(1) & \cdots & x_1(n-N) \\ x_1(2) & \cdots & x_1(n-N+1) \\ \vdots & \ddots & \vdots \\ x_1(N) & \cdots & x_1(n) \\ x_2(1) & \cdots & x_2(n-N) \\ \vdots & \ddots & \vdots \\ x_2(N) & \cdots & x_2(n) \\ \vdots & \ddots & \vdots \\ x_K(1) & \cdots & x_K(n-N) \\ \vdots & \ddots & \vdots \\ x_K(N) & \cdots & x_K(n) \end{bmatrix}$$

Resaltamos aquí el "parecido" entre esta matriz propuesta por Brauckhoff *et al* [250] y la que posteriormente propondrían Jiang *et al* [270]:

$$\mathbf{X'}^{\mathsf{T}} = \begin{bmatrix} x_1(1) & \cdots & x_1(t) & \cdots & x_1(n-N+1) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_1(N) & \cdots & x_1(t+N-1) & \cdots & x_1(n) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_p(1) & \cdots & x_p(t) & \cdots & x_p(n-N+1) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_p(N) & \cdots & x_p(t+N-1) & \cdots & x_p(n) \end{bmatrix}$$

A partir de esta nueva matriz **X** Brauckhoff *et al* [250] obtienen una nueva matriz de covarianzas espacio-temporales:

$$\Sigma = \frac{1}{n-N-1}\mathbf{X}\mathbf{X}^{\mathsf{T}}$$

A esta matriz se le aplica el ACP "tradicional", obteniendo los correspondientes autovectores y autovalores. Los autores declaran en repetidas ocasiones en su artículo que la utilización de esta nueva matriz de datos, más compleja que la inicial tradicionalmente utilizada, es inevitable cuanto hay que tratar con observaciones que están correladas [250].

Como ya indicamos, la matriz que proponen Brauckhoff *et al* [250] para la obtención de la matriz que se someterá al ACP presenta también notables similitudes con la matriz de trayectorias del algoritmo MSSA (*Multichannel Singular Spectrum Analysis*) [243], [210]:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}(1) & \mathbf{x}(2) & \cdots & \mathbf{x}(N-L+1) \\ \mathbf{x}(2) & \mathbf{x}(3) & \cdots & \mathbf{x}(N-L+2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}(L) & \mathbf{x}(L+1) & \cdots & \mathbf{x}(N) \end{bmatrix}$$

En este caso se trata de la matriz de trayectorias del SSA para un único proceso estocástico y que se yuxtapone en vertical para incorporar otros procesos estocásticos si los hubiese, resultando dicha última matriz similar a la propuesta por Brauckhoff *et al* [250] y la misma matriz que presentan en el anexo de su artículo Jiang *et al* [270].

Para clarificar el problema puesto de manifiesto por Brauckhoff *et al* [250], la imposibilidad de detectar algunos tipos de correlación mediante la matriz de varianzas/covarianzas "tradicional", revisaremos brevemente algunos principios y definiciones básicas en el procesado de señales discretas o de series temporales.

6.3. DEFINICIÓN Y TRATAMIENTOS BÁSICOS DE SERIES TEMPORALES

Definiremos una serie temporal **Y** de longitud N como un vector unidimensional de N componentes:

$$\mathbf{Y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_N \end{pmatrix}$$

Este vector puede ser interpretado como las N diferentes medidas de una determinada variable en otros tantos lugares, tiempos, personas o, en general, diferentes unidades taxonómicas.

Lógicamente, cada uno de los elementos de la serie numérica podemos referenciarlo por un índice numérico entero:

$$y[1] = y_1 \quad y[2] = y_2 \quad \ldots \quad y[n] = y_n$$

Consideremos, por ejemplo, que se trata de diferentes medidas de una variable obtenidas en N tiempos 't', sean estos años, días, etc.

Definiremos a continuación algunos operadores "básicos" para esta series temporales.

6.3.1. Operador Convolución Discreta

Se define la convolución discreta **W** de dos series numéricas (temporales) $\mathbf{Y}_1$ e $\mathbf{Y}_2$ como [282]:

$$w[n] = \mathbf{Y_1} * \mathbf{Y_2} = y_1[n] * y_2[n] = \sum_{k=-\infty}^{+\infty} y_1[k] y_2[n-k]$$

La interpretación de ese operador es sencilla. Supongamos que tenemos dos series numéricas concretas:

$$\mathbf{Y_1} = (1 \quad 2 \quad 3 \quad 4) \qquad \mathbf{Y_2} = (5 \quad 6 \quad 7)$$

La convolución **W** entre ambas series se obtiene de la siguiente manera:

1) Invertimos el orden de la segunda serie numérica

$$\mathbf{Y_2'} = (7 \quad 6 \quad 5)$$

2) La serie segunda se desplaza "sobre" la primera multiplicando las componentes concordantes:

Y_1	0	0	0	1	2	3	4	0	0	0
Y'_2	0	7	6	5	0	0	0	0	0	0
W	0	0	0	5	0	0	0	0	0	0

Y_1	0	0	0	1	2	3	4	0	0	0
Y'_2	0	0	7	6	5	0	0	0	0	0
W	0	0	0	6	10	0	0	0	0	0

Y_1	0	0	0	1	2	3	4	0	0	0
Y'_2	0	0	0	7	6	5	0	0	0	0
W	0	0	0	7	12	15	0	0	0	0

Y_1	0	0	0	1	2	3	4	0	0	0
Y'_2	0	0	0	0	7	6	5	0	0	0
W	0	0	0	0	14	18	20	0	0	0

$\mathbf{Y_1}$	0	0	0	1	2	3	4	0	0	0
$\mathbf{Y'_2}$	0	0	0	0	0	7	6	5	0	0
W	0	0	0	0	0	21	24	0	0	0

$\mathbf{Y_1}$	0	0	0	1	2	3	4	0	0	0
$\mathbf{Y'_2}$	0	0	0	0	0	0	7	6	5	0
W	0	0	0	0	0	0	28	0	0	0

3) El resultado final se obtiene sumando los diferentes productos obtenidos, así el resultado de la convolución entre $\mathbf{Y}_1$ e $\mathbf{Y}_2$ es:

$$\mathbf{W} = \begin{pmatrix} 5 & 6+10 & 7+12+15 & 7+12+13 & 14+18+20 & 21+24 & 28 \end{pmatrix}$$

$$\mathbf{W} = \begin{pmatrix} 5 & 16 & 34 & 32 & 52 & 45 & 28 \end{pmatrix}$$

Este operador de convolución tiene gran importancia en procesado de señales y sistemas, ya que determina la respuesta de un sistema lineal discreto a la entrada de una señal discreta cualquiera. Existe, lógicamente, un equivalente en el mundo de las funciones continuas, en donde la convolución se define a partir de una integral:

$$w(t) = \int_{-\infty}^{+\infty} y_1(\tau) y_2(t-\tau)\, d\tau$$

El operador de convolución presenta las siguientes propiedades matemáticas:

- Conmutativa: $\mathbf{Y_1} * \mathbf{Y_2} = \mathbf{Y}_2 * \mathbf{Y}_1$
- Asociativa: $\mathbf{Y_1} * \left[\mathbf{Y}_2 * \mathbf{Y}_3\right] = \left[\mathbf{Y}_1 * \mathbf{Y}_2\right] * \mathbf{Y}_3$
- Distributiva: $\mathbf{Y_1} * \left[\mathbf{Y}_2 + \mathbf{Y}_3\right] = \left[\mathbf{Y}_1 * \mathbf{Y}_2\right] + \left[\mathbf{Y}_1 * \mathbf{Y}_3\right]$

Y es muy interesante comprobar que la convolución discreta puede ser también interpretada en términos de un producto de matrices en donde una de las series se ha transformado en una matriz Toeplitz [303].

$$\mathbf{W} = w[n] = \mathbf{Y_1} * \mathbf{Y_2} = y_1[n] * y_2[n] = \begin{bmatrix} w[1] \\ w[2] \\ w[3] \\ \vdots \\ w[n] \end{bmatrix} = \begin{bmatrix} y_1[1] & 0 & \cdots & 0 & 0 \\ y_1[2] & y_1[1] & \cdots & \vdots & \vdots \\ y_1[3] & y_1[2] & \cdots & 0 & 0 \\ \vdots & y_1[3] & \cdots & y_1[1] & 0 \\ y_1[m-1] & \vdots & \cdots & y_1[2] & y_1[1] \\ y_1[m] & y_1[m-1] & \ddots & \vdots & y_1[2] \\ 0 & y_1[m] & \cdots & y_1[m-2] & \vdots \\ 0 & 0 & \cdots & y_1[m-1] & y_1[m-2] \\ \vdots & \vdots & \ddots & y_1[m] & y_1[m-1] \\ 0 & 0 & 0 & \cdots & y_1[m] \end{bmatrix} \begin{bmatrix} y_2[1] \\ y_2[2] \\ y_2[3] \\ \vdots \\ y_2[n] \end{bmatrix}$$

6.3.2. Función de correlación "temporal" (y autocorrelación)

La función de correlación entre dos series temporales (señales discretas) se define como [282]:

$$\Phi_{xy}[n] = \sum_{m=-\infty}^{+\infty} x[m+n]y[m]$$

Y, análogamente, la función de autocorrelación como:

$$\Phi_{xx}[n] = \sum_{m=-\infty}^{+\infty} x[m+n]x[m]$$

Como puede apreciarse, la función de correlación temporal puede calcularse a partir de la convolución de manera sencilla:

$$\Phi_{xy}[n] = x[n] * y[-n]$$

Esto es, la función de correlación temporal se obtiene convolucionando las dos series, pero invirtiendo previamente una de ellas. Considerando que la convolución ya invierte una, la correlación se calcula igual que la convolución, pero sin invertir la serie.

Por ejemplo, calculemos la correlación de dos series de igual longitud:

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix}$$

La correlación entre dichas series se obtendría como:

$$\Phi_{xy} = \begin{pmatrix} y_1x_3 & y_1x_2 + y_2x_3 & y_1x_1 + y_2x_2 + y_3x_3 & y_2x_1 + y_3x_2 & y_3x_1 \end{pmatrix}$$

Fijémonos en el término central de la función de correlación:

$$\Phi_{xy}\left[0\right] = y_1x_1 + y_2x_2 + y_3x_3$$

La expresión se corresponde con el producto escalar de los vectores formados por las series numéricas:

$$\Phi_{xy}\left[0\right] = y_1x_1 + y_2x_2 + y_3x_3 = \mathbf{xy}$$

Una interesante propiedad es que la función de autocorrelación presenta su máximo en el origen [282]

$$\Phi_{xx}\left[n\right] \leq \Phi_{xx}\left[0\right] \quad \forall n$$

Pero esto no es siempre así para el resto de casos de correlaciones entre series (temporales) diferentes, en las que la función de correlación temporal es capaz de "detectar", a través de sus máximos, la existencia de similitudes entre series que se encuentran situadas fuera del valor central "tradicional". Así, es capaz de detectar desplazamientos temporales que de otro modo podrían pasar desapercibidos.

6.3.3. LIMITACIONES DEL ACP EN PRESENCIA DE CORRELACIONES TEMPORALES Y ESPACIALES

La conclusión que se extrae, de manera preliminar, es que en el trabajo de Lakhina *et al* [212] y en cualquier otro que utilizase el ACP (o la DVS) de la manera tradicional, se estarían posiblemente infravalorando los valores de las correlaciones/covarianzas entre las series temporales (variables) que presentasen, entre otras situaciones, un comportamiento de similitud fuera del valor central. Esto estaría en consonancia con lo expuesto en los trabajos anteriormente reseñados que exponen limitaciones en la aplicación del ACP a este tipo de conjunto de datos [214], [227], [250], [270]

Al igual ocurriría, por lo tanto y en principio, en el análisis que se realizase con cualquiera de los tres métodos Biplot "tradicionales" propuestos en la literatura (GH, JK, HJ). Esto es así ya que, como hemos visto, uno de los caminos para obtener la representación Biplot de una matriz de datos **X** parte de la obtención de las matrices de varianzas y covarianzas de sus m vectores fila (unidades taxonómicas: individuos, tiempos,…) y de sus n vectores columnas (variables, series,…), esto es $\mathbf{XX}^T$ y $\mathbf{X}^T\mathbf{X}$**.** Posteriormente se obtienen sus autovalores (iguales para ambos casos) y sus

autovectores para calcular los marcadores fila y marcadores columna en función del tipo de Biplot a considerar, o bien se obtienen desde la DVS de la matriz **X.**

$$\mathbf{X} = \mathbf{U}\,\Sigma\,\mathbf{V}^T$$

Marcadores	Dimensiones	Denominación	Definición
Fila	m × n	**G**	**U**
Columna	n × n	**H**	**ΣV**
Fila	m × n	**J**	**UΣ**
Columna	n × n	**K**	**V**

Como hemos visto las matrices $\mathbf{XX}^T$ y $\mathbf{X}^T\mathbf{X}$ no son más que el resultado de los productos interiores (escalares) entre los respectivos vectores fila ($\mathbf{XX}^T$) y vectores columna ($\mathbf{X}^T\mathbf{X}$) y estos productos interiores de filas por filas y columnas por columnas para la obtención de ambas matrices no son más que una medida de la covarianza entre las respectivas series temporales de filas y columnas de **X** [212]. Pero como podemos ver, hay en esta aplicación de cualquiera de los métodos Biplot "tradicionales" una diferencia con relación a los trabajos en los que se aplica el ACP, la incorporación de la matriz que tiene en cuenta la correlación/covarianza de la matriz de los datos traspuesta y por tanto de la otra "dimensión" en juego, temporal o espacial. Esta consideración no se tiene en cuenta en los trabajo que solo utilizan el ACP, por lo que es de esperar que los métodos Biplot presenten mejoras a la hora de considerar las correlaciones espacio-temporales que dificultan la aplicación del ACP citadas por Brauckhoff *et al* [250].

La Descomposición en Valores Singulares (DVS) [158] es el resultado de la aplicación del teorema de descomposición según el cual cualquier matriz real n×m **X** de rango *p* puede ser descompuesta como

$$\mathbf{X}=\mathbf{U}\,\Sigma\,\mathbf{V}^T$$

En donde las columnas de **U** son los autovectores de $\mathbf{XX}^T$, las columnas de **V** son los autovectores de $\mathbf{X}^T\mathbf{X}$ y los elementos σ_i que constituyen la diagonal de **Σ** son las raíces cuadradas de los autovalores o valores propios λ_i de $\mathbf{XX}^T$ o $\mathbf{X}^T\mathbf{X}$ (son iguales los diferentes de 0) y se denominan "valores singulares" de **X**. Pero también esta descomposición puede ser vista como una serie

$$\mathbf{X} = \sum_{i=1}^{p} \lambda_i^{0.5} \mathbf{u}_i \mathbf{v}_i^T$$

Donde el producto $\mathbf{u}_i\mathbf{v}_i^T$ es una matriz n × m resultado del producto de la matriz n×1 $\mathbf{u}_i$ y la matriz 1×m $\mathbf{v}_i^T$. Esta ecuación muestra que la DVS constituye una descomposición de la matriz n×m **X** en una suma de matrices separables de rango 1. Por lo que estaríamos en presencia de una descomposición en serie de la matriz **X** original.

El problema con el que nos encontramos en este punto es la posibilidad, al menos teóricamente, de utilizar una función que permita "convertir" la serie numérica obtenida mediante la correlación temporal de filas o columnas de la matriz **X**, en un único valor que pueda establecerse en las respectivas "nuevas" matrices de varianzas y covarianzas, en una suerte de formulación paralela a la realizada por otros autores [214], [227], [250], [270]. En el caso "tradicional" esa función se correspondería con el valor de la correlación temporal en el origen. Posteriormente a la obtención de esas nuevas matrices de correlación se calcularía la descomposición en autovectores y autovalores de ambas matrices, para obtener desde ahí los respectivos marcadores fila y columna, que constituirían en potencia nuestro "nuevo" Biplot. Como ya hemos dicho si, por ejemplo, el valor que utilizamos es el valor en el origen, nos encontraríamos en el caso tradicional, si bien estaríamos en presencia de los problemas ya expuestos anteriormente y otros añadidos derivados de la posible obtención de unas matrices para cada caso análogo a las previas "$\mathbf{XX}^T$" y "$\mathbf{X}^T\mathbf{X}$" que conllevasen la existencia de distintos autovalores para ambos casos, algo que no sucede en la DVS original.

No obstante nuestro planteamiento será más "convencional" dentro de la metodología Biplot, de la que ya existen precedentes de su aplicación al estudio de matrices de tráfico. En el año 2012 Roughan *et al* [267] publicaron un artículo en el que proponían, con otro nombre y sin aplicar realmente muchas de sus propiedades, la utilización del SQRT-Biplot para el análisis de matrices de tráfico. Como sabemos existen otros biplots alternativos a esta elección. En el capítulo dedicado a detección de anomalías se encuentra un amplio resumen del contenido de dicho artículo y de sus conclusiones. Otra aplicación de los métodos Biplot al estudio de series temporales se encuentra en la tesis doctoral "Biplot Dinámico" [304]. En ella se propone una extensión de los métodos Biplot al tratamiento de los datos de tres dimensiones, una de las cuales puede ser el tiempo. El Biplot Dinámico se realiza en dos etapas, ofreciendo una visión estática y otra dinámica de los datos, representando sobre la

situación elegida la evolución de las diferentes situaciones y posicionando en el mismo gráfico simultáneamente las trayectorias tanto de los individuos como de las variables. El análisis Biplot de la situación de referencia presenta todas las propiedades de la factorización elegida (GH, HJ, JK) y los elementos proyectados en este conservan propiedades similares respecto a la situación de referencia. Se parte de una matriz de tres vías: individuos para las filas, variables para las columnas y situaciones para las diferentes ocasiones. Se fija una referencia temporal/situacional como la base a partir de la que se va a estudiar la dinámica y sobre ese estado inicial se representa la evolución del resto de situaciones temporales.

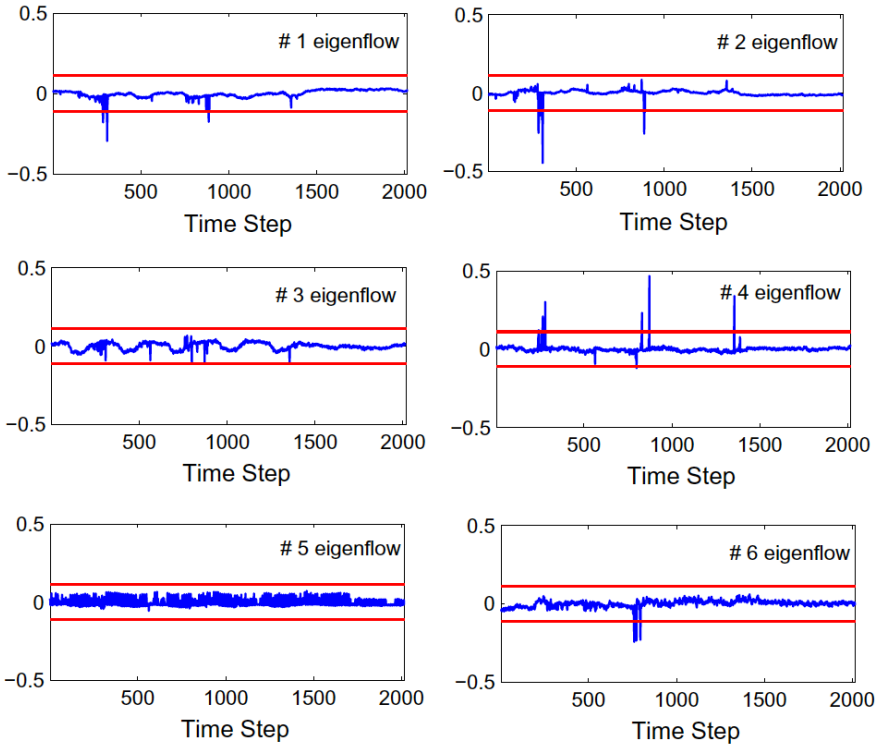
6.4. REPRESENTACIÓN BIPLOT DE LOS AUTOFLUJOS

Con todos los condicionantes expuestos anteriormente, volvamos a la propuesta de Lakhina *et al* [212] a través de un artículo posterior de otros autores diferentes [265] que utiliza también el ACP para el análisis de la matriz de tráfico y cuyo planteamiento nos servirá como una especial "piedra de Rosetta" para el arranque de nuestro análisis.

6.4.1. EL JUEGO DE DATOS

En el año 2012 Wang *et al* [265] retoman el planteamiento de Lakhina *et al* [212], pero utilizando una técnica de descomposición diferente del ACP. No obstante, y esta es la parte que utilizaremos, inicialmente replican el trabajo de Lakhina *et al* [212], ofreciendo la representación gráfica temporal de los autoflujos obtenidos del análisis de un juego de datos que está a disposición de los investigadores en Internet [305].

Así, los autores parten de la matriz **X01** que contiene datos de tráfico Origen-Destino (OD) de la red Abilene durante un tiempo de 1 semana con intervalos de medida de 5 minutos, resultando en una matriz de 144 columnas (flujos OD) y 2016 filas (intervalos temporales). Si bien tras eliminar las columnas correspondientes al tráfico entrante o saliente del nodo ATLA-M5, al establecer que son inestables, se quedan finalmente con 121 columnas y las mencionadas 2016 filas. Ofrecen en su artículo las representaciones de los seis primeros autoflujos de esta nueva matriz **X01** tras su preproceso. Estos resultados nos permitirán validar nuestro punto de partida.



Seis primeros autoflujos de la matriz **X01** de Wang *et al*. [265]

Como hemos indicado Zhang tiene publicado en Internet [305] un valioso juego de datos consistente en 24 matrices de tráfico semanal de la red Abilene, comenzando la primera matriz el 1 de marzo de 2004 y la última el 4 de septiembre de 2004. Cada matriz está dispuesta en un fichero de texto "X*ab*" que contiene 2016 matrices de tráfico, correspondientes a siete días consecutivos en los que constan lecturas de volumen de tráfico cursado cada 5 minutos (12×24×7) con una tasa de muestreo de 100.

F. Inicio	Archivo
2004-03-01	X01
2004-03-08	X02
2004-04-02	X03
2004-04-09	X04
2004-04-22	X05
2004-05-01	X06
2004-05-08	X07
2004-05-15	X08

F. Inicio	Archivo
2004-05-22	X09
2004-05-29	X10
2004-06-05	X11
2004-06-12	X12
2004-06-19	X13
2004-06-26	X14
2004-07-03	X15
2004-07-10	X16

F. Inicio	Archivo
2004-07-17	X17
2004-07-24	X18
2004-07-31	X19
2004-08-07	X20
2004-08-13	X21
2004-08-21	X22
2004-08-28	X23
2004-09-04	X24

Información sobre el juego de datos de la red Abilene publicado por Y. Zhang [305]

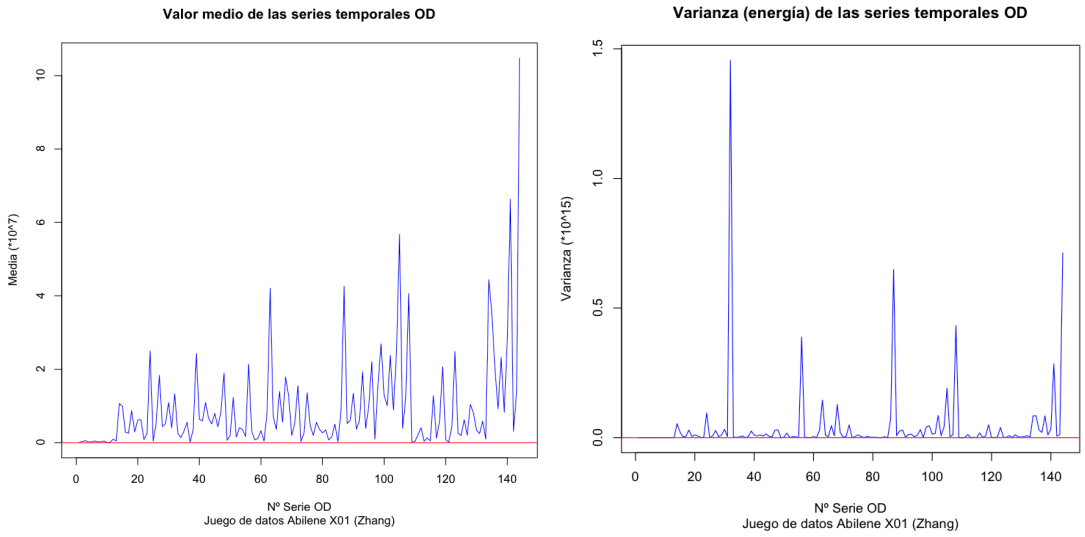
Cada línea del fichero "X*ab*" se corresponde a una matriz de tráfico O-D y contiene 720 valores, correspondientes a las 144 rutas estudiadas (12x12 nodos de la red Abilene) y a 5 valores organizados de la siguiente manera:

<realOD_1> \
<simpleGravityOD_1> \
<simpleTomogravityOD_1> \
<generalGravityOD_1 \
<generalTomogravityOD_1> \
...
<realOD_1> \
<simpleGravityOD_1> \
<simpleTomogravityOD_1> \
<generalGravityOD_1 \
<generalTomogravityOD_1> \

En nuestro caso nos interesan solo los datos de tráfico reales identificados como <realOD_n> que están normalizados por 100 bytes. Para más detalles sobre la información contenida en los archivos se puede consultar la descripción de los mismos ofrecida por el investigador Zhang junto con los juegos de datos [305].

En primer lugar vamos a preprocesar la matriz **X01**, correspondiente a la semana de los días del 1 al 7 de marzo de 2004, ambos incluidos. Como hemos visto el juego de datos publicado por Zhang [305] contiene mucha más información, además del tráfico en bruto de la red, información esa que no es relevante para nuestro trabajo, y que en primer lugar eliminaremos.

Representemos los valores medios y las varianzas (energía) de las series temporales del volumen de tráfico tal cual se encuentran publicadas.

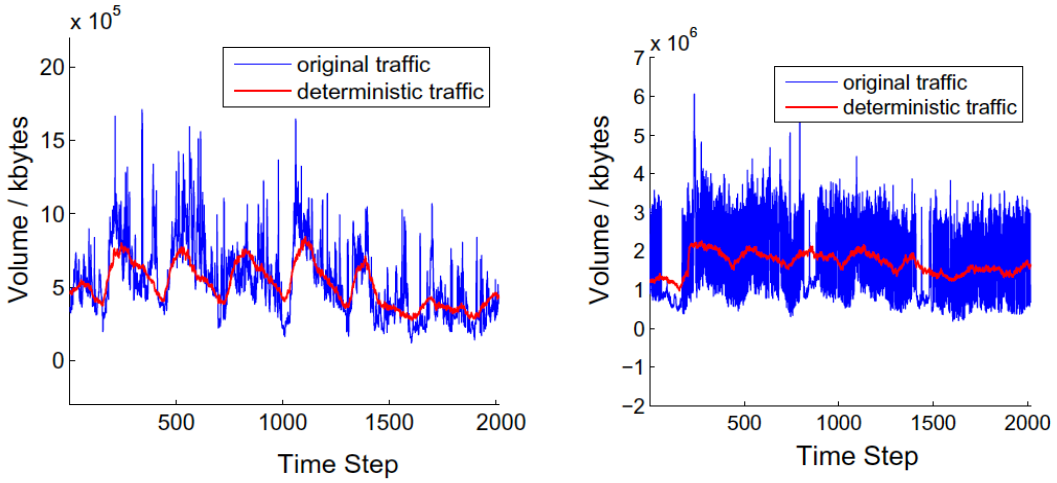


Valor medio y varianza de las series temporales del juego de datos de la red Abilene

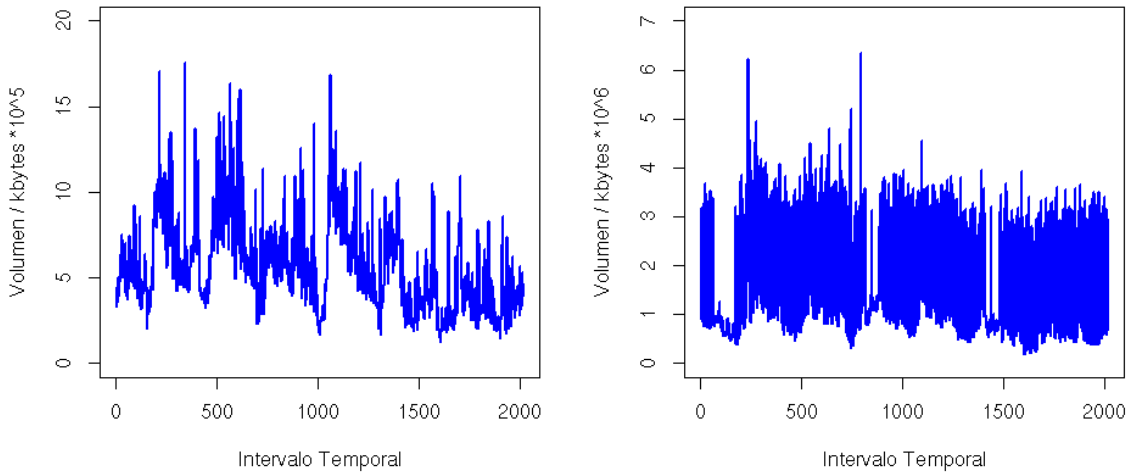
Sobre los datos de tráfico en bruto, Wang *et al* [265] retiran las columnas correspondientes al tráfico entrante y saliente del nodo ATLA-M5 (columnas 1-13, 25, 37, 49, 61, 73, 85, 97, 109, 121, 133) ya que, justifican, contienen un porcentaje muy elevado de ceros. Exponen que esta situación generalmente significa que estos flujos

OD no son estables. En las representaciones anteriores de medias y varianzas que incluyen todas las series iniciales podemos ratificar la situación manifestada por los autores, por lo que retiraremos también de nuestro estudio las columnas indicadas correspondientes al tráfico entrante-saliente de ATLA-M5.

En las figuras 11 y 12 del artículo de Wang *et al* [265] se encuentran representadas las series temporales números 50 y 51, lo que nos permite verificar, con cierta seguridad, que la matriz de datos de que disponemos en este punto es la misma que la utilizada por los autores en su trabajo.



Representación de las series nº 50 y 51 extraídas del trabajo de Wang *et al*.

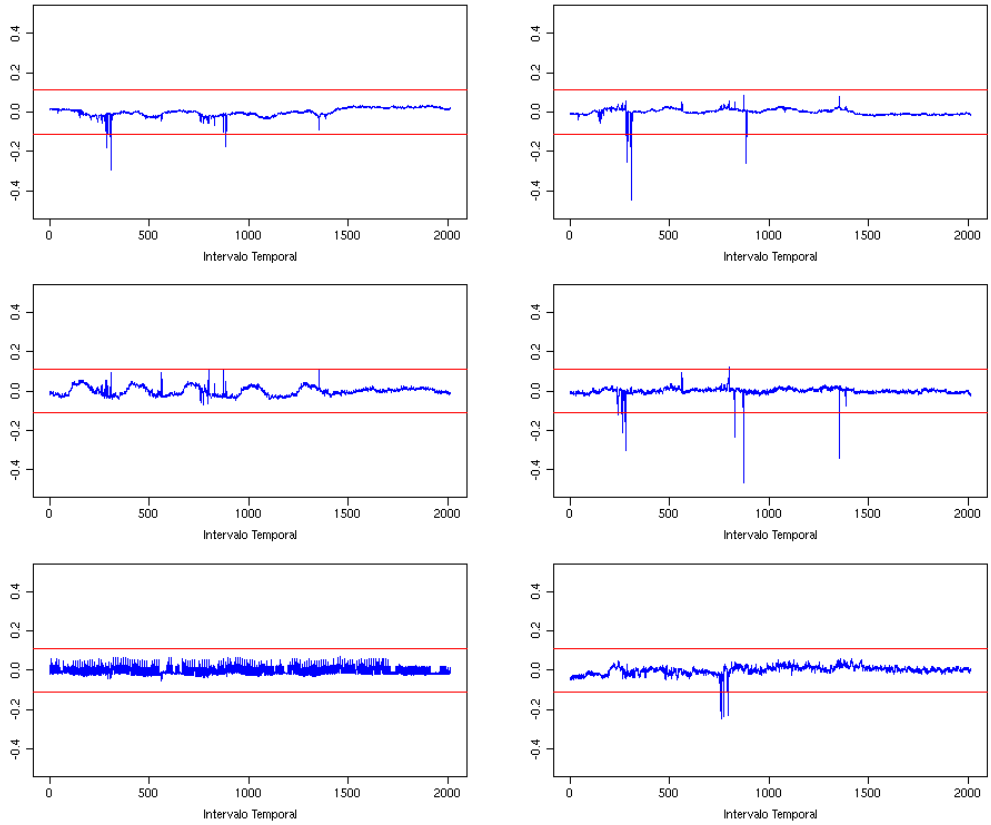


Representación de las series nº 50 y 51 obtenidas en nuestro trabajo.

Así pues, disponemos en este momento del mismo juego de datos que Wang *et al* [265] y podemos abordar la replicación de la parte inicial de su trabajo.

Los autores indican que antes de someter la matriz al ACP la centran por columnas. Posteriormente realizaremos un ACP a través de la DVS de la matriz **X01** centrada.

Finalmente representamos los 6 primeros autoflujos (*eigenflows*) **u**. Al igual que en el trabajo de Wang *et al* [265] las líneas horizontales representan el rango ±5σ para facilitar la comparación entre ambas representaciones.



Representación de los 6 primeros *eigenflows* $\mathbf{u}_i$ (marcadores **G**) obtenidos en nuestro trabajo.

6.4.2. EIGENFLOWS Y BIPLOTS

Como podemos comprobar a simple vista (ver figuras anteriores), se trata de las mismas representaciones que las obtenidas por Wang *et al* en su trabajo [265], que de facto son los marcadores fila **G** de un GH-Biplot.

En efecto si la DVS es $\mathbf{X}=\mathbf{U}\Sigma\mathbf{V}^T$ se puede demostrar que

$$\mathbf{U} = \frac{1}{\Sigma}\mathbf{XV} = \mathbf{G}$$

Lo que coincide son la expresión propuesta por Lakhina *et al* [212] para sus *autoflujos* que se expuso anteriormente:

$$\mathbf{u}_i = \frac{\mathbf{X}\mathbf{v}_i}{\sigma_i} \qquad i = 1,...,p$$

Donde, recordemos, $\sigma_i = \lambda_i^{0.5}$ siendo λ_i los autovalores de $\mathbf{X}^T\mathbf{X}$, esto es σ_i son los valores singulares de **X**, raíces cuadradas de los autovalores de $\mathbf{X}^T\mathbf{X}$ o $\mathbf{X}\mathbf{X}^T$. Y, por

supuesto, **Σ** es la matriz diagonal de valores singulares de **X** en la Descomposición en Valores Singulares de la matriz $\mathbf{X}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Por lo que ambas expresiones coinciden.

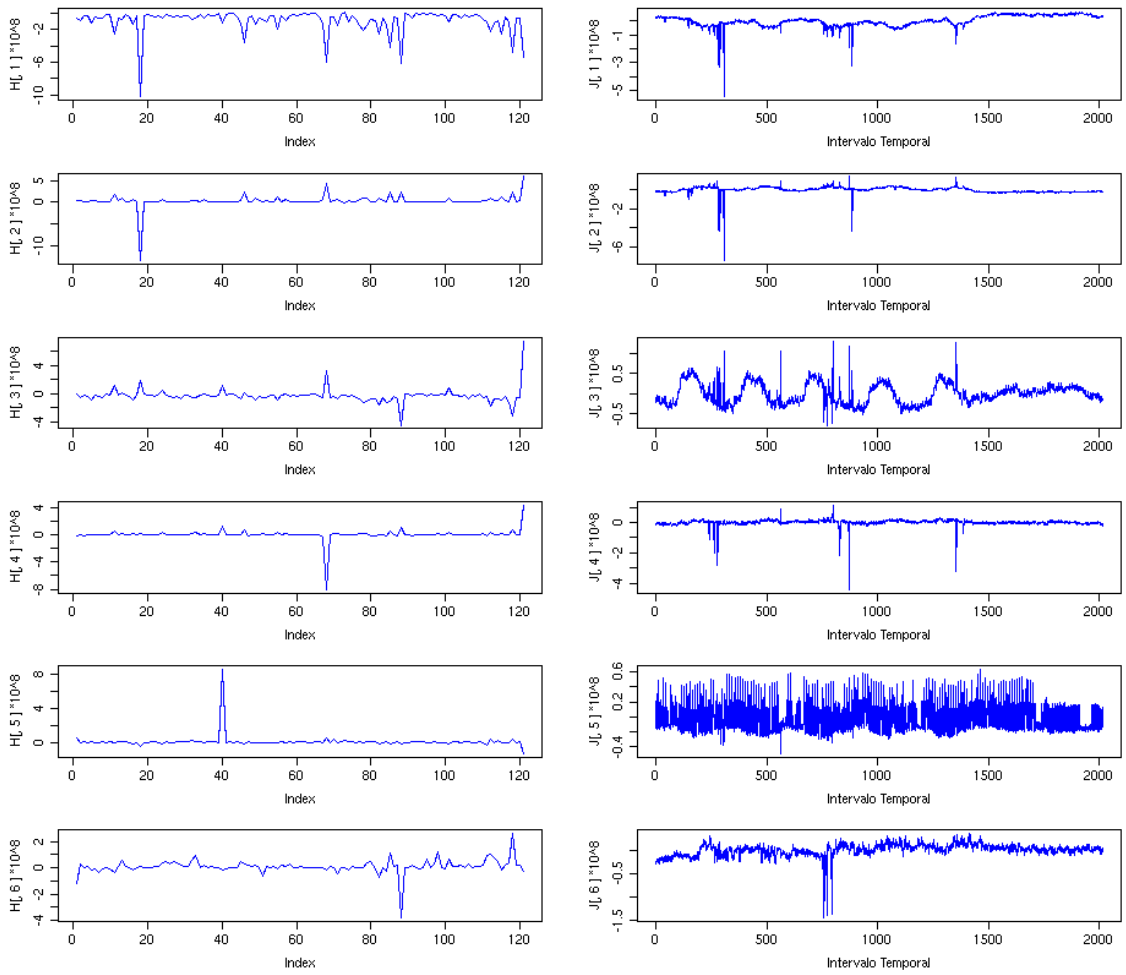
Así pues, Lakhina *et al* [212] y Wang *et al* [265] estaban descomponiendo la matriz de tráfico formada por las series temporales de flujos Origen-Destino, en una suerte de "Biplot temporal GK" siendo **G** y **K** los marcadores correspondientes a los GH y JK Biplots "tradicionales", de Gabriel [157].

6.4.3 APLICACIÓN DEL HJ-BIPLOT AL JUEGO DE DATOS X01

Los marcadores **J** y **H** se corresponden respectivamente con los marcadores **G** y **K** ponderados por los valores singulares de la DVS de la matriz **X**, que son las raíces cuadradas de los autovalores o valores propios λ_i de $\mathbf{X}^T\mathbf{X}$ o $\mathbf{X}\mathbf{X}^T$.

Obtenemos los marcadores **H** y **J** del modo tradicional a partir de nuestra matriz de datos **X01**, conteniendo las series temporales OD centradas de la red Abilene.

Representemos las primeras 4 series temporales de marcadores columna **H** y fila **J**:

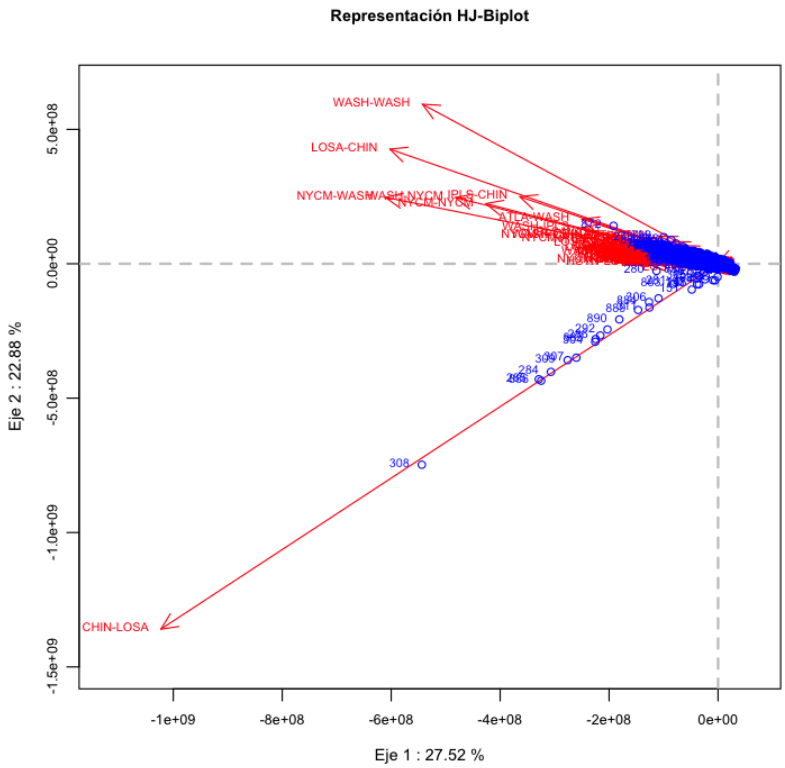


Representación de las 6 primeras series temporales de marcadores H y J de la matriz **X01** centrada.

Como podemos comprobar visualmente los “perfiles” de las representaciones de los *eigenflows* obtenidos anteriormente por Wang *et al* [265] coinciden fielmente con los mostrados de los marcadores fila **J**, diferenciándose exclusivamente en su escalado (el coeficiente del valor propio). Así pues, igualmente la taxonomía que propone Lakhina *et al* [212] sobre sus autoflujos es también plenamente aplicable a nuestros marcadores fila **J** del HJ-Biplot.

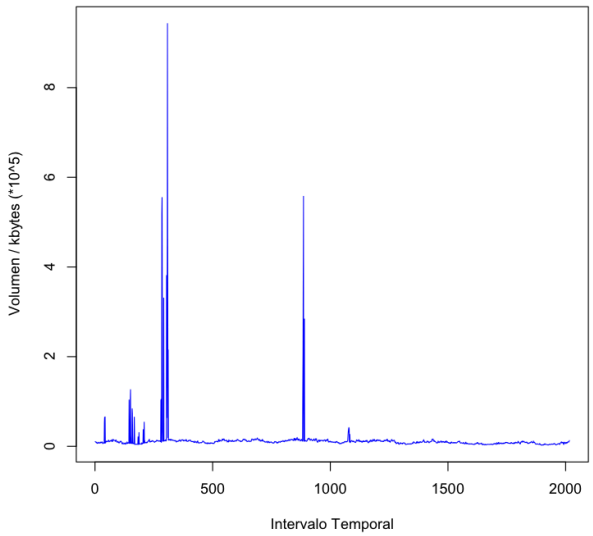
Análogamente los marcadores columna **H** se corresponden con los vectores propios del trabajo de Lakhina *et al* [212], salvo el escalado, y por lo tanto su interpretación es similar a la que se realiza en aquel.

Pero la representación más “interesante” de los resultados obtenidos del análisis de la matriz **X01** es la del HJ-Biplot, tal y como la conocemos [160] y que constituye una de nuestras aportaciones al trabajo realizado por Lakhina *et al* [212]. Como hemos visto la metodología Biplot en general permite, en cierto modo, soslayar los problemas derivados de la aplicación directa del ACP al juego de datos bajo estudio al considerar en la representación su correlación temporal y espacial simultáneamente. Representemos en primer lugar el HJ-Biplot del primer plano factorial con los 2016 marcadores correspondientes a los sucesivos intervalos temporales y los 121 marcadores correspondientes a los diferentes emparejamientos Origen-Destino considerados.



Representación HJ-Biplot de la matriz de datos de tráfico de Abilene **X01**

Veamos qué información podemos extraer de esta nueva representación simultánea de los marcadores fila y columna de la matriz **X01** bajo estudio. Se aprecia claramente en este primer plano factorial, con un 50,4% de varianza acumulada, que la serie correspondiente a la ruta Chicago-Los Ángeles (CHIN-LOSA) es claramente "diferente" a las demás. Centrémonos en ella, por ahora. Esta serie temporal tiene la siguiente representación:

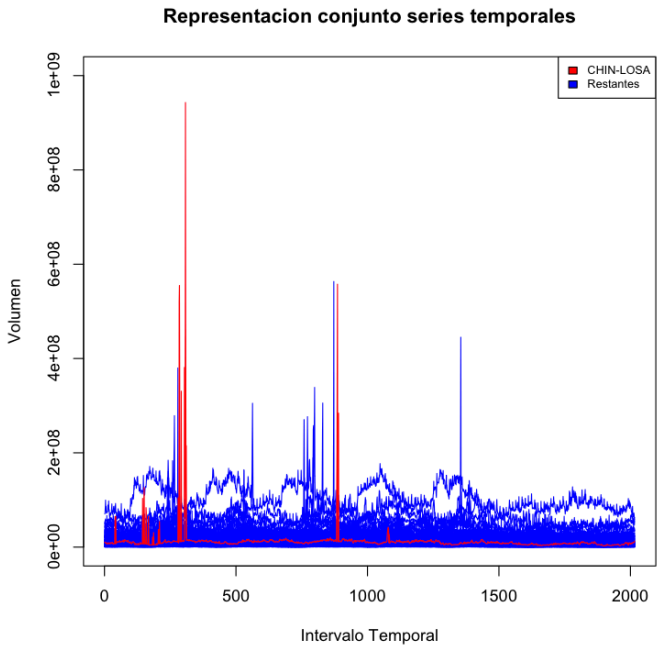


Representación de la serie temporal CHIN-LOSA

Así vista, no podemos concluir si efectivamente esta serie temporal es "diferente" por algún motivo a las otras 120 series temporales de la matriz **X01**. Solo vemos que presenta varios picos muy pronunciados sobre unos restantes valores más reducidos.

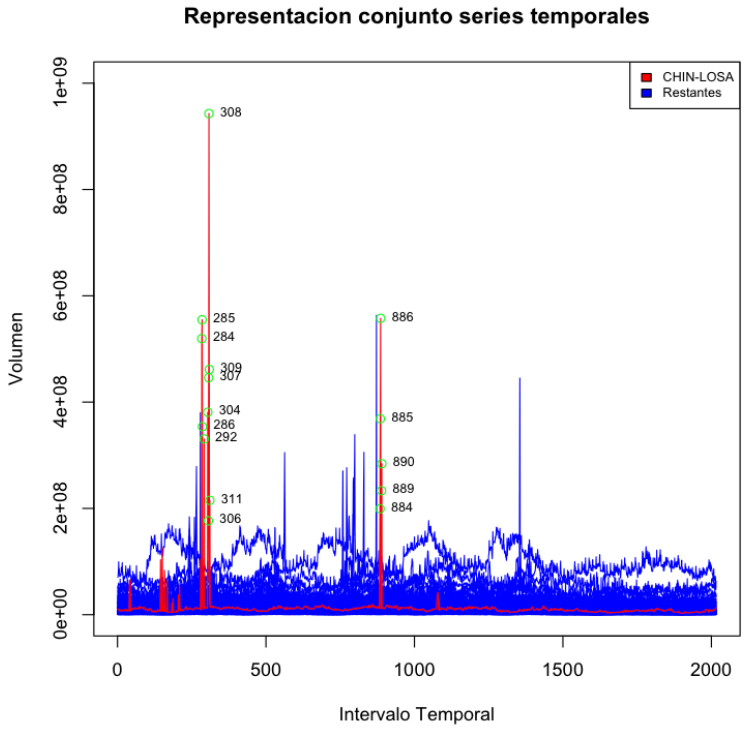
Representemos ahora superpuestas las 121 series de la matriz **X01**, pero destacando la serie correspondiente a la ruta CHIN-LOSA para compararlas visualmente. Vemos que, por simple inspección, la serie CHIN-LOSA presenta algunos máximos que destacan sobre las restantes series y que, muy posiblemente, su valor fuera de esos máximos pueda considerarse bajo en comparación a las restantes series temporales.

Si volvemos a la representación del HJ-Biplot comprobamos, mediante la proyección de los marcadores correspondientes, que en los intervalos temporales identificados como 308, 886, 285, 284, 309, 307, 304, 885, 286, 292, 890, 889, 311, 884, 306 esta serie CHIN-LOSA presenta valores elevados.



Representación simultánea de las 121 series temporales de **X01**. En rojo la serie CHIN-LOSA.

Si marcamos directamente esos puntos en la serie temporal CHIN-LOSA para ver a qué puntos concretos se corresponden, obtendremos que esos intervalos temporales se corresponden con máximos de la serie temporal bajo estudio, como podemos comprobar a continuación.

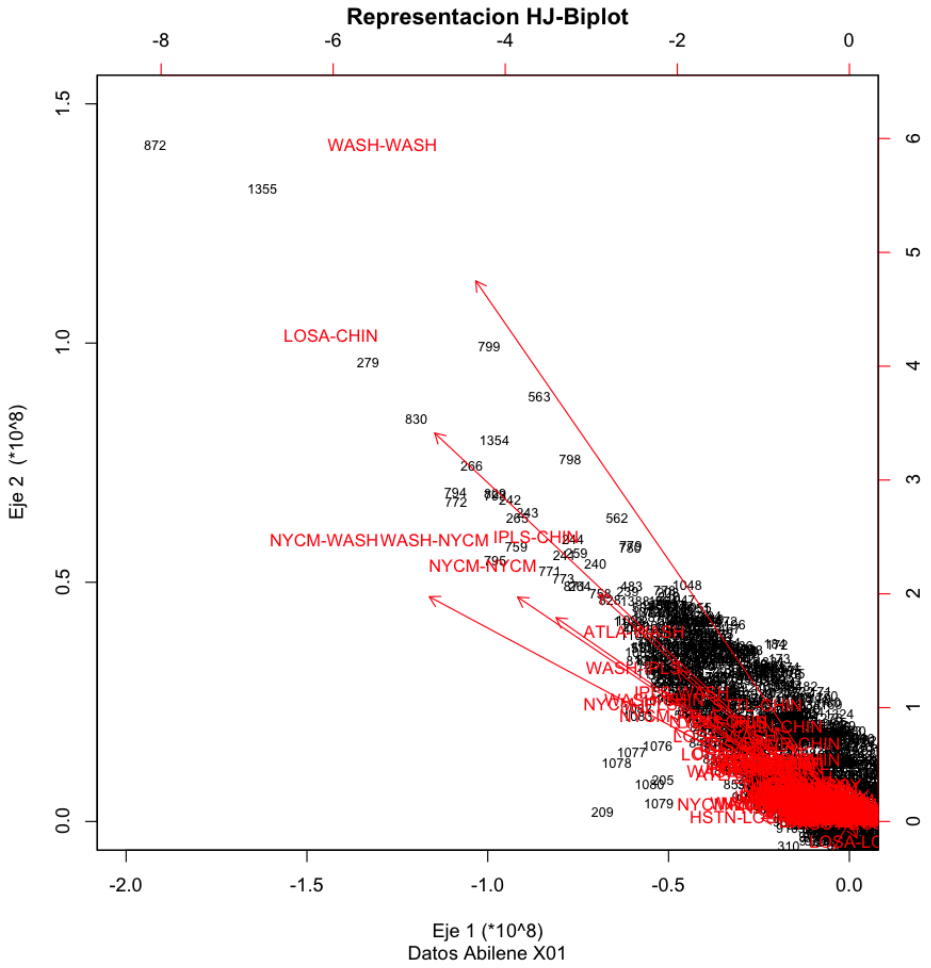


Representación simultánea de las 121 series temporales de **X01**. En rojo la serie CHIN-LOSA. En verde los marcadores **J** que presentan proyecciones elevadas sobre el marcador **H** correspondiente a CHIN-LOSA.

Es evidente que la serie CHIN-LOSA presenta los picos más pronunciados de las 121 series incluidas en la matriz **X01**, situación que hemos podido detectar simplemente inspeccionando el HJ-Biplot obtenido de la matriz de tráfico **X1** bajo estudio, sin tener que recurrir a la representación temporal específica.

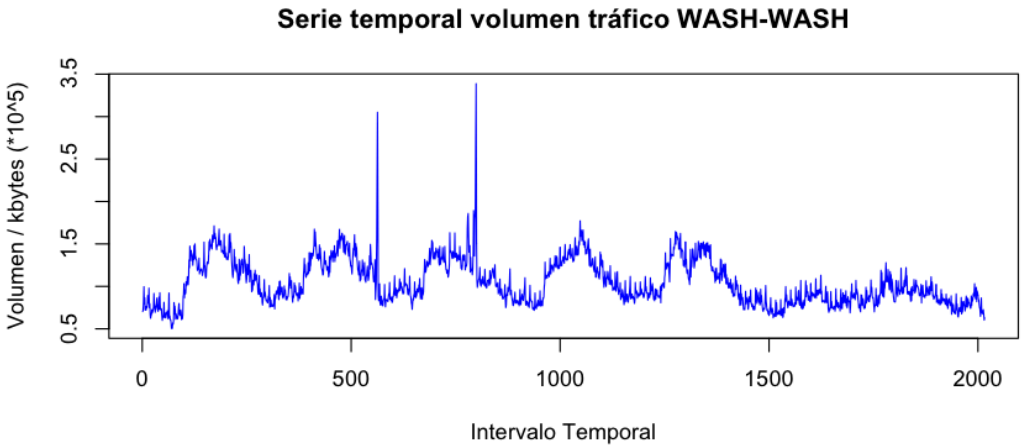
Para analizar con más detalle este primer plano factorial vamos a realizar una ampliación del segundo cuadrante de la representación HJ-Biplot anterior. Para facilitar su interpretación vamos a utilizar dos juegos de ejes diferentes, uno para los marcadores fila y otros para los marcadores columna, lo que es equivalente a un escalado de los mismos, que suele llevarse a cabo en algunos casos precisamente para facilitar esta comparación entre marcadores de diferentes magnitudes.

En esta representación HJ-Biplot otro marcador columna **H** que destaca en ese primer plano factorial es el correspondiente a la columna/serie correspondiente con la ruta Washington-Washington (WASH-WASH).



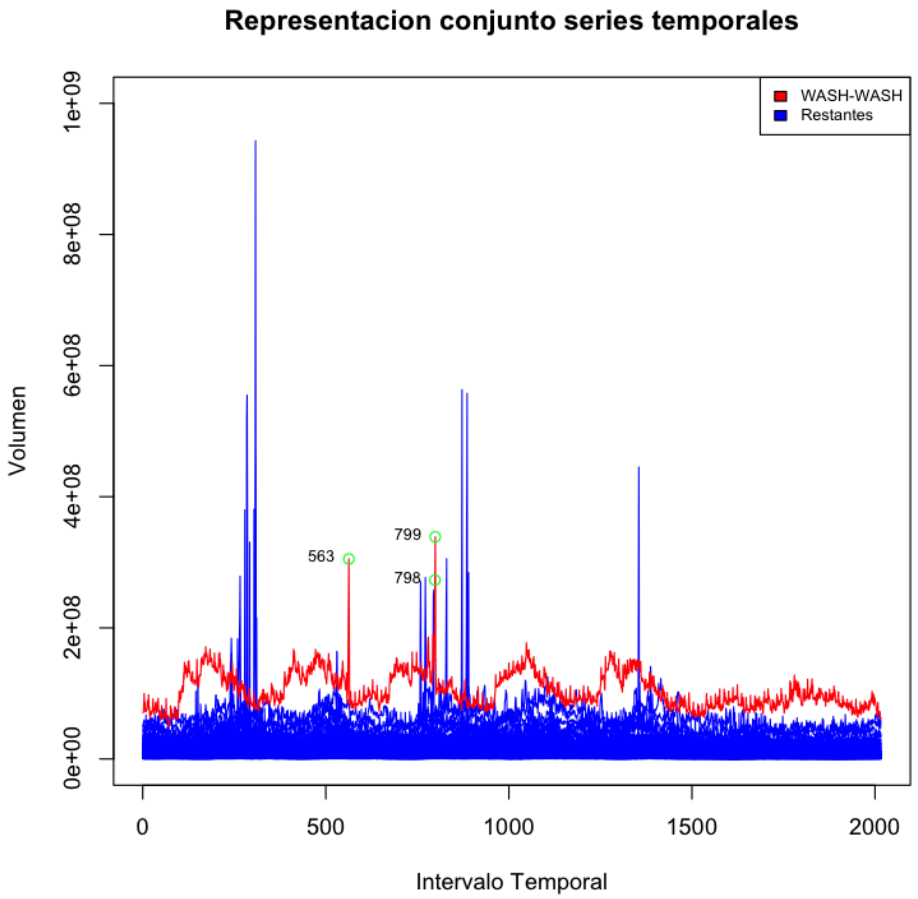
Ampliación segundo cuadrante del primer plano factorial HJ-Biplot de la matriz **X01**

Podemos representar gráficamente la serie temporal WASH-WASH de manera aislada, pero no ofrece motivos para justificar sea separación de los otros marcadores.



Representación de la serie temporal WASH-WASH (121) cuyo marcador aparece destacado en el HJ-Biplot

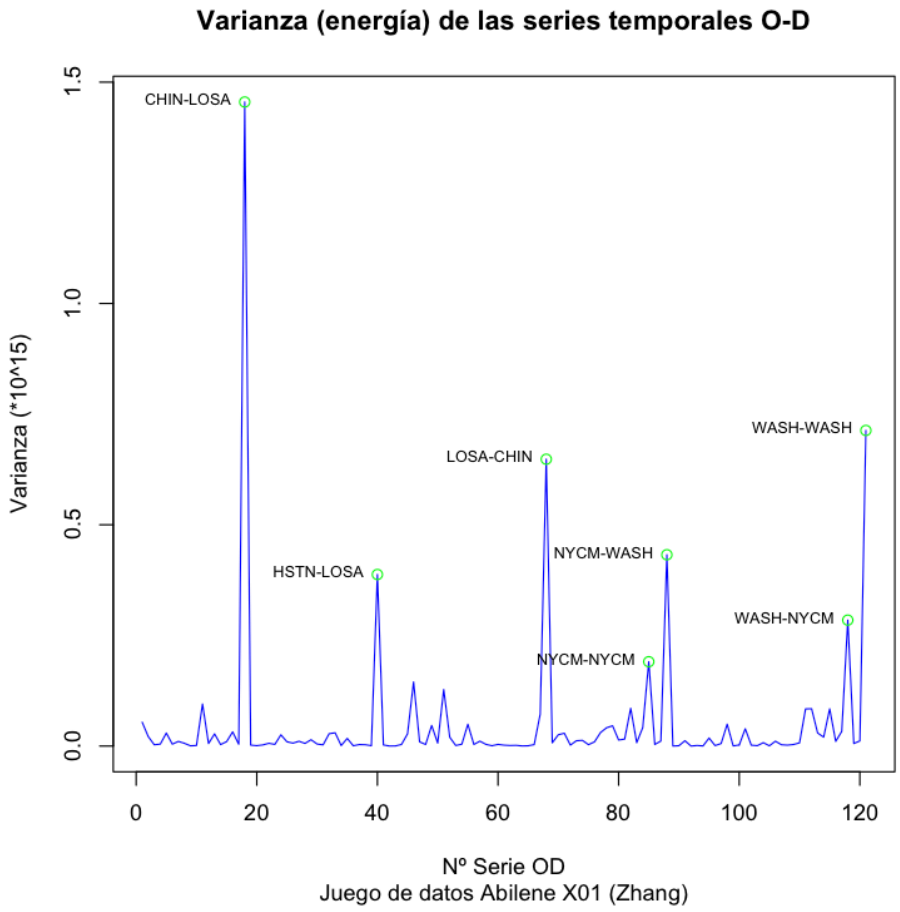
O podemos representarla conjuntamente con las 120 series temporales restantes para ver si efectivamente su comportamiento es en algún aspecto "diferente" al de las restantes.



Representación serie temporal WASH-WASH (rojo) simultáneamente a las restantes que forman la matriz **X01**.

Efectivamente comprobamos como el perfil de esta serie temporal de la ruta WASH-WASH sobresale por encima de las restantes con claridad.

Si representamos de nuevo la varianza (energía) de las 121 series con las que finalmente estamos trabajando, tras la eliminación de las 23 con elevado número de "ceros", obtendremos la representación siguiente, en la que hemos marcado los máximos locales e identificado las series a las que corresponden: CHIN-LOSA (18), WASH-WASH (121), LOSA-CHIN (68), NYCM-WASH (88), HSTN-LOSA (40) y WASH-NYCM (118), NYCM-NYCM (85).



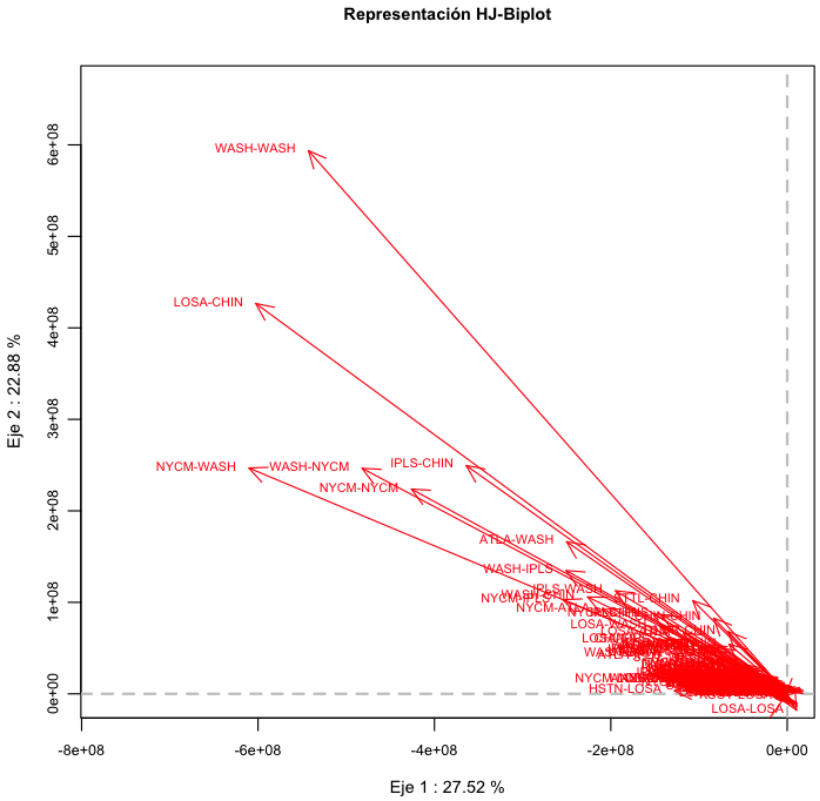
Representación de la varianza de las series temporales de la matriz **X01** identificando los máximos locales

Para facilitar la identificación de los marcadores columna en el HJ-Biplot de la matriz **X01** vamos a ocultar los marcadores fila y a ampliar el segundo cuadrante. Identificamos los marcadores columna **H** que presentan módulos más elevados en este primer plano como CHIN-LOSA (18), WASH-WASH (121), LOSA-CHIN (68), NYCM-WASH(88), WASH-NYCM (118), NYCM-NYCM (85), IPLS-CHIN (46).

La principal diferencia entre ambas relaciones es la aparición en el listado de series con mayores energía de la ruta HSTN-LOSA (20) que no se detecta entre los vectores

con mayores módulos de este primer plano. Retengamos esta situación incoherente, a la que encontraremos explicación en el siguiente punto de nuestro análisis.

Así pues, como es lógico por las propiedades del HJ-Biplot, es posible identificar visualmente las series temporales con mayores energías (varianzas) a través del módulo del vector que las representa en el HJ-Biplot.

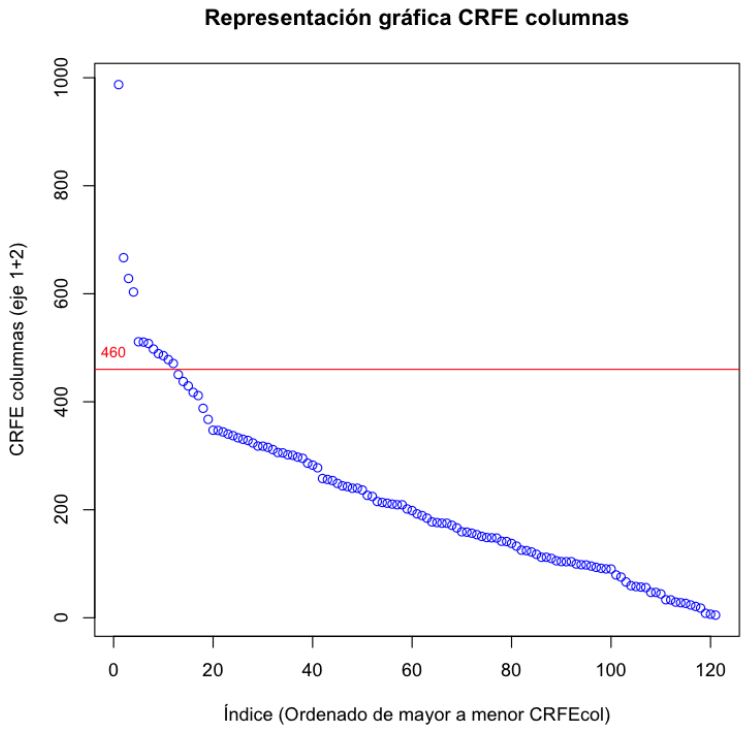


Representación de los marcadores columna del HJ-Biplot de la matriz **X01**.

Un aspecto importante que se ha considerado muy por encima hasta esta fase de nuestro análisis es la "calidad de la representación" obtenida. El ACP ofrece como único indicador destacable de la "calidad de representación" el porcentaje de la inercia/varianza absorbida, en el primer plano factorial, por ejemplo. La representación Biplot también nos permite utilizar, como indicativo de la calidad de representación, la suma de los porcentajes de inercia absorbida en los ejes factoriales participantes en el gráfico como medida global de la calidad de la representación obtenida. Sin embargo, y este elemento constituye otra de las mejoras de nuestra propuesta, la representación Biplot, como vimos en su momento, nos permite obtener un indicador de la calidad de representación o "bondad de ajuste" individual de cada uno de los marcadores fila y columna que la componen. Por si eso no fuera suficiente, además la representación HJ-Biplot, de entre los biplots "clásicos", ofrece la máxima calidad de representación conjuntamente para ambos marcadores fila y columna. Esta calidad de representación

individual de cada marcador se obtiene, como se expuso en su momento, a través de la denominada Contribución Relativa del Factor al Elemento (CRFE).

Comencemos por analizar las CRFE de los marcadores correspondientes a las columnas (series temporales) del primer plano factorial, como suma de las CRFE de dos ejes con mayores inercia absorbida, y ordenados de mayor a menor:



Representación gráfica CRFE de las columnas ordenado de mayor a menor.

Observamos un pequeño salto en el nivel 460 que, en principio, estableceremos como punto de corte, ya que si bien aparece un salto previo en el nivel 600, ese posible punto de corte nos deja tan solo con 4 series temporales (CHIN-LOSA, IPLS-CHIN, NYCM-ATLA y NYCM-NYCM) con una CRFE superior a dicho nivel.

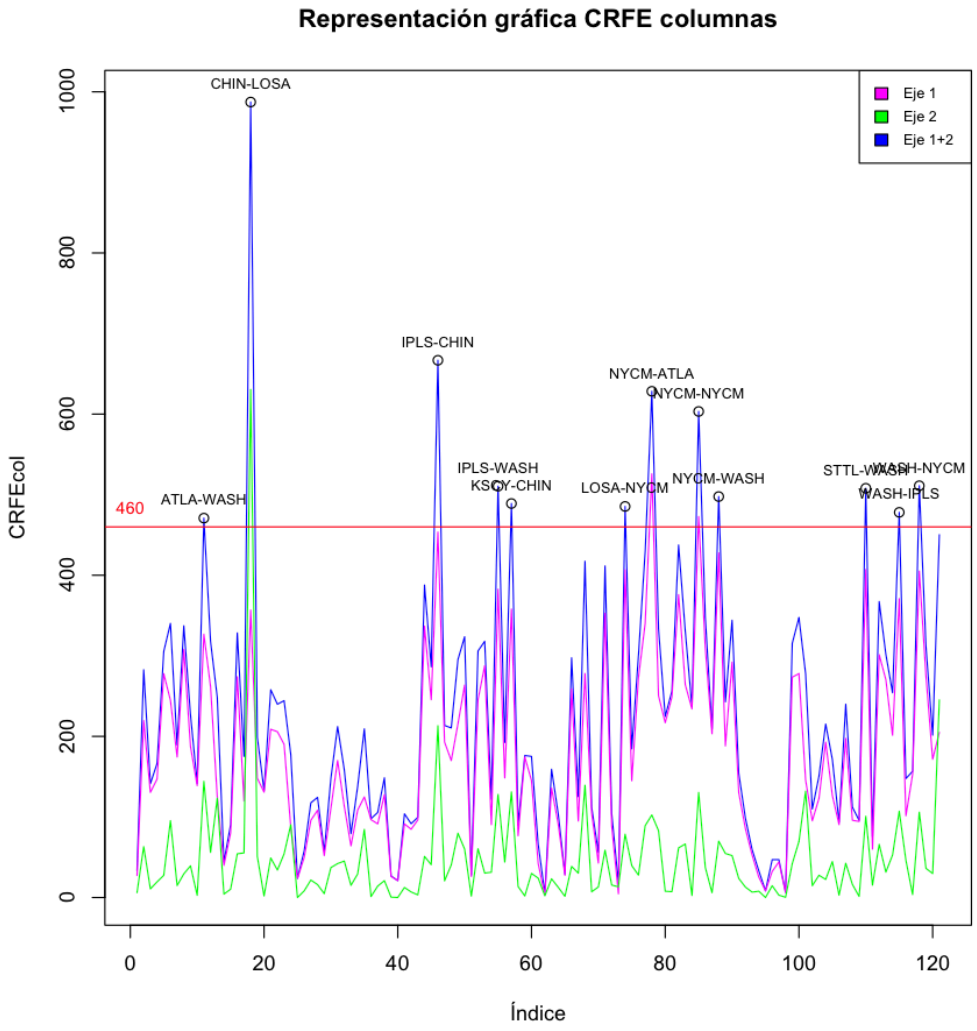
Así, los 12 marcadores columna con CRFE totales en el primer plano factorial superiores a 460 son los correspondientes a las siguientes series:

Columna	CRFEcol eje 1	CRFEcol eje 2	CRFEcol eje 1+2
CHIN-LOSA	357	631	987 (*)
IPLS-CHIN	454	213	667
NYCM-ATLA	526	102	628
NYCM-NYCM	473	130	603
WASH-NYCM	405	106	511
IPLS-WASH	383	128	511
STTL-WASH	407	101	508
NYCM-WASH	428	70	498
KSCY-CHIN	358	131	489
LOSA-NYCM	407	79	485 (*)

Columna	CRFEcol eje 1	CRFEcol eje 2	CRFEcol eje 1+2
WASH-IPLS	371	107	478
ATLA-WASH	327	144	471

(*) Sujeto a redondeo decimal

Podemos representar los valores de las CRFE del primer y segundo eje factorial, así como del primer plano factorial, para analizar su perfil e identificar los marcadores con contribuciones superiores al umbral establecido en 460.



Representación gráfica de la CRFE de las columnas con identificación de marcadores más representativos.

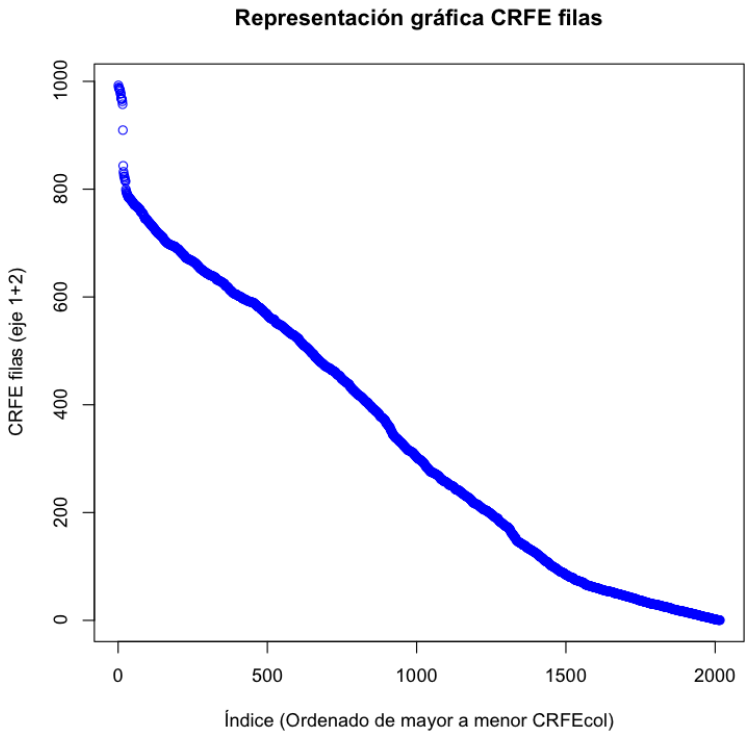
Si recuperamos la tabla anterior en la que identificábamos en el primer plano factorial las series cuyos marcadores aparecían con comportamientos “diferentes” y la cruzamos con las series que presentan una “buena” calidad de representación (establecida arbitrariamente en el nivel 460) obtenemos que los marcadores correspondientes a las series LOSA-CHIN y WASH-WASH tienen una Calidad de Representación CRFE de 417 y 450 respectivamente, y se encuentran en las posiciones 16ª y 13ª (recordemos que hemos dado por “buenas” las Calidades de Representación de los marcadores correspondientes a las 12 primeras series

temporales). Así pues, las series identificadas como “extrañas” presentan una calidad de representación “buena” o “aceptable” en el primer plano factorial bajo estudio.

Identificadas visualmente	**CRFE “buena” (CRFE)**
CHIN-LOSA (#18)	CHIN-LOSA (987)
IPLS-CHIN (#46)	IPLS-CHIN (667)
	NYCM-ATLA (628)
NYCM-NYCM (#85)	NYCM-NYCM (603)
WASH-NYCM (#118)	WASH-NYCM (511)
	IPLS-WASH (511)
	STTL-WASH (508)
NYCM-WASH (#88)	NYCM-WASH (498)
	KSCY-CHIN (489)
	LOSA-NYCM (485)
	WASH-IPLS (478)
	ATLA-WASH (471)
WASH-WASH (#121)	*WASH-WASH (450) Posición #13*
LOSA-CHIN (#68)	*LOSA-CHIN (417) Posición #16*

Recuperemos en este punto la situación que se nos presentó cuando analizábamos la varianza/energía de las series bajo estudio y su relación con los módulos de los marcadores Biplot correspondientes: la ruta HSTN-LOSA (20), que presentaba unos valores de energía elevados no se detectaba como tal en la representación HJ-Biplot. Ahora sabemos el motivo de esa incoherencia: su calidad de representación, en ese primer plano factorial del HJ-Biplot, su CRFE presenta un valor tan reducido como 20.

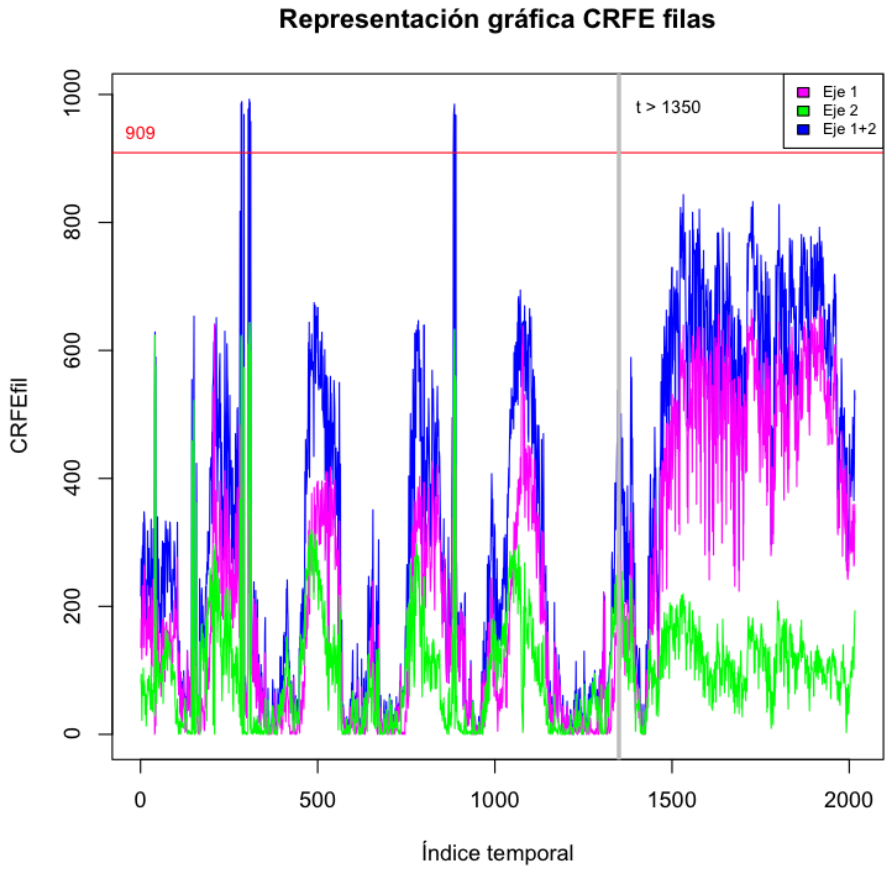
El análisis de la Calidad de Representación para los marcadores correspondientes a las filas **J** (intervalos temporales) es un poco menos aclaratorio que el precedente, fruto del elevado número de elementos a considerar. Si representamos las CRFE de las filas ordenadamente obtenemos el siguiente gráfico:



Representación gráfica de la CRFE de las filas ordenadas de mayor a menor.

La elección, siempre subjetiva, de un nivel de corte para determinar la “buena” calidad de representación es aquí aún más complicada. Solo se aprecia un “codo” alrededor del nivel 800 y un salto en el marcador correspondiente al intervalo temporal 883 con una CRFE de 910. No obstante es importante destacar el elevado nivel de CRFE que estaríamos considerando, en comparación con el criterio establecido para los marcadores columna (460).

Si representamos la CRFE para las filas, con los índices ordenados temporalmente obtenemos la siguiente representación:

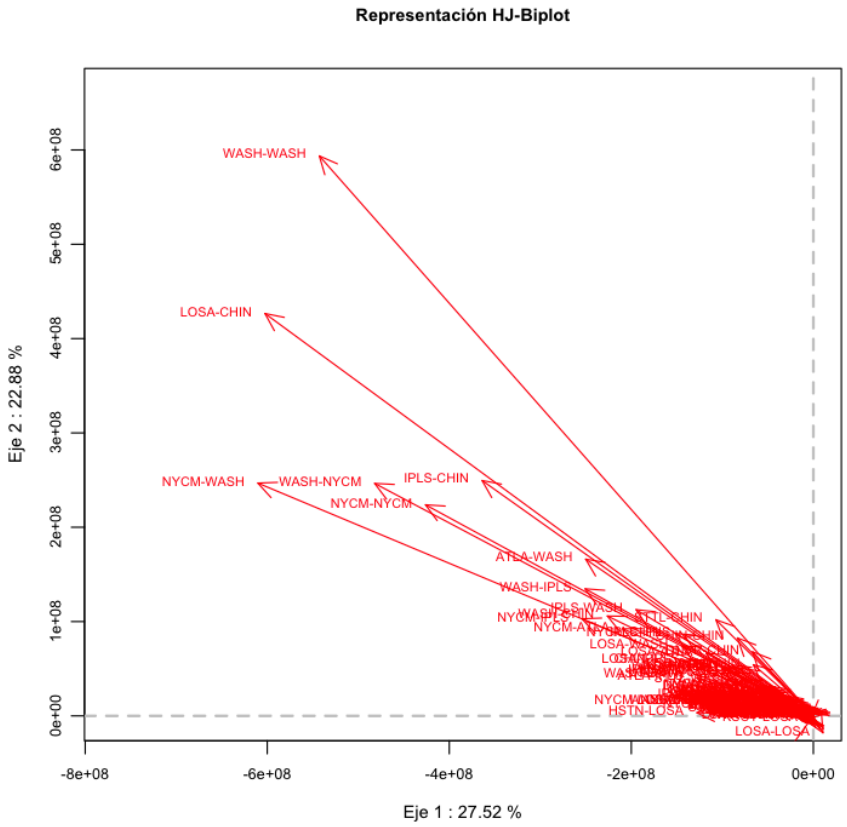


Representación gráfica de la CRFE de las filas identificando los marcadores más representativos.

Con el nivel de corte en 909 obtendríamos solo 16 marcadores fila (de un total de 2016) con una "buena" calidad de representación. Si establecemos el nivel en 460, como en el criterio adoptado en el caso de las columnas, el número de marcadores con "buena" calidad de representación ascendería a 728.

De la representación gráfica de la CRFE de la filas llaman la atención visualmente dos efectos: en primer lugar una aparente "periodicidad" hasta, aproximadamente, el intervalo temporal nº 1350, y en segundo lugar la ruptura de dicha periodicidad a partir de ese punto, acentuándose además el carácter preponderante del primer eje factorial en dicha fase temporal.

Pero la representación HJ-Biplot nos aporta aún más información sobre las series temporales, reflejando la correlación/covarianza entre ellas. Si recuperamos una representación anterior, en la que se mostraba la ampliación del segundo cuadrante del primer plano factorial del HJ-Biplot de la matriz X01, vemos como existe un ángulo pequeño entre los marcadores de las series LOSA-CHIN (68) y IPLS-CHIN (46) y los marcadores de las series WASH-NYCM (118) y NYCM-NYCM (85).

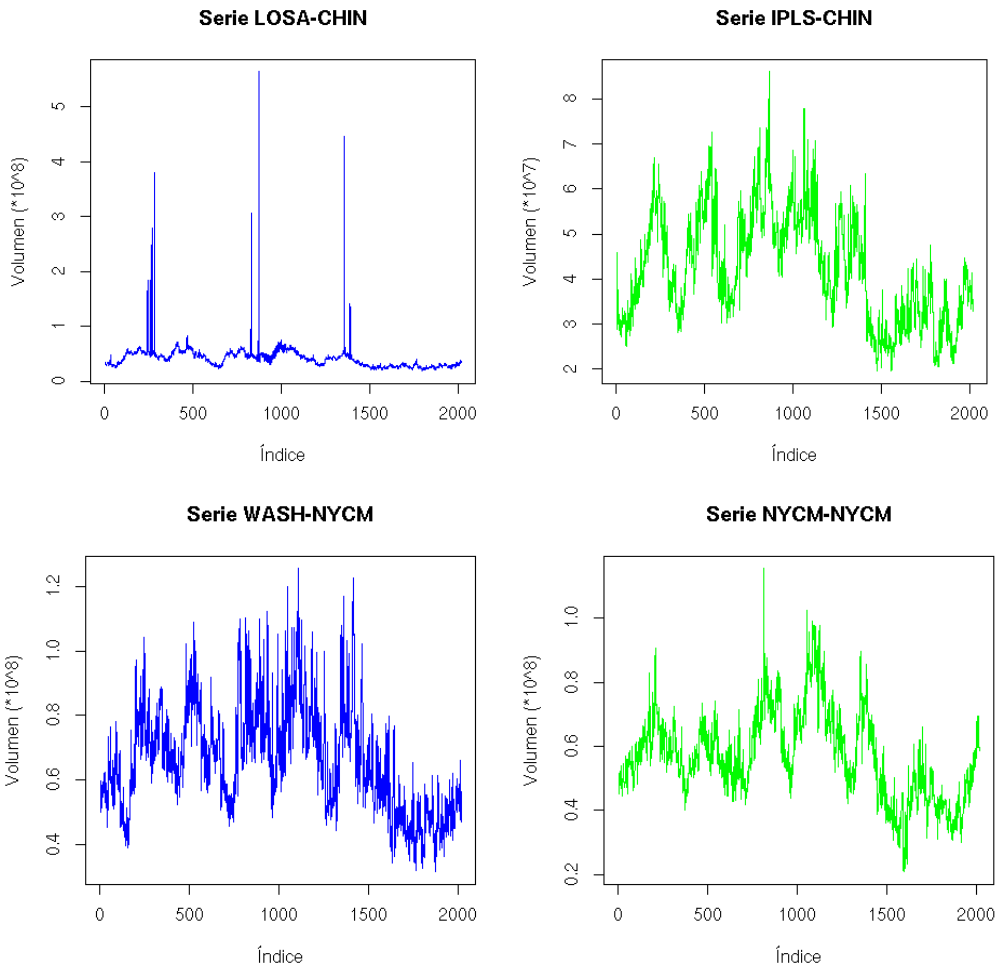


Representación HJ-Biplot (ampliación del segundo cuadrante)

Como se vio en el apartado en el que se expusieron las propiedades del HJ-Biplot, la covariación entre las variables (series temporales) puede estimarse observando en ángulo que forman los respectivos marcadores, estando el coseno del ángulo que forman en relación directa con la covariación entre dichas variables. Esto es, a menor ángulo entre los marcadores, mayor covariación entre las variables correspondientes.

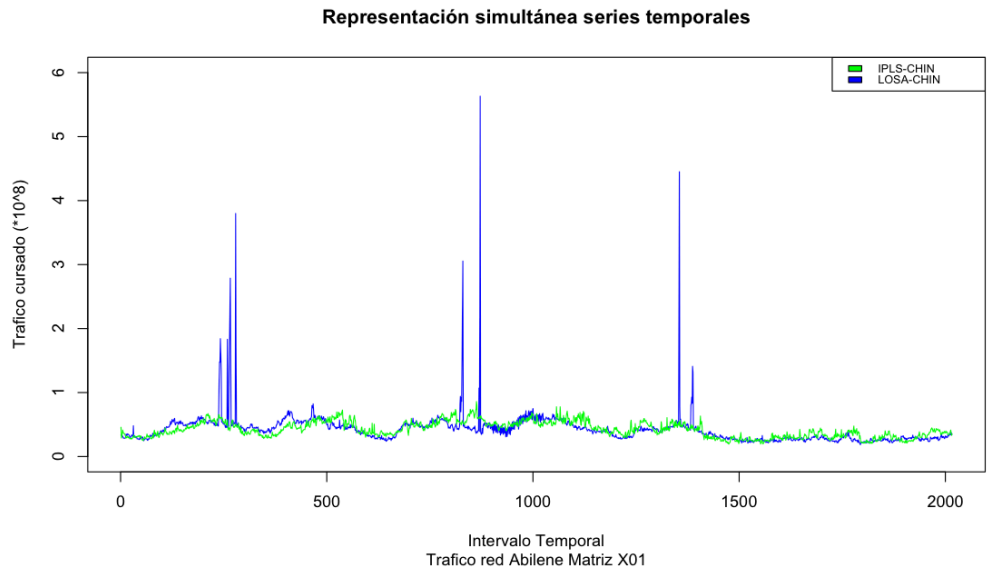
Como ya se ha expuesto anteriormente existe una clara diferencia conceptual entre la correlación/covariación “estadística” entre las variables y la correlación como función temporal entre las series temporales. Este último concepto implica, como hemos visto, una mayor capacidad para “detectar” parecidos entre series temporales, por ejemplo en el caso de desplazamientos temporales. Así pues, quizá una buena manera de validar nuestra hipótesis de “similitud” entre las series temporales con mayor covariación (menor ángulo entre marcadores) sea procediendo a su representación.

A continuación se representan individualmente las series temporales correspondientes a las rutas LOSA-CHIN, IPLS-CHIN, WASH-NYCM y NYCM-NYCM.

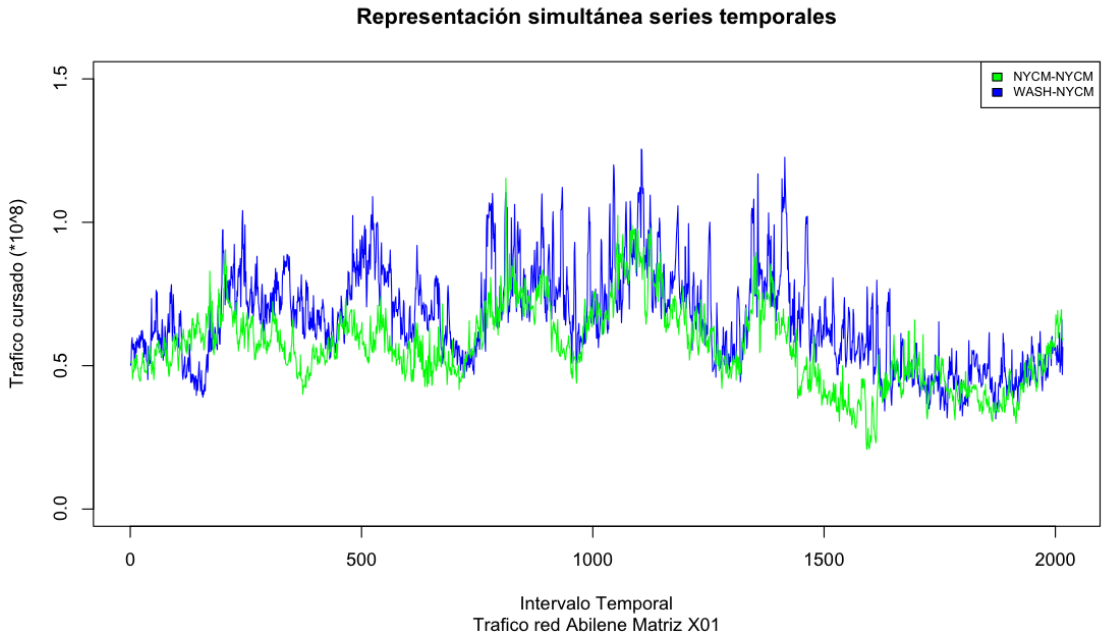


Representación de las series temporales LOSA-CHIN, IPLS-CHIN, WASH-NYCM, NYCM-NYCM.

Pero la mejor representación que nos permite efectuar la comparación buscada entre las dos parejas de series temporales es, probablemente, su representación conjunta:

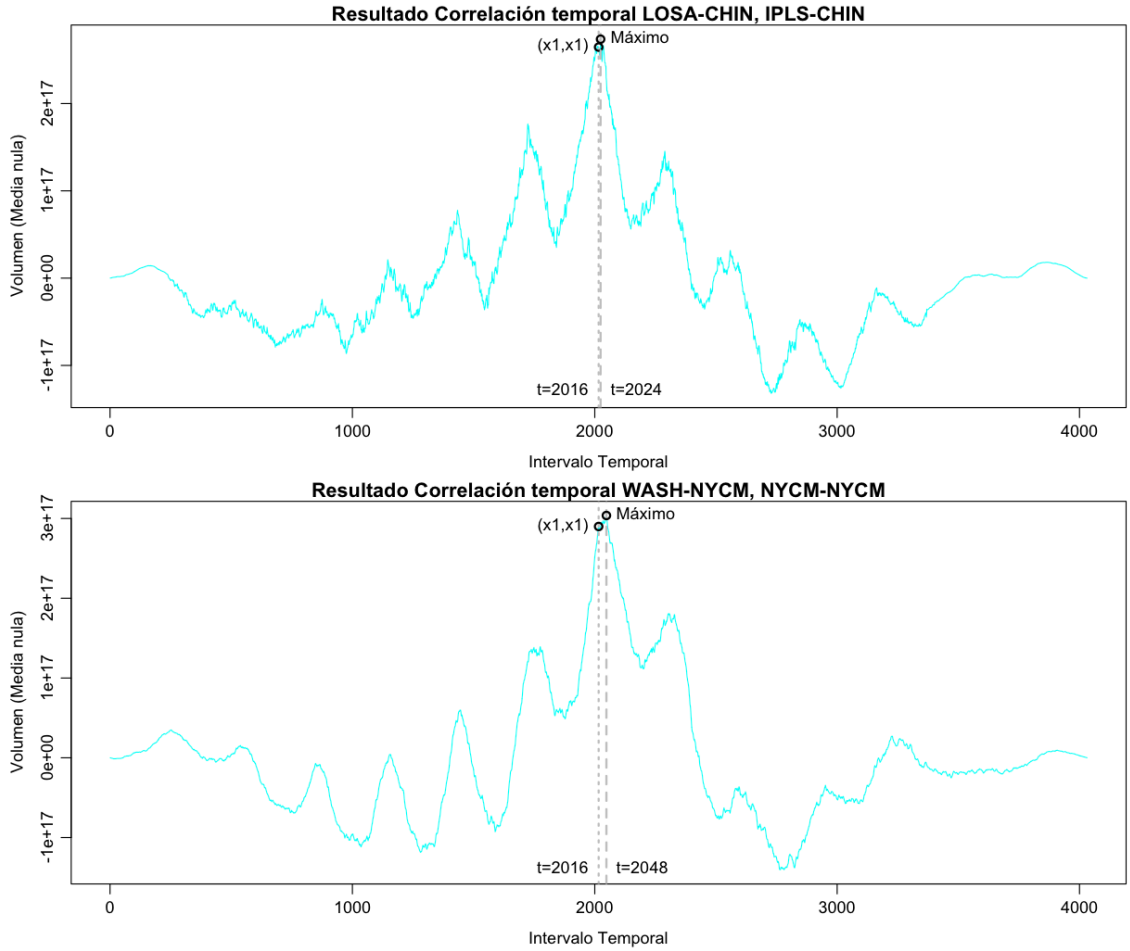


Representación conjunta de las series temporales IPLS-CHIN y LOSA-CHIN.



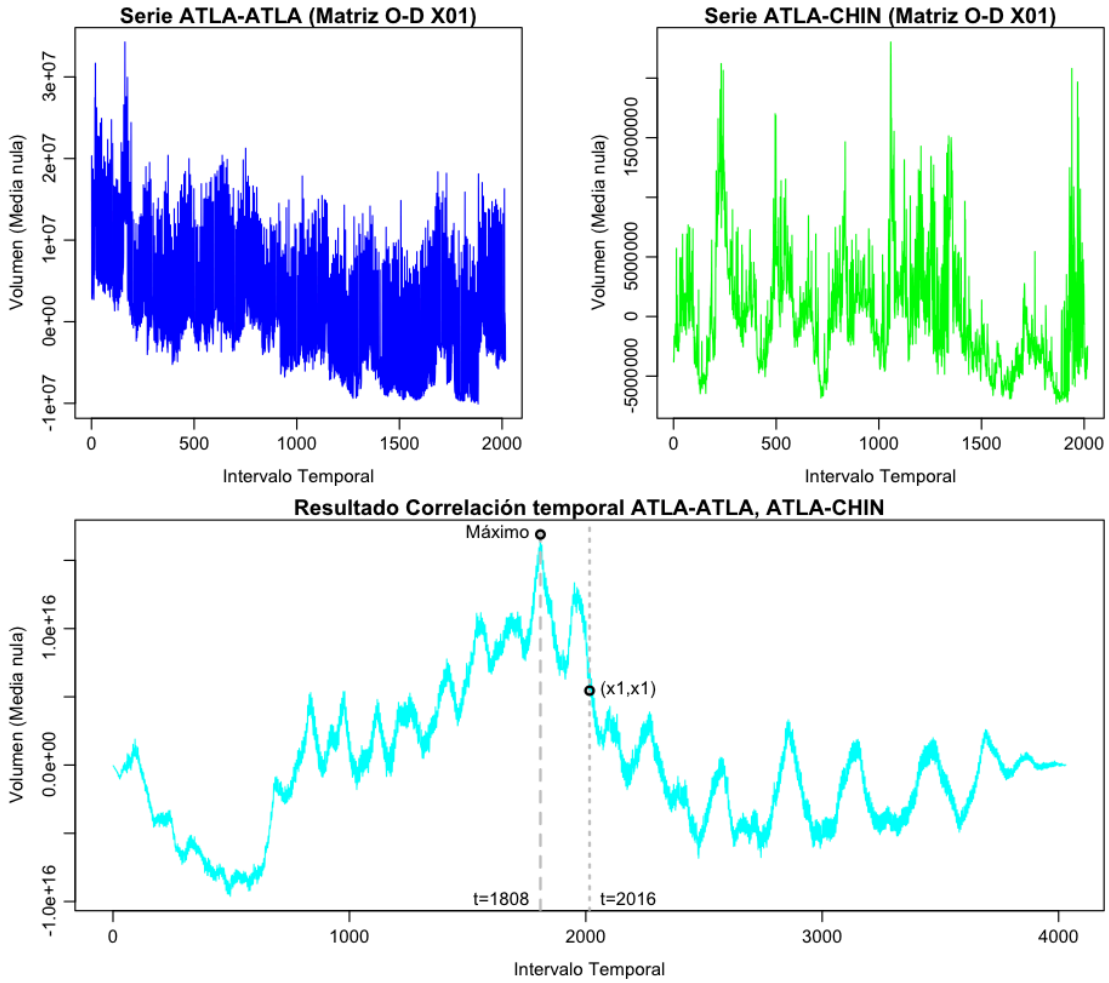
Representación conjunta de las series temporales NYCM-NYCM y WASH-NYCM.

En el primer caso, a excepción de los picos de la serie LOSA-CHIN, ambas series temporales se asemejan visualmente bastante. Al igual sucede en el segundo caso.



Resultado correlaciones temporales entre las series LOSA-CHIN, IPLS-CHIN y WASH-WASH, NYCM-NYCM

La representación de las correlaciones temporales entre las series LOSA-CHIN, IPLS-CHIN y WASH-WASH, NYCM-NYCM nos permite visualizar el efecto al que hacíamos referencia: el máximo de la correlación temporal, que identifica las similitudes entre las series temporales no coincide con el valor del producto interior de las referidas series. El valor del producto interior se corresponde con el punto central de la correlación entre las series, pero el máximo de dicha correlación no tiene porque estar en esa posición, excepto en el caso de las autocorrelaciones. Aunque en los dos casos visualizados ni la situación de los máximos ni los mismos valores máximos se alejan demasiado, esto puede no ser siempre así.

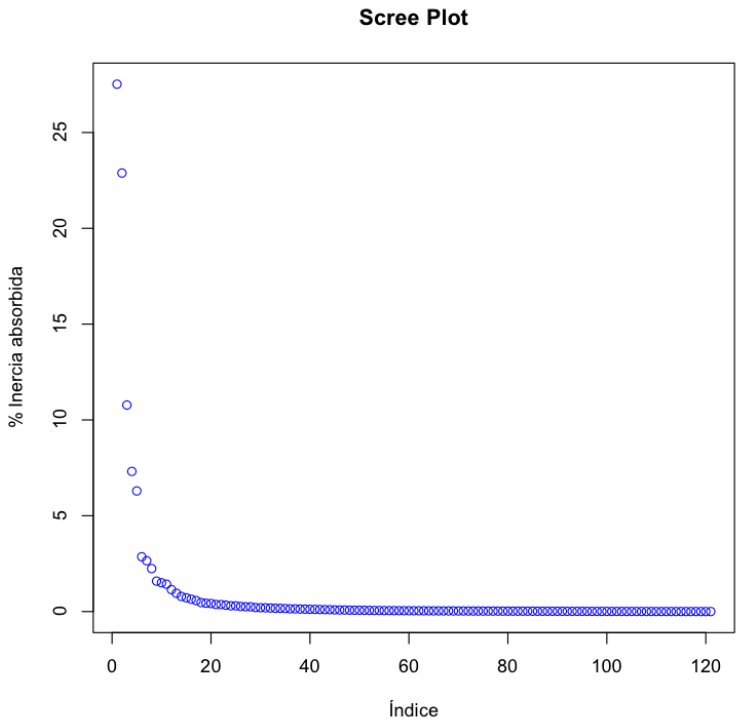


Representación de las series temporales ATLA-ATLA y ATLA-CHIN y su correlación temporal

Este planteamiento podría darnos lugar a pensar en una nueva descomposición de la matriz de datos utilizando el máximo de esta función de correlación, calculando la matriz de correlación temporal para $\mathbf{X}$ y $\mathbf{X}^T$ y para $\mathbf{X}^T$ y $\mathbf{X}$, y siguiendo el método original obtendríamos la descomposición en autovalores y autovectores de cada una de la matrices. El problema que existe es que nada nos garantiza que los autovalores en ambos casos sean los mismos, por lo que el planteamiento general se viene abajo.

Vamos a analizar ahora brevemente qué sucede en otros planos factoriales.

Como ya hemos indicado en el primer plano factorial se absorbe el 50,41% de la inercia total. El *scree plot* representado a continuación nos indica que con las 5 primeras dimensiones retendríamos casi el 75% de la inercia total.



Scree-plot de la matriz de tráfico **X01** de la red Abilene

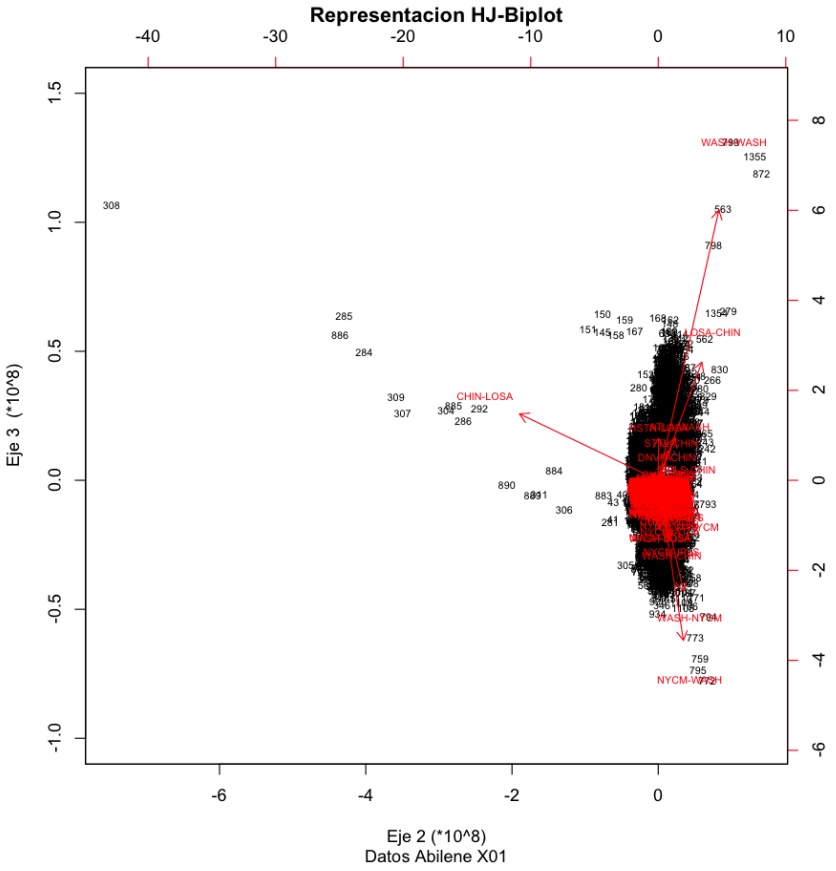
Vamos a representar el HJ-Biplot de los planos factoriales 2-3 y 3-4.

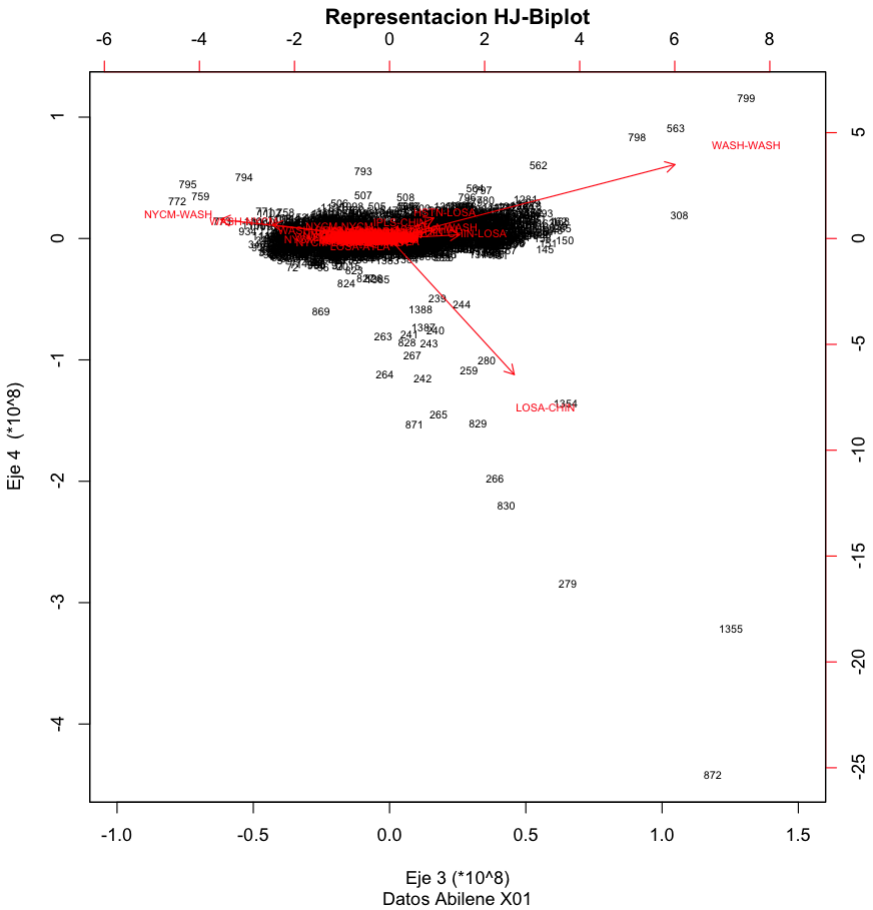
En la representación del plano factorial 2-3 se aprecia un comportamiento diferente del resto de los marcadores en los correspondientes a las serie WASH-WASH, CHIN-LOSA, LOSA-CHIN, WASH-NYCM y NYCM-WASH. Mientras que en el plano 3-4 un diferente comportamiento puede ser atribuido a las series temporales correspondientes al tráfico en las rutas NYCM-WASH, WASH-WASH y LOSA-CHIN.

Plano 1-2	Plano 2-3	Plano 3-4
CHIN-LOSA (18)	CHIN-LOSA	
WASH-WASH (121)	WASH-WASH	WASH-WASH
LOSA-CHIN (68)	LOSA-CHIN	LOSA-CHIN
NYCM-WASH (88)	NYCM-WASH	NYCM-WASH
WASH-NYCM (118)	WASH-NYCM	
NYCM-NYCM (85)		
IPLS-CHIN (46)		

Tabla identificando las series temporales con comportamientos “diferentes” en los 3 primeros planos factoriales

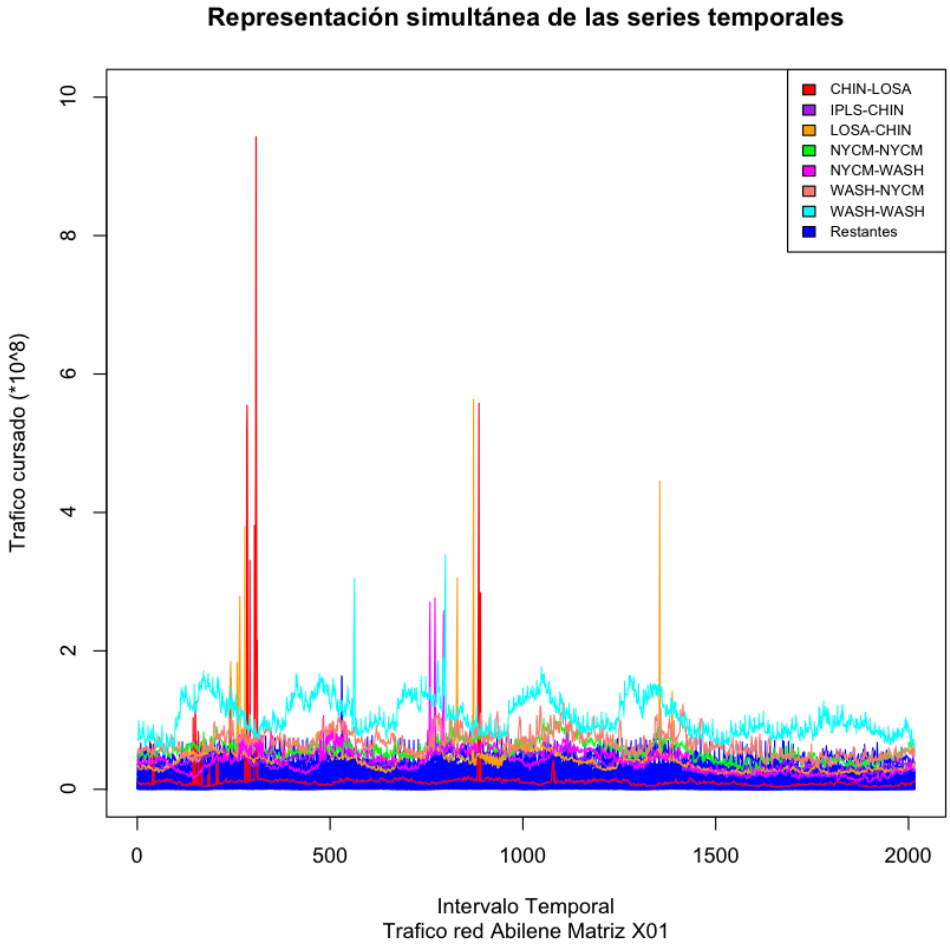
Así pues, la información que nos ofrecen los planos 2-3 y 3-4 no es muy diferente de la que ya nos había aportado el primer plano factorial estudiado, 1-2.





Representaciones HJ-Biplot de los planos factoriales 2-3 y 3-4

Vamos a representar ahora todas las series temporales conjuntamente, pero destacando aquellas que hemos identificado como “diferentes” en el primer plano factorial del HJ-Biplot.

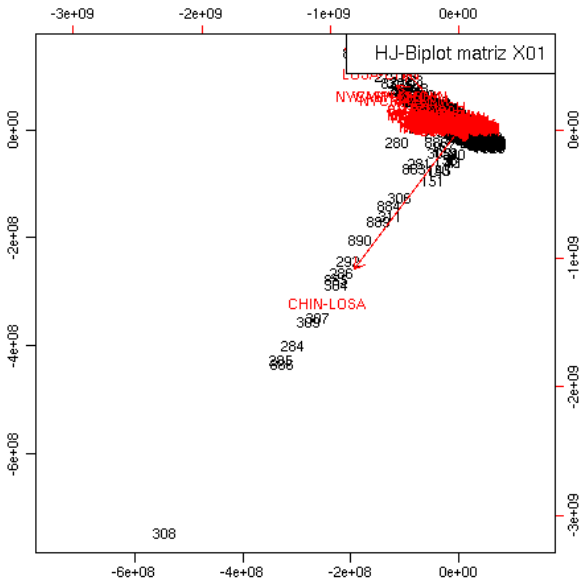


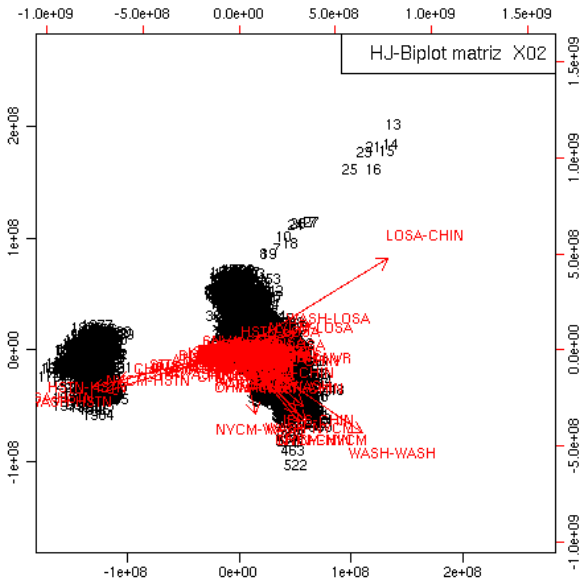
Representación de las series temporales con comportamientos identificados como diferentes en el HJ-Biplot.

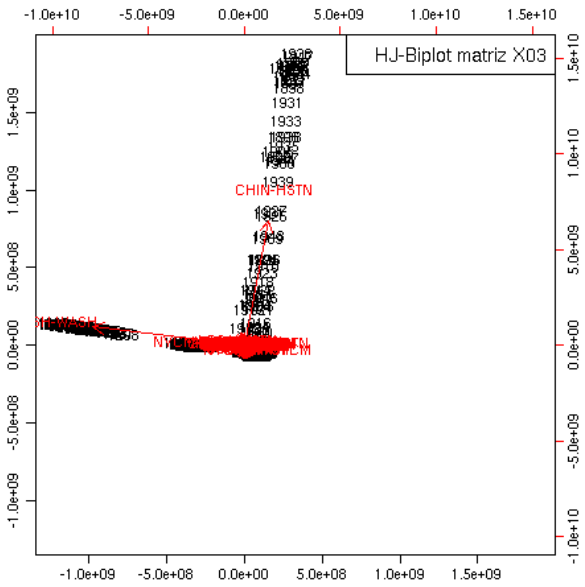
Visualmente podemos comprobar, con las obvias limitaciones del procedimiento, que las series identificadas con comportamientos diferentes en la representación HJ-Biplot sí que “destacan” de alguna manera sobre las restantes: bien por sus valores máximos, bien por los picos que presentan, o bien por su energía. Así pues, es posible concluir que la representación HJ-Biplot puede constituir una positiva ayuda en el estudio de series temporales, permitiendo de forma sencilla identificar diferentes situaciones en las mismas, tales como picos, máximos y energía contenida en la serie.

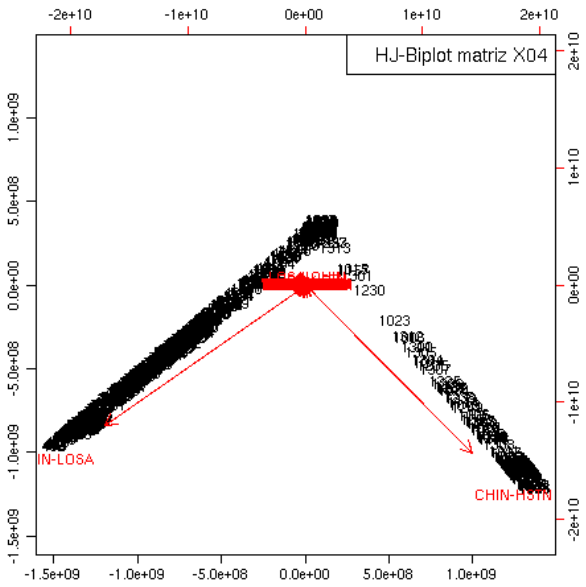
6.4.4 APLICACIÓN DEL HJ-BIPLOT A LOS 24 JUEGOS DE DATOS

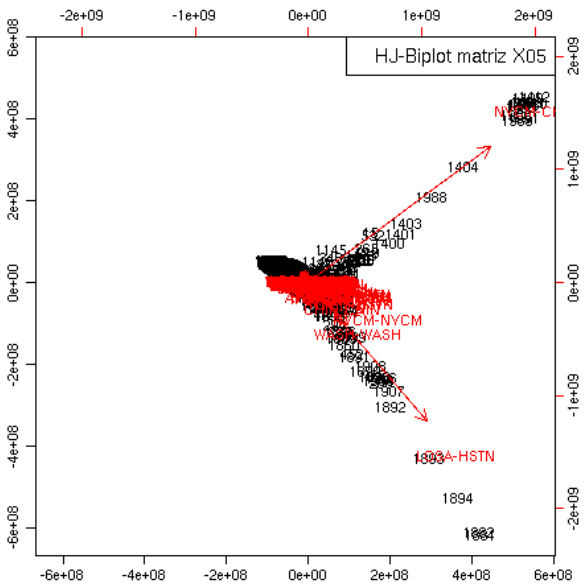
Como expusimos anteriormente, Zhang [305] publicó 24 matrices con la características descritas. Vamos a realizar, de manera sistemática, un HJ-Biplot de todas ellas para ver si es posible extraer alguna conclusión sobre el comportamiento de la red Abilene a lo largo del tiempo.

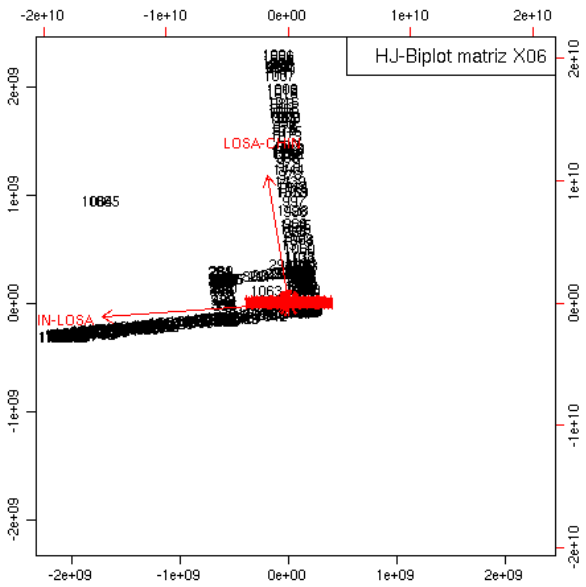

HJ-Biplot matriz X01
CHIN-LOSA
308

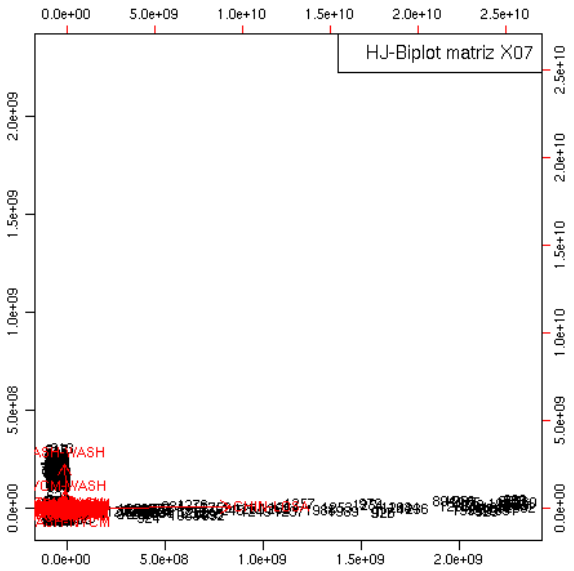

HJ-Biplot matriz X02
LOSA-CHIN
WASH-WASH

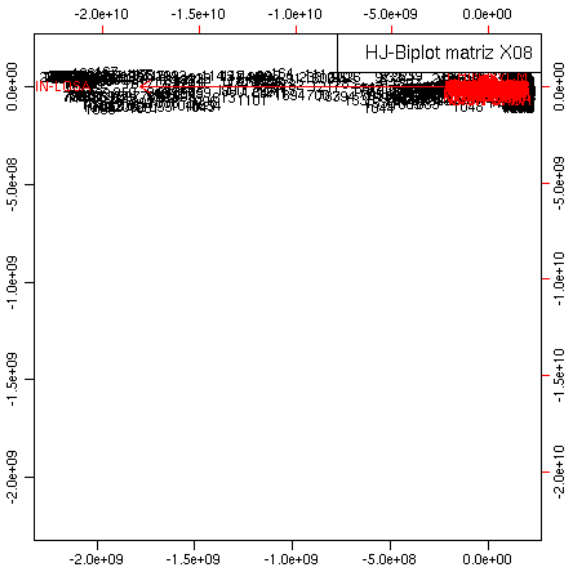

HJ-Biplot matriz X03
CHIN-HSTN

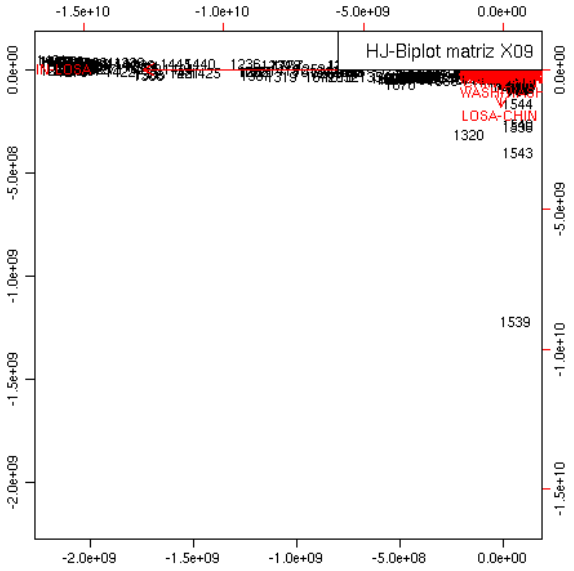

HJ-Biplot matriz X04
IN-LOSA
CHIN-HSTN

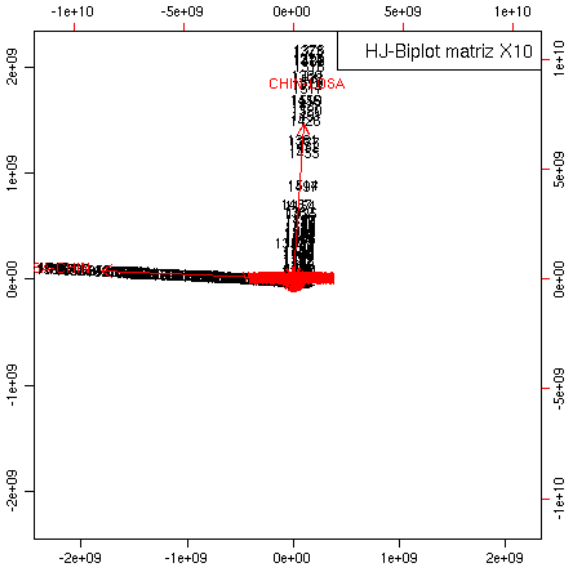

HJ-Biplot matriz X05
CM-NYCM
WASH
A-HSTN

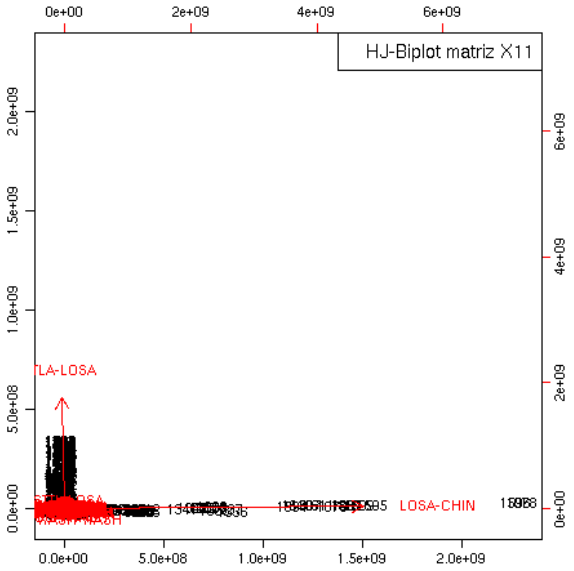

HJ-Biplot matriz X06
IN-LOSA

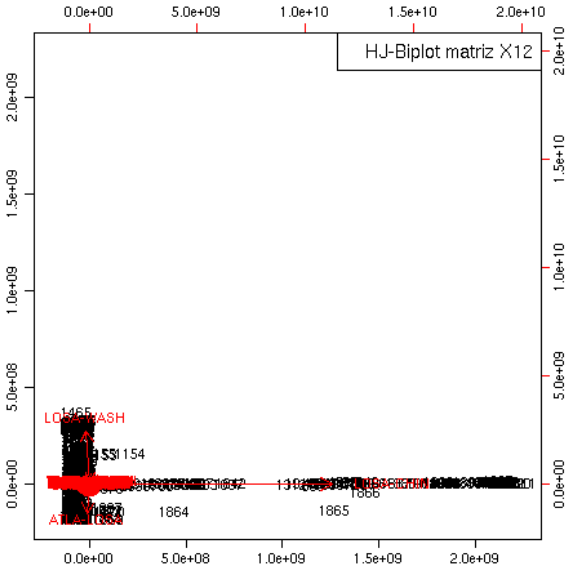

HJ-Biplot matriz X07
0.0e+00
5.0e+09
1.0e+10
1.5e+10
2.0e+10
2.5e+10
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09

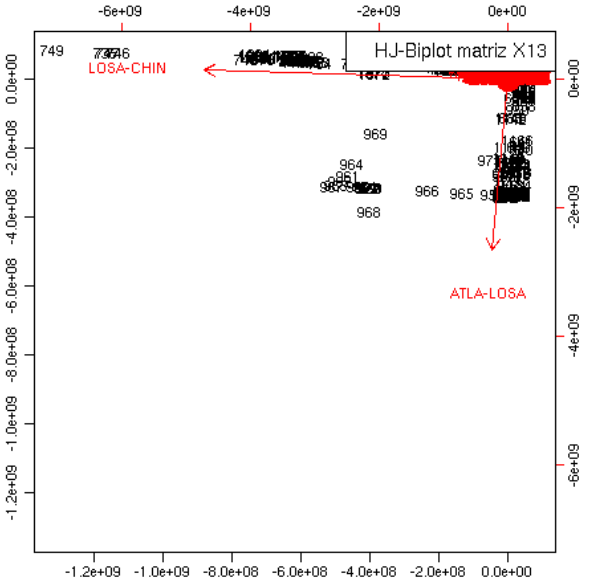

HJ-Biplot matriz X08
-2.0e+10
-1.5e+10
-1.0e+10
-5.0e+09
0.0e+00
-2.0e+09
-1.5e+09
-1.0e+09
-5.0e+08
0.0e+00

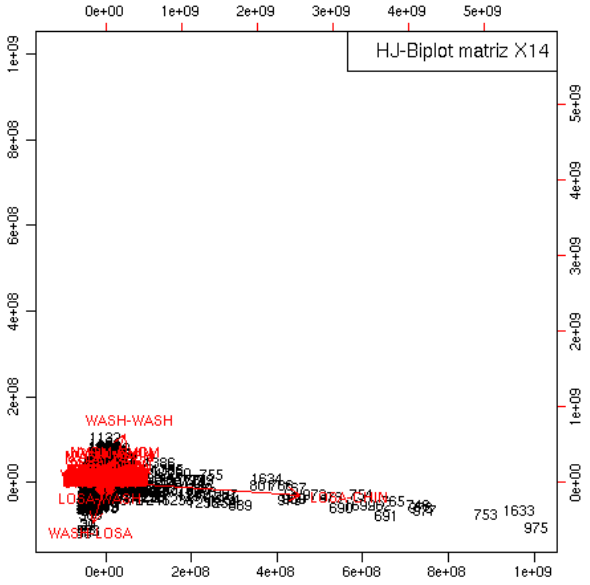

HJ-Biplot matriz X09
-1.5e+10
-1.0e+10
-5.0e+09
0.0e+00
-2.0e+09
-1.5e+09
-1.0e+09
-5.0e+08
0.0e+00
LOSA-CHIN
1320
1543
1539

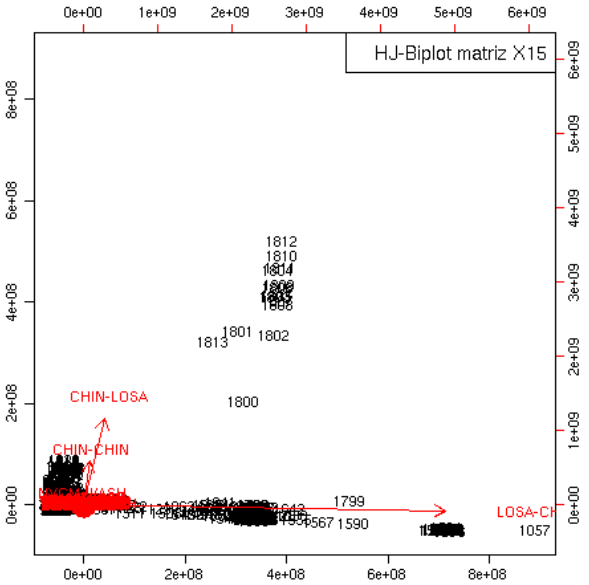

HJ-Biplot matriz X10
-1e+10
-5e+09
0e+00
5e+09
1e+10
-2e+09
-1e+09
0e+00
1e+09
2e+09

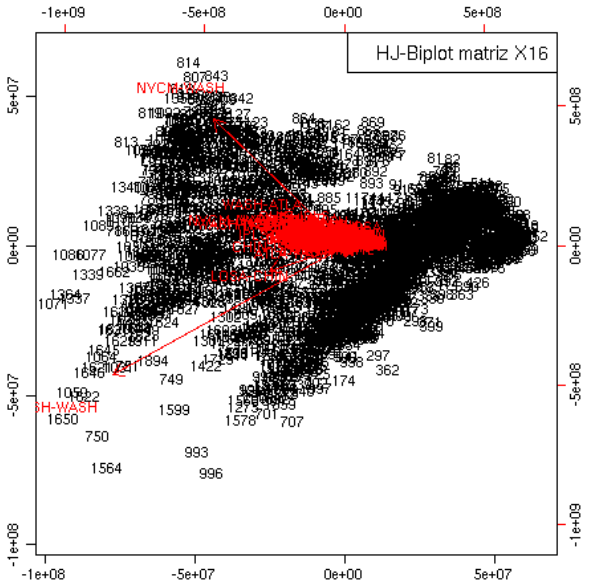

HJ-Biplot matriz X11
0e+00
2e+09
4e+09
6e+09
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
LOSA-CHIN

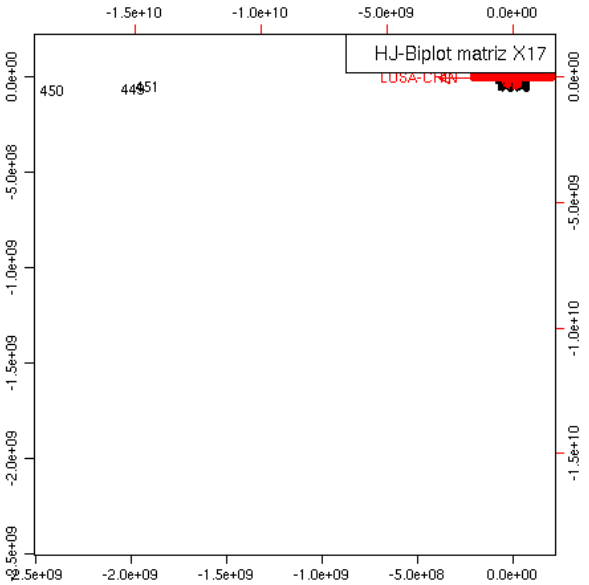

HJ-Biplot matriz X12
0.0e+00
5.0e+09
1.0e+10
1.5e+10
2.0e+10
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
1864
1865

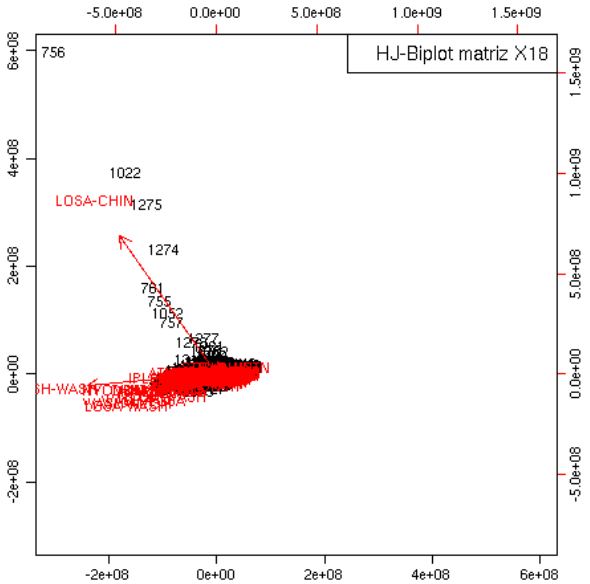

HJ-Biplot matriz X13
LOSA-CHIN
ATLA-LOSA

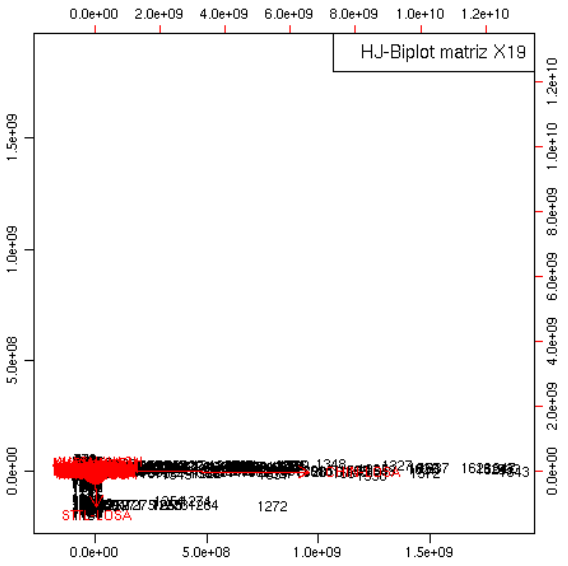

HJ-Biplot matriz X14
WASH-WASH

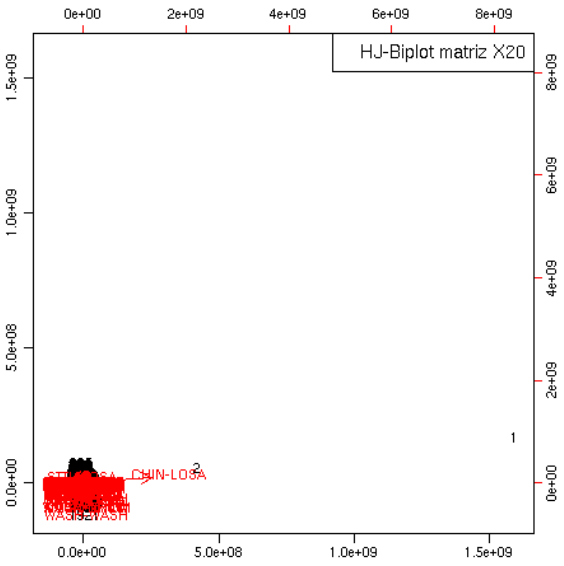

HJ-Biplot matriz X15
CHIN-LOSA
CHIN-CHIN

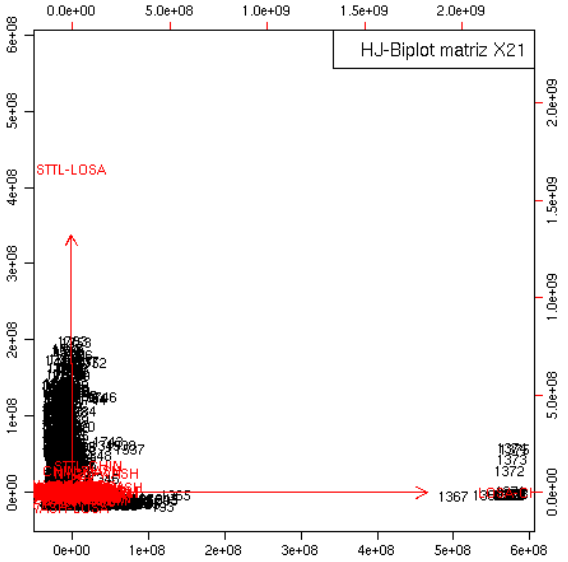

HJ-Biplot matriz X16

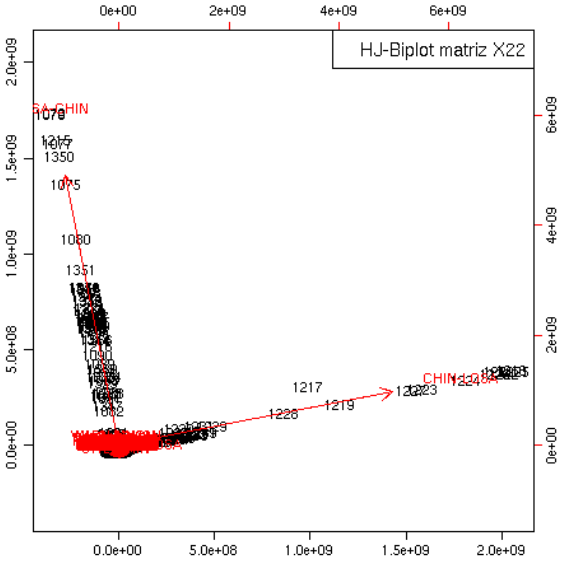

HJ-Biplot matriz X17

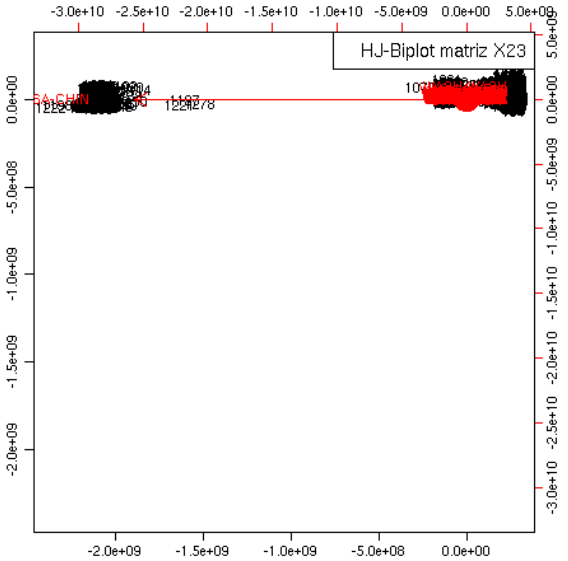

HJ-Biplot matriz X18
LOSA-CHIN

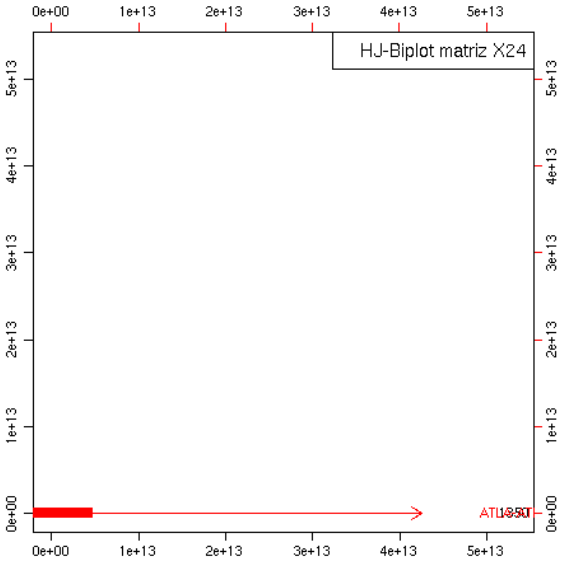

HJ-Biplot matriz X19
0.0e+00 2.0e+09 4.0e+09 6.0e+09 8.0e+09 1.0e+10 1.2e+10
0.0e+00 5.0e+08 1.0e+09 1.5e+09
1272


HJ-Biplot matriz X20
0e+00 2e+09 4e+09 6e+09 8e+09
0.0e+00 5.0e+08 1.0e+09 1.5e+09
CHIN-LOSA
1
2


HJ-Biplot matriz X21
0.0e+00 5.0e+08 1.0e+09 1.5e+09 2.0e+09
0e+00 1e+08 2e+08 3e+08 4e+08 5e+08 6e+08
STTL-LOSA
1367
1372


HJ-Biplot matriz X22
0e+00 2e+09 4e+09 6e+09
0.0e+00 5.0e+08 1.0e+09 1.5e+09 2.0e+09
1079
1350
1075
1080
1351
1217
1219
1228


HJ-Biplot matriz X23
-3.0e+10 -2.5e+10 -2.0e+10 -1.5e+10 -1.0e+10 -5.0e+09 0.0e+00 5.0e+09
-2.0e+09 -1.5e+09 -1.0e+09 -5.0e+08 0.0e+00

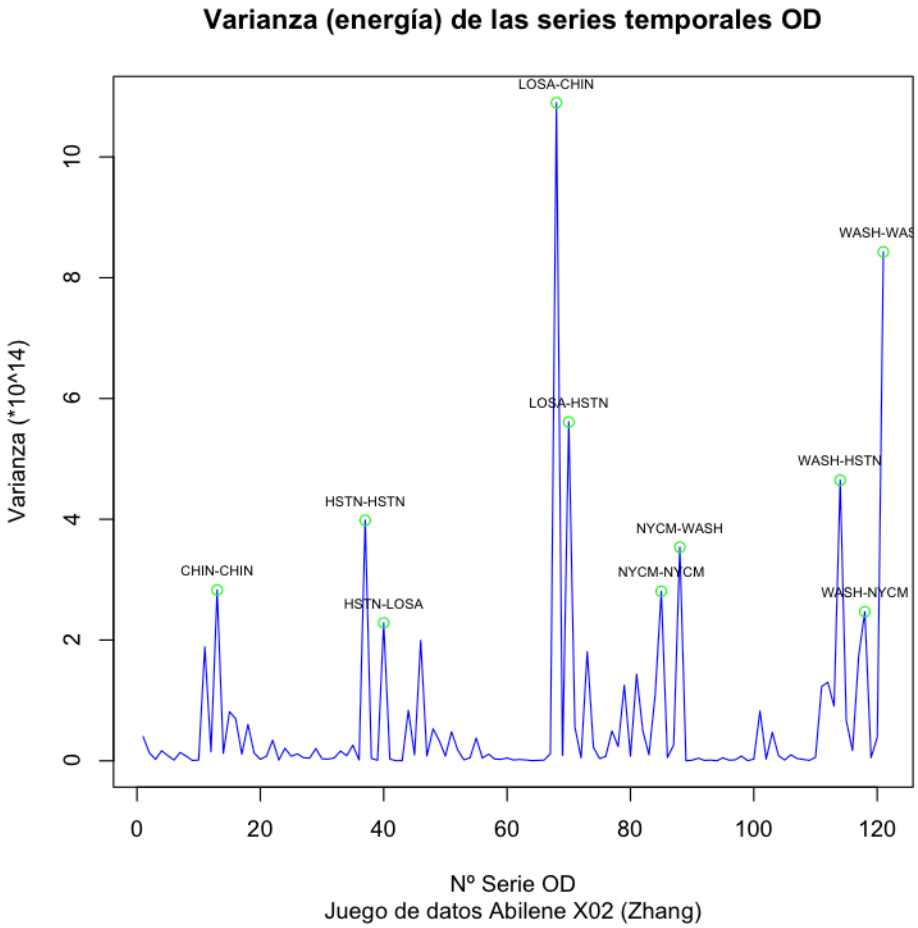

HJ-Biplot matriz X24
0e+00 1e+13 2e+13 3e+13 4e+13 5e+13

De las 24 representaciones mostradas observamos tres que nos resultan especialmente interesantes. Se trata de las correspondientes a las matrices **X02**, **X23** y **X24**. De manera muy evidente en estas tres representaciones HJ-Biplot aparecen dos nubes de puntos clarísimamente diferenciadas correspondientes a marcadores fila **J** (intervalos temporales).

6.4.5. APLICACIÓN DEL HJ-BIPLOT AL JUEGO DE DATOS X02

Vamos a proceder a representar el HJ-Biplot correspondiente a la matriz **X02** para analizar con más detalle este comportamiento.

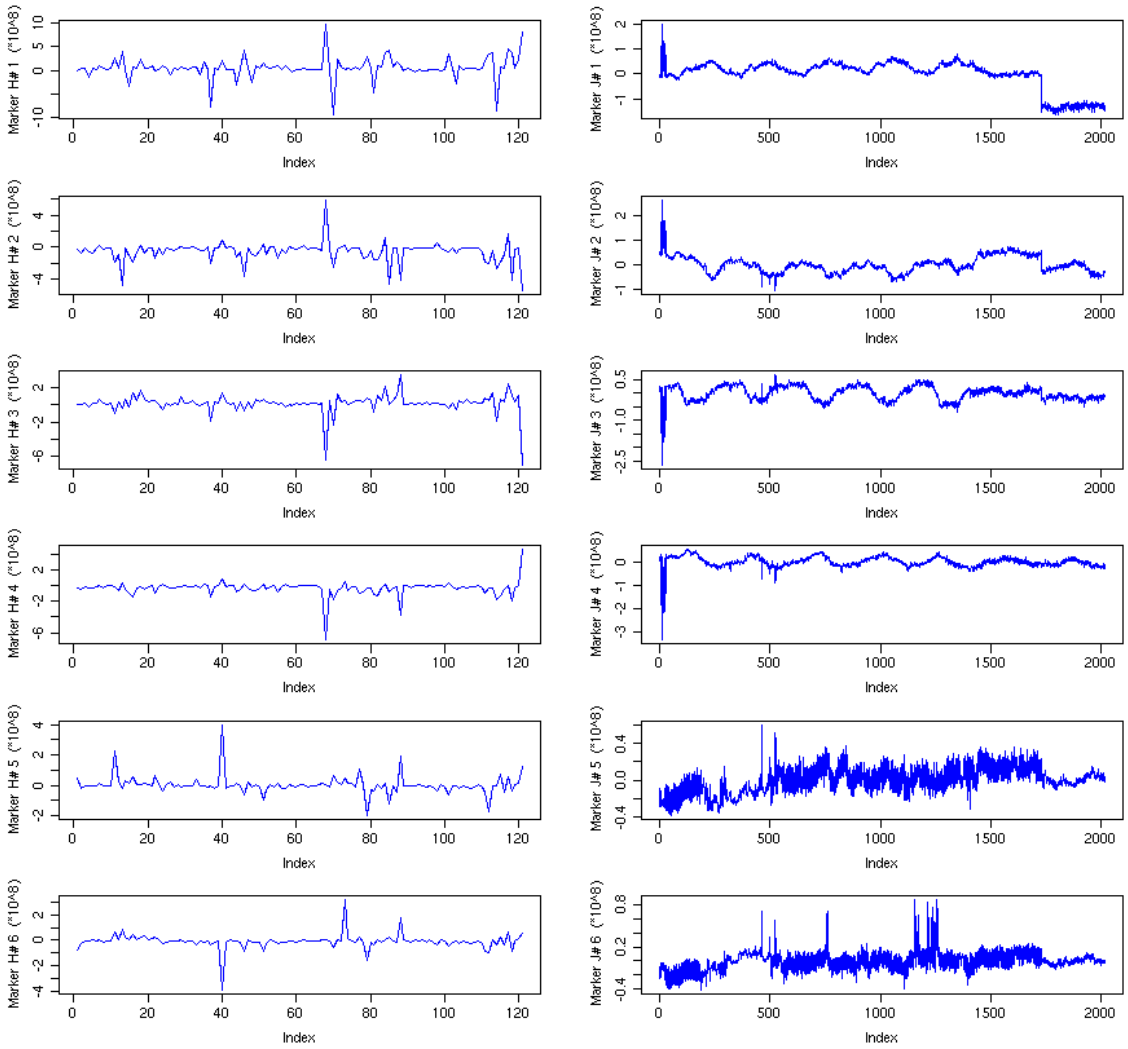
Procedemos como en el caso anterior de la matriz de tráfico **X01**, retirando las series correspondientes tráfico con origen o destino en ATLA-M5, que al igual que antes presenta un elevado número de "ceros". Representemos la varianza de las series temporales restantes de la matriz **X02**.



Representación de la varianza de las series temporales OD del juego de datos X02

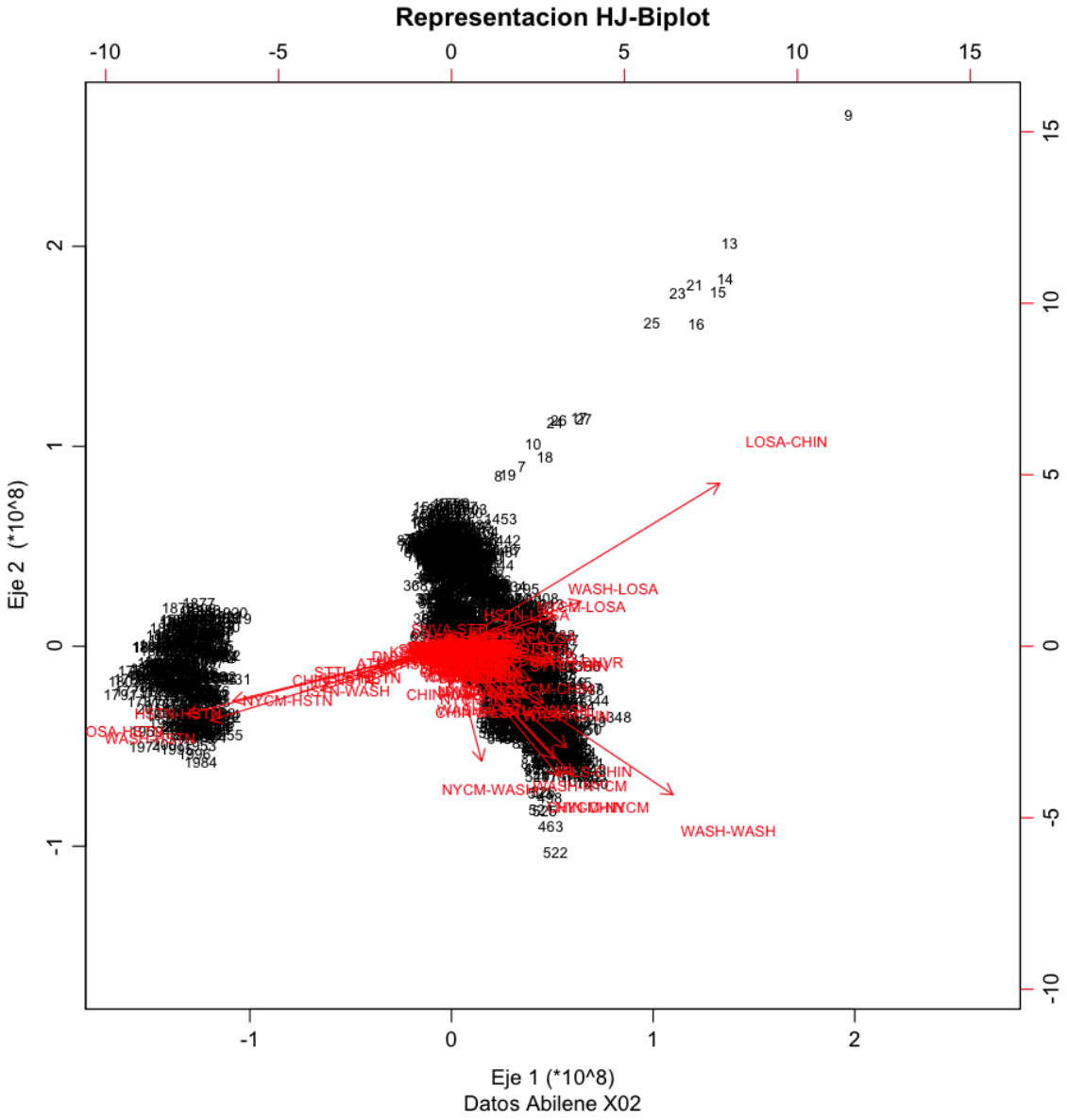
Las 10 series temporales que presentan mayor varianza (energía) son las correspondientes a las rutas LOSA-CHIN, WASH-WASH, LOSA-HSTN, WASH-HSTN, HSTN-HSTN, NYCM-WASH, CHIN-CHIN, NYCM-NYCM, HSTN-LOSA y WASH-NYCM.

Tras este estudio previo de los datos y el preprocesado de los mismos como consecuencia de la estructura descriptiva detectada en las series temporales, podemos representar los 6 primeros *eigenflows* englobados dentro de las series de marcadores columna **H** y fila **J**.

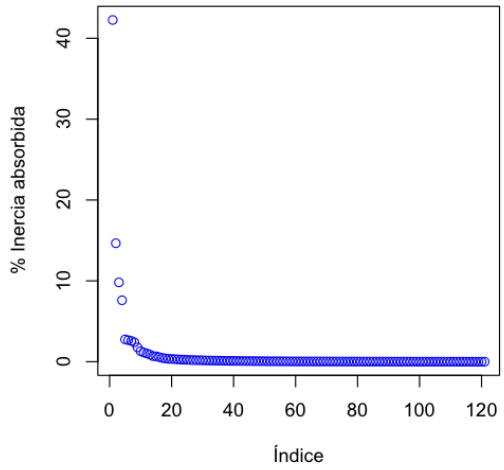


Representación de las seis primeras series de marcadores **H** y **J**.

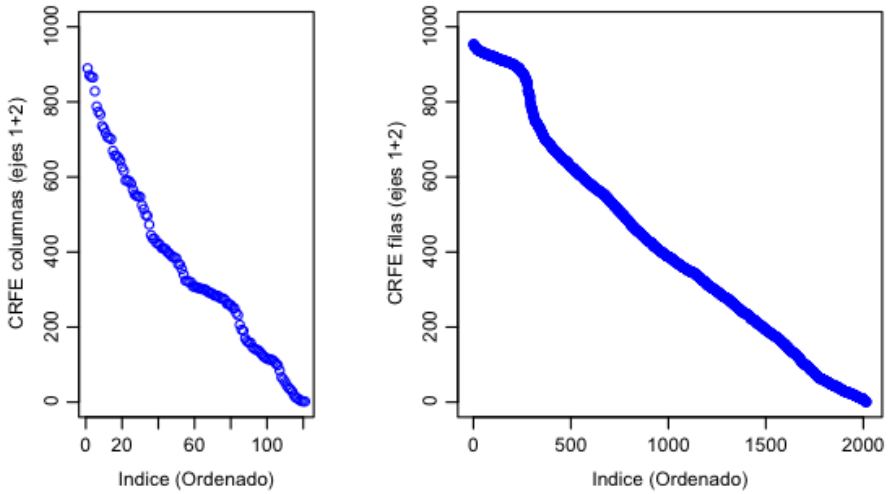
Se procede a representar a continuación el HJ-Biplot obtenido al procesar esta matriz de tráfico **X02**, tras centrar por columnas las series temporales que la forman.



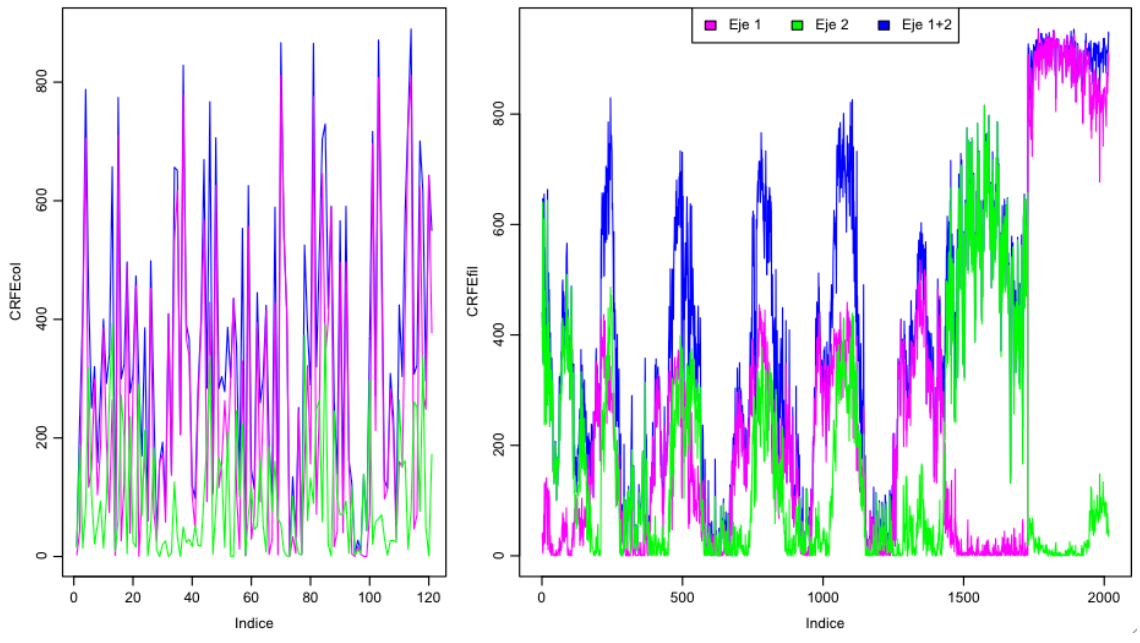
La inercia absorbida en el primer plano factorial alcanza el 56,9%. El *scree plot* nos muestra el codo en la quinta componente de las 121 totales.



Podemos efectuar un estudio sobre la calidad de representación / bondad de ajuste para los marcadores en este primer plano factorial. No se aprecia realmente un punto de corte o solución de continuidad que nos permita aportar un punto de corte a la CRFE.



Si representamos conjuntamente la CRFE de los ejes 1 y 2 y la suma la CRFE de ambas vemos que por lo que respecta a la CRFE para las columnas aparece un efecto similar al ya observado en el caso del estudio de la matriz de tráfico **X01** precedente: una "periodicidad" hasta un intervalo temporal que desaparece a favor de una preponderancia del eje factorial primero. En el caso de las columnas la situación es más heterogénea y habrá que analizar la bondad de ajuste de los marcadores en este primer plano caso a caso.



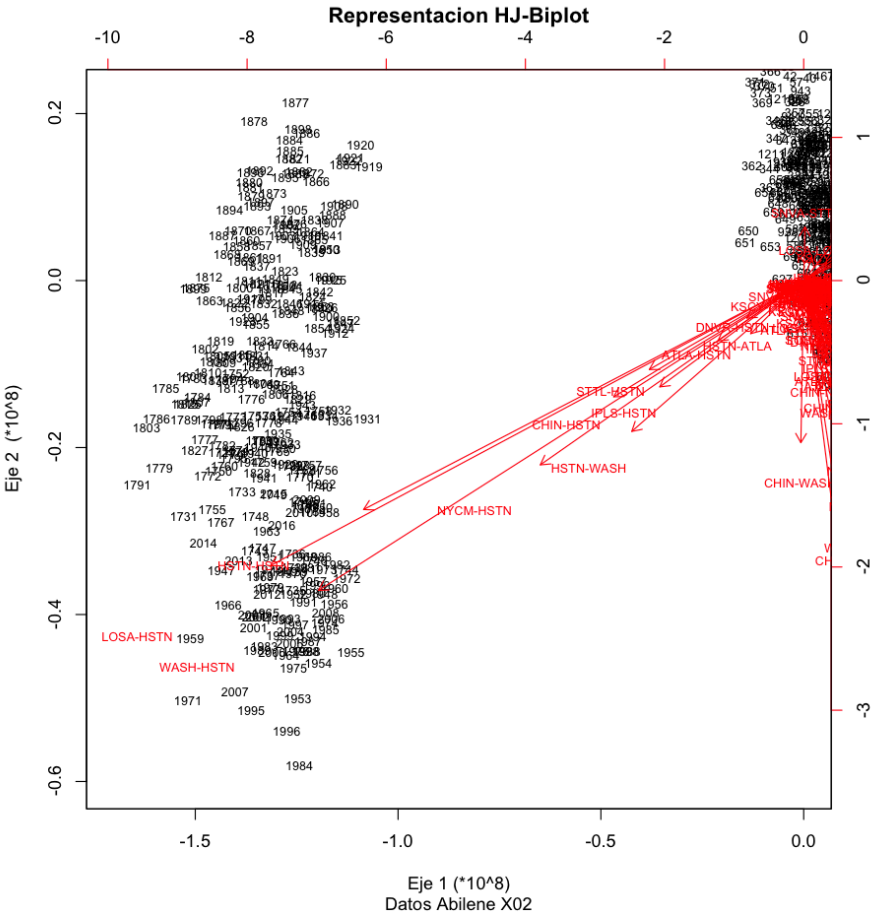
Se realiza a continuación una ampliación de la zona correspondiente a la nube de puntos situada a la izquierda, para poder observar con mayor detalle su composición (ver página siguiente). Si analizamos los marcadores fila **J** que componen esta nube veremos que se encuentran en esta nube "satélite" los marcadores correspondientes a los intervalos temporales situados entre los valores 1729 y 2016 (el final).

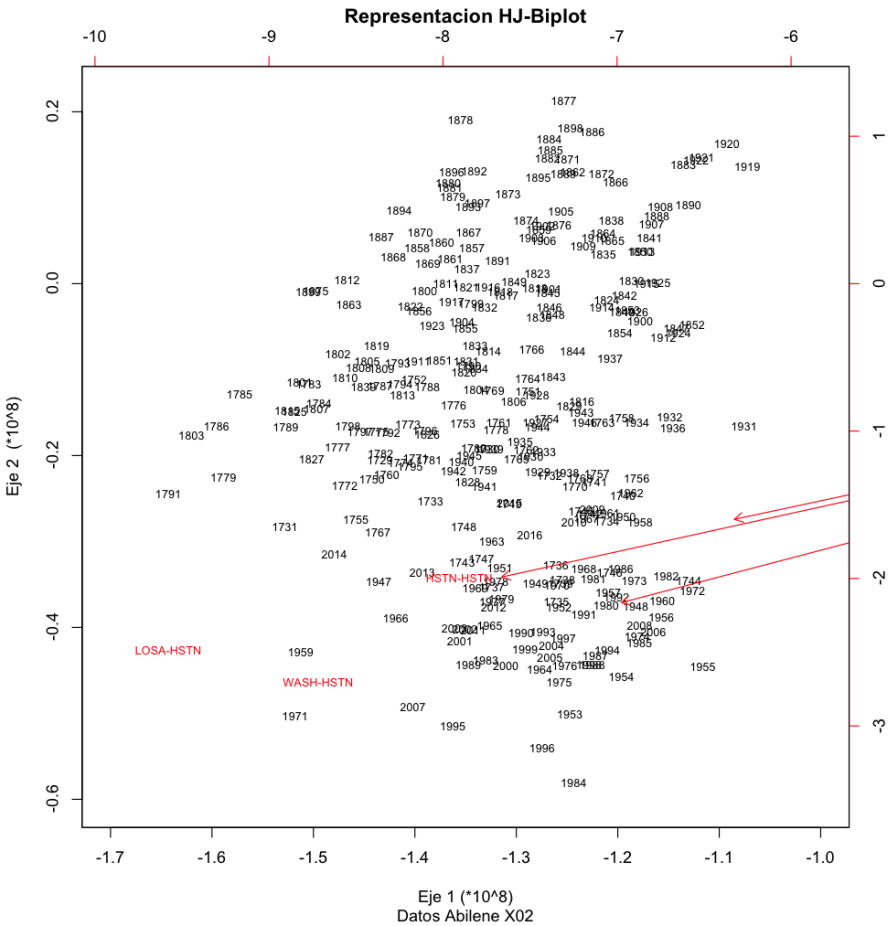
Los vectores correspondientes a marcadores columna **H** que apuntan a esta nube de puntos son los correspondientes a las siguientes trece series temporales: ATLA-HSTN (4), CHIN-HSTN (15), DNVR-HSTN (26), HSTN-ATLA (34), HSTN-HSTN (37), HSTN-WASH (44), IPLS-HSTN (48), KSCY-HSTN (59), LOSA-HSTN (70), NYCM-HSTN (81), SNVA-HSTN (92), STTL-HSTN (103) y WASH-HSTN (114).

Si obtenemos la CRFE de estos marcadores columna **H** obtendremos dos conclusiones: son preponderantemente de primer eje y la bondad de ajuste es muy buena en la mayoría de los casos y suficientemente buena en todos ellos.

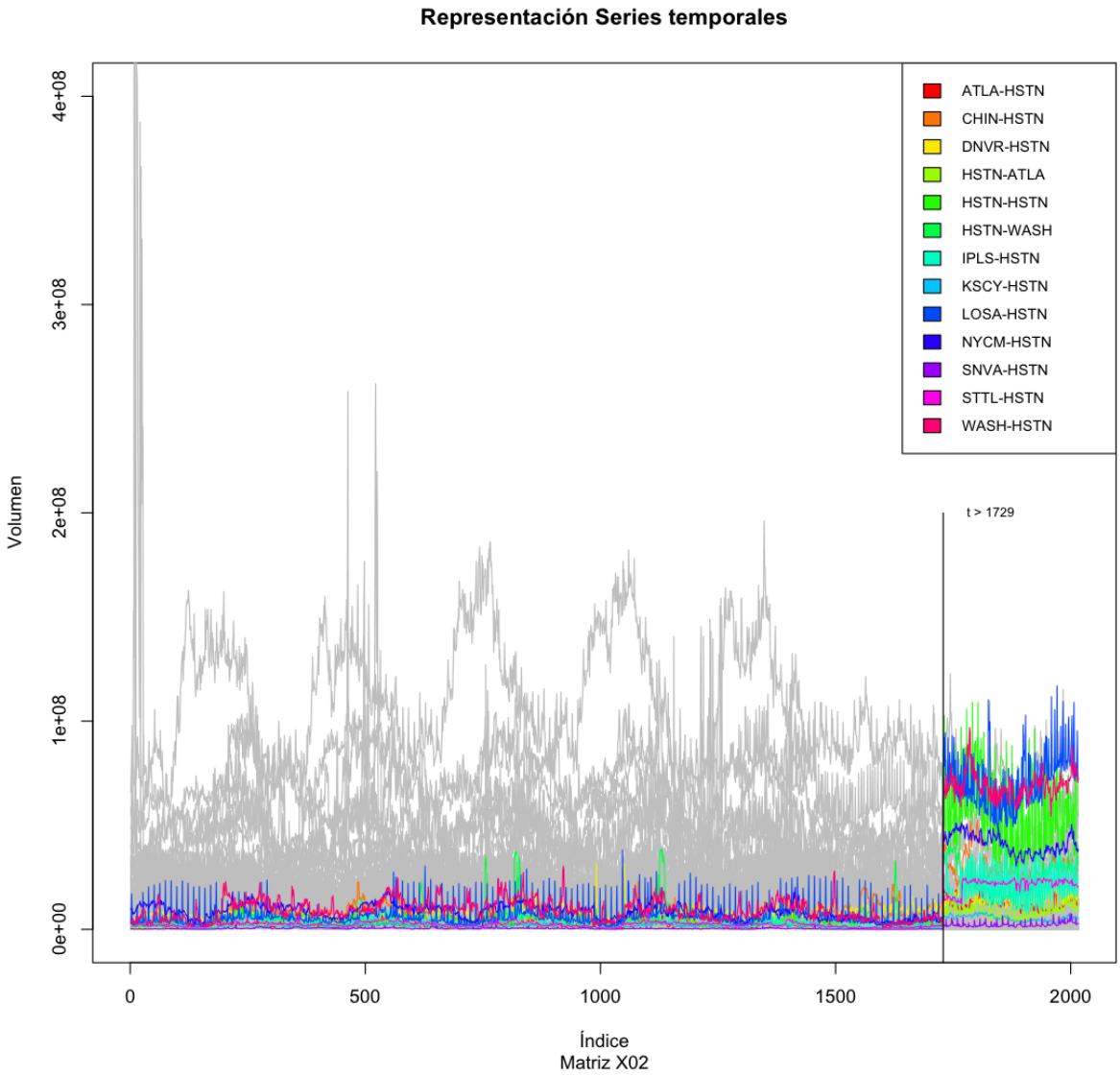
Serie O-D (# flow/col)	CRFE Eje1	CRFE Eje2	CRFE Eje1+2
ATLA-HSTN (4)	705	82	788 (*)
CHIN-HSTN (15)	711	63	774
DNVR-HSTN (26)	452	46	499 (*)
HSTN-ATLA (34)	531	126	656 (*)
HSTN-HSTN (37)	779	50	828 (*)
HSTN-WASH (44)	567	103	670
IPLS-HSTN (48)	626	81	706 (*)
KSCY-HSTN (59)	556	69	625
LOSA-HSTN (70)	812	55	866 (*)
NYCM-HSTN (81)	776	90	865 (*)
SNVA-HSTN (92)	496	94	591 (*)
STTL-HSTN (103)	808	63	871
WASH-HSTN (114)	811	78	890 (*)

(*) Sujeto a redondeo decimal


Representacion HJ-Biplot
Eje 2 (*10^8)
Eje 1 (*10^8)
Datos Abilene X02
LOSA-HSTN
WASH-HSTN
NYCM-HSTN
HSTN-WASH
CHIN-HSTN
IPLS-HSTN
STTL-HSTN
CHIN-WAS

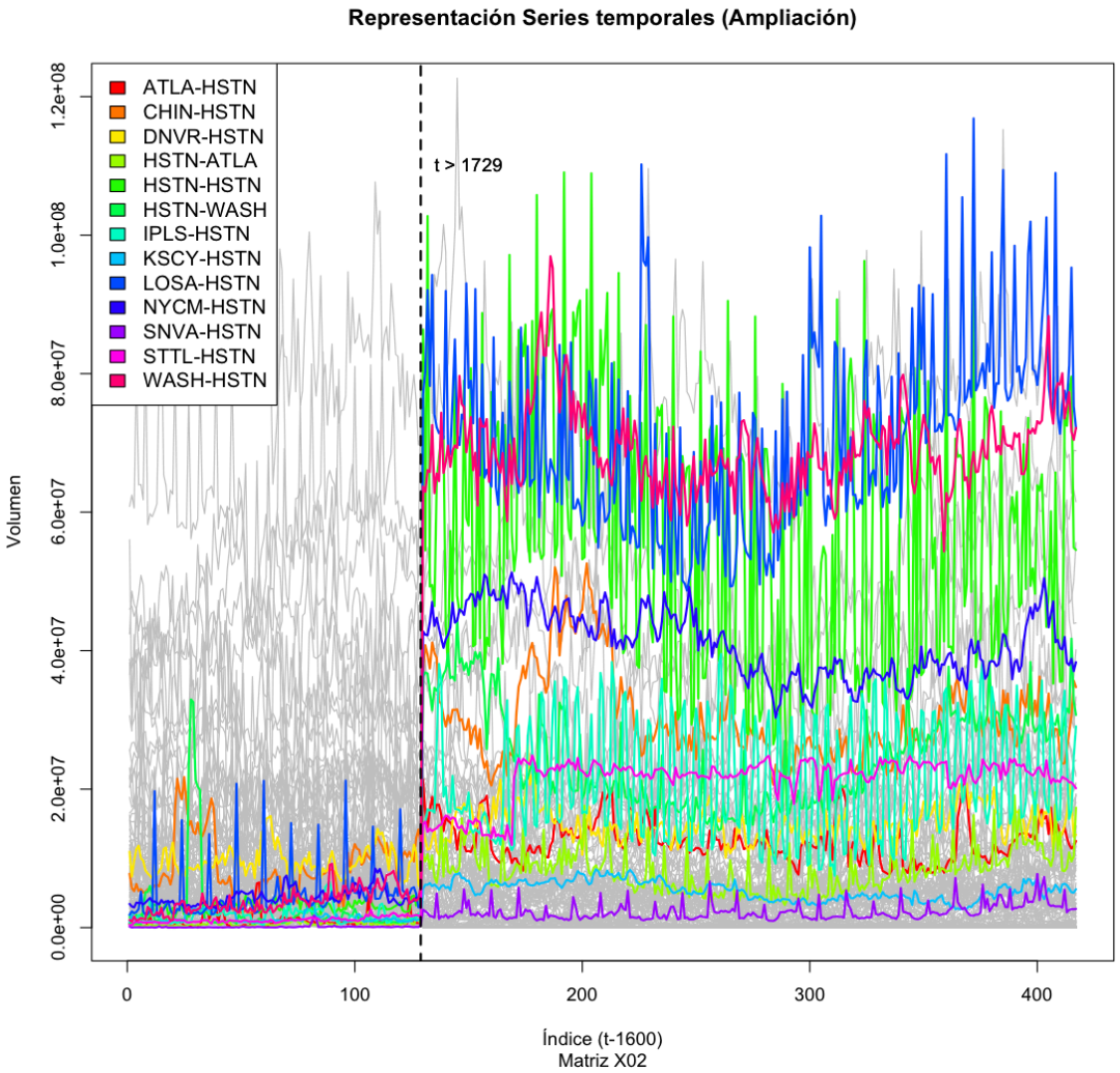

Representacion HJ-Biplot
Eje 2 (*10^8)
Eje 1 (*10^8)
Datos Abilene X02
LOSA-HSTN
WASH-HSTN
HSTN-HSTN

Si representamos las todas las series temporales, destacando aquellas cuyos marcadores están claramente orientados hacia la segunda nube de puntos obtenemos un interesante resultado:



Esto es, la representación HJ-Biplot y concretamente la identificación de los marcadores correspondientes a las series temporales orientados hacia la segunda nube de puntos se corresponden con series temporales que experimentan un aumento importante de valor en los intervalos temporales correspondientes a los marcadores fila que conforman la segunda nube de puntos.

Si se representan individualmente las 121 series que conforman la matriz de datos de tráfico **X02** ampliando la zona temporal de interés se comprueba que estas series, y solo estas, presentan dicho comportamiento en la parte final del intervalo temporal bajo estudio.



En resumen, para identificar el comportamiento detectado en esas series del modo "tradicional" deberíamos representar individualmente las 121 series temporales y analizar por separado su perfil.

Utilizando en su lugar una única representación HJ-Biplot para analizar el conjunto de las 121 series temporales a lo largo de los 2016 intervalos temporales bajo estudio hemos obtenido los mismos resultados de una manera más rápida y por lo tanto eficiente.

Nuestra propuesta nos ha permitido también corregir la problemática presente en los estudios que utilizan ACP [227] consistente en la imposibilidad de identificar la localización espacial original de la anomalía. El HJ-Biplot nos identifica la serie temporal, y por lo tanto la localización espacial, que es responsable de la situación detectada.

7. REDES Y GRAFOS PARA EL ANÁLISIS DE TOPOLOGÍAS

7. REDES Y GRAFOS PARA EL ANÁLISIS DE TOPOLOGÍAS

7.1. INTRODUCCIÓN A LAS REDES Y A LOS GRAFOS

Algunos autores [306] se refieren al término "*red*" (*network*) para reflejar el concepto que describe una entidad compuesta de diferentes elementos y las interacciones o conexiones entre dichos elementos. Básicamente una red consiste en vértices y las aristas que los interconectan, también denominados respectivamente nodos y enlaces. Estos nodos y enlaces pueden representar objetos físicos, tales como personas, equipos informáticos o ciudades, o incluso objetos no físicos, o inmateriales, como reuniones, eventos o páginas de hipertexto [307]. Matemáticamente los datos que describen cuantitativamente esas redes consisten típicamente en un conjunto de nodos y un vínculo de relación entre cada par de esos nodos [308]. Las interacciones pueden estar constituidas por mediciones en bruto, tales como la localización de individuos, pertenencias a grupos, o contadores numéricos y también pueden ser calculadas analíticamente de manera secundaria. Así, se pueden asociar diferentes estadísticos tanto a los nodos como a los enlaces, y esos estadísticos pueden ser tanto fijos en el tiempo como variables [307]. Incluso se puede pensar en la argumentación inversa: matrices de datos apoyadas en técnicas de análisis multivariante, tales como Análisis de Componentes Principales, análisis de clusters, escalado multidimensional o biplots, que son representadas en forma de grafos en los que los nodos son conceptos indivisibles (atómicos) y un conjunto de aristas formando un camino que los unen contituyen una observación de la tabla original, que puede ser incluso multivía. Estos grafos de intersección constituyen otra técnica posible para análisis multivariante de datos [309].

Este tipo de planteamientos presentan múltiples aplicaciones, incluyendo el estudio de guerras, el comercio, comportamiento de epidemias, la conectividad en la World Wide Web y los patrones de llamadas telefónicas [308], por citar solo algunos ejemplos generales, que serán ampliados posteriormente.

El análisis de redes trabaja con los datos que describen el conjunto de relaciones entre los miembros de un sistema. El objetivo de estos análisis es obtener, a partir de los datos de relaciones de bajo nivel, una descripción más elevada de la estructura del sistema que identifique tipos de patrones entre el conjunto de las relaciones existentes. Estos patrones estarán basados en la manera en la que los nodos (individuos) se relacionen con otros nodos de la red. Algunas aproximaciones al

análisis de redes buscan agrupaciones o *clusters* de nodos que tengan similares patrones de relación con el resto de la red. Otros métodos, por el contrario, no buscan nada en particular, en su lugar construyen una representación multidimensional de la red en la que las coordenadas de los vértices puedan ser analizadas más profundamente para obtener varios tipos de información sobre ellos y su relación con el resto de la red [310]. Esta información buscada incluye la comprensión de la propia estructura de la red, los flujos de tráfico entre vértices, sus variaciones y/o la importancia relativa de los nodos y enlaces clave. Hay tareas más sutiles en el análisis de redes, tales como el aislamiento de "signos relevantes" entre la masiva cantidad de actividades de fondo, y que con frecuencia se identifican como *outliers* [307].

Otro de los principales temas de análisis de redes compete a los caminos que unen pares de nodos. Este tipo de problemas derivan de manera natural en una gran variedad de aspectos, tales como, el estudio de los procesos de comunicación en redes, el flujo de información en computación en paralelo y el análisis de algoritmos de encaminamiento en circuitos integrados [311].

El campo de la "minería de grafos" ha sido solo desarrollado recientemente por la comunidad de "minería de datos", pero fue estudiado antes bajo diferentes perspectivas por otros grupos de investigadores, más notablemente por sociólogos. Diferentes trabajos culminaron con la aceptación y el uso del Análisis de Redes Sociales como una herramienta para investigar la estructura de grupos sociales y organizaciones [312]. Se han aplicado estos conceptos incluso al análisis de redes terroristas [313], [106] que presentan como retos a las herramientas de análisis utilizadas el que se trate de redes sobre las que se dispone un conocimiento incompleto, sus contornos son difusos y son además dinámicas [312].

La manera natural de modelar redes de manera formal es a través de la noción de *grafo* [306], por lo que a continuación se expondrán sus principios más relevantes y la nomenclatura que se empleará en el presente texto, ya que en algunos casos existen diferentes terminologías en la literatura que complican la comprensión de los diferentes conceptos implicados en su estudio [314].

7.1.1. GENERALIDADES DE LOS GRAFOS

En su forma más simple un grafo es una estructura G(V,E) que representa una relación binaria E (por *edges,* aristas) sobre un conjunto de entidades V (por *vertex*, vértices) [315]. En este grafo G(V,E) el conjunto de elementos V denominados vértices

o nodos se representan como puntos y el conjunto de aristas o enlaces E se representan como líneas [306], [310].

El cardinal de V es usualmente denotado por *n*, el cardinal de E por *m* [306] y al número de vértices/nodos *n* de un grafo se le denomina “orden de G” [310].

Dos vértices unidos por una arista se denominan *vértices finales*. Si dos vértices están unidos por una arista, serán *adyacentes* y los calificaremos como *vecinos*.

Los grafos pueden ser no-dirigidos o dirigidos. En los grafos no-dirigidos, el orden de los vértices finales de una arista es irrelevante. En los grafos dirigidos, cada arista con dirección tiene un origen y un destino. Para un grafo dirigido, el “grafo no dirigido subyacente” es el grafo no dirigido que, con los mismos vértices que el grafo dirigido, tiene vértices no dirigidos entre dos de esos vértices si en el grafo dirigido tiene un vértice en cualquiera de los dos sentidos. Dicho de otro modo los vértices dirigidos se convierten en no-dirigidos en el “grafo no dirigido subyacente” [306]. Incluso se pueden definir grafos mixtos, con unas aristas dirigidas y otras no [316].

Un grafo es denominado simple si solo puede contener una arista en *E* una única vez, esto es, no existen aristas paralelas entre dos mismos nodos. Si existen aristas paralelas se denominará grafo multígrafo. Una arista uniendo un vértice consigo mismo se denomina lazo. Un grafo se dice libre de lazos si no tiene lazos [306].

A veces es útil asociar valores numéricos (pesos) con las aristas o con los vértices de un grafo G(V,E). A este grafo se le denomina grafo con pesos o grafo ponderado. El caso más habitual son los pesos en las aristas. Estos pesos se representan como una función $w : E \rightarrow \Re$ que asigna cada arista $e \in E$ un valor w_e. Dependiendo del contexto los pesos de las aristas pueden describir varias propiedades, tales como, “coste” (por ejemplo, tiempo de viaje o distancia), capacidad, fuerza de interacción, o similaridad. Asumimos que:

$$w_{ij} \geq 0$$

$w_{ij} = 0$ para cualquier par de nodos i,j no adyacentes

En la mayoría de los casos un grafo sin pesos es equivalente a un grafo con pesos en el que todos sus pesos son iguales a la unidad [306].

El grado de un vértice *v* en un grafo no dirigido G(V,E), denotado por d_v y también denominado valencia [317], es el número (cardinal) de aristas de *E* que tienen en *v* su vértice final [306].

En los grafos con pesos en lugar de tratar con los cardinales de cada subconjunto de aristas trataríamos con las sumas de los pesos de las respectivas aristas [306]. Así, se define el grado del nodo i-ésimo en un grafo ponderado como [315]:

$$d_i=\sum_j w_{ij}.$$

El grado medio se define como [306], [318]

$$\bar{d}(G)=\frac{1}{|V|}\sum_{v\in V} d_v \qquad \text{o más abreviadamente} \qquad \bar{d}=\frac{1}{n}\sum_{i\in V} d_i$$

Un grafo no dirigido G(V,E) se dice *regular* si todos su vértices tienen el mismo grado, y *r-regular* si ese grado es igual a *r* [306].

Un paseo (*walk*) de x_0 a x_k en un grafo G(V,E) es una secuencia alternada x_0, e_1, x_1, e_2, x_2,..,x_{k-1}, e_k, x_k de vértices y aristas, donde $e_i=\{x_{i-1}, x_i\}$ en el caso no-dirigido y $e_i=(x_{i-1}, x_i)$ en el caso dirigido. La longitud del paseo se define como el número de aristas en el paseo. El paseo se denomina ruta o camino si $e_i \neq e_j$ para $i\neq j$, esto es, un paseo se denomina camino si no repite aristas. Un camino es simple si $x_i \neq x_j$ para $i\neq j$. Una ruta con $x_0=x_k$ es un ciclo. Un ciclo es simple si $x_i \neq x_j$ para $0\leq i<j\leq k-1$ [306]. Para un camino *p* en un grafo G(V,E) con pesos en los aristas ω, el peso del camino, denotado por w(*p*) se define como la suma de los pesos de los aristas de *p*. Un camino de *u* a *v* en G es el camino más corto (respecto a w) si su peso es el menor posible entre todos los camino posibles entre *u* y *v*. La longitud del camino más corto entre *u* y *v* se denota $d_{G,W}(u,v)$ [306]. El diámetro de la red N es el camino más largo posible entre cualquier par de nodos [310].

Un grafo no dirigido G(V,E) se dice que está conectado, o que es conexo, si cada vértice puede ser alcanzado desde cualquier otro vértice, esto es, si hay un ruta entre un vértice y cualquier otro vértice. Los grafos no conectados se denominan desconectados, o inconexos [306].

Un grafo dirigido G(V,E) está fuertemente conectado si hay un camino (*path*) directo entre cada vértice y cada otro vértice. Un grafo dirigido se dice débilmente conectado si su grafo no dirigido subyacente es conectado [306].

Un vértice de corte es aquel cuya eliminación junto con sus aristas incidentes parte los grafos remanentes en dos o más componentes desconectadas [319].

Una red de flujo esta definida [306] por un grafo dirigido G(V,E), una función u:E→$\Re$ que asigna capacidades no-negativas a las aristas, y dos vértices diferentes s,$t \in$ V designados como fuente y sumidero, respectivamente. Un flujo f desde s a t, o abreviadamente un flujo-s-t, es una función f:E→$\Re$ que satisface las siguientes restricciones:

- Restricciones de capacidad: $\forall$ e ∈ E : 0 ≤ f(e) ≤ u(e)
- Condición de balance: $\forall$ v ∈ V \ {s,t} : $\Sigma_{e \in \Gamma^-(v)}$ f(e) = $\Sigma_{e \in \Gamma^+(v)}$ f(e)
- El valor del flujo f se define como $\Sigma_{e \in \Gamma^+(s)}$ f(e) - $\Sigma_{e \in \Gamma^-(s)}$ f(e).

Siendo $\Gamma^+(v)$ el conjunto de aristas con origen en v.
$\Gamma^-(v)$ el conjunto de aristas con destino en v.

7.1.1.1. Redes bipartitas

Un red bipartita es aquella que puede ser dividida de manera que nodos en una parte de la misma tengan conexiones solo con nodos situados en la otra parte [310]. De manera formal [320] un grafo G(V,E) es bipartito con dos clases de vértices X e Y si V=X∪Y con X∩Y=ϕ y cada arista en E tiene un extremo en X y otro extremo en Y.

Como ejemplo de red bipartita [320] se puede considerar un grafo ponderado constituido por X∪Y vértices, tales que X representa el conjunto de términos e Y representa el conjunto de documentos, y la ponderación w_{ij} puede ser definido como el número de veces que el término i-ésimo aparece en el documento j-ésimo.

También, en el caso de una red de telecomunicación se podría considerar cada nodo de comunicaciones como constituido por dos subnodos diferenciados, un subnodo transmisor y otro subnodo receptor de información, lo que daría lugar a un grafo bipartito para modelar esa la red.

7.1.2. GENERALIDADES DE MÉTODOS ESPECTRALES Y GRAFOS

El espectro de una matriz **X** = (x_{ij}) ∈ C se define como el conjunto todos los autovalores de **X**, donde la multiplicidad de un autovalor λ es su multiplicidad como raíz del polinomio característico de **X** [318].

Como veremos, un grafo puede ser asociado con diferentes formulaciones de matrices cuyos autovalores reflejarán propiedades estructurales del propio grafo. Pero ¿cómo puede ser utilizado el espectro para analizar un grafo? En particular pueden ser de interés las siguientes cuestiones [318]:

- ¿Cómo puede ayudarnos el espectro a clasificar grafos?
- ¿Pueden estar ciertos autovalores relacionados con otros estadísticos globales (denominados parámetros del grafo), tales como, por ejemplo, su diámetro?
- ¿Qué puede decirnos el espectro de los subgrafos?
- ¿Se puede probar la existencia o inexistencia de ciertos subgrafos en un grafo analizando el espectro?

Los métodos espectrales han sido parte de la teoría de grafos durante un siglo. Los investigadores han utilizado métodos espectrales de manera explícita o implícita desde 1960 [310], [311]. En sus inicios se utilizaba análisis matricial y álgebra lineal para analizar determinadas matrices relacionadas con los grafos. Así como, por ejemplo, los astrónomos estudian el espectro estelar para determinar la composición de estrellas distantes, uno de los principales objetivos en la teoría espectral de grafos es deducir las propiedades principales y la estructura de un grafo a partir del espectro del grafo. El espectro de un grafo revela propiedades fundamentales del mismo [311], [153]. La teoría espectral de grafos enlaza, a través de sus autovalores, con teoría y aplicaciones en telecomunicaciones [311]. Los autovalores están íntimamente ligados a casi la mayor parte de invariantes de un grafo y no hay duda alguna de que dichos autovalores juegan un papel central en la comprensión de los grafos [311].

Los autovalores de una red están formalmente relacionados a importantes características topológicas de la red, tales como la máxima distancia a través de la red (diámetro), presencia de clusters, rutas y cuellos de botella, y a cuánto de aleatorio es el grafo. Los autovectores asociados pueden ser utilizados como un sistema de coordenadas natural para la visualización del grafo y también facilitan el descubrimiento de clusters y otras características locales. Cuando se combinan con otros valores, obtenidos fácilmente de la propia red (por ejemplo, el grado del nodo), pueden ser utilizados para describir varias propiedades de la red, tales como, su robustez, así como a otras propiedades estructurales y la relación entre estas propiedades y las características de nodos o enlaces en redes grandes, complejas y multivariantes [310].

7.1.2.1. CONDICIONES DE CONTORNO GENERALES HABITUALES EN LOS ESTUDIOS SOBRE GRAFOS

Las condiciones de contorno generales que más habitualmente se consideran en los diferentes estudios sobre grafos son las siguientes (por ejemplo [310], [321], [322]):

- Grafos no dirigidos: los enlaces son “simétricos”.
- Grafos sin autolazos: No existen aristas que unan un vértice consigo mismo.
- Grafos simples: No existen aristas múltiples conectando dos mismos nodos.
- Tienen una sola componente: hay un conjunto de aristas conectando dos nodos cualquiera.

Además, en algunos estudios ni siquiera se permite que la red presente una ponderación o pesos en los enlaces (por ejemplo [310]), aunque en otros estudios sí que admiten esa circunstancia (por ejemplo [321]).

La conectividad en grafos dirigidos no es simétrica [323] y esa es la razón por la que muchos algoritmos para grafos no-dirigidos no pueden aplicarse al caso de grafos dirigidos y se dispone de algoritmos específicos para analizar grafos dirigidos.

Estas condiciones de contorno habituales en el estudio de grafos que aparecen muy frecuentemente en la literatura son, no obstante, bastante restrictivas en el ámbito de las redes de telecomunicación, principalmente por los siguientes motivos:

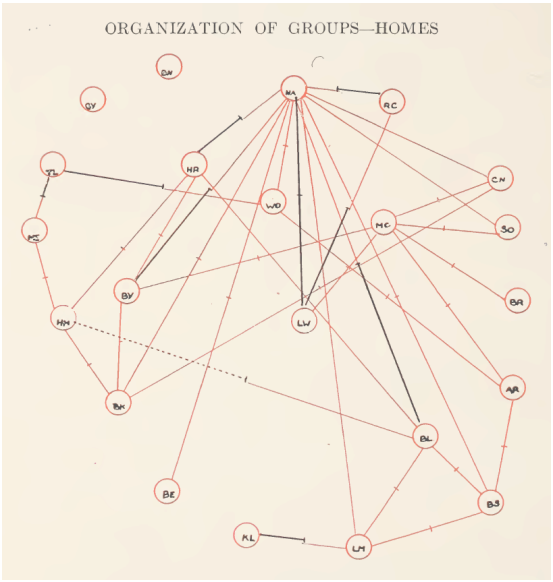
1. El refuerzo de la supervivencia o resiliencia, esto es, la tolerancia a la retirada de nodos o enlaces de las redes en general, y de las de telecomunicación en particular, aconsejan no solo el establecimiento de posibles rutas alternativas entre nodos, sino incluso la existencia de enlaces redundantes entre dos nodos vecinos. Esta última medida se refleja en el grafo como una arista múltiple, si bien podría también modelarse como una arista con un peso múltiplo del existente en presencia de un enlace no redundado para el caso de redes “no ponderadas” (aunque el grafo obtenido sí sería ahora ponderado).

2. La existencia y utilización en redes de telecomunicación de tecnologías de transmisión típicamente asimétricas (como por ejemplo el ADSL, enlaces vía satélite o redes MPLS) hace que el grafo resultante del modelado de la red sea “no simétrico”; o con mayor propiedad, que los enlaces del grafo resultante no sean simétricos y deban ser representados individualmente como una arista

dirigida y ponderada para simbolizar de una manera precisa las diferentes capacidades bidireccionales de los enlaces de la red de comunicación.

7.1.3. VISUALIZACIÓN DE GRAFOS

Los modelos y métodos para visualizar grafos han mejorado fuertemente desde que Moreno introdujo en 1934 el "sociograma" [152]. En aquel momento la idea básica era representar los actores sociales por círculos y las relaciones entre ellos por flechas conectando los respectivos círculos. En tiempos más recientes la barrera entre visualización y modelado formal ha ido desapareciendo conforme se han desarrollado técnicas de representación más potentes [324]. Este nuevo planteamiento, equiparando técnicas de visualización con el propio modelado de la red, es muy interesante ya que nos ofrece en nuestro estudio un vínculo real entre las representaciones visuales que se obtengan de las redes y el modelado más formal de las mismas, que puede ser de evidente utilización en el diseño de redes y su gestión.



Ejemplo de "sociograma" (grafo) [152].

En un sentido general, se espera que la representación de un grafo capture su estructura inherente [315]. Hay muchos planteamientos de problemas que pueden ser representados como grafos y analizados mediante su "simple" visualización [307]. No obstante, a veces el tamaño de los conjuntos de datos representados sobrepasa con facilidad las capacidades de visualización disponibles con lo que la imagen obtenida se torna visualmente confusa y contaminada (*clutter*). Entre los muchos tipos de problemas que implican la comprensión de los datos de las redes que representan se encuentran la monitorización de redes de telecomunicación, el seguimiento de flujos

económicos, la comprensión de patrones de desplazamientos o el análisis de contactos personales, entre otros posibles.

7.2. MATRICES ASOCIADAS A LOS GRAFOS

A partir de un grafo dado pueden formularse distintas matrices asociadas al mismo. A continuación definiremos varias de ellas y enumeraremos algunas de sus propiedades principales e interrelaciones.

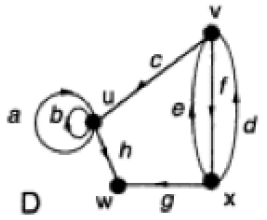
7.2.1. MATRIZ INCIDENCIA

La "*matriz de incidencia*" o "matriz de incidencia vértice-arista" de un grafo dirigido no ponderado G(V,E) con vértices (nodos) V=$\{v_1,\ldots, v_n\}$ y aristas (enlaces) E= $\{e_1,\ldots, e_m\}$ es una matriz **Q**=**Q**(G) con *n* filas y *m* columnas cuyas entradas $q_{i,j}$ satisfacen [306]

$$q_{i,j} = \begin{cases} -1 & \text{si } v_i \text{ es el origen de } e_j \\ 1 & \text{si } v_i \text{ es el destino de } e_j \\ 0 & \text{en otro caso} \end{cases}$$

En el caso de encontrarnos en presencia de un grafo no dirigido, será necesario considerar una orientación arbitraria para las aristas del grafo G y así poder construir la matriz de incidencia vértice-arista según la definición anterior [318], [321], [325], [326]. El signo positivo o negativo de la entrada de la matriz, si el enlace parte o llega al nodo/arista correspondiente, también depende del autor considerado: algunos le otorgan el signo positivo cuando el enlace parte del nodo correspondiente [318], otros cuando llega a ese nodo [306], [326] y otros lo dejan completamente al azar [321], [325]. Este último planteamiento (al azar), está refrendado por el hecho de que por lo general las redes estudiadas serán "no dirigidas" y que, por lo tanto, aún considerando las posibles orientaciones de las aristas, en la mayoría de los casos estas orientaciones ficticias tendrán que ser asignadas aleatoriamente por el investigador.

La anterior definición de matriz de incidencia es compatible con la presencia de aristas múltiples (multígrafos) y también con la existencia de lazos/autolazos. En este último supuesto la matriz de incidencia presentará todos los elementos nulos en la columna correspondiente a la arista con el lazo, excepto el número ±2 en la fila del vértice con la presencia del lazo [327].



$$I_D = \begin{array}{c} \\ u \\ v \\ w \\ x \end{array} \begin{array}{c} \begin{array}{cccccccc} a & b & c & d & e & f & g & h \end{array} \\ \begin{pmatrix} 2 & 2 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 & -1 & 1 & -1 & 0 \end{pmatrix} \end{array}$$

Ejemplo de matriz de incidencia para un grafo [327].

La matriz de incidencia **Q** ya fue estudiada por Poincarè en el año 1900 (Traducción en [328])

Unos autores denominan a la matriz de incidencia **B** [306], [318], otros **Q** [321], [325], otros **C** [326] o incluso **I** [327]. En nuestro caso la denominaremos **Q** para no equivocarla con otras matrices que definiremos posteriormente.

Algunos autores, por ejemplo [321], definen también la matriz de incidencia vértice-arista sin signo, que se puede definir y obtener de manera trivial como |**Q**| y que también podría representar directamente a redes no dirigidas [327]. También se pueden definir matrices de incidencia para grafos mixtos, que mezclan aristas dirigidas y no dirigidas [316].

En el caso de redes ponderadas la matriz de incidencia generalizada se define como [329]:

$$q_{i,j} = \begin{cases} -\sqrt{w_{ij}} & \text{si } v_i \text{ es el origen de } e_j \\ +\sqrt{w_{ij}} & \text{si } v_i \text{ es el destino de } e_j \\ 0 & \text{en otro caso} \end{cases}$$

7.2.2. MATRIZ DE ADYACENCIA

Otra matriz que puede y suele ser utilizada para “modelar” un grafo es la denominada “*matriz de adyacencia”*. La definición concreta de la matriz de adyacencia dependerá de las características específicas del grafo al que represente: si es ponderado, si es simple, si es dirigido, si tiene lazos/autolazos,… la matriz presentará unas u otras características, como veremos.

En un caso general, la matriz de adyacencia del grafo G será una matriz de dimensiones n × n que se denota como

$$\mathbf{A}=\mathbf{A}(G)=[a_{ij}]$$

Si estamos en presencia de un grafo dirigido simple se tiene que [306]

$$a_{ij} = \begin{cases} 1 & \text{si } (v_i, v_j) \in E \\ 0 & \text{en otro caso} \end{cases}$$

siendo v_i el vértice origen y v_j el vértice destino de la arista. Esto es, en la matriz de adyacencia aparece un "1" y hay una arista orientada con origen el vértice i-ésimo y destino el vértice j-ésimo, y cero en cualquier otro caso. Es una matriz asimétrica, si las aristas lo son. Expresado de una manera más funcional para redes no dirigidas, esta definición es similar a establecer [310], [330]:

$$a_{ij} = \begin{cases} 1 & \text{si existe una arista entre los vértices i y j (los nodos i y j están conectados)} \\ 0 & \text{en otro caso} \end{cases}$$

Obsérvese que en cualquier caso la matriz de Adyacencia **A** es siempre positiva, sin elementos negativos, independientemente de que sea orientada o no, o la orientación específica de las aristas.

En el caso de redes bipartitas con m nodos de "tipo 1" y n nodos de "tipo 2" el grafo equivalente puede ser representado por un caso particular de matriz de adyacencia de dimensiones $m \times n$, con la misma definición anterior [330].

Si el grafo es ponderado entonces [306], [326], [330] :

$$a_{ij} = w_{ij}$$

siendo w_{ij} valor del peso/ponderación asignado a la arista. Se puede utilizar el mismo nombre para la matriz de adyacencia **A** bien sea para la matriz de adyacencia de grafos ponderados o para no ponderados, y en general lo haremos así, salvo que induzca a error.

Se considera que el grafo bipartito es ponderado G(X,Y,W) con $\mathbf{A}=(w_{ij})$ si $w_{ij}>0$ indica la ponderación de la arista entre el vértices i y j, y con $w_{ij}=0$ si no hay aristas entre los vértices i y j [320].

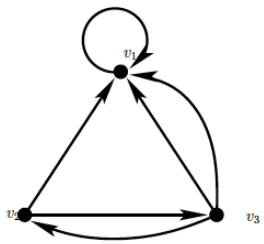
En general estos pesos o ponderaciones de las aristas satisfacen las siguientes condiciones [326]

1. $a_{ij} = a_{ij}$, con $v_i, v_j \in V$, si el grafo es no dirigido (simetría)
2. $a_{ij} \neq 0$, si y solo si v_i, v_j son adyacentes en G

3. Por lo general $a_{ij} \geq 0$, $v_i, v_j \in V$

Los grafos no ponderados (aunque no necesariamente simples) pueden ser vistos como un caso especial de grafos ponderados en los que el peso $a_{ij}=w_{ij}$ es igual al número de aristas entre los vértices v_i, v_j [326]. Si estamos en presencia de un multígrafo entonces a_{ij} será igual al número de aristas/arcos, la multiplicidad de la arista, originados en el vértice v_i y terminados en el vértice v_j. En cualquier caso, dos vértices de G se dice que son adyacentes si están conectados por una arista o arco [318], [322].

Evidentemente, si el grafo no presenta autolazos, entonces $a_{ii} = 0$, esto es, la diagonal de la matriz de adyacencia **A** será nula.



$$A = \begin{pmatrix} 100 \\ 101 \\ 210 \end{pmatrix}$$

Ejemplo de la matriz de adyacencia de un grafo. [318]

7.2.2.1. Propiedades básicas de la matriz de adyacencia

El número *n* de filas y columnas de **A** (en general se trata de una matriz cuadrada) es igual al *orden de G* (número de nodos) [310].

La matriz de adyacencia **A** depende de la ordenación seguida en la numeración asignada a los nodos [318].

Si G es un grafo no orientado su matriz de adyacencia **A** es simétrica [306].

En el caso de grafos no orientados, simples y sin ponderar, la matriz de adyacencia **A** será por lo tanto simétrica y binaria, donde $a_{ij}=1$ si el nodo i está conectado al nodo j, y 0 en el caso contrario [153].

La traza de una matriz es la suma de los valores de su diagonal y es invariante a las rotaciones ortogonales de la misma. Si en un grafo no hay nodos con enlaces a si mismos (autolazos) la traza de **A** es igual a 0 [310].

En el caso de redes no ponderadas el número de aristas conectadas a un determinado vértice i, que como hemos visto se denomina *grado del vértice* o también *valencia* [317] y se puede obtener a partir de la matriz de adyacencia **A** como [331]:

$$d_i = \sum_{\forall j} a_{ij}$$

Para la mayoría de las redes "reales" la matriz de adyacencia **A** consiste en su mayor parte de "0"s, esto es, estará escasamente poblada [310].

Existe una relación intrínseca entre las características combinatorias de un grafo y las propiedades algebraicas de su matriz de adyacencia [153]. Una importante característica de una red o grafo es el conjunto de distancias entre cualquier par de nodos i y j [310] y que será el menor número de enlaces existentes entre cualquier par de nodos i y j. Para el supuesto de que la matriz **A** sea binaria, esto es, correspondiente a un grafo simple y sin ponderar, esta distancia se puede calcular mediante potencias de la matriz **A** [312], [322], [332].

1ª potencia	**A** = **A**	por definición es la matriz de todos los pares de nodos que están vinculados/conectados/enlazados entre sí.
2ª potencia	**AA**	tiene valores no nulos en la posición i,j si el nodo j está a dos saltos del nodo i. Dado que i estará siempre a dos saltos de si mismo, la diagonal i,i cuenta el número de estos dos saltos.
3ª potencia	**AAA**	tiene un valor no nulo en la posición i,j si el nodo j está a tres saltos del nodo i.
Nª potencia	$\mathbf{A}^N$	todas las entradas son no-nulas, implicando que cada nodo ha sido alcanzado desde otro nodo.

Se define así N como el diámetro de la red, esto es, el camino más largo posible entre cualquier par de nodos que forman la red. Evidentemente calcular el diámetro de una red de a través de la potencia de una matriz es muy ineficiente: requiere la ejecución de muchos cálculos y un gran espacio de almacenamiento [310].

La matriz de adyacencia **A** es una medida local y pura de los vértices vecinos. Según algunos autores [330] esta medida local es insuficiente para proveer información sobre la estructura global del grafo al que representa.

7.2.2.2. Propiedades espectrales de la matriz de adyacencia

El espectro de la matriz de adyacencia **A** de un grafo G se denomina espectro de G [306], [318].

Aunque, como se ha dicho, **A** depende del orden de la numeración asignada a los nodos/vértices, su espectro no: el orden de la numeración equivale a intercambiar filas en la matriz de adyacencia lo que, lógicamente, no afecta ni a su determinante ni a su polinomio característico, por lo que su espectro no varía [332]. De una manera más formal [322], los autovectores de **A** son los números λ que satisfacen $\mathbf{Ax} = \lambda\mathbf{x}$ para determinados vectores no nulos $\mathbf{x} \in \Re^n$. Cada uno de esos vectores **x** se denomina autovector de la matriz **A** (o del grafo G) asociado al autovalor λ. Si λ es un autovalor de **A** entonces el conjunto $\{\mathbf{x} \in \Re^n : \mathbf{Ax} = \lambda\mathbf{x}\}$ es un subespacio de $\Re^n$ denominado autoespacio de λ o autoespacio de G. Luego renumerar los vértices de G resultará en una permutación de las coordenadas de los autovectores (y de los autoespacios).

Los grafos con el mismo espectro se denominan isoespectrales o coespectrales. Se denominan con el acrónimo inglés *PING* (*pair of isospectral non-isomorphic graphs*, par de grafos no isomórficos coespectrales) a dos grafos con los mismos autovalores pero diferente forma. Un grafo G decimos que está caracterizado por su espectro si el único grafo coespectral con G son aquellos isomórficos a G [322].

Como hemos visto, si un grafo G no tiene nodos con enlaces a si mismos (autolazos) entonces **A** tiene su traza igual a 0 y por lo tanto la suma de todos los autovalores de **A** es también 0 [310].

Algunos autovectores de la matriz de adyacencia son también muy útiles para conocer importantes propiedades de los grafos [318], [322]. Los autovalores de la matriz de adyacencia contienen información entre otras características del diámetro de la red, el grado máximo y la conectividad del grafo [310], [153].

Si se consideran grafos no dirigidos y sin lazos, la matriz de adyacencia **A** será, como hemos visto, una matriz binaria, simétrica y con un espectro real de n autovalores λ_i, donde se asume, por conveniencia, que $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. [306], [318]

Mientras que los autovalores por si solos, en general, no determinan la estructura de un grafo (como se demuestra con los grafos coespectrales), los autovalores más los autovectores SI que determinan completamente la estructura de un grafo [318]: Si

$\mathbf{u}_1,\ldots,\mathbf{u}_n$ son los autovectores linealmente independientes de la matriz de adyacencia **A** de un grafo G correspondientes a $\lambda_1,\ldots,\lambda_n$ respectivamente, entonces $\mathbf{A}=\mathbf{U}\Lambda\mathbf{U}^{-1}$ donde $\mathbf{U}=[\mathbf{u}_1|\ \mathbf{u}_2\ldots,|\mathbf{u}_n]$ es la matriz con los vectores $\mathbf{u}_i$ en columnas y Λ es la matriz diagonal con entradas λ_i esto es $\Lambda=\text{diag}(\lambda_1,\ \lambda_2,..\ \lambda_n)$. Si además la base $\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_n$ es ortonormal entonces $\mathbf{U}^{-1}=\mathbf{U}^T$ y $\mathbf{A}=\mathbf{U}\Lambda\mathbf{U}^T$ [322].

El mayor autovalor λ_n de la matriz de adyacencia **A** de un grafo G se denomina *índice de G* y dado que las matrices de adyacencia son no-negativas hay un correspondiente autovector cuyas componentes son todas no-negativas [322]. Pero no solamente eso, un grafo es conexo si y solo si su índice es un autovalor simple, con un autovector positivo [322]. El mayor autovalor λ_n de **A** es también conocido como *radio espectral* y captura la conectividad del grafo, su valor es superior al grado medio. El autovalor λ_n en términos de medida de la conectividad se comporta como el grado medio, pero está ponderado para considerar todos las longitudes de los posibles caminos. Así, cuanto mejor conectado esté un grafo, mayor será λ_n [312].

En este mismo sentido, si $\mathbf{A} \in \Re^{n\ x\ n}$ es la matriz de adyacencia de un grafo conexo y no dirigido G, entonces [332]:

- El mayor autovalor λ_n de **A** es simple.
- Todas las componentes del autovector correspondiente a λ_n son del mismo signo y no iguales a cero.

El grado medio de un grafo no dirigido y sin lazos está relacionado con el mayor autovalor de su matriz de adyacencia **A** de manera que [318]

$$\overline{d} \leq \lambda_n$$

Sean λ_i los autovalores de **A** y $\mathbf{x}_i$ los correspondientes autovectores. Cuando $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ la descomposición espectral de **A** es [153],

$$\mathbf{A} = \sum \lambda_i \mathbf{x}_i \mathbf{x}_i^T$$

Esta descomposición espectral puede utilizarse para calcular de manera más eficiente las potencias de **A** que si se realizase directamente a partir de la matriz **A**.

Sea G un grafo conexo, las siguientes afirmaciones son equivalentes [322]

1. G es bipartito
2. Si λ es un autovalor de G, entonces -λ es también un autovalor de G con la misma multiplicidad.
3. Si r es el mayor autovalor de G, entonces –r es un autovalor de G.

Si x_{ij} la j-ésima componente de $\mathbf{x}_i$ y tenemos la siguiente matriz de adyacencia:

$$\alpha_u \rightarrow \begin{bmatrix} x_{11} & \cdots & x_{i1} & \cdots & x_{k1} & \cdots & x_{n1} \\ \vdots & & \vdots & & \vdots & & \vdots \\ x_{1u} & \cdots & x_{iu} & \cdots & x_{ku} & \cdots & x_{nu} \\ \vdots & & \vdots & & \vdots & & \vdots \\ x_{1n} & \cdots & x_{in} & \cdots & x_{kn} & \cdots & x_{nn} \end{bmatrix}$$
$$\mathbf{x}_1 \qquad \mathbf{x}_i \qquad \mathbf{x}_k \qquad \mathbf{x}_n$$

Donde $\mathbf{x}_i$ es un vector columna, entonces el vector fila $\alpha_u \rightarrow (x_{1u}, x_{2u}, \ldots, x_{nu})$ representa las coordenadas del nodo u en el espacio espectral n-dimensional.

Si G(V,E) es un grafo con n vértices con matriz de adyacencia **A** y autovalores $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, si $G=G_1 \cup G_2$ es la unión de dos grafos disjuntos G_1 y G_2 entonces el espectro de G es igual a la unión de los espectros de G_1 y G_2 [318].

Sea G(V,E) un grafo con n vértices con matriz de adyacencia **A** y autovalores $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. G es bipartito si y solo si los autovalores de G ocurren en pares λ, λ' tales que $\lambda=-\lambda$' [310], [318].

7.2.3. MATRIZ DE GRADOS

Se define la matriz de grados **D** como la matriz diagonal de grados de los vértices donde [321], [326]:

$$d_{uv}=d_v.$$

También puede expresarse como la matriz $\mathbf{D}=[d_{ij}]$ en la que [330]

$$d_{ij} = \begin{cases} d_i & \text{grado del nodo i cuando i = j} \\ 0 & \text{cuando i} \neq \text{j} \end{cases}$$

En el caso de redes no ponderadas la matriz **D** describe cuántas conexiones tiene cada nodo [310].

Esta matriz **D** está asociada a la matriz de adyacencia **A** ya que **D** es una matriz diagonal con la suma por filas de la matriz **A** en su diagonal [310]. Matemáticamente la relación entre **D** y **A** se expresaría como [326]

$$d_u = \sum_u a_{uv}$$

Siendo d_u el grado del vértice $u \in V(G)$

7.2.3.1. Propiedades de la matriz de grados

La matriz de grados **D** está relacionada con la matrices de adyacencia **A**, como hemos visto, y también con la matriz de incidencia **Q** [322].

7.2.4. MATRIZ LAPLACIANA

La matriz Laplaciana de una red o grafo fue originalmente formulada por Kirchhoff en el ámbito de la teoría de circuitos [333]. Por eso es también denominada matriz de Kirchhoff o matriz de admitancia (la admitancia es la conductividad eléctrica, esto es, el recíproco de la impedancia eléctrica) [326]. No obstante la matriz Laplaciana recibe otros nombres en diferentes campos: así, es también denominada matriz de información, matriz de Zimm, matriz de Rouse-Zimm, matriz de conectividad y matriz de incidencia vértice-vértice [321].

Se denomina matriz Laplaciana ya que se corresponde con la analogía discreta del operador diferencial de Laplace, esto es, esta matriz se puede obtener cuando se discretiza el operador Laplaciano continuo ∇^2 [321], [325], [334]. Con esta relación, la matriz orientada de incidencia **Q** se corresponde con el operador gradiente ∇ y la ecuación $\mathbf{L}(G)=\mathbf{Q}\mathbf{Q}^T$ tiene entonces obviasignificación física [326].

La matriz Laplaciana de un grafo se puede obtener a través de dos formulaciones diferentes, una directamente desde las características del grafo y otra desde una expresión matemática a partir de las matrices de grado y de adyacencia:

Así, la matriz Laplaciana, **L**, es una matriz simétrica n x n asociada con un grafo ponderado de grado n y definida como [315], [335]:

$$L_{ij} = \begin{cases} d_i & i = j \\ -w_{ij} & i \neq j \end{cases} \qquad \text{para i,j=1,…,n con } d_i = \sum_j w_{ij}$$

Si G es un grafo simple, no dirigido, sin ponderar, entonces La matriz Laplaciana **L**=(l_{ij}) se define como [318], [330]:

$$l_{ij} = \begin{cases} d_i & si\ i = j \\ -1 & si\ \{i, j\} \in E \\ 0 & en\ otros\ casos \end{cases}$$

Pero la matriz Laplaciana también se puede definir como [325], [334], [336], [321], [322], [310], [330], [337], [319], [306] :

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

Por último, como hemos indicado antes, es posible obtener la matriz Laplaciana **L** a partir de la matriz de incidencia [306], [325], [322], [326], [321], [318]

$$\mathbf{L}(G)=\mathbf{Q}\mathbf{Q}^T$$

Esta última expresión es independiente de la orientación considerada, si el grafo no tiene orientación y se ha seleccionado una aleatoriamente para la determinación de **Q** [321], [326], [318]. Para el caso de considerar una matriz de incidencia vértice-arista **Q** orientada trinaria {0,+1,-1}, habitual en grafos orientados (o no) pero NO ponderados.

Para el caso de grafos ponderados la matriz Laplaciana **L** "generalizada" [338]–[340] puede obtenerse a partir de la matriz de incidencia vértice-arista binaria con signo como [339]–[344]:$\mathbf{L}=\mathbf{Q}\mathbf{W}\mathbf{Q}^T$

Siendo **W** la matriz diagonal con los pesos de las aristas del grafo, también denominada matriz de conductancia [339]–[341].

La matriz Laplaciana **L** generalizada de un grafo ponderado a partir de la matriz vértice-arista puede ser también formulada de la siguiente manera [329], [340]:

$$\mathbf{L}=(\mathbf{Q}\mathbf{W}^{0.5})(\mathbf{W}^{0.5}\mathbf{Q})^T$$

Lo que en realidad, como se ha visto, nos lleva a la consideración de una nueva matriz de incidencia vértice-arista "normalizada" ***Q*** [329], [340]:

$$\mathbf{L}=\boldsymbol{Q}\boldsymbol{Q}^T$$

Con

$$\boldsymbol{Q} = \left[q_{ij}\right] = \begin{cases} -w_{ij}^{0.5} & \text{la arista j de peso } w_{ij} \text{ sale del vértice i} \\ +w_{ij}^{0.5} & \text{la arista j de peso } w_{ij} \text{ entra en el vértice i} \\ 0 & \text{la arista j no entra/sale de vértice i} \end{cases}$$

Lo que permite obtener una matriz Laplaciana "generalizada" para grafos ponderados también a partir de la matriz de incidencia vértice-arista en esos casos. Como en el caso de la matriz de incidencia binaria "con signo", el signo en el caso de grafos ponderados no dirigidos se establece fijando un criterio cualquiera.

La Laplaciana se utiliza, por ejemplo y entre otras aplicaciones, para la partición de matrices rectangulares en el contexto del balanceo de carga en el diseño de algoritmos de multiplicación de vectores y matrices para cálculo en paralelo [320] y en la asignaciones de frecuencias en redes [329]. También aparece en múltiples problemas físicos y químicos, entre ellos como se ha dicho, en teoría de circuitos: la matriz de incidencia **Q** y la matriz Laplaciana **L** se pueden encontrar en las leyes de Kirchhoff [326], [333].

7.2.4.1. Propiedades de la matriz Laplaciana

A continuación se enumeran algunas de las propiedades de la matriz Laplaciana. A pesar de la aparente exhaustividad de la relación mostrada se trata "solamente" de algunas de las muchas propiedades que presenta la matriz Laplaciana. En las referencias indicadas pueden encontrarse más propiedades, que presentan una menor relevancia para nuestro trabajo y por ello no se exponen explícitamente aquí.

1. La suma de cada fila (y columna) de la matriz Laplaciana **L** es cero [331].
2. La matriz Laplaciana **L** es una matriz real, simétrica y sus filas suman 0 [319].
3. La matriz Laplaciana **L** es semidefinida positiva y simétrica. [336], [321].
4. La matriz Laplaciana **L** es singular y semidefinido positiva [337].
5. La existencia de posibles lazos en el grafo no tiene efecto sobre **L** [326].

Otra característica de la matriz Laplaciana es que la forma cuadrática asociada con ella es la suma ponderada de todas las distancias al cuadrado, esto es [315]

$\mathbf{x}^T\mathbf{L}\mathbf{x} = \frac{1}{2}\sum_{i,j} w_{ij}\left(\mathbf{x}_i - \mathbf{x}_j\right)^2$ A esta expresión se denomina "energía" de Hall [315] por su similitud en el aspecto con expresiones de "energía" tales como por ejemplo, la energía cinética:

$$E_c = \frac{1}{2}mv^2$$

7.2.4.2. Algunas propiedades espectrales de la matriz Laplaciana

La matriz Laplaciana **L** de un grafo y sus autovalores $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ pueden ser utilizadas en diversas áreas de investigación y presentan interpretaciones físicas en varios escenarios [326]. La matriz Laplaciana **L** depende obviamente de G y también de la ordenación (arbitraria) de sus vértices. No obstante, matrices Laplacianas obtenidas de diferentes ordenaciones de los vértices de un mismo grafo son permutaciones-similares. Esto es, los grafos G_1 y G_2 son isomórficos si y solo si existe una matriz de permutación **P** tal que $L(G_2)=\mathbf{P}^T L(G_1)\mathbf{P}$ [321].

Sean los autovalores de la matriz Laplaciana $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ con n=|V(G)|, numerados en orden creciente y repetidos conforme a su multiplicidad. Denotaremos $\lambda_k(G)$ el k−ésimo autovalor de G (considerando multiplicidades). La letra n indica el orden de G, así $\lambda_n(G)$ será el mayor autovalor de **L**(G) [326].

Sea G un grafo (ponderado) no dirigido, con todos sus pesos no-negativos, entonces [326], [345], [336], [319], [322], [315], [306]:

1. **L**(G) tiene solo autovalores reales
2. **L**(G) es semidefinida positiva
3. **L**(G) es simétrica
4. El menor autovalor es $\lambda_1=0$ y su correspondiente autovector es $(1,1,..,1)^T$. La multiplicidad de 0 como autovalor de **L**(G) es igual al número de componentes de G.
5. $\lambda_1=0$ y $\lambda_2>0$ si y solo si G está conectada (es conexo).

Un grafo conexo **L** tiene solo un autovalor nulo con autovector $\mathbf{1}=(1,1,\ldots,1)^T$ [315], [331] u ortonormal $\mathbf{v}_1=(1/\sqrt{n})\mathbf{1}$ [315] así la matriz **L** tiene rango n-1, siendo n el número de aristas del grafo [346].

Para la matriz Laplaciana los más importantes autovectores corresponden con los menores autovalores [310].

Todos los autovalores de la matriz Laplaciana son positivos, ya que es simétrica e igual al cuadrado de la matriz de incidencia, por lo que sus autovalores son todos el cuadrado de "vectores" [sic] reales [331]. Si G es un grafo simple conectado con λ ($0<\lambda<1$) un autovalor de L(G), entonces el diámetro de G es al menos 3 [319].

Sea G un grafo simple conectado de orden n (>2). Si G tiene vértices pendientes, entonces el menor autovalor no nulo es menor o igual que 1. Es más, el menor autovalor no nulo es estrictamente menor que 1 si el vértice pendiente no es adyacente al vértice de mayor grado [319].

Un grafo G consiste en k componentes conectados si y solo si $\lambda_1(L) = \ldots = \lambda_k(L)=0$ y $\lambda_{k+1}(L) > 0$ [318], [334]

Los autovalores de **L** recogen información sobre la estructura de árbol de G. El espectro de **L** contiene un 0 por cada componente conectado. No hay la misma manera de descubrir el número de componentes de una red a partir del espectro de **A** [310].

Hay mucha literatura sobre segmentación espectral de grafos, en la que se relacionan propiedades de los grafos/redes con el espectro de la matriz Laplaciana [331].

Se consideran grafos simples no dirigidos. Sea G un grafo conectado con n vértices y sea $0=\lambda_1<\lambda_2\leq\ldots\leq\lambda_n$ los autovalores de su matriz Laplaciana, entonces la distancia media entre pares de vértices será [345]

$$\overline{d}_G \geq \frac{2}{n-1}\sum_{i=2}^{n}\frac{1}{\lambda_i}$$

Cumpliéndose la igualdad si y solo si G es un árbol.

La matriz Laplaciana de un grafo G está relacionada con los *spanning trees* de la siguiente manera [318]:

1. Para cada $i \in \{1,\ldots,n\}$ el número de *spanning trees* en G es igual al $|\det(\mathbf{L}_i)|$, donde $\mathbf{L}_i$ se obtiene eliminando la fila y columna i de la matriz Laplaciana **L**.

2. El número de *spanning trees* es igual a (1/n) $\prod_{i \geq 2} \lambda_i(L)$

Se han determinado los siguientes límites para los autovalores, sea G un grafo de orden n, entonces [326]:

1. $\lambda_2 \leq \frac{n}{n-1} \min\{d_v; v \in V(G)\}$

2. $\lambda_n \leq \max\{d_u + d_v; uv \in E(G)\}$

3. Si G es un grafo simple, entonces $\lambda_n \leq n$, cumpliéndose la igualdad si y solo si el complemento de G es no conectado.

Sea G un grafo conectado con diámetro d. Supongamos que **L**(G) tiene exactamente k distintos autovalores, entonces d+1 ≤ k.

7.2.4.3. CONECTIVIDAD ALGEBRAICA Y VALORACIÓN CARACTERÍSTICA

Supongamos que **L=L**(G)=**QQ**^{T}, con **Q** la matriz de incidencia vértice-arista, tiene autovalores 0= $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Fiedler denominó *conectividad algebraica* a(G) al segundo autovalor menor de **L** [317]

$$\lambda_2 = a(G)$$

La definición completa que Fiedler expone es [317]: Sea G(V,E) un grafo finito no dirigido, sin lazos ni aristas múltiples. Sea la matriz cuadrada **L** cuyos elementos fuera de la diagonal son $a_{jk}=a_{kj}=-1$ si entre los vértices j,k existe una arista y $a_{jk}=a_{kj}=0$ si no y cuyos elementos de la diagonal principal a_{ii} son iguales a las valencias del vértice i. Esta matriz **L**(G) es frecuentemente utilizada para enumerar los árboles de expansión del grafo G, es simétrica, singular (todas las filas suman 0) y semidefinido positiva. **L**(G)=**QQ**^{T} con **Q** la matriz (0,1,-1) de adyacencia vértice-arista de un grafo G arbitrariamente dirigido. Sean $0=\lambda_1 \leq \lambda_2=a(G) \leq \lambda_2 \leq \ldots \leq \lambda_n$ los autovalores de la matriz **L**(G). a(G) es cero sí y solo sí el grafo G está no conectado (inconexo). El segundo menor autovalor a(G) de la matriz **L**(G) se denomina *conectividad algebraica* del grafo G. Este autovalor es el más importante de la matriz Laplaciana de un grafo G y, como veremos, está relacionado con el diámetro y la distancia media del grafo, entre otras características del mismo. [326].

Si λ_2 es próximo a 0 el grafo está casi desconectado, mientras que si $\lambda_2 >> \lambda_1$ (salto en los autovalores) el diámetro del grafo es pequeño [310].

Las propiedades que presenta la *conectividad algebraica* son numerosas [317], [347], [334], [326], [318], [325], [321]. A continuación se muestran las propiedades más representativas de la *conectividad algebraica*:

a) Para un grafo completo K_n con n vértices, se tiene que $a(K_n)=n$

b) Si G es un grafo con n vértices que no es completo, entonces $a(G) \leq n-2$

c) $a(G) \geq 0$, $a(G)=0$ si y solo si G es no conexo.

d) Si $G_1(V,E_1)$, $G_2(V,E_2)$ y $E_1 \subset E_2$, entonces $a(G_1) \leq a(G_2)$.

e) $G_1(V,E_1)$, $G_2(V,E_2)$ y $E_1 \cap E_2 = \varnothing$, entonces $a(G_1)+a(G_2) \leq a(G_3)$ con $G_3(V, E_1 \cup E_2)$

f) Si G_1 se obtiene de G eliminando k vértices (y las aristas incidentes), entonces $a(G_1) \geq a(G)-k$

g) Si m es la valencia mínima de un grafo G no completo, entonces $a(G) \leq m$

h) $0 \leq a(G) \leq n$, con n en número de vértices del grafo G(V,E)

i) Si $G_1(V,E_1)$ es un subgrafo de G, entonces $a(G_1) \leq a(G)$

j) Si G es un árbol, entonces $0 < a(G) \leq 1$, y $a(G)=1$ si G es la estrella $K_{1,n-1}$

k) Si $G \neq K_n$ entonces $a(G) \leq v(G) \leq e(G)$ donde v(G) es la conectividad de los vértices de G y e(G) es la conectividad de las arista de G, esto es el mínimo número de nodos y vértices, respectivamente, que hay que quitar para hace G desconectado

l) Sea G un grafo conectado con diámetro d, entonces $a(G) \geq 4/dn$

m) Sea G un grafo conectado con máximo grado de un vértices d_1, entonces $[2d_1/a(G)]^{1/2}\log_2(n^2)$ es un límite superior para el diámetro de G.

La conectividad algebraica está relacionada con los valores singulares de la matriz de incidencia [347]. Hemos visto que $\mathbf{L}(G)=\mathbf{QQ}^T$ siendo **Q** la matriz cuyas filas

corresponden a los vértices de G y cuyas columnas corresponden a las e_s aristas dirigidas con q_{is}=1, q_{js}=-1 con e_s una arista de G y e_s=(i,j).

Se puede también definir la matriz $\mathbf{K}(G)=\mathbf{Q}^T\mathbf{Q}$ que es de nuevo simétrica y semidefinido positiva y cuyas filas y columnas corresponden a aristas de G [347]. Esta matriz depende de la orientación de las aristas: un cambio en la orientación resulta en la multiplicación de la correspondiente columna de **K** por -1. También un cambio de numeración conlleva una permutación simultánea de filas y columnas de **K**.

Es conocido que las matrices $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$ tienen los mismos autovalores no nulos incluyendo multiplicidades. Como sabemos, las raíces cuadradas de los autovalores de $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$ se denominan valores singulares de la matriz **Q** [158].

Análogamente la conectividad algebraica a(G) de un grafo conexo es igual al menor autovalor positivo de la matriz $\mathbf{K}(G)=\mathbf{Q}^T\mathbf{Q}$ e igual al cuadrado del menor valor singular positivo de la matriz de incidencia **Q**, para cualquier orientación de G [347].

Si consideramos grafos ponderados el segundo menor autovalor de la matriz Laplaciana es análogamente denominada conectividad algebraica de G [338].

Valoración Característica

Fiedler también acuñó el término *valoración característica* [317] en referencia al autovector $\mathbf{x_2}$ de **L**(G) correspondiente a λ_2 = a(G), que es el primer autovector no trivial de **L**, reflejando el hecho de que un autovector es una función o una etiqueta de los vértices de G [334]. La estructura del autovector $\mathbf{x_2}$ de λ_2 también es especial, siendo principalmente interesante la estructura de sus signos [326]. Algunos autores denominan a la valoración característica también *vector de Fiedler* [331].

En un artículo del año 1975 Fiedler presenta con más detalle la definición de *valoración característica* [338]. Bajo las mismas condiciones anteriores (grafos finitos, no dirigidos, sin lazos ni múltiples aristas) en este artículo se centra en el autovector $\mathbf{x_2}$ asociado a la *conectividad algebraica* que como hemos indicado denomina *valoración característica*. Las coordenadas de este autovector $\mathbf{x_2}$ se asignan a los vértices de G de manera natural y pueden ser considerados una "valoración" de los vértices de G. Se denomina a esta valoración "*valoración característica de G*" y es siempre no-nulo.

Fiedler sugirió que el autovector $\mathbf{x_2}$ asociado con el segundo menor autovalor λ_2 puede utilizarse para resolver el problema denominado de "*min-cut*": separar la red en dos

conjuntos de nodos aproximadamente iguales con el menor número de conexiones entre ellos, basándose en los signos de las componentes de $\mathbf{x_2}$ [310].

La forma cuadrática asociada con la Laplaciana que vimos anteriormente como la suma ponderada de todas las distancias al cuadrado, esto es [315]

$$\mathbf{x}^T\mathbf{L}\mathbf{x} = \frac{1}{2}\sum_{i,j} w_{ij}\left(\mathbf{x}_i - \mathbf{x}_j\right)^2$$

que se denomina Energía de Hall toma su valor mínimo precisamente para la valoración característica $\mathbf{x} = \mathbf{x_2}$ [347]

7.2.5. MATRIZ LAPLACIANA SIN SIGNO

La matriz Laplaciana sin signo se define como [330], [319]:

$$|\mathbf{L}|=\mathbf{D}+\mathbf{A}$$

O bien

$$\left|L\right|_{ij} = \begin{cases} w_i & \text{cuando i = j} \\ 1 \quad \text{o } w_{ij} & \text{cuando i} \neq \text{j y i es adyacente a j con } w_i = \sum_j w_{ij} \\ 0 & \text{en otro caso} \end{cases}$$

7.2.5.1. Propiedades de la matriz Laplaciana sin signo

Si G es un grafo conectado entonces $|\mathbf{L}|(G)$ es una matriz no negativa, simétrica e irreducible [319].

La matriz Laplaciana sin signo permite la existencia de autolazos: si un nodo está conectado a si mismo, esto se reflejará como un incremento de 2 en su correspondiente grado para una red no dirigida. Las representaciones que permiten autolazos son importantes en algunos campos, por ejemplo neurología [330].

La principal razón para utilizar la matriz Laplaciana en cualquiera de sus versiones (con o sin signo) es la manera en la que sus elementos codifican la conectividad: nos permite formular una medida de la fuerza relativa de asociación entre nodos [330].

7.2.5.2. Propiedades espectrales de la matriz Laplaciana sin signo

Todos los autovalores de $|\mathbf{L}|(G)$ son no negativos. Sean $0 \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n$ los autovalores de $|\mathbf{L}|(G)$ [319].

Hay una conexión entre los autovalores de $|\mathbf{L}|(G)$ y los autovalores de $\mathbf{A}(G)$ [319]

Supongamos que G es un grafo conectado y que μ_n es el mayor autovalor de $|\mathbf{L}|(G)$. Es bien conocido que todos las componentes del autovector correspondiente al autovalor μ_n de $|\mathbf{L}|(G)$ tienen el mismo signo (y son no nulas). Podemos asumir que son todas positivas. En la literatura se puede encontrar limites para los autovalores de los grafos. Aquí caracterizaremos los grafos por su mayor autovalor μ_n de $|\mathbf{L}|(G)$ [319].

Sea $\mathbf{x}=(x_1,x_2,\ldots,x_n)^T$ un autovector correspondiente al autovalor μ_n de $|\mathbf{L}|(G)$. Si x_i es la mayor componente de dicho autovector, entonces el grado del vértice v_i es mayor o igual a $\mu_n/2$ [319].

Si $\mu_n=d_n+d_{n-1}$ (con $d_n \neq d_{n-1}$) entonces [319]

1. Los vértices correspondientes a la mayor y segunda mayor componente del autovector, son adyacentes.

2. La segunda mayor componente es mayor o igual a $(d_{n-1}/d_n)x_i$ con x_i la mayor componente.

Sea G un grafo conectado, entonces $\mu_n=d_n+d_{n-1}$ si y solo si G es una estrella o un grafo regular [319].

Sea G un grafo conectado, entonces $\mu_n \leq \max\{d_i+m_i \mid v_i \in V\}$ verificándose la igualdad si y solo si G es un grafo regular o un grafo semiregular bipartito [319].

7.2.6. MATRIZ LAPLACIANA NORMALIZADA

La Laplaciana normalizada de G es una matriz $n \times n$ definida como [306], [337]

$$\mathscr{L} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$$

donde $\mathbf{D}^{-1/2}$ es la matriz diagonal donde la i-ésima entrada de la diagonal es 0 si $d_i=0$ y $1/\sqrt{d_i}$ en otro caso.

Esta matriz se denomina Laplaciana normalizada $\mathscr{L}=\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$. Para grafos simples $\mathscr{L}=(l_{ij})$ verifica [318]

$$\bar{l}_{ij}=\begin{cases}1 & \text{si } i=j \text{ y } d(i)>0\\ \dfrac{-1}{\sqrt{d_i d_j}} & \text{si } \{i,j\}\in E\\ 0 & \text{en otro caso}\end{cases}$$

También se la denomina como **M** o matriz de correlación de un grafo [321] ya que una matriz simétrica semidefinido positiva es una *matriz de correlación* si sus elementos de la diagonal son todos iguales a 1 [321]]

$$\mathbf{M}=M(G)=\mathbf{D}(G)^{-1/2}\mathbf{L}(G)\mathbf{D}(G)^{-1/2}$$

Análogamente se tiene que para un grafo cualquiera si **A** es la matriz de adyacencia, entonces

$$\mathscr{L}=\mathbf{I}-\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$$

Igualmente, si **S** es la matriz cuyas filas están indexadas por los vértices y cuyas columnas están indexadas por las aristas de G, de tal manera que cada columna correspondiente a una arista $e=\{u,v\}$ tiene el valor $1/\sqrt{d_u}$ en la fila correspondiente a u, un valor $-1/\sqrt{d_v}$ en la fila correspondiente a v, y cero en los otras posiciones, se tiene:

$$\mathscr{L}=\mathbf{S}\mathbf{S}^{T}$$

Si $\mathscr{L}$ es simétrica, sus autovalores son todos reales y no-negativos. Se puede igualmente probar que 0 es un autovalor de $\mathscr{L}$. Los autovalores de $\mathscr{L}$ se denotan como $0=\lambda_0\leq\lambda_{01}\leq\ldots\leq\lambda_{n-1}$. Al conjunto de los λ_i se le denomina espectro de $\mathscr{L}$.

7.2.6.1. Algunas propiedades espectrales de la matriz Laplaciana Normalizada

El espectro de la matriz Laplaciana normalizada nos permite reconocer tanto estructuras bipartitas como componentes conectados [318]

De nuevo $\mathscr{L}$ es simétrica con valores reales, y es posible ordenar sus n autovalores en una secuencia $\lambda_1(\mathscr{L}) \leq \lambda_2(\mathscr{L}) \leq \ldots \leq \lambda_n(\mathscr{L})$ [306], [318].

Si G es un grafo con matriz Laplaciana normalizada $\mathscr{L}$ se tiene que

1. $\lambda_1(\mathscr{L}) = 0$ y $\lambda_n(\mathscr{L}) \leq 2$

2. G es bipartito si y solo si para cada $\lambda(\mathscr{L})$, el valor $2-\lambda(\mathscr{L})$ es también un autovalor de $\mathscr{L}$

3. Si $\lambda_1(\mathscr{L}) = \ldots = \lambda_i(\mathscr{L}) = 0$ y $\lambda_{i+1}(\mathscr{L}) \neq 0$ entonces G tiene exactamente i componentes conectadas.

Así pues, el espectro de la matriz de adyacencia, Laplaciana, normalizada Laplaciana de un grafo no-dirigido G(V,E) contendrá n valores reales.

7.2.7. MATRIZ DE ADYACENCIA NORMALIZADA

Si **A** es la matriz de adyacencia y **D** es la matriz diagonal de grados entonces se define la matriz de adyacencia normalizada $\mathscr{A}$ como [337]

$$\mathscr{A} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$$

7.2.7.1. Propiedades de la matriz de adyacencia normalizada

Se tiene que

$$\chi = \mathbf{D}^{-1/2}\,\mathscr{A}\,\mathbf{D}^{-1/2}$$

con lo que $\chi^2 = \Sigma_j v_i^2 / \Sigma_i a_{ij}$ omitiendo el término $v_0 = 1$ esperado para $\mathbf{n}_i = \mathbf{1}$ [310].

7.2.8. MATRIZ NORMAL

La matriz Normal se define como [310]

$$\mathbf{N} = \mathbf{D}^{-1}\mathbf{A}$$

Siendo **D** la matriz de grados. La matriz **N** es la Laplaciana de una red de nodos ponderada por su grado [310].

$\mathbf{N}^2$ es similar a la matriz χ^2 , tratando **A** como tabla de contingencia [310].

7.2.8.1. Algunas propiedades de la matriz normal

Las filas de **N** suman 1 (matriz estocástica) [310].

La matriz **N** es una Laplaciana generalizada, con una definición de ortonormalidad diferente [310].

7.2.8.2. Algunas propiedades espectrales de la matriz normal

A continuación se exponen alguna propiedades espectrales de la matriz Normal [310]:

- Los autovalores de **N** encierran información sobre la estructura de G.
- Presenta un autovector trivial n_0 con autovalor v_0=**1**
- El espectro de **N** está acotado $1 = v_0 \geq v_1 \geq v_{m-1} \geq -1$
- El espectro de N contiene un 1 por cada componente conectado
- El autovalor -1 solo ocurre si G es bipartito, en cuyo caso los autovalores de **N** ocurren en pares.

7.2.9. MATRIZ DE ENRUTAMIENTO

Sea G(V,E) un grafo dirigido que representa una red, con V el conjunto de los N_v =n nodos de la red y E el conjunto de los N_e =m enlaces de la misma. Sea P el conjunto de todos los N_p caminos posibles en G. Sea **B** la matriz de enrutamiento, formada en sus filas por los enlaces de la red (N_e), y sus columnas las parejas origen-destino (N_p), de forma que, si solo hay una ruta para ir de *i* a *j* se tiene que: [348]

$$\mathbf{B}_{e;ij} = \begin{cases} 1, & \text{si el enlace } e \text{ es atravesado para ir de } i \text{ a } j \\ 0, & \text{en otro caso} \end{cases}$$

En el supuesto de que más de un enlace pueda ser utilizado de forma simultánea para una misma ruta, el valor $\mathbf{B}_{e;ij}$ será el porcentaje en tanto por uno del tráfico cursado por cada uno de esos enlaces.

7.2.9.1. Algunas propiedades de la matriz de enrutamiento

Sea **y** una métrica medida sobre los N_p diferentes caminos existentes en G. Consideremos que el valor de esta métrica está linealmente relacionado con **x**, que es

el valor de esa misma métrica, pero medida sobre los N_e diferentes enlaces. Bajo este supuesto se tiene que [348]:

$$\mathbf{y} = \mathbf{Bx}$$

Son métricas que presentan esa característica de aditividad tanto el retardo, como el tráfico, y por lo tanto podrán ser **y**.

Incluso bajo determinados supuestos de independencia entre enlaces, la probabilidad de pérdida de paquetes también verifica la igualdad **y** = **Bx**, pero en este caso en escala logarítmica:

$$y_i = \log(1-\alpha_i) \qquad x_j = \log(1-\beta_j)$$

Con α la probabilidad de pérdida de paquetes en caminos.
β la probabilidad de pérdida de paquetes en enlaces.

Así pues, las métricas más importantes utilizadas en redes de telecomunicación se pueden considerar dentro de este supuesto, ya que verifican la condición de aditividad entre enlaces antes expuesta [349].

Para analizar la relación que juegan algunos enlaces sobre otros en el enrutamiento de las diferentes rutas o caminos se puede observar que los elementos de la diagonal de la matriz $\mathbf{B}^T\mathbf{B}$ son precisamente el número de rutas cursadas sobre el respectivo enlace y por lo tanto constituye un posible indicador de la centralidad de dicho enlace en la red. Los elementos fuera de la diagonal miden el número de rutas cursadas simultáneamente sobre pares de enlaces y podría ser interpretado como una medida de "co-centralidad" [291].

El rango de la matriz de enrutamiento **B,** tal y como ha sido definida anteriormente, es generalmente igual al número de rutas independientes en la red, que suele ser mucho menor que el número total de rutas posibles en la red N_p [348]. No se conoce muy bien los motivos de ese fenómeno, pero en todo caso será importante conocer si esta propiedad es "robusta", esto es, cómo se verá afectada por, al menos, dos fenómenos [348]:

- Si un enlace falla, ¿cómo se verá modificará la compartición de enlaces por las rutas?

- Si las métricas de los enlaces no presentasen igual varianza (homocedasticidad), ¿cómo se reducen las ventajas de la compartición de los enlaces?. El método propuesto se basa en la hipótesis de homocedasticidad de las métricas de los enlaces, pero esta hipótesis no siempre se cumple en la práctica, por lo que será necesario conocer sus efectos.

7.2.9.2. Algunas propiedades espectrales de la matriz de enrutamiento

Podemos aplicar la Descomposición en Valores Singulares [158] a la matriz de enrutamiento **B** como [350]

$$\mathbf{B}^T=\mathbf{U}\Lambda^{0.5}\mathbf{V}^T$$

Con **V** matriz de autovectores de $\mathbf{BB}^T$ ortogonal de dimensiones N_e x N_e

Λ matriz diagonal N_e x N_e conteniendo los autovalores de $\mathbf{BB}^T$ (o $\mathbf{B}^T\mathbf{B}$)

U matriz de dimensión IJ x N_e tal que $\mathbf{U}^T\mathbf{U}=\mathbf{I}$ con IJ el número de pares OD.

Sabemos que la ecuación anterior también puede ser escrita como [350]

$$\mathbf{B}^T = \sum_{k=1}^{N_e} \lambda_k^{1/2}\mathbf{u}_k\mathbf{v}_k^T$$

con λ_k la k-ésima entrada mayor de la diagonal de Λ y $\mathbf{u}_k$ y $\mathbf{v}_k$ la k-ésima columna de **U** y **V**, respectivamente. Formalmente descompone una matriz de rango N_e en la suma de N_e matrices de rango 1.

Si los valores menores de λ_k son “pequeños” en comparación con los mayores, el sumatorio anterior puede recortarse, obteniendo una aproximación de la matriz $\mathbf{B}^T$ (o de **B**) de dimensión reducida [159].

La dependencia entre las rutas en G puede estudiarse a partir de los vectores singulares $\mathbf{v}_i$ de la matriz de enrutamiento **B** de una red ya que cada ruta de G es una suma ponderada de diferentes $\mathbf{v}_i$. Como además los diferentes $\mathbf{v}_i$ son ortogonales capturan los patrones independientes entre las diferentes rutas en G con el respectivo valor singular λ_i representando la contribución de cada ruta al conjunto de todas las rutas. Es más, cada componente del vector singular está asociado a un enlace, así que es posible, por ejemplo, su representación georeferenciada como veremos posteriormente [348].

La matriz $\mathbf{B}^T\mathbf{B}$ es simétrica y sus autovalores son todos positivos y el rango de **B** es igual a número de caminos considerados independientes [349], [348]. Un número reducido de autovalores de la matriz $\mathbf{B}^T\mathbf{B}$ serán mucho mayores que el resto. En el espectro de $\mathbf{B}^T\mathbf{B}$ puede aparecer un patrón de "emparejamiento" entre los dos autovalores mayores, el emparejamiento aparece si la matriz de enrutamiento es simétrica [349].

En un artículo de 2006 Chua *et al* [291] apuntan a la asociación entre las características espectrales de la red y algunos parámetros de conectividad de los nodos en la red, como por ejemplo, con la centralidad. En concreto apuntan la vinculación entre la caída del espectro de la matriz de enrutamiento **B** y determinadas métricas de la estructura topológica de la red. Demuestran que los saltos espectrales entre valores singulares de la matriz $\mathbf{B}^T\mathbf{B}$ en el peor de los casos igualan la caída en la conectividad de los vértices:

$$\frac{\lambda_k}{\lambda_1} \leq \frac{\left[\mathbf{B}^T\mathbf{B}\right]_{k,k}}{\left[\mathbf{B}^T\mathbf{B}\right]_{1,1}} \mathrm{diam}\left(G\right)$$

Donde diam(G) es el diámetro del grafo de red G, $[\mathbf{B}^T\mathbf{B}]_{i,i}$ es el elemento i-ésimo de la diagonal de $\mathbf{B}^T\mathbf{B}$ y asumimos que las aristas del grafo (enlaces) han sido ordenados de forma que $[\mathbf{B}^T\mathbf{B}]_{1,1} \geq \ldots \geq [\mathbf{B}^T\mathbf{B}]_{Ne,Ne}$

Además se verifica que

$$\lambda_k \leq \left[\mathbf{B}^T\mathbf{B}\right]_{k,k} \mathrm{diam}\left(G\right)$$

Chua *et al* [349] estudian también el efecto de la desigualdad de varianza entre los diferentes enlaces de la red (hipótesis de homocedasticidad). El método propuesto utiliza la descomposición

$$\mathbf{BC} = \mathbf{U}\boldsymbol{\Lambda}^{0.5}\mathbf{V}^T$$

En donde, recordemos, las columnas de la matriz **U** forman una base ortogonal de **GC** con la importancia relativa de cada columna indicada por la magnitud del correspondiente valor singular en la matriz diagonal $\boldsymbol{\Delta}$. Estos valores singulares son, como sabemos, la raíz cuadrada de los autovalores de $(\mathbf{BC})^T(\mathbf{BC})$. **C** es una matriz no-singular derivada de la factorización $\boldsymbol{\Sigma} = \mathbf{CC}^T$ con $\boldsymbol{\Sigma}$ la matriz de covarianzas de **x**. Así

BC corresponde con los enlaces de la red reescalados por la variabilidad del correspondiente enlace (si $\boldsymbol{\Sigma}$ es una matriz diagonal, **C** también lo es).

En el espectro de $\mathbf{B}^T\mathbf{B}$ puede aparecer un patrón de "emparejamiento" entre los dos autovalores mayores, patrón que no aparece en el espectro de $(\mathbf{BC})^T(\mathbf{BC})$ en el caso de que $\boldsymbol{\Sigma}$ sea una matriz diagonal. El emparejamiento aparece si la matriz de enrutamiento es simétrica y la desaparición en el segundo supuesto se debe a la desigualdad de las varianzas entre los diferentes enlaces [349].

7.2.10 MATRIZ DE ORIGEN-DESTINO O MATRIZ DE TRÁFICO

Se define la matriz OD (Origen-Destino) Z también denominada "matriz de tráfico" como [348]:

$$\mathbf{Z} = (Z_{ij})$$

Siendo Z_{ij} el volumen total de tráfico cursado entre el origen *i* y el destino *j*.

Se pueden también definir:

$Z_{i+} = \sum_{\forall j} Z_{ij}$ como el flujo saliente del nodo *i*.

$\mathbf{X} = (\mathbf{x}_e)$ con $\mathbf{x}_e$ el vector que representa el flujo total sobre el enlace *e.*

Con lo que se tiene que

$$\mathbf{X} = \mathbf{BZ}$$

Por lo general **Z** es variable en el tiempo, pero **B** se considera habitualmente fija.

En [350] pueden encontrarse referencias sobre estimación de la matriz de tráfico.

7.2.11 OTRAS MATRICES DE INTERÉS

Hemos visto anteriormente que se define la matriz

$$\mathbf{K}=\mathbf{K}(G)=\mathbf{Q}^T\mathbf{Q}$$

que puede ser vista como la *versión de aristas de la matriz Laplaciana* [321]. Ambas matrices **L** y **K** tienen capacidad para explicar ciertas propiedades estáticas y dinámicas de algunas moléculas. **K** depende de la orientación que se le de al grafo.

Se ha demostrado que existe una conexión entre $\mathbf{L}(G)=\mathbf{Q}\mathbf{Q}^T$ y $\mathbf{K}(G)=\mathbf{Q}^T\mathbf{Q}$ que explica simultáneamente las propiedades estadísticas y dinámicas de determinados redes de polímeros [325], [334].

Por razones obvias, tanto **L**(G) como **K**(G) comparten los mismos autovalores no nulos [325], [334].

La *matriz de distancias* de un grafo conexo G(V,E) es

$$\Delta(G)=[d(v_i,v_j)]$$

de dimensiones $n\times n$ y cuya entrada (i,j) es precisamente la distancia entre v_i y v_j.

$\Delta(G)$ es simétrica con ceros en su diagonal principal. Esta matriz tiene aplicaciones químicas directas [321], [334] y fue estudiada por Cayley ya en el siglo XIX [351].

7.3. COMPARATIVA ENTRE LAS PRINCIPALES MATRICES

Seary y Richards en 2003 [310] introducen tres tipos de análisis espectral de grafos en función de si operan sobre **A**, **L**, o **N** y describen sus propiedades matemáticas así como las fortalezas y debilidades de cada uno.

Muchas propiedades relativas al espectro de **A** cuando G es k-regular, son ciertas para **L** y **N**, en algunos casos incluso cuando G no es regular. No obstante **L** y **N** son más representativas de los grafos ya que para grafos que no sean k-regulares los autovalores-autovectores de **A** no proporcionan una buena representación del grafo.

La Laplaciana **L** proporciona una buena representación visual de grafos que sean productos cartesianos (tales como cuadrículas e hipercubos), mientras que la Normal **N** provee una buena representación para grafos que sean productos de Kronecker (grafos consistentes en bloques) [310].

La matriz **A** puede ser utilizada para descubrir particiones de clusters/agrupaciones de nodos altamente conectados, pero estos métodos no son tan generales o claros como aquellos derivados de **L** o **N**.

Los autovectores de **A** están muy localizados en nodos con elevados grados, y sugiere que este efecto puede ser utilizado para distinguir cierto tipo de redes. Este efecto no sucede en **L** o **N** dado que se introduce cierto control sobre el grado, y por lo tanto no se pueden utilizar estos espectros para distinguir aquellas redes. El mayor problema que presentan **L** y **N** es su sensibilidad a las "rutas largas" y especialmente a árboles colgados del cuerpo principal de la red. Para **N**, esto puede ser interpretado como nodos que son de alcance difícil (distantes) en un paseo aleatorio. Para rutas largas internas a la red, este efecto es realmente una ventaja, dado que estos ciclos se detectan como "localmente bipartitos" y enfatizan autovectores importantes. Nodos de dichos caminos tienen efectos importantes en propiedades globales tales como el diámetro.

No obstante, mientras que el espectro de la matriz Laplaciana **L** tiene la ventaja sobre el espectro de la matriz de adyacencia **A** para identificar el número de componentes conectadas del grafo, falla para identificar estructuras bipartitas [318].

Sea G un grafo con matriz de adyacencia **A** y matriz Laplaciana **L**. Si Δ y δ son el grado máximo y mínimo de los vértices de G, entonces el k-ésimo menor autovalor $\lambda_k(A)$ de **A** y el k-ésimo mayor autovalor $\lambda_{n+1+k}(L)$ de **L** están relacionados por [318]

$$\delta - \lambda_k(A) \leq \lambda_{n+1+k}(L) \leq \Delta - \lambda_k(A)$$

En el caso de gráficos regulares, existe una relación directa entre el espectro de la matriz Laplaciana **L** y el espectro de la matriz de adyacencia **A**, ya que los polinomios característicos de ambas matrices están relacionados. También entre grafos no regulares puede obtenerse una relación entre ambos espectros [326].

7.4. COMPARATIVA CON TABLAS CONTINGENCIA Y ANALISIS DE CORRESPONDENCIAS

El Análisis de Correspondencias [352], [353] está específicamente diseñado para el análisis de relaciones entre dos modos, así existe una conexión entre el análisis de grafos en general y bipartitos en particular y el análisis de correspondencias [354], [320], [355].

Las redes de dos modos, que combinan dos tipos diferentes de nodos y conexiones, pueden ser representadas como redes bipartitas para las que el uso de las matrices **A** y **N** está más ajustado. Estas matrices tienen espectros simétricos al ocurrir los autovalores en pares con signos opuestos. En este caso de redes de dos modos no precisamos trabajar con la matriz completa, podemos operar con la representación rectangular e inferir las partes perdidas de la descomposición. Por ejemplo, si asumimos m_1 personas y m_2 eventos, los autovectores resultantes consistirán en m_1 componentes para las personas seguidas de m_2 componentes para los eventos. Los bloques resultantes estarán estrictamente fuera de la diagonal y de nuevo los autovectores de **N** permitirán una solución superior maximizando χ^2. De hecho esta solución es idéntica a la obtenida por Análisis de Correspondencias, técnica estadística utilizada para descubrir patrones en datos de dos modos [310].

La matriz χ^2 se define en términos de las marginales (suma) de las filas y columnas en la que un elemento típico es [310]

$$\chi^2_{ij} = \frac{\left(\text{Observada}_{ij} - \text{Esperada}_{ij}\right)^2}{\text{Esperada}_{ij}}$$

Para una matriz poco poblada **A** se puede considerar χ que tiene como elemento típico

$$\chi_{ij} = \frac{\text{Observada}_{ij} - \text{Esperada}_{ij}}{\sqrt{\text{Esperada}_{ij}}} \quad \text{con} \quad \text{Esperada}_{ij} = \frac{deg(i)\,deg(j)}{\sum deg(i)}$$

lo que mantiene la nueva matriz como poco poblada [310].

$$\chi = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$$

con lo que

$$\chi^2 = \Sigma_j v_i^2 / \Sigma_i a_{ij}$$

omitiendo el término $v_0=1$ esperado para $\mathbf{n}_i=\mathbf{1}$.

Esta ecuación muestra cuánto contribuye cada dimensión a χ^2 ,lo que es una medida de la dependencia entre filas y columnas. En esta interpretación, si v_1 es pequeña ($v_1 << v_0=1$) entonces χ^2 es también pequeña: no existe relación entre filas y columnas de **A** y no hay “señal” por encima del “fondo” esperado. Si v_1 es próxima a 1, entonces

χ^2 es grande y hay relación entre filas y columnas de **A**, con el primer autovector apuntando en la dirección de máxima variabilidad en χ^2. Si v_2, v_3,…,v_k son también grandes, entonces necesitaremos k+1 autovectores para describir los patrones en la matriz χ^2. La función $\chi^2=\Sigma_j v_i^2/\Sigma_i a_{ij}$ nos indicará cuantos autovectores necesitaremos para explicar la mayor parte de χ^2 de la red. Mientras que el ACP nos indica cuánta de la varianza explica cada dimensión, los autovalores de la Normal nos indican cuánto de la χ^2 explica cada dimensión.

El cuadrado de la matriz Normal $\mathbf{N}^2$ es similar a la matriz χ^2 , tratando **A** como tabla de contingencia [310].

7.5. ALGUNAS APLICACIONES ESPECÍFICAS DE LA TEORÍA DE GRAFOS

Se han propuesto multitud de aplicaciones de la teoría de grafos a la resolución de problemas. En [356] pueden encontrarse relacionadas y brevemente comentadas algunas de esas aplicaciones. A continuación expondremos brevemente aquellas que destacan, en nuestra opinión por su interés o temática.

7.5.1. TEORÍA DE CIRCUITOS: ANÁLISIS SISTEMÁTICO DE REDES ELECTRICAS

Aunque esta aplicación fue formulada inicialmente por Kirchhoff en [333] veremos la exposición realizada en [306]. Una red eléctrica se define por un grafo simple, conectado y no dirigido G(V,E), junto con una función de conductancia $c:E\rightarrow\Re$. La corriente eléctrica externa entra y sale de esta red, lo que se especifica por una función de suministro $b:V\rightarrow\Re$. Valores positivos de b representan corriente entrante, valores negativos representan corriente que sale de la red, y las cantidades de corriente entrante y saliente deben ser iguales $\Sigma_{v\in V}b(v)=0$. Dada la conveniencia de hablar de dirección de la corriente en un grafo no dirigido, cada arista $e \in E$ es arbitrariamente orientada para obtener una arista orientada $\underline{e}$ que conlleva un conjunto de aristas orientadas $\underline{E}$.

Una función $x : \underline{E} \rightarrow\Re$ se denomina una corriente (eléctrica) en N=(G(V,E),c) si

$$\sum_{(v,w)\in\vec{E}} x(v,w) - \sum_{(w,v)\in\vec{E}} x(w,v) = b(v) \quad \text{para todo } v \in V$$

$$\sum_{e\in C} x(\vec{e}) = 0 \quad \text{para cada ciclo } C \subseteq E$$

La primera ecuación es la Ley de la corriente de Kirchhoff y la segunda la Ley del potencial de Kirchhoff. Valores negativos de x se interpretan como corriente fluyendo contra la dirección de la arista orientada.

Alternativamente a la corriente x, un flujo eléctrico puede ser representado por potenciales. Una función p : V→ℜ es un potencial (eléctrico) si p(v)-p(w)=x(v,w)/c(v,w) para todo (v,w) ∈ E. Como una red eléctrica N=(G,c) tiene una única corriente x para cada fuente b, se tiene que el potencial p es único salvo un factor aditivo.

Así, se define la matriz Laplaciana **L**=**L**(N) de una red eléctrica N como

$$L_{vw} = \begin{cases} \sum_{e \ni v} c(e) & \text{si } v = w \\ -c(e) & \text{si } e = \{v,w\} \\ 0 & \text{en otro caso} \end{cases}$$

Entonces, un potencial **p** para una red eléctrica N=(G,c) y una fuente **b** puede ser obtenido resolviendo el sistema lineal

$$\mathbf{Lp} = \mathbf{b}.$$

Finalmente, para el propósito de establecer centralidades basadas en corrientes eléctricas, se define una unidad fuente-s-t b_{st} como una fuente de una unidad que entra en la red en s y sale de ella en t, esto es, $b_{st}(s)=1$, $b_{st}(t)=-1$ y $b_{st}(v)=0$ $\forall v \in V\backslash\{s,t\}$.

7.5.2. RESILIENCIA DE UNA RED

La resiliencia de una red, definida como la tolerancia de su conectividad a la retirada de nodos o enlaces de la red, ha sido estudiada a través de la remoción progresiva de nodos mediante dos métodos diferentes [357]:

1) aleatoriamente, lo que modela una vulnerabilidad incidental;
2) basándonos en medidas de centralidad, lo que modela un ataque más estratégico.

Estos análisis de redes pueden ser utilizados alternativamente para minar o para fortificar estructuras de redes existentes que sean necesarias para las fuerzas del orden y también aquellas relativas a la seguridad nacional, lo que en España se denominan “Infraestructuras Críticas” [358], [359].

La Eficiencia de una red se define como:

$$E = \frac{\sum_{i=1}^{n}\sum_{j>i}^{n} 1/\Delta_{ij}}{n(n-1)}$$

Donde E es la "eficiencia" de una red de comunicación;

Δ_{ij} es la distancia entre los vértices i y j

n es número de nodos de la red.

Menores distancias implican comunicaciones más efectivas y redes más eficientes. Cuando una red incluye más de una componente tal que i y j no están conectados, d_{ij} no está definido. El recíproco de la distancia entre nodos es cero para vértices en componentes separadas y se aproxima a uno cuando la distancia entre dos nodos disminuye. El denominador crea una métrica que varía entre 0 y 1, donde 1 representa conectividad total, todos los nodos están conectados directamente a otros nodos, y cero indica la disolución de la red en componentes separadas.

La vulnerabilidad E_A es una medida de la disminución media en la eficiencia de una red tras un ataque y tiene en cuenta la historia entera del ataque y la rapidez de la caída.

$$E_A = \frac{\sum_{i=1}^{K} E_i}{K}$$

Siendo $E_1, E_2, \ldots, E_K$ la eficiencia de la red después de la eliminación de 1,2, hasta K nodos.

Será interesante conocer cómo los autovalores de un grafo son afectados por cambios en la estructura del grafo. La teoría de perturbaciones de grafos está relacionada principalmente con cambios de autovalores como resultado de modificaciones en el grafo tales como adiciones o eliminaciones de arcos o vértices. Para varias de estas modificaciones, los autovalores del grafo perturbado están determinados a partir del grafo primario [322].Con aplicaciones en el campo de la resiliencia se formula el Problema del jugador clave (*Key Player Problem*) [360], [312] con dos enunciados diferentes:

1. **KPP-1**: consistente en descubrir un conjunto de k nodos de manera que su remoción desconecte máximamente la red. Estos individuos deberían ser el

objetivo para prevenir, por ejemplo, que una infección se torne en una epidemia.

2. **KPP-2**: consistente en descubrir un conjunto de k nodos que estén máximamente conectados al resto de la red. Estos individuos deberían ser el objetivo, por ejemplo, si se desea difundir una información en una red social en el menor tiempo posible.

7.5.3. EL PROBLEMA DEL MIN-CUT

El problema del Min-Cut se formula en cómo separar un grafo en dos conjuntos de nodos aproximadamente iguales con el menor número de aristas entre ellos. Se han propuesto múltiples métodos para buscar buenos puntos por donde separar un grafo, y el clustering espectral es uno de ellos muy satisfactorio. Este método utiliza los primeros vectores singulares de la matriz de adyacencia **A** o de su Laplaciana **L** para partir el grafo [312]. En [323] puede consultarse la evolución histórica del algoritmo.

Como hemos visto anteriormente, Fiedler sugirió que el autovector $\mathbf{x_2}$ denominado valoración característica, asociado con el segundo menor autovalor λ_2, o conectividad algebraica, puede utilizarse para resolver el problema *min-cut* [310] y se realiza basándose en los signos de las componentes de $\mathbf{x_2}$. Otros investigadores han utilizado $\mathbf{x_3}$, $\mathbf{x_4}$ y otros autovectores superiores para producir particiones multi-vía (*multi-way partitions*).

La valoración característica, o autovector correspondiente al autovalor λ_2 provee una buena alternativa heurística para particionar los vértices de un grafo en dos partes con escasa interferencia (pocas aristas entre las dos partes). El procedimiento sería ordenar los vértices de G en orden creciente de sus coordenadas en $\mathbf{x_2}$, si $u \leq v$ entonces $x^{(2)}_u \leq x^{(2)}_v$. Sea $A_u = \{v \in V(G) \mid v \geq u\}$ y sea $B_u = V(G) \backslash A_u$ para $u \in V(G)$. Estas particiones ofrecen resultados satisfactorios [336].

También podemos añadir restricciones adicionales al problema de *min-cut*, imponiendo que el número de aristas en cada parte sea aproximadamente igual, ponderando los nodos por su grado total. Esta partición se basa en $\mathbf{n_1}$ de **N**, dado que $\mathbf{n_1}$ apunta en la dirección de máxima variabilidad de χ^2. Particiones sobre $\mathbf{n_2}$, $\mathbf{n_3}$,… también producen conjuntos de nodos con un gran número de aristas en común (tan grandes como ν_2, ν_3,.. contribuyan a χ^2). Particiones basadas en autovalores positivos producirán bloques en la diagonal de **A** de aristas asociados con cada conjunto de

nodos, mientras que aquellos basados en autovalores negativos generarán casi bloques bipartitos fuera de la diagonal, que además ocurrirán en pares si la red es simétrica) [310].

En general el k-ésimo autovector divide la red en no más de k+1 componentes desconectadas [310]. Otros autores [331] proponen un método de particionado que es similar en concepto al ACP tradicional.

7.5.4. PROPAGACIÓN DE VIRUS E INMUNIZACIÓN

Es posible utilizar también la teoría de grafos para modelar la propagación de virus (biológicos o informáticos) y así estudiar posibles estrategias de inmunización.

Para varios modelos de propagación de virus y operen sobre un grafo no dirigido con matriz de adyacencia **A**, el umbral epidémico depende solo del mayor autovalor λ_n de **A** y una constante C que es determinada por el propio modelo de propagación [361], [312].

Inmunizar nodos equivale a eliminarlos del grafo de contacto, lo que hace a su vez disminuir λ_n. El criterio óptimo de inmunización consiste en minimizar el autovalor de la matriz tras la inmunización.

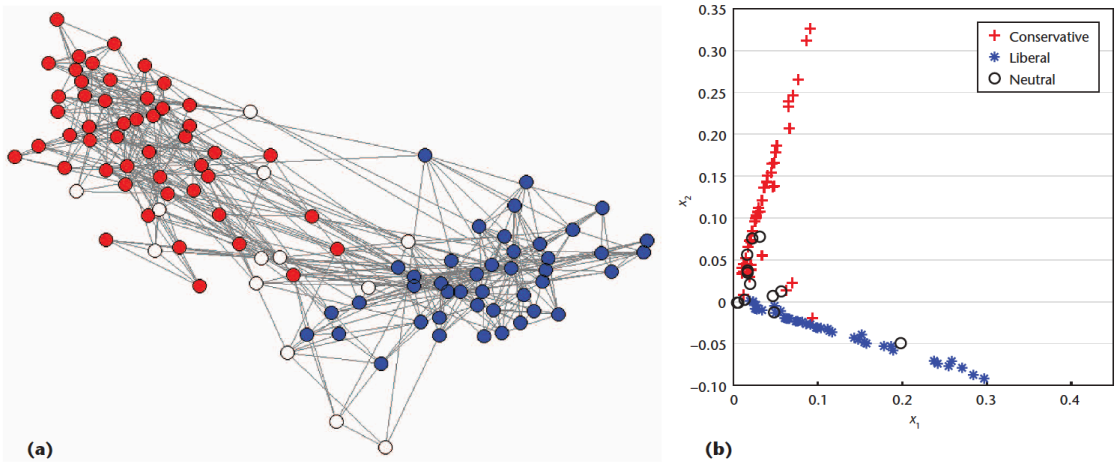
7.5.5. DETECCIÓN DE CLUSTERS

Dado un grafo dirigido G(V,E) deseamos dividir el conjunto de vértices en “clusters” de vértices similares. Dado que toda nuestra información está en el grafo, definimos que dos vértices son “similares” si comparten gran parte de vértices. En general asumiremos, en principio, que toda la información relevante está capturada por la matriz de adyacencia **A** del grafo G [362].

Algunos autores [362] afirman que la introducción de la DVS para agrupar clusters fue propuesta por Kleinberg en 1999 [363], que argumentaba que dada la matriz de adyacencia de un subgrafo de un grafo de la web, donde los nodos corresponden con páginas web que son retornadas por una herramienta de búsqueda de una consulta específica, es deseable buscar los mayores vectores singulares de la matriz. Cada vector corresponde con diferentes significados de la búsqueda.

Para un grafo con k comunidades (colecciones de nodos relacionadas con mayor preponderancia), la coordenada del nodo u en el espacio k-dimensional $\alpha_u=(x_{1u}, x_{2u},$

…,x_{nu}) ∈ $\Re^{1xk}$ indica la probabilidad de que el nodo u esté conectado a esa k comunidad [153]. Los puntos de los nodos de una comunidad forman una línea que pasa a través del origen en el espacio k-dimensional. Nodos en k comunidades forman k líneas cuasi-ortogonales en el espacio espectral.



Ejemplo de proyección de red en el espacio espectral 2D. [153]

En 2010 Fay *et al* [364] publican un artículo [364] en el que utilizan el espectro del grafo asociado a una red de comunicaciones para la comparación y comprensión de la topología de Internet. Afirman que el espectro de los grafos han sido utilizados para gran variedad de propósitos, además de la caracterización de topologías de Internet. Se han utilizado los autovectores de los k mayores autovalores de las matrices de adyacencia para la comparación coeficientes de clustering de grafos. También se han utilizados los autovalores de la matriz Laplaciana normalizada para representar la topología de la red o, como hemos visto, su robustez. Y, relacionado con esta aplicación concreta de detección de clusters, se afirma que el primer menor autovalor no nulo de la matriz Laplaciana normalizada y su autovector asociado se pueden utilizar para la detección de los principales clusters en los datos vinculados a redes.

No obstante, a nuestro modo de ver, el artículo más interesante para la aplicación de detección de clusters en grafos es el propuesto por Sarkar y Dong en 2011 [330], cuya argumentación expondremos a continuación detalladamente por la relevancia que, como veremos, tiene para nuestra propuesta por su contenido.

El problema de identificar comunidades o “clusters” en investigación de redes complejas está ampliamente estudiado y sigue abierto, según los autores [330]. Sarkar y Dong presentan un método relacionado con la DVS de la matriz del grafo, agrupando posteriormente la estructura en baja dimensión resultante para detectar comunidades. Utiliza dos representaciones matriciales diferentes, la Laplaciana sin signo para grafos

unipartitos y la matriz rectangular de adyacencia adaptada para grafos bipartitos. Las mayores contribuciones del artículo son:

- Identifica correctamente si existe o no en la red una estructura de comunidades.
- Opera tanto en grafos unipartitos como bipartitos y ponderados o no ponderados.
- Detecta comunidades disjuntas y/o solapadas simultáneamente, incluyendo casos en los que un grafo pueda contener una mezcla ya que los vértices no están restringidos a pertenecer a una u otra clase.
- Muestra organizaciones jerárquicas y modulares.
- No impone el número o tamaño de las comunidades, solapamientos, número de niveles jerárquicos, como parámetros externos.
- El algoritmo es computacionalmente tan eficiente como los algoritmos existente para calcular la DVS.

En general todos los métodos para realizar particiones espectrales de grafos utilizan la información contenida en los autovectores y autovalores de una representación matricial apropiada del grafo.

A partir de la matriz de adyacencia del grafo definida como:

$$A_{ij} = \begin{cases} 1 & \text{si existe una arista entre los nodos i y j} \\ 0 & \text{en otro caso} \end{cases}$$

y que para un grafo no dirigido, **A** es simétrica, y de la matriz de grado **D** definida como

$$D_{ij} = \begin{cases} d_i & \text{grado del nodo i cuando } i = j \\ 0 & \text{cuando } i \neq j \end{cases}$$

Se tiene que la matriz Laplaciana puede obtenerse como

$$\mathbf{L = D - A}$$

con

$$L_{ij} = \begin{cases} d_i & \text{cuando } i = j \\ -1 & \text{cuando } i \neq j \text{ y } i \text{ es adyacente a } j \\ 0 & \text{en otro caso} \end{cases}$$

Las aproximaciones espectrales operan sobre **L** (o variantes de **L**) para partir el grafo de manera recursiva, buscando cada vez una bisección óptima. El grafo es primero partido en dos módulos con respecto a una función de optimización, continuando posteriormente un proceso recursivo. El algoritmo espectral de particionado tiene como objetivo minimizar el "tamaño de corte" R, definido como el número de aristas que unen los dos grupos de vértices en los que se divide el grafo. Para ello se puede concluir que se debe operar con el autovector correspondiente al segundo menor autovalor de **L** que como hemos visto se denomina conectividad algebraica de [338].

En el supuesto de encontrarnos ante un grafo con aristas ponderadas, tenemos que

$$A_{ij} = w_{ij} \geq 0$$

La matriz de adyacencia es una medida local y pura de los vértices vecinos que es, afirman los autores, insuficiente para proveer información sobre la estructura global del grafo. Dado que la mayoría de las matrices de adyacencia suelen estar poco pobladas, el calculo de productos internos entre filas y columnas de **A** no suelen revelar información trascendente.

Consideraremos en lugar de la matriz Laplaciana, la matriz Laplaciana sin signo

$$|\mathbf{L}| = \mathbf{D} + \mathbf{A}$$

con

$$L_{ij} = \begin{cases} w_i & \text{cuando } i = j \\ 1 & \text{o } w_{ij} \text{ cuando } i \neq j \text{ y } i \text{ es adyacente a } j \\ 0 & \text{en otro caso} \end{cases} \qquad \text{con } w_i = \sum_j w_{ij}$$

La principal razón para utilizar la matriz Laplaciana, en cualquiera de sus versiones con o sin signo, es la manera en la que sus elementos codifican la conectividad: nos permite formular una medida de la fuerza relativa de la asociación entre nodos. Como veremos posteriormente, la modularidad se expresa espacialmente en el espacio vectorial, donde la posición de cada nodo es representativa de la conectividad con todos los otros nodos. Este medida espacial sobre un vector no solo depende de la

fuerza individual de las conexiones entre nodos, también depende de cuántos nodos están conectados a un nodo en particular. La Laplaciana sin signo nos permite capturar esta fuerza en las entradas de su diagonal, mientras que la matriz de adyacencia habitual no.

Los autores proponen para el caso de redes bipartitas la utilización de una matriz de incidencia vértice-vértice (sería de adyacencia en el caso general), así, una matriz **A'** m×n representará los m nodos en V_1 y los n nodos en V_2, donde para i=1,…,m y j=1,…,n

$$a'_{ij} = \begin{cases} 1 & \text{o } w_{ij} \text{ si existe una arista entre i y j} \\ 0 & \text{en otro caso} \end{cases}$$

Se puede representar **A'** en una forma cuadrada similar a |**L**| con la matriz (m+n)×(m+n) **A**= [**0 A'**;**A'**^{T} **0**]. La correspondiente matriz de grados será **D**=[$\mathbf{D}_1$ **0**;**0** $\mathbf{D}_2$] donde $\mathbf{D}_1$ es la matriz de grados para los nodos en V_1 y $\mathbf{D}_2$ es la matriz de grados para los nodos en V_2. Así, la matriz Laplaciana sin signo para el caso bipartito será

$$|\mathbf{L}| = \begin{bmatrix} \mathbf{D}_1 & \mathbf{A}' \\ \mathbf{A}'^T & \mathbf{D}_2 \end{bmatrix}$$

Esta representación causa una repetición de los datos y es por lo tanto computacionalmente ineficiente. El método que se propone opera directamente sobre la matriz **A'** para el caso de grafos bipartitos: la matriz de adyacencia rectangular para un grafo bipartito presenta en su diagonal la conectividad entre elementos, capturando de alguna manera la fuerza relativa asociación entre nodos directamente sin necesidad de utilizar el grado u otra información.

Así pues, los autores proponen operar con |**L**| para grafos unipartitos y con **A'** para grafos bipartitos, siendo **A'** una matriz de adyacencia vértice-vértice anteriormente definida.

Se muestra la aplicación del método que proponen los autores para la matriz **A'**, pero podría aplicarse de la misma manera a la matriz |**L**|. En **A'**, el i-ésimo vector fila en $\Re^n$ muestra los vecinos del vértice i de tipo 1. Similarmente, el j-ésimo vector fila en $\Re^m$ muestra los vecinos del vértice j de tipo 2. Una DVS de la matriz causa una transformación lineal que diagonaliza la matriz [158]

$$\mathbf{A'}=\mathbf{U}\Lambda^{0.5}\mathbf{V}^T$$

Siendo **U** es una base m×m ortonormal en $\Re^m$, **V** es una base n×n ortonormal en $\Re^n$, la matriz diagonal $\mathbf{\Lambda}^{0.5}$ contiene los valores singulares que representan la información de escalado de los vectores cuando van de $\Re^n$ a $\Re^m$ y están ordenados decrecientemente. El número de valores singulares es igual al rango r de la matriz de adyacencia **A**. Si ignoramos el espacio nulo, **U** es una matriz m×r, $\mathbf{\Lambda}^{0.5}$ es una matriz r×r, y $\mathbf{V}^T$ es una matriz r×n. Así **U** y **V** representan conjuntos de bases de autovectores para $\Re^m$ a $\Re^n$ respectivamente, donde la información original de correlación entre los vértices originales está diagonalizada y expresada en términos de vectores independientes e incorrelados. La información de acoplamiento local vértice-vértice en **A'** está descompuesta: los vectores pueden ser ahora expresados como combinaciones lineales de las bases ortogonales y los valores principales.

Escogiendo las bases ortonormales $\mathbf{V}=(\mathbf{v}_1,\ldots,\mathbf{v}_r)$ para el espacio de las filas y $\mathbf{U}=(\mathbf{u}_1,\ldots,\mathbf{u}_r)$ para el espacio de las columnas, de forma que $\mathbf{A'v}_i$ esté en la dirección de $\mathbf{u}_i$, siendo s_i el factor de escala, obtendríamos

$$\mathbf{A'v}_i = s_i\mathbf{u}_i$$

O escrito de otra manera

$$\mathbf{A'V} = \mathbf{U\Lambda}^{0.5}$$

La columna de **A'** muestra los vecinos de un vértice de tipo 1. El producto escalar entre una columna de **A'** y el i-ésimo autovector de **V** es una medida de cuanto un vector de un vértice apunta en la misma dirección que el autovector $\mathbf{v}_i$. Todos los vectores con vecinos comunes apuntan en la misma dirección del espacio.

Similarmente

$$\mathbf{B}^T\mathbf{U} = \mathbf{V\Lambda}^{0.5}$$

Así, los productos $\mathbf{U\Lambda}^{0.5}$ y $\mathbf{V\Lambda}^{0.5}$ proveen un nuevo modo abstracto de describir cada uno de los m vértices de Tipo 1 y n vértices de Tipo 2, respectivamente, como una combinación lineal de las correspondientes bases ortogonales y valores singulares. Esto posibilita la representación de un vértice como un vector en el espacio, donde su posición es representativa de su relación con otros vértices, en términos de pertenencia a un autovector. En realidad, los autores están realizando un HJ-Biplot [160], como podemos comprobar si comparamos las expresiones anteriores con las de

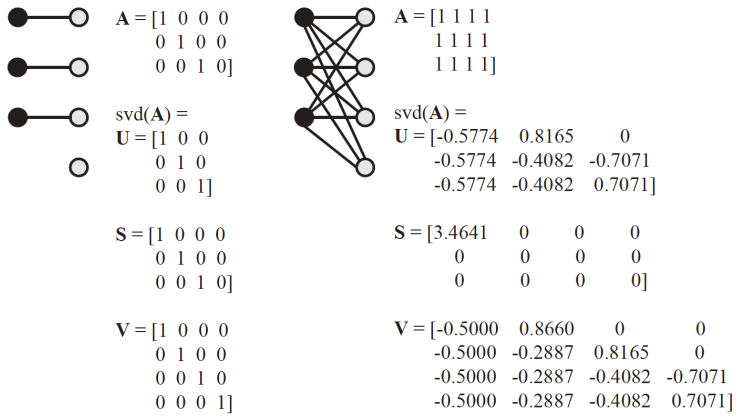
los marcadores columna **H** y marcadores fila **J** de la definición del HJ-Biplot, y aunque utiliza algunas de sus propiedades, no lo aplica en toda su extensión, ni extrae de la representación gráfica toda la información que los marcadores **H** y **J** aportan.

Grupos de vértices altamente conectados, por ejemplo un cluster, apuntarán en la misma dirección del espacio, representando similares relaciones con los autovectores. Productos interiores entre vectores ofrecerán una medida de cuánto cercanos están acoplados en el gráfico, lo que es una medida de la pertenencia a una "comunidad" o "cluster". Cuánto más ortogonales sean los vectores, menor posibilidad de que pertenezcan a la misma comunidad, cuánto más paralelos será más probable que pertenezcan a la misma comunidad. La componente de cada autovector no contribuye igual a la definición de pertenencia, aquellas que corresponde a autovalores mayores tienen mayor contribución.

En el caso de grafos unipartitos, la matriz |**L**| es simétrica, y los productos $\mathbf{\Lambda}^{0.5}\mathbf{U}$ y $\mathbf{\Lambda}^{0.5}\mathbf{V}$ colapsan en uno. En el caso de grafos unimodales la matriz es no-simétrica. Los autores dejan para futuros estudios la extensión del método a grafos dirigidos.

Para obtener la clasificación de nodos los autores analizan la información contenida en la dimensión, concluyendo que no todas las dimensiones son importantes o necesarias para la clasificación. Consideran la relación entre el número de vértices, el número de comunidades y el rango de la matriz **A'**. Si dos vértices comparten el mismo conjunto de vecino, entonces existirá dependencia entre filas y columnas de **A'**. En el supuesto de tal redundancia, será suficiente un número menor de dimensiones para capturar que esos dos vértices pertenecen a la misma comunidad. Esto tiene como consecuencia que el rango r será mucho menor que n y entonces habrá como mucho r agrupaciones.

Para ilustrar el método los autores proponen inicialmente dos ejemplos extremos de grafos bipartitos.



Ejemplos demostrativos propuestos por Sarkar *et al*. [330]

Consideran en primer lugar un grafo en el que cada vértice de un tipo este unido a uno y solo uno del otro tipo y puedan existir nodos sin vértices. Es evidente que cada pareja será una comunidad, y el número de comunidades será igual al número de parejas existentes. La DVS mostrará que son necesarios todos los vectores y valores singulares para detectar el número de comunidades: todas las columnas de **A'** son independientes y no es posible reducción de la dimensionalidad sin pérdida de información. Todos los m o n valores singulares son igualmente importantes. Hay r comunidades distintas, donde r es el rango de la matriz, con una pareja de vértices perteneciente a cada comunidad.

En segundo lugar consideran el caso extremo, un grafo con todos los vértices conectados entre si. En este caso el número de comunidades es igual a 1. La DVS muestra que solo un vector y valor singular es necesario para detectar el número exacto de comunidades: todas las columnas de la matriz original son dependientes entre si, ya que son iguales. Debido a esta redundancia en los datos, solo una dimensión es suficiente para detectar la solución exacta. El rango de la matriz es 1, y todos los vértices están representados por una única dimensión.

Así, el número de comunidades será siempre menor o igual al rango de la matriz. Pero los autores llegan más allá, demostrando que no solo será menor, sino mucho menor que el rango de la matriz, basándose precisamente en la observación de que la DVS como descomposición de la matriz original en una serie de matrices de rango 1 que puede ser aproximada por solo los primeros términos para obtener una aproximación optima de la matriz original en términos de mínimos cuadrados.

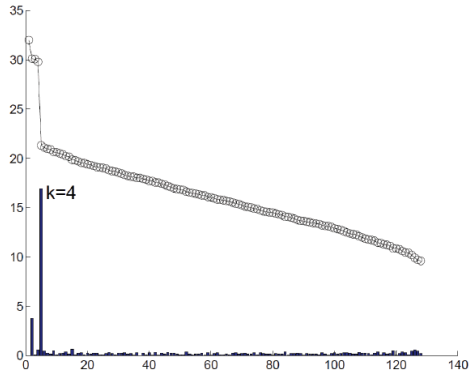
Continúan los autores analizando la relación de su propuesta con el ACP, ya que la matrices **U** y **V** contienen los autovectores o componentes principales de las matrices $\mathbf{A'A'}^T$ y $\mathbf{A'}^T\mathbf{A'}$ respectivamente. Esto es, calculando la DVS de **A'** y $\mathbf{A'}^T$ tendríamos

$$\mathbf{A'A'}^T=\mathbf{U}\Lambda^{0.5}\Lambda^{0.5T}\mathbf{U}^T$$

$$\mathbf{A'}^T\mathbf{A'}=\mathbf{V}\Lambda^{0.5T}\Lambda^{0.5}\mathbf{V}^T$$

Dicho de otro modo, en la aproximación de rango reducido los productos internos entre vértices que compartan vecinos se incrementarán, y viceversa, ya que los autovectores menos importantes serán descartados en la descomposición. Esto es, los vértices que comparten muchos vecinos estarán más orientados en la misma dirección del espacio.

La identificación de comunidades es trivial en este punto. Simplemente habrá que representar en el espacio de dimensión reducida los productos $\mathbf{U}\Lambda^{0.5}$ y $\mathbf{V}\Lambda^{0.5}$ como vectores. Un coseno más elevado entre dos vectores, mayor probabilidad de que dichos vértices pertenezcan a la misma comunidad. La identificación de agrupaciones puede llevarse a cabo mediante un algoritmo de clustering K-medias, por ejemplo, utilizando el coseno entre vectores como distancia. Para obtener el número óptimo de agrupaciones proponen un *scree plot*. (por cierto en el *scree plot* plantean que se superponga el salto entre valores como columnas)



Screeplot para detección del número de comunidades. [330]

A continuación los autores verifican su propuesta con cinco redes disponibles en la literatura.

Finalmente exponen posibles continuaciones de su trabajo, a redes dirigidas y ponderadas. Exponen una limitación a su propuesta en el caso de solapamientos entre

agrupaciones. Proponen también el estudio sobre la matriz Laplaciana u otras formulaciones matriciales relacionadas con los grafos estudiados.

7.5.6. REDUCCIÓN DEL NÚMERO DE MEDIDAS PARA ESTIMAR VARIABLES EN LA RED COMPLETA

En el año 2005 Chua *et al* [348] presentan en la conferencia IEEE INFOCOM un método para la monitorización eficiente de las propiedades extremo a extremo de redes. La pregunta que se proponen responder es la siguiente: ¿es posible obtener una buena estimación global de las propiedades extremo a extremo de una red a partir de la realización de mediciones en un número reducido de rutas? Si, por ejemplo, deseamos conocer el retardo entre todos los nodos de una red, será preciso realizar a priori un número de medidas igual al cuadrado del número de nodos. Chua *et al* ya exponen que diversos estudios previos [365]–[367] muestran que es posible reducir este número de medidas, y aún obtener el conjunto de medidas deseadas de manera exacta. En este trabajo Chua *et al* proponen un método para obtener el conjunto de medidas perseguido de manera aproximada, pero reduciendo fuertemente el número de medidas necesarias para ello.

Además las preguntas a las que deben responderse serán:

1. ¿En cuántas rutas será necesario monitorizar un métrica **y** para obtener una estimación aceptable de dicha métrica en las restantes rutas?
2. ¿Cuáles serán las rutas en las que será preciso monitorizar la métrica **y**?

Los autores desarrollan un modelo estadístico para la predicción de los valores de la métrica **y** en los enlaces no monitorizados, a partir de las mediciones en determinados enlaces de la red. Afirman que, en realidad, se trata en una versión del problema de tomografía de la red y, en cierto sentido, del planteamiento inverso del bien conocido problema de estimación de la matriz de tráfico de una red [350].

Afirman los autores que para el objetivo propuesto no serían necesarias mediciones de una métrica aditiva **y** en más de N_e rutas de la red bajo estudio, siendo N_e el rango máximo de la matriz de enrutamiento **B** de dicha red. Esto es, para disponer de manera exacta de la información de una métrica **y** que sea aditiva sobre cualquier ruta de la red, será suficiente el conocimiento de dicha métrica sobre las rutas independientes de la red, ya que así se puede reconstruir la métrica sobre cualquier otra ruta de la red. Este planteamiento, como ya se ha indicado, se apoya en los

estudios de Shavitt *et al* presentados en 2001 [365] y posteriormente publicados en 2004 [367].

Los autores demuestran en primer lugar que es posible la reconstrucción aproximada de la matriz correspondiente a la métrica aditiva considerada para todas las rutas posibles, a partir de un conjunto mediciones sobre algunas de la rutas. En segundo lugar determinan cuáles deberán ser las rutas sobre las que efectuar la medición para obtener una mejor estimación. Especialmente interesante, como veremos, es el procedimiento aplicación para este segundo paso. Afirman los autores que el problema puede formularse como sigue: Sea una matriz **M** y un número entero *k*, estos métodos buscan un subconjunto de *k* columnas (o filas) linealmente independientes de **M** que aproximen de manera precisa sus *k* primeras dimensiones y formen una matriz bien condicionada. Utilizan para ello la Descomposición en Valores Singulares $\mathbf{M} = \mathbf{U}\Lambda^{0.5}\mathbf{V}^{T}$.

Unos meses después, en ese año 2005, los mismos autores Chua *et al* presentan en otro congreso un nuevo artículo [349] con planteamientos muy similares al precedente [348]. Introducen, no obstante, alguna novedad. Siendo **B** la matriz de enrutamiento, analizan, por ejemplo, los autovalores de la matriz $\mathbf{B}^{T}\mathbf{B}$, que es simétrica y cuyos autovalores son todos positivos y su rango es igual a número de enlaces considerados independientes. Así pues, no será necesario obtener más que ese número de medidas de la métrica aditiva para recuperar la métrica sobre todas las rutas posibles. Además, una implicación de que las matrices de enrutamiento presenten un rango efectivo menor que el rango máximo esperado, es que, como ya expusimos, un número reducido de autovalores de la matriz $\mathbf{B}^{T}\mathbf{B}$ serán mucho mayores que el resto [349]. Esta situación permitirá recuperar de manera aproximada la matriz de estimaciones de la métrica sobre todas las rutas a partir de un número reducido de mediciones.

Demuestran los autores que el rango reducido de la matriz se mantiene de manera suficientemente estable bajo la situación de pérdida o caída de enlaces, aunque en algunos casos sea necesaria la consideración de otras medidas diferentes para poder recalcular todas las estimaciones.

En este artículo los autores proponen como posibles aplicaciones de su propuesta la monitorización de los valores medios de la red completa, la detección de anomalías y la comparación entre subredes.

A finales de 2006, de nuevo Chua *et al* publican un artículo sobre lo que denominan *Network Kriging* [291]. Parten los autores del clásico problema de predicción de las características de una población a partir de una muestra. Concretan un tipo de problema más específico, el formulado en geociencia bajo el nombre de *kriging* [368], [290] en el que, por ejemplo, se toman medidas en catas distribuidas espacialmente para predecir la concentración de petróleo en el subsuelo. Asimilan esta situación que con la que se proponen resolver: monitorizar de manera precisa (pero no exacta) determinadas propiedades en toda una red a partir de mediciones en un conjunto reducido de rutas, y de ahí el nombre de *network kriging*, planteamiento similar al expuesto anteriormente.

El contenido del artículo publicado [291] fue presentado parcialmente en los dos previamente ya comentados [348], [349] por lo que coinciden algunas de sus conclusiones:

1. Las matrices de enrutamiento presentan un rango efectivo menor que el rango esperado por sus dimensiones.
2. Es posible la predicción de las propiedades extremo a extremo entre todos los nodos de la red a partir de la medición en un número reducido de ellas.
3. Las características anteriores pueden ser utilizadas para la detección de anomalías o la caracterización global de una red, por ejemplo.

Se ha propuesto en la literatura otras propuesta similares para resolver este problema.

En el año 2009 Kolaczyk, coautor de los artículos antes reseñados [291], [348], [349] publica en su libro *Statistical Analysis of Network Data* [350] otra aproximación ligeramente distinta a la anterior para resolver el mismo problema.

Kunegis y Lommantzsch proponen en 2009 [337] la utilización de técnicas de reducción de rango de la matriz de adyacencia y de la matriz Laplaciana para implementar una predicción del peso de las aristas de un grafo.

Buza y Galambos proponen en 2013 [369] la utilización de la factorización de matrices en la predicción de enlaces en redes bipartitas, si bien proponen una descomposición del tipo $\mathbf{X} \approx \mathbf{UV}$ en lugar de la DVS, esto es, $\mathbf{X} \approx \mathbf{U}\Lambda^{0.5}\mathbf{V}$.

7.6. CONCEPTO DE CENTRALIDAD EN LOS GRAFOS

Los índices de centralidad sirven para cuantificar una cuestión intuitiva, que en la mayoría de las redes algunos vértices o aristas son más “centrales” o “importantes” que otros [332], [356], [370]. Existen índices de centralidad para vértices y aristas y en general no todos los índices sirven para cualquier aplicación, por lo que han aparecido decenas de ellos a lo largo del tiempo. El primer obstáculo que aparece es que, en general, el término “centralidad” no está claramente definido. Por ejemplo, en algunos problemas un vértice puede entenderse como más “central” en la red cuantas más aristas entrantes presente. En otras situaciones, sin embargo, puede establecerse que un nodo es más “central” cuanto más necesario sea para establecer caminos entre otros nodos. Incluso en otros casos puede determinarse que un nodo es más central, cuanto más centrales sean sus nodos adyacentes [324].

Análogamente a la centralidad de los nodos, puede hablarse con la misma propiedad de la centralidad de las aristas. Valga como ejemplo de este último escenario, básico en nuestro contexto, el modelado de una red de comunicaciones a través de un grafo, en el que los enlaces pueden presentar de manera obvia una determinada “centralidad” en la arquitectura de la red.

7.6.1. Definición y propiedades mínimas de un índice de centralidad

Como se exponía anteriormente no existe una definición comúnmente aceptada para los índices de centralidad y prácticamente cada autor introduce su índice de centralidad, sin disponer de una definición estricta de centralidad en general. No obstante es posible establecer unas propiedades mínimas para cualquier índice de centralidad. Intuitivamente un índice de centralidad implica un orden de importancia en los vértices o aristas de un grafo, asignándoles a los mismos un número real. Signifique ese concepto de “importancia” lo que corresponda en cada situación o problema considerado. Esas propiedades mínimas consisten en que el índice de centralidad sea un “índice estructural”, lo que a la inversa no siempre tiene porque cumplirse (no todo índice estructural es un índice de centralidad).

En primer lugar establezcamos que dos grafos $G_1(V_1,E_1)$ y $G_2(V_2,E_2)$ son *isomórficos* ($G_1 \cong G_2$) si existe un mapeado uno-a-uno $\phi: V_1 \rightarrow V_2$ tal que (u,v) es una arista en E_1 si y solo si (ϕ(u), ϕ(v)) es una arista en E_2.

Sea G(V,E) un multígrafo, dirigido o no, con pesos, y sea X el conjunto de vértices (o aristas) de G. Una función real s se dice que es un "índice estructural" sí y solo sí se satisface la siguiente condición:

$$\forall x \in X : G \cong H \Rightarrow s_G(x) = s_H(\phi(x))$$ donde $s_G(x)$ representa el valor de s(x) en G.

Esto es, un índice de centralidad debe representar al menos una semi-ordenación en el conjunto de vértices/aristas, dicho de otro modo, x será al menos tan central como y, respecto a un mismo índice de centralidad c, si $c(x) \geq c(y)$.

Además el índice de centralidad debe ser invariante ante isomorfismos, y en particular el índice de centralidad debe ser también invariante frente a automorfismos. Existen diferentes conceptos en los que se puede basar un índice de centralidad que examinaremos brevemente a continuación.

7.6.2. Algunas propuestas de centralidades

Centralidad de Grado: el grado de centralidad de un nodo se define en este caso como el número de aristas incidentes sobre ese nodo [354]. Nodos con mayor grado se consideran más centrales. No obstante esto puntúa a un nodo solo por sus vecinos inmediatos, sin tener en cuenta aquellos que se encuentran a dos o tres saltos[312].

Centralidad de Proximidad: la proximidad de un nodo se define aquí como la inversa de la distancia total geodésica desde el nodo a todos los otros nodos de la red, siendo la distancia geodésica como el número de aristas del camino más corto entre dos vértices [354]. Es la inversa normalizada de la distancia métrica media, aquellos nodos que están a corta distancia de todos los demás nodos tienen más elevada la centralidad de proximidad [312].

Centralidad de intermediación (*betweenness*): los nodos con elevados valores de este indicador están presentes en más caminos, y son presumiblemente más importantes [312]. Se define como el número de caminos geodésicos que pasan a través de un nodo dado, ponderado inversamente por el número total de caminos equivalentes que pasan a través de los mismos dos nodos, incluyendo aquellos que no pasan a través del nodo considerado [354].

Centralidad de flujo [312]: similar a la de intermediación, excepto que en lugar de considerar las rutas más cortas, considera todas las rutas posibles.

Centralidad de Bonacich: Este indicador utiliza también aquellos vecinos más alejados, utilizando las componentes del primer autovector del grafo [312]. Por su relevancia será revisado posteriormente con mayor detalle.

7.6.3. Estructura interna de los índices de centralidad

El análisis de las diferentes posibilidades de medir la centralidad hace pensar en la clasificación de sus índices en cuatro categorías de acuerdo a su modelo de cálculo [371]. Cada uno de estos modelos está representado por un término básico, y otros tres adicionales, el término de operador, la personalización y normalización, estas dimensiones son "independientes" entre sí.

- **Término básico**: Constituye la primera dimensión a considerar y clasifica los índices de centralidad en cuatro categorías.
    - **Alcanzabilidad (*Reachability*)**: Un vértice se supone "central" si alcanza muchos otros vértices. Estas centralidades se sustentan sobre el concepto de distancia entre dos vértices.
    - **Cantidad de flujo (*Amount of flow*)**: Esta segunda categoría se basa en la cantidad de flujo $f_{st}(x)$ desde el vértice s al vértice t que fluye a través del vértice o arista x. En el fondo no solo están basadas en el proceso de flujo, también se apoyan en las rutas más cortas.
    - **Vitalidad (*Vitality*)**: La tercera categoría se sustenta sobre la medida de la vitalidad, que tienen en consideración la importancia de los vértices o aristas en un grafo.
    - **Retorno (*Feedback*)**: Esta última categoría se basa en la definición implícita de centralidad, en la que la centralidad de un determinado vértice v_i depende de los valores de la centralidad en todos los vértices $v_1,\ldots v_n$.
- **Término de operador**: Esta segunda dimensión se determina por la función de operación sobre la que se calcula la centralidad. Así, por ejemplo, se puede seleccionar que la centralidad es el máximo de la centralidad de todos los vértices, o su suma, o su distancia media, o incluso su varianza.

- **Personalización**: La tercera dimensión tiene que ver con la personalización de los índices de centralidad, esto es, la posibilidad de asignar un vector de ponderaciones a los vértices o aristas o incluso a considerar solo un conjunto de vértices sobre los que se calcula la centralidad.

- **Normalización**: La última dimensión es la posible normalización de los índices de centralidad, por ejemplo, por el máximo valor de centralidad.

Utilizando estos conceptos puede seguirse el siguiente diagrama para adaptar o construir un índice de centralidad apropiado a una aplicación concreta.

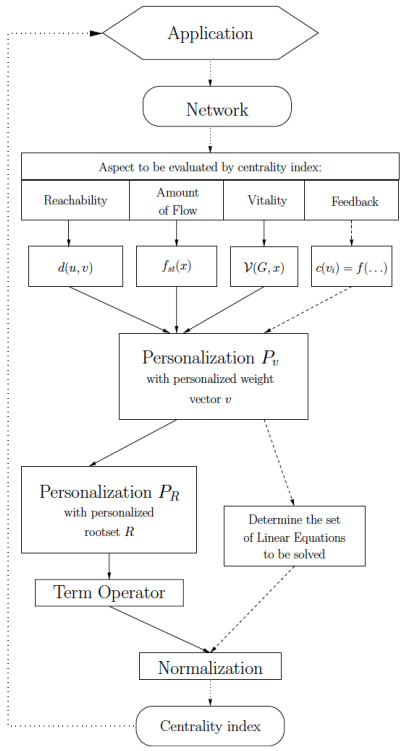
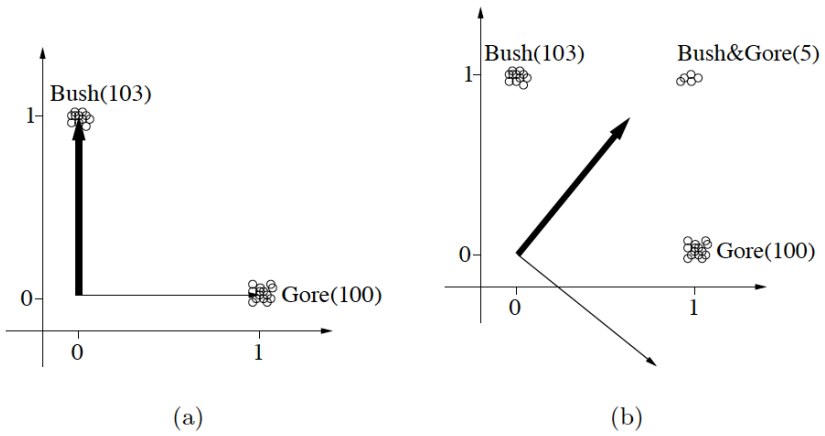


Diagrama de flujo para definir un índice de centralidad [372].

7.6.4. Estabilidad de los índices de centralidad

Ng et al [373] (citado por [372]) presentan un ejemplo de cómo un autovector puede cambiar considerablemente incluso si el grafo varía solo ligeramente. El ejemplo considera un conjunto de páginas web donde 100 de ellas están enlazadas al dominio *algore.com* y las otras 103 están enlazadas al dominio *georgewbush.com*. Se muestran en la figura (a) los primeros dos autovectores, más propiamente la proyección sobre sus componentes no nulas. En la figura (b) se muestra el cambio si se incluyen solo 5 nuevas páginas enlazadas a ambos dominios.



Estabilidad de los autovectores en el ejemplo de Ng *et al*.

7.6.5. Centralidad del autovector o de Bonacich

En 1972 Bonacich [374] introdujo una medida de centralidad basada en los autovectores de la matriz de adyacencia **A** [332]. Se asume que el grafo es no dirigido, conectado, sin lazos, simple y sin ponderaciones. Como el grafo es no dirigido y sin lazos, la matriz de adyacencia es simétrica y su diagonal esta formada por ceros. La centralidad de Bonacich de un nodo se corresponde valor absoluto de la primera componente del autovector correspondiente a ese nodo [312]. Más concretamente la Centralidad autovectorial o de Bonacich [354], [374], [356], [370] se define como el autovector principal de la matriz de adyacencia de un grafo. Si **A** es la matriz de adyacencia de un conjunto de nodos, entonces los nodos más centrales son aquellos con los valores más elevados de las componentes a_i, donde $\mathbf{a}=(a_1,\ldots,a_N)$ es el vector singular derecho de la DVS de **A** [312].

Esta centralidad también puede ser vista como una medida ponderada del grado en el que la centralidad de un nodo es proporcional a la suma de las centralidades de los nodos a los que es adyacente [354]. Esta interpretación es idéntica (y las puntuaciones son proporcionales) al primer factor resultante de la Descomposición en Valores Singulares de los datos brutos de la matriz de incidencia de dos modos, que es la aproximación tomada por Bonacich en 1991 [375], [354]. Esta formulación es también equivalente a calcular los autovectores de $\mathbf{X}\mathbf{X}^T$ y $\mathbf{X}^T\mathbf{X}$ donde **X** es de nuevo la matriz en bruto de “incidencia” (sic) de dos modos representando, por ejemplo, la pertenencia de miembros a grupos.

Pero a la matriz de incidencia vértice-arista **Q** también puede serle de aplicación dicha definición de centralidad del autovector. De facto la matriz de incidencia vértice-arista **Q** puede ser interpretada igualmente como una matriz de adyacencia **A** de dos modos

para una red bipartita, en la que un conjunto de los vértices serán los vértices del grafo inicial y el otro grupo de vértices representan las aristas del grafo inicial. Así las nuevas aristas de este grafo serán las relaciones existentes entre nodos y aristas de la red inicial. El signo negativo de algunos elementos tampoco constituye problema alguno, Bonacich en 2007 [376] plantea matrices de adyacencia con signo, para grafos dirigidos, como modo, por ejemplo, para representar un "me gusta"/"no me gusta" en las relaciones. En ese mismo artículo del año 2007 Bonacich [376] expuso también nuevas propiedades del índice de centralidad del autovector.

Recordemos que de facto, al calcular la DVS de la matriz de incidencia **Q** para obtener las centralidades del grafo, o equivalentemente de la matriz Adyacencia **A** de dos modos para la red bipartita equivalente como hemos definido antes, en realidad estaremos obteniendo, una vez más, los autovalores y autovectores de la matriz Laplaciana $\mathbf{L}=\mathbf{QQ}^T$ y de su versión de aristas $\mathbf{K}=\mathbf{Q}^T\mathbf{Q}$. Tampoco se trata éste de un planteamiento demasiado extraño ya que varios autores [377] [378] han propuesto diferentes medidas de centralidades basadas precisamente en la matriz Laplaciana **L**, aunque ciertamente no bajo esta misma formulación que proponemos.

Así pues, sabemos que la matriz Laplaciana **L** contiene en sus autovectores y autovalores múltiple información clave sobre la estructura del grafo del que parte. Dicha matriz Laplaciana **L** puede obtenerse a partir de la matriz de incidencia vértice-arista **Q** como $\mathbf{L}=\mathbf{QQ}^T$ por lo que si calculamos la DVS estaremos obteniendo los autovalores y autovectores de $\mathbf{QQ}^T$, o sea de **L** (con sus propiedades) y $\mathbf{Q}^T\mathbf{Q}$ que resulta ser otra matriz **K** de propiedades relevantes en sus autovalores y autovectores. Así, el HJ-Biplot, obtenido a partir de la DVS de la matriz de incidencia vértice-arista **Q** contendrá la información de **L** y **K**.

Recordemos que de las propiedades de los Biplot en general y del HJ-Biplot, en particular se tiene que

$$\mathbf{L} = \mathbf{JJ}^T = \mathbf{QQ}^T$$
$$\mathbf{K} = \mathbf{HH}^T = \mathbf{Q}^T\mathbf{Q}$$

Hay tres métodos para el cálculo de la medida de centralidad de Bonacich y las tres alternativas permiten obtener similares resultados [332]:

1. Análisis factorial
2. Convergencia de una secuencia infinita
3. Resolución de un sistema de ecuaciones

Nos centraremos en el primer método, considerando que los tres obtienen la misma valoración para los vértices y los vectores difieren solo en un factor constante.

Estamos interesados en un vector $\mathbf{p} \in \Re^n$, tal que su i-ésima componente p_i recoja la interacción del vértice *i*. Se tiene que $p_i p_j$ debe ser próximo a a_{ij} y se interpreta el problema como la minimización de la diferencia de mínimos cuadrados, así pues estamos interesados en un vector **p** que minimice la siguiente expresión

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(p_i p_j - a_{ij}\right)^2$$

Se establece que la centralidad, independientemente del método de cálculo, es:

$$c_{EV} = \frac{|\mathbf{p}|}{\|\mathbf{p}\|}$$

Este índice tiene un problema: en general cada autovector tenderá a corresponderse con las cargas (*loadings*) en cada cluster o agrupación, por lo que no será un índice satisfactorio para caracterizar la centralidad sobre rutas (*walk-based centrality*). Por otro lado, para medir cómo de concentrado es un grafo o, propiamente dicho, como de cercano está el grafo a la estructura ideal núcleo-periferia, en la que el núcleo se corresponde con un subgrafo completo y los nodos en la periferia no interaccionan entre sí (¡una estrella!), se propone calcular el autovector principal de la matriz de adyacencia y obtener los valores de c_{EV}, los más "nucleares" tendrán valores altos, mientras que aquellos más periféricos tenderán a cero.

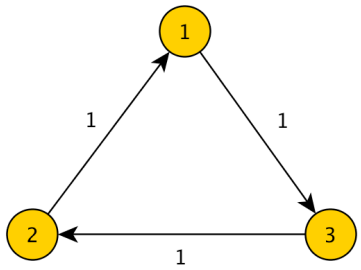
Hemos comprobado en esta revisión bibliográfica, como el análisis espectral de grafos permite extraer valiosa información de la estructura de los mismos, con lo que una representación de una red que incorpore dicho análisis espectral, en el fondo estará encerrando toda esa información sobre la estructura de interna de la red en una única representación gráfica.

7.7. INSPECCIÓN HJ-BIPLOT DE GRAFOS Y REDES

7.7.1. TOY PROBLEMS: PRUEBAS SOBRE GRAFOS BÁSICOS.

Vamos a analizar unos grafos simples, que nos permitan comprobar el potencial de la aplicación del HJ-Biplot a la matriz de incidencia de un grafo.

Comencemos por un grafo de 3 nodos y 3 aristas sin ponderar. Este grafo elemental es completo. Las aristas ha sido dotadas de una orientación arbitraria para construir la matriz de incidencia con signo.



Grafo completo con 3 nodos y 3 aristas, sin ponderar

La correspondiente matriz de incidencia es

Q	1-2	1-3	2-3
1	+1	-1	0
2	-1	0	+1
3	0	+1	-1

En donde se ha aplicado el signo positivo para aristas entrantes al nodo correspondiente y el signo negativo para aristas salientes del nodo correspondiente.

Por el orden del grafo (número de vértices) y su topología al efectuar la DVS de la matriz **Q** el 100% de la inercia está absorbida en el primer plano factorial.

Obtengamos las matrices $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$, que en este caso son iguales.

$$\mathbf{QQ}^T = \begin{bmatrix} +2 & -1 & -1 \\ -1 & +2 & -1 \\ -1 & -1 & +2 \end{bmatrix} = \mathbf{L} \qquad \mathbf{Q}^T\mathbf{Q} = \begin{bmatrix} +2 & -1 & -1 \\ -1 & +2 & -1 \\ -1 & -1 & +2 \end{bmatrix} = \mathbf{K} \qquad \mathbf{QQ}^T = \mathbf{L} = \mathbf{Q}^T\mathbf{Q} = \mathbf{K}$$

Si obtenemos para nuestro grafo la matriz **A** de Adyacencia y de grado **D** podemos calcular la matriz Laplaciana como **L** = **D**-**A**, que coincide con la matriz $\mathbf{QQ}^T = \mathbf{Q}^T\mathbf{Q}$

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad \mathbf{L} = \mathbf{D} - \mathbf{A} = \begin{bmatrix} +2 & -1 & -1 \\ -1 & +2 & -1 \\ -1 & -1 & +2 \end{bmatrix}$$

Calculemos la Descomposición en Valores Singulares de la matriz de incidencia **Q**. Los valores singulares de **Q**, raíces cuadradas de los autovalores de $\mathbf{QQ}^T$ o $\mathbf{Q}^T\mathbf{Q}$, son

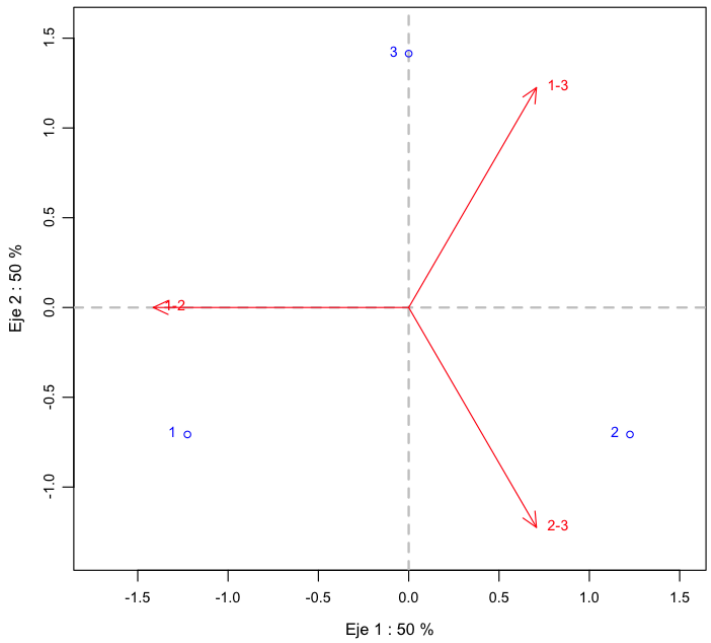
$$\sqrt{3} \qquad \sqrt{3} \qquad 0$$

Esto cumple una de las propiedades de la matriz Laplaciana **L** = **QQ**^{T} que indica que para un grafo completo de n vértices, la conectividad algebraica del grafo, su menor autovalor no nulo de **L**, es n en nuestro caso 3.

Y los autovectores por la derecha y por la izquierda correspondientes con la descomposición en valores singulares **U** y **V** son

$$\mathbf{U} = \begin{bmatrix} \frac{-\sqrt{2}}{2} & \frac{-1}{\sqrt{6}} & \frac{+\sqrt{3}}{3} \\ \frac{+\sqrt{2}}{2} & \frac{-1}{\sqrt{6}} & \frac{+\sqrt{3}}{3} \\ 0 & \frac{+2}{\sqrt{6}} & \frac{+\sqrt{3}}{3} \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} \frac{-2}{\sqrt{6}} & 0 & \frac{-\sqrt{3}}{3} \\ \frac{+1}{\sqrt{6}} & \frac{+\sqrt{2}}{2} & \frac{-\sqrt{3}}{3} \\ \frac{+1}{\sqrt{6}} & \frac{-\sqrt{2}}{2} & \frac{-\sqrt{3}}{3} \end{bmatrix}$$

Obtengamos ahora la representación HJ-Biplot de la matriz de incidencia **Q**

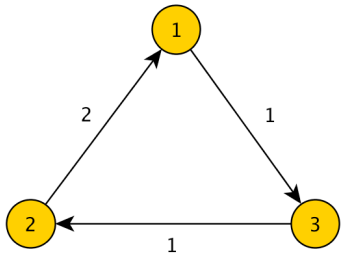


HJ-Biplot grafo completo de 3 nodos y 3 aristas.

Todos los marcadores, **H** y **J**, se encuentran sobre una circunferencia de radio √2 a causa a la "simetría" del grafo (todos los nodos y aristas son "iguales"). Debido al orden del grafo y la absorción del 100% de la variabilidad (inercia) en este primer plano, las CRFE suman 1000 para todos los marcadores, esto es, están todos perfectamente representados en este primer plano factorial.

En el siguiente paso a nuestro grafo de 3 nodos y 3 aristas completo le ponderamos una arista (la 1-2, por ejemplo, y sin pérdida de generalidad) con valor/peso 2.

Mantenemos la misma orientación arbitraria de las aristas que se asignó anteriormente para construir la matriz de incidencia con signo.



Grafo completo con 3 nodos y 3 aristas, con una arista ponderada.

La correspondiente matriz de incidencia es

Q	1-2	1-3	2-3
1	+2	-1	0
2	-2	0	+1
3	0	+1	-1

Como estamos en presencia de una red con ponderación debemos utilizar la matriz de incidencia generalizada **Q**

Q'	1-2	1-3	2-3
1	+√2	-√1	0
2	-√2	0	+√1
3	0	+√1	-√1

En donde, como anteriormente, se ha aplicado el signo positivo para aristas entrantes al nodo correspondiente y el signo negativo para aristas salientes del nodo correspondiente.

Por el orden del grafo (número de vértices) al efectuar la DVS de la matriz **Q** el 100% de la inercia está absorbida en el primer plano factorial, como en el ejemplo previo.

Obtengamos las matrices $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$ que ahora ya no serán iguales

$$\mathbf{QQ}^T = \begin{bmatrix} +3 & -2 & -1 \\ -2 & +3 & -1 \\ -1 & -1 & +2 \end{bmatrix} = \mathbf{L} \qquad \mathbf{Q}^T\mathbf{Q} = \begin{bmatrix} +4 & -\sqrt{2} & -\sqrt{2} \\ -\sqrt{2} & +2 & -1 \\ -\sqrt{2} & -1 & +2 \end{bmatrix} = \mathbf{K}$$

Si obtenemos para nuestro grafo la matriz **A** de Adyacencia, de grado **D** y la matriz Laplaciana como **L** = **D**-**A** con lo que se verifica que $\mathbf{QQ}^T = \mathbf{L}$

$$\mathbf{A}=\begin{bmatrix}0 & 2 & 1\\ 2 & 0 & 1\\ 1 & 1 & 0\end{bmatrix} \qquad \mathbf{D}=\begin{bmatrix}3 & 0 & 0\\ 0 & 3 & 0\\ 0 & 0 & 2\end{bmatrix} \qquad \mathbf{L}=\mathbf{D}-\mathbf{A}=\begin{bmatrix}+3 & -2 & -1\\ -2 & +3 & -1\\ -1 & -1 & +2\end{bmatrix}$$

Calculamos la Descomposición en Valores Singulares de **Q**. Los valores singulares de **Q**, raíces cuadradas de los autovalores de $\mathbf{QQ}^T$ o $\mathbf{Q}^T\mathbf{Q}$, son

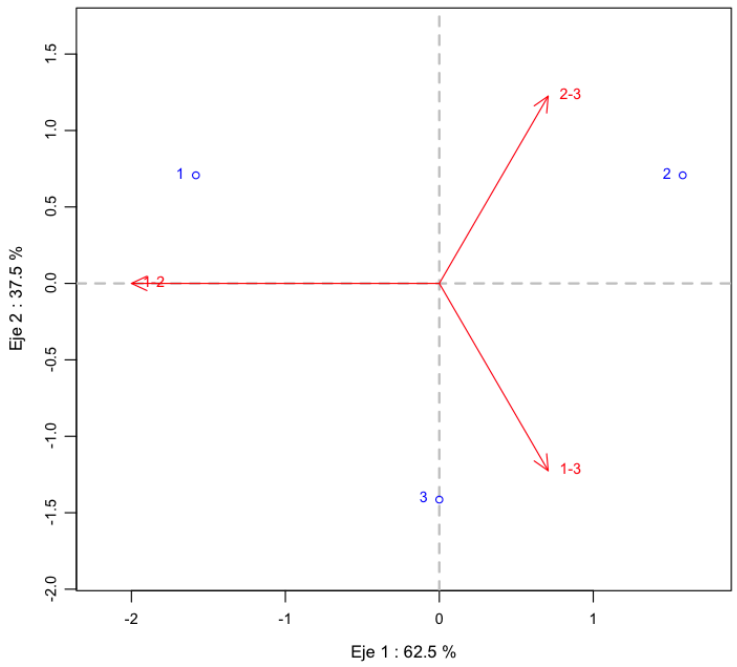
$$\sqrt{5} \qquad \sqrt{3} \qquad 0$$

Como vemos se verifica también de nuevo la propiedad de la conectividad algebraica que indica que para un grafo completo de n vértices, la valoración característica del grafo, menor autovalor (cuadrado del valor singular) no nulo de **L**, es n en nuestro caso 3.

Y los autovectores por la derecha y por la izquierda correspondientes con la descomposición en valores singulares **U** y **V** son

$$\mathbf{U}=\begin{bmatrix}\frac{-\sqrt{2}}{2} & \frac{+1}{\sqrt{6}} & \frac{+\sqrt{3}}{3}\\ \frac{+\sqrt{2}}{2} & \frac{+1}{\sqrt{6}} & \frac{+\sqrt{3}}{3}\\ 0 & \frac{-2}{\sqrt{6}} & \frac{+\sqrt{3}}{3}\end{bmatrix} \qquad \mathbf{V}=\begin{bmatrix}\frac{-2}{\sqrt{5}} & 0 & \frac{+1}{\sqrt{5}}\\ \frac{+1}{\sqrt{10}} & \frac{-\sqrt{2}}{2} & \frac{+2}{\sqrt{10}}\\ \frac{+1}{\sqrt{10}} & \frac{+\sqrt{2}}{2} & \frac{+2}{\sqrt{10}}\end{bmatrix}$$

Representemos el HJ-Biplot



HJ-Biplot grafo completo de 3 nodos y 3 aristas, la arista 1-2 ponderada el doble.

Como antes, debido al orden del grafo y la absorción del 100% de la variabilidad en este primer plano, las CRFE suman 1000 para todos los marcadores, esto es, de nuevo están todos los marcadores perfectamente representados en este primer plano factorial.

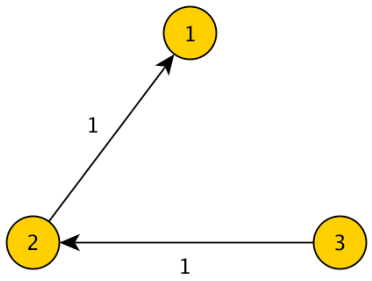
Ahora, lógicamente, los marcadores ya no se encuentran situados en una circunferencia, ya que hay características diferentes.

Marcador H (arista)	1-2	2-3	1-3
Módulo (longitud)	2	√2	√2

Marcador J (vértice)	1	2	3
Módulo (longitud)	√3	√3	√2

La arista con mayor peso, presenta una longitud mayor en su correspondiente marcador, al igual que sucede con los nodos que constituyen los extremos de dicha arista. Esto es, el peso de la arista afecta tanto a la longitud de marcador de la arista como, aparentemente, a la de los marcadores correspondientes a los nodos extremos de dicha arista.

Como siguiente “experimento” ahora a nuestro grafo de 3 nodos y 3 aristas, sin ponderar, le retiramos una de las aristas, por ejemplo y sin pérdida de generalidad, eliminamos la arista 1-3. Mantenemos la misma orientación arbitraria de las aristas que se asignó anteriormente para construir la matriz de incidencia con signo.



Grafo con 3 nodos y 2 aristas, sin ponderar.

La correspondiente matriz de incidencia es

Q	1-2	2-3
1	+1	0
2	-1	+1
3	0	-1

En donde, como anteriormente, se ha aplicado el signo positivo para aristas entrantes al nodo correspondiente y el signo negativo para aristas salientes del nodo correspondiente.

Como hasta ahora, por el orden del grafo (número de vértices) al efectuar la DVS de la matriz **Q** el 100% de la inercia está absorbida en el primer plano factorial, como en el ejemplo previo.

Obtengamos las matrices $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$

$$\mathbf{QQ}^T = \begin{bmatrix} +1 & -1 & 0 \\ -1 & +2 & -1 \\ 0 & -1 & +1 \end{bmatrix} = \mathbf{L} \qquad \mathbf{Q}^T\mathbf{Q} = \begin{bmatrix} +2 & -1 \\ -1 & +2 \end{bmatrix} = \mathbf{K}$$

Si obtenemos para nuestro grafo la matriz **A** de Adyacencia, de grado **D** y la matriz Laplaciana como **L = D-A** para verificar que $\mathbf{QQ}^T$ **= L.**

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{L} = \mathbf{D} - \mathbf{A} = \begin{bmatrix} +1 & -1 & 0 \\ -1 & +2 & -1 \\ 0 & -1 & +1 \end{bmatrix}$$

Los valores singulares de **Q**, raíces cuadradas de los autovalores de $\mathbf{QQ}^T$ o $\mathbf{Q}^T\mathbf{Q}$, son

$$\sqrt{3} \qquad 1 \qquad (0)$$

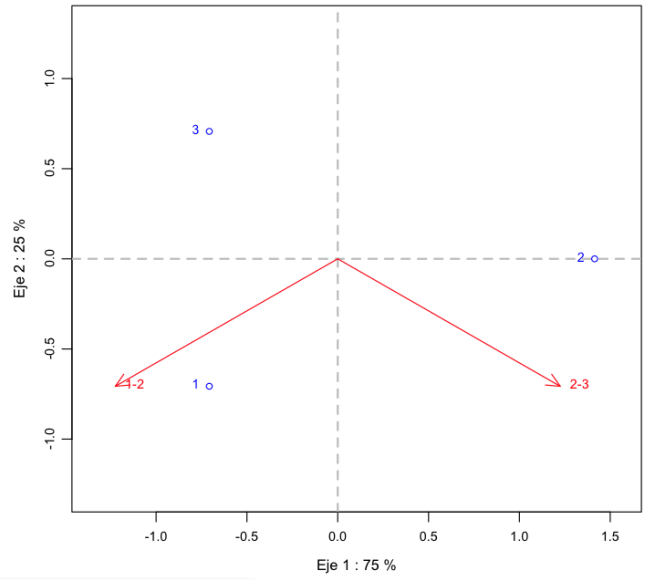
Recordemos que $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$ tienen los mismos autovalores no nulos (el autovalor nulo se obtiene para producto de mayor dimensión). Dado las dimensiones de ambas matrices, el número de autovalores de cada una es diferente, pero se cumple que tienen los mismos no nulos. En este caso, al estar en presencia de un grafo no completo se cumple que la conectividad algebraica del grafo es menor o igual que n-2, siendo n el número de vértices del grafo. En nuestro caso n=3, y λ_2=1. También se cumple que siendo m el grado (o valencia) menor de los nodos $\lambda_2 \leq m=1$. Al desaparecer una arista del grafo, la "conectividad" se reduce y esa situación se detecta en la reducción del valor de la conectividad algebraica o de Fiedler.

Los autovectores por la derecha y por la izquierda correspondientes con la descomposición en valores singulares **U** y **V** son

$$\mathbf{U} = \begin{bmatrix} \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \\ \frac{+2}{\sqrt{6}} & 0 \\ \frac{-1}{\sqrt{6}} & \frac{+1}{\sqrt{2}} \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{+1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$$

Aquí tiene sentido explorar la centralidad y su efecto en los marcadores: el primer autovector de **U** tiene la componente más elevada (en valor absoluto) para el nodo "2" que es evidentemente el más central de los tres que componen este grafo trivial.

Obtengamos ahora el HJ-Biplot



HJ-Biplot grafo de 3 nodos y 2 aristas, no ponderadas (desaparece arista 1-2).

Como ya se ha expuesto, debido al orden del grafo y la absorción del 100% de la variabilidad en este primer plano, las CRFE suman 1000 para todos los marcadores, esto es, están todos los marcadores perfectamente representados en este primer plano factorial.

Pero ahora los marcadores tampoco se encuentran situados sobre una circunferencia.

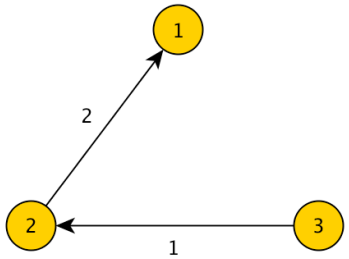
Marcador H (arista)	1-2	2-3
Módulo (longitud)	√2	√2

Marcador J (vértice)	1	2	3
Módulo (longitud)	1	√2	1

Las dos aristas tienen la misma longitud, realmente tienen las mismas características en el grafo, al igual que sucede con los nodos que se encuentran en los extremos del

grafo. El nodo “central” presenta no obstante una longitud mayor que los otros dos, que además es la misma que la de los marcadores correspondientes a las aristas. Esto podría representar, y ya lo hemos intuido anteriormente, que quizá el módulo del marcador **J** podría tener alguna relación con algún aspecto de “centralidad” de la arista/nodo correspondiente. Recordemos que la segunda componente del autovector por la izquierda correspondiente al autovalor/valor propio más elevado tenía el valor más alto de las tres.

Para finalizar la casuística con nuestro grafo de 3 nodos y 2 aristas, le ponderamos una arista (la 1-2, por ejemplo, y sin pérdida de generalidad) con valor 2. Mantenemos la misma orientación arbitraria de las aristas que se asignó anteriormente para construir la matriz de incidencia con signo.



Grafo con 3 nodos y 2 aristas, con una arista ponderada.

La correspondiente matriz de incidencia es

Q	1-2	2-3
1	+2	0
2	-2	+1
3	0	-1

Como estamos en presencia de una red con ponderación debemos utilizar la matriz de incidencia generalizada **Q** considerando la matriz de pesos **W**.

Q	1-2	2-3
1	+√2	0
2	-√2	+√1
3	0	-√1

En donde, como anteriormente, se ha aplicado el signo positivo para aristas entrantes al nodo correspondiente y el signo negativo para aristas salientes del nodo correspondiente.

Por el orden del grafo (número de vértices) al efectuar la DVS de la matriz **Q** el 100% de la inercia está absorbida en el primer plano factorial, como en el ejemplo previo.

Obtengamos las matrices $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$

$$\mathbf{QQ}^T = \begin{bmatrix} +2 & -2 & 0 \\ -2 & +3 & -1 \\ 0 & -1 & +1 \end{bmatrix} = \mathbf{L} \qquad \mathbf{Q}^T\mathbf{Q} = \begin{bmatrix} +4 & -\sqrt{2} \\ -\sqrt{2} & +2 \end{bmatrix} = \mathbf{K}$$

Una vez más obtenemos para nuestro grafo la matriz **A** de Adyacencia, de grado **D** y la matriz Laplaciana como **L = D**-**A**

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & 0 \\ 2 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{L} = \mathbf{D} - \mathbf{A} = \begin{bmatrix} +2 & -2 & 0 \\ -2 & +2 & -1 \\ 0 & -1 & +1 \end{bmatrix}$$

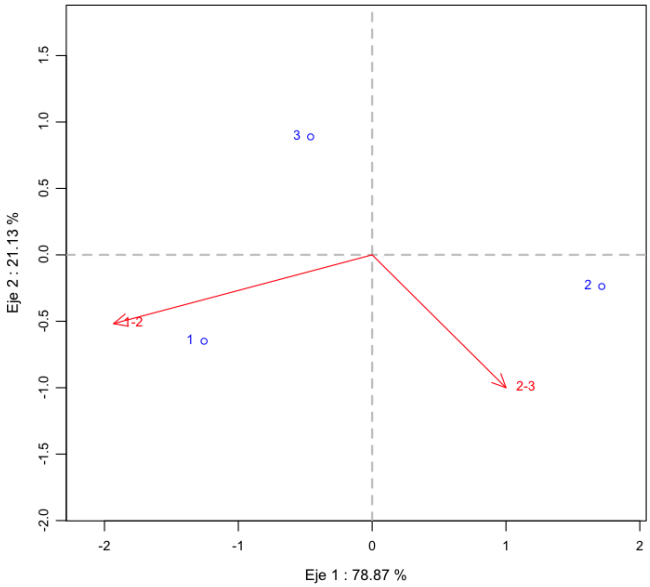
Los valores singulares de **Q**, raíces cuadradas de los autovalores de $\mathbf{QQ}^T$ o $\mathbf{Q}^T\mathbf{Q}$, son

2.175328 1.126033 (0)

Como siempre $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$ tienen los mismos autovalores no nulos, y como en ocasiones anteriores el autovalor 0 se corresponde con el producto correspondiente a la Laplaciana. Los autovectores por la derecha y por la izquierda correspondientes con la descomposición en valores singulares **U** y **V** son

$$\mathbf{U} = \begin{bmatrix} \frac{-1}{\sqrt{3}} & \frac{-1}{\sqrt{3}} \\ +0.7886751 & -0.2113249 \\ -0.2113249 & +0.7886751 \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} -0.8880738 & -0.4597008 \\ -0.4597008 & -0.8880738 \end{bmatrix}$$

Obtengamos el HJ-Biplot



HJ-Biplot grafo completo de 3 nodos y 2 aristas, arista 1-2 ponderada el doble.

Como hasta ahora, debido al orden del grafo y la absorción del 100% de la variabilidad en este primer plano, las CRFE suman 1000 para todos los marcadores, esto es, están todos perfectamente representados en este primer plano factorial.

Ahora, lógicamente, tampoco los marcadores se encuentran situados sobre una circunferencia.

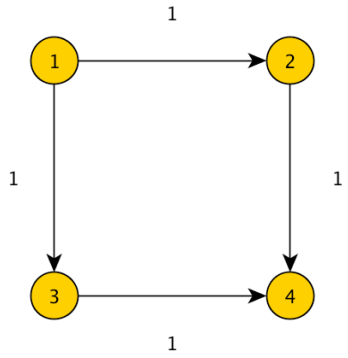
Marcador H (arista)	1-2	2-3
Módulo (longitud)	2	√2

Marcador J (vértice)	1	2	3
Módulo (longitud)	√2	√3	1

La arista con mayor peso, presenta una longitud mayor en su correspondiente marcador, al igual que sucede con los nodos que constituyen los extremos de dicha arista y aún presenta mayor longitud en el caso del nodo “central”. Esto último refuerza una vez más la idea de que estos marcadores quizá podrían asociarse a la “centralidad” en algún sentido, del nodo, tal y como ya intuíamos en el caso anterior y observando el primer autovector por la izquierda de la DVS de **Q**.

Por lo visto hasta el momento, la longitud de los marcadores asociados a las aristas **H** tienen una relación directa con el peso de las aristas, mientras que las longitudes de los marcadores asociados a los nodos **J** aparentan tener relación directa con los pesos de las aristas que los unen y probablemente también con la posición del nodo en el grafo (centralidad).

Compliquemos un poco más nuestro grafo “elemental” añadiéndole un cuarto nodo y una cuarta arista. Empecemos, como antes, por un grafo que una los 4 nodos (en este caso NO es completo, ya que faltarían por unir los nodos opuestos).



Grafo con 4 nodos y 4 aristas, sin ponderación.

La correspondiente matriz de incidencia es

Q	1-2	1-3	2-4	3-4
1	-1	-1	0	0
2	+1	0	-1	0
3	0	+1	0	-1
4	0	0	+1	1

En donde, como anteriormente, se ha aplicado el signo positivo para aristas entrantes al nodo correspondiente y el signo negativo para aristas salientes del nodo correspondiente.

Al estar ahora en presencia de un grafo de orden 4 (número de vértices) al efectuar la DVS de la matriz **Q** ya el 100% de la inercia probablemente no estará absorbida en el primer plano factorial, de hecho ahora el primer plano solo explica el 75% de la variabilidad de los datos.

Obtengamos las matrices $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$

$$\mathbf{QQ}^T = \begin{bmatrix} +2 & -1 & -1 & 0 \\ -1 & +2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix} = \mathbf{L} \qquad \mathbf{Q}^T\mathbf{Q} = \begin{bmatrix} +2 & +1 & -1 & 0 \\ +1 & +2 & 0 & -1 \\ -1 & 0 & +2 & +1 \\ 0 & -1 & +1 & +2 \end{bmatrix} = \mathbf{K}$$

Si obtenemos para nuestro grafo la matriz **A** de Adyacencia, de grado **D** y comprobemos que la matriz Laplaciana como **L** = **D**-**A** coincide con $\mathbf{QQ}^T$

$$\mathbf{A}=\begin{bmatrix}0&1&1&0\\1&0&0&1\\1&0&0&1\\0&1&1&0\end{bmatrix} \qquad \mathbf{D}=\begin{bmatrix}2&0&0&0\\0&2&0&0\\0&0&2&0\\0&0&0&2\end{bmatrix} \qquad \mathbf{L}=\mathbf{D}-\mathbf{A}=\begin{bmatrix}+2&-1&-1&0\\-1&+2&0&-1\\-1&0&+2&-1\\0&-1&-1&+2\end{bmatrix}$$

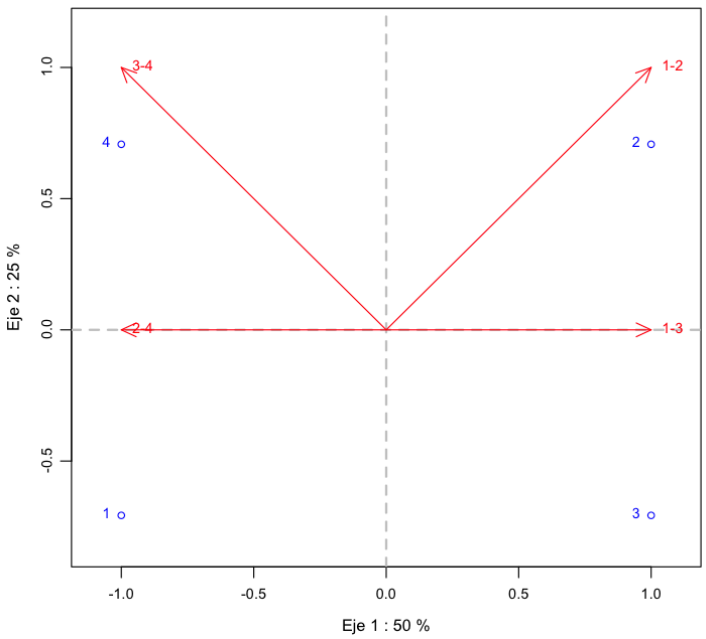
Los valores singulares de **Q**, raíces cuadradas de los autovalores de $\mathbf{QQ}^T$ o $\mathbf{Q}^T\mathbf{Q}$, son

$$2 \qquad \sqrt{2} \qquad \sqrt{2} \qquad 0$$

El grafo no es completo y la conectividad algebraica es menor o igual a n-2, esto es 2. Y los autovectores por la derecha y por la izquierda correspondientes con la descomposición en valores singulares **U** y **V** son

$$\mathbf{U}=\begin{bmatrix}-1/2&-1/2&-1/2&+1/2\\+1/2&+1/2&-1/2&+1/2\\+1/2&-1/2&+1/2&+1/2\\-1/2&+1/2&+1/2&+1/2\end{bmatrix} \qquad \mathbf{V}=\begin{bmatrix}+1/2&\frac{+1}{\sqrt{2}}&0&+1/2\\+1/2&0&\frac{+1}{\sqrt{2}}&-1/2\\-1/2&0&\frac{+1}{\sqrt{2}}&+1/2\\-1/2&\frac{+1}{\sqrt{2}}&0&-1/2\end{bmatrix}$$

Como es obvio no tiene aquí sentido analizar la centralidad de vértices o aristas del grafo. Obtengamos la representación HJ-Biplot en el primer plano factorial



HJ-Biplot grafo completo de 4 nodos y 4 aristas, sin ponderar.

La absorción de solo el 75% de la variabilidad en este primer plano también se pone de manifiesto en las CRFE que ya no suman 1000 para todos los marcadores, esto es, no todos los marcadores están ya perfectamente representados en este primer plano factorial. En una representación tridimensional, sí estarían de nuevo perfectamente representados.

Marcador	PC1	PC2	PC3
1	500	250	250
2	500	250	250
3	500	250	250
4	500	250	250

1-2	500	500	0
1-3	500	0	500
2-4	500	0	500
3-4	500	500	0

Esto tiene una consecuencia directa en la interpretación de la representación HJ-Biplot del grafo: La longitud de los marcadores, como ya sucedía en el primer ejemplo, es igual en todos los casos a √2 considerando todo el espacio vectorial, y no solo el plano de dimensión reducida en el que esta igualdad entre todos los marcadores no se presenta. Esto quiere decir que analizados los marcadores **H** y **J** para el espacio tridimensional en el que se enmarcan (debido al grado 4 del grafo bajo estudio) podemos concluir que todos ellos presentan características equivalentes, conclusión que no podemos concluir analizando la proyección de baja dimensión correspondiente al primer plano factorial.

	Espacio tridimensional (100%)			
Marcador H (arista)	1-2	1-3	2-4	3-4
Módulo (longitud)	√2	√2	√2	√2

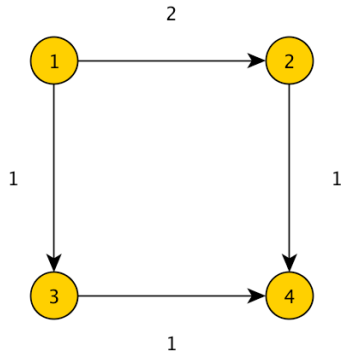
Espacio bidimensional (75%)			
1-2	1-3	2-4	3-4
√2	1	1	√2

Marcador J (vértice)	1	2	3	4
Módulo (longitud)	√2	√2	√2	√2

1	2	3	4
√(3/2)	√(3/2)	√(3/2)	√(3/2)

Los marcadores correspondientes a los vértices tienen igual calidad de representación todos ellos en el primer plano factorial, por lo que sus posiciones pueden seguir siendo interpretables, al estar todas “penalizadas” por igual. Pero no sucede así en el caso de los marcadores correspondientes a las aristas, cuyas posiciones inducen a error al interpretarse características diferentes entre ellas, cuando en realidad no es así.

Continuamos, como antes, modificando nuestro grafo para que uno de las 4 aristas tenga un peso mayor que las restantes (sin pérdida de generalidad consideremos la arista 1-2 como ponderada con peso 2).



Grafo con 4 nodos y 4 aristas, con ponderación.

La correspondiente matriz de incidencia es

Q	1-2	1-3	2-4	3-4
1	-2	-1	0	0
2	+2	0	-1	0
3	0	+1	0	-1
4	0	0	+1	1

En donde, como anteriormente, se ha aplicado el signo positivo para aristas entrantes al nodo correspondiente y el signo negativo para aristas salientes del nodo correspondiente.

Como estamos en presencia de una red con ponderación debemos utilizar la matriz de incidencia generalizada **Q**, considerando la matriz de ponderaciones **W**.

Q	1-2	1-3	2-4	3-4
1	-√2	-√1	0	0
2	+√2	0	-√1	0
3	0	+√1	0	-√1
4	0	0	+√1	+√1

Obtengamos las matrices $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$

$$\mathbf{QQ}^T = \begin{bmatrix} +3 & -2 & -1 & 0 \\ -2 & +3 & 0 & -1 \\ -1 & 0 & +2 & -1 \\ 0 & -1 & -1 & +2 \end{bmatrix} = \mathbf{L} \qquad \mathbf{Q}^T\mathbf{Q} = \begin{bmatrix} +4 & +\sqrt{2} & -\sqrt{2} & 0 \\ +\sqrt{2} & +2 & 0 & -1 \\ -\sqrt{2} & 0 & +2 & +1 \\ 0 & -1 & +1 & +2 \end{bmatrix} = \mathbf{K}$$

Si obtenemos para nuestro grafo la matriz **A** de Adyacencia, de grado **D** y la matriz Laplaciana como **L** = **D**-**A**

$$\mathbf{A}=\begin{bmatrix}0&2&1&0\\2&0&0&1\\1&0&0&1\\0&1&1&0\end{bmatrix}\qquad \mathbf{D}=\begin{bmatrix}3&0&0&0\\0&3&0&0\\0&0&2&0\\0&0&0&2\end{bmatrix}\qquad \mathbf{L}=\mathbf{D}-\mathbf{A}=\begin{bmatrix}+3&-2&-1&0\\-2&+3&0&-1\\-1&0&+2&-1\\0&-1&-1&+2\end{bmatrix}$$

De nuevo, al estar ahora en presencia de un grafo de orden 4 (número de vértices) al efectuar la DVS de la matriz **Q** ya el 100% de la inercia no está absorbida en el primer plano factorial, ahora el primer plano solo explica el 80% de la variabilidad de los datos.

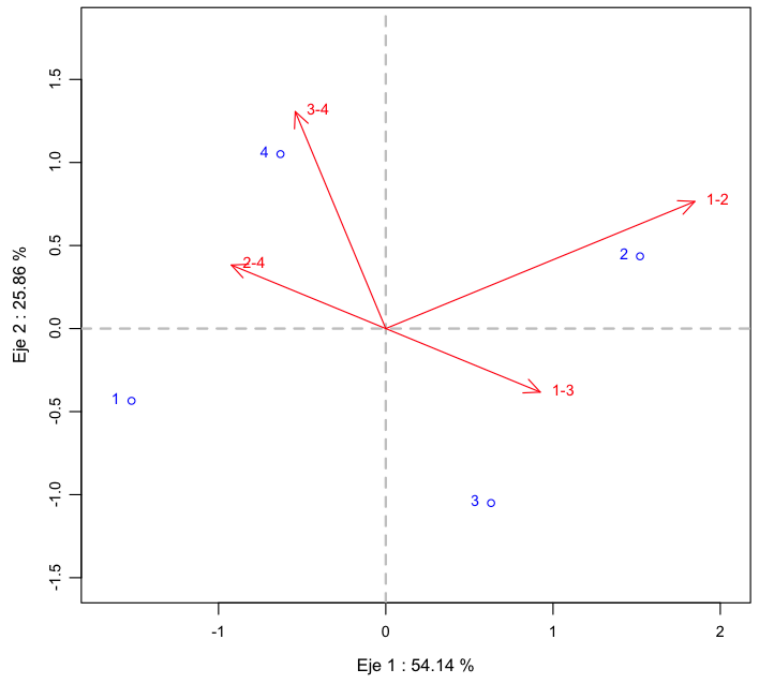
Los valores singulares de **Q**, raíces cuadradas de los autovalores de $\mathbf{QQ}^{\mathsf{T}}$ o $\mathbf{Q}^{\mathsf{T}}\mathbf{Q}$, son

$$2.326846 \qquad 1.608038 \qquad \sqrt{2} \qquad 0$$

Y los autovectores por la derecha y por la izquierda correspondientes con la descomposición en valores singulares **U** y **V** son

$$\mathbf{U}=\begin{bmatrix}-0.6532815&-0.2705981&+1/2&+1/2\\+0.6532815&+0.2705981&+1/2&+1/2\\+0.2705981&-0.6532815&-1/2&+1/2\\-0.2705981&+0.6532815&-1/2&+1/2\end{bmatrix}\quad \mathbf{V}=\begin{bmatrix}+0.7941045&+0.4759631&0&-0.3779645\\+0.3970522&-0.2379816&\frac{-1}{\sqrt{2}}&+0.5345225\\-0.3970522&+0.2379816&\frac{-1}{\sqrt{2}}&-0.5345225\\-0.2325878&+0.8125199&0&+0.5345225\end{bmatrix}$$

Obtengamos la representación HJ-Biplot en el primer plano factorial



HJ-Biplot grafo completo de 4 nodos y 4 aristas ponderado.

La absorción de solo el 80% de la variabilidad en este primer plano también se pone de manifiesto de nuevo en las CRFE que ya no suman 1000 para todos los marcadores, esto es, no todos los marcadores están ya perfectamente representados en este primer plano factorial.

Marcador	PC1	PC2	PC3
1	770.2	63.1	166.7
2	770.2	63.1	166.7
3	198.2	551.8	250.0
4	198.2	551.8	250.0

1-2	853.6	146.4	0
1-3	426.8	73.2	500
2-4	426.8	73.2	500
3-4	146.4	853.6	0

Esto afecta a la interpretación del HJ-Biplot obtenido en el espacio de baja dimensión, ya que al no tener todos los marcadores máxima calidad de representación en el plano algunos no están suficientemente bien representados. Por ejemplo, considerando el espacio tridimensional completo el marcador correspondiente a la arista 1-2 presenta longitud 2, mientras que las restantes solo tienen longitud √2, justificándose la situación en la ponderación mayor de dicha arista. Al igual que sucede con los marcadores de los nodos 1 y 2 que presenta mayor longitud que los correspondientes a los nodos 3 y 4, en línea con lo visto en un ejemplo anterior.

	Espacio tridimensional (100%)			
Marcador H (arista)	1-2	1-3	2-4	3-4
Módulo (longitud)	2	√2	√2	√2

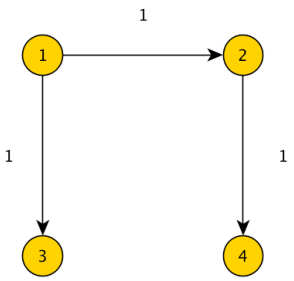
Espacio bidimensional (80%)			
1-2	1-3	2-4	3-4
2	1	1	√2

Marcador J (vértice)	1	2	3	4
Módulo (longitud)	√3	√3	√2	√2

1	2	3	4
√(5/2)	√(5/2)	√(3/2)	√(3/2)

Los marcadores fila y columna tienen diferentes calidades de representación en el primer plano factorial, por lo que sus posiciones no pueden ser interpretables, al estar “penalizadas” de diferente manera, por lo que sus posiciones inducen a error al interpretarse características diferentes, cuando en realidad no es así.

Vamos ahora a retirar una arista en nuestro grafo de 4 nodos.



Grafo con 4 nodos y 3 aristas, sin ponderación.

La correspondiente matriz de incidencia es

Q	1-2	1-3	2-4
1	-1	-1	0
2	+1	0	-1
3	0	+1	0
4	0	0	+1

En donde, como anteriormente, se ha aplicado el signo positivo para aristas entrantes al nodo correspondiente y el signo negativo para aristas salientes del nodo correspondiente.

Obtengamos las matrices $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$

$$\mathbf{QQ}^T = \begin{bmatrix} +2 & -1 & -1 & 0 \\ -1 & +2 & 0 & -1 \\ -1 & 0 & +1 & 0 \\ 0 & -1 & 0 & +1 \end{bmatrix} = \mathbf{L} \qquad \mathbf{Q}^T\mathbf{Q} = \begin{bmatrix} +2 & +1 & -1 \\ +1 & +2 & 0 \\ -1 & 0 & +2 \end{bmatrix} = \mathbf{K}$$

Si obtenemos para nuestro grafo la matriz **A** de Adyacencia, de grado **D** y la matriz Laplaciana como **L = D-A**

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{L} = \mathbf{D} - \mathbf{A} = \begin{bmatrix} +2 & -1 & -1 & 0 \\ -1 & +2 & 0 & -1 \\ -1 & 0 & +1 & 0 \\ 0 & -1 & 0 & +1 \end{bmatrix}$$

Al estar ahora en presencia de un grafo de orden 4 (número de vértices) al efectuar la DVS de la matriz **Q** ya el 100% de la inercia no está absorbida en el primer plano factorial, ahora el primer plano explica el 90.23% de la variabilidad de los datos.

Los valores singulares de **Q**, raíces cuadradas de los autovalores de $\mathbf{QQ}^T$ o $\mathbf{Q}^T\mathbf{Q}$, son

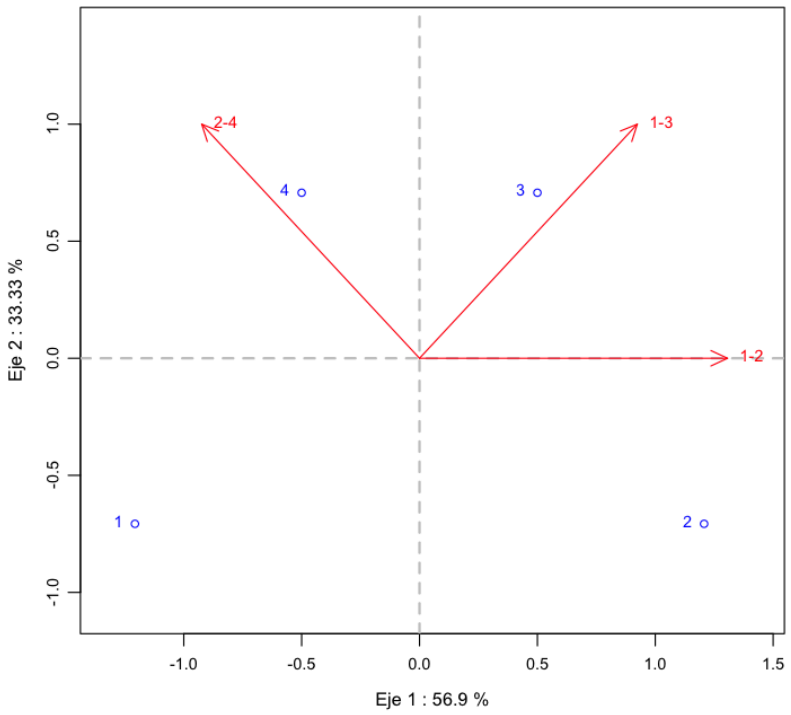
$$1.847759 \qquad \sqrt{2} \qquad 0.7653669 \qquad (0)$$

La conectividad algebraica pasa de $\sqrt{2}$ a 0.7653669, como consecuencia de la eliminación de una arista en el grafo y la consiguiente pérdida de “conectividad” en el mismo, como era de esperar. Los autovectores por la derecha y por la izquierda correspondientes con la descomposición en valores singulares **U** y **V** son:

$$\mathbf{U} = \begin{bmatrix} -0.6532815 & -1/2 & -0.2705981 \\ +0.6532815 & -1/2 & +0.2705981 \\ +0.2705981 & +1/2 & -0.6532815 \\ -0.2705981 & +1/2 & +0.6532815 \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} \frac{+1}{\sqrt{2}} & 0 & \frac{+1}{\sqrt{2}} \\ \frac{+1}{2} & \frac{+1}{\sqrt{2}} & \frac{-1}{2} \\ \frac{-1}{2} & \frac{+1}{\sqrt{2}} & \frac{+1}{2} \end{bmatrix}$$

El primer vector de la matriz **U** nos ofrece información sobre la centralidad de los nodos. Si consideramos el valor absoluto de las componentes podemos deducir la mayor centralidad de los vértices 1 y 2 y menor de los 3 y 4, como efectivamente sucede a la vista del grafo. La matriz **U** nos ofrece información sobre la centralidad de las aristas, reflejándose de nuevo la mayor centralidad de la arista 1-2 como es obvio a la vista del grafo bajo estudio.

Obtengamos la representación HJ-Biplot en el primer plano factorial



HJ-Biplot grafo completo de 4 nodos y 3 aristas, sin ponderar.

La absorción del 90.23% de la variabilidad en este primer plano representa un porcentaje, en principio, muy elevado, pero también se pone de manifiesto en las CRFE que ya no suman 1000 para todos los marcadores, esto es, no todos los marcadores están ya perfectamente representados en este primer plano factorial.

Marcador	PC1	PC2	PC3
1	728.6	250	21.4
2	728.6	250	21.4
3	250.0	500	250
4	250.0	500	250

1-2	853.6	0	146.4
1-3	426.8	500	73.2
2-4	426.8	500	73.2

Esto tiene, de nuevo, consecuencias directas en la interpretación de la representación HJ-Biplot del grafo: En el espacio tridimensional (100 de absorción de inercia) todos los marcadores fila (aristas) tienen igual longitud √2, mientras que los marcadores columna correspondientes a los nodos 1 y 2 (más centrales) tiene un valor √2 superior al de los nodos 3 y 4 (nodos extremos) que es igual a 1.

	Espacio tridimensional (100%)		
Marcador H (arista)	1-2	1-3	2-4
Módulo (longitud)	√2	√2	√2

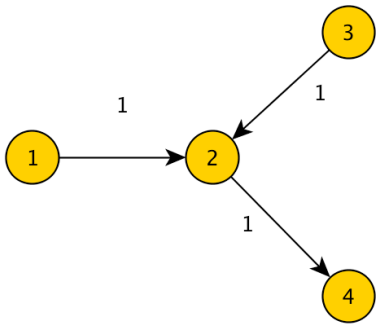
Espacio bidimensional (75%)		
1-2	1-3	2-4
1.307	1.361	1.361

Marcador J (vértice)	1	2	3	4
Módulo (longitud)	√2	√2	1	1

1	2	3	4
1.399	1.399	0.866	0.866

Los marcadores correspondientes a los vértices mantienen el emparejamiento observado en el espacio tridimensional, por lo que sus posiciones pueden seguir siendo interpretables, al estar todas "penalizadas" por igual. Pero no sucede así en el caso de los marcadores correspondientes a las aristas, cuyas posiciones inducen a error al interpretarse características diferentes entre ellas, cuando en realidad no es así, aunque las diferencias en este caso sean pequeñas entre ellas.

Como en algún ejemplo anterior hemos creído interpretar el mayor módulo de algún marcador motivada por la "centralidad" del nodo o arista correspondiente vamos a estudiar un caso particular de red de cuatro nodos y tres aristas: una estrella, en este caso sin ponderar, para intentar poner de manifiesto el efecto solo de la centralidad del nodo.



Grafo en estrella con 4 nodos y 3 aristas, sin ponderación.

La correspondiente matriz de incidencia es

Q	1-2	2-3	2-4
1	-1	0	0
2	+1	+1	+1
3	0	-1	0
4	0	0	-1

En donde, como anteriormente, se ha aplicado el signo positivo para aristas entrantes al nodo correspondiente y el signo negativo para aristas salientes del nodo correspondiente.

Obtengamos las matrices $\mathbf{QQ}^T$ y $\mathbf{Q}^T\mathbf{Q}$

$$\mathbf{QQ}^T = \begin{bmatrix} +1 & -1 & 0 & 0 \\ -1 & +3 & -1 & -1 \\ 0 & -1 & +1 & 0 \\ 0 & -1 & 0 & +1 \end{bmatrix} = \mathbf{L} \qquad \mathbf{Q}^T\mathbf{Q} = \begin{bmatrix} +2 & +1 & +1 \\ +1 & +2 & +1 \\ +1 & +1 & +2 \end{bmatrix} = \mathbf{K}$$

Si obtenemos para nuestro grafo la matriz **A** de Adyacencia, de grado **D** y la matriz Laplaciana como **L = D-A**

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{L} = \mathbf{D} - \mathbf{A} = \begin{bmatrix} +1 & -1 & 0 & 0 \\ -1 & +3 & -1 & -1 \\ 0 & -1 & +1 & 0 \\ 0 & -1 & 0 & +1 \end{bmatrix}$$

Una vez más al estar ahora en presencia de un grafo de orden 4 (número de vértices) al efectuar la DVS de la matriz **Q** ya el 100% de la inercia no está absorbida en el primer plano factorial, ahora el primer plano explica el 83.33% de la variabilidad de los datos.

Los valores singulares de **Q**, raíces cuadradas de los autovalores de $\mathbf{QQ}^T$ o $\mathbf{Q}^T\mathbf{Q}$, son

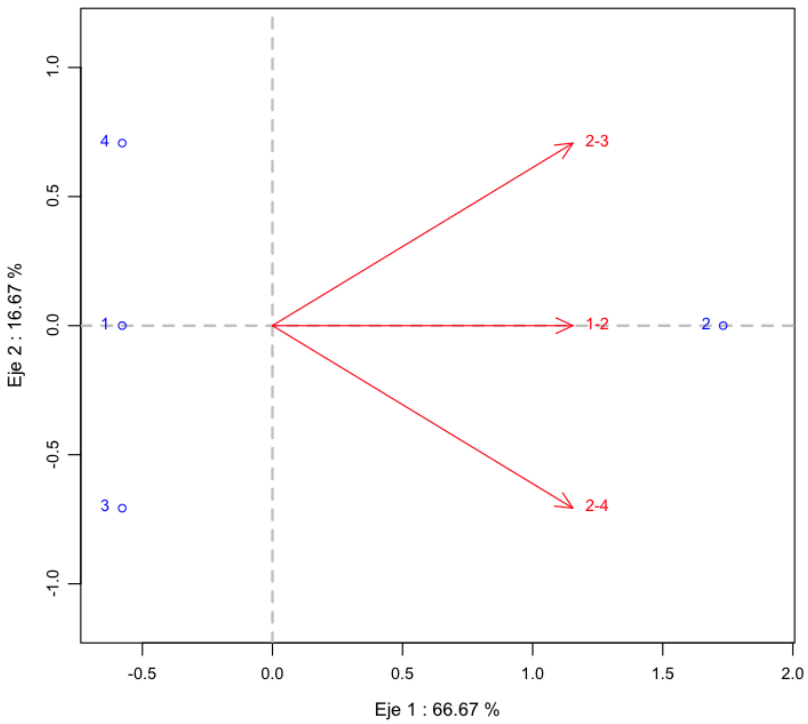
2 1 1 (0)

Se cumple de nuevo la propiedad de la conectividad algebraica que dice que para una estrella la conectividad algebraica toma valor 1. Y los autovectores por la derecha y por la izquierda correspondientes con la descomposición en valores singulares **U** y **V** son

$$\mathbf{U} = \begin{bmatrix} \frac{-1}{\sqrt{12}} & 0 & -\sqrt{\frac{2}{3}} \\ \frac{\sqrt{3}}{2} & 0 & 0 \\ \frac{-1}{\sqrt{12}} & \frac{-1}{\sqrt{2}} & \frac{+1}{\sqrt{6}} \\ \frac{-1}{\sqrt{12}} & \frac{-1}{\sqrt{2}} & \frac{+1}{\sqrt{6}} \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} \frac{+1}{\sqrt{3}} & 0 & \sqrt{\frac{2}{3}} \\ \frac{+1}{\sqrt{3}} & \frac{+1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{+1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \end{bmatrix}$$

Del primer vector de la matriz **U** obtenemos que "previsiblemente" el nodo "2" será más central en el grafo y los restantes jugarán un papel "similar". Del estudio del primer vector que compone la matriz **V** vemos que las tres aristas juegan un papel similar, en lógica trasposición de su posición en el grafo bajo estudio.

Obtengamos la representación HJ-Biplot en el primer plano factorial



HJ-Biplot grafo en estrella de 4 nodos y 3 aristas, sin ponderar.

La absorción del 83.33% de la variabilidad en este primer plano representa un porcentaje, en principio, muy elevado, pero también se pone de manifiesto en las CRFE que ya no suman 1000 para todos los marcadores, esto es, no todos los marcadores están ya perfectamente representados en este primer plano factorial.

Marcador	PC1	PC2	PC3
1	333.3	0	666.7
2	1000	0	0
3	333.3	500	166.7
4	333.3	500	166.7

1-2	666.7	0	333.3
1-3	666.7	250	83.3
2-4	666.7	250	83.3

Esto tiene, de nuevo, consecuencias directas en la interpretación de la representación HJ-Biplot del grafo: En el espacio tridimensional (100 de absorción de inercia) todos los marcadores fila (aristas) tienen igual longitud √2, en perfecta congruencia con sus características y posiciones en el grafo, mientras que los marcadores columna correspondientes a los nodos 1 y 2 (más centrales) tienen un valor √2 superior al de los nodos 3 y 4 (nodos extremos) que es igual a 1.

	Espacio tridimensional (100%)		
Marcador H (arista)	1-2	2-3	2-4
Módulo (longitud)	√2	√2	√2

Espacio bidimensional (83.3%)		
1-2	2-3	2-4
1.54701	1.354	1.354

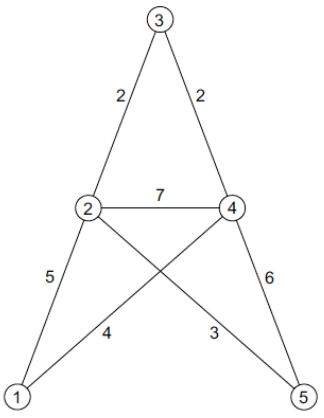
Marcador J (vértice)	1	2	3	4
Módulo (longitud)	1	√3	1	1

1	2	3	4
0.578	1.732	0.913	0.913

En el espacio de dimensión reducida la situación varía: Los marcadores **H** presentan diferencias entre sí y al igual sucede con los marcadores **J** correspondientes a los nodos no centrales de la estrella. Se mantiene que el marcador correspondiente al nodo central presenta una longitud superior a los otros tres de los restantes vértices.

7.7.2. Otro Ejemplo: Grafo "Torre Eiffel"

Koren *et al* [315] plantean como un "*toy problem*" el análisis de un grafo que, por razones "evidentes", denominan "grafo de la torre Eiffel" y que se muestra a continuación:



$$L = \begin{pmatrix} 9 & -5 & 0 & -4 & 0 \\ -5 & 17 & -2 & -7 & -3 \\ 0 & -2 & 4 & -2 & 0 \\ -4 & -7 & -2 & 19 & -6 \\ 0 & -3 & 0 & -6 & 9 \end{pmatrix}$$

Grafo denominado "Torre Eiffel" y matriz Laplaciana asociada. [315]

Se trata de un grafo de 5 nodos/vértices, conectado por 7 aristas con ponderaciones. Los autores proponen en primer lugar un método de representación a partir de la autodescomposición de la matriz Laplaciana del grafo y la proyección de los resultados en el plano correspondiente a los dos autovalores menores excepto el nulo.

En primer lugar, a partir del grafo construyamos la matriz Laplaciana, para ello obtengamos en primer lugar la matriz de adyacencia **A** del grafo:

$$\mathbf{A} = \begin{bmatrix} 0 & 5 & 0 & 4 & 0 \\ 5 & 0 & 2 & 7 & 3 \\ 0 & 2 & 0 & 2 & 0 \\ 4 & 7 & 2 & 0 & 6 \\ 0 & 3 & 0 & 6 & 0 \end{bmatrix}$$

Podemos obtener la matriz **D** fácilmente desde el grafo o a partir de la matriz **A:**

$$d_i = \sum_{\forall j} a_{ij} \qquad \mathbf{D} = \begin{bmatrix} 9 & 0 & 0 & 0 & 0 \\ 0 & 17 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 19 & 0 \\ 0 & 0 & 0 & 0 & 9 \end{bmatrix}$$

Y finalmente calculamos la matriz Laplaciana mediante la fórmula **L=D**-**A**.

$$\mathbf{L} = \mathbf{D} - \mathbf{A} = \begin{bmatrix} 9 & -5 & 0 & -4 & 0 \\ -5 & 17 & -2 & -7 & -3 \\ 0 & -2 & 4 & -2 & 0 \\ -4 & -7 & -2 & 19 & -6 \\ 0 & -3 & 0 & -6 & 9 \end{bmatrix}$$

Que coincide, lógicamente, con lo obtenido por Koren *et al* [315].

También podríamos haber obtenido la matriz Laplaciana obteniendo la matriz de incidencia, y concretamente la matriz de incidencia "normalizada" [329], utilizando la matriz **W** debido a que el grafo es ponderado para que $(\mathbf{QW}^{1/2})\ (\mathbf{QW}^{1/2})^{T}$ [341]:

$$\mathbf{Q} = \begin{bmatrix} +1 & +1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & +1 & +1 & +1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & +1 & 0 \\ 0 & -1 & 0 & -1 & 0 & -1 & +1 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 \end{bmatrix} \qquad \mathbf{W} = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 6 \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} +\sqrt{5} & +\sqrt{4} & 0 & 0 & 0 & 0 & 0 \\ -\sqrt{5} & 0 & +\sqrt{2} & +\sqrt{7} & +\sqrt{3} & 0 & 0 \\ 0 & 0 & -\sqrt{2} & 0 & 0 & +\sqrt{2} & 0 \\ 0 & -\sqrt{4} & 0 & -\sqrt{7} & 0 & -\sqrt{2} & +\sqrt{6} \\ 0 & 0 & 0 & 0 & -\sqrt{3} & 0 & -\sqrt{6} \end{bmatrix} \qquad \text{de donde} \qquad \mathbf{L} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}}$$

Siguiendo los pasos de Koren *et al* [315] obtenemos la descomposición en autovalores y autovectores de **L** con R:

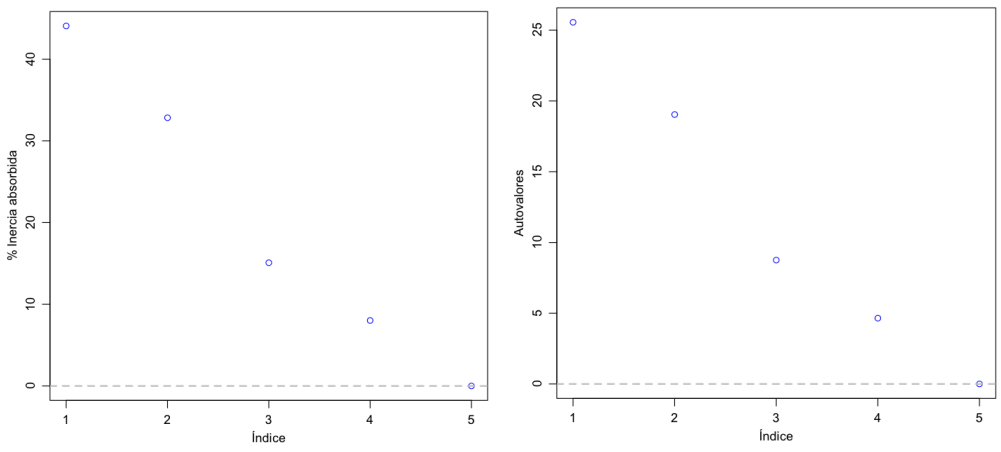
Autovalores de **L**: {25.5583, 19.0415, 8.7512, 4.6491, 0.0000}

Como cabía esperar el menor autovalor (cuadrado del valor singular) es nulo, ya que **L** es singular. El menor autovalor no nulo es 4.6491. Este autovalor sería también la conectividad algebraica o autovalor de Fiedler aunque la matriz sea ponderada [338].

También hemos obtenido de la descomposición los siguientes autovectores de **L**

$$\mathbf{V} = \begin{bmatrix} -0.0205 & 0.4777 & 0.6961 & 0.2947 & -0.4472 \\ -0.5713 & -0.6677 & 0.0968 & 0.1354 & -0.4472 \\ -0.0211 & 0.1373 & -0.0080 & -0.8835 & -0.4472 \\ 0.7988 & -0.3646 & -0.0777 & 0.1513 & -0.4472 \\ -0.1860 & 0.4173 & -0.7071 & 0.3021 & -0.4472 \end{bmatrix}$$

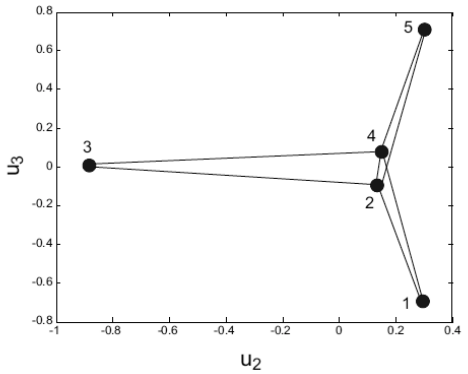
Podemos representar en un *scree-plot* los autovalores obtenidos y el porcentaje de variabilidad explicada.



Representación de la inercia absorbida y autovalores por eje factorial del grafo “torre Eiffel”.

La inercia absorbida/explicada en el primer plano factorial es del 77%, que en general consideraríamos un “buen resultado”.

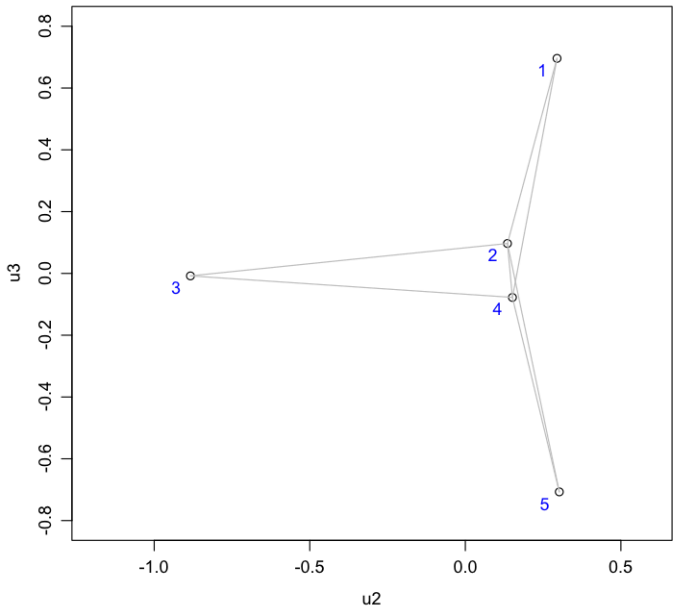
El método comentado por Koren *et al* [315] es la representación de los autovectores correspondientes a los MENORES autovalores, excluyendo aquel autovector correspondiente al autovalor nulo (esto es, utiliza el correspondiente a Fiedler y el anterior). El resultado que obtiene Koren *et al* [315] es el siguiente:



$$u_2 = \begin{pmatrix} 0.2947 \\ 0.1354 \\ -0.8835 \\ 0.1513 \\ 0.3021 \end{pmatrix} \qquad u_3 = \begin{pmatrix} -0.6961 \\ -0.0968 \\ 0.0080 \\ 0.0777 \\ 0.7071 \end{pmatrix}$$

Representación de ejemplo de la Torre Eiffel obtenida por solución directa del problema de autoproyección y vectores asociados a la solución. [315]

Si representamos los resultados de nuestra réplica del análisis de Koren *et al* [315] obtenemos:



$$u_2 = \begin{pmatrix} 0.2947 \\ 0.1354 \\ -0.8835 \\ 0.1513 \\ 0.3021 \end{pmatrix} \qquad u_3 = \begin{pmatrix} 0.6961 \\ 0.0968 \\ -0.0080 \\ -0.0777 \\ -0.7071 \end{pmatrix}$$

Resultado obtenido de la representación de los autovectores del grafo Eiffel.

Que coincide, salvo una reflexión especular en el eje x (cambios de signo en el vector u3, como podemos comprobar), con lo expuesto por Koren *et al* [315].

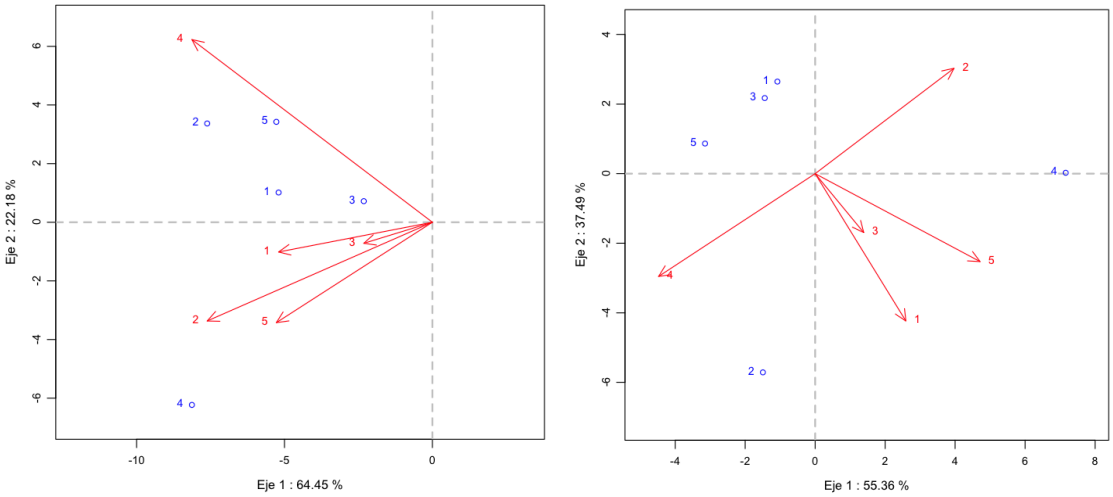
La aplicación del HJ-Biplot a este juego de datos presenta varias alternativas diferentes que exploraremos para analizar sus resultados.

En primer lugar podemos obtener un HJ-Biplot de la matriz de Adyacencia (con los pesos de las aristas incluidos) de los datos “Torre Eiffel”.

Recordemos que la matriz de adyacencia **A** del grafo Eiffel es

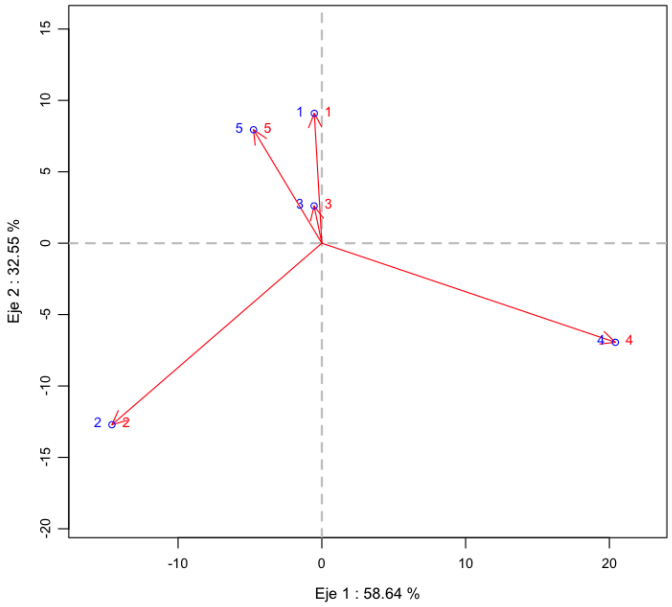
$$\mathbf{A} = \begin{bmatrix} 0 & 5 & 0 & 4 & 0 \\ 5 & 0 & 2 & 7 & 3 \\ 0 & 2 & 0 & 2 & 0 \\ 4 & 7 & 2 & 0 & 6 \\ 0 & 3 & 0 & 6 & 0 \end{bmatrix}$$

Obtendremos un HJ-Biplot sin normalizar y normalizando [271] la matriz de adyacencia **A.** La inercia explicada por el primer plano factorial es muy elevada en ambos casos (86,63% y 92.85%, respectivamente). En el primer caso, y dado que la matriz **A** es simétrica, los marcadores para las filas y las columnas son los mismos, salvo una reflexión sobre el eje x. En el segundo, la normalización (para hacer la media nula por columnas) rompe la simetría de la matriz sometida al HJ-Biplot y los marcadores fila **J** y columna **H** son diferentes. No obstante, en nuestra opinión, de cualquiera de ambas representaciones no se extrae demasiada información útil para este caso concreto.



Representaciones HJ-Biplot de la matriz **A** sin normalizar [izq.] y normalizada por columnas (media nula) [der.]

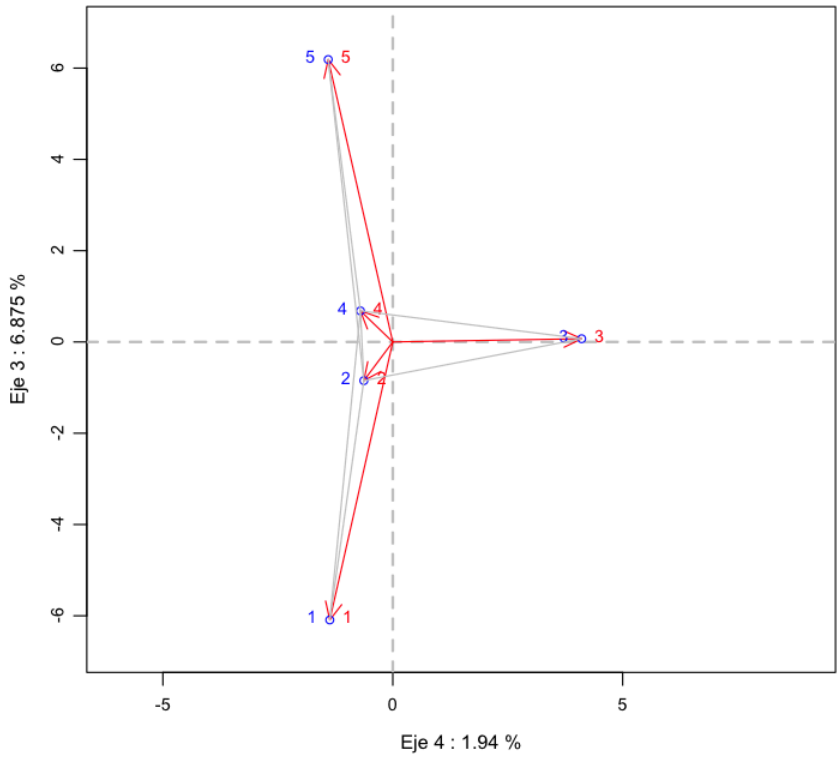
Efectuemos ahora el análisis HJ-Biplot para la matriz Laplaciana **L**. En este caso no tiene sentido la normalización (por columnas) de la matriz **L**, ya que las columnas (y las filas) de la Laplaciana suman cero. En un HJ-Biplot habitualmente representaremos los dos primeros ejes, que son los que acumulan la mayor parte de la inercia, y por lo tanto suponemos más representativos.



Representación HJ-Biplot de la matriz Laplaciana de *toy problem* Torre Eiffel

De nuevo, como en el caso anterior, al tratarse **L** de una matriz simétrica, los marcadores fila y columna coinciden. La varianza explicada en este primer plano factorial 1-2 es notable, alcanzando el 91%. La CRFE es igual para filas y columnas (obviamente, dada la simetría de **L**) y excepto para el nodo "3" obtiene valores totales de la CRFE para este primer plano factorial superiores a 680. Pero no observamos, a primera vista, nada "especial" en este primer plano factorial. Los nodos 1,3 y 5 forman un cluster frente a los nodos 2 y 4, pero es que efectivamente presenta un comportamiento bastante similar: los tres vértices 1,2 y 3 disponen de aristas (con diferentes pesos) con los nodos 2 y 4. El nodo 3 tiene la longitud menor y presenta la menor valencia (o grado). Los vértices 2 y 4 tienen los marcadores con mayor módulo y en el grafo se observa que tienen la mayor valencia.

Representemos ahora el último plano factorial 3-4 (excluyendo el último eje, 5º, con autovalor nulo, al ser **L** singular). La varianza explicada en este "penúltimo" plano factorial es muy pequeña (8,81%) y la representación que se obtiene es la siguiente:



Representación HJ-Biplot del segundo plano factorial de la matriz **L** de la Torre Eiffel

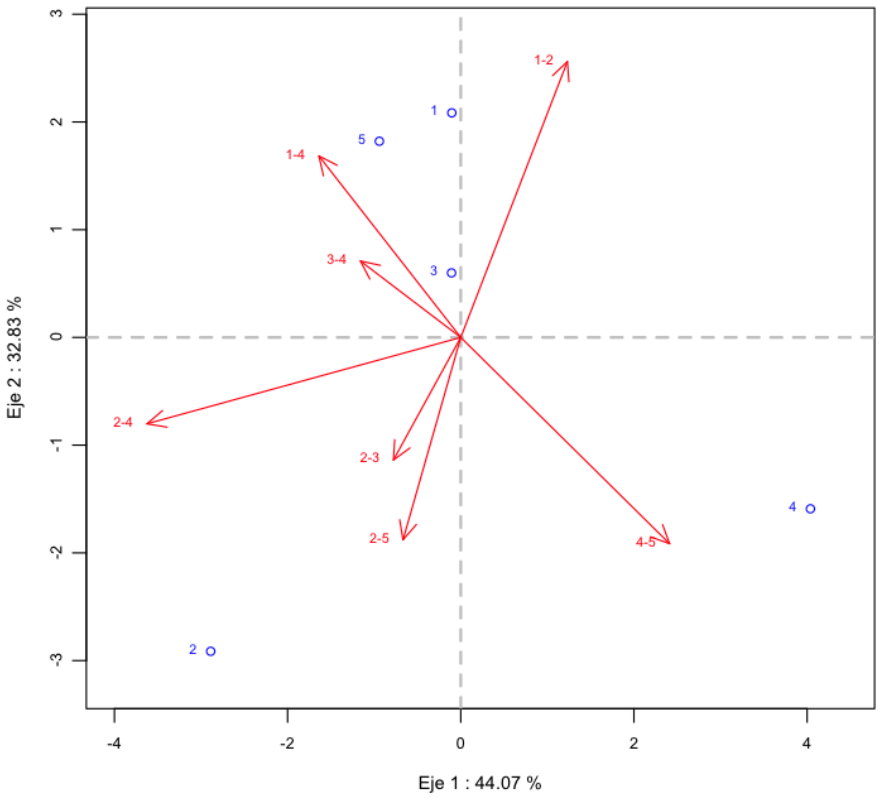
Si bien la representación de los dos primeros ejes factoriales NO se asemeja en demasía a la representación del grafo Eiffel inicial, sin embargo, el plano 3-4 del HJ-Biplot sí que se asemeja más al grafo Eiffel inicial. De hecho se trata de la misma representación obtenida por Koren *et al* [315], escalada en cada eje por los respectivos valores singulares. La calidad de representación CRFE en este plano factorial puede considerarse "aceptable" para tres nodos, y "baja" para otros dos.

Nodo	CRFE eje 3	CRFE eje 4	CRFE ejes 3 y 4
3	0.2	703.0	703.2
5	303.9	15.6	319.6
1	304.1	15.4	319.5
2	1.9	1.1	3
4	1	1.1	2.1

CRFE HJ-Biplot último plano factorial

Hasta aquí realmente la propuesta expuesta que realizamos ofrece pocas novedades con relación a la existente (descomposición en autovalores/autovectores), como podemos observar, salvo la novedosa aplicación del HJ-Biplot en este escenario. Realmente el hecho de operar sobre una matriz simétrica, como **L,** hace que los métodos Biplot en general, y el HJ-Biplot en particular, no ofrezcan mucha más información que la ya disponible a partir de la matriz Laplaciana con otros análisis ya aplicados, en cuanto a elementos fila y columna.

Obtengamos el HJ-Biplot de esa matriz de incidencia vértice-arista **Q** para grafos ponderados, "generalizada" o "normalizada" por la matriz $\mathbf{W}^{0.5}$ de pesos de las aristas ($\mathbf{QW}^{0.5}$) para el primer plano factorial. La inercia/variabilidad absorbida en este primer plano factorial asciende al 77%.



Representación HJ-Biplot de la matriz de incidencia del grafo Eiffel

La representación de las filas es estructuralmente idéntica a la obtenida para las filas y las columnas de la aplicación del HJ-Biplot a la matriz Laplaciana, antes vista, precisamente por la manera de obtener la matriz Laplaciana a partir de la matriz de incidencia y la forma de obtener la DVS a partir de la que se obtiene el HJ-Biplot. Con ello es válido todo lo expuesto anteriormente relativo a la relación entre los módulos de los marcadores **J** correspondientes a los vértices y los grados/valencias respectivos .

Pero además de la coincidencia estructural del resultado obtenido para las filas, en este caso obtenemos la situación de las aristas del grafo (columnas) en el mismo sistema de representación que el de las filas, algo que hasta ahora no se obtenía. Esto constituye una novedad, y habrá que analizar la información que aporta a la representación.

Como podemos comprobar la inercia absorbida en este primer plano es elevada (76.9%). Además ahora disponemos de CRFE para las filas y las columnas:

Vértice	CRFE eje 1	CRFE eje 2	CRFE eje 1 + eje 2
4	858.4	133.2	991.6
2	490.7	499.4	990.2
1	1.2	482.8	484.0
5	98.2	368.5	466.7
3	2.8	89.7	92.5

Arista	CRFE eje 1	CRFE eje 2	CRFE eje 1 + eje 2
2-4	938.7	46.0	984.6
1-2	151.7	656.0	807.8
4-5	484.9	305.7	790.6
1-4	335.6	354.7	690.3
2-5	74.3	588.7	662.9
2-3	151.4	324.0	475.4
3-4	336.2	125.9	462.1

Excepto en el caso del vértice “3”, las calidades de representación para este plano son buenas o razonablemente aceptables. Pero, ¿qué información aporta la inclusión de las aristas en la representación conjunta?

En primer lugar la orientación de los marcadores asociados a las aristas (**H**) está aparentemente vinculada a la orientación (real o adoptada) de la arista: si nos fijamos están orientadas hacia el nodo origen, que se corresponde con el valor positivo en la matriz **Q**, y desde el nodo de destino, valor negativo en la matriz **Q**: por ejemplo, la arista 1-2, que ha sido definida desde el nodo 1 hacia el nodo 2, tiene un marcador columna orientado desde 2 hacia 1. Y así en todos los casos, al menos en este *toy problem*.

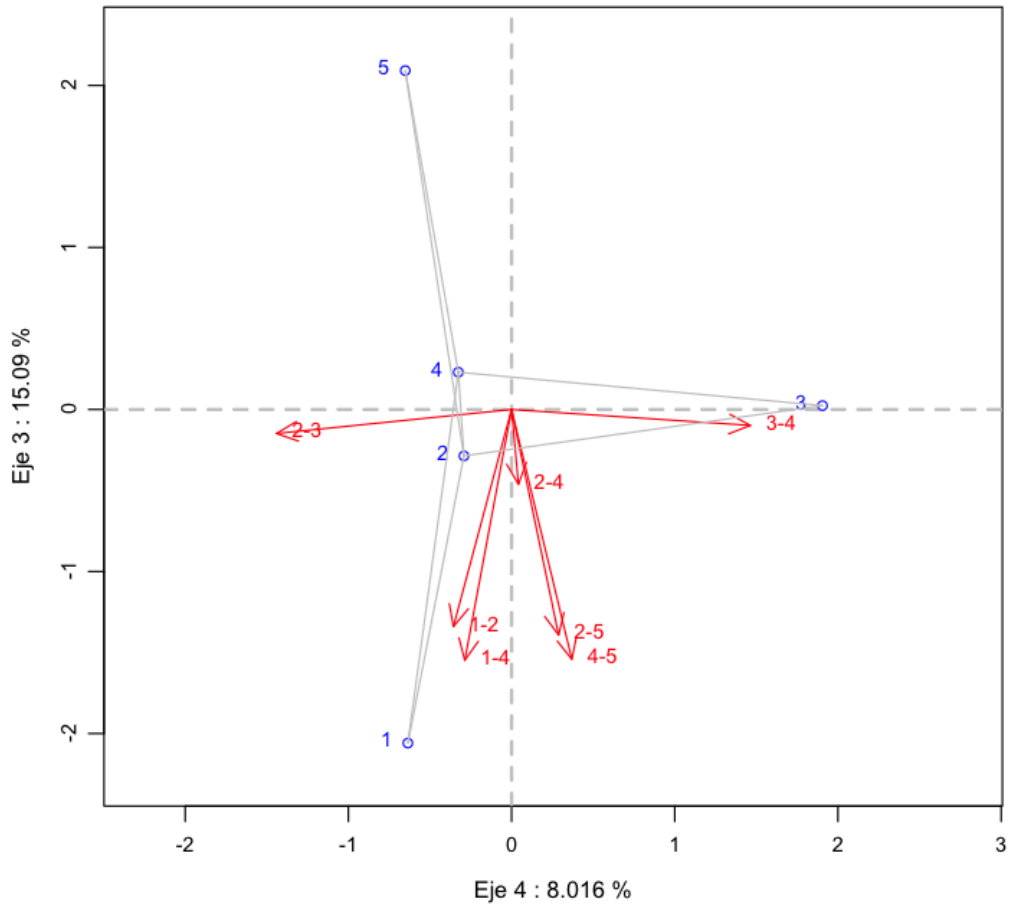
Consideremos las ponderaciones de las aristas del grafo y las longitudes de los marcadores columna **H** en este primer plano factorial,

Arista	Ponderación	Longitud
2-4	7	3.71
4-5	6	3.08
1-2	5	2.84
1-4	4	2.35
2-5	3	1.99
2-3	2	1.38
3-4	2	1.36

Como se puede comprobar, al menos en este caso, la longitud de los marcadores columna **H** en este primer plano factorial presenta la misma ordenación que las ponderaciones asignadas a esas aristas. De hecho la ponderación de cada arista puede obtener de manera exacta como:

Diagonal(**W**) = (1/2)diagonal($\mathbf{HH}^T$) Recordemos, una vez más, que de las propiedades de los Biplot en general y del HJ-Biplot, en particular se tiene que $\mathbf{L=JJ}^T\mathbf{=QQ}^T$ y $\mathbf{K=HH}^T\mathbf{=Q}^T\mathbf{Q}$.

Trabajemos ahora con el último plano factorial del HJ-Biplot. En primer lugar la inercia absorbida en este plano es sensiblemente superior a la obtenida en el último plano factorial del HJ-Biplot de **L**, antes teníamos un 8.81% y ahora para el último plano factorial del HJ-Biplot de **Q** tenemos un 23.1%. La representación gráfica de los marcadores **J** correspondientes a los nodos/vértices sigue "pareciéndose" mucho al grafo Eiffel de partida, de hecho de nuevo se trata de la misma representación mostrada por Koren *et al* [315] con la ponderación de los valores singulares. Y los marcadores **H** de las columnas, que se corresponden con las aristas, siguen con la orientación dada a las aristas, no obstante en este caso hemos perdido la ordenación de las longitudes de las aristas en función de los pesos de las mismas.



Representación HJ-Biplot, ejes factoriales 3 y 4 de la matriz de incidencia del grafo Eiffel

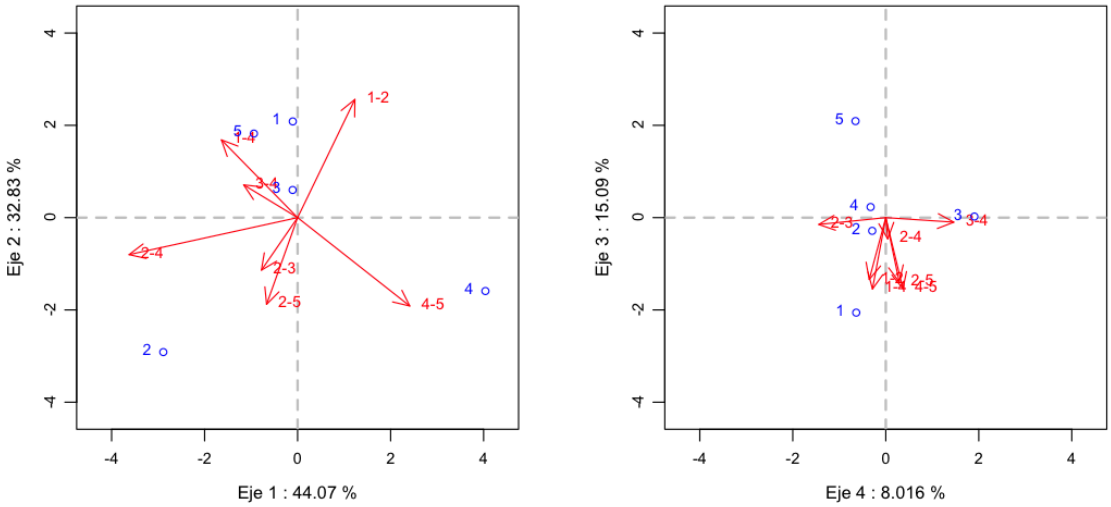
Obtengamos la CRFE para las filas y columnas en este último plano factorial:

Vértice	CRFE eje 3	CRFE eje 4	CRFE eje 3 + eje 4
3	0.1	907.3	907.5
5	486.2	47.1	533.3
1	471.1	44.9	516
2	4.8	5.0	9.8
4	2.8	5.6	8.4

Arista	CRFE eje 3	CRFE eje 4	CRFE eje 3 + eje 4
3-4	2.4	535.5	537.9
2-3	5.5	519.1	524.6
2-5	323.1	13.9	337.1
1-4	299.4	10.3	309.7
4-5	198.1	11.4	209.4
1-2	179.5	12.7	192.2
2-4	15.2	0.1	15.4

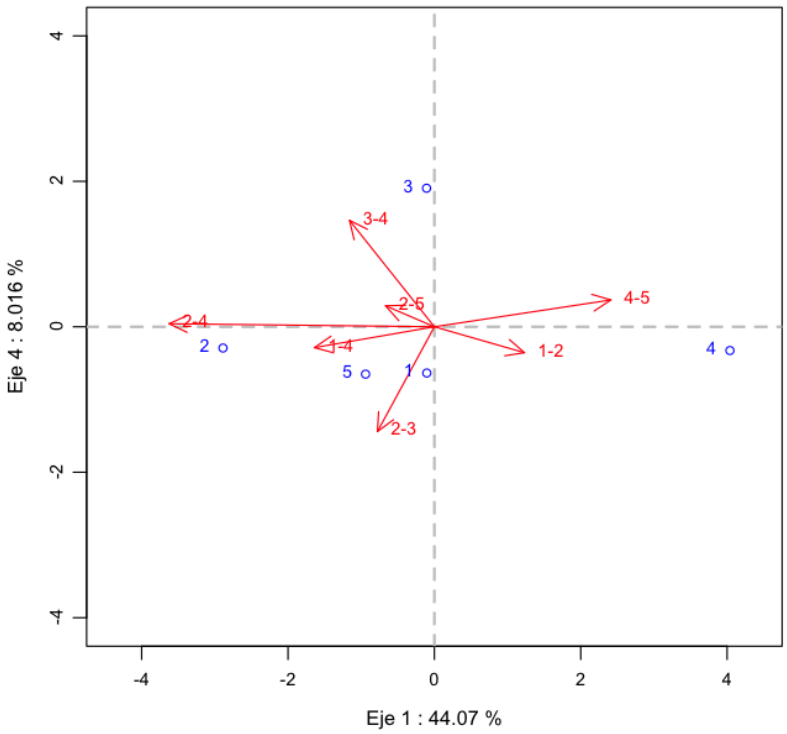
Para los vértices, los “2” y “4” obtienen una CRFE bastante baja, mientras que para las aristas la “2-4” (precisamente) obtiene una CRFE bastante baja y las “4-5” y “1-2” es baja. Solo en el caso del vértices “3” se puede considerar el resultado de CRFE alto.

Representación de los 4 ejes factorial con autovalores no nulos.



Representación HJ-Biplot. 4 ejes factoriales correspondientes a autovalores no nulos de la matriz **Q**’

Representemos los ejes teóricamente más “representativos”, el correspondiente al mayor valor propio (con información sobre “centralidad vectorial” o de Bonacich) y el del menor valor propio no nulo (el de conectividad algebraica y valoración característica o de Fiedler). Esta representación encerrará toda la información que el análisis espectral de grafos obtienen del modelo de la red de comunicación.



HJ-Biplot "Torre Eiffel" ejes 1 y 4.

7.7.3. HJ-BIPLOT DE LA MATRIZ DE ENRUTAMIENTO DE LA RED ABILENE

7.7.3.1. Planteamientos preliminares

Tenemos la matriz de enrutamiento de la red Abilene, obtenida del fichero SAND.mat disponible, entre otros lugares, en la página web del autor [379]. La matriz incluida en el fichero original (data$A) contiene 30 filas y 121 columnas. Las filas corresponden con los 15 enlaces de la red Abilene, considerando cada sentido del enlace, y los 121 pares origen-destino, correspondientes a los 11 nodos (11 x 11 =121) de la red Abilene.

Mapa de la red Abilene [348], [349]

Como en la matriz de enrutamiento están también las columnas correspondientes a la diagonal de la matriz Origen-Destino (OD), que se correspondería en el grafo con lazos, son columnas con ceros en todos sus elementos y los eliminamos de la matriz de enrutamiento **B**.

Así obtenemos la matriz de enrutamiento **B** de 30 x 110 (30 enlaces dirigidos en sus filas, y 110 rutas OD en sus columnas), sobre la que trabajaremos y que se corresponde con la utilizada en los trabajos de Chua *et al* [291], [348], [349] y Kolaczyk [350].

	ATLA-CHIN	ATLA-DNVR	ATLA-HSTN	ATLA-IPLS	ATLA-KSCY	ATLA-LOSA	ATLA-NYCM	…
atla-hstn	0	0	1	0	0	1	0	…
atla-ipls	1	1	0	1	1	0	0	…
atla-wash	0	0	0	0	0	0	1	…
chin-ipls	0	0	0	0	0	0	0	…
chin-nycm	0	0	0	0	0	0	0	…
dnvr-kscy	0	0	0	0	0	0	0	…
dnvr-snva	0	0	0	0	0	0	0	…
dnvr-sttl	0	0	0	0	0	0	0	…
hstn-atla	0	0	0	0	0	0	0	…
hstn-kscy	0	0	0	0	0	0	0	…
hstn-losa	0	0	0	0	0	1	0	…
ipls-atla	0	0	0	0	0	0	0	…
ipls-chin	1	0	0	0	0	0	0	…
ipls-kscy	0	1	0	0	1	0	0	
kscy-dnvr	0	1	0	0	0	0	0	…
kscy-hstn	0	0	0	0	0	0	0	…
kscy-ipls	0	0	0	0	0	0	0	…
kscy-snva	0	0	0	0	0	0	0	…
losa-hstn	0	0	0	0	0	0	0	…
losa-snva	0	0	0	0	0	0	0	…
nycm-chin	0	0	0	0	0	0	0	…
nycm-wash	0	0	0	0	0	0	0	…
snva-dnvr	0	0	0	0	0	0	0	…
snva-kscy	0	0	0	0	0	0	0	…
snva-losa	0	0	0	0	0	0	0	…
snva-sttl	0	0	0	0	0	0	0	…
sttl-dnvr	0	0	0	0	0	0	0	…
sttl-snva	0	0	0	0	0	0	0	…
wash-atla	0	0	0	0	0	0	0	…
wash-nycm	0	0	0	0	0	0	1	…

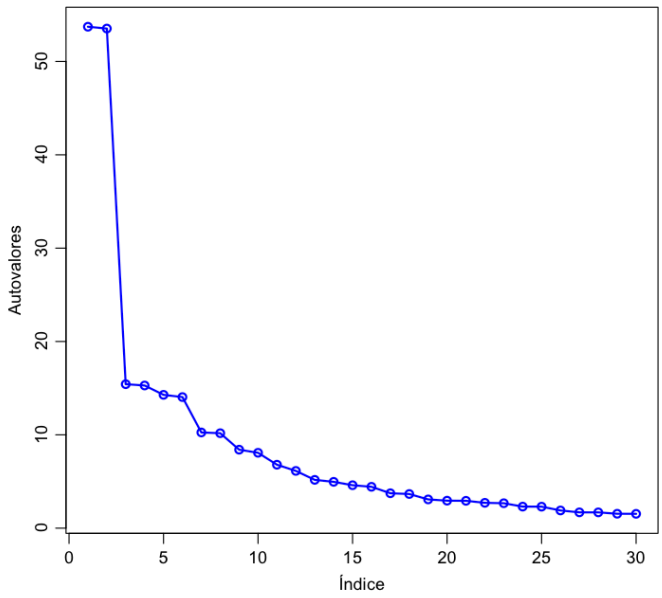
Primeras 7 columnas (de las 110 totales) de la matriz de enrutamiento de Abilene.

En filas están los enlaces unidireccionales y en columnas las rutas unidireccionales. Es fácil comprobar las rutas compuestas por un único enlace (por ejemplo Atlanta-

Houston y Atlanta-Indianápolis en la parte de la matriz mostrada anteriormente) ya que solo utilizan el enlace obvio entre el origen y destino establecido. Este hecho es el que hemos aprovechado para comprobar que la información colgada por Kolaczyk en su página web [379] es parcialmente incorrecta: el orden real de las filas de la matriz de enrutamiento de Abilene que proporciona el autor no es el que tiene el vector de etiquetas que igualmente pone a disposición de la comunidad ("*The ordering of pairs of names in `odnames' and `edgenames' corresponds to the orderings in X and A*"), situación que hemos corregido en nuestro análisis.

Obtenemos la descomposición en valores singulares de la matriz $\mathbf{B}^T$. Operaremos con $\mathbf{B}^T$ por seguir los pasos del autor [379] pero evidentemente no habría inconveniente en trabajar directamente sobre la matriz **B.**

Representamos los autovalores de $\mathbf{B}\mathbf{B}^T$ o $\mathbf{B}^T\mathbf{B}$ que son los cuadrados de los valores singulares de $\mathbf{B}^T$ obtenidos en la DVS.



Autovalores de la matriz de enrutamiento de la red Abilene

El *screeplot,* o espectro de autovalores, obtenido de la matriz de enrutamiento de Abilene coincide con los resultados publicados por Kolaczyk [350] y Chua [291], [348], [349]. Tal y como indican los autores, el salto entre el segundo y tercer autovalor, y el codo así formado en la representación son indicativos de la relación existente entre las filas de la matriz $\mathbf{B}^T$ y permite aventurar que mediciones que se realicen en entre 5 y 10 rutas, y quizá solo de 2, podrían ser suficientes para obtener información sobre todas las rutas posibles en la red Abilene. Recordemos que

$$||\mathbf{B}^T - \mathbf{B}^T_{(K)}|| = \lambda_{K+1} + \ldots + \lambda_{rank(\mathbf{B})}$$

Esto es, la suma de los autovalores nos permite obtener la diferencia de valores cuadráticos entre las matrices de enrutamiento reales y aproximada [350]. La asociación entre los dos primeros autovalores también es habitual en esta aplicación [350].

Recordemos que Chua *et al* [291] asocian además esta caída del autovalor de la matriz de enrutamiento con la caída de la distribución de los valores de "intermediación" (*betweenness*) conectando estos planteamientos con el concepto de centralidad.

Consideremos ahora los autovectores **V** de la matriz de enrutamiento $\mathbf{B}^T$ obtenidos en la DVS. En este escenario, los autovectores puede ser vistos como "meta-rutas" linealmente independientes, en el espacio de los enlaces. Estas "meta-rutas" disponen de un orden dado por sus correspondientes autovalores, que esencialmente indicada para cada una de esas "meta-rutas" la proporción relativa de rutas en la red que representan. Consideremos los primeros 4 autovectores, ordenados por el valor absoluto de su correspondiente componente (sin repetir los enlaces bidireccionales que obtienen el mismo resultado en módulo, el signo cambia):

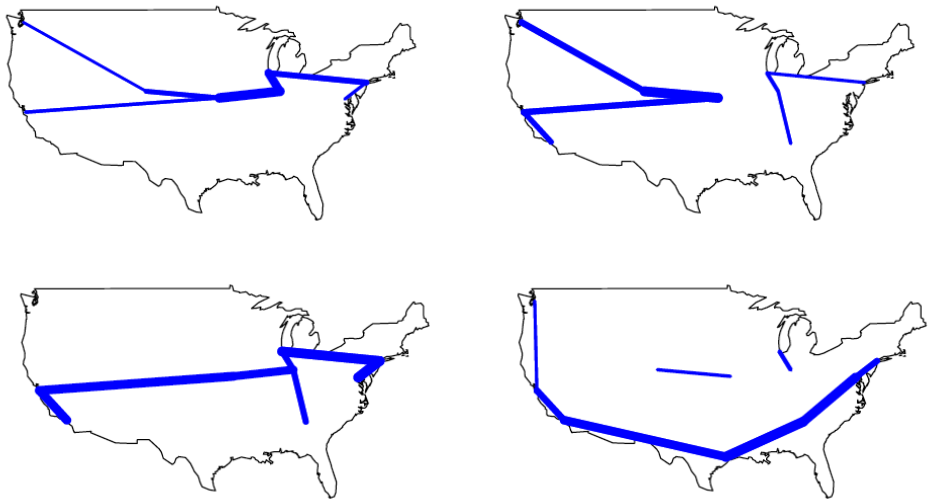
Primera componente	
kscy-ipls	0.442
ipls-chin	0.365
chin-nycm	0.255
dnvr-kscy	0.214
kscy-snva	0.150
wash-nycm	0.130
dnvr-sttl	0.113
snva-losa	0.065
atla-ipls	0.059
kscy-hstn	0.029
atla-wash	0.006
hstn-atla	0.003
hstn-losa	0.003
sttl-snva	0.003
snva-dnvr	0.001

Segunda componente	
ipls-kscy	0.443
chin-ipls	0.365
nycm-chin	0.256
kscy-dnvr	0.213
snva-kscy	0.151
nycm-wash	0.131
sttl-dnvr	0.113
losa-snva	0.066
ipls-atla	0.044
hstn-kscy	0.021
wash-atla	0.006
atla-hstn	0.003
losa-hstn	0.003
snva-sttl	0.003
snva-dnvr	0.001

Tercera componente	
dnvr-kscy	0.440
snva-kscy	0.279
sttl-dnvr	0.276
losa-snva	0.244
chin-ipls	0.162
chin-nycm	0.146
atla-ipls	0.145
nycm-wash	0.089
kscy-ipls	0.082
losa-hstn	0.060
sttl-snva	0.044
hstn-kscy	0.038
atla-wash	0.031
hstn-atla	0.029
dnvr-snva	0.018

Cuarta componente	
kscy-dnvr	0.434
kscy-snva	0.317
snva-losa	0.278
dnvr-sttl	0.273
ipls-chin	0.126
chin-nycm	0.125
ipls-atla	0.098
nycm-wash	0.070
hstn-losa	0.067
kscy-ipls	0.052
snva-sttl	0.051
kscy-hstn	0.036
atla-wash	0.028
atla-hstn	0.028
snva-dnvr	0.021

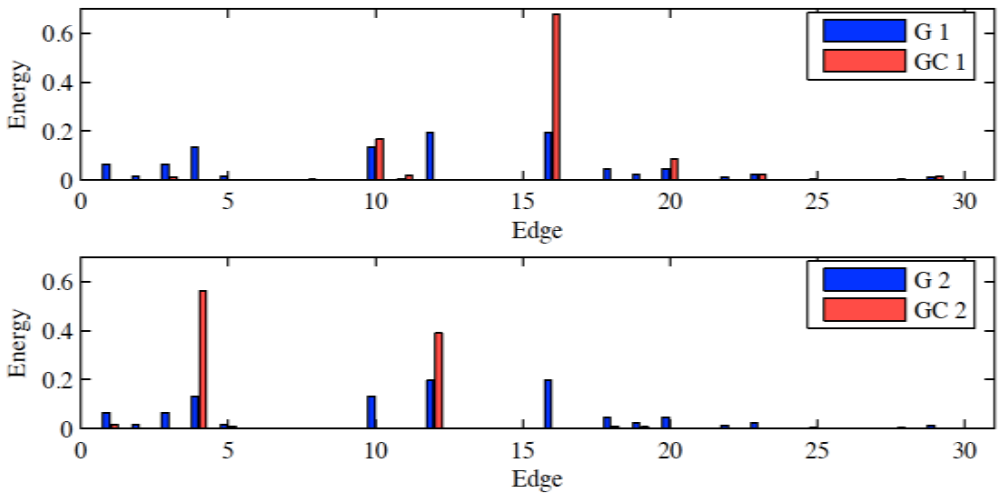
Se pueden representar sobre el mapa de la red Abilene los autovectores obtenidos [291], [348], [350]:



Representación geoposicionada de los primeros autovectores de la matriz de enrutamiento de Abilene [291], [348], [350].

El primer autovector representa una “meta-ruta” este-oeste en la parte norte de la red. Los restantes segundo y tercer autovectores inciden en esta ruta, mientras que el cuarto representa otra “meta-ruta” este-oeste, pero a través de la parte sur de la red Abilene. Esta aplicación esta, a nuestro entender, claramente en consonancia con las teorías de centralidad del autovector o de Bonacich [374]–[376], [380].

Chua *et al* [349] proporcionan otra representación de los autovectores equiparándola al concepto de “energía” subyacente en la representación espectral (de señales) y calculada como el cuadrado del valor de cada componente del autovector.



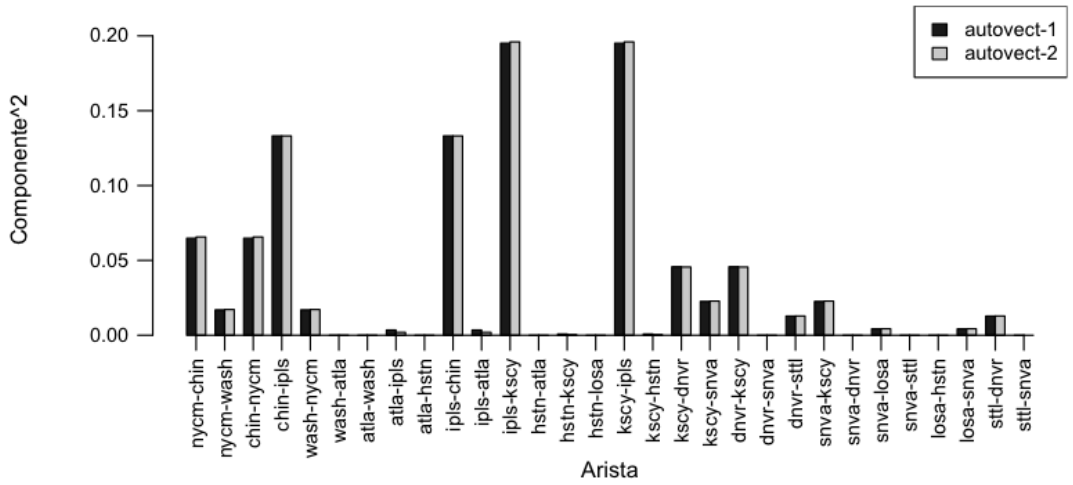
Comparación de la distribución de energía de los primeros dos autovectores de la matriz de enrutamiento y de la matriz de enrutamiento ponderada por la varianza del enlace (arriba primeros autovectores, abajo segundos autovectores) [349]

El autor proporciona en el anexo la lista de bordes (enlaces, "*edges*") utilizada para poder asociar cada columna con su correspondiente enlace en la red Abilene. Así, tal y como se ha indicado, en el caso del primer autovalor de la matriz de enrutamiento (en gris oscuro en la gráfica superior) los enlaces con mayor concentración de "energía" serían, por este orden, los números 12, 16, 10, 4, 1, 3, 18, 20 correspondientes con los enlaces

Nº Componente del autovector	Identificación enlace red (Origen-Destino)	Identificación abreviada del enlace
12	Indianápolis - Kansas City	ipls-kscy
16	Kansas City - Indianápolis	kscy-ipls
10	Indianápolis - Chicago	ipls-chin
4	Chicago - Indianápolis	chin-ipls
1	Nueva York - Chicago	nycm-chin
3	Chicago - Nueva York	chin-nycm
18	Kansas City - Denver	kscy-dnvr
20	Denver - Kansas City	dnvr-kscy

Relación las componentes de mayor valor del primer autovector
de la matriz de enrutamiento de Abilene (extraída de Fig. 3 de [349]).

En el caso del segundo autovector de la matriz de enrutamiento (en gris oscuro en la gráfica) los enlaces que acumulan la mayor concentración de "energía" serían los mismos. La representación que obtenemos de los dos primeros autovectores de la matriz de enrutamiento de Abilene coincide fielmente con lo expuesto por Chua *et al* [349].



Representación obtenida correspondiente a los dos primeros autovectores de la matriz de enrutamiento de Abilene.

Así pues, una vez que somos capaces de reproducir el análisis efectuado sobre la matriz de enrutamiento de Abilene realizado por Chua *et al* [291], [348], [350] mediante la Descomposición en Valores Singulares, obteniendo información sobre su estructura a partir del espectro de autovalores y de su funcionamiento estudiando los vectores propios, apliquemos el HJ-Biplot para comparar los resultados y analizar posibles coincidencias y mejoras.

7.7.3.2. ANÁLISIS HJ-BIPLOT DE LA MATRIZ DE ENRUTAMIENTO

Vamos a aplicar el análisis HJ-Biplot a la matriz de enrutamiento. Recordemos que está formada en sus filas por los enlaces de la red (N_e), y sus columnas las parejas origen-destino (N_p), de forma que, si solo hay una ruta para ir de *i* a *j* se tiene que:

$$\mathbf{B}_{e;ij} = \begin{cases} 1, & \text{si el enlace } e \text{ es atravesado para ir de } i \text{ a } j \\ 0, & \text{en otro caso} \end{cases}$$

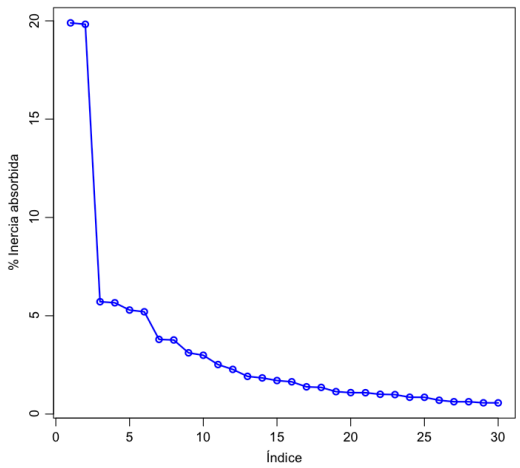
En el supuesto de que más de un enlace pueda ser utilizado de forma simultánea para una misma ruta, el valor $\mathbf{B}_{e;ij}$ será el porcentaje en tanto por uno del tráfico cursado por cada uno de esos enlaces.

En primer lugar no se normalizará [271] la matriz de enrutamiento antes de ser sometida al análisis HJ-Biplot. Las características de las variables intervinientes son homogéneas y la principal cuestión que se debe destacar es que, por lo general, se tratará de matrices poco pobladas, por lo que la normalización puede ayudar a obtener un resultado más estable. No obstante, este aspecto será reconsiderado más adelante.

Sin normalización de la matriz de enrutamiento

A continuación procederemos a aplicar el análisis HJ-Biplot [160] a la matriz de enrutamiento en bruto, esto es, sin normalizar. Obtendremos unos marcadores **H** para las columnas (rutas) y unos marcadores **J** para los enlaces (rutas).

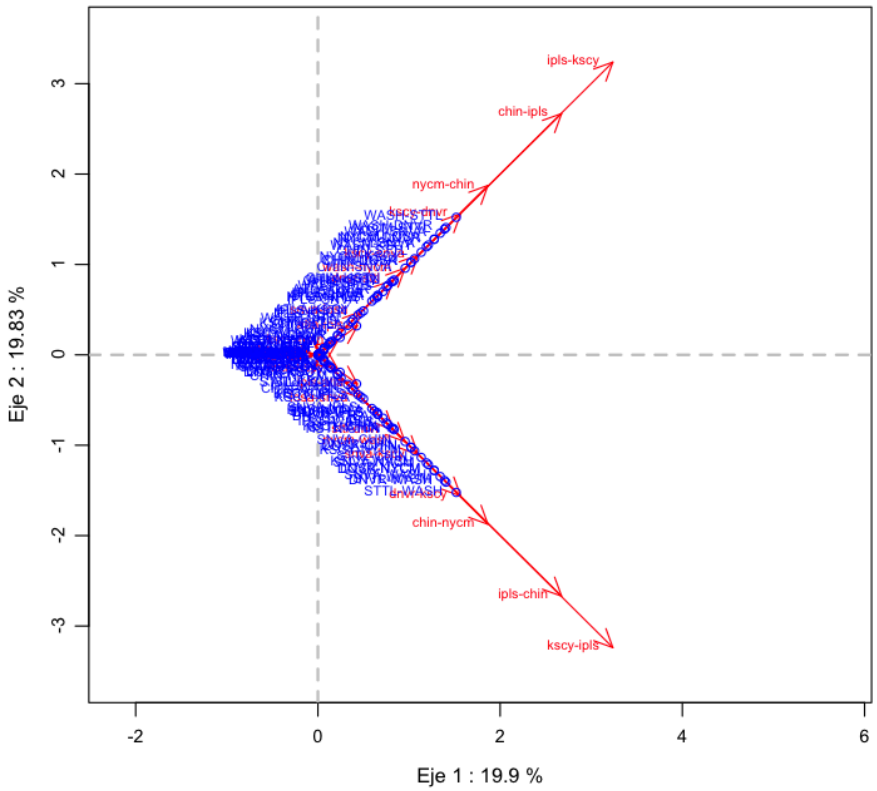
La varianza acumulada en dos primeros ejes (primer plano) alcanza el 39,7%. Representemos de nuevo el *screeplot* o espectro de los autovalores obtenidos, esta vez en forma de inercia absorbida



Representación espectral obtenida en HJ-Biplot sobre matriz enrutamiento Abilene.

Como vimos esto coincide con los resultados publicados por Kolaczyk [350] y Chua [291], [348], [349] incluyendo el emparejamiento conocido de los dos primeros autovalores. Este efecto se debe a la simetría de la matriz de enrutamiento [349].

Representemos, inicialmente, los resultados obtenidos por el HJ-Biplot.



Representación 1er plano factorial HJ-Biplot de la matriz enrutamiento de Abilene (sin normalizar)

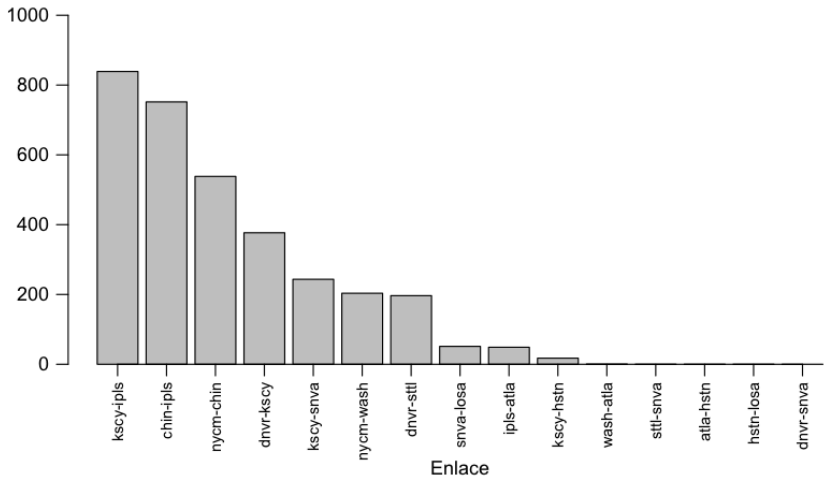
Resalta la alineación tanto de rutas como enlaces a lo largo de dos direcciones casi ortogonales (ángulo 1.574102 radianes) en el primer y cuarto cuadrante de este primer plano factorial. Por los vectores que se apoyan en cada dirección, claramente esta disposición tiene que ver con el sentido de las rutas/enlaces. En cada dirección se sitúan las rutas/enlaces en un sentido.

La longitud de los vectores correspondientes a los enlaces de la red (etiquetas en minúsculas) se corresponde con la información obtenida en los trabajos anteriores de Kolaczyk [350] y Chua [291], [348], [349], que justificaban por la mayor utilización de dichos enlaces y de nuevo surge la idea del concepto de “centralidad”.

Analicemos la CRFE de los enlaces (columnas, marcadores **H**) para este primer plano factorial y como anteriormente, solo se muestra uno de los sentidos ya que el resultado para el sentido inverso es el mismo:

Enlace / CRFE	Eje 1	Eje 2	Acumulada
kscy-ipls	419,27	419,64	838,91
chin-ipls	376,62	375,12	751,74
chin-nycm	267,99	270,33	538,32
dnvr-kscy	189,17	187,63	376,79
kscy-snva	121,43	122,07	243,50
nycm-wash	101,28	102,25	203,53
sttl-dnvr	98,23	98,35	196,58
snva-losa	25,51	25,77	51,28
ipls-atla	31,34	17,37	48,71
kscy-hstn	11,62	5,86	17,48
atla-wash	0,34	0,34	0,68
sttl-snva	0,12	0,13	0,25
atla-hstn	0,11	0,11	0,21
losa-hstn	0,10	0,10	0,19
snva-dnvr	0,04	0,04	0,09

Comprobamos que las CRFE acumuladas se encuentran emparejadas entre los enlaces unidireccionales que se corresponden con el mismo enlace bidireccional. Representemos estos valores para intentar establecer un "punto de corte" en la calidad de representación.



Representación de la CRFE acumulada de los ejes 1 y 2 para enlaces bidireccionales de la Red Abilene (HJ-Biplot).

Vista la gráfica podemos establecer dos posibles "puntos de corte", uno relativamente exigente, en 400 (incluyendo 4 enlaces bidireccionales, hasta Denver-Kansas City) y otro más laxo en 200, aproximadamente (incluyendo 7 enlaces bidireccionales, hasta Seattle-Denver). Con el punto de corte "habitual" en 600 solo tendríamos dos enlaces.

Analicemos la CRFE de las rutas también para este primer plano factorial, de nuevo al encontrarse emparejadas por ruta, sin que el sentido tenga efecto, mostramos solo uno de los sentidos:

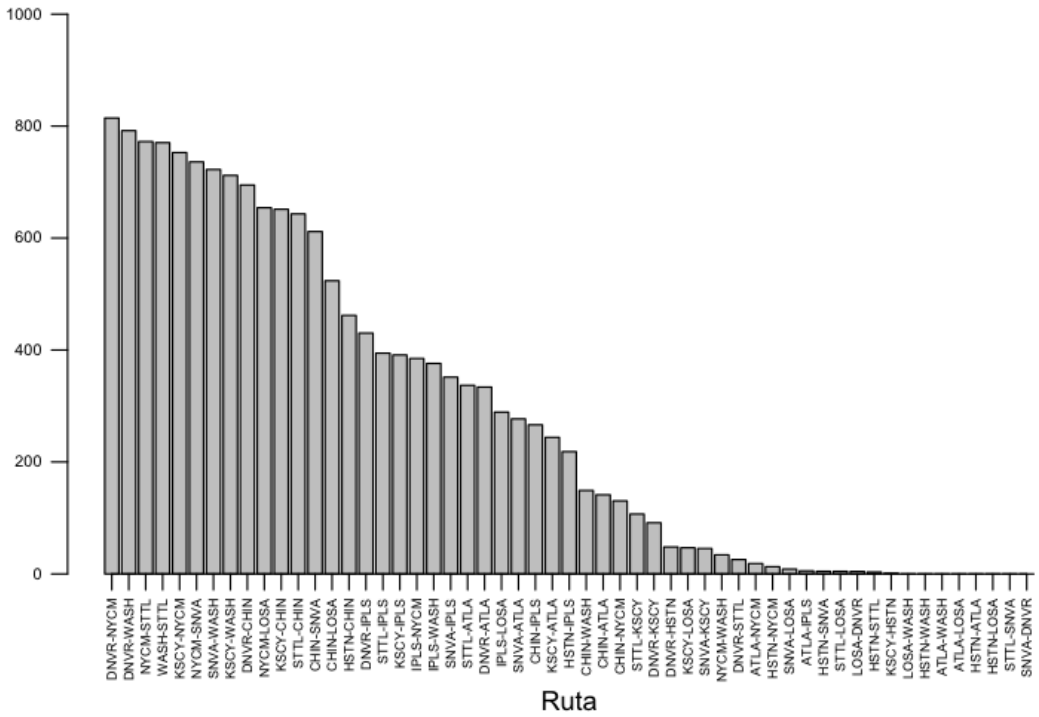
Ruta / CRFE	Eje 1	Eje 2	Acumulada
DNVR-NYCM	406,59	407,88	814,46
DNVR-WASH	395,11	396,73	791,84
STTL-NYCM	385,54	386,82	772,36
STTL-WASH	384,40	386,00	770,39
KSCY-NYCM	375,47	377,25	752,72
NYCM-SNVA	367,04	368,97	736,00
WASH-SNVA	360,16	362,33	722,48
WASH-KSCY	354,97	356,98	711,95
DNVR-CHIN	347,23	347,53	694,76
NYCM-LOSA	326,17	328,02	654,19
KSCY-CHIN	325,36	326,11	651,48
STTL-CHIN	321,36	321,76	643,12
CHIN-SNVA	305,30	306,32	611,61
CHIN-LOSA	261,32	262,37	523,69
HSTN-CHIN	233,01	228,83	461,84
DNVR-IPLS	214,95	215,28	430,23
STTL-IPLS	197,02	197,42	394,44
KSCY-IPLS	195,11	195,99	391,10
NYCM-IPLS	191,96	192,90	384,86
WASH-IPLS	187,44	188,63	376,06
SNVA-IPLS	175,27	176,25	351,52
STTL-ATLA	171,38	165,53	336,91
DNVR-ATLA	170,32	163,47	333,80
IPLS-LOSA	144,07	145,00	289,07
SNVA-ATLA	141,36	135,62	276,98
CHIN-IPLS	133,20	133,15	266,35
KSCY-ATLA	125,44	118,50	243,94
HSTN-IPLS	110,98	107,48	218,46

Ruta / CRFE	Eje 1	Eje 2	Acumulada
IPLS-HSTN	110,98	107,48	218,46
CHIN-WASH	74,08	75,02	149,10
ATLA-CHIN	89,94	51,45	141,39
CHIN-NYCM	64,85	65,65	130,50
KSCY-STTL	53,49	53,42	106,92
KSCY-DNVR	45,78	45,57	91,34
HSTN-DNVR	29,61	18,53	48,15
LOSA-KSCY	23,27	23,51	46,78
SNVA-KSCY	22,60	22,81	45,41
NYCM-WASH	16,97	17,19	34,16
DNVR-STTL	12,80	12,86	25,66
NYCM-ATLA	9,23	9,35	18,58
NYCM-HSTN	6,44	6,53	12,97
LOSA-SNVA	4,27	4,33	8,61
ATLA-IPLS	3,50	1,95	5,45
HSTN-SNVA	2,34	2,37	4,71
LOSA-STTL	2,31	2,35	4,66
LOSA-DNVR	2,22	2,25	4,47
HSTN-STTL	1,68	1,70	3,38
HSTN-KSCY	0,86	0,44	1,30
LOSA-WASH	0,05	0,05	0,09
HSTN-WASH	0,04	0,04	0,08
ATLA-WASH	0,03	0,03	0,06
LOSA-ATLA	0,02	0,02	0,04
ATLA-HSTN	0,01	0,01	0,02
LOSA-HSTN	0,01	0,01	0,02
SNVA-STTL	0,01	0,01	0,01
SNVA-DNVR	0,00	0,00	0,00

Podemos colocar esta información en forma matriz cuadrada

ORI DST	ATLA	CHIN	DNVR	HSTN	IPLS	KSCY	LOSA	NYCM	SNVA	STTL	WASH
ATLA		141,39	333,80	0,02	5,45	243,94	0,04	18,58	276,98	336,91	0,06
CHIN	141,39		694,76	461,84	266,35	651,48	523,69	130,50	611,61	643,12	149,10
DNVR	333,80	694,76		48,15	430,23	91,34	4,47	814,46	0,00	25,66	791,84
HSTN	0,02	461,84	48,15		218,46	1,30	0,02	12,97	4,71	3,38	0,08
IPLS	5,45	266,35	430,23	218,46		391,10	289,07	384,86	351,52	394,44	376,06
KSCY	243,94	651,48	91,34	1,30	391,10		46,78	752,72	45,41	106,92	711,95
LOSA	0,04	523,69	4,47	0,02	289,07	46,78		654,19	8,61	4,66	0,09
NYCM	18,58	130,50	814,46	12,97	384,86	752,72	654,19		736,00	772,36	34,16
SNVA	276,98	611,61	0,00	4,71	351,52	45,41	8,61	736,00		0,01	722,48
STTL	336,91	643,12	25,66	3,38	394,44	106,92	4,66	772,36	0,01		770,39
WASH	0,06	149,10	791,84	0,08	376,06	711,95	0,09	34,16	722,48	770,39	

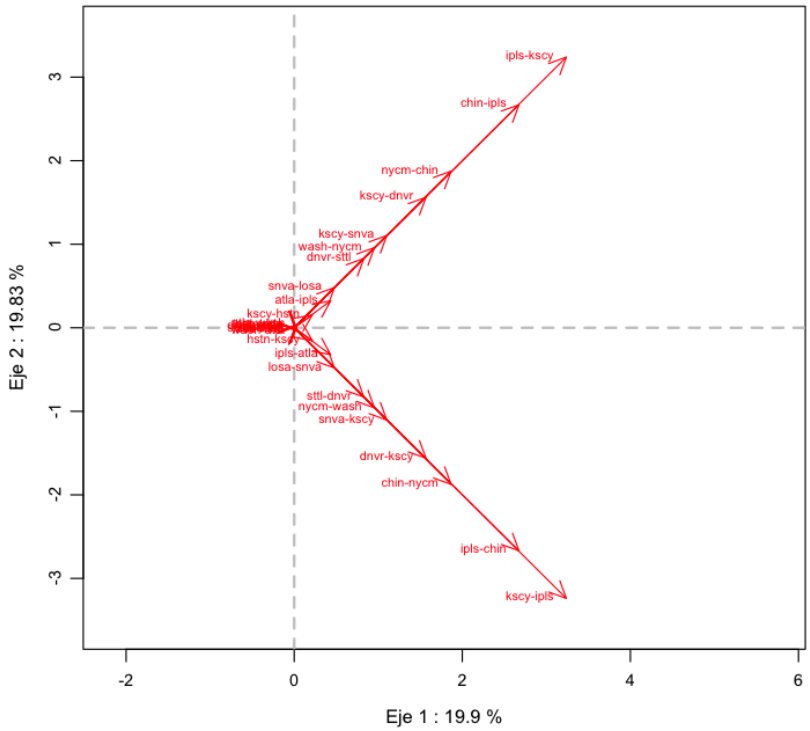
Y como en el caso anterior, las CRFE están emparejadas por rutas bidireccionales, y también podemos representarlas gráficamente (como antes, solo se representa un sentido de la ruta).



Representación de la CRFE acumulada de los ejes 1 y 2 para rutas bidireccionales de la Red Abilene (HJ-Biplot).

En este caso observamos un “codo” entre el nivel 600 y 400, lo que nos dejaría con 13 rutas bidireccionales (hasta Chicago-Sunnyvale) o 16 (hasta Denver-Indianápolis).

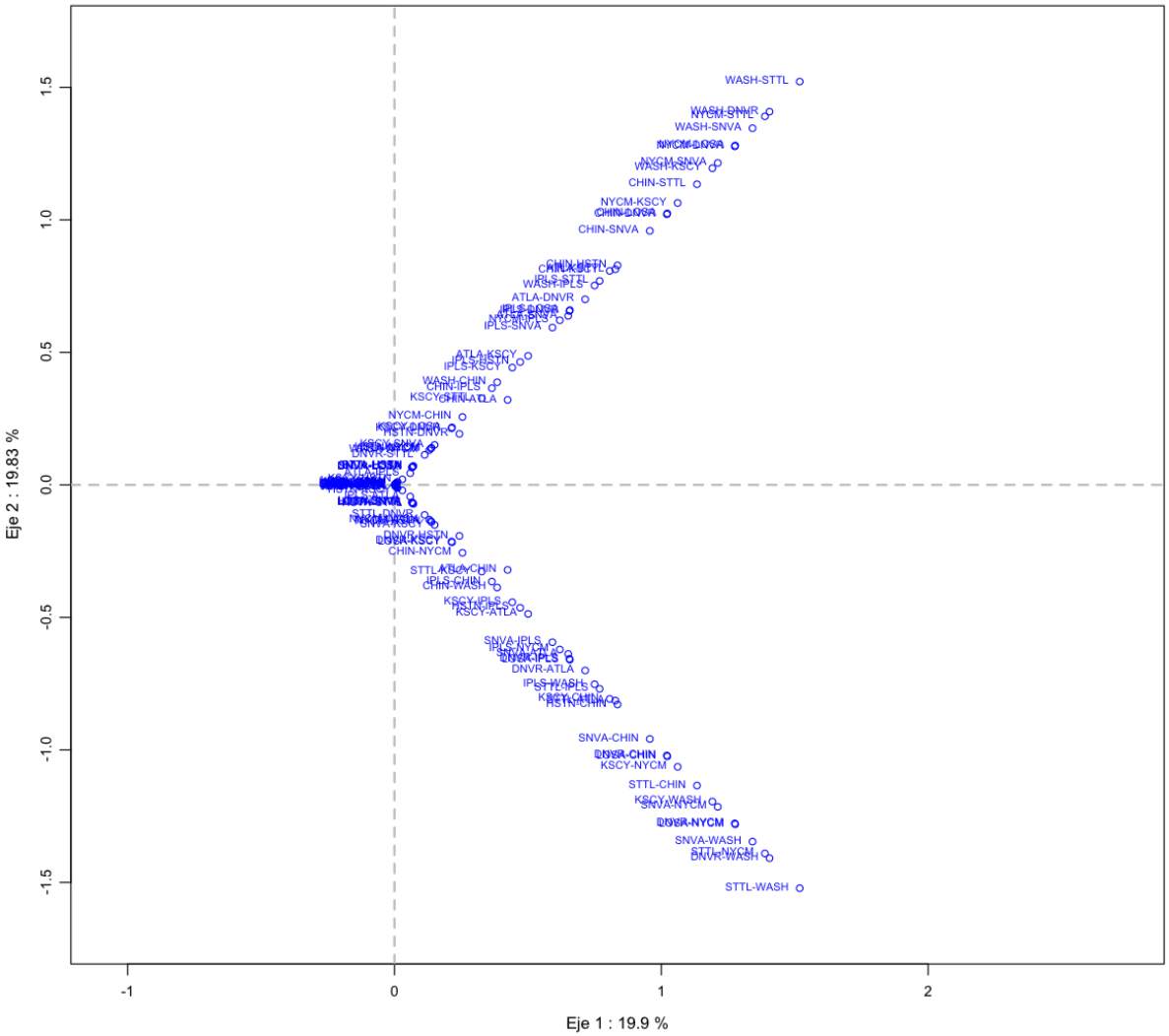
Representemos aisladamente los vectores correspondientes a los marcadores de los enlaces:



Representación de los marcadores HJ-Biplot de los enlaces de la matriz de enrutamiento de Abilene.

Aparecen los vectores agrupados en torno a dos direcciones principales, separadas casi 90º, indicando que están muy poco correladas. Curiosamente cada dirección de un mismo enlace aparece en una dirección distinta. El "orden" de los vectores se corresponde, lógicamente, con lo expuesto por Chua [349] relativo a la "energía" de los autovectores de la matriz de enrutamiento de Abilene.

Si representamos los marcadores obtenidos en el HJ-Biplot para las rutas, el resultado es similar: aparecen alineadas en las mismas dos direcciones que los enlaces, y también cada sentido de la ruta se "apoya" en el vector "contrario".

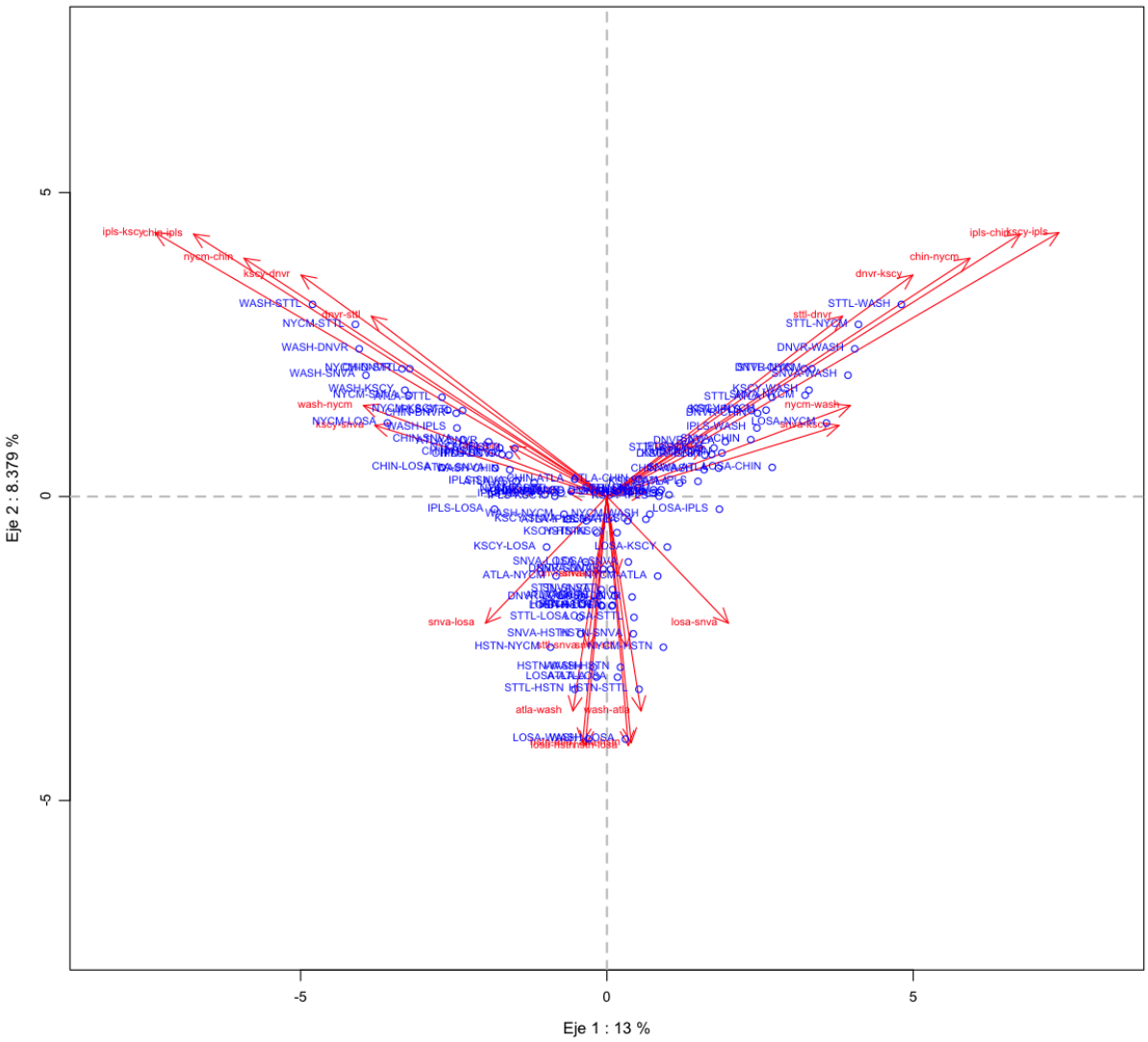


Representación de los marcadores HJ-Biplot de las rutas de la matriz de enrutamiento de Abilene.

Con normalización de la matriz de enrutamiento

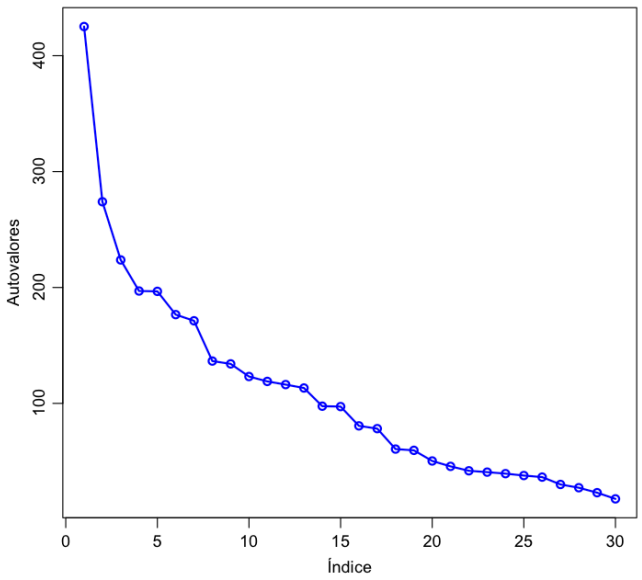
Repitamos ahora el estudio anterior, **pero normalizando la matriz de enrutamiento** previamente a someterla al HJ-Biplot, esto es, las variables se centran y escalan para que su varianza sea la unidad.

Lo primero que se comprueba en la representación es una mayor "riqueza", por diversidad, en el análisis obtenido. La inercia acumulada en los dos primeros ejes es del 21,4%, y si añadimos el tercero asciende solo hasta el 28,2%. Cuando no normalizábamos la matriz se obtenía una inercia explicada en el primer plano factorial del 39,7%, con la normalización hemos perdido calidad en este primer plano.



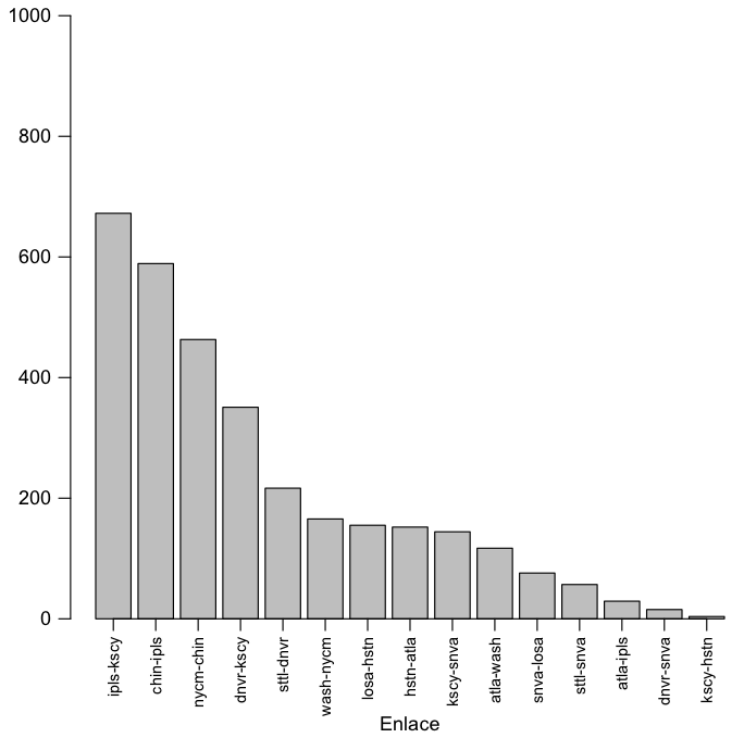
HJ-Biplot de la matriz de enrutamiento normalizada de la red Abilene.

Si representamos el espectro de los autovalores se comprueba que ha desaparecido el efecto de los dos primeros autovalores “asociados” (por llamarlo de alguna manera).

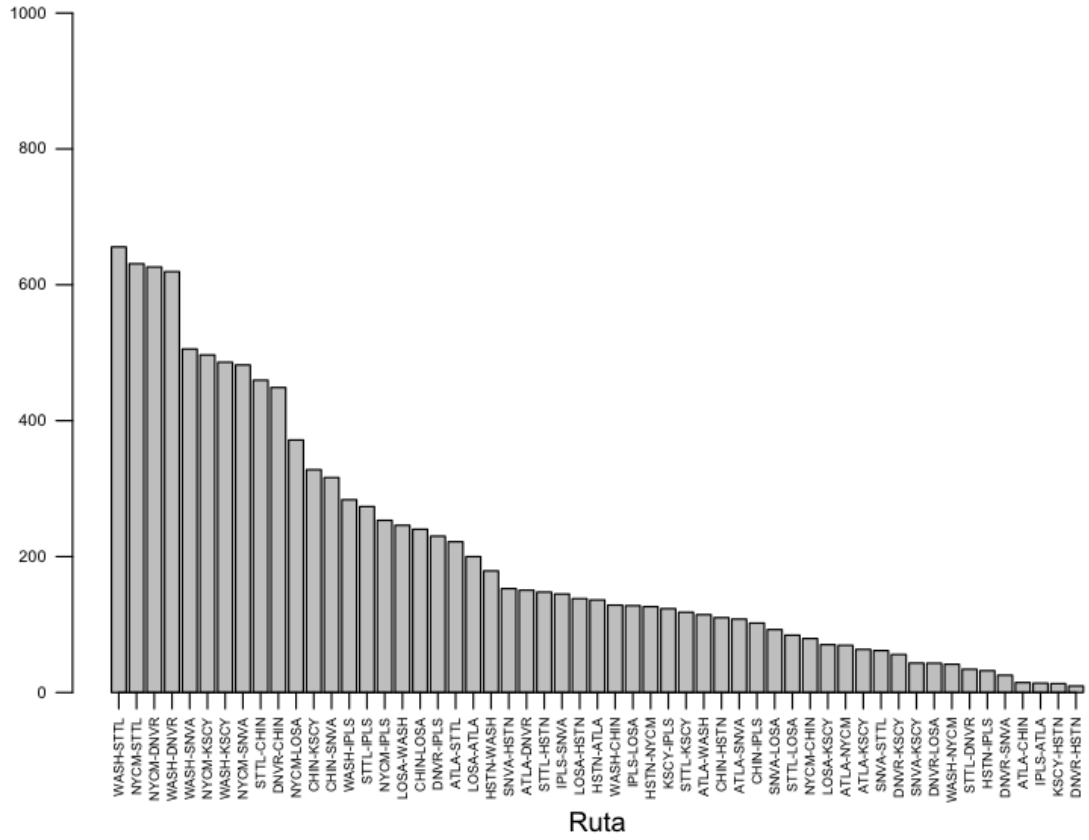


Screeplot de la matriz de enrutamiento de Abilene normalizada.

Representemos la CRFE para las filas y columnas en este nuevo supuesto



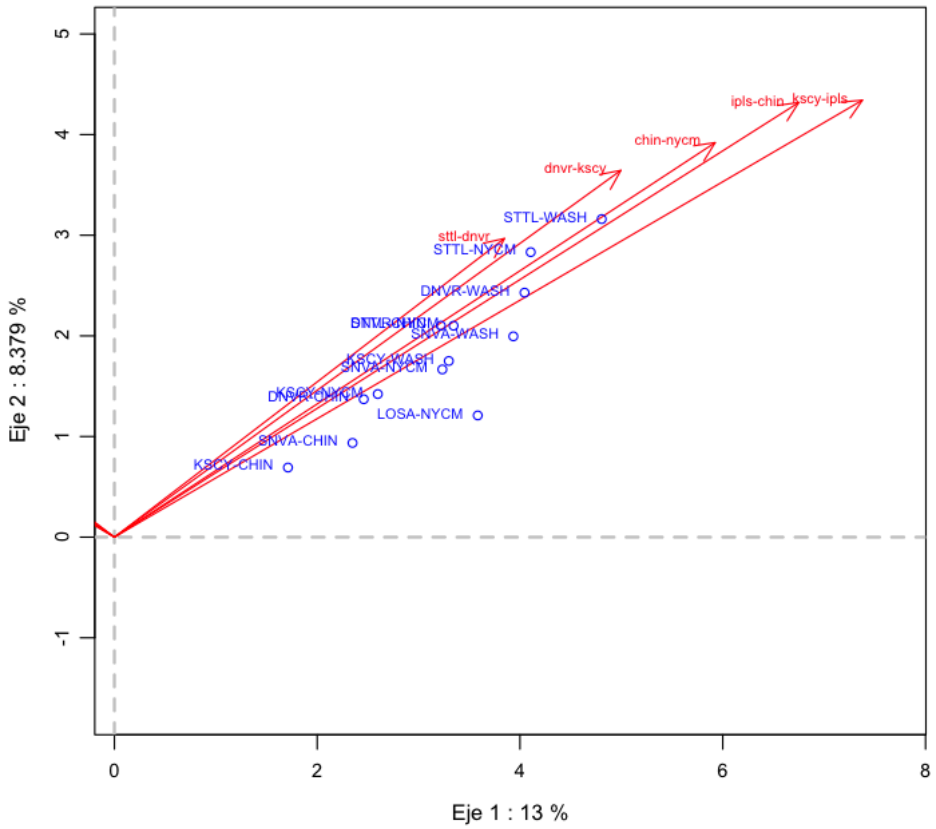
CRFE para las columnas (enlaces) de la matriz de enrutamiento de Abilene estandarizada.



CRFE para las filas (rutas) de la matriz de enrutamiento de Abilene estandarizada

Las CRFE se presenta por parejas para ambos sentidos de rutas y enlaces, debido a la simetría, por lo que se representa solo uno de los sentidos. Son muy pocos los marcadores con una calidad de representación “aceptable” en este primer plano factorial del HJ-Biplot, tanto para las rutas como para los enlaces.

Vista la reflexión en el eje Y así como la calidad suficiente de representación de solo algunos marcadores en este plano, representemos el primer cuadrante del HJ-Biplot con los marcadores fila **J** (rutas) con CRFE mayores de 300 y marcadores columna **H** (enlaces) con CRFE mayores de 200.



HJ-Biplot Matriz enrutamiento Abilene normalizada, marcadores con CRFE más elevadas.

Los marcadores fila con mayores CRFE se corresponden con las rutas de extremo a extremo del continente (STTL-WASH, STTL-NYCM, DNVR-WASH, SNVA-WASH,…). Los marcadores columna con mayores CRFE se corresponden con enlaces asociados también con una ruta de extremo a extremo (STTL-NYCM). Esto coincide con lo obtenido por Chua. Pero además en nuestro caso se muestran gráficamente las relaciones entre filas (rutas) y columnas (enlaces) mostrando los enlaces principales que forman parte precisamente de esas rutas.

7.7.4. ANÁLISIS HJ-BIPLOT DE LA MATRIZ DE ADYACENCIA DE LA RED ABILENE

La matriz de adyacencia de la red Abilene es la siguiente:

A	ATLA	CHIN	DNVR	HSTN	IPLS	KSCY	LOSA	NYCM	SNVA	STTL	WASH
atla	0	0	0	1	1	0	0	0	0	0	1
chin	0	0	0	0	1	0	0	1	0	0	0
dnvr	0	0	0	0	0	1	0	0	1	1	0
hstn	1	0	0	0	0	1	1	0	0	0	0
ipls	1	1	0	0	0	1	0	0	0	0	0
kscy	0	0	1	1	1	0	0	0	1	0	0
losa	0	0	0	1	0	0	0	0	1	0	0
nycm	0	1	0	0	0	0	0	0	0	0	1
snva	0	0	1	0	0	1	1	0	0	1	0
sttl	0	0	1	0	0	0	0	0	1	0	0
wash	1	0	0	0	0	0	0	1	0	0	0

Que se corresponde con el siguiente grafo de la red Abilene, en el que se han mantenido las “posiciones” geográficas aproximadas de los nodos para facilitar su comprensión, a pesar de su irrelevancia en el estudio.

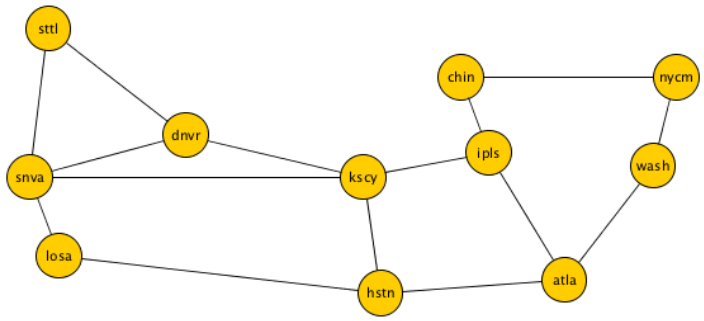


Gráfico geoposicionado de la red Abilene elaborado con yEd [381]

7.7.4.1. Análisis espectral de la matriz de adyacencia de Abilene

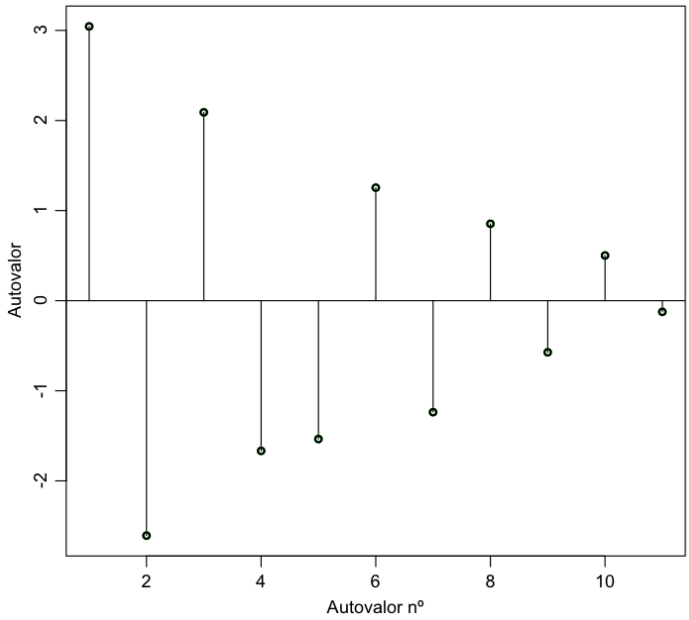
Comenzaremos realizando un análisis espectral de la matriz de adyacencia de Abilene

Los autovalores obtenidos son:

3.044 -2.608 2.091 -1.667 -1.535 1.254 -1.237 0.853 -0.573 0.502 -0.124

Se cumple que la suma de todos los autovalores de la matriz simétrica de Adyacencia de un grafo sin autolazos suman cero, ya que el valor de su traza es igualmente nulo.

Representamos los autovalores (espectro) obtenidos de la matriz de adyacencia



Representación de los autovalores ordenados de la matriz de adyacencia de la red Abilene

El mayor autovalor 3.044 se denomina índice del grafo o radio espectral, y contiene información sobre el grado de conectividad del grafo.

Y obtenemos también los siguientes autovectores

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
atla	-0.221	0.373	-0.447	-0.254	-0.377	-0.247	-0.194	0.114	-0.030	-0.435	0.322
chin	-0.109	0.150	-0.344	0.481	-0.242	0.400	-0.227	-0.239	-0.223	0.445	0.204
dnvr	-0.404	-0.134	0.280	-0.111	0.179	0.257	-0.715	-0.013	-0.182	-0.234	-0.175
hstn	-0.313	-0.443	-0.182	0.123	-0.045	-0.525	0.100	0.111	-0.561	0.145	-0.149
ipls	-0.265	-0.389	-0.373	-0.208	0.376	0.058	0.117	-0.533	0.273	-0.060	0.276
kscy	-0.477	0.491	0.011	0.120	0.042	-0.080	0.277	-0.330	0.096	-0.040	-0.560
losa	-0.257	0.290	0.055	-0.072	0.404	-0.331	-0.207	0.310	0.255	0.547	0.256
nycm	-0.067	-0.003	-0.346	-0.593	-0.004	0.444	0.165	0.329	-0.145	0.283	-0.301
snva	-0.468	-0.314	0.298	-0.004	-0.575	0.110	0.156	0.154	0.415	0.130	0.118
sttl	-0.286	0.172	0.276	0.069	0.258	0.292	0.452	0.165	-0.407	-0.208	0.464
wash	-0.095	-0.142	-0.379	0.508	0.248	0.157	0.023	0.519	0.306	-0.303	-0.166

Aunque el primer autovector tenga todas sus componentes negativas, se debe a un artefacto del algoritmo y la descomposición se cumple invirtiendo el signo de todos los elementos de la matriz. Así pues, se cumple también que asociado al mayor autovalor de la matriz de adyacencia se encuentra un autovector con todas sus componentes positivas (realmente del mismo signo) y ninguna es nula. Recordemos que este autovector contiene información sobre la “centralidad” de los nodos (centralidad del autovector o de Bonacich). Así obtenemos la siguiente ordenación de los nodos de la red Abilene según la centralidad del autovector.

kscy	snva	dnvr	hstn	sttl	ipls	losa	atla	chin	wash	nycm
0.477	0.468	0.404	0.313	0.286	0.265	0.257	0.221	0.109	0.095	0.067

7.7.4.2. Análisis HJ-Biplot de la matriz de adyacencia de Abilene

En primer lugar calcularemos la descomposición en valores singulares. Veamos los valores propios de **A** que se obtienen, y que son las raíces cuadradas de los autovalores de $\mathbf{AA}^T$ o $\mathbf{A}^T\mathbf{A}$:

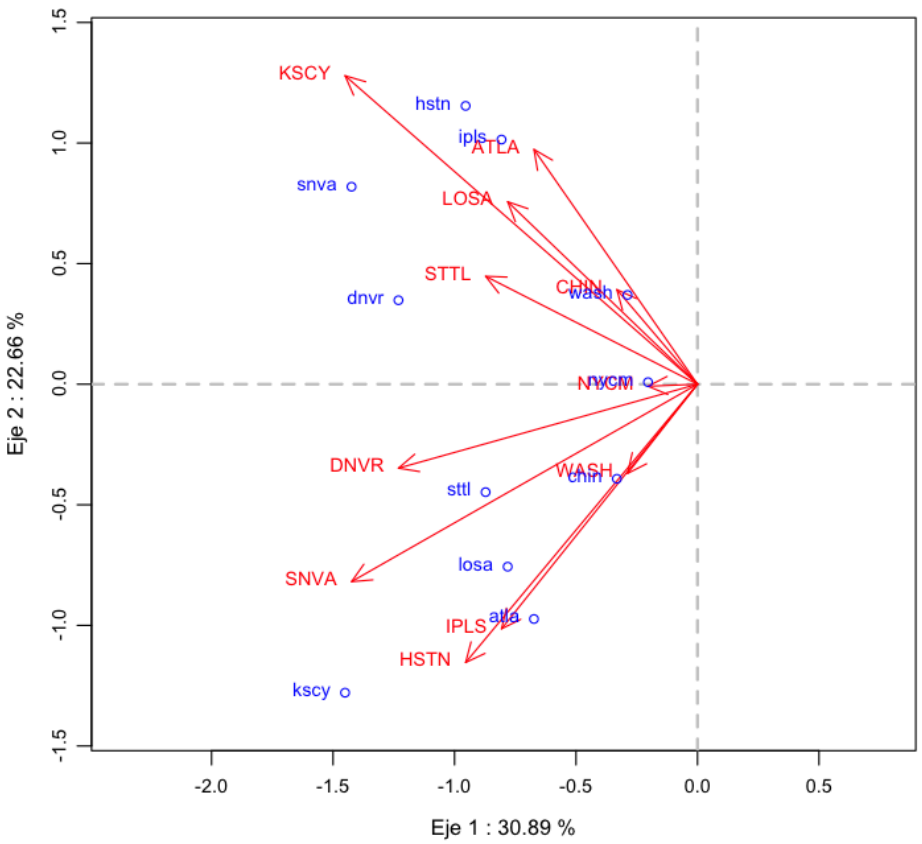
3.044 2.608 2.091 1.667 1.535 1.254 1.237 0.853 0.573 0.502 0.124

Comparémoslos con los autovalores obtenidos de la descomposición espectral

3.044 -2.608 2.091 -1.667 -1.535 1.254 -1.237 0.853 -0.573 0.502 -0.124

Si tomamos su valor absoluto y los ordenamos en sentido decreciente obtenemos los mismos resultados. Esto es, son los mismos valores que en el caso de la descomposición espectral.Vamos a obtener una representación HJ-Biplot de la matriz de adyacencia de Abilene.

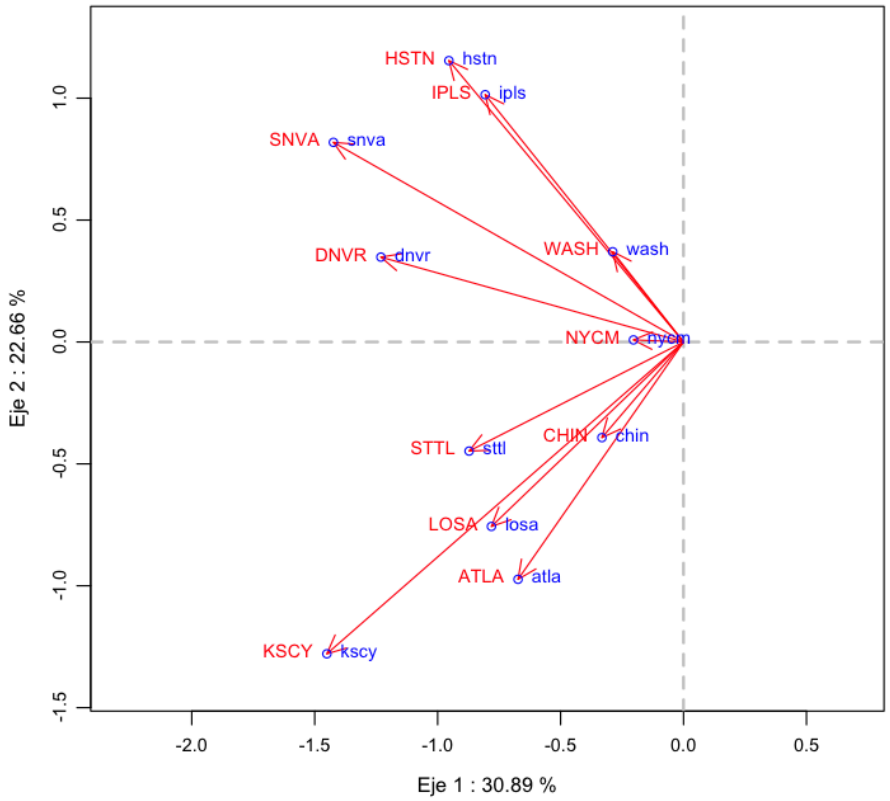
La calidad de representación para el primer plano (ejes factoriales 1 y 2) alcanza el 53,6%, y para los tres primeros ejes factoriales sube hasta el 68,1%.



Representación HJ-Biplot de la matriz de Adyacencia de la red Abilene (sin normalizar)

Examinando la representación HJ-Biplot se extraen algunas consideraciones de carácter general para la representación del primer plano factorial:

- La representación presenta una simetría sobre el eje X: Las coordenadas X (primer eje factorial) de los marcadores **H** y **J** son las mismas para las mismas filas/columnas. Las coordenadas Y (segundo eje factorial) de los marcadores **H** y **J** se sitúan en semiplanos diferentes; si la coordenada H está en el semiplano superior (y>0), la coordenada J está en plano inferior (y<0) para el mismo "nodo" de la red/grafo. Esto es evidentemente así por la simetría de la matriz de adyacencia **A**.



Representación HJ-Biplot primer plano factorial con inversión del eje Y para marcadores columna

- $\mathbf{HH}^T$ y $\mathbf{JJ}^T$ reproducen en sus respectivas diagonales los grados o valencias de los nodos de la red, como ya vimos en ejemplos anteriores.

	Atla	Chin	Dnvr	Hstn	Ipls	Kscy	Losa	Nycm	Snva	Sttl	Wash
diag(**D**)	3	2	3	3	3	4	3	2	4	2	2
diag($\mathbf{HH}^T$)	3	2	3	3	3	4	3	2	4	2	2
diag($\mathbf{JJ}^T$)	3	2	3	3	3	4	3	2	4	2	2

Analicemos la calidad de representación individual de cada elemento, calculando la CRFE para cada grupo de marcadores. Evidentemente, dada la simetría de las representación para los marcadores fila **J** y columna **H**, la CRFE serán iguales.

J	PC1	PC2	CRFE12
kscy	-1.451	-1.279	935
hstn	-0.954	1.154	747
snva	-1.424	0.819	675
losa	-0.781	-0.757	591
ipls	-0.807	1.014	560
dnvr	-1.231	0.348	545
sttl	-0.872	-0.448	480
atla	-0.673	-0.973	467
chin	-0.332	-0.392	132
wash	-0.288	0.370	110
nycm	-0.204	0.009	21

H	PC1	PC2	CRFE12
KSCY	-1.451	1.279	935
HSTN	-0.954	-1.154	747
SNVA	-1.424	-0.819	675
LOSA	-0.781	0.757	591
IPLS	-0.807	-1.014	560
DNVR	-1.231	-0.348	545
STTL	-0.872	0.448	480
ATLA	-0.673	0.973	467
CHIN	-0.332	0.392	132
WASH	-0.288	-0.370	110
NYCM	-0.204	-0.009	21

Podemos considerar que CRFE superiores a 400-500 pueden corresponderse con marcadores suficientemente bien representados. De esta manera solo dejaríamos fuera los nodos CHIN, WASH y NYCM. P

Podemos intentar extraer alguna información de carácter más “local” sobre los nodos y sus conexiones con nodos vecinos. Si comparamos el grafo de la red Abilene (por comprensión mantendremos, por ahora, las posiciones geográficas de los nodos), con la representación HJ-Biplot anterior observamos que:

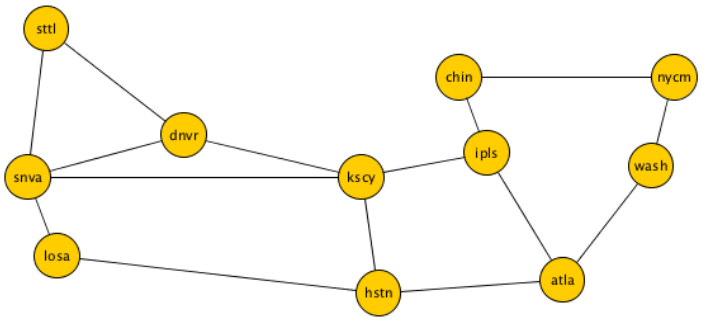


Gráfico geoposicionado de la red Abilene elaborado con yEd [381]

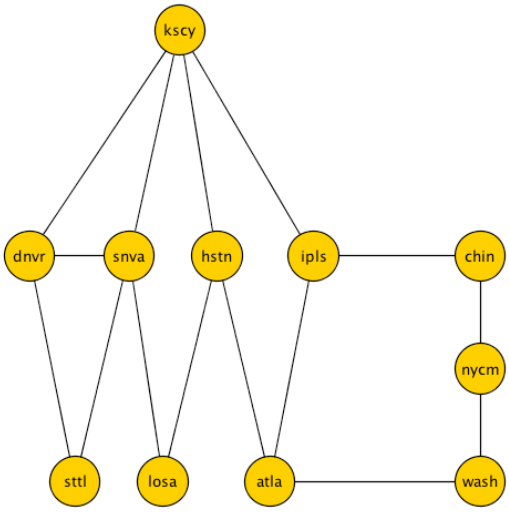


Grafico de la red Abilene reposicionado.

En la primera componente del HJ-Biplot en este primer plano factorial se reproduce el orden de la centralidad del autovector o de Bonacich.

KSCY	SNVA	DNVR	HSTN	STTL	IPLS	LOSA	ATLA	CHIN	WASH	NYCM
-1.451	-1.424	-1.231	-0.954	-0.872	-0.807	-0.781	-0.673	-0.332	-0.288	-0.204

7.7.5. ANÁLISIS HJ-BIPLOT DE LAS MATRICES LAPLACIANA Y DE INCIDENCIA DE LA RED ABILENE

Partimos de nuevo del esquema de la red Abilene, para la que construiremos la matriz de incidencia vértice-arista **Q**: Las filas corresponden 11 nodos y las columnas con los 15 enlaces.

Mapa de la red Abilene [348], [349]

Obtendremos la siguiente matriz de incidencia:

Q	atla-hstn	atla-ipls	atla-wash	chin-ipls	chin-nycm	dnvr-kscy	dnvr-snva	dnvr-sttl	hstn-kscy	hstn-losa	ipls-kscy	kscy-snva	losa-snva	nycm-wash	snva-sttl
atla	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
chin	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
dnvr	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
hstn	-1	0	0	0	0	0	0	0	1	1	0	0	0	0	0
ipls	0	-1	0	-1	0	0	0	0	0	0	1	0	0	0	0
kscy	0	0	0	0	0	-1	0	0	-1	0	-1	1	0	0	0
losa	0	0	0	0	0	0	0	0	0	-1	0	0	1	0	0
nycm	0	0	0	0	-1	0	0	0	0	0	0	0	0	1	0
snva	0	0	0	0	0	0	-1	0	0	0	0	-1	-1	0	1
sttl	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	-1
wash	0	0	-1	0	0	0	0	0	0	0	0	0	0	-1	0

Matriz de incidencia binaria de la red Abilene

Vamos a construir la matriz de adyacencia **A** también a partir del grafo de Abilene:

A	atla	chin	dnvr	hstn	ipls	kscy	losa	nycm	snva	sttl	wash
atla	0	0	0	1	1	0	0	0	0	0	1
chin	0	0	0	0	1	0	0	1	0	0	0
dnvr	0	0	0	0	0	1	0	0	1	1	0
hstn	1	0	0	0	0	1	1	0	0	0	0
ipls	1	1	0	0	0	1	0	0	0	0	0
kscy	0	0	1	1	1	0	0	0	1	0	0
losa	0	0	0	1	0	0	0	0	1	0	0
nycm	0	1	0	0	0	0	0	0	0	0	1
snva	0	0	1	0	0	1	1	0	0	1	0
sttl	0	0	1	0	0	0	0	0	1	0	0
wash	1	0	0	0	0	0	0	1	0	0	0

Matriz de adyacencia binaria de la red Abilene

Desde aquí construir la matriz diagonal de grados/valencias **D** es sencillo, como hemos visto anteriormente

$$k_i = \sum_{\forall j} A_{ij}$$

Luego la matriz **D** es igual a:

D	atla	chin	dnvr	hstn	ipls	kscy	losa	nycm	snva	sttl	wash
atla	3	0	0	0	0	0	0	0	0	0	0
chin	0	2	0	0	0	0	0	0	0	0	0
dnvr	0	0	3	0	0	0	0	0	0	0	0
hstn	0	0	0	3	0	0	0	0	0	0	0
ipls	0	0	0	0	3	0	0	0	0	0	0
kscy	0	0	0	0	0	4	0	0	0	0	0
losa	0	0	0	0	0	0	2	0	0	0	0
nycm	0	0	0	0	0	0	0	2	0	0	0
snva	0	0	0	0	0	0	0	0	4	0	0
sttl	0	0	0	0	0	0	0	0	0	2	0
wash	0	0	0	0	0	0	0	0	0	0	2

Matriz diagonal de grados de la red Abilene

Estamos en este momento en disposición de calcular la matriz Laplaciana de la red Abilene a través de dos expresiones diferentes

$$\mathbf{L} = \mathbf{Q}\mathbf{Q}^T = \mathbf{D} - \mathbf{A}$$

En ambos casos obtenemos para **L** el siguiente resultado

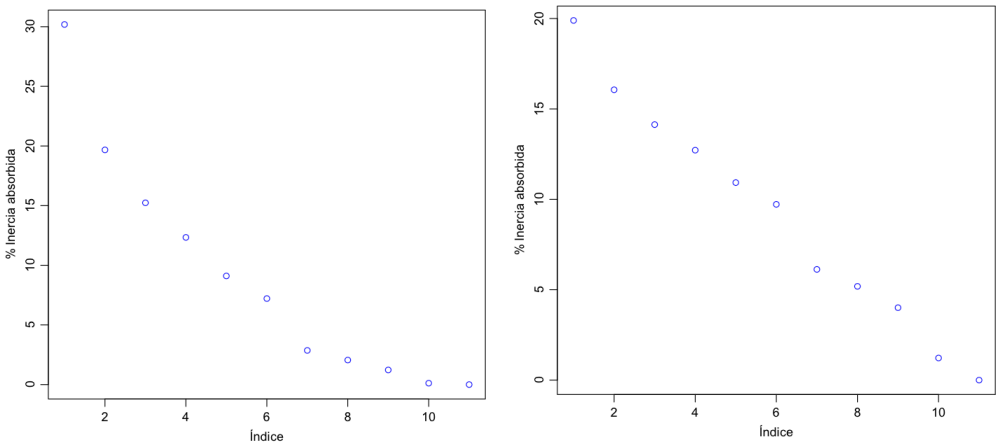
L	atla	chin	dnvr	hstn	ipls	kscy	losa	nycm	snva	sttl	wash
atla	3	0	0	-1	-1	0	0	0	0	0	-1
chin	0	2	0	0	-1	0	0	-1	0	0	0
dnvr	0	0	3	0	0	-1	0	0	-1	-1	0
hstn	-1	0	0	3	0	-1	-1	0	0	0	0
ipls	-1	-1	0	0	3	-1	0	0	0	0	0
kscy	0	0	-1	-1	-1	4	0	0	-1	0	0
losa	0	0	0	-1	0	0	2	0	-1	0	0
nycm	0	-1	0	0	0	0	0	2	0	0	-1
snva	0	0	-1	0	0	-1	-1	0	4	-1	0
sttl	0	0	-1	0	0	0	0	0	-1	2	0
wash	-1	0	0	0	0	0	0	-1	0	0	2

Matriz Laplaciana de la red Abilene

Obtenemos la DVS de las matrices Laplaciana **L** e incidencia **Q**. Los valores singulares de la matriz **Q** son, obviamente, iguales a la raíz cuadrada de los autovalores de la matriz **L** por la forma en que está construida $\mathbf{L} = \mathbf{Q}\mathbf{Q}^T$. Y los valores singulares de **L** son iguales al cuadrado de los valores singulares de **Q**.

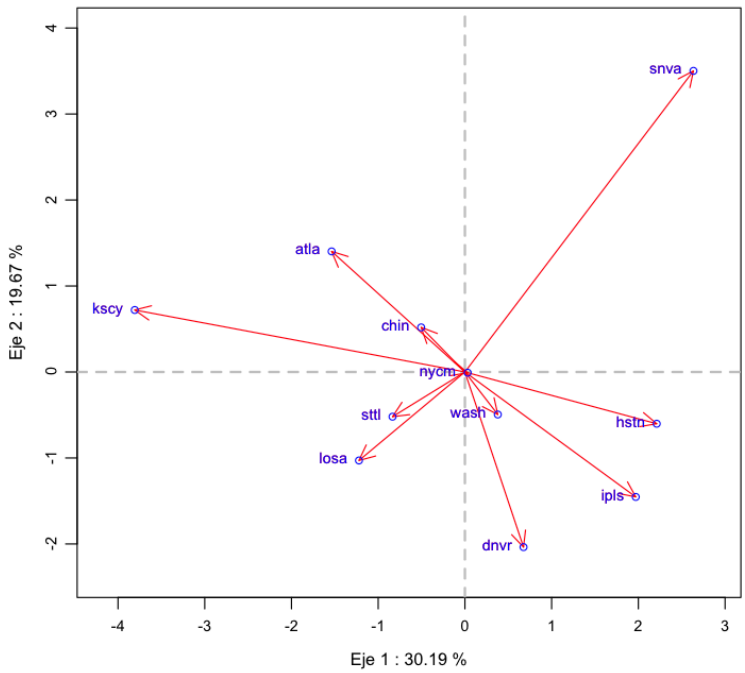
En un planteamiento habitual, los dos primeros ejes factoriales serían los candidatos a retener, por absorber el mayor porcentaje de inercia, para cualquiera de ambas matrices, tanto de incidencia **Q** como Laplaciana **L**.

Ambas matrices presentan un autovalor/valor singular nulo. Representemos las inercias absorbidas por cada ejes para ambos casos **L** y **Q**:

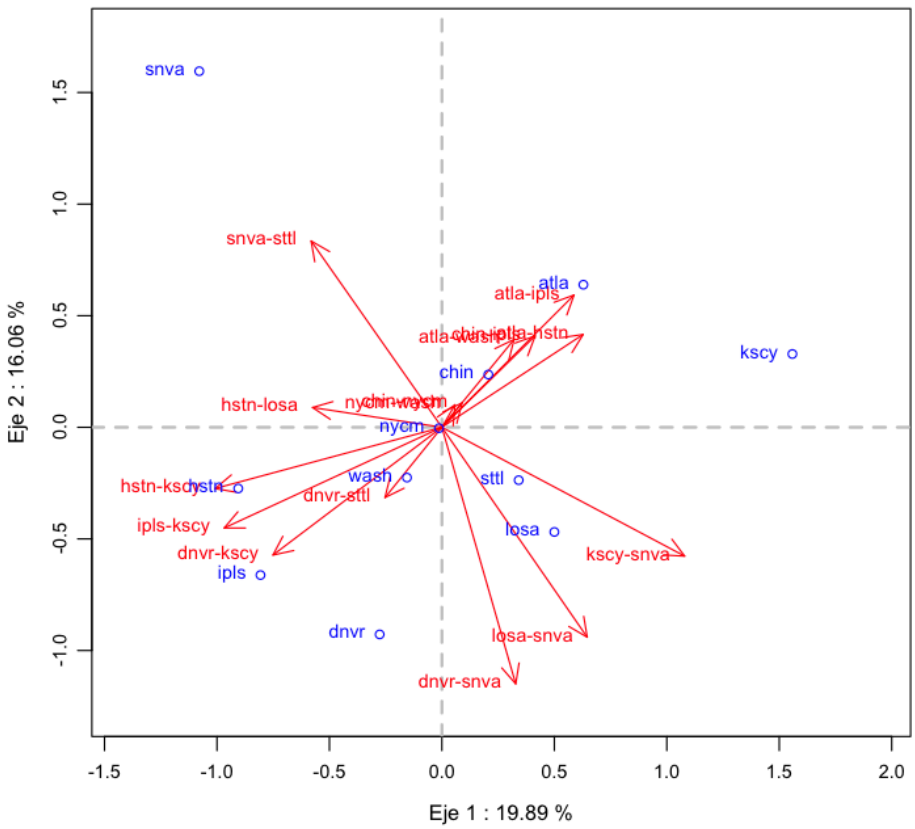


Representación autovalores de la matriz Laplaciana **L** [izq.] e Incidencia **Q** [der.] de la red Abilene.

Representemos un HJ-Biplot de la matriz Laplaciana **L** y de la matriz de incidencia **Q** para el primer plano factorial (clásico):



Representación HJ-Biplot (plano 1-2) de la matriz Laplaciana **L** de la red Abilene

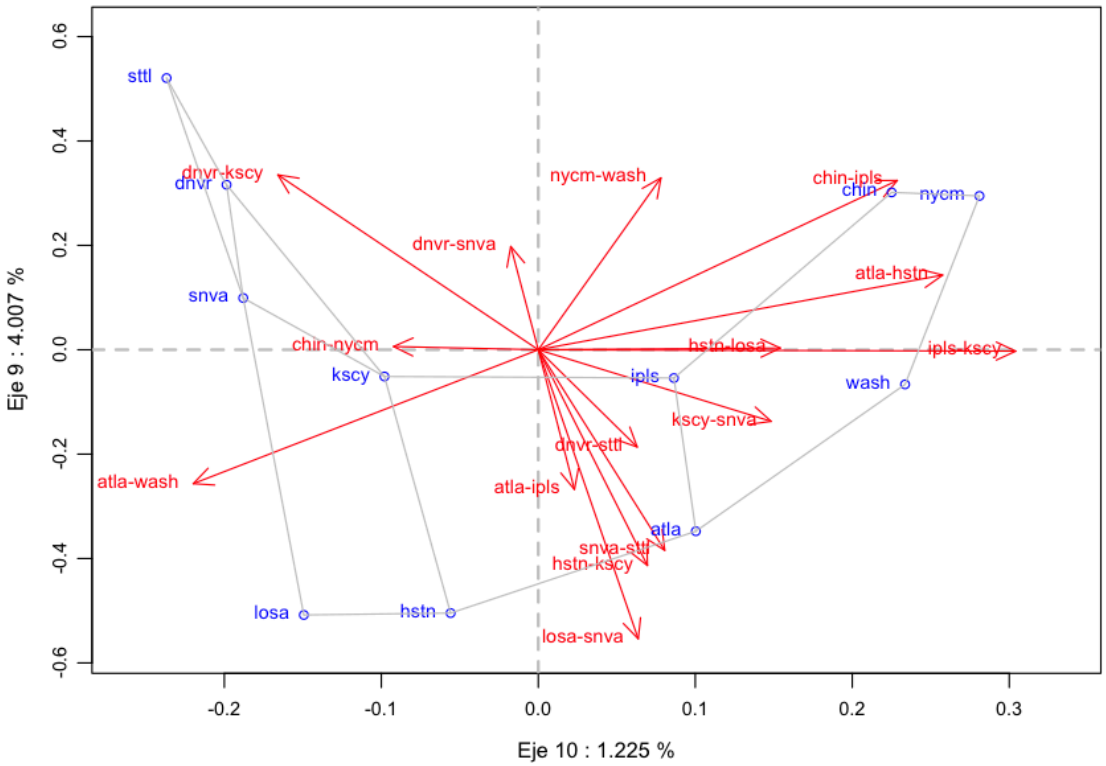


Representación HJ-Biplot (plano 1-2) de la matriz de incidencia **Q** de la red Abilene

Como ya se ha expuesto anteriormente, la simetría de la matriz Laplaciana **L** hace que los marcadores fila **J** y columna **H** sean iguales y que, por lo tanto, la representación simultánea no ofrezca mucha información adicional en comparación con el análisis más tradicional.

Sin embargo la representación del primer plano factorial del HJ-Biplot de la matriz de incidencia **Q** sí que ofrece algo más de información. Los marcadores fila **J** y columna **H** ya no son iguales (la matriz **Q** lógicamente no es simétrica) y representan, respectivamente, los nodos de la red (filas, marcadores **J**) y los enlaces que los unen (columnas, marcadores **H**). Como también se indicó anteriormente, los vectores **H** de los enlaces están orientados desde el nodo destino hacia el nodo origen, dependiendo de los signos positivos o negativos dados en la matriz de incidencia (se orientan hacia el valor positivo). Las posiciones relativas de los marcadores **J** (nodos de Abilene) son las mismas que las obtenidas para los marcadores **H** y **J** para el HJ-Biplot de la matriz Laplaciana.

Obtengamos la representación del último plano factorial de HJ-Biplot de la matriz de incidencia **Q** de la red Abilene, en nuestro caso el plano 9-10.



Representación HJ-Biplot plano 10-9 de la matriz de incidencia **Q** de la red Abilene.

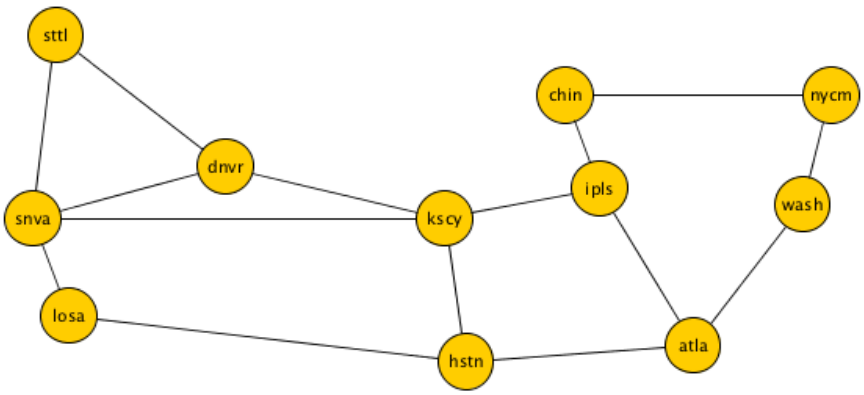
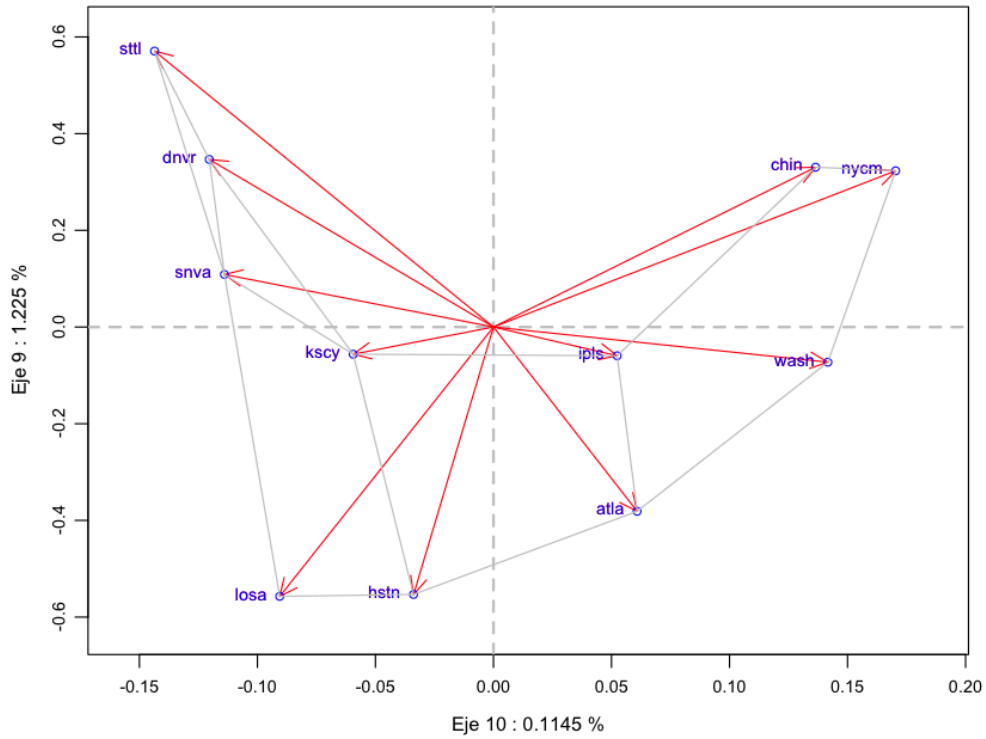


Gráfico geoposicionado de la red Abilene elaborado con yEd [381]

Es innegable la similitud entre la ubicación de los nodos (marcadores **J** del HJ-Biplot en la representación) y las respectivas localizaciones en el grafo de la red Abilene. Ciertamente, para obtener este "parecido" visual hemos tenido que "invertir" el eje y de la representación. La representación de ese mismo plano factorial de la representación HJ-Biplot de la matriz Laplaciana **L** nos permite obtener ese mismo resultado:



Representación HJ-Biplot plano 10-9 de la matriz Laplaciana **L** de la red Abilene.

Recordemos que en ningún momento se ha incluido en los datos información georeferenciada sobre los nodos de la red y que la única información sobre la que se ha aplicado el método ha sido la matriz binaria (sin información, por ejemplo, sobre distancias entre nodos) de incidencia. Así pues el parecido obtenido con la escasa información aportada al análisis es destacable, no obstante hasta aquí hay, como hemos indicado, pocas novedades con la aplicación de Koren *et al* [315].

En el primer plano factorial del HJ-Biplot de la matriz de incidencia **Q** de la red Abilene, con una representación que no se ajusta al grafo analizado, la inercia absorbida alcanza el 36%, mientras que en el último plano factorial (9-10) la inercia es de solo el 5.2%, con una representación que se aproxima bastante a la georeferenciada de la red Abilene.

Recordemos que a partir de las propiedades del HJ-Biplot y de las propiedades de la matriz de incidencia **Q** se tiene que:

$$\mathbf{H}\mathbf{H}^{T} = \mathbf{Q}^{T}\mathbf{Q} = \mathbf{K}$$

$$\mathbf{J}\mathbf{J}^{T} = \mathbf{Q}\mathbf{Q}^{T} = \mathbf{L}$$

Calculemos ahora las CRFE para la representación HJ-Biplot de la matriz de incidencia **Q** de la red Abilene.

CRFEcol	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
atla-hstn	197.2	86.5	333.7	59.3	112.8	62.9	89.8	14.4	10.2	33.2	0
atla-ipls	172.6	175.6	268.2	100.3	89.1	16.3	1.0	140.7	35.8	0.3	0
atla-wash	51.5	77.5	402.0	57.1	143.0	25.2	54.9	131.6	32.9	24.2	0
chin-ipls	86.0	83.7	13.2	533.0	23.7	11.8	158.2	11.5	52.6	26.2	0
chin-nycm	4.0	6.0	8.8	464.4	82.2	138.7	30.4	261.2	0.0	4.3	0
dnvr-kscy	282.1	163.8	397.2	52.7	6.3	3.2	2.4	22.3	56.2	13.8	0
dnvr-snva	53.9	660.7	130.6	57.2	5.5	45.1	26.9	0.4	19.6	0.2	0
dnvr-sttl	32.0	49.6	258.3	79.2	82.0	467.0	1.8	10.7	17.4	2.0	0
hstn-kscy	508.5	37.7	11.4	54.2	201.9	77.2	16.0	5.1	85.5	2.4	0
hstn-losa	165.5	3.9	62.0	82.2	430.8	48.9	194.7	0.0	0.0	11.9	0
ipls-kscy	468.5	101.8	27.8	107.1	34.2	160.6	20.0	33.7	0.0	46.3	0
kscy-snva	582.6	166.6	72.3	0.1	23.6	72.4	45.5	16.5	9.4	11.0	0
losa-snva	208.8	442.2	7.5	2.0	2.9	52.9	125.8	2.6	153.4	2.0	0
nycm-wash	1.7	5.1	105.5	257.1	270.8	54.3	136.5	111.7	54.2	3.1	0
snva-sttl	169.0	348.4	21.6	1.8	129.9	221.9	14.8	15.5	73.9	3.2	0
	2984.1	2409.0	2120.0	1907.7	1638.8	1458.4	918.8	778.1	601.1	183.8	0

CRFEfilas	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
atla	131.9	136.0	470.3	4.8	23.6	66.0	116.1	7.5	40.3	3.4	0
chin	21.3	27.9	0.1	522.2	5.4	79.3	27.2	245.8	45.5	25.3	0
dnvr	25.5	287.0	353.7	98.3	17.3	160.8	9.0	1.9	33.3	13.1	0
hstn	273.2	25.1	81.5	101.8	422.7	8.6	0.1	1.3	84.9	1.0	0
ipls	216.8	146.1	34.2	329.9	82.7	6.2	93.5	87.3	1.0	2.5	0
kscy	607.1	27.0	162.2	2.1	0.1	173.0	16.7	8.8	0.7	2.4	0
losa	125.0	109.9	26.6	28.7	153.7	69.8	344.7	1.3	129.1	11.2	0
nycm	0.1	0.0	41.3	370.3	198.7	125.6	160.9	20.1	43.4	39.5	0
snva	290.9	636.3	0.1	3.5	43.6	8.3	5.7	0.4	2.5	8.8	0
sttl	58.3	28.1	101.2	27.5	127.7	456.9	3.4	33.4	135.5	28.0	0
wash	12.1	25.4	216.9	145.9	246.3	1.9	9.9	312.2	2.2	27.3	0
	1762.3	1448.7	1488.0	1634.8	1321.9	1156.4	787.2	719.9	518.2	162.6	0

Recuperemos las primeras componentes de las matrices de marcadores **H** y **J** ordenadas por el valor absoluto de sus componentes:

kscy-snva	hstn-kscy	ipls-kscy	dnvr-kscy	losa-snva	atla-hstn	atla-ipls	snva-sttl	hstn-losa	chin-ipls	dnvr-snva	atla-wash	dnvr-sttl	chin-nycm	nycm-wash
1.08	-1.01	-0.97	-0.75	0.65	0.63	0.59	-0.58	-0.57	0.41	0.33	0.32	-0.25	0.09	0.06

kscy	snva	hstn	ipls	atla	losa	sttl	dnvr	chin	wash	nycm
1.558	-1.079	-0.905	-0.807	0.629	0.500	0.342	-0.277	0.206	-0.155	-0.013

Recordemos los resultados que obtuvimos para la centralidad del autovector o de Bonacich obtenida a partir de la matriz de Adyacencia **A**

kscy	snva	dnvr	hstn	sttl	ipls	losa	atla	chin	wash	nycm
0.477	0.468	0.404	0.313	0.286	0.265	0.257	0.221	0.109	0.095	0.067

Los resultados son bastante comparables entre sí y se podría concluir que el primer eje principal de podría estar relacionado directamente con la centralidad del elemento del grafo, bien se trate de un nodo/vértice o enlace/arista.

7.7.6. ANÁLISIS HJ-BIPLOT DE UNA MATRIZ DE INCIDENCIA CON INFORMACIÓN DE TRÁFICO DE LA RED ABILENE

Como último ejemplo obtendremos un HJ-Biplot de una matriz de incidencia, en la que en lugar de incorporar exclusivamente información sobre topología, colocaremos información relativa al tráfico cursado, en una suerte de matriz de tráfico. Operaremos por lo tanto sobre la matriz de incidencia generalizada.

La matriz de incidencia vértice-arista **Q** de la red, tiene en sus filas los nodos de la red y en sus columnas los enlaces. Cuando operamos con información sobre topología, los elementos de la matriz son la raíz cuadrada de las ponderaciones de las aristas entre nodos, con el signo apropiado, como en nuestro caso operaremos con datos de tráfico será el tráfico entrante o saliente del nodo correspondiente, con el signo correspondiente. Vamos a utilizar la primera matriz de tráfico del conjunto de datos X01 de entre las publicadas por Zhang [305].

Utilizaremos la propiedad de la matriz de tráfico **Z** para obtener el flujo en cada enlace **X** de forma que **X=BZ** y a partir de esa matriz de tráfico en cada enlace obtendremos la matriz **Q** de incidencia de tráfico, en donde cada elemento es el tráfico entrante o saliente del nodo para ese enlace.

La matriz de tráfico **Z** que vamos a estudia la obtenemos del primer intervalo temporal publicado por Zhang. Eliminaremos como en casos anteriores el tráfico con origen o destino ATLA-M5 por estimarse inestable, pero mantendremos el tráfico “local” correspondiente a tráfico con origen y destino en el mismo nodo, que en casos previos retiramos. Esta matriz de tráfico puede ser expresa en forma de una matriz 11x11 con 121 elemento, de la forma más tradicional, pero la utilizaremos en forma de un vector :

ATLA-ATLA	ATLA-CHIN	...	ATLA-WASH	CHIN-ATLA	...	WASH-WASH
30997781	6106169	...	19405592	4775747	...	70370056

El vector **Z** tiene 121 filas (11 nodos Origen x 11 nodos Destino) y 1 columnas.

Por lo que respecta a la matriz de enrutamiento **B** es la utilizada en casos anteriores, pero aquí mantendremos las rutas locales que se representarían en un grafo como

autolazos, con todos sus elemento a cero, por consistencia con la matriz **Z** antes definida:

	ATLA-ATLA	ATLA-CHIN	ATLA-DNVR	ATLA-HSTN	ATLA-IPLS	ATLA-KSCY	ATLA-LOSA	ATLA-NYCM	...
atla-hstn	0	0	0	1	0	0	1	0	...
atla-ipls	0	1	1	0	1	1	0	0	...
atla-wash	0	0	0	0	0	0	0	1	...
chin-ipls	0	0	0	0	0	0	0	0	...
chin-nycm	0	0	0	0	0	0	0	0	...
dnvr-kscy	0	0	0	0	0	0	0	0	...
dnvr-snva	0	0	0	0	0	0	0	0	...
dnvr-sttl	0	0	0	0	0	0	0	0	...
hstn-atla	0	0	0	0	0	0	0	0	...
hstn-kscy	0	0	0	0	0	0	0	0	...
hstn-losa	0	0	0	0	0	0	1	0	...
ipls-atla	0	0	0	0	0	0	0	0	...
ipls-chin	0	1	0	0	0	0	0	0	...
ipls-kscy	0	0	1	0	0	1	0	0	
kscy-dnvr	0	0	1	0	0	0	0	0	...
kscy-hstn	0	0	0	0	0	0	0	0	...
kscy-ipls	0	0	0	0	0	0	0	0	...
kscy-snva	0	0	0	0	0	0	0	0	...
losa-hstn	0	0	0	0	0	0	0	0	...
losa-snva	0	0	0	0	0	0	0	0	...
nycm-chin	0	0	0	0	0	0	0	0	...
nycm-wash	0	0	0	0	0	0	0	0	...
snva-dnvr	0	0	0	0	0	0	0	0	...
snva-kscy	0	0	0	0	0	0	0	0	...
snva-losa	0	0	0	0	0	0	0	0	...
snva-sttl	0	0	0	0	0	0	0	0	...
sttl-dnvr	0	0	0	0	0	0	0	0	...
sttl-snva	0	0	0	0	0	0	0	0	...
wash-atla	0	0	0	0	0	0	0	0	...
wash-nycm	0	0	0	0	0	0	0	1	...

Primeras 8 columnas (de las 121 totales) de la matriz de enrutamiento de Abilene.

Esta matriz **B** tiene por lo tanto 121 columnas (como antes, 11 nodos Origen x 11 nodos Destino) y 30 filas (los 30 enlaces bidireccionales 15x2 de la red Abilene).

Para obtener el vector **X** de 30 filas y 1 columna que representa el flujo total sobre cada enlace efectuamos

$$\mathbf{X}_{30x1} = \mathbf{B}_{30x121}\,\mathbf{Z}_{121x1}$$

Con lo que obtenemos el tráfico sobre cada uno de los 30 enlaces que componen la red Abilene para el primer intervalo temporal contenido en la matriz **X01**.

Enlace	Trafico cursado
atla-hstn	57474764
atla-ipls	24010295
atla-wash	51018938
chin-ipls	185064202
chin-nycm	93783390
dnvr-kscy	60789848
dnvr-snva	13410387
dnvr-sttl	34535033
hstn-atla	35193626
hstn-kscy	13821972
hstn-losa	80081414
ipls-atla	17136696
ipls-chin	211250865
ipls-kscy	163087895
kscy-dnvr	68724465
kscy-hstn	21333370
kscy-ipls	167101500
kscy-snva	56766744
losa-hstn	36334235
losa-snva	93462732
nycm-chin	191411432
nycm-wash	99496462
snva-dnvr	7292360
snva-kscy	83783897
snva-losa	69659509
snva-sttl	11645512
sttl-dnvr	33018785
sttl-snva	12684055
wash-atla	111723110
wash-nycm	143812713

Podemos comprobar en la matriz de OD que se obtendría no es simétrica, ya que los volúmenes de tráfico cursado son diferentes en cada sentido de los enlaces. Esta matriz sería equivalente a la matriz de adyacencia **A** ponderada:

O / D	ATLA	CHIN	DNVR	HSTN	IPLS	KSCY	LOSA	NYCM	SNVA	STTL	WASH
atla	0	0	0	57474764	24010295	0	0	0	0	0	51018938
chin	0	0	0	0	185064202	0	0	93783390	0	0	0
dnvr	0	0	0	0	0	60789848	0	0	13410387	34535033	0
hstn	35193626	0	0	0	0	13821972	80081414	0	0	0	0
ipls	17136696	211250865	0	0	0	163087895	0	0	0	0	0
kscy	0	0	68724465	21333370	167101500	0	0	0	56766744	0	0
losa	0	0	0	36334235	0	0	0	0	93462732	0	0
nycm	0	191411432	0	0	0	0	0	0	0	0	99496462
snva	0	0	7292360	0	0	83783897	69659509	0	0	11645512	0
sttl	0	0	33018785	0	0	0	0	0	12684055	0	0
wash	111723110	0	0	0	0	0	0	143812713	0	0	0

Y a partir de aquí podemos construir una matriz de incidencia “generalizada” **Q** del grafo de la red Abilene considerando que el peso de cada arista se corresponde con la raíz cuadrada del tráfico cursado en el sentido correspondiente de cada enlace. Consideraremos en este caso el tráfico saliente de cada nodo como positivo y

recíprocamente el entrante como negativo. Con esto obtenemos (se muestra $\mathbf{Q}^T$ por razones de espacio)

$\mathbf{Q}^T$	atla	chin	dnvr	hstn	ipls	kscy	losa	nycm	snva	sttl	wash
atla-hstn	7581.21	0.00	0.00	-5932.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00
atla-ipls	4900.03	0.00	0.00	0.00	-4139.65	0.00	0.00	0.00	0.00	0.00	0.00
atla-wash	7142.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-10569.92
chin-ipls	0.00	13603.83	0.00	0.00	-14534.47	0.00	0.00	0.00	0.00	0.00	0.00
chin-nycm	0.00	9684.18	0.00	0.00	0.00	0.00	0.00	-13835.15	0.00	0.00	0.00
dnvr-kscy	0.00	0.00	7796.78	0.00	0.00	-8290.02	0.00	0.00	0.00	0.00	0.00
dnvr-snva	0.00	0.00	3662.02	0.00	0.00	0.00	0.00	0.00	-2700.44	0.00	0.00
dnvr-sttl	0.00	0.00	5876.65	0.00	0.00	0.00	0.00	0.00	0.00	-5746.20	0.00
hstn-kscy	0.00	0.00	0.00	3717.79	0.00	-4618.81	0.00	0.00	0.00	0.00	0.00
hstn-losa	0.00	0.00	0.00	8948.82	0.00	0.00	-6027.79	0.00	0.00	0.00	0.00
ipls-kscy	0.00	0.00	0.00	0.00	12770.59	-12926.77	0.00	0.00	0.00	0.00	0.00
kscy-snva	0.00	0.00	0.00	0.00	0.00	7534.37	0.00	0.00	-9153.35	0.00	0.00
losa-snva	0.00	0.00	0.00	0.00	0.00	0.00	9667.61	0.00	-8346.23	0.00	0.00
nycm-wash	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9974.79	0.00	0.00	-11992.19
snva-sttl	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3412.55	-3561.47	0.00

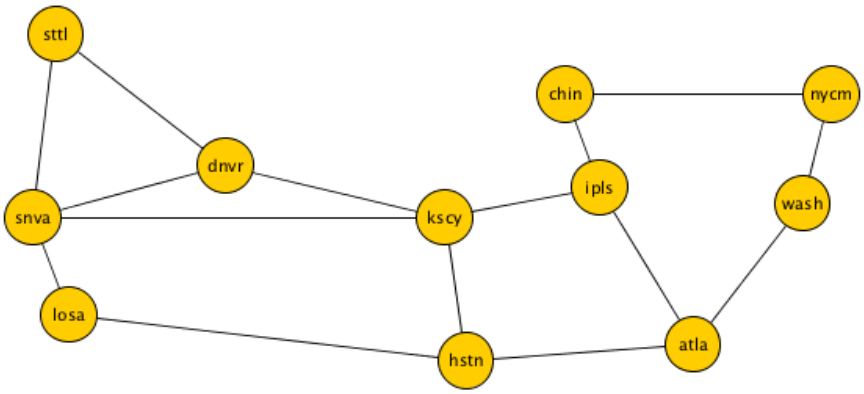
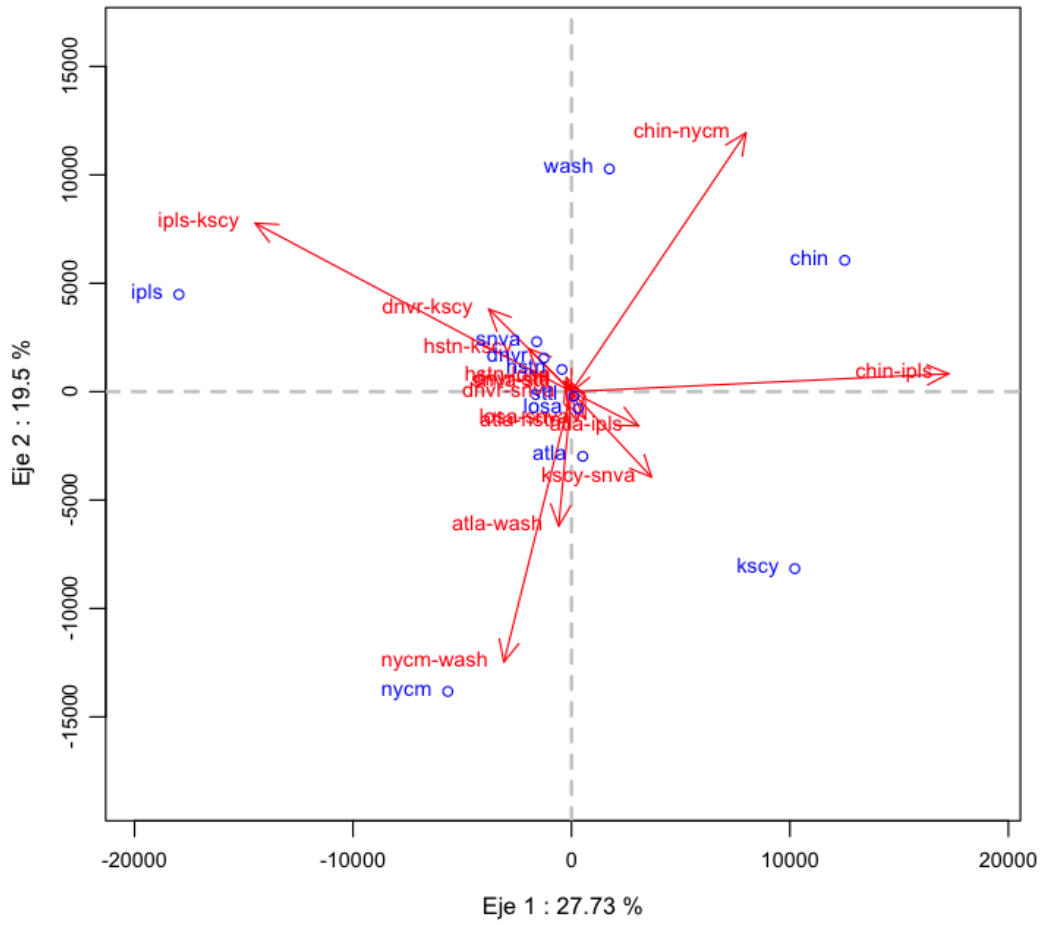


Gráfico geoposicionado de la red Abilene elaborado con yEd [381]

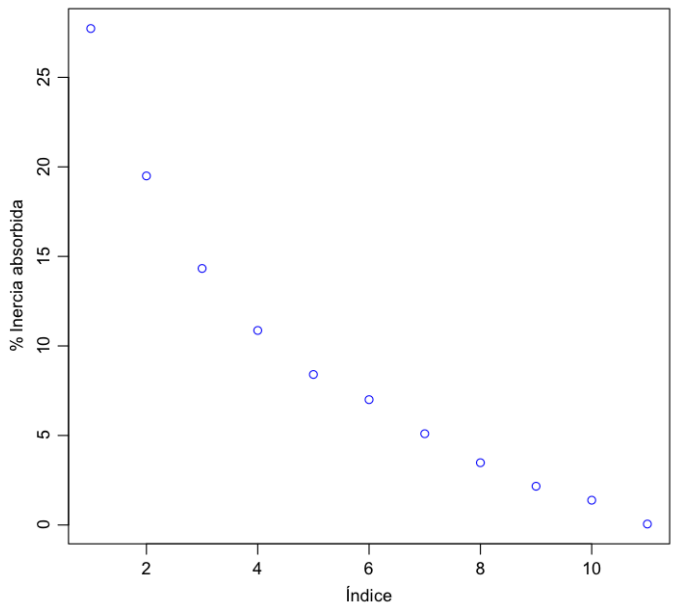
Suponemos que $\mathbf{L}=\mathbf{Q}\mathbf{Q}^T$, no obstante observamos que, a pesar de que el grafo está compuesto por una sola componente el menor autovalor de **L** no es nulo, al igual que la suma por columnas de la matriz tampoco es nula, debido precisamente a la no simetría de matriz de tráfico. Esto tiene como consecuencia que no se respetan las reglas de un grafo de flujo, y en concreto que la suma de tráfico entrante no es igual que la de tráfico saliente de un nodo, lo que es consecuente con lo esperado en una red de datos, pero que tiene consecuencias en el análisis de grafos. Evidentemente la definición de la matriz Laplaciana $\mathbf{L}=\mathbf{D}-\mathbf{A}$ tampoco se cumple, porque hay dos matrices **D** diferentes, para tráfico entrante y saliente. Sí que se cumple que los autovalores no nulos de $\mathbf{L}=\mathbf{Q}\mathbf{Q}^T$ y $\mathbf{K}=\mathbf{Q}^T\mathbf{Q}$ son reales e iguales, por construcción y al tratarse de matrices de Gram.

No obstante lo anterior, continuaremos con la metodología propuesta de análisis de tráfico basado en la matriz **Q** de incidencia vértice-arista y en la aplicación del HJ-Biplot a esta matriz obtenemos la siguiente representación, para el primer plano factorial y operando con los datos en bruto (sin normalizar) :



HJ-Biplot, primer plano factorial matriz **Q** de tráfico de la red Abilene, primer intervalo temporal de la matriz **X01**

Observamos que la inercia absorbida para este primer plano factorial alcanza el 47,22%. Si representamos la inercia absorbida para cada eje factorial, obtenemos:



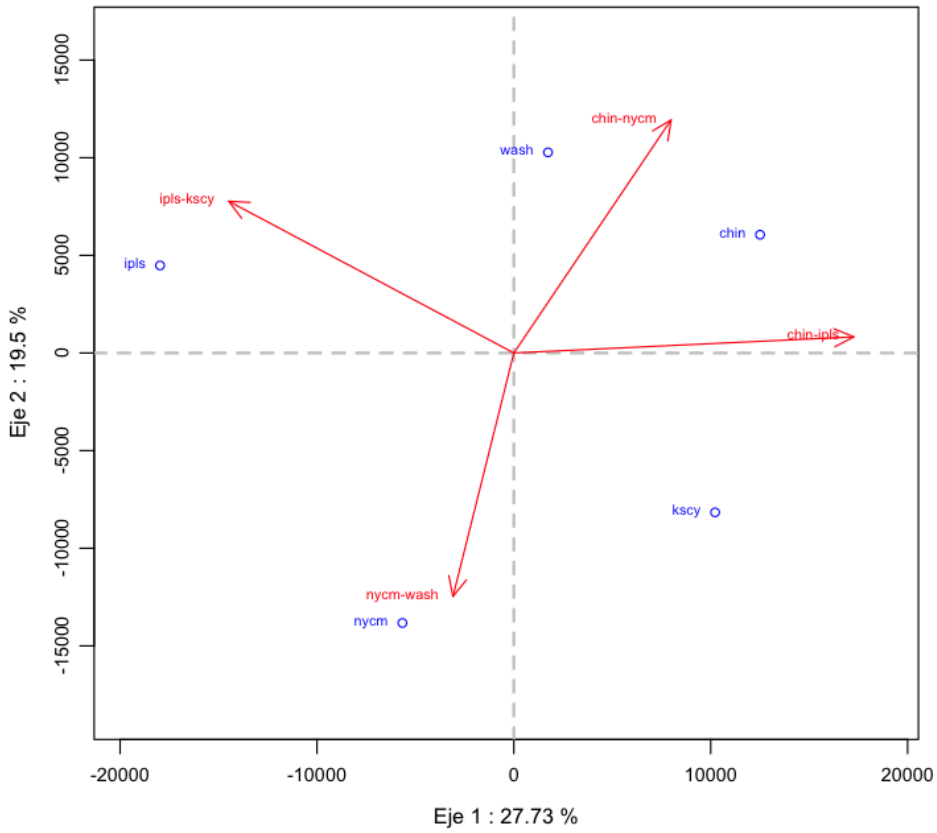
Inercia absorbida por eje factorial

Analicemos la calidad de representación de cada marcador del HJ-Biplot obtenido:

CRFEcol	PC1	PC2	PC1+PC2
ipls-kscy	635.1	182.9	818.1
chin-ipls	752.8	1.7	754.5
chin-nycm	223.7	499.7	723.3
nycm-wash	39.3	639.2	678.5
atla-ipls	230.6	61.0	291.5
atla-wash	2.1	236.5	238.6
dnvr-kscy	110.7	112.1	222.8
hstn-kscy	108.8	111.8	220.5
kscy-snva	95.7	110.7	206.4
atla-hstn	0.7	20.2	21.0
losa-snva	2.6	10.0	12.6
snva-sttl	2.2	6.9	9.2
dnvr-sttl	1.5	3.6	5.1
hstn-losa	0.5	3.8	4.2
dnvr-snva	0.0	0.0	0.0

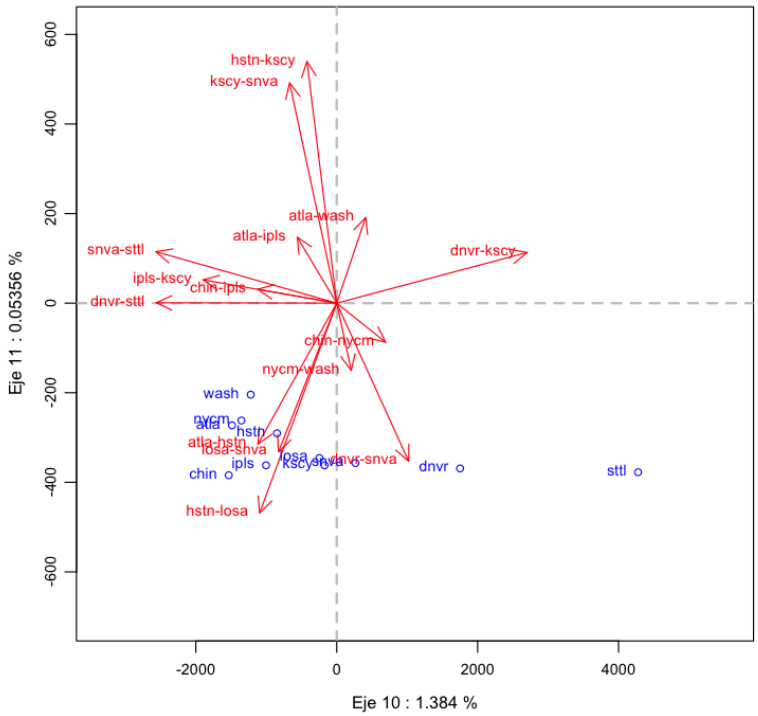
CRFEfil	PC1	PC2	PC1+PC2
ipls	824.9	51.3	876.2
nycm	110.2	657.3	767.4
chin	561.0	131.5	692.5
kscy	332.9	212.3	545.1
wash	11.8	413.0	424.8
atla	2.0	67.1	69.1
snva	14.7	30.8	45.5
dnvr	14.6	22.2	36.7
hstn	1.5	8.1	9.6
losa	0.7	4.7	5.4
sttl	0.3	0.9	1.2

Si representamos el primer plano factorial del HJ-Biplot con solo aquellos marcadores con CRFE>400, que parece un punto de corte natural, obtenemos:



HJ-Biplot, primer plano factorial matriz **Q** de tráfico de la red Abilene, primer intervalo temporal de la matriz **X01** para CRFE de los marcadores respectivos superiores a 400

Representemos ahora el último plano factorial (10-11) que aunque no tendrá gran absorción de inercia (1.44%) sabemos que para este tipo de datos será muy representativa.



Representación HJ-Biplot (último plano factorial 10-11) red Abilene.

La representación HJ-Biplot proporciona una herramienta de visualización que incorpora toda la información de la matriz Laplaciana **L** y de su contraparte de aristas **K**, que como sabemos contienen importantes características de los grafos: diámetro, conectividad, agrupaciones, centralidad, ponderaciones, topologías....

Las propiedades de ambas matrices **L** y **K** pasan por su obtención a partir de la matriz de incidencia vértice-arista **Q**. El cálculo del HJ-Biplot una matriz **X**, al realizarse a partir de la DVS de la matriz **X** y esto a su vez corresponderse con el cálculo de los autovalores y autovectores de $\mathbf{XX}^T$ y $\mathbf{X}^T\mathbf{X}$ resulta en la obtención de las matrices $\mathbf{L}=\mathbf{QQ}^T$ y $\mathbf{K}=\mathbf{Q}^T\mathbf{Q}$ cuando calculamos el HJ-Biplot de la matriz de incidencia **Q**.

La matriz de incidencia **Q** presenta una definición compatible no solo con los tipos de redes/grafos más habituales (sin ponderar, simples, simétricas, no dirigidas,...) sino que permite su obtención para redes dirigidas, asimétricas y ponderadas, lo que encaja perfectamente con los requerimientos más generales que precisa el estudio de los grafos asociados a las redes de telecomunicación más generales que dan soporte, por ejemplo, a la red Internet.

Estas representaciones visuales puede ser perfectamente incorporadas tanto en las fases de gestión de las redes, con datos de monitorización de la red, para la detección de anomalías o modelado, como de diseño de la red, en estudios del tipo "que pasaría si...", en las que se visualizasen los cambios de los marcadores del HJ-Biplot.

8. CONCLUSIONES

8. CONCLUSIONES

De la exhaustiva revisión bibliográfica realizada concluimos:

1. Las técnicas gráficas multivariantes que hoy se reconocen como *Visual Analytics* han sido ampliamente utilizadas en la inspección de tráfico de redes.

2. Se han identificado tres grupos de aplicaciones en las que los métodos estadísticos multivariantes han sido aplicados, detección de anomalías, análisis de series temporales y análisis de topología con datos de tráfico cursado, pero los métodos utilizados no van más allá del Análisis de Componentes Principales y el Análisis de Cluster.

3. Las modernas técnicas multivariantes para la inspección de tablas de dos y de k-vías no han llegado a este campo.

Teniendo en cuenta estos resultados,

4. Proponemos una metodología que combina los métodos Biplot, los métodos Statis y la Teoría de Grafos para la detección y diagnosis de anomalías de volumen en redes Ethernet y en Internet.

5. En el caso de las redes Ethernet se ha detectado mediante la aplicación del HJ-Biplot una tormenta de difusión y en el caso de Internet un ataque de negación de servicio distribuido.

6. La aplicación del HJ-Biplot al estudio de matrices de tráfico ha permitido corregir la problemática presente en los estudios que utilizan el Análisis de Componentes Principales, al permitir identificar la localización espacial de las posibles anomalías y los puntos temporales en las que se producen.

7. El estudio de matrices de tráfico Origen-Destino ha permitido identificar visualmente puntos temporales con valores de interés en los flujos de tráfico bajo estudio, así como posibles correlaciones entre dichos flujos.

8. En el área de análisis espectral de grafos, se ha demostrado la relación existente entre la HJ-bigeometría de la matriz de incidencia de un grafo y el análisis espectral de la matriz Laplaciana de ese mismo grafo.

9. El Análisis Espectral de Grafos nos ha permitido probar que hay información relevante en planos factoriales residuales de la representación HJ-Biplot de la matriz Laplaciana.

10. La representación en coordenadas principales de los marcadores correspondientes a los enlaces entre nodos y las rutas origen-destino, que forman la matriz de enrutamiento, nos ha permitido extraer información operativa sobre las relaciones subyacentes entre rutas y enlaces

11. El análisis espectral de la matriz de adyacencia de la red Abilene nos ha permitido demostrar que en el primer plano factorial de la representación HJ-Biplot se reproduce el orden de la centralidad del autovector de Bonacich.

12. La representación HJ-Biplot de la matriz Laplaciana de la red Abilene ha puesto de manifiesto que el primer eje principal está relacionado directamente con la centralidad del elemento del grafo, bien se trate de un nodo/vértice o enlace/arista

13. El método STATIS permite el análisis conjunto de matrices de tráfico de una red capturadas en diferentes intervalos temporales, identificando aquellos intervalos con perfiles de tráfico anómalos.

14. La metodología propuesta proporciona mejoras en las fases de diseño y de posterior gestión de las principales redes de comunicación actualmente utilizadas incorporando técnicas de analítica visual multivariante a los procedimientos habitualmente empleados.

9. BIBLIOGRAFÍA

8. BIBLIOGRAFÍA

[1] V. E. Al Gore, «Discurso ante la Unión Internacional de Telecomunicaciones». 21-mar-1994.

[2] L. Bernstein y C. M. Yuhas, «Network architectures for the 21st century», *IEEE Communications Magazine*, vol. 34, n.º 1, pp. 24 -28, ene. 1996.

[3] D. O. Awduche, «MPLS and traffic engineering in IP networks», *IEEE Communications Magazine*, vol. 37, n.º 12, pp. 42 -47, dic. 1999.

[4] R. Boutaba, K. El Guemioui, y P. Dini, «An outlook on intranet management», *Communications Magazine, IEEE*, vol. 35, n.º 10, pp. 92–99, 1997.

[5] B. Khasnabish y R. Saracco, «Intranets: technologies, services and management», *Communications Magazine, IEEE*, vol. 35, n.º 10, pp. 84–91, 1997.

[6] H.-G. Hegering, S. Abeck, y R. Wies, «A corporate operation framework for network service management», *Communications Magazine, IEEE*, vol. 34, n.º 1, pp. 62–68, 1996.

[7] L. Z. Granville, D. M. da Rosa, A. Panisson, C. Melchiors, M. J. B. Almeida, y L. M. R. Tarouco, «Managing computer networks using peer-to-peer technologies», *Communications Magazine, IEEE*, vol. 43, n.º 10, pp. 62–68, 2005.

[8] Bruce Boardman, «When It Comes to Statistics, the Devil's in the Details», *Network Magazine*, 14-may-2001.

[9] O. Niggemann, B. Stein, y J. Tolle, «Visualization of traffic structures», en *IEEE International Conference on Communications, 2001. ICC 2001*, 2001, vol. 5, pp. 1516 -1521 vol.5.

[10] D. Rosenbluth y M. Pucci, «Network Perception using Data Imaging and Image Analysis», en *2005 9th IFIP/IEEE International Symposium on Integrated Network Management, 2005. IM 2005*, 2005, pp. 1241-1244.

[11] S. S. Kim y A. L. N. Reddy, «Image-Based Anomaly Detection Technique: Algorithm, Implementation and Effectiveness», *IEEE Journal on Selected Areas in Communications*, vol. 24, n.º 10, pp. 1942-1954, oct. 2006.

[12] T. Samak, A. El-Atawy, E. Al-Shear, y M. Ismail, «A novel visualization approach for efficient network-wide traffic monitoring», en *End-to-End Monitoring Techniques and Services, 2007. E2EMON'07. Workshop on*, 2007, pp. 1–7.

[13] S. Lee y H. Kim, «Correlation, visualization, and usability analysis of routing policy configurations», *IEEE Transactions on Network and Service Management*, vol. 7, n.º 1, pp. 28-41, mar. 2010.

[14] P. Baran, «On Distributed Communications Networks», RAND Corporation, Santa Monica, CA, USA, P-2626, sep. 1962.

[15] P. Baran, «On Distributed Communications Networks», *IEEE Transactions on Communications Systems*, vol. 12, n.º 1, pp. 1-9, 1964.

[16] V. G. Cerf, «In memoriam: Paul Baran», *IEEE Network*, vol. 25, n.º 3, pp. 2-4, 2011.

[17] P. Baran, «The beginnings of packet switching: some underlying concepts», *Communications Magazine, IEEE*, vol. 40, n.º 7, pp. 42–48, 2002.

[18] «"Para que el creador del email me atendiese solo le dije «calabacín»" - Tecnología - ElConfidencial.com». [En línea]. Disponible en: http://www.elconfidencial.com/tecnologia/2013/05/27/para%2Dque%2Del%2Dcreador%2Ddel%2Demail%2Dme%2Datendiese%2Dsolo%2Dle%2Ddije%2Dcalabacin%2D4954/. [Accedido: 19-jun-2013].

[19] A. S. Tanenbaum, *Redes de computadoras*, 4ª ed. México: Pearson Educación, 2003.

[20] O. B. Akan, J. Fang, y I. F. Akyildiz, «TP-Planet: A Reliable Transport Protocol for Interplanetary Internet», *IEEE Journal on Selected Areas in Communications*, vol. 22, n.º 2, pp. 348-361, feb. 2004.

[21] J. D'Ambrosia, «The next generation of Ethernet», *Communications Magazine, IEEE*, vol. 46, n.º 2, pp. S8–S15, 2008.

[22] S. Melle, J. Jaeger, D. Perkins, y V. Vusirikala, «Market drivers and implementation options for 100-GBE transport over the WAN», *Communications Magazine, IEEE*, vol. 45, n.º 11, pp. 18–24, 2007.

[23] A. Meddeb, «Why ethernet WAN transport?», *Communications Magazine, IEEE*, vol. 43, n.º 11, pp. 136–141, 2005.

[24] A. F. Benner, P. K. Pepeljugoski, y R. J. Recio, «A roadmap to 100G ethernet at the Enterprise data center», *Communications Magazine, IEEE*, vol. 45, n.º 11, pp. 10–17, 2007.
[25] C. Spurgeon, *Ethernet configuration guidelines: a quick reference guide to the official Ethernet (IEEE 802.3) configuration rules*. San Jose, Calif: Peer-to-Peer Communications, 1996.
[26] «Infonetics: 10G, 40G, 100G network port shipments up 62% in 2012, revenue to double by 2017». [En línea]. Disponible en: http://www.infonetics.com/pr/2013/2H12-Networking-Ports-Market-Highlights.asp. [Accedido: 18-jun-2013].
[27] X. Liu, H. Wang, y Y. Ji, «Resilient burst ring: extend IEEE 802.17 to WDM networks», *IEEE Communications Magazine*, vol. 46, n.º 11, pp. 74-81, nov. 2008.
[28] M. Batayneh, B. Mukherjee, D. A. Schupke, M. Hoffmann, y A. Kirstaedter, «Carrier-grade ethernet: etherpath protection vs. Ethertunnel protection», *Network, IEEE*, vol. 23, n.º 3, pp. 10–17, 2009.
[29] K. Fouli y M. Maier, «The road to carrier-grade Ethernet», *IEEE Communications Magazine*, vol. 47, n.º 3, pp. S30-S38, 2009.
[30] H. Kuwahara y J. Theodoras, «ET TU, Ethernet/IP?», *Communications Magazine, IEEE*, vol. 44, n.º 11, pp. 74–76, 2006.
[31] J. M. Cioffi, «Mining copper and ether [Technology Leaders Forum]», *Communications Magazine, IEEE*, vol. 45, n.º 6, pp. 18–20, 2007.
[32] J. D'Ambrosia y P. Mooney, «400 Gb/s Ethernet: Why Now?» Ethernet Alliance, abr-2013.
[33] R. Merritt, «The road to 400G Ethernet», 08-abr-2013. [En línea]. Disponible en: http://www.eetimes.com/design/communications-design/4411382/Slideshow--The-road-to-400G-Ethernet. [Accedido: 06-may-2013].
[34] «Event - Bob Metcalfe Leads a Celebration of 40 Years of Ethernet Innovation - PARC, a Xerox company». [En línea]. Disponible en: http://www.parc.com/event/1912/bob-metcalfe-leads-a-celebration-of-40-years-of-ethernet-innovation.html. [Accedido: 18-jun-2013].
[35] R. M. Metcalfe y D. R. Boggs, «Ethernet: distributed packet switching for local computer networks», *Communications of the ACM*, vol. 19, n.º 7, pp. 395–404, 1976.
[36] A. A. Michelson y E. W. Morley, «XXXVI. On the relative motion of the earth and the luminiferous Æther», *The American Journal of Science*, vol. XXXIV, n.º 203, pp. 333-345, nov. 1887.
[37] Hewlett-Packard, «HP Completes Acquisition of 3Com Corporation, Accelerates Converged Infrastructure Strategy». 12-abr-2010.
[38] C. Babla, «Addressing challenges in serial 10 Gb/s multimode fiber enterprise networks», *Communications Magazine, IEEE*, vol. 43, n.º 2, pp. S22–S28, 2005.
[39] Institute of Electrical and Electronics Engineers y IEEE-SA Standards Board, *IEEE standard for information technology - telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements. Part 3, Amendment 4, Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications. Media access control parameters, physical layers, and management parameters for 40 Gb/s and 100 Gb/s operation*. New York: Institute of Electrical and Electronics Engineers, 2010.
[40] Y. J. Won, M.-J. Choi, J. W.-K. Hong, M.-S. Kim, H. Hwang, J.-H. Lee, y S.-G. Lee, «Fault Detection and Diagnosis in IP-Based Mission Critical Industrial Process Control Networks», *IEEE Communications Magazine*, vol. 46, n.º 5, pp. 172-180, may 2008.
[41] R. A. Maxion, «Anomaly detection for diagnosis», en *Fault-Tolerant Computing, 1990. FTCS-20. Digest of Papers., 20th International Symposium*, 1990, pp. 20-27.
[42] F. Feather, D. Siewiorek, y R. Maxion, «Fault detection in an ethernet network using anomaly signature matching», en *ACM SIGCOMM Computer Communication Review*, 1993, vol. 23, pp. 279–288.
[43] L. Kleinrock, «History of the Internet and its flexible future», *Wireless Communications, IEEE*, vol. 15, n.º 1, pp. 8–18, 2008.
[44] L. Kleinrock, «An early history of the internet [History of Communications]», *Communications Magazine, IEEE*, vol. 48, n.º 8, pp. 26–36, 2010.
[45] D. W. Davies, K. A. Bartlett, R. A. Scantlebury, y P. T. Wilkinson, «A digital communication network for computers giving rapid response at remote terminals», en *Proceedings of the first ACM symposium on Operating System Principles*, 1967, pp. 2–1.
[46] S. Loreto, V. K. Gurbani, y J. Ott, «Web-based communications [Guest editorial]», *Communications Magazine, IEEE*, vol. 51, n.º 4, pp. 18–19, 2013.
[47] «GPS Selective Availability». [En línea]. Disponible en: http://www.navcen.uscg.gov/?pageName=gpsSelectiveAvailability. [Accedido: 19-jun-2013].

[48] K. L. Calvert, M. B. Doar, y E. W. Zegura, «Modeling internet topology», *Communications Magazine, IEEE*, vol. 35, n.º 6, pp. 160–163, 1997.
[49] «RFC1918.TXT». .
[50] M. Á. Ruiz-Sánchez, E. W. Biersack, y W. Dabbous, «Survey and taxonomy of IP address lookup algorithms», *Network, IEEE*, vol. 15, n.º 2, pp. 8–23, 2001.
[51] T. Berners-Lee, «Information Management: A Proposal», CERN, mar. 1989.
[52] A. Lakhina, M. Crovella, y C. Diot, «Characterization of network-wide anomalies in traffic flows», en *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, 2004, pp. 201–206.
[53] K. Crisler, T. Turner, A. Aftelak, M. Visciola, A. Steinhage, M. Anneroth, M. Rantzer, B. Von Niman, A. Sasse, M. Tscheligi, y others, «Considering the user in the wireless world», *Communications Magazine, IEEE*, vol. 42, n.º 9, pp. 56–62, 2004.
[54] AENOR Comité AEN/CTN 81, «UNE-EN ISO 13407: Procesos de diseño para sistemas interactivos centrados en el operador humano». AENOR, jun-2000.
[55] R. T. Allen, S. E. Arkin, y R. E. Lawrence, «Moving naval C4I into the next century», *IEEE Communications Magazine*, vol. 33, n.º 10, pp. 96-105, oct. 1995.
[56] K. A. Delic y U. Dayal, Eds., «Network Management Analytics», en *Integrated network management IX: managing new networked worlds: 2005 9th IFIP/IEEE International Symposium on Integrated Network Management (IM 2005)*, Piscataway, NJ, 2005.
[57] H. I. Fahmy y C. Douligeris, «END: an expert network designer», *IEEE Network*, vol. 9, n.º 6, pp. 18-27, nov. 1995.
[58] D. L. Spohn, *Data network design*, 2nd ed. New York: McGraw-Hill, 1997.
[59] A. Kind, X. Dimitropoulos, S. Denazis, y B. Claise, «Advanced network monitoring brings life to the awareness plane», *Communications Magazine, IEEE*, vol. 46, n.º 10, pp. 140–146, 2008.
[60] P. Trimintzios, G. Pavlou, P. Flegkas, P. Georgatsos, A. Asgari, y E. Mykoniati, «Service-driven traffic engineering for intradomain quality of service management», *IEEE Network*, vol. 17, n.º 3, pp. 29-36, may 2003.
[61] L. Dimopoulou, E. Nikolouzou, P. Sampatakos, y L. S. Venieris, «QMTool: an XML-based management platform for QoS-aware IP networks», *IEEE Network*, vol. 17, n.º 3, pp. 8 - 14, jun. 2003.
[62] F. Dupuy, G. Nilsson, y Y. Inoue, «The tina consortium: Toward networking telecommunications information services», *Communications Magazine, IEEE*, vol. 33, n.º 11, pp. 78–83, 1995.
[63] L. Lewis, «Implementing policy in enterprise networks», *Communications Magazine, IEEE*, vol. 34, n.º 1, pp. 50–55, 1996.
[64] Unión Internacional de las Telecomunicaciones, «UIT-T Rec. M.3400 (02/2000) Funciones de gestión de la red de gestión de las telecomunicaciones.», feb. 2000.
[65] R. Boutaba y A. Polyrakis, «Projecting advanced enterprise network and service management to active networks», *Network, IEEE*, vol. 16, n.º 1, pp. 28–33, 2002.
[66] L. B. Ruiz, J. M. Nogueira, y A. A. Loureiro, «Manna: A management architecture for wireless sensor networks», *Communications Magazine, IEEE*, vol. 41, n.º 2, pp. 116–125, 2003.
[67] N. Samaan y A. Karmouch, «Towards Autonomic Network Management: an Analysis of Current and Future Research Directions», *IEEE Communications Surveys & Tutorials*, vol. 11, n.º 3, pp. 22-36, 2009.
[68] A. S. Tanenbaum, *Redes de ordenadores*, 2ª ed. México: Prentice Hall Hispanoamericana, 1991.
[69] D. Alderson, Lun Li, W. Willinger, y J. C. Doyle, «Understanding Internet topology: principles, models, and validation», *IEEE/ACM Transactions on Networking*, vol. 13, n.º 6, pp. 1205-1218, dic. 2005.
[70] S. Lee, K. Levanti, y H. S. Kim, «Network monitoring: Present and future», *Computer Networks*, vol. 65, pp. 84-98, jun. 2014.
[71] D. Applegate y E. Cohen, «Making Routing Robust to Changing Traffic Demands: Algorithms and Evaluation», *IEEE/ACM Transactions on Networking*, vol. 14, n.º 6, pp. 1193-1206, dic. 2006.
[72] A. Callado, C. Kamienski, G. Szabo, B. Gero, J. Kelner, S. Fernandes, y D. Sadok, «A Survey on Internet Traffic Identification», *IEEE Communications Surveys & Tutorials*, vol. 11, n.º 3, pp. 37-52, 2009.
[73] A. Pras, J. Schonwalder, M. Burgess, O. Festor, G. M. Perez, R. Stadler, y B. Stiller, «Key research challenges in network management», *Communications Magazine, IEEE*, vol. 45, n.º 10, pp. 104–110, 2007.
[74] J. Tukey, *Exploratory Data Analysis*. Pearson, 1977.

[75] E. R. Tufte, *The visual display of quantitative information*, 1.ª ed. Cheshire, Conn.: Graphics Press, 1983.
[76] E. R. Tufte, *The visual display of quantitative information*, 2.ª ed. Cheshire, Conn.: Graphics Press, 2001.
[77] F. J. Anscombe, «Graphs in Statistical Analysis», *The American Statistician*, vol. 27, n.º 1, p. 17-21, 1973.
[78] R. A. Becker, S. G. Eick, E. O. Miller, y A. R. Wilks, «Dynamic graphics for network visualization», en *Visualization, 1990. Visualization'90., Proceedings of the First IEEE Conference on*, 1990, pp. 93–96.
[79] V. Paxson, «Growth trends in wide-area TCP connections», *Network, IEEE*, vol. 8, n.º 4, pp. 8-17, 1994.
[80] R. A. Becker, S. G. Eick, y A. R. Wilks, «Visualizing network data», *Visualization and Computer Graphics, IEEE Transactions on*, vol. 1, n.º 1, pp. 16-28, 1995.
[81] R. A. Becker, S. G. Eick, y A. R. Wilks, «Graphical methods to analyze network data», en *Communications, 1993. ICC 93. Geneva. Technical Program, Conference Record, IEEE International Conference on*, 1993, vol. 2, pp. 946–951.
[82] Large Scale Networking y Next Generation Internet Implementation Team, «Next Generation Internet - Implementation Plan Draft». jul-1997.
[83] CAIDA, «ISMA Apr '99 - Report from the ISMA Network Visualization Workshop», UCSD, La Jolla, abr. 1999.
[84] B. Huffaker, E. Nemeth, y k. claffy, «Otter: A general-purpose network visualization tool», en *International Networking Conference (INET) '99*, San Jose, CA, 1999.
[85] Network Associates, *Sniffer*. Network Associates, 1998.
[86] Raytheon, *SilentRunner*. Raytheon Systems Company, 1999.
[87] J. A. Brown, A. J. McGregor, y H. W. Braun, «Network performance visualization: Insight through animation», en *PAM2000 Passive and Active Measurement Workshop, Apr*, 2000, pp. 33–41.
[88] C.-I. W. Team y others, «Internet service performance: Data analysis and visualization», 2000.
[89] T. Munzner, «Interactive visualization of large graphs and networks», Tesis Doctoral, Stanford University, 2000.
[90] R. F. Erbacher, «Visual traffic monitoring and evaluation», en *Proceedings of the Conference on Internet Performance and Control of Network Systems II*, 2001, pp. 153–160.
[91] K. Nyarko, T. Capers, C. Scott, y K. Ladeji-Osias, «Network intrusion visualization with NIVA, an intrusion detection visual analyzer with haptic integration», en *Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2002. HAPTICS 2002. Proceedings. 10th Symposium on*, 2002, pp. 277–284.
[92] D. A. Keim, «Information visualization and visual data mining», *Visualization and Computer Graphics, IEEE Transactions on*, vol. 8, n.º 1, pp. 1–8, 2002.
[93] A. Wood, «Intrusion detection: Visualizing attacks in ids data», *GIAC GCIA Practical*, 2003.
[94] B. K. Sheffler, «Intrusion detection in Depth», *GIAC GCIA Practical*, 2002.
[95] W. Yurcik, K. Lakkaraju, W. Y. K. L. J. Barlow, y J. Rosendale, «A Prototype Tool for Visual Data Mining of Network Traffic for Intrusion Detection», en *Workshop on Data Mining for Computer Security*, 2003, p. 67.
[96] J. Zachary, J. McEachen, y D. Ettlich, «Conversation exchange dynamics for real-time network monitoring and anomaly detection», en *Information Assurance Workshop, 2004. Proceedings. Second IEEE International*, 2004, pp. 59–70.
[97] H. Kim, I. Kang, y S. Bahk, «Real-time visualization of network attacks on high-speed links», *Network, IEEE*, vol. 18, n.º 5, pp. 30–39, 2004.
[98] J. R. Goodall, «User requirements and design of a visualization for intrusion detection analysis», en *Information Assurance Workshop, 2005. IAW'05. Proceedings from the Sixth Annual IEEE SMC*, 2005, pp. 394–401.
[99] J. J. Van Wijk, «The value of visualization», en *Visualization, 2005. VIS 05. IEEE*, 2005, pp. 79–86.
[100] K. Abdullah, C. Lee, G. Conti, y J. A. Copeland, «Visualizing network data for intrusion detection», en *Information Assurance Workshop, 2005. IAW'05. Proceedings from the Sixth Annual IEEE SMC*, 2005, pp. 100–108.
[101] E. R. Gansner, Y. Koren, y S. C. North, «Topological fisheye views for visualizing large graphs», *Visualization and Computer Graphics, IEEE Transactions on*, vol. 11, n.º 4, pp. 457–468, 2005.
[102] R. F. Erbacher, K. Christensen, y A. Sundberg, «Designing visualization capabilities for ids

challenges», en *Visualization for Computer Security, 2005.(VizSEC 05). IEEE Workshop on*, 2005, pp. 121–127.
[103] A. Inselberg, «The plane with parallel coordinates», *The Visual Computer*, vol. 1, n.º 2, pp. 69–91, 1985.
[104] Kemal A. Delic y Umeshwar Dayal, «Network Management Analytics», presentado en 2005 9th IFIP/IEEE International Symposium on Integrated Network Management, 2005. IM 2005, Nice, France, 15-may-2005.
[105] S. Krasser, G. Conti, J. Grizzard, J. Gribschaw, y H. Owen, «Real-time and forensic network data analysis using animated and coordinated visualization», en *Information Assurance Workshop, 2005. IAW'05. Proceedings from the Sixth Annual IEEE SMC*, 2005, pp. 42–49.
[106] J. J. Thomas y K. A. Cook, *Illuminating the path: the research and development agenda for visual analytics*, 1st ed. Los Alamitos, CA: IEEE, 2005.
[107] D. A. Keim, M. Sips, y M. Ankerst, «Visual Data-Mining Techniques», en *The visualization handbook*, Burlington, MA: Elsevier Butterworth-Heinemann, 2005, pp. 831-843.
[108] C. D. Hansen y C. R. Johnson, Eds., *The visualization handbook*. Amsterdam ; Boston: Elsevier-Butterworth Heinemann, 2005.
[109] D. A. Keim, F. Mansmann, J. Schneidewind, y T. Schreck, «Monitoring network traffic with radial traffic analyzer», en *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, 2006, pp. 123–128.
[110] T. Kisner, A. Essoh, y F. Kaderali, «Visualisation of network traffic using dynamic co-occurrence matrices», en *Internet Monitoring and Protection, 2007. ICIMP 2007. Second International Conference on*, 2007, pp. 7–7.
[111] M. Oka, Y. Oyama, H. Abe, y K. Kato, «Anomaly detection using layered networks based on eigen co-occurrence matrix», en *Recent Advances in Intrusion Detection*, 2004, pp. 223–237.
[112] F. Mansmann, D. A. Keim, S. C. North, B. Rexroad, y D. Sheleheda, «Visual analysis of network traffic for resource planning, interactive monitoring, and interpretation of security threats», *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, n.º 6, pp. 1105–1112, 2007.
[113] M. Withall, I. Phillips, y D. Parish, «Network visualisation: a review», *IET Communications*, vol. 1, n.º 3, p. 365, 2007.
[114] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, y H. Ziegler, «Visual analytics: Scope and challenges», *Lecture Notes in Computer Science (LNCS)*. Springer, pp. 76-90, 2008.
[115] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, y G. Melan\ccon, «Visual analytics: Definition, process, and challenges», *Lecture Notes in Computer Science (LNCS)*. Springer, pp. 154-175, 2008.
[116] A. L. da S. Kauer, B. S. Meiguins, R. M. C. do Carmo, M. de B. Garcia, y A. S. G. Meiguins, «An Information Visualization Tool with Multiple Coordinated Views for Network Traffic Analysis», 2008, pp. 151-156.
[117] L. Han y J. van Hemert, «A Novel Visual Discriminator for Network Traffic Patterns», 2008, pp. 141-146.
[118] T. Nayak, A. Neogi, y R. Kothari, «Visualization and Analysis of System Monitoring Data using Multi-resolution Context Information», *IEEE Transactions on Network and Service Management*, vol. 5, n.º 3, pp. 168-177, sep. 2008.
[119] T. Kohonen, «The self-organizing map», *Proceedings of the IEEE*, vol. 78, n.º 9, pp. 1464–1480, 1990.
[120] T. Kohonen, *Self-Organizing Maps*, 3rd ed., vol. 30. Springer, 2001.
[121] E. M. Salvador y L. Z. Granville, «An investigation of visualization techniques for SNMP traffic traces», en *Network Operations and Management Symposium, 2008. NOMS 2008. IEEE*, 2008, pp. 887–890.
[122] E. M. Salvador y L. Z. Granville, «Using visualization techniques for snmp traffic analyses», en *Computers and Communications, 2008. ISCC 2008. IEEE Symposium on*, 2008, pp. 806–811.
[123] H. Tong, S. Papadimitriou, J. Sun, P. S. Yu, y C. Faloutsos, «Colibri: fast mining of large static and dynamic graphs», en *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 686–694.
[124] M. W. Mahoney y P. Drineas, «CUR matrix decompositions for improved data analysis», *Proceedings of the National Academy of Sciences*, vol. 106, n.º 3, pp. 697-702, ene. 2009.
[125] J. Sun, Y. Xie, H. Zhang, y C. Faloutsos, «Less is more: Compact matrix decomposition for large sparse graphs», 2007.
[126] J. R. Goodall, «Visualization is better! a comparative evaluation», en *Visualization for Cyber Security, 2009. VizSec 2009. 6th International Workshop on*, 2009, pp. 57–68.

[127] D. W. H. Ten, S. Manickam, S. Ramadass, y H. A. A. Bazar, «Study on Advanced Visualization Tools In Network Monitoring Platform», 2009, pp. 445-449.
[128] R. Fontugne, T. Hirotsu, y K. Fukuda, «A visualization tool for exploring multi-scale network traffic anomalies», en *Performance Evaluation of Computer & Telecommunication Systems, 2009. SPECTS 2009. International Symposium on*, 2009, vol. 41, pp. 274–281.
[129] Y. Jin, E. Sharafuddin, y Z.-L. Zhang, «Unveiling core network-wide communication patterns through application traffic activity graph decomposition», en *Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems*, 2009, pp. 49–60.
[130] M. Iliofotou, P. Pappu, M. Faloutsos, M. Mitzenmacher, S. Singh, y G. Varghese, «Network monitoring using traffic dispersion graphs (tdgs)», en *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 2007, pp. 315–320.
[131] C. Ding, T. Li, W. Peng, y H. Park, «Orthogonal nonnegative matrix t-factorizations for clustering», 2006, p. 126.
[132] H. Kim, I. Lee, J. Cho, y J. Moon, «Visualization of network components for attack analysis», en *Computational Intelligence in Cyber Security, 2009. CICS'09. IEEE Symposium on*, 2009, pp. 1–8.
[133] H. Read, A. Blyth, y I. Sutherland, «A unified approach to network traffic and network security visualisation», en *Communications, 2009. ICC'09. IEEE International Conference on*, 2009, pp. 1–6.
[134] A. Shahrestani, M. Feily, R. Ahmad, y S. Ramadass, «Architecture for Applying Data Mining and Visualization on Network Flow for Botnet Traffic Detection», 2009, pp. 33-37.
[135] M. P. Singh y N. Subramanian, «Visualization of flow data based on clustering technique for identifying network anomalies», en *Industrial Electronics & Applications, 2009. ISIEA 2009. IEEE Symposium on*, 2009, vol. 2, pp. 973–978.
[136] A. Jin, D. Hwang, G. Kim, H. Chang, S. Lee, y H. Choi, «Abnormal pattern detection based on visualization», en *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, 2010, vol. 4, pp. 363–366.
[137] P. T. Barbosa y L. Z. Granville, «Interactive SNMP traffic analysis through information visualization», en *Network Operations and Management Symposium (NOMS), 2010 IEEE*, 2010, pp. 73–79.
[138] M. Celenk, T. Conley, J. Willis, y J. Graham, «Predictive Network Anomaly Detection and Visualization», *IEEE Transactions on Information Forensics and Security*, vol. 5, n.º 2, pp. 288-299, jun. 2010.
[139] N. Promrit, A. Mingkhwan, S. Simcharoen, y N. Namvong, «Multi-dimensional visualization for network forensic analysis», en *Networked Computing (INC), 2011 The 7th International Conference on*, 2011, pp. 68–73.
[140] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, y D. W. Fellner, «Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges», *Computer Graphics Forum*, vol. 30, n.º 6, pp. 1719-1749, sep. 2011.
[141] J. R. Goodall, F. Mansmann, y J. Gerth, «Computer network visualization [Guest Editorial]», *Network, IEEE*, vol. 26, n.º 6, pp. 4–5, 2012.
[142] L. Harrison y A. Lu, «The future of security visualization: Lessons from network visualization», *Network, IEEE*, vol. 26, n.º 6, pp. 6–11, 2012.
[143] M. Dumas, J.-M. Robert, y M. J. McGuffin, «Alertwheel: radial bipartite graph visualization applied to intrusion detection system alerts», *Network, IEEE*, vol. 26, n.º 6, pp. 12–18, 2012.
[144] M. Couture, F. Massicotte, H. Normandin, y M. Letourneau, «Navigating and visualizing the malware intelligence space», *Network, IEEE*, vol. 26, n.º 6, pp. 19–25, 2012.
[145] S. Knight, N. Falkner, H. X. Nguyen, P. Tune, y M. Roughan, «I can see for miles: Re-visualizing the internet», *Network, IEEE*, vol. 26, n.º 6, pp. 26–32, 2012.
[146] E. Biersack, Q. Jacquemart, F. Fischer, J. Fuchs, O. Thonnard, G. Theodoridis, D. Tzovaras, y P.-A. Vervier, «Visual analytics for BGP monitoring and prefix hijacking identification», *Network, IEEE*, vol. 26, n.º 6, pp. 33–39, 2012.
[147] U. Sedlar, M. Volk, J. Sterle, A. Kos, y R. Sernec, «Contextualized monitoring and root cause discovery in IPTV systems using data visualization», *Network, IEEE*, vol. 26, n.º 6, pp. 40–46, 2012.
[148] J. Davey, F. Mansmann, J. Kohlhammer, y D. Keim, «Visual analytics: towards intelligent interactive internet and security solutions», en *The Future Internet*, Springer, 2012, pp. 93–104.
[149] M. J. McGuffin, «Simple algorithms for network visualization: A tutorial», *Tsinghua Science and Technology*, vol. 17, n.º 4, pp. 383–398, 2012.

[150] Y. Liu, W. Chen, y Y. Guan, «Monitoring Traffic Activity Graphs with low-rank matrix approximation», 2012, pp. 59-67.
[151] J. Pfeffer, «Fundamentals of visualizing communication networks», *Communications, China*, vol. 10, n.º 3, pp. 82–90, 2013.
[152] J. L. Moreno, *Who Shall Survive?*, vol. 58. Nervous and Mental Disease Publishing Co., 1934.
[153] X. Hu, A. Lu, y X. Wu, «Spectrum-Based Network Visualization for Topology Analysis», *Computer Graphics and Applications, IEEE*, vol. 33, n.º 1, pp. 58-68, ene. 2013.
[154] G.-D. Sun, Y.-C. Wu, R.-H. Liang, y S.-X. Liu, «A Survey of Visual Analytics Techniques and Applications: State-of-the-Art Research and Future Challenges», *Journal of Computer Science and Technology*, vol. 28, n.º 5, pp. 852-867, sep. 2013.
[155] P. Joia, F. V. Paulovich, D. Coimbra, J. A. Cuminato, y L. G. Nonato, «Local affine multidimensional projection», *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, n.º 12, pp. 2563–2571, 2011.
[156] C. Turkay, P. Filzmoser, y H. Hauser, «Brushing dimensions-a dual visual analysis model for high-dimensional data», *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, n.º 12, pp. 2591–2599, 2011.
[157] K. R. Gabriel, «The biplot graphic display of matrices with application to principal component analysis», *Biometrika*, vol. 58, n.º 3, pp. 453–467, 1971.
[158] C. Eckart y G. Young, «A principal axis transformation for non-Hermitian matrices», *Bulletin of the American Mathematical Society*, vol. 45, n.º 2, pp. 118–121, 1939.
[159] C. Eckart y G. Young, «The approximation of one matrix by another of lower rank», *Psychometrika*, vol. 1, n.º 3, pp. 211-218, 1936.
[160] M. P. Galindo-Villardon, «Una alternativa de representación simultánea: HJ-Biplot», *Qüestiió*, vol. 10, n.º 1, pp. 13-23, mar. 1986.
[161] K. R. Gabriel, *Biplot Display of Multivariate Matrices for Inspection of Data and Diagnosis*. Defense Technical Information Center, 1980.
[162] N. Baccala, «Contribuciones al Análisis de Matrices de Datos Multivia», Universidad de Salamanca, 2004.
[163] Z. del C. Osuna Marín, «Contribuciones al Análisis de datos Textuales», Universidad de Salamanca, 2006.
[164] M. Marcos Hidalgo, «HJ-Biplot Aumentado». 2011.
[165] E. Frutos, M. P. Galindo, y V. Leiva, «An interactive biplot implementation in R for modeling genotype-by-environment interaction», *Stochastic Environmental Research and Risk Assessment*, vol. 28, n.º 7, pp. 1629-1641, oct. 2014.
[166] A. Orfao, M. Gonzalez, J. F. San Miguel, M. C. Cañizo, P. Galindo, M. D. Caballero, R. Jimenez, y A. L. Borrasca, «Clinical and immunological findings in large B-cell chronic lymphocytic leukemia», *Clinical immunology and immunopathology*, vol. 46, n.º 2, pp. 177–185, 1988.
[167] E. Galante, M. García-Román, I. Barrera, y P. Galindo, «Comparison of spatial distribution patterns of dung-feeding scarabs (Coleoptera: Scarabaeidae, Geotrupidae) in wooded and open pastureland in the Mediterranean Dehesa area of the Iberian Peninsula», *Environmental Entomology*, vol. 20, n.º 1, pp. 90–97, 1991.
[168] C. Santos, S. S. Munoz, Y. Gutierrez, E. Hebrero, J. L. Vicente, P. Galindo, y J. C. Rivas, «Characterization of young red wines by application of HJ biplot analysis to anthocyanin profiles», *Journal of Agricultural and food chemistry*, vol. 39, n.º 6, pp. 1086–1090, 1991.
[169] J. C. Rivas-Gonzalo, Y. Gutierrez, A. M. Polanco, E. Hebrero, J. L. Vicente, P. Galindo, y C. Santos-Buelga, «Biplot Analysis Applied to Enological Parameters in the Geographical Classification of Young Red Wines», *Am. J. Enol. Vitic.*, vol. 44, n.º 3, pp. 302-308, ene. 1993.
[170] K. Hron, M. Jelínková, P. Filzmoser, R. Kreuziger, P. Bednář, y P. Barták, «Statistical analysis of wines using a robust compositional biplot», *Talanta*, vol. 90, pp. 46-50, feb. 2012.
[171] J. Garcia-Talegon, M. A. Vicente, E. Molina-Ballesteros, y S. Vicente-Tavera, «Determination of the origin and evolution of building stones as a function of their chemical composition using the inertia criterion based on an HJ-biplot», *Chemical Geology*, vol. 153, n.º 1-4, pp. 37-51, ene. 1999.
[172] A. Iñigo, F. López-Moro, S. Vicente-Tavera, y V. Rives, «Monitoring of Origin and Evolution of Building Stones through Their Major Components», *Journal of Materials in Civil Engineering*, vol. 17, n.º 4, pp. 440-446, 2005.
[173] M. J. Varas, S. Vicente-Tavera, E. Molina, y J. L. Vicente-Villardón, «Role of canonical biplot method in the study of building stones: an example from Spanish monumental heritage», *Environmetrics*, vol. 16, n.º 4, pp. 405-419, jun. 2005.

[174] G. Correa Londoño, L. L. Lavalett Oñate, M. P. Galindo Villardón, y L. Afanador Kafuri, «Use of multivariate methods for grouping strains of colletotrichum spp. Based on Cultural and morphological characters», *Revista Facultad Nacional de Agronomía, Medellín*, vol. 60, n.º 1, pp. 3671–3690, 2007.
[175] C. Theoharatos, G. Economou, y S. Fotopoulos, «Semantic Mapping of Image Databases using Perceptual Similarity», en *Image Analysis for Multimedia Interactive Services, 2007. WIAMIS'07. Eighth International Workshop on*, 2007, pp. 43–43.
[176] W. Yan, M. S. Kang, B. Ma, S. Woods, y P. L. Cornelius, «GGE Biplot vs. AMMI Analysis of Genotype-by-Environment Data», *Crop Science*, vol. 47, n.º 2, p. 643, 2007.
[177] C. Theoharatos, S. Fotopoulos, E. Panagiotopoulos, I. Boniatis, y G. Panayiotakis, «Visual organization of hip joint osteoathritis data in low-dimensional biplots», en *IEEE International Workshop on Imaging Systems and Techniques, 2008. IST 2008*, 2008, pp. 187-192.
[178] N. Cukrov, N. Tepic, D. Omanovic, S. Lojen, E. Bura-Nakic, V. Vojvodic, y I. Pizeta, «Anthropogenic and natural influences on the Krka River (Croatia) evaluated by multivariate statistical analysis», en *Information Technology Interfaces, 2009. ITI'09. Proceedings of the ITI 2009 31st International Conference on*, 2009, pp. 219–224.
[179] S. Mendes, M. J. Fernández-Gómez, M. P. Galindo-Villardón, F. Morgado, P. Maranhão, U. M. Azeiteiro, y P. Bacelar-Nicolau, «The study of bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) by applying the HJ-biplot method», 2009.
[180] F. M. Afendi, L. K. Darusman, A. Hirai, M. Altaf-Ul-Amin, H. Takahashi, K. Nakamura, y S. Kanaya, «System Biology Approach for Elucidating the Relationship Between Indonesian Herbal Plants and the Efficacy of Jamu», 2010, pp. 661-668.
[181] E. Castela y P. Galindo, «Ecological Inference for the characterization of electoral turnout: The Portuguese Case», CIEO-Research Centre for Spatial and Organizational Dynamics, University of Algarve, 2010.
[182] P. V. Vicente Galindo, T. de N. Vaz, y P. Nijkamp, «Institutional capacity to dynamically innovate: An application to the Portuguese case», *Technological Forecasting and Social Change*, vol. 78, n.º 1, pp. 3-12, ene. 2011.
[183] I.-M. García-Sánchez, J.-V. Frías-Aceituno, y L. Rodríguez-Domínguez, «Determinants of corporate social disclosure in Spanish local governments», *Journal of Cleaner Production*, vol. 39, pp. 60-72, ene. 2013.
[184] L. Costa y P. Oliveira, «Biplots in offline multiobjective reduction», en *Evolutionary Computation (CEC), 2010 IEEE Congress on*, 2010, pp. 1–8.
[185] G. Costagliola y V. Fuccella, «CyBiS: A Novel Interface for Searching Scientific Documents», 2011, pp. 276-281.
[186] I. Gallego-Álvarez y J. L. Vicente-Villardón, «Analysis of environmental indicators in international companies by applying the logistic biplot», *Ecological Indicators*, vol. 23, pp. 250-261, dic. 2012.
[187] «The STATIS Method». SAGE Publications, Thousand Oaks, Calif, 2007.
[188] H. Abdi, L. J. Williams, D. Valentin, y M. Bennani-Dosse, «STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling», *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, n.º 2, pp. 124-167, mar. 2012.
[189] H. L'Hermier des Plantes, «Structuration des tableaux à trois indices de la statistique: théorie et application d'une méthode d'analyse conjointe», Université des sciences et techniques du Languedoc, [Montpellier], 1976.
[190] P.-A. Jaffrenou, «Sur l'analyse des familles finies de variables vectorielles: bases algébriques et application à la description statistique», s.n.], S.l., 1978.
[191] C. Lavit, *Analyse conjointe de tableaux quantitatifs*. 1988.
[192] J. Thioulouse y D. Chessel, «Les analyses multitableaux en ecologie factorielle. I : de la typologie d'etat a la typologie de fonctionnement par l'analyse triadique», *Acta Oecologica Oecologia Generalis*, vol. 8, pp. 463-480, 1987.
[193] I. Stanimirova, B. Walczak, D. L. Massart, V. Simeonov, C. A. Saby, y E. Di Crescenzo, «STATIS, a three-way method for data analysis. Application to environmental data», *Chemometrics and Intelligent Laboratory Systems*, vol. 73, n.º 2, pp. 219-233, oct. 2004.
[194] H. A. Kiers, «Towards a standardized notation and terminology in multiway analysis», *Journal of chemometrics*, vol. 14, n.º 3, pp. 105–122, 2000.
[195] Y. Escoufier, «Le Traitement des Variables Vectorielles», *Biometrics*, vol. 29, n.º 4, p. 751, dic. 1973.
[196] J. B. Evanowsky, «Information for the Warrior», *Communications Magazine, IEEE*, vol. 33, n.º 10, pp. 106–112, 1995.

[197] M. D. Wonacott, K. Eluthesen, y R. S. Braudy, «Distributed enterprise information networking: A case example», *IEEE Communications Magazine*, vol. 34, n.º 1, pp. 38 -43, ene. 1996.
[198] L. L. Ho, D. J. Cavuto, S. Papavassiliou, y A. G. Zawadzki, «Adaptive and automated detection of service anomalies in transaction-oriented WANs: network analysis, algorithms, implementation, and deployment», *Selected Areas in Communications, IEEE Journal on*, vol. 18, n.º 5, pp. 744–757, 2000.
[199] M. Thottan y C. Ji, «Proactive anomaly detection using distributed intelligent agents», *Network, IEEE*, vol. 12, n.º 5, pp. 21–27, 1998.
[200] V. Chandola, A. Banerjee, y V. Kumar, «Anomaly detection: A survey», *ACM Comput. Surv.*, vol. 41, n.º 3, pp. 15:1–15:58, jul. 2009.
[201] T. J. Pincince, D. Goodtree, y C. Barth, «Network Strategy Service: The Full Service Intranet», *The Forrester Report*, vol. 10, n.º 4, mar. 1996.
[202] «The birth of the web». [En línea]. Disponible en: http://home.web.cern.ch/about/birth-web. [Accedido: 12-jun-2013].
[203] I. Katzela y M. Schwartz, «Schemes for fault identification in communication networks», *Networking, IEEE/ACM Transactions on*, vol. 3, n.º 6, pp. 753–764, 1995.
[204] C. S. Hood y C. Ji, «Proactive network fault detection», en *, Proceedings IEEE INFOCOM '97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution*, 1997, vol. 3, pp. 1147 -1155 vol.3.
[205] M. Thottan y C. Ji, «Statistical detection of enterprise network problems», *Journal of Network and Systems Management*, vol. 7, n.º 1, pp. 27–45, 1999.
[206] P. Akritas, P. G. Akishin, I. Antoniou, A. Y. Bonushkina, I. Drossinos, V. V. Ivanov, Y. L. Kalinovsky, V. V. Korenkov, y P. V. Zrelov, «Nonlinear analysis of network traffic», *Chaos, Solitons & Fractals*, vol. 14, n.º 4, pp. 595–606, 2002.
[207] P. Barford, J. Kline, D. Plonka, y A. Ron, «A signal analysis of network traffic anomalies», en *Internet Measurement Conference: Proceedings of the 2 nd ACM SIGCOMM Workshop on Internet measurement*, 2002, vol. 6, pp. 71–82.
[208] C. Manikopoulos y S. Papavassiliou, «Network intrusion and fault detection: a statistical anomaly approach», *Communications Magazine, IEEE*, vol. 40, n.º 10, pp. 76–82, 2002.
[209] I. Antoniou, V. V. Ivanov, V. V. Ivanov, y P. V. Zrelov, «On a statistical model of network traffic», *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 502, n.º 2-3, pp. 768-771, abr. 2003.
[210] N. Golyandina, V. Nekrutkin, y A. A. Zhigljavsky, *Analysis of Time Series Structure: SSA and related techniques*, 1.ª ed. Chapman and Hall/CRC, 2001.
[211] M. Shyu, S. Chen, K. Sarinnapakorn, y L. Chang, «A novel anomaly detection scheme based on principal component classifier», en *in Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM'03*, 2003, pp. 172–179.
[212] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, y N. Taft, «Structural analysis of network traffic flows», en *Proceedings of the joint international conference on Measurement and modeling of computer systems*, New York, NY, USA, 2004, pp. 61–72.
[213] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, y N. Taft, «Analysis of Origin Destination Flows (Raw Data)». Boston University, nov-2003.
[214] O. Bouhaddou, C. H. Obled, y T. P. Dinh, «Principal component analysis and interpolation of stochastic processes: methods and simulation», *Journal of Applied Statistics*, vol. 14, n.º 3, pp. 251-267, ene. 1987.
[215] A. Lakhina, M. Crovella, y C. Diot, «Diagnosing network-wide traffic anomalies», en *ACM SIGCOMM Computer Communication Review*, 2004, vol. 34, pp. 219–230.
[216] A. Lakhina, M. Crovella, y C. Diot, «Mining anomalies using traffic feature distributions», en *ACM SIGCOMM Computer Communication Review*, 2005, vol. 35, pp. 217–228.
[217] Y. Tang, E. S. Al-Shaer, y R. Boutaba, «Active integrated fault localization in communication networks», en *Integrated Network Management, 2005. IM 2005. 2005 9th IFIP/IEEE International Symposium on*, 2005, pp. 543–556.
[218] Y. Tang, E. Al-Shaer, y R. Boutaba, «Efficient fault diagnosis using incremental alarm correlation and active investigation for internet and overlay networks», *IEEE Transactions on Network and Service Management*, vol. 5, n.º 1, pp. 36-49, mar. 2008.
[219] G. Androulidakis, V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, y V. Maglaris, «Understanding and evaluating the impact of sampling on anomaly detection techniques», en *Military Communications Conference, 2006. MILCOM 2006. IEEE*, 2006, pp. 1–7.
[220] S. Farraposo, P. Owezarski, y E. Monteiro, «Contribution of anomalies detection and analysis on

traffic engineering», en *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, 2006, pp. 1–2.

[221] J. Jiang y S. Papavassiliou, «Enhancing network traffic prediction and anomaly detection via statistical network traffic separation and combination strategies», *Computer Communications*, vol. 29, n.º 10, pp. 1627-1638, jun. 2006.

[222] R. Kwitt y U. Hofman, «Robust Methods for Unsupervised PCA-based Anomaly Detection», en *Proceedings of the IEEE / IST Workshop on Monitoring, Attack Detection and Mitigation*, Tuebingen, Germany, 2006.

[223] E. Kim y S. Kim, «Anomaly detection in network security based on nonparametric techniques», en *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, 2006, pp. 1–2.

[224] S. Agrawal, K. V. M. Naidu, y R. Rastogi, «Diagnosing link-level anomalies using passive probes», en *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, 2007, pp. 1757–1765.

[225] Y. Vardi, «Network tomography: Estimating source-destination traffic intensities from link data», *Journal of the American Statistical Association*, vol. 91, n.º 433, pp. 365-377, 1996.

[226] G. Munz y G. Carle, «Real-time analysis of flow data for network attack detection», en *Integrated Network Management, 2007. IM'07. 10th IFIP/IEEE International Symposium on*, 2007, pp. 100–108.

[227] H. Ringberg, A. Soule, J. Rexford, y C. Diot, «Sensitivity of PCA for traffic anomaly detection», *ACM SIGMETRICS Performance Evaluation Review*, vol. 35, n.º 1, pp. 109–120, 2007.

[228] T. Ahmed, M. Coates, y A. Lakhina, «Multivariate online anomaly detection using kernel recursive least squares», en *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, 2007, pp. 625–633.

[229] T. Ahmed, S. Ahmed, S. Ahmed, y M. Motiwala, «Real-Time Intruder Detection in Surveillance Networks Using Adaptive Kernel Methods», en *Communications (ICC), 2010 IEEE International Conference on*, 2010, pp. 1–5.

[230] L. Huang, X. L. Nguyen, M. Garofalakis, J. M. Hellerstein, M. I. Jordan, A. D. Joseph, y N. Taft, «Communication-efficient online detection of network-wide anomalies», en *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, 2007, pp. 134–142.

[231] P. Du, S. Abe, Y. Ji, S. Sato, y M. Ishiguro, «Detecting and tracing traffic volume anomalies in SINET3 backbone network», en *Communications, 2008. ICC'08. IEEE International Conference on*, 2008, pp. 5833–5837.

[232] P. Li, M. Salour, y X. Su, «A survey of internet worm detection and containment», *Communications Surveys & Tutorials, IEEE*, vol. 10, n.º 1, pp. 20–35, 2008.

[233] K. V. M. Naidu, D. Panigrahi, y R. Rastogi, «Detecting anomalies using end-to-end path measurements», en *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, 2008, pp. 1849–1857.

[234] S. S. Kim y A. L. N. Reddy, «Statistical Techniques for Detecting Traffic Anomalies Through Packet Header Data», *IEEE/ACM Transactions on Networking*, vol. 16, n.º 3, pp. 562-575, jun. 2008.

[235] N. Samaan y A. Karmouch, «Network anomaly diagnosis via statistical analysis and evidential reasoning», *IEEE Transactions on Network and Service Management*, vol. 5, n.º 2, pp. 65-77, jun. 2008.

[236] P. Chhabra, C. Scott, E. D. Kolaczyk, y M. Crovella, «Distributed spatial anomaly detection», en *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, 2008, pp. 1705–1713.

[237] H. Sun, Y. Zhaung, y H. J. Chao, «A principal components analysis-based robust DDoS defense system», en *Communications, 2008. ICC'08. IEEE International Conference on*, 2008, pp. 1663–1669.

[238] I. C. Paschalidis y G. Smaragdakis, «Spatio-Temporal Network Anomaly Detection by Assessing Deviations of Empirical Measures», *IEEE/ACM Transactions on Networking*, vol. 17, n.º 3, pp. 685-697, jun. 2009.

[239] P. Velarde-Alvarado, C. Vargas-Rosales, D. Torres-Roman, y A. Martinez-Herrera, «Detecting anomalies in network traffic using the method of remaining elements», *IEEE Communications Letters*, vol. 13, n.º 6, pp. 462-464, jun. 2009.

[240] A. Botta, A. Dainotti, A. Pescape, y G. Ventre, «Reducing Network Traffic Data Sets», en *Communications, 2007. ICC'07. IEEE International Conference on*, 2007, pp. 350–356.

[241] P. Velarde-Alvarado, C. Vargas-Rosales, D. Torres-Roman, y A. Martinez-Herrera, «An Architecture for Intrusion Detection Based on an Extension of the Method of Remaining

Elements», *Journal of applied research and technology*, vol. 8, n.º 2, pp. 159–174, 2010.
[242] Yi Xie y Shun-Zheng Yu, «Monitoring the Application-Layer DDoS Attacks for Popular Websites», *IEEE/ACM Transactions on Networking*, vol. 17, n.º 1, pp. 15-25, feb. 2009.
[243] I. T. Jolliffe, *Principal component analysis*, 2nd ed. New York: Springer, 2002.
[244] A. Kind, M. Stoecklin, y X. Dimitropoulos, «Histogram-based traffic anomaly detection», *IEEE Transactions on Network and Service Management*, vol. 6, n.º 2, pp. 110-121, jun. 2009.
[245] T. Qin, X. Guan, W. Li, y P. Wang, «Monitoring abnormal traffic flows based on independent component analysis», en *Communications, 2009. ICC'09. IEEE International Conference on*, 2009, pp. 1–5.
[246] Y. Himura, K. Fukuda, K. Cho, y H. Esaki, «An automatic and dynamic parameter tuning of a statistics-based anomaly detection algorithm», en *Communications, 2009. ICC'09. IEEE International Conference on*, 2009, pp. 1–6.
[247] C.-K. Han y H.-K. Choi, «Effective discovery of attacks using entropy of packet dynamics», *Network, IEEE*, vol. 23, n.º 5, pp. 4–12, 2009.
[248] G. Androulidakis, V. Chatzigiannakis, y S. Papavassiliou, «Network anomaly detection and classification via opportunistic sampling», *Network, IEEE*, vol. 23, n.º 1, pp. 6–12, 2009.
[249] N. Duffield, P. Haffner, B. Krishnamurthy, y H. Ringberg, «Rule-based anomaly detection on IP flows», en *INFOCOM 2009, IEEE*, 2009, pp. 424–432.
[250] D. Brauckhoff, K. Salamatian, y M. May, «Applying PCA for traffic anomaly detection: Problems and solutions», en *INFOCOM 2009, IEEE*, 2009, pp. 2866–2870.
[251] A. Abdelkefi, Y. Jiang, W. Wang, A. Aslebo, y O. Kvittem, «Robust traffic anomaly detection with principal component pursuit», en *Proceedings of the ACM CoNEXT Student Workshop*, 2010, p. 10.
[252] E. J. Candès, X. Li, Y. Ma, y J. Wright, «Robust principal component analysis?», *arXiv preprint arXiv:0912.3599*, 2009.
[253] E. J. Candès, X. Li, Y. Ma, y J. Wright, «Robust principal component analysis?», *J. ACM*, vol. 58, n.º 3, pp. 11:1–11:37, jun. 2011.
[254] Q. Ke y T. Kanade, «Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming», en *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 1, pp. 739–746.
[255] F. Raimir Holanda y J. E. B. Maia, «Network traffic prediction using PCA and K-means», en *Network Operations and Management Symposium (NOMS), 2010 IEEE*, 2010, pp. 938–941.
[256] W. E. Leland, M. S. Taqqu, W. Willinger, y D. V. Wilson, «On the self-similar nature of Ethernet traffic (extended version)», *IEEE/ACM Transactions on Networking*, vol. 2, n.º 1, pp. 1-15, 1994.
[257] V. Paxson y S. Floyd, «Wide area traffic: the failure of Poisson modeling», *IEEE/ACM Transactions on Networking*, vol. 3, n.º 3, pp. 226-244, 1995.
[258] Y. Kanda, K. Fukuda, y T. Sugawara, «Evaluation of anomaly detection based on sketch and pca», en *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, 2010, pp. 1–5.
[259] C. Callegari, L. Gazzarrini, S. Giordano, M. Pagano, y T. Pepe, «A Novel PCA-Based Network Anomaly Detection», en *2011 IEEE International Conference on Communications (ICC)*, 2011, pp. 1-5.
[260] K. Nyalkalkar, S. Sinhay, M. Bailey, y F. Jahanian, «A comparative study of two network-based anomaly detection methods», en *INFOCOM, 2011 Proceedings IEEE*, 2011, pp. 176–180.
[261] G. Cormode, F. Korn, S. Muthukrishnan, y D. Srivastava, «Diamond in the rough: finding hierarchical heavy hitters in multi-dimensional data», en *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, 2004, pp. 155–166.
[262] Y. Zhang, S. Singh, S. Sen, N. Duffield, y C. Lund, «Online identification of hierarchical heavy hitters: algorithms, evaluation, and applications», en *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, 2004, pp. 101–114.
[263] G. Thatte, U. Mitra, y J. Heidemann, «Parametric Methods for Anomaly Detection in Aggregate Traffic», *IEEE/ACM Transactions on Networking*, vol. 19, n.º 2, pp. 512-525, abr. 2011.
[264] C. Pascoal, M. R. de Oliveira, R. Valadas, P. Filzmoser, P. Salvador, y A. Pacheco, «Robust feature selection and robust PCA for Internet traffic anomaly detection», en *INFOCOM, 2012 Proceedings IEEE*, 2012, pp. 1755–1763.
[265] Z. Wang, K. Hu, K. Xu, B. Yin, y X. Dong, «Structural analysis of network traffic matrix via relaxed principal component pursuit», *Computer Networks*, vol. 56, n.º 7, pp. 2049-2067, may 2012.
[266] P. Szilagyi y S. Novaczki, «An Automatic Detection and Diagnosis Framework for Mobile

Communication Systems», *IEEE Transactions on Network and Service Management*, vol. 9, n.º 2, pp. 184-197, jun. 2012.
[267] M. Roughan, Yin Zhang, W. Willinger, y Lili Qiu, «Spatio-Temporal Compressive Sensing and Internet Traffic Matrices (Extended Version)», *IEEE/ACM Transactions on Networking*, vol. 20, n.º 3, pp. 662-676, jun. 2012.
[268] F. J. Delgado Alvarez, «Una aplicación del método HJ-Biplot a la diagnósis de redes de área local basadas en la norma Ethernet (ISO 8802-3)», en *XXV Congreso Nacional de Estadística e Investigación Operativa*, Vigo, 2000, pp. 289-290.
[269] F. J. Delgado y P. Galindo, «Detección de un ataque de negación de servicio distribuido (DDoS) basada en STATIS», en *XI Conferencia española y primer encuentro iberoamericano de biometría: CEIB*, Salamanca, 2007, pp. 179-180.
[270] R. Jiang, H. Fei, y J. Huan, «A Family of Joint Sparse PCA Algorithms for Anomaly Localization in Network Data Streams», *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, n.º 11, pp. 2421-2433, 2013.
[271] R. Bro y A. K. Smilde, «Centering and scaling in component analysis», *Journal of Chemometrics*, vol. 17, n.º 1, pp. 16-33, ene. 2003.
[272] R. Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2012.
[273] S. Dray, A.-B. Dufour, y others, «The ade4 package: implementing the duality diagram for ecologists», *Journal of statistical software*, vol. 22, n.º 4, pp. 1–20, 2007.
[274] «MIT Lincoln Laboratory: Cyber Systems & Technology: DARPA Intrusion Detection». [En línea]. Disponible en: http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/index.html. [Accedido: 25-jun-2013].
[275] «MIT Lincoln Laboratory: Communication Systems and Cyber Security: Cyber Systems and Technology: DARPA Intrusion Detection Evaluation». [En línea]. Disponible en: http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/2000data.html. [Accedido: 25-jun-2013].
[276] «CERT Incident Note IN-2000-05: mstream Distributed DoS». [En línea]. Disponible en: http://www.cert.org/incident_notes/IN-2000-05.html. [Accedido: 29-jun-2013].
[277] David Dittrich, George Weaver, Sven Dietrich, y Neil Long, «The "mstream" distributed denial of service attack tool». [En línea]. Disponible en: http://staff.washington.edu/dittrich/misc/mstream.analysis.txt. [Accedido: 29-jun-2013].
[278] S. Savage, D. Wetherall, A. Karlin, y T. Anderson, «Network support for IP traceback», *Networking, IEEE/ACM Transactions on*, vol. 9, n.º 3, pp. 226–237, 2001.
[279] A. Yaar, A. Perrig, y D. Song, «StackPi: New Packet Marking and Filtering Mechanisms for DDoS and IP Spoofing Defense», *IEEE Journal on Selected Areas in Communications*, vol. 24, n.º 10, pp. 1853-1863, oct. 2006.
[280] «MIT Lincoln Laboratory: Communication Systems and Cyber Security: Cyber Systems and Technology: DARPA Intrusion Detection Evaluation: LLS-DDOS 2.0.2». [En línea]. Disponible en: http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/2000/LLS_DDOS_2.0.2.html. [Accedido: 29-jun-2013].
[281] «TCPDUMP/LIBPCAP public repository». [En línea]. Disponible en: http://www.tcpdump.org/. [Accedido: 25-jun-2013].
[282] A. V. Oppenheim, A. S. Willsky, y I. T. Young, *Signals and systems*. London [u.a.: Prentice Hall, 1983.
[283] D. Middleton, IEEE Communications Society, y IEEE Information Theory Society, *An introduction to statistical communication theory*. Piscataway, NJ: IEEE Press, 1996.
[284] G. Stefanou, «The stochastic finite element method: Past, present and future», *Computer Methods in Applied Mechanics and Engineering*, vol. 198, n.º 9-12, pp. 1031-1051, feb. 2009.
[285] M. Kirby y L. Sirovich, «Application of the Karhunen-Loeve procedure for the characterization of human faces», *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, n.º 1, pp. 103–108, 1990.
[286] K. Pearson, «On lines and planes of closest fit to systems of points in space», *Philosophical Magazine*, vol. 2, n.º 6, pp. 559-572, 1901.
[287] H. Hotelling, «Analysis of a complex of statistical variables into principal components.», *Journal of educational psychology*, vol. 24, n.º 6, p. 417, 1933.
[288] A. Papoulis y S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th Revised edition. McGraw Hill Higher Education, 2002.

[289] J.-C. Deville, «Méthodes statistiques et numériques de l'analyse harmonique», en *Annales de l'INSEE*, 1974, pp. 3–101.
[290] G. Matheron, *The Theory of Regionalized Variables and Its Applications*. École national supérieure des mines, 1971.
[291] D. B. Chua, E. D. Kolaczyk, y M. Crovella, «Network Kriging», *IEEE Journal on Selected Areas in Communications*, vol. 24, n.º 12, pp. 2263-2272, dic. 2006.
[292] D. Dong, P. Fang, Y. Bock, F. Webb, L. Prawirodirdjo, S. Kedar, y P. Jamason, «Spatiotemporal filtering using principal component analysis and Karhunen-Loeve expansion approaches for regional GPS network analysis», *Journal of Geophysical Research*, vol. 111, n.º B3, 2006.
[293] K. V. M. Fernando y H. Nicholson, «Karhunen-Loève expansion with reference to singular-value decomposition and separation of variables», *Control Theory and Applications, IEE Proceedings D*, vol. 127, n.º 5, pp. 204 -206, sep. 1980.
[294] J. J. Gerbrands, «On the relationships between SVD, KLT and PCA», *Pattern Recognition*, vol. 14, n.º 1-6, pp. 375-381, ene. 1981.
[295] Jianbo Gao, Yinhe Cao, Wen-wen Tung, y Jing Hu, *Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond*, 2007.ª ed. John Wiley & Sons. Inc., 2007.
[296] E. Biglieri y K. Yao, «Some properties of singular value decomposition and their applications to digital signal processing», *Signal Processing*, vol. 18, n.º 3, pp. 277-289, 1989.
[297] K. Ozeki, «A coordinate-free theory of eigenvalue analysis related to the method of principal components and the Karhunen—Loève expansion», *Information and Control*, vol. 42, n.º 1, pp. 38–59, 1979.
[298] E. Stone y A. Cutler, «Archetypal analysis of spatio-temporal dynamics», *Physica D: Nonlinear Phenomena*, vol. 90, n.º 3, pp. 209-224, feb. 1996.
[299] M. D. Graham y I. G. Kevrekidis, «Alternative approaches to the Karhunen-Loève decomposition for model reduction and data analysis», *Computers & Chemical Engineering*, vol. 20, n.º 5, pp. 495-506, may 1996.
[300] A. J. Newman, «Model reduction via the Karhunen-Loeve expansion part i: an exposition», Institute for Systems Research, University of Maryland, Technical Research Report 96-32, 1996.
[301] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Segunda. San Diego, CA, USA: Academic Press Professional, Inc., 1990.
[302] K. Fukunaga y W. L. Koontz, «Application of the Karhunen-Lo? ve Expansion to Feature Selection and Ordering», *IEEE Transactions on Computers*, vol. 19, n.º 4, pp. 311–318, 1970.
[303] G. H. Golub y C. F. Van Loan, *Matrix computations*, Fourth edition. Baltimore: The Johns Hopkins University Press, 2013.
[304] J. Egido, «Biplot Dinámico», Universidad de Salamanca, Salamanca, 2015.
[305] Yin Zhang, «Abilene Traffic Matrix Dataset 2004», 22-mar-2015. [En línea]. Disponible en: http://www.cs.utexas.edu/~yzhang/research/AbileneTM/. [Accedido: 22-mar-2015].
[306] U. Brandes y T. Erlebach, «Cap. 2 Fundamentals», en *Network analysis methodological foundations*, Berlin; New York: Springer, 2005.
[307] S. G. Eick, «Scalable Network Visualization», en *The Visualization Handbook*, Burlington, MA: Elsevier Butterworth-Heinemann, 2005, p. 819.
[308] C. R. Palmer, P. B. Gibbons, y C. Faloutsos, «Data Mining on Large Graphs», en *Dynamic Social Network Modeling and Analysis workshop summary and papers*, Washington, D.C.: National Academies Press, 2003, pp. 265-288.
[309] V. M. Vairinhos, V. Lobo, y P. G. Villardón, «Data Analysis with Intersection Graphs», *Procedia Computer Science*, vol. 18, pp. 60-69, 2013.
[310] A. Seary y W. Richards, «Spectral methods for analyzing and visualizing networks: an introduction», en *Dynamic Social Network Modeling and Analysis workshop summary and papers*, Washington, D.C.: National Academies Press, 2003, pp. 209–228.
[311] F. R. K. Chung, *Spectral Graph Theory*. Providence, R.I: American Mathematical Society, 1996.
[312] D. Chakrabarti y C. Faloutsos, *Graph Mining: Laws, Tools, and Case Studies*, Edición: New. San Rafael, Calif.: Morgan & Claypool Publishers, 2012.
[313] G. Ackerman, «Modeling Terrorists», *Spectrum, IEEE*, vol. 43, n.º 9, pp. 26–34, 2006.
[314] D. Chakrabarti, «Tools for large graph mining», DTIC Document, 2005.
[315] Y. Koren, L. Carmel, y D. Harel, «Drawing huge graphs by algebraic multigrid optimization», *Multiscale Modeling & Simulation*, vol. 1, n.º 4, pp. 645–673, 2003.
[316] X.-D. Zhang y R. Luo, «The Laplacian eigenvalues of mixed graphs», *Linear Algebra and its Applications*, vol. 362, pp. 109-119, mar. 2003.
[317] M. Fiedler, «Algebraic connectivity of graphs», *Czechoslovak Mathematical Journal*, vol. 23, n.º

2, pp. 298–305, 1973.
[318] Andreas Baltz y Lasse Kliemann, «Cap. 14 Spectral Analysis», en *Network analysis methodological foundations*, Berlin; New York: Springer, 2005.
[319] K. C. Das, «The Laplacian spectrum of a graph», *Computers & Mathematics with Applications*, vol. 48, n.º 5-6, pp. 715-724, sep. 2004.
[320] H. Zha, X. He, C. Ding, H. Simon, y M. Gu, «Bipartite Graph Partitioning and Data Clustering», en *Proceedings of the Tenth International Conference on Information and Knowledge Management*, New York, NY, USA, 2001, pp. 25–32.
[321] R. Merris, «Laplacian matrices of graphs: a survey», *Linear Algebra and its Applications*, vol. 197-198, pp. 143-176, ene. 1994.
[322] D. M. Cvetković, *Eigenspaces of graphs*. Cambridge ; New York: Cambridge University Press, 1997.
[323] F. Kammer y H. Täubig, «Cap. 7 Connectivity», en *Network analysis methodological foundations*, Berlin; New York: Springer, 2005.
[324] R. L. Breiger, «Emergent Themes in Social Network Analysis: Results, Challenges, Opportunities», en *Dynamic Social Network Modeling and Analysis workshop summary and papers*, Washington, D.C.: National Academies Press, 2003, pp. 19-38.
[325] R. Grone, R. Merris, y V. S. Sunder, «The Laplacian spectrum of a graph», *SIAM Journal on Matrix Analysis and Applications*, vol. 11, n.º 2, pp. 218–238, 1990.
[326] B. Mohar y Y. Alavi, «The Laplacian spectrum of graphs», *Graph theory, combinatorics, and applications*, vol. 2, pp. 871–898, 1991.
[327] J. L. Gross y J. Yellen, *Graph Theory and Its Applications, Second Edition (Discrete Mathematics and Its Applications)*. Chapman & Hall/CRC, 2005.
[328] H. Poincaré, *Papers on Topology: Analysis Situs and Its Five Supplements*, vol. 37. American Mathematical Soc., 2010.
[329] J. Van Den Heuvel y S. Pejić, «Using Laplacian eigenvalues and eigenvectors in the analysis of frequency assignment problems», *Annals of Operations Research*, vol. 107, n.º 1-4, pp. 349–368, 2001.
[330] S. Sarkar y A. Dong, «Community detection in graphs using singular value decomposition», *Physical Review E*, vol. 83, n.º 4, p. 046114, 2011.
[331] M. E. Newman, «Finding community structure in networks using the eigenvectors of matrices», *Physical review E*, vol. 74, n.º 3, p. 036104, 2006.
[332] Dirk Koschützki, Katharina Anna Lehmann, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl, y Oliver Zlotowski, «Cap. 3 Centrality Indices», en *Network analysis methodological foundations*, Berlin; New York: Springer, 2005.
[333] G. Kirchhoff, «On the solution of the equations obtained from the investigation of the linear distribution of galvanic currents», *Circuit Theory, IRE Transactions on*, vol. 5, n.º 1, pp. 4–7, 1958.
[334] R. Grone, «On the geometry and Laplacian of a graph», *Linear Algebra and its Applications*, vol. 150, pp. 167-178, may 1991.
[335] R. Agaev y P. Chebotarev, «On the spectra of nonsymmetric Laplacian matrices», *Linear Algebra and its Applications*, vol. 399, pp. 157-168, abr. 2005.
[336] B. Mohar, «Laplace eigenvalues of graphs—a survey», *Discrete Mathematics*, vol. 109, n.º 1-3, pp. 171-183, nov. 1992.
[337] J. Kunegis y A. Lommatzsch, «Learning spectral graph transformations for link prediction», en *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 561–568.
[338] M. Fiedler, «A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory», *Czechoslovak Mathematical Journal*, vol. 25, n.º 4, pp. 619-633, 1975.
[339] S. Guattery, «Graph embeddings, symmetric real matrices, and generalized inverses», DTIC Document, 1998.
[340] S. Guattery y G. Miller, «Graph Embeddings and Laplacian Eigenvalues», *SIAM. J. Matrix Anal. & Appl.*, vol. 21, n.º 3, pp. 703-723, ene. 2000.
[341] P. Barooah y J. P. Hespanha, «Graph effective resistance and distributed control: Spectral properties and applications», en *Decision and control, 2006 45th IEEE conference on*, 2006, pp. 3479–3485.
[342] S. Y. Shafi, M. Arcak, y L. E. Ghaoui, «Designing node and edge weights of a graph to meet Laplacian eigenvalue constraints», en *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, 2010, pp. 1016–1023.
[343] J. Gallier, «Notes on Elementary Spectral Graph Theory. Applications to Graph Clustering

Using Normalized Cuts», *arXiv preprint arXiv:1311.2492*, 2013.
[344] D. Zelazo y M. Bürger, «On the definiteness of the weighted Laplacian and its connection to effective resistance», en *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, 2014, pp. 2895–2900.
[345] S. Sivasubramanian, «Average distance in graphs and eigenvalues», *Discrete Mathematics*, vol. 309, n.º 10, pp. 3458-3462, may 2009.
[346] O. Knill, «Cauchy-Binet for Pseudo-Determinants», *arXiv:1306.0062 [math]*, may 2013.
[347] M. Fiedler, «Laplacian of graphs and algebraic connectivity», *Banach Center Publications*, vol. 25, n.º 1, pp. 57–70, 1989.
[348] D. B. Chua, E. D. Kolaczyk, y M. Crovella, «Efficient Monitoring of End-to-End Network Properties», en *In Proc. of IEEE INFOCOM*, Miami, EEUU, 2005, pp. 1701–1711.
[349] D. Chua, E. D. Kolaczyk, y M. Crovella, «A Statistical Framework for Efficient Monitoring of End-to-end Network Properties», en *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, Banff, Alberta, Canada, 2005, pp. 390–391.
[350] E. D. Kolaczyk, *Statistical analysis of network data: methods and models*. New York ; [London]: Springer, 2009.
[351] A. Cayley, «A theorem in the geometry position», *Cambridge Mathematical Journal*, vol. II, pp. 267-271, 1841.
[352] J. P. Benzécri, *L'analyse des données: L'analyse des correspondances*. Dunod, 1973.
[353] M. J. Greenacre, *La Práctica del análisis de correspondencias*. Barcelona: Fundación BBVA, 2008.
[354] S. P. Borgatti y M. G. Everett, «Network analysis of 2-mode data», *Social Networks*, vol. 19, n.º 3, pp. 243-269, ago. 1997.
[355] W. Richards y A. Seary, «Eigen analysis of networks», *Journal of Social Structure*, vol. 1, n.º 2, pp. 1–17, 2000.
[356] S. P. Borgatti, «Centrality and network flow», *Social Networks*, vol. 27, n.º 1, pp. 55-71, ene. 2005.
[357] E. J. Bienenstock y P. Bonacich, «Balancing Efficiency and Vulnerability in Social Networks», en *Dynamic Social Network Modeling and Analysis workshop summary and papers*, Washington, D.C.: National Academies Press, 2003.
[358] Jefatura del Estado, *Ley 8/2011, de 28 de abril, por la que se establecen medidas para la protección de las infraestructuras críticas.*, vol. 102. 2011, p. 43370.
[359] Ministerio del Interior, *Real Decreto 704/2011, de 20 de mayo, por el que se aprueba el Reglamento de protección de las infraestructuras críticas.*, vol. 121. 2011, p. 50808.
[360] S. P. Borgatti, «The Key Player Problem», en *Dynamic social network modeling and analysis: Workshop summary and papers*, 2003, p. 241.
[361] B. A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, y C. Faloutsos, «Got the flu (or mumps)? check the eigenvalue!», *arXiv preprint arXiv:1004.0060*, 2010.
[362] P. Drineas, A. Frieze, R. Kannan, S. Vempala, y V. Vinay, «Clustering large graphs via the singular value decomposition», *Machine learning*, vol. 56, n.º 1-3, pp. 9–33, 2004.
[363] J. M. Kleinberg, «Authoritative sources in a hyperlinked environment», *Journal of the ACM (JACM)*, vol. 46, n.º 5, pp. 604–632, 1999.
[364] D. Fay, H. Haddadi, A. Thomason, A. W. Moore, R. Mortier, A. Jamakovic, S. Uhlig, y M. Rio, «Weighted Spectral Distribution for Internet Topology Analysis: Theory and Applications», *IEEE/ACM Transactions on Networking*, vol. 18, n.º 1, pp. 164-176, feb. 2010.
[365] Y. Shavitt, X. Sun, A. Wool, y B. Yener, «Computing the unmeasured: an algebraic approach to Internet mapping», en *IEEE INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings*, 2001, vol. 3, pp. 1646-1654 vol.3.
[366] Y. Chen, D. Bindel, y R. H. Katz, «Tomography-based overlay network monitoring», en *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, 2003, pp. 216–231.
[367] Y. Shavitt, X. Sun, A. Wool, y B. Yener, «Computing the Unmeasured: An Algebraic Approach to Internet Mapping», *IEEE Journal on Selected Areas in Communications*, vol. 22, n.º 1, pp. 67-78, ene. 2004.
[368] N. Cressie, *Statistics for Spatial Data*, Revised Edition edition. New York: Wiley-Interscience, 1993.
[369] K. Buza y I. Galambos, «An Application of Link Prediction in Bipartite Graphs: Personalized Blog Feedback Prediction», en *8th Japanese-Hungarian Symposium on Discrete Mathematics and Its Applications June 4-7, 2013, Veszprém, Hungary*, 2013, p. 8.

[370] S. P. Borgatti y M. G. Everett, «A Graph-theoretic perspective on centrality», *Social Networks*, vol. 28, n.º 4, pp. 466-484, oct. 2006.
[371] Dirk Koschützki, Katharina Anna Lehmann, Dagmar Tenfelde-Podehl, y Oliver Zlotowski, «Cap. 5 Advanced Centrality Concepts», en *Network analysis methodological foundations*, Berlin; New York: Springer, 2005.
[372] U. Brandes y T. Erlebach, *Network analysis methodological foundations*. Berlin; New York: Springer, 2005.
[373] A. Y. Ng, A. X. Zheng, y M. I. Jordan, «Link analysis, eigenvectors and stability», en *International Joint Conference on Artificial Intelligence*, 2001, vol. 17, pp. 903–910.
[374] P. Bonacich, «Factoring and weighting approaches to status scores and clique identification», *The Journal of Mathematical Sociology*, vol. 2, n.º 1, pp. 113-120, ene. 1972.
[375] P. Bonacich, «Simultaneous group and individual centralities», *Social Networks*, vol. 13, n.º 2, pp. 155-168, jun. 1991.
[376] P. Bonacich, «Some unique properties of eigenvector centrality», *Social Networks*, vol. 29, n.º 4, pp. 555-564, oct. 2007.
[377] P.-Y. Chen y A. O. Hero, «Local Fiedler vector centrality for detection of deep and overlapping communities in networks», en *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 1120–1124.
[378] X. Qi, E. Fuller, Q. Wu, Y. Wu, y C.-Q. Zhang, «Laplacian centrality: A new centrality measure for weighted networks», *Information Sciences*, vol. 194, pp. 240-253, jul. 2012.
[379] E. D. Kolaczyk, *Abilene Datasets*. 2009.
[380] P. Bonacich, «Power and Centrality: A Family of Measures», *American Journal of Sociology*, vol. 92, n.º 5, pp. 1170-1182, mar. 1987.
[381] yWorks, *yEd Graph Editor*. yWorks GmbH, 2014.