Métodos y fuentes en la sociología computacional: La exploración de los *big data* bibliográficos y de bibliotecas mediante análisis de redes

Luis Martínez Uribe
Departamento de Sociología y Comunicación

Tesis para el grado de
Doctor

Director de tesis
D. Rafael Modesto Escobar Mercado
2022

La presente tesis titulada “Métodos y fuentes para la sociología computacional: La exploración de los *big data* bibliográficos y de bibliotecas mediante análisis de redes” ha sido financiada por la Fundación Juan March. Además, ha contado con financiación de los siguientes proyectos de investigación del Ministerio de Ciencia, Innovación y Universidades con fondos del Programa FEDER de la Unión Europea por los proyectos de referencia CSO2013-49278-EXP y PGC2018-093755-B-I0.

El contenido de esta tesis doctoral se corresponde a un compendio de los siguientes trabajos previamente publicados:

Modesto, Escobar; Martínez-Uribe, Luis (2020). “Network Coincidence Analysis: The netCoin R Package.” *Journal of Statistical Software*, 93 (11),1-32. doi: https://doi.org/10.18637/jss.v093.i11

Martínez-Uribe, Luis (2018). “Digital Archives as Big Data.” *Mathematical Population Studies*, 26(2), 69-79. doi: https://doi.org/10.1080/08898480.2017.1418116

Martínez-Uribe, Luis (2022). “La sociología a través de sus publicaciones en revistas de impacto mediante el uso de big data.” *Empiria. Revista de Metodología en Ciencias Sociales. (53), 53-89. doi:* https://doi.org/10.5944/empiria.53.2022.32612

D. Rafael Modesto Escobar Mercado, Catedrático de Sociología de la Universidad de Salamanca.

CERTIFICA: Que el trabajo doctoral por compendio de artículos o publicaciones realizado bajo su dirección por D. Luis Martínez Uribe, titulado “Métodos y fuentes para la sociología computacional La exploración de los *big data* bibliográficos y de bibliotecas mediante análisis de redes” reúne las condiciones de originalidad requeridas para optar al grado de Doctor en Ciencias Sociales por la Universidad de Salamanca.

Y para que así conste, firma la presente certificación en Salamanca, a 11de junio de 2022.

Agradecimientos

Con la presentación de esta tesis alcanzo un hito y termino una etapa tras la que considero importante hacer una reflexión sobre las motivaciones que me llevaron a embarcarme en el doctorado, el proceso de investigación conducido durante estos seis años y el resultado final obtenido.

Empezar una tesis a los cuarenta años y compatibilizarlo con un trabajo a tiempo completo y una familia con niños es un ejercicio que requiere de una motivación especial. Esta motivación surge de la educación que me han dado mis padres. Ellos han sabido fomentar mi formación y me han inculcado el valor del esfuerzo y el trabajo a través de su propio ejemplo. Realmente, la idea de hacer un doctorado empiezo a considerarla tras terminar el máster en Londres en 2007, pero nunca surgía el momento adecuado. Ya en Madrid, Paz Fernández, amiga y directora durante diez años en la Biblioteca de la Fundación Juan March, me animó a que me embarcase en el proyecto para alimentar mi curiosidad por la investigación y disponer del título de doctor de cara a futuro. En 2015 Modesto Escobar me invita a que realice la tesis bajo su dirección en el departamento de Sociología y Comunicación de la Universidad de Salamanca, la idea finalmente cristaliza y la Fundación Juan March generosamente financia los estudios. Estos años de doctorando han estado llenos de momentos especiales donde he podido disfrutar de dedicar tiempo a leer acerca de temáticas apasionantes, he colaborado en el desarrollado de herramientas de software y metodologías analíticas que además he podido aplicar a diversos conjuntos de datos, he asistido a congresos y talleres presentando los avances y obteniendo el comentario de otros investigadores para finalmente publicar los tres artículos requeridos en revistas académicas especializadas. Estoy contento con el resultado final, aunque quizás me hubiera gustado haber realizado una estancia en algún centro de investigación en el extranjero para centrarme únicamente en la investigación de la tesis durante algún tiempo, pero la pandemia lo hizo imposible.

Por supuesto, la realización de la tesis no habría sido posible sin haber contado con el apoyo de personas a las estoy inmensamente agradecido. Ante todo, quiero agradecer de manera especial a mis padres su amor incondicional, su dedicación absoluta y su ayuda constante. Sé que se alegran tanto o más que yo por la consecución de este doctorado. A Modesto que es mucho más que un director de tesis. Es un amigo del que he aprendido mucho sobre cómo llevar adelante un trabajo de investigación, pero que sobre todo es un ejemplo de generosidad y pasión, entrega y compromiso con la labor de investigación y docencia. A Paz, por empujarme a hacerlo, por hacer posible que pudiese compatibilizar la vida laboral, la de estudiante y la familiar sin volverme loco, por ayudarme con innumerables revisiones de todo tipo y por animarme e inspirarme en todo momento de forma tan cariñosa. A Pablo Cabrera, compañero de doctorado con el que he compartido muy buenos ratos durante los congresos y viajes varios. A los compañeros de la Biblioteca por su interés en los avances y sus continuas palabras de apoyo. A la Fundación Juan March por financiar el doctorado y alentar mi formación personal. A mi mujer Ruth y a mis hijos por ser comprensivos y apoyarme para seguir adelante a pesar de los planes sacrificados durante fines de semana y vacaciones. Espero que para Jaime y Alicia sirva de inspiración para conseguir lo que se propongan a base de trabajo y tesón. Al resto de mi familia, a todos los amigos y compañeros de trabajo que se han interesado y con los que he podido desconectar y disfrutar de otras cosas durante todo este tiempo. Para acabar, me gustaría dedicar de forma especial esta tesis a mi amigo Stuart Macdonald que nos dejó en 2020, al cual echo mucho de menos y recuerdo a menudo.

Métodos y fuentes para la sociología computacional: La exploración de los *big data* bibliográficos y de bibliotecas mediante análisis de redes

Resumen

Los orígenes de la sociología computacional, según Cioffi-Revilla (2014), se remontan a la invención de la computadora digital tras la Segunda Guerra Mundial aunque es sin duda la aparición del fenómeno *big data* lo que realmente ha supuesto un punto de inflexión comparable a los momentos de revolución científica y cambio de paradigma de Khun (1962). Desde hace unos años la sociología se ha relacionado con las nuevas fuentes de datos denominados *big data* de una manera excepcional, incorporándose en multitud de estudios de diversa índole, combinando las técnicas estadísticas habituales con nuevas metodologías cuantitativas de la ciencia de datos y la inteligencia artificial. Esta nueva subdisciplina de la sociología tiene ante sí retos sustanciales como son el acceso a los nuevos datos que, en la mayoría de las ocasiones pertenecen a empresas privadas, o la formación tecnológica necesaria de los sociólogos computacionales que les permitan almacenar, consultar, limpiar, explorar y analizar grandes conjuntos de datos.

El presente trabajo de investigación en formato de tesis por compendio de publicaciones se inscribe en la tarea de aportar a la sociología computacional metodologías y herramientas que faciliten el trabajo con grandes cantidades de información. Para ello esta tesis intenta descubrir y aplicar herramientas y estrategias novedosas con las que enfrentarse a los *big data* y conseguir de ellos una correcta interpretación. Además, esta tesis tiene por objetivo ilustrar a través de ejemplos cómo los sociólogos pueden disponer y tratar los datos generados y gestionados por las bibliotecas y los archivos. Estas instituciones preservan la memoria que posibilita un análisis de las circunstancias políticas, económicas, culturales y sociales a lo largo de la historia. Para ello, esta tesis se compone de tres artículos: el primer artículo presenta la herramienta netCoin de código abierto para la exploración de estructuras de datos que implementa el análisis de coincidencias y que combina análisis multivariante y de redes. En el segundo artículo se trata con la herramienta netCoin una colección de datos abiertos de 3 millones de registros bibliográficos de la *British Library*. El tercer artículo analiza la disciplina sociológica en los últimos años, a partir de los 300 millones de publicaciones de los datos de *Microsoft Academic Graph*, aplicando varias metodologías y estrategias que permiten reducir la dimensionalidad.

Los resultados de las investigaciones realizadas ofrecen varias conclusiones. En primer lugar, el análisis de coincidencias aporta un método de detección de patrones aplicable a colecciones de grandes y pequeños datos que ayuda a entenderlos mejor. Por otro lado, la herramienta desarrollada ofrece la utilización de redes interactivas que permiten realizar análisis exploratorios y presentar resultados finales. Además, se han presentado varias técnicas que reducen la dimensionalidad de los datos convirtiéndolos en datos más sencillos de manejar y analizar. Finalmente, los datos bibliográficos y de bibliotecas proporcionan una fuente de calidad y de fácil acceso para los sociólogos computacionales.

Palabras clave: sociología computacional, *big data*, datos bibliográficos, datos de bibliotecas, métodos cuantitativos, análisis de coincidencias, análisis exploratorios, análisis de redes sociales, técnicas de reducción de dimensionalidad

Tabla de contenidos

1.Introducción

Las ciencias sociales se originan en la era precomputacional, siendo Auguste Comte una de las primeras figuras que, durante el siglo XVII, habla de una ciencia natural de los sistemas sociales (Cioffi-Revilla 2014). Desde entonces, y durante años, las ciencias sociales han desarrollado un corpus extenso de conocimiento sobre los fenómenos sociales a través del establecimiento de subdisciplinas, conocidas como *Big Five,* que incluyen a la antropología, la economía, la ciencia política, la psicología y la sociología (Bernard 2012, Horowitz 2006). Khun (1962) establece que las disciplinas científicas se forman y evolucionan a través de un compendio compartido de teorías, métodos y problemas a resolver. De forma periódica - Khun argumenta - se suceden los momentos de revolución científica, momentos de cambio de paradigma, en los que nuevas formas de pensamiento o herramientas desafían los enfoques tradicionales.

A comienzos de este siglo Savage y Burrows (2007) alertaron en un artículo de gran impacto acerca de la crisis potencial que se cernía sobre la sociología empírica al darse cuenta de que en el ámbito privado se estaban llevado a cabo estudios sociológicos utilizando nuevas fuentes de información. Esas fuentes de información contenían enormes cantidades de datos transaccionales registrados por compañías privadas en el ejercicio de su actividad. Los sociólogos ni siquiera estaban considerando estos datos como fuente para su investigación acostumbrados a realizar complejos, y muchas veces costosos, estudios de encuesta para capturar información. A través de ese artículo Savage y Burrows (2007) hicieron una llamada a reconceptualizar la sociología empírica, fomentando la relación e interés con las fuentes de datos que estaban apareciendo y mediante una crítica a las prácticas de captura y uso de los datos sociales. Unos años más tarde Burrows y Savage (2014), se dieron cuenta de que realmente se habían topado con el fenómeno del *big data,* al cual, sin saberlo, se habían referido en su publicación original. Los datos transaccionales a los que hacían alusión eran solo una parte de un espectro mayor datos que eran ya una realidad. Y quizás, como planteaba Kitchin (2014b), propiciando un cambio de paradigma en multitud de disciplinas científicas.

Desde entonces la sociología, y las ciencias sociales en su conjunto, se han relacionado con estas nuevas fuentes de datos de una manera excepcional. Se han analizado los beneficios potenciales y los posibles problemas metodológicos y epistemológicos (Aragona 2022, Boyd y Crawford 2012). Las nuevas fuentes de datos se han incorporado a las fuentes habituales en multitud de estudios de diversa índole combinando las técnicas estadísticas habituales con nuevas metodologías cuantitativas de la ciencia de datos y la inteligencia artificial (Edelmann et al. 2020). Así, se han dedicado a la temática números especiales de revistas[1] e incluso revistas especializadas[2], se han incorporado como materia específica en los planes de estudios de grado y

[1] Como ejemplo dos de los artículos de esta tesis son parte de números de revistas dedicados al fenómeno del *big data*. El artículo *Digital Archives as Big Data* es parte de *Methods for Big Data in Social Science* en la revista *Mathematical Population Studies* *https://www.tandfonline.com/toc/gmps20/26/2* y el artículo La sociología a través de sus publicaciones en revista de impacto mediante el uso de big data forma parte del número El Big Data en las ciencias sociales de la revista Empiria http://revistas.uned.es/index.php/empiria

[2] Algunos ejemplos para resaltar incluyen la revista *Big Data & Society* *https://journals.sagepub.com/home/bds o la revista Journal of Computational Social Science* https://www.springer.com/journal/42001

de postgrado[3] e incluso se han creado departamentos dedicados al tema[4]. Toda esta actividad frenética en pocos años ha dado lugar a una nueva rama de la sociología denominada sociología computacional, un campo emergente, interdisciplinar, que por medio de nuevas herramientas y fuentes de datos se extiende a la sociología tradicional (Edelmann et al. 2020, Evans y Foster 2019, Gualda 2022a).

La sociología computacional tiene ante sí retos sustanciales. Uno de ellos es el acceso a estos nuevos datos. Cuando un sociólogo pretende utilizar estas fuentes, ha de enfrentarse a complicaciones derivadas del acceso a los datos. Los datos de las fuentes habituales de *big data* son en muchas ocasiones propiedad de empresas privadas y para poder disponer de ellos se requiere algún tipo de asociación con las empresas, lo cual limita el acceso a grupos de investigación con cierto estatus (King y Persily 2019). A nivel técnico y metodológico los *big data* presentan problemáticas tales como el almacenamiento y la consulta, el filtrado de la información útil, la reducción de la dimensionalidad o la manera en que se pueden realizar análisis exploratorios que permitan detectar aquellas relaciones intrínsecas y patrones que puedan esconderse en los datos. Por tanto, es fundamental la formación tecnológica apropiada de los sociólogos computacionales, imprescindible para que puedan trabajar con herramientas con las que almacenar, consultar, limpiar, explorar y analizar grandes conjuntos de datos (Bail 2014, Edelmann et al. 2020, Evans y Foster 2019).

Por otro lado, durante estos años ha ido cobrando fuerza el movimiento de la ciencia abierta, (*open science*) con el propósito de hacer accesible la investigación y sus datos (Ayris y Ignat 2018). Uno de los pilares básicos de este movimiento son los datos abiertos (*open data*) surgido de los mismos fundamentos que las corrientes de código abierto (*open source*) y de acceso abierto, (*open access*) (Murray-Rust 2008). Sus tres pilares son la transparencia, la participación y la colaboración (White House 2009). La *Open Knowledge Foundation (2022)* define los datos abiertos como aquellos que pueden ser utilizados, reutilizados y redistribuidos por cualquiera con la única condición de citar las fuentes y compartirlos. Muchos gobiernos y administraciones públicas se han sumado al movimiento publicando sus datos a través de portales de datos abiertos desde los que hacen accesible y reutilizable su información. Agencias de financiación a nivel internacional han incorporado políticas de datos que exigen planes de gestión de datos y la publicación abierta de los mismos. Multitud de editoriales y revistas comienzan a exigir que los artículos se acompañen de los datos de investigación subyacentes (Hrynaszkiewicz et al. 2020). Las bibliotecas no han querido quedarse atrás y han buscado formas de participar en los nuevos ecosistemas científicos formados por investigadores, editores, instituciones y agencias de financiación (MacDonald y Martínez-Uribe 2008). Una de las maneras en las que las bibliotecas están queriendo participar en este contexto es preparando sus colecciones digitales para uso computacional (Padilla et al. 2019).

Todo lo anterior ha ocurrido motivado por los drásticos cambios producidos por los avances tecnológicos relacionados con el almacenamiento y la organización de los datos. En unas pocas décadas se ha pasado de almacenar datos de forma analógica, en tarjetas perforadas y cintas magnéticas, a almacenamientos distribuidos y computación en la nube en poderosos centros de

[3] Ejemplos de estudios de postgrrado incluyen *el* master de la Universidad de Lucerne https://www.unilu.ch/en/study/study-programmes/masters-degrees/faculty-of-humanities-and-social-sciences/lucerne-master-in-computational-social-sciences-lumacss/#section=c77943, el de la Universidad de Chicago https://macss.uchicago.edu/ o el de la Universidad Carlos III de Madrid https://www.uc3m.es/master/computational-social-science#:~:text=The%20Master%20in%20Computational%20Social,understand%20society%20and%20human%20behavior.

[4] Un ejemplo es The Institute for Analytical Sociology de la Universidad de Linkoping https://liu.se/en/organisation/liu/iei/ias

datos (Farber et al. 2013, Kaur, Kumar y Singh 2014). Con las bases de datos en las que se estructura la información ha ocurrido algo parecido. Driscoll (2012) repasa su evolución desde los inicios donde se utilizaban sistemas ineficientes de ficheros jerárquicos, analiza el paso fundamental que suponen las bases de datos relacionales que permiten una gestión más eficiente y que además cuentan con un lenguaje de consulta como es SQL y finalmente como son complementadas con bases de datos NOSQL capaces de tratar grandes cantidades de información no estructurada.

El presente trabajo de investigación en formato de tesis por compendio de publicaciones se engloba en la tarea de aportar a la sociología computacional metodologías y herramientas que faciliten el trabajo con grandes cantidades de información. Para ello esta tesis intenta descubrir y aplicar herramientas y estrategias novedosas para enfrentarse a los *big data* y conseguir de ellos una correcta interpretación. En particular se presenta un paquete del lenguaje de programación estadística R llamado netCoin que consta del poderoso aparato estadístico del análisis de coincidencias, detallado en el segundo capítulo de la tesis, para identificar una serie de eventos que tienden a aparecer juntos en un conjunto de espacios. La librería hace uso del análisis de redes sociales y permite generar redes dinámicas e interactivas con las que realizar análisis exploratorios para profundizar en las relaciones entre las distintas variables. Son técnicas de analítica visual que aplican los principios de análisis exploratorio de Tukey (1977), pero incorporan otras herramientas desarrolladas con posterioridad. Para poder hacer uso de ello es fundamental reducir la cantidad de información que se analiza. A tal efecto es crucial disponer de métodos que permitan filtrar la información de utilidad en ocasiones valiéndonos de fuentes de datos externas o de técnicas de reducción de dimensionalidad para poder hacer la selección de los datos a utilizar en el análisis.

Además, esta tesis tiene por objetivo ilustrar a través de ejemplos presentados en los capítulos 3 y 4 de este trabajo, cómo los sociólogos tienen a su alcance datos generados y gestionados por las bibliotecas y los archivos de potencial interés para la disciplina. Estos datos son relevantes para la sociología ya que las bibliotecas y archivos son instituciones que preservan la memoria a partir de la cual se puede interpretar las circunstancias políticas, económicas y sociales a lo largo del tiempo. Con esta finalidad se presentan ejemplos de la aplicación de diversas técnicas de análisis de grandes colecciones de libros y artículos científicos. Los datos que actualmente gestionan las bibliotecas y archivos son heterogéneos en cuanto a sus formatos y contenidos. Los propios catálogos bibliográficos que contienen la información de los volúmenes físicos que se almacenan en las bibliotecas pueden, por ejemplo, poner de relevancia colecciones de investigación especializadas en dominios específicos. Las bibliografías nacionales, es decir las colecciones con información sobre todos los libros publicados en un país, muestran la historia editorial, la evolución de temáticas de interés y la especialización y la importancia de los grupos editoriales. Otros datos habitualmente relacionados con las bibliotecas son los bibliográficos. Estos datos incluyen información de publicaciones científicas, congresos, tesis y libros que permiten medir el impacto y la productividad científica, así como estudiar la evolución de los dominios científicos. A estas fuentes hay que añadirles la información recogida en los archivos que describen colecciones históricas o legados personales, que incluyen archivos fotográficos, correspondencia, recortes de prensa y demás materiales que permiten una investigación histórica del pasado que ayude a entender el presente y el futuro (Moore et al. 2016).

Esta tesis está organizada del siguiente modo: el capítulo introductorio comienza enmarcando el problema de investigación tratado en el campo de la sociología computacional. Tras esto se presenta el propósito de la investigación repasando los objetivos de cada una de las tres publicaciones. A continuación, se profundiza en conceptos como la sociología computacional, el fenómeno de *big data*, los datos bibliográficos y de bibliotecas. Tras lo cual se revisa la tradición de metodologías cuantitativas en sociología. A continuación, se incluyen los tres artículos. El

primer artículo presenta un paquete del lenguaje estadístico R denominado netCoin que permite explorar estructuras de datos, detectar patrones y generar redes interactivas usando análisis multivariante y análisis de redes sociales. El segundo artículo hace uso de una colección de *big data* abiertos de la *British Library* a la que se aplican técnicas de análisis de redes sociales, análisis de coincidencias, analítica visual y técnicas de reducción de dimensionalidad. El tercer artículo describe la sociología de los últimos años a través de las publicaciones en revistas de impacto usando el *Microsoft Academic Graph dataset,* una fuente de *big data* bibliográfica con la que se estudia la evolución del número de citas, coautoría y género de los autores para terminar con una red de afiliación entre autores y revistas, que muestra la sociología como agrupaciones temáticas y geográficas interrelacionadas. La tesis termina planteando una serie de conclusiones y haciendo explícitas las limitaciones del trabajo y sugiriendo posibles avenidas de investigación para el futuro.

1.1 Descripción del problema de investigación

Este trabajo se enmarca en el campo de la sociología computacional, un campo interdisciplinar que surge a partir del aumento de los datos digitales disponibles para la investigación social fruto de las nuevas fuentes de información en Internet, de los medios sociales y de los proyectos de digitalización masiva llevados a cabo en archivos y bibliotecas (Edelmann et al. 2020, Evans y Foster 2019, Gualda 2022a). Como consecuencia se plantea el problema de la gestión de estos nuevos datos, denominados *big data,* que aportan perspectivas nuevas y complementarias a aquellas resultantes de los datos convencionales utilizados por los científicos sociales. Entre las oportunidades que aparecen encontramos la capacidad de usar datos sobre fenómenos sociales no disponibles anteriormente, la disponibilidad de datos discretos (o no invasivos) y de poblaciones completas o en tiempo real (Boyd y Crawford 2012, Espeland y Stevens 2008, Manovich 2015, Martinho 2018, Tinati et al. 2014). Del mismo modo, estas nuevas fuentes de información tienen asociadas una serie de retos metodológicos y teóricos al encontrarnos en ocasiones con datos que no han sido capturados para un análisis sociológico y pueden ser parciales, erróneos o sin información de procedencia (Halford y Savage 2017, Lazer y Radford 2017, McFarland y McFarland 2015).

Revisiones de la literatura sobre la sociología computacional como las de Edelman et al. (2020) o la de Lazer and Radford (2017) presentan multitud de estudios y publicaciones sociológicas llevadas a cabo durante los últimos años que utilizan fuentes de *big data.* Estas revisiones permiten identificar los tipos de fuentes de *big data* más utilizadas. Los estudios que utilizan como fuente los medios sociales como Twitter o Facebook son los más habituales. En ellos se explotan estos datos para analizar temas como la formación de grupos, los comportamientos colectivos o diversos elementos de la sociología cultural cultural (Bail 2016, Bail, Brown y Mann 2017, Flores 2017, Golder Scott y Macy Michael 2011, Lewis, Gray y Meierhenrich 2014, Park, Baek y Cha 2014, Tufekci y Wilson 2012, Vasi et al. 2015). Por otro lado, abundan los estudios bibliográficos centrados en fuentes como ISI Web of Science[5], JSTOR[6] o Medline[7]. Estas publicaciones tratan temas de impacto y prestigio académico, el dominio de los grupos científicos en la producción de conocimiento o las desigualdades de género relacionadas con las citas o la autoría (Adams y Light 2015, Evans y Aceves 2016, King et al. 2017, Leahey y Moody 2014, Rzhetsky et al. 2015, Uzzi

[5] ISI Web of Science es un servicio de acceso a información bibliográfica que cubre gran variedad de disciplinas científicas https://clarivate.com/webofsciencegroup/solutions/web-of-science/

[6] JSTOR es una biblioteca de publicaciones académicas de humanidades y ciencias sociales https://www.jstor.org/

[7] Medline es una base de datos de bibliografía médica producida por la Biblioteca Nacional de Medicina de los Estados Unidos https://medlineplus.gov/

et al. 2013, West et al. 2013, Wuchty, Jones Benjamin y Uzzi 2007). También encontramos estudios sociológicos que utilizan datos generados a partir de experimentos a través de simulaciones (Helbing, Farkas y Vicsek 2000) o con juegos *online* (Centola 2010, Shirado y Christakis 2017). Existen además publicaciones que emplean datos de telefonía y de dispositivos móviles utilizando datos de las llamadas telefónicas (Toole et al. 2015), de utilización de WIFI (Shirado et al. 2019) o de Bluetooth (Eagle, Pentland y Lazer 2009) o de mensajería instantánea (Saavedra, Duch y Uzzi 2011). Asimismo, abundan los estudios que utilizan información de distintos servicios que se ofrecen a través de Internet como Spotify para la reproducción de música música (Askin y Mauskapf 2017), Airbnb para alquileres vacacionales (Edelman y Luca 2014) o páginas de citas (Potârcă y Mills 2015). Por supuesto no faltan estudios que utilicen la variedad de información que almacena Google a través de las consultas a su buscador (Bail, Brown y Wimmer 2019), el texto completo de los libros en *Google Books* (Shor et al. 2015) o su servicio de *Street View (Gebru et al. 2017*) Otras fuentes utilizadas que merecen ser destacadas incluyen los artículos de la Wikipedia (Wagner et al. 2016), o la información de préstamos en bibliotecas durante décadas (Hoffman 2019).

Esta revisión de las fuentes de datos de *big data* utilizados en estudios de sociología computacional, sin llegar a ser exhaustiva,muestra una importante omisión de fuentes de datos *big data* de bibliotecas y archivos. Es precisamente en este espacio donde esta tesis pretende aportar a la literatura ya existente relacionada con la sociología computacional. Para ello se presentan dos ejemplos del tratamiento de *big data* bibliográfico y de bibliotecas mediante análisis de redes. En estos ejemplos se utilizan metodologías y estrategias que permiten explorar visualmente grandes cantidades de información para descubrir, y poder interpretar las relaciones que aparecen en estas colecciones de datos de bibliotecas. La aportación de este trabajo puede ser relevante ya que los *big data* de bibliotecas aportan beneficios importantes en términos de calidad e integridad de la información, así como de facilidad de acceso y utilización de estos.

1.2 Objetivos de la investigación

El principal objetivo de este trabajo consiste en realizar una contribución metodológica al campo de la sociología computacional mediante una herramienta analítica que puede emplearse para estudiar de modo exploratorio la forma en que los datos están relacionados entre sí. Para ello formulamos estrategias y proporcionamos ejemplos para el tratamiento de *big data* basado en análisis de redes sociales. El objetivo secundario de la tesis consiste en mostrar mediante dos casos de uso concreto como acceder y tratar los *big data* bibliográficos y aquellos disponibles en bibliotecas y así exponerlos como una fuente de investigación de interés para los sociólogos por su capacidad para analizar las condiciones sociales, culturales e históricas a lo largo del tiempo.

A continuación, se presentan los objetivos de los tres artículos científicos. El primer artículo *Network Coincidence Analysis: The netCoin R package* presenta una herramienta de código abierto realizada con el lenguaje de programación estadística R que implementa el análisis de coincidencias el cual combina análisis multivariante y de redes para la exploración de estructuras de datos:

- Presenta la metodología del análisis de coincidencias para detectar cuándo una serie de eventos tienden a ocurrir de forma conjunta en un conjunto finito de determinados escenarios.

- Representa los resultados del análisis de coincidencias mediante redes permitiendo tanto la presentación de resultados finales como una exploración visual de datos previa de forma dinámica e interactiva.

- Aporta una herramienta de código abierto para aplicar el análisis de coincidencias y generar redes sociales dinámicas e interactivas, mostrando ejemplos diversos de su aplicación en distintas disciplinas.

En el segundo artículo *Digital Archives as Big Data* se analiza una colección de datos abiertos de 3 millones de registros bibliográficos de la *British Library* con la herramienta netCoin:

- Presenta los *big data* de bibliotecas fruto de las labores de limpieza y enriquecimiento habituales y de los proyectos de digitalización masiva de los últimos años como una tipología más de *big data* de utilidad para la investigación.

- Ofrece ejemplos de diversas técnicas de reducción de dimensionalidad y de filtrado de información mediante la combinación con datos externos.

- Aplica la metodología del análisis de coincidencias y el uso del análisis de redes dinámicas e interactivas como método de exploración visual de datos.

El tercer artículo *´La sociología a través de sus publicaciones en revistas de impacto mediante el uso de big data´* ahonda en los intereses de la disciplina sociológica en los últimos veinte años combinando los 300 millones de publicaciones de los datos de *Microsoft Academic Graph* con la información de revistas incluidas en el Journal Citations Report en la categoría de sociología. En este artículo, además, se aplican varias metodologías y estrategias que permiten reducir la dimensionalidad en la exploración de grandes cantidades de datos:

- Analiza la sociología concebida como un sistema complejo de relaciones sociales entre investigadores, instituciones, asociaciones y editoriales utilizando las publicaciones en revistas de impacto de los últimos años.

- Presenta el fenómeno del *big data* con sus expectativas y retos para después ceñirse a los *big data* bibliográficos y su aportación al estudio de las disciplinas científicas.

- Aplica varias metodologías de análisis y tratamiento de datos, entre ellas: el análisis descriptivo de las revistas JCR de sociología; el de filtrado de entre los millones de artículos de *Microsoft Academic Graph* utilizando las revistas JCR de sociología; la agrupación de revistas en cuatro tipologías y su comparación a nivel de citas, coautoría y distribución de género. Por último, genera una red bimodal de afiliación entre revistas y autores que es después proyectada a una red unimodal reduciendo la dimensionalidad, para poder analizar la relación entre las revistas.

Tabla 1. Resumen de las cuestiones de investigación planteadas, fuentes de datos y métodos utilizados

Artículo	Objetivos	Fuentes de datos	Métodos
Network Coincidence Analysis: the netCoin R Package	Presentar una herramienta de código abierto compuesta por una metodología estadística para el análisis de coincidencias que se apoya en el análisis de redes sociales para generar redes dinámicas e interactivas que permiten la exploración visual de los datos	• Matrimonios entre familias italianas • Especies de pájaros en las Islas Galápagos • Fichero de encuesta con preguntas multirrespuesta	• El análisis de coincidencias • Métodos multivariantes • Análisis de redes sociales • Medidas de proximidad
Digital Archives as Big Data	Presentar los datos de bibliotecas como *big data*. Mostrar el uso de técnicas de reducción de dimensionalidad y de exploración visual de *big data*	• El conjunto de datos abiertos de la Biblioteca Británica *The British National Bibliography* con 3 millones de registros de todas las publicaciones realizadas en Reino Unido e Irlanda desde 1962.	• Reducción de dimensionalidad mediante la proyección de una red bimodal • Filtrado de información usando datos externos • Exploración visual usando netCoin y análisis de redes sociales
La sociología a través de sus publicaciones en revistas de impacto mediante el uso de big data	Caracterizar la sociología a través de sus publicaciones en revistas de impacto. Presentar datos bibliográficos como *big data*. Mostrar el uso de técnicas de reducción de dimensionalidad y de exploración visual de *big data*	• El *Microsoft Academic Graph* dataset con información sobre más de 300 millones de publicaciones y congresos científicos. • Las revistas JCR de sociología con información enriquecida	• Infraestructura de *big data* en la nube de Microsoft • Reducción de dimensionalidad mediante la proyección de una red de afiliación bimodal • Filtrado de información usando datos externos • Exploración visual usando netCoin y análisis de redes sociales

1.3 La sociología computacional, el *big data* y los datos bibliográficos y de bibliotecas como oportunidad

Cioffi-Revilla (2014) realiza una revisión histórica notable del origen y la evolución de las ciencias sociales computacionales. Para Cioffi-Revilla, los orígenes se remontan al periodo de la revolución científica ocurrida entre el final del Renacimiento y el comienzo de la Ilustración. Es entonces cuando las ciencias sociales adaptan los principios de la metodología científica positivista. El comienzo más estricto, según Cioffi-Revilla (2014), lo marca la invención de la computadora digital surgida tras la Segunda Guerra Mundial. Esta computadora digital transforma las ciencias sociales como instrumento que permite por primera vez analizar datos, validar hipótesis novedosas y explorar nuevas dimensiones sociales. Por otro lado, la computadora digital también sirve de inspiración a nuevos conceptos, principios, teorías y modelos sobre los sistemas, procesos y componentes del universo social.

Es hoy en día cuando somos conscientes de que gran parte de nuestras interacciones diarias quedan registradas digitalmente y que Internet es una de las fuentes principales de generación de información (Gualda 2022a). Esta información se acumula de manera incesante, en particular por su gran valor comercial, recogiendo los rastros de la evolución de nuestras relaciones y comunicaciones como individuos, grupos, organizaciones y sociedades. El avance en la capacidad de capturar, almacenar y analizar datos está teniendo un efecto transformador en nuestra sociedad. Esta transformación se ve reflejada en muchas disciplinas científicas dando lugar a la llamada investigación basada en datos - *data driven research.* A los dominios científicos más pioneros en este ámbito, como la biología, las ciencias de la salud o la física, se le han sumado las humanidades a través del área de las humanidades digitales o las ciencias sociales mediante las denominadas ciencias sociales computacionales.

Para Cioffi-Revilla (2014), las ciencias sociales computacionales se basan en el paradigma del procesamiento de información de la sociedad. Según este paradigma, la información juega un papel clave a la hora de entender cómo funcionan los sistemas sociales. Cioffi-Revilla identifica cuatro áreas - formadas cada una por conceptos, teorías y métodos- , que forman la base de las ciencias sociales computacionales. Estas áreas, según el autor mencionado, incluyen la extracción automatizada de información social, el análisis de redes sociales, la teoría de la complejidad social y el modelado de la simulación social.

a) La extracción automatizada de información social se refiere a las metodologías para generar información de utilidad para el análisis social de las fuentes de datos brutas.
b) El análisis de redes sociales tiene una amplia tradición en las ciencias sociales dada la abundancia e importancia de su estudio.
c) La teoría de la complejidad social entiende la sociedad como un sistema adaptativo complejo que cambia de estado como respuesta a las condiciones.
d) Finalmente, el modelado de la simulación social, también con una amplia tradición, incluye diversos modelos de simulación como los de dinámicas de sistema, los autómatas celulares o los sistemas multiagente.

Dentro de las ciencias sociales computacionales se encuentra la rama de la sociología computacional, un campo que según Evans y Foster (2019) requiere aplicar el concepto de la imaginación social de Mills (1959) para alcanzar su máximo potencial. Edelman et al. (2020) la definen como un campo interdisciplinar en el que se aplican técnicas computacionales a los *big data* extraídos de medios sociales, Internet o grandes archivos con el fin de avanzar las teorías sociales del comportamiento humano. Esta definición recalca la importancia del uso de la teoría sociológica pues inicialmente, el análisis de *big data* fue abordado por ingenieros, informáticos y matemáticos gracias a sus habilidades técnicas para acceder y computar estas grandes cantidades

de información (Burrows y Savage 2014, McFarland, Lewis y Goldberg 2016, Savage y Burrows 2007). Sin embargo, para profundizar y mejorar en el análisis de estas nuevas fuentes de información es crucial la participación de los sociólogos para que con su experiencia y conocimiento puedan interpretar la manera en que se estructura lo social (Tubaro 2014). Los datos por sí mismos no son suficientes para abordar los problemas sociales más acuciantes y el trabajo con *big data* continúa sometido a interpretación, parte fundamental del análisis social (Boyd y Crawford 2012). De igual forma la sociología podría enriquecer el campo de los *big data* reforzando su enfoque para centrarse en aquello que produce mayor beneficio social (Gualda 2022b).

El término *big data* suele estar asociado a la sociología computacional y al resto de investigación basada en datos, refiriéndose a aquellas nuevas fuentes de datos de gran tamaño al alcance del investigador. Originalmente, este término apareció en un artículo de investigadores de la NASA en 1997 al encontrarse con unos datos que debido a su tamaño les suponía un problema de memoria, a esto le llamaron el problema del *big data (Press 201*4). Unos años después se convertiría en una palabra de moda, *buzzword,* y aparecería regularmente en periódicos como *The New York Times, Financial Times* y en revistas científicas como *Nature, Science,* o *The Economist (Kitchin 2014a).).* Hoy en día, el entusiasmo que alimenta el *big data* procede del éxito que han alcanzado compañías tecnológicas como *Google* o *Amazon*. Las tipologías comunes de datos dentro de *big data* incluyen aquellos datos generados en los medios sociales como puedan ser Twitter, Instagram o Facebook, los datos producidos por dispositivos conectados a Internet como los acelerómetros o los sensores de posicionamiento, los datos transaccionales generados por ejemplo con los gastos de las tarjetas de crédito y los datos administrativos capturados y almacenados por los gobiernos. Muchos de estos datos están gestionados por grandes compañías privadas y presentan retos importantes para su acceso. Sin embargo, una tipología de datos que se suele ignorar es la de aquellos datos gestionados por las bibliotecas y archivos fruto de las labores masivas de digitalización y curación de datos realizadas comúnmente en este tipo de unidades de información. Estas grandes colecciones de datos digitales incluyen vídeos, grabaciones sonoras, imágenes, documentos y datos bibliográficos fácilmente accesibles (Martinez-Uribe y Fernandez 2015).

Las bibliotecas y archivos son instituciones que conservan la inquietud, el devenir, las circunstancias políticas, económicas, culturales, sociales, la historia de los individuos y sociedades a lo largo del tiempo, por ello sus datos son cruciales para el análisis diacrónico y multilateral, temática y geográficamente. Desde hace años estas instituciones se están sumando al movimiento de ciencia abierta (Murray-Rust 2008), al abrir el acceso a grandes conjuntos de datos bibliográficos, catalogados y cuidadosamente enriquecidos durante años por los bibliotecarios para su análisis computacional por parte de los investigadores (Deliot 2014, Padilla et al. 2019). Así, la British Library publicó en 2010 dieciséis millones de registros de su catálogo en abierto y hoy en día tiene ciento cincuenta conjuntos de datos publicados[8]. Otros ejemplos destacables incluyen a *HathiTrust Research Center Analytics*[9], la Biblioteca Nacional Holandesa[10] o el servicio *Chronicling America*[11] de *Library of Congress*. De esta forma las bibliotecas pasan a formar parte de las infraestructuras de datos de investigación existentes (Lauriault et al. 1969) junto con los centros nacionales de datos, data archives, (como the *UK Data Archive* en Reino Unido o *The Data Archiving and Network Services* en los Países Bajos) y los portales de datos de comunidades científicas específicas (como *The Protein Data Bank* en

[8] En la página *British Library Datasets* publican los datos en abierto https://data.bl.uk/bl_labs_datasets/

[9] El *HathiTrust Research Center Analytics da soporte al análisis computacional de sus colecciones digitales https://analytics.hathitrust.org/*

[10] La Biblioteca Nacional Holandesa tiene un API para acceder a los datos y metadatos de sus colecciones digitales https://www.kb.nl/en/resources-research-guides/data-services-apis

[11] *Chronicling america* da acceso a periódicos históricos americanos de 1777 a 1963 https://chroniclingamerica.loc.gov/about/api/

biología, *The National Virtual Observatory* en astronomía o *The Archive of World Music* en musicología).

Los conjuntos de datos que custodian y sirven las bibliotecas incluyen información detallada sobre libros, artículos y otros tipos de publicaciones, detallando sus autores, temas y contenidos, así como el editor, la ubicación y el momento de la publicación. En algunos casos, estos catálogos son colecciones completas de grupos de investigación en ámbitos especializados o de toda la producción editorial del país. Otros datos gestionados por las bibliotecas son las colecciones históricas y legados personales custodiadas por los archivos e incluyen información detallada de materiales como libros, fotografías, mapas, correspondencia, recortes de prensa y muchos otros documentos diversos. Los datos bibliográficos se han analizado tradicionalmente en el ámbito de la bibliometría, la rama de la biblioteconomía que se ocupa de la aplicación de la estadística al estudio de los datos bibliográficos para, por ejemplo, evaluar la investigación o medir la producción científica (Battisti y Salini 2012).). Pero el análisis de datos bibliográficos no se limita a la bibliometría y en los últimos años ha habido una proliferación de estudios multidisciplinares con datos bibliométricos (González-Alcaide 2021). En concreto, han aparecido multitud de estudios enmarcados dentro de la sociología de la ciencia que describen la evolución de las disciplinas científicas o estudian el proceso de generación y consumo de conocimiento a través de sus publicaciones (Evans y Foster 2011, Su y Lee 2010, Vanderstraeten 2010). No solamente esto, la aplicación de las metodologías propias de la sociología a los datos de las bibliotecas y archivos tienen el potencial de explicar comportamientos y la evolución de la sociedad.

Entre las ventajas de los datos de las bibliotecas están su alto nivel de calidad y su facilidad de acceso. La normalización y la estandarización de la información son elementos fundacionales de la disciplina de las ciencias de la información. Entre las estrategias utilizadas para garantizar la calidad de la información está la utilización de vocabularios y tesauros con términos geográficos, temáticos o temporales como el *Library of Congress Subject Headings*[12] creado en 1898 o el *Tesauro de Arte y Arquitectura* del *Getty*[13] de los años 70. Además, se ha desarrollado un control de autoridades de tipo personal e institucional integrado a nivel internacional, como el *Fichero de Autoridades Virtual Internacional* (VIAF) [14]. Las bibliotecas además se han sumado a iniciativas como el movimiento *de datos abiertos* enlazados para asegurar que los datos puedan interconectarse entre fuentes (Berners-Lee 2009, Hallo et al. 2016), a la agenda de ciencia abierta que tiene por objetivo hacer accesibles la investigación y los datos a todos los niveles que la sociedad necesita (Ayris y Ignat 2018). Así como a un uso responsable de las tecnologías de inteligencia artificial para optimizar procesos o añadir nuevas dimensiones a la gestión del conocimiento (IFLA 2020).

1.4 Metodologías cuantitativas en la sociología. De la estadística a la ciencia de datos.

Ronald A. Fisher comienza su ya clásico e influyente libro *Statistical Methods for Research Workers* (1925) definiendo la ciencia estadística como una rama de la matemática aplicada a los datos observacionales y argumenta que su utilización en los estudios sociales hace que estos puedan ser elevados al rango de ciencias. Figuras prominentes en la fundación de la disciplina como Marx, Durkheim y Weber se dieron cuenta de la importancia de la obtención y el análisis estadístico de la información cuantitativa relativa a los fenómenos sociales para construir una ciencia de la sociedad (García Ferrando 1980). Las ciencias sociales, y en particular la sociología, han utilizado metodologías cuantitativas desde sus inicios aportando rigor a la disciplina.

[12] *Library of Congress Subject Headings* es el tesauro de la Biblioteca del Congreso de los Estados Unidos https://id.loc.gov/authorities/subjects.html

[13] El tesauro de Arte y Arquitectura del *The Getty Research Institute* está disponible online en https://www.getty.edu/research/tools/vocabularies/aat/

[14] El Fichero de Autoridades Virtual Internacional se puede consultar en http://viaf.org/

Asimismo, en muchos casos la sociología ha contribuido de manera importante al desarrollo de métodos estadísticos (Clogg, 1992).

En la revisión del uso de la estadística en la sociología de 1950 al año 2000 realizada por Raftery (2000) se presenta una clasificación de los métodos estadísticos de la sociología de acuerdo con los tipos de datos usados. Así, una primera generación de métodos aparece a partir de la Segunda Guerra Mundial donde la mayor parte de los datos vienen de tabulaciones cruzadas de encuestas y censos, y con un número pequeño de variables. Los análisis se centran en las medidas de asociación usando modelos log lineales y contrastes de hipótesis. Tras los años 60, una segunda generación comienza a tener a su disposición microdatos de encuesta en tablas en forma de matriz con casos independientes. Los métodos que resultan efectivos con estos datos son los modelos de regresión lineal, ecuaciones estructurales, modelos lineales generalizados y los modelos de registro de eventos. A partir de los años 80 comienzan a aparecer datos que van más allá de las tabulaciones y encuestas y que incluyen, por ejemplo, datos textuales, redes sociales o datos espaciales. Tal variedad de datos conlleva un rápido desarrollo en el uso de metodologías para su análisis que abarcan el análisis de redes sociales, el análisis espacial, el análisis de contenidos, el análisis textual o los modelos de simulación.

En los últimos años, y con la proliferación de la disponibilidad de datos, muchos de los métodos mencionados anteriormente han seguido utilizándose y ampliándose con nuevas aproximaciones computacionales para tratar las nuevas fuentes de información. Surge la ciencia de datos, definida por Donoho (2017) como la ciencia para el aprendizaje de los datos donde el paradigma estadístico evoluciona de un modelado generativo, donde el objetivo es la inferencia, a un modelado predictivo donde la finalidad es la predicción (Breiman 2001). Hace ya sesenta años que Tukey (1962) identificó la aparición de una nueva ciencia que iba más allá de la estadística formada por las teorías estadísticas, los avances tecnológicos, el aumento de las cantidades de datos y el énfasis en la cuantificación de muchas disciplinas.

De esta forma, metodologías tradicionales como el análisis de contenido y el análisis textual han sido complementadas por las metodologías de procesamiento del lenguaje natural, mucho más potentes y precisas, permitiendo identificar diferencias, preferencias y disposiciones de textos y contenidos relacionados (Evans y Aceves 2016). Comienza también a abundar el uso de las técnicas de aprendizaje automático (*machine learning)* en la sociología (Molina y Garip 2019). Los métodos supervisados, donde se cuenta con un conjunto de datos para entrenar el modelo, permiten predecir los *outputs*, usar las predicciones como punto de partida para entender las relaciones sociales o para mejorar las técnicas estadísticas clásicas. Por otro lado, el aprendizaje no supervisado puede ser usado para describir y clasificar *inputs* y para conceptualizar basándose en las descripciones.

Asimismo, existe una corriente dentro de la sociología que aboga por un mayor uso de la visualización de datos tanto para la presentación de resultados finales como para exploración de los datos como parte del proceso de limpieza y comprobación rutinario (Healy y Moody 2014).

1.5 Capacidades y beneficios del análisis de coincidencias para el análisis de *big data*

En esta sección se explica el análisis de coincidencias, marco analítico utilizado en los tres artículos de la tesis. El desarrollo de la notación matemática que explica el aparato estadístico del análisis de coincidencias se encuentra al completo en el artículo del capítulo 2 y no es necesario volver a repetirlo, pero si es preciso detallar las capacidades y beneficios que aporta y que hace de esta metodología cuantitativa una herramienta de gran utilidad para los sociólogos computacionales.

El análisis de coincidencias es un marco estadístico que tiene por misión identificar objetos, atributos, características o eventos que tienden a aparecer juntos en un conjunto finito de

escenarios (Escobar 2015). Así, el propósito de este análisis es encontrar el subconjunto de pares de objetos, atributos, características o eventos que no son independientes en el conjunto de escenarios.

Para ello se genera una matriz binaria de incidencias con los escenarios en las filas y los eventos en las columnas, indicando los eventos que suceden en cada escenario. Al multiplicar la matriz de incidencias por la matriz traspuesta se obtiene una matriz simétrica donde cada uno de los eventos aparecen en las filas y las columnas y con los valores denotando el número de veces que dos eventos cualesquiera han coincidido en los escenarios con los que contábamos. Con esta información se utiliza la teoría de probabilidad para generar distintas métricas de coincidencias basadas en medidas de proximidad. Como consecuencia, se puede generar una red donde los nodos son los eventos y el peso de las aristas viene determinado por la métrica de coincidencias seleccionada.

Con esta metodología disponemos de la siguientes capacidades y estrategias para tratar y analizar *big data*:

- En primer lugar, la capacidad de detectar las coincidencias entre distintos eventos en un número limitado de espacios y cuantificarlas usando diferentes medidas de distancia aporta información relevante para identificar patrones y relaciones dentro de los conjuntos de datos. En el segundo artículo de esta tesis, capítulo 3, se aplica el análisis de coincidencias a los datos del *British National Bibliography* donde los espacios son los libros y los eventos son las materias temáticas, temporales y geográficas junto con las décadas de publicación y las editoriales. El análisis permite detectar patrones que muestran la evolución de las temáticas en el tiempo, así como las diferencias entre las editoriales más prolíficas. En el tercer artículo, capítulo 4, al analizar las publicaciones en sociología se utilizan los autores como los espacios y los eventos son las revistas. Esta red de afiliación permite identificar las relaciones que se establecen entre las revistas debidas a las temáticas tratadas, las metodologías utilizadas o la localización geográfica de los autores.

- La combinación del análisis de coincidencias con el análisis de redes sociales permite representar los resultados de las coincidencias tanto para realizar exploraciones de los datos como para presentar los resultados finales. Este marco analítico ha sido desarrollado en el lenguaje de programación R a través del paquete netCoin. Como aplicación nos permite realizar análisis exploratorios de los resultados a través de redes interactivas en las que se puede jugar con el tamaño, los colores, las formas y las etiquetas de los nodos y aristas, así como con toda una serie de filtros para seleccionar los nodos y aristas más relevantes. El uso del análisis de redes de manera interactiva hace posible la exploración inicial de datos, un paso fundamental en el análisis. Ello se hace aún más importante cuando se trabaja con *big data*. Además, extiende el análisis exploratorio de Tukey (1977) con la idea de observar los datos a través de métodos visuales para visualizar lo que parecen mostrar y así después utilizar otras técnicas que nos confirmen o no los resultados. En los artículos presentados en los capítulos 3 y 4 se ha utilizado de manera extensiva el análisis exploratorio mediante las redes sociales construidas con el análisis de coincidencias. Estas exploraciones iniciales han permitido enfocar las investigaciones en determinados aspectos posteriormente. Además, los resultados finales que presentan ambos artículos utilizan representaciones gráficas de las redes para poder interpretar de manera visual las conclusiones.

- Por último, el análisis de coincidencias posibilita la capacidad de reducir la dimensionalidad de *big data* de varias formas. Por un lado, los eventos que se consideran pueden ser de distinto tipo y al representarlos como nodos de una red se pueden utilizar los colores y las formas para diferenciar las distintas tipologías. Esto permite la representación de relaciones entre las distintas dimensiones de datos multivariantes. No solo eso, en aquellos casos en los

que contamos con un número alto de eventos o dimensiones, el cálculo de las métricas de coincidencias permite cuantificar la importancia de las relaciones para así poder reducir el número de eventos a aquellos con las relaciones de mayor importancia. En el artículo del capítulo 3 se utilizan estas dos estrategias. Los eventos que se incluyen en el análisis son diversos e incluyen temáticas, décadas de publicación o editoriales. Estas tipologías se representan como nodos de una red con distintas formas. Al calcular el peso de las aristas, es decir, la fuerza de la coincidencia, se eliminan de las representaciones gráficas aquellos nodos con menor grado de coincidencia para enfocar el análisis en las relaciones de mayor alcance.

- Otras de las técnicas de reducción de dimensionalidad que utiliza el análisis de coincidencias consiste en partir de una red bimodal que se proyecta a una red de modo uno. La red bimodal es la matriz de incidencias inicial de espacios y eventos. Se trata de redes con dos tipos distintos de nodos donde las relaciones solo existen entre nodos de distinto tipo. Al multiplicar esta matriz por su traspuesta, la red se proyecta a una red de modo uno donde solo quedan uno de los tipos de nodos. Esta técnica de reducción de dimensionalidad se utiliza en los artículos de los capítulos 3 y 4. En el artículo donde se utilizan los datos del *British National Bibliography* se parte de una red bipartita con más de dos millones de espacios, las publicaciones, y casi trescientos mil eventos, las materias, décadas y editoriales. Al proyectar la red nos quedamos con las relaciones entre los eventos que después son filtradas para quedarnos con aquellas de mayor fuerza. En el artículo del capítulo 4 se aplica la técnica a los datos del *Microsoft Academic Graph.* Como se muestra en la figura 1, partimos de una red bipartita formada por más de ciento veinticinco mil espacios, los autores, y ciento cincuenta eventos que son las revistas. Al proyectar a una red de modo uno la red queda formada únicamente por las revistas y las relaciones entre dos revistas son más fuertes cuantos más autores comunes tengan las revistas.

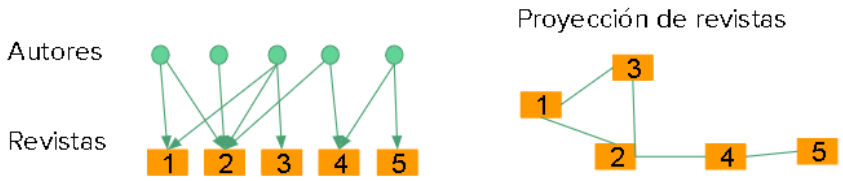


Figura 1. Red de afiliación entre autores y revistas y su proyección a una red de revistas

1.6 Presentación de los artículos publicados

Tras esta introducción a la sociología computacional y las metodologías cuantitativas de la sociología, en la siguiente sección se presentan los tres artículos publicados como parte de la tesis. En el primero se describe la metodología del análisis de coincidencias y se presenta la herramienta netCoin que permite aplicar el análisis y generar redes interactivas. El análisis de coincidencias hace posible identificar patrones en los datos y reducir la dimensionalidad. Las redes interactivas son útiles para realizar análisis exploratorios iniciales y presentar los resultados finales visualmente. El segundo artículo presenta los datos de bibliotecas como una tipología más de *big data* y aplica el análisis de coincidencias aportando un ejemplo de cómo se pueden tratar estos conjuntos de datos. Finalmente, en el último artículo se analiza una colección de grandes datos bibliográficos a través de distintas metodologías, incluido el análisis de coincidencias, para presentar la disciplina de la sociología como un sistema complejo de relaciones sociales entre investigadores, instituciones, asociaciones y editoriales.

Journal of Statistical Software

May 2020, Volume 93, Issue 11. *doi: 10.18637/jss.v093.i11*

Network Coincidence Analysis: The netCoin R Package

Modesto Escobar
Universidad de Salamanca

Luis Martinez-Uribe
Universidad de Salamanca

Abstract

The aim of the R package **netCoin** is to explore data structures using a number of statistical techniques that share the handling of interdependent variables. The main objective of this analysis is to detect events, characters, objects, attributes or characteristics that tend to appear together within a given set of scenarios. Its most notable feature is the combination of traditional multivariate statistical analysis and network analysis supported by topological graph theory. In addition, **netCoin** produces HTML graphs using the **D3.js** visualization library to support the interactive exploration of networked data. Among its many applications, **netCoin** can be used for the analysis of multiple responses in questionnaires to explore relevant associations, for the development of textual networks, for the study of ecological communities, for audience analysis, for mining large databases or for basket market analysis.

Keywords: co-occurrence analysis, social network analysis, multivariate analysis, interactive graphs.

1. Introduction

netCoin (Escobar, Barrios, Prieto, and Martinez-Uribe 2020) is an R package which performs network coincidence analysis, whose aim is to find out the structure and the degree to which a series of events (subjects, objects or characteristics) tends to occur together within certain limits called scenarios. To discover these patterns, this package generates visualizations of the coincidences through interactive network graphs via a web browser.

Graphs represent elements (nodes) that may or may not be connected (edges). Coincidence graphs consist of two types of information: a set of nodes or vertices (events), $N = (n_1, n_2, \ldots, n_J)$, and a set of lines, links or edges (coincidences), $L = (l_1, l_2, , \ldots, l_L)$ (Wasserman and Faust 1994).

The interactive web graphs produced by **netCoin** allow modification of their elements and their features (such as size, color or position). In addition, data about nodes and edges is displayed below the graph, and both the data and the graphs can be downloaded onto every computer connected to the graph via an Internet browser.

Among the several interactive elements available, the following are key:

a. The label, size, color and shape of the events or nodes based on their properties. It is also possible to represent groups of nodes with similar characteristics as well as using images to depict them.

b. The width, weight, color and any text of the edges that represent the coincidences between the events based on the edges' properties, such as their frequency, degree of coincidence or statistical significance.

c. Nodes can be filtered manually or dynamically by either the value of their attributes or their connections.

d. Edges can also be filtered dynamically by the value of their attributes.

The starting point for these graphs is the incidence matrix, made up of two dimensions: The rows that contain the scenarios where the coincidences are to be detected and the columns which contain the elements whose coincidences are of interest for this particular study.

An application for this analysis arises from the complexity of working with multiple-choice questionnaires. To illustrate this use, we may consider a simple question about job hunting strategies that unemployed people might use to find a job. For this question, a survey could include responses such as family, friends, sending résumés to companies, job ads in newspapers or job centers. Therefore, what is the best way to code and save this information? One column is insufficient, as we might be dealing with multiple alternative answers.

However, two solutions might be applied: firstly, using as many columns as there are possible responses. The question may ask respondents to provide their top three job hunting strategies. In this case, three columns could be enough, providing a different strategy in each one. Nonetheless, if the number of responses is open, the number of columns needed could be codified using the multiple mode (from 1 to 5 in the case of 5 possible responses) or in a dichotomous fashion using one column for each response and codifying them as one for those selected and as zero for those not selected.

Another use for the network coincidence analysis is content analysis. A survey, apart from multiple-choice questions, may also include open-ended questions. The text from the responses to those open-ended questions may be divided into words or phrases whose coincidences may also be the subject of analysis. Plenty of specialized software could be used (N-Vivo, Atlas-Ti, QDA Miner, MaxQDa) whose main objective is to enable the classification of large text corpora such as transcripts of focus groups and interviews. In addition to this, algorithms are emerging to perform thematization (Corman, Kuhn, McPhee, and Dooley 2002; Blei, Ng, and Jordan 2003; Feinerer, Hornik, and Meyer 2008; Van Attenveld 2008; Grimmer and Stewart 2013; Roberts *et al.* 2014; Lucas, Nielsen, Roberts, Stewart, Storer, and Tingley 2015) and sentiment analysis (Young and Soroka 2012), which could use graph representation.

Due to **netCoin**'s core objective to produce graphs, it turns this into a fit-for-purpose methodology for the study of bimodal networks, which present two sets not internally connected but

interconnected between them. One of the typical applications of these structures are affiliation networks, which represent the connections between an actor with a set of social situations (Wasserman and Faust 1994). For example, bimodal networks could help study the membership of an executive group in a company or the events that the inhabitants of a certain village attend.

Those relationships can be studied using bipartite graphs or hypergraphs as well as dual hypergraphs, although in most cases the representation of only one of the sets is of interest (such as the actors or the inhabitants in the previous examples), and any bimodal network can be transformed into a unimodal one, which leads to the preference for the co-participation matrix. Precisely, the main operation behind **netCoin** is generating a coincidence matrix (unimodal) from an incidence matrix (bimodal) to convert the former into a graph.

Another application of network coincidence analysis is the study of species within different ecosystems. The coexistence of bird species in the Galápagos Islands is one highly popular case among biologists (Sanderson 2000). For this study, a probabilistic co-occurrence method based on the hypergeometric distribution, which is also included in **netCoin**, has been developed (Veech 2013).

2. Similar software

There are a variety of tools within the statistics and data analysis domain that perform similar operations to **netCoin** to visually analyze the structure of binary data.

It is also possible to find common ground with machine learning techniques, especially the association rules (Borgelt 2012) that have binary matrices as their starting point. In contrast, **netCoin** focuses on the associations between pairs of events while `apriori()` and `eclat()` procedures seek for higher order connections available through the package **arules** (Hahsler, Grün, and Hornik 2005).

The comparative qualitative analysis (QCA; Ragin 1987, 2000) has a similar input, i.e., a matrix made up of zeros and ones, although it is based on different algorithms using Boolean logic. Coincidence analysis (Baumgartner 2009) is derived from the QCA and they both have R packages: **QCA** (Wickham and Miller 2019) and **cna** (Baumgartner and Thiem 2015).

It is also worth mentioning some packages associated with text analysis, like **tm** (Feinerer 2019; Feinerer and Hornik 2019), **RTextTools** (Jurka, Collingwood, Boydstun, Grossman, and van Atteveldt 2014), **textometry** (Loiseau, Vaudor, Decorde, and Heiden 2015), **lda** (Chang 2015), **stm** (Roberts, Stewart, and Tingley 2019a,b) and **tidytext** (Robinson and Silge 2020).

Tools with a specific focus on network analysis and visualization include four major packages: **igraph** (Csardi and Nepusz 2006; Csardi 2020), **network** by Butts (2008, 2019), the graphical complement **networkD3** (Grandrud, Allaire, and Rusell 2016; Allaire, Grandrud, Rusell, and Yetman 2015) and **visNetwork** by Almende, Benoit, and Titouan (2019). The first two are powerful tools for analyzing networks and can represent them in a non-interactive way, unless they are used in conjunction with **tcltk2** (Grosjean 2014), but they lack the analytic instruments to study coincidences and the ability to create HTML graphs. Another similar package is **RJSplot** (Prieto and Barrios 2017), which produces interactive and dynamic graphics widely used in DNA structure data analysis. The last three are more similar to **netCoin**. However, they lack statistical tools to produce the coincidence graphs. Outside of the R environment, a variety of social network analysis tools exist such as **Gephi** (Bastian,

Heymann, and Jacomy 2009), **Pajek** (Batagelj and Mrvar 1998) or the Phyton (Van Rossum *et al.* 2011) package **NetworkX** (Schult and Swart 2008).

It is important to mention those packages which specialize in co-occurrence in community structures. Griffith, Veech, and Marsh (2016a) created the **cooccur** package (Griffith, Veech, and Marsh 2016b), with incidence matrices similar to those of **netCoin**, but only using the hypergeometric distribution. They refer to other packages to detect pairs of species that share some space with one another such as **picante** (Kembel *et al.* 2010), **spaa** (Zhang 2016) and **vegan** (Oksanen *et al.* 2019).

In terms of similarity and distance calculations, packages like **stats** (R Core Team 2020), **proxy** (Meyer and Buchta 2019) and even **parallelDist** (Eckert 2018) cover most of the metrics that **netCoin** calculates. However, they do not include some of the coincidence analysis metrics needed, such as frequency, conditional frequency or statistical significance. In addition to this, **netCoin** allows the calculation of more than one metric at the same time just by calling one function. This reduces calculation time and thus improves performance.

A similar package to **netCoin** is **qgraph** (Epskamp, Cramer, Waldorp, Schmittmann, and Borsboom 2012; Epskamp, Costantini, Haslbeck, and Isvoranu 2020), which provides an interface to visualize data through network modeling techniques. However, **qgraph** is intended to represent a correlation matrix or a factor analysis statically, while **netCoin** is specialized in the representation of qualitative variables transformed into dichotomies and its parameters can be interactively changed through a web page.

In sum, despite the fact that there are many packages and software tools to analyze binary metrics and represent networks, **netCoin** adds value by providing the possibility to efficiently calculate a series of distance and similarity measures, including their statistical significance, and allowing the generation of interactive graphic output in HTML.

3. Coincidence analysis

Co-occurrences have been widely studied in many fields, especially in the content analysis of texts (Carley 1993; Lund and Burgess 1996; Popping 2000, 2003; Matsuo and Ishizuka 2004) and in the study of ecological communities (Diamond and Gilpin 1982; Connor and Simberloff 1983; Veech 2013). In addition to this, there is extensive literature that focuses on applications and many R packages that facilitate their analysis as seen in the previous section.

netCoin focuses on a particular form of dealing with co-occurrences, which is called coincidence analysis, and whose aim is to detect which people, subjects, objects, attributes or events tend to appear at the same time in different limited spaces (Diaconis and Mosteller 1989; Baumgartner 2009; Escobar 2015).

An event (j) is a potential outcome of a random experiment. The set of possible outcomes is called a sample space and is composed of a series of elementary mutually exclusive events.

A scenario (i) is each one of the results of a complex experiment made up of a set of events (X_j) with varying degrees of dependence between each other. A scenario can also be defined as a spatial and temporal set in which the researcher collects information on the events that take place.

Since the events of the scenarios are not mutually exclusive, they can be represented using

Scenarios	Head	Tail
I	1	0
II	1	1
III	1	1
IV	0	1

Table 1: Incidence matrix with 4 scenarios after tossing 2 coins.

Scenarios	Head	Tail
I	2	0
II	1	1
III	1	1
IV	0	2

Table 2: Occurrence matrix with 4 scenarios after tossing 2 coins.

	Head	Tail
Head	3	
Tail	2	3

Table 3: Coincidence matrix of the 4 scenarios of 2 coins.

dichotomous vectors (they can either occur or not) or vectors containing natural numbers (number of times each event occurs in a given scenario).

Therefore, the set of observed n scenarios can be represented as an *incidence matrix* ($\mathbf{I} = (x_{ij})$). In one dimension (generally the rows) the matrix contains the scenarios (i) and in the other dimension (commonly the columns) it contains the events (j). This matrix consists of 1s and 0s indicating if the events occurred or not, respectively, within the scenario. Alternatively, the occurrence matrix, which records the number of appearances of the event in every scenario, can be employed.

This distinction will be better understood with this simple example: If two coins are tossed four times, each toss represents a scenario where the events heads and tails are of interest. The three possible results for each toss of the two coins are: a) two heads, no tails; b) a head and a tail, and c) two tails and no head. The incidence matrix can be presented as shown in Table 1. On the other hand, the occurrence matrix must reflect the two heads or two tails obtained when the result is not head and tail (see Table 2).

Coincidence and co-occurrence matrices can be calculated from the incidence and occurrence matrices.

Definition. *Two coincident events (j and k) are those which occur together in the same scenario i.*

$$(x_{ij} > 0 \wedge x_{ik} > 0) \Rightarrow f_{ijk} = 1$$

Along with the basic coincidence in a given scenario i, when considering whether two events coincide in a multiple set of scenarios, the total number of coincidences of the events j and k can be obtained.

$$f_{jk} = \sum_{i=1}^{I} f_{ijk}$$

In addition, we can distinguish different degrees of coincidences. Thus, the most basic coincidence classification would distinguish between:

a. **No coincidence**: Two events that never occur in the same scenario, i.e., they are mutually exclusive ($f_{jk} = 0$).

b. **Simple coincidence**: Two events are merely coincident if they occur together in at least one scenario ($f_{jk} > 0$).

c. **Total coincidence**: Two events that always occur together in the same scenarios. If one of them occurs, then the other does too ($f_{jk} = f_{jj} = f_{kk}$). A special case is the subtotal coincidence in which the other event occurs only if the first occurs and not vice versa ($f_{jk} = f_{jj} > f_{kk}$), i.e., the occurrence of the more frequent event (k) does not necessarily imply the occurrence of the less frequent event (j).

From the incidence matrix, the *coincidence matrix* $\mathbf{F} = (f_{ij})$ can be calculated using this expression: $\mathbf{F} = \mathbf{I}^\top \mathbf{I}$. This is an example of how to project a bimodal network to a unimodal one. The elements of this matrix are either univariate (f_{jj}) or bivariate (f_{jk}) frequencies of the different events in the set of scenarios (i) contained in the rows of $\mathbf{I}$.

From the coincidence matrix ($\mathbf{F}$) three probabilistic measures can be derived:

a. The ***marginal probability*** of X_j, denoted as $\mathsf{P}(X_j)$, can be obtained by dividing the frequencies of each event (f_{jj}) by the total number of scenarios (n) in which it could have occurred:

$$\mathsf{P}(X_j) = \frac{f_{jj}}{n}.$$

b. The ***joint probability*** of two events X_j and X_k, expressed as $\mathsf{P}(X_{jk})$ is given by the frequency of occurrence in the same scenario divided by the set of scenarios considered in a given set:

$$\mathsf{P}(X_{jk}) = \frac{f_{jk}}{n}.$$

c. The ***conditional probability***, denoted as $\mathsf{P}(X_j|X_k)$, expresses the possibility that a certain event occurs when the second event has already occurred. It is obtained by dividing the joint probability by the marginal probability of the conditional event:

$$\mathsf{P}(X_j|X_k) = \frac{\mathsf{P}(X_{jk})}{\mathsf{P}(X_k)} = \frac{f_{jk}}{f_{kk}}.$$

With the conditional probability, we can create a coincidence gradient, the ***probable coincidence***, between two events when their conditional probability is greater than 50%:

$$\mathsf{P}(X_j|X_k) > 0.5.$$

When working with samples of scenarios instead of the whole universe, the upper limit of the confidence interval can be estimated under the alternative hypothesis of $\mathsf{P}(X_j|X_k) < 0.5$ using the formula

$$L_{\text{sup}} = \frac{f_{jk}}{f_{kk}} + \frac{t_{\alpha, f_{kk}-1}}{2\sqrt{f_{kk}}},$$

Type of coincidence	Definition	Asymmetric	Statistical test
Null	$f_{jk} = 0$	No	No
Simple	$f_{jk} > 0$	No	No
Probable	$f_{jk}/f_{kk} > 0.5$	Yes	Yes
Conditional	$f_{jk} > f^*_{jk}$	No	Yes
Subtotal	$f_{jk} = f_{jj} < f_{kk}$	Yes	No
Total	$f_{jk} = f_{jj} = f_{kk}$	No	No

Table 4: Types of coincidences.

where $t_{\alpha,f_{kk}-1}$ is the value of the Student distribution for $f_{kk} - 1$ degrees of freedom with a significance level of α.

The ***conditional coincidence*** is another coincidence gradient. It is derived from the concept of independence of events. Two events are independent if Equation 1 is true:

$$\mathsf{P}(X_j) = \mathsf{P}(X_j|X_k) \Longleftrightarrow \frac{f_{jj}}{n} = \frac{f_{jk}}{f_{kk}}. \tag{1}$$

Therefore, for that condition to be met, the following condition needs to be verified:

$$f^*_{jk} = \frac{f_{jj} f_{kk}}{n}.$$

From this equation, two events have a conditional coincidence when their frequency is greater than the expected (f^*_{jk}) under the assumption of independence:

$$f_{jk} > \frac{f_{jj} f_{kk}}{n} = f^*_{jk}.$$

It is also known (Haberman 1973) that the difference between f_{jk} and f^*_{jk} assumes asymptotically a normal distribution with the following standard error:

$$\sqrt{f^*_{jk}(1 - f_{jj}/n)(1 - f_{kk}/n)},$$

which can be used to standardize (r_{jk}) the difference between the empirical frequency of coincident events (f_{jk}) and the expected frequency (f^*_{jk}) under the assumption of mutual independence:

$$r_{jk} = \frac{f_{jk} - f^*_{jk}}{\sqrt{f^*_{jk}(1 - f_{jj}/n)(1 - f_{kk}/n)}}.$$

For small samples, the one-sided Fisher exact test, which employs the hypergeometric distribution should be used instead (Fisher 1935; Finney 1948).

The degrees of coincidence that can be detected between each pair of events is summarized in Table 4.

3.1. Coincidence metrics

In addition to classifying coincidences into different types, they can be measured using binary proximity metrics (Hubálek 1982; Gower 1985). These measures have a maximum value of

	Event X_k	
Event X_j	Present	Absent
Present	a	b
Absent	c	d

Table 5: Contingency table.

one when there is total coincidence between two dichotomous events and a value of 0 when there is total independence between them. Some of them can take negative values, in which case the minimum value could be -1 when two incompatible events are implied.

For the calculation of these metrics each element (f_{jk}) of the coincidence matrix can be split into the following system equivalences:

$$
\begin{aligned}
a &= f_{jk} \\
b &= f_{jj} - f_{jk} \\
c &= f_{kk} - f_{jk} \\
d &= n - f_{jj} - f_{kk} + f_{jk}
\end{aligned}
$$

Therefore, for each pair of events, Table 5 can be elaborated. With these four figures (a, b, c, d), representing the frequencies of the four states of presence/absence of two events in the set of scenarios studied, binary proximity measures are obtained.

These coefficients or binary proximity metrics can be classified into four types: The *first* one includes metrics that are similar to that of *matching* (Rogers and Tanimoto 1960; also known as Rogers and Tanimoto). They are the result of divisions with a numerator with both positive coincidences (the two events occur in the same scenario) and negative coincidences (the two events are absent in the same scenario), and a denominator where all scenarios are considered with different weights. The metrics belonging to this category are *Rogers* (Rogers and Tanimoto 1960), *Sneath* (Sneath and Sokal 1962), *Anderberg* (1973) and *Gower* (1985). These measurements should be used when considering coincidence both when two events are present in the same scenario, as well as when both are not present.

$$
\begin{aligned}
Matching &= \frac{a+d}{a+b+c+d} \\
Rogers &= \frac{a+d}{(a+d)+2(b+c)} \\
Sneath &= \frac{2(a+d)}{2(a+d)+(b+c)} \\
Anderberg &= \left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{c+d} + \frac{d}{b+d}\right)/4 \\
Gower &= \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}
\end{aligned}
$$

In the *second* type of metrics there is *Jaccard* (1901). Here, scenarios where neither of the two events whose coincidence degree we intend to measure (d) are excluded. Therefore, neither the numerator nor the denominator include those scenarios without any of the two events.

Metrics of this type also include *Dice* (Jaccard 1901), *Antidice* (Anderberg 1973), *Ochiai* (1957) and *Kulczynski* (1927). In this case, events that are not present in the same scenario are not considered to be coincident, and only those scenarios where at least one event has occurred are coincident.

$$
\begin{aligned}
Jaccard &= \frac{a}{a+b+c} \\
Dice &= \frac{2a}{2a+b+c} \\
Antidice &= \frac{a}{2+2(b+c)} \\
Ochiai &= \frac{a}{\sqrt{(a+b)(a+c)}} \\
Kulczynski &= (\frac{a}{a+b} + \frac{a}{a+c})/2
\end{aligned}
$$

The *third* type of similarity metrics for binary data only includes *Russell and Rao* (1940). It only considers those scenarios to be similar in which both events occur. It excludes from the numerator those in which none of the events occurs, considering that this does not indicate that the scenarios are similar. However, unlike the similarity metrics such as *Jaccard*'s, all the possible scenarios are present in the denominator of the equation. This coincidence measure only takes into account coincident events and contemplates all scenarios, including those in which both events are not present. Logically, if there are no scenarios where neither of the two is present, then both are equal. However, if within an infinite number of scenarios neither of the two events existed, the value of Russell and Rao would be zero, while Jaccard would be 1 by convention.

$$
Russell\ and\ Rao = \frac{a}{a+b+c+d}
$$

Finally, in the *fourth* type we may include all metrics in which frequencies of coincidences (whether the events occur or not) are compared (subtracted) with frequencies of no coincidences (scenarios where an event occurs but the other one does not). Thus, these measurements can be positive if coincident events predominate, or negative otherwise, i.e., when the scenarios in which the events do not coincide predominate. Metrics of this type include *Pearson* (1900), *Yule* (1900) and *Hamann* (1961). This modality is similar to the correlation coefficients and has the advantage of presenting both positive and negative values. Positive values imply that whenever an event is present, the other is as well; while negative ones evidence that in most cases, the presence of an event implies the absence of the other.

$$
\begin{aligned}
Pearson &= \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \\
Yule &= \frac{ad-bc}{ad+bc} \\
Hamann &= \frac{(a+d)-(b+c)}{a+b+c+d}
\end{aligned}
$$

All the previous expressions are called similarity metrics. To turn them into distance measurements, the following expression can be used $distance = 1 - similarity$. If the metric has a range between 0 and 1, then these limits are preserved, although with a different meaning, as

Type	Measures (abbreviation for procedures)
Frequencies	Frequencies (f), Relative frequencies (x), Conditional frequencies (i, ii)*
Degrees	Coincidence degree (cc), Probable degree (cp)
Expected values	Expected (e), Confidence interval (con)
Matching	*Matching* (m), *Rogers* (t), *Gower* (g), *Sneath* (s), *Anderberg* (and)
Jaccard	*Jaccard* (j), *Dice* (d), *antiDice* (a), *Ochiai* (o), *Kulczynski* (k)
Russell	*Russell* (r)
Pearson	*Pearson* (p), *Haberman* (h), *Yule* (y), *Hamann* (ham), *odds ratio* (od)
Probabilistic	p value of Haberman (z), hypergeometric p greater value (hyp)

Table 6: Similarity measures (* i: conditioned by the source frequency; ii: conditioned by the target frequency).

the 0 indicates complete coincidence. Nevertheless, if the metric range is between -1 and $+1$, the new similarity metric will be between 0 and 2, with 1 indicating complete independence and higher values meaning that two events coincide less often than by mere chance.

An outline of these measures and the abbreviations to obtain them with **netCoin** can be found in Table 6.

3.2. Adjacency matrix

Coincidence and distance matrices have been covered. Both types can be transformed into adjacency matrices. An adjacency matrix connects each pair of events depending on whether their coincidence metric is above a certain value. Thus, it is a square matrix with as many rows and columns as the number of events being studied, and formed by elements representing the number of coincidences between every pair of events. Using all the previous metrics, adjacency matrices can be formed in the following ways:

a. With the simple coincidences so that there will be a connection between two events provided that they have coincided in a single scenario.

Frequency matrix

	odd	even	small	large
odd	54			
even	0	46		
small	41	13	54	
large	13	33	0	46

Adjacency matrix

	odd	even	small	large
odd	–			
even	0	–		
small	1	1	–	
large	1	1	0	–

b. With total or subtotal coincidences so that two completely overlapping events will be connected. In the first category, it will be a symmetrical connection, and in the case of subtotal coincidences, it will only connect the less frequent category and the most frequent ones.

Conditional frequencies

	odd	even	small	large
odd	100.0			
even	0.0	100.0		
small	75.9	28.3	100.0	
large	24.1	71.7	0.0	100.0

Adjacency matrix

	odd	even	small	large
odd	–			
even	0	–		
small	0	0	–	
large	0	0	0	–

c. With the probable or conditional coincidences, connecting events with more than 50% probability in the first case and a positive residual (r_{jk}).

Standardized residuals (r_{jk})

	odd	even	small	large
odd	100.0			
even	−10.0	100.0		
small	4.8	−4.8	100.0	
large	−4.8	4.8	0.0	100.0

Adjacency matrix

	odd	even	small	large
odd	–			
even	0	–		
small	1	0	–	
large	0	1	0	–

d. With the statistical tests applied to the probable or conditional coincidences, in which case we could have statistically significant coincidences with different degrees or levels of significance (0.05, 0.01, 0.001, 0.0001, ...).

Significance of r_{jk}

	odd	even	small	large
odd	–			
even	1.0e+00	–		
small	3.2e–06	1.0e+00	–	
large	1.0e+00	3.2e–06	1.0e+00	–

Adjacency matrix

	odd	even	small	large
odd	–			
even	0	–		
small	1	0	–	
large	0	1	0	–

e. With the coincidence metrics, in which case one of the 14 possible coincidences must be chosen, setting a threshold (0.50, for instance) from which it can be considered that two events are coincident.

Jaccard's similarity

	odd	even	small	large
odd	1.00			
even	0.00	1.00		
small	0.61	0.15	1.00	
large	0.15	0.56	0.00	1.00

Adjacency matrix

	odd	even	small	large
odd	–			
even	0	–		
small	1	0	–	
large	0	1	0	–

3.3. Layouts

The same way that a series of coincidences can become an adjacency matrix, the latter can be converted into a graph. As previously said, a graph $\mathcal{G}$ consists of "two sets of information: a set of nodes (events), $\mathcal{N}= \{n_1, n_2, \ldots, n_g\}$, and a set of lines (coincidences), $\mathcal{L}= \{l_1, l_2, \ldots, l_L\}$ between a pair of nodes"(Wasserman and Faust 1994).

Layout	Argument	Abbreviation
Random disposition of vertices	`"layout.random"`	`"ra"`
Rectangular grid disposition	`"layout.grid"`	`"gr"`
Circle distributed vertexes	`"layout.circle"`	`"ci"`
Star disposition of vertices	`"layout.star"`	`"st"`
Fruchterman and Reingold	`"layout.fruchterman.reingold"`	`"fr"`
Kamada and Kawai	`"layout.kamada.kawai"`	`"ka"`
Forced directed layout (GEM)	`"layout.gem"`	`"ge"`
Simulated annealing algorithm	`"layout.davidson.harel"`	`"da"`
Multidimensional scaling coordinates	`"layout.mds"`	`"md"`
Tidy arrangement of vertices	`"layout.reingold.tilford"`	`"re"`
Layered directed acyclic graphs	`"layout.sugiyama"`	`"su"`
Large scale graphs	`"layout.drl"`	`"dr"`
Large graph layout	`"layout.lgl"`	`"lg"`

Table 7: Layout algorithms. References: fr, Fruchterman and Reingold (1991); ka, Kamada and Kawai (1989); ge, Frick *et al.* (1995); da, Newman (2006); md, Cox and Cox (2001); re, Reingold and Tilford (1981); su, Sugiyama *et al.* (1981); dr, Martin *et al.* (2008); lg, Martin *et al.* (2008).

An additional problem is where to draw each node, i.e., the spatial distribution of the nodes. Thanks to **igraph**, **netCoin** can be laid out according to the criteria in Table 7.

If none of these layouts are indicated, **netCoin** uses a dynamic Fruchterman-Reingold algorithm by default. Alternatively, the user can provide a matrix with two columns indicating the coordinates of those nodes that are going to be fixed in the representation. Leftover nodes should be stated as `NA` and would be placed according to a forced directed mechanism.

3.4. Communities

Cluster analysis is "a set of methods for constructing a (hopefully) sensible and informative classification of an initially unclassified set of data, using the variable values observed on each individual" (Everitt 2003). In agglomerative hierarchical clustering methods, there are various procedures to join cases using dendrograms: single, complete, average, median, Ward, etc. In the coincidence analysis, clustering could be useful to classify events according to their concurrences, using the Haberman residuals (r_{jk}) or another distance matrix (geodesic, matching, Jaccard, ...) as inputs to the clustering method.

Events j and k are structurally equivalent if, for all events, $l = 1, 2, \ldots, g\ (l \neq j, k)$, and for all associations $r = 1, 2, \ldots, R$, event j has a relation to l if and only if k also has a relation to l. Consequently, structurally equivalent events are those that have identical edges with the rest of events. Structural equivalence can imply "community", but it does not have to (e.g., if each community consists of a standard set of hierarchical actors), and community does not have to imply structural equivalence. Events can be partitioned into subsets of structural equivalence using a *hierarchical clustering* or a similar algorithm of classification. **netCoin** allows us to obtain the **igraph** procedures listed in Table 8.

Community	Argument	Abbreviation
Edge-betweenness	`"cluster_edge-betweenness"`	`"ed"`
Fast-greedy	`"cluster_fast_greedy"`	`"fa"`
Label propagation	`"cluster_label_prop"`	`"la"`
Leading eigenvector	`"cluster_leading_eigen"`	`"le"`
Louvain	`"cluster_louvain"`	`"lo"`
Optimal modularity	`"cluster_optimal"`	`"op"`
Sping glass	`"cluster_spinglass"`	`"sp"`
Walktrap	`"cluster_walktrap"`	`"wa"`

Table 8: Communities algorithms. References: ed, Girvan and Newman (2002); fa, Wakita and Tsurumi (2007); la, Raghavan *et al.* (2007); le, Newman (2006); lo, Blondel *et al.* (2008); op, Good *et al.* (2009); sp, Reichardt and Bornholdt (2006); wa, Pons and Latapy (2006).

4. The R package netCoin

Some of **netCoin**'s statistical and graphical features were originally implemented in Stata (StataCorp 2019) as the **coin** ado program (Escobar 2015). This initial Stata program lacked the graphical interactivity which provides agile data exploratory capabilities. That is the main reason why R was chosen to generate an extended version of the original **coin** program.

Firstly, the **shiny** (Chang, Cheng, Allaire, Xie, and McPherson 2020) and **igraph** packages were used to achieve graph results, but what provided the solution to accomplish the desired interactivity was the integration with the **D3.js** data visualization library (Bostock, Ogievetsky, and Heer 2011). In addition to this, R code has been written to obtain the coincidence metrics and their significance.

4.1. Installation

The **netCoin** package is available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=netCoin and has dependencies on three other R packages **igraph** (Csardi and Nepusz 2006), **Matrix** (Bates and Mächler 2019) and **haven** (Dusa and Thiem 2020) which are loaded with **netCoin**.

```
R> install.packages("netCoin")
R> library("netCoin")
```

4.2. Overview with three simple examples

The **netCoin** package incorporates every coincidence analysis element detailed in Section 3. The functions included help the analyst convert the data into an incidence matrix that is suitable for the analysis, produce the coincidence matrix, calculate all the statistical indicators, generate the nodes and edges of the graph, produce interactive network visualizations and export those networks as '`igraph`' objects.

The basic input is an incidence binary matrix, which can be obtained with the function `dichotomize()` in case of absence. This function can be applied to both character variables and factor variables. In addition to this, among the former it is able to split fragments separated by a constant chain, whose default value is the null character (`""`).

Argument	Meaning
`sep = ""`	The separator in case that the variables are composed.
`min = 1`	Minimum frequency of the value of a variable to be considered as an event.
`length = Inf`	Maximum number of events to be considered.
`values = NULL`	Events to be converted into dichotomies (not for multiple composed variables).
`sparse = FALSE`	Produce a sparse matrix instead of a data frame.
`add = TRUE`	Add the new columns to the original data frame.
`sort = TRUE`	Order the new columns by their frequencies.

Table 9: Arguments of function `dichotomize`.

In addition to the data frame and the variable or variables to be dichotomized, the arguments of this function are given in Table 9.

The simplest example can be applied to the `dice` data frame included in the package:

```
R> data("dice", package = "netCoin")
R> events <- dichotomize(dice, "dice", add = FALSE, sort = FALSE)
R> head(events)

   1 2 3 4 5 6 dice:None
V1 1 0 0 0 0 0         0
V2 0 1 0 0 0 0         0
V3 0 0 0 0 1 0         0
V4 0 0 0 1 0 0         0
V5 0 1 0 0 0 0         0
V6 0 0 0 0 1 0         0
```

Thus, a new data frame with 6 columns corresponding to the six possible events of throwing a dice would be obtained.

We would have to add the argument `sep =` in case of factor variables composed of several events. As a second example, imagine that we tossed two coins in unison ten times into the air. The results could be `"H,H"`, `"T,H"`, `"H,T"`, `"T,T"`, each with the same probability. Therefore, to convert the events of each toss into elementary events, we use `dichotomize()` with the argument `sep = ","`.

```
R> set.seed(10)
R> coins <- data.frame(coin = cut(runif(10), c(0, 0.25, 0.50, 0.75, 1),
+    labels = c("H,H", "T,H", "H,T", "T,T")))
R> events <- dichotomize(coins, "coin", sep = ",")
R> events

    coin H T coin:None
V1   H,T 1 1         0
V2   T,H 1 1         0
```

```
V3   T,H 1 1         0
V4   H,T 1 1         0
V5   H,H 1 0         0
V6   H,H 1 0         0
V7   T,H 1 1         0
V8   T,H 1 1         0
V9   H,T 1 1         0
V10  T,H 1 1         0
```

Once we have an incidence matrix, we obtain a '`coin`' object, a list composed by the number of events and the coincidence matrix, with the function `coin()`. Then, the function `edgeList()` converts a '`coin`' object into a data frame containing an edge list with the similarity measures stated in the procedure argument. By default, `edgeList()` produces Haberman residuals with their p values. The third example considers the presence of three people (`"Man"`, `"Woman"` and `"Undet."`) in four different scenarios.

```
R> frame <- data.frame(A = c("Man; Woman", "Woman; Woman", "Man; Man",
+    "Undet.; Woman; Man"))
R> data <- dichotomize(frame, "A", sep = "; ")[2:4]
R> coin <- coin(data)
R> coin

n= 4
       Man Woman Undet.
Man      3
Woman    2     3
Undet.   1     1      1

R> edges <- edgeList(coin)
R> edges

  source target  Haberman         Z
3    Man Undet. 0.6666667 0.2707349
6  Woman Undet. 0.6666667 0.2707349
```

Finally, the function `netCoin()` can mix the nodes (extracted from the '`coin`' object) with the edge list data frames in order to produce a '`netCoin`' object, and if the argument `dir = "directory"` is used, a directory will be created with a graph within a web page whose main file is named `index.html`.

The '`netCoin`' object has three methods: `print()` shows a sample (until 6) of nodes and links with their attributes, `summary()` shows the basic statistics of the nodes, and `plot()` shows the corresponding graph in the computer's default browser.

```
R> nodes <- asNodes(coin)
R> netCoin(nodes, edges)
R> (net <- netCoin(nodes, edges))
R> print(net)
```

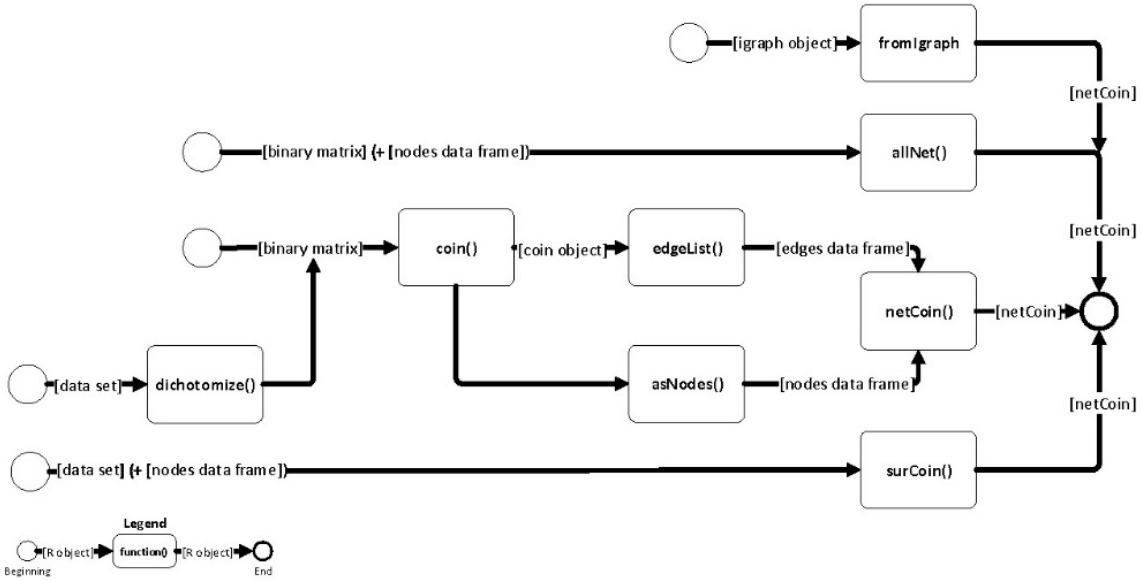


Figure 1: **netCoin** processes to create a graph.

```
Nodes(3):
         name frequency
Man       Man         3
Woman   Woman         3
Undet. Undet.         1

Links(2):
  source target  Haberman         Z
3    Man Undet. 0.6666667 0.2707349
6  Woman Undet. 0.6666667 0.2707349
```

```
R> summary(net)

3 nodes and 2 links.
frequency distribution of nodes:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   2.333   3.000   3.000
Haberman's distribution of links:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6667  0.6667  0.6667  0.6667  0.6667  0.6667
```

```
R> plot(net)
```

Alternatively, the '`netCoin`' object could be obtained directly from a binary incidence matrix with the `allNet()` function, where more than 40 arguments can be controlled, although the only required argument is the incidence matrix. However, if we want to obtain the directory

with the graph, we must add the `dir = "directory"` argument. Other sources to obtain a '`netCoin`' object are an '`igraph`' object with the function `fromIgraph`, and a data set with factor variables with the function `surCoin()`. See the four functions that obtain a '`netCoin`' object in Figure 1.

```
R> frame <- data.frame(A = c("Man; Woman", "Woman; Woman", "Man; Man",
+    "Undet.; Woman; Man"))
R> data <- dichotomize(frame, "A", sep = "; ")[2:4]
R> allNet(data)
```

Using the previous `data` data frame, a set of coincidence measures and their significance can be printed with the `edgeList` function, whose input must be a '`netCoin`' object.

```
R> edgeList(coin(data),
+    proc = c("frequency", "Jaccard", "Pearson", "Haberman", "Z", "fisher"),
+    criteria = "fisher", max = 1)
```

```
  Source Target coincidences  Jaccard   Pearson  Haberman     p(Z) p(Fisher)
1    Man  Woman            2 0.500000 -0.333333 -0.666667 0.729265      1.00
2    Man Undet.            1 0.333333  0.333333  0.666667 0.270735      0.75
3  Woman Undet.            1 0.333333  0.333333  0.666667 0.270735      0.75
```

4.3. Other examples

Multigraph coincidence analysis with data of families from Renaissance Italy

The following example uses data about families from Renaissance Italy from Padgett and Ansell (1993). It consists of a data frame (families) with information about Italian families of the Renaissance, and another data frame (links) with the marriage and business bonds between families.

```
R> data("families", package = "netCoin")
R> data("links", package = "netCoin")
```

The previous `coin()`, `edgeList()`, `asNodes()` and `netCoin()` functions can be executed together with the `allNet()` function where several arguments can be specified (see Table 10).

Two networks are generated representing the business and marriage bonds between the two families with the following commands.

```
R> G <- allNet(incidence = links[links$link == "Marriage", -17],
+    nodes = families, layout = "md", criteria = "f", minL = 1,
+    size = "frequency", color = "seat",
+    main = "Marriage links between Italian families",
+    note = "Data source: Padgett & Ansell (1983)")
```

Argument	Meaning
`incidence`	A data frame that contains the incidence matrix.
`nodes`	A data frame with at least one vector of names.
`layout`	The algorithm selected for the network topology.
`criteria`	The statistical criteria to be used for the selection of the edges.
`minL`	Minimum value of the statistic to represent the edge in the graph.
`size`	Name of the vector with size in the nodes data frame.
`color`	Name of the vector with color variable in the nodes data frame.
`main`	Upper title of the graph.
`note`	Lower title of the graph.

Table 10: Arguments of function `allNet`.

Function	Description
`dichotomize`	Function to convert factor or character column(s) in a data frame into a set of dichotomous columns. Their names will correspond to the labels or text of every category.
`coin`	This function generates a '`coin`' object from an incidence matrix data frame. A '`coin`' object consists of a list with two elements: the number of scenarios, and a coincidence matrix of events, whose main diagonal figures are the frequency of events and outside the said diagonal there are conjoint frequencies of these events
`asNodes`	From a '`coin`' object, this function generates a data frame of nodes.
`edgeList` (`sim`)	Function to convert a coincidence matrix into an edge list calculating a variety of coincidence (proximity) metrics. The `sim` function produces the same information, but as a list of proximity matrices instead.
`netCoin`	The `netCoin` function produces an interactive '`netCoin`' object from two data frames: one including nodes with attributes, and another one containing edges also with its own attributes.
`multigraphCreate`	This function produces an interactive multinetwork with several '`netCoin`' objects.
`fromIgraph`	From an '`igraph`' object, this function generates a '`netCoin`' object.
`toIgraph`	With this function an '`igraph`' object is generated from a '`netCoin`' object.
`allNet`	Produces a '`netCoin`' object from a data frame or a matrix with dichotomous values.
`surCoin`	Produces a '`netCoin`' object from a data frame with factor variables accepting also a '`tbl_df`' class (see package **haven**).

Table 11: **netCoin** main functions.

```
R> H <- allNet(incidence = links[links$link == "Business", -17],
+    nodes = families, layout = "md", criteria = "f", minL = 1,
+    size = "frequencb", color = "seat",
+    main = "Marriage links between Italian families",
+    note = "Data source: Padgett & Ansell (1983)")
```

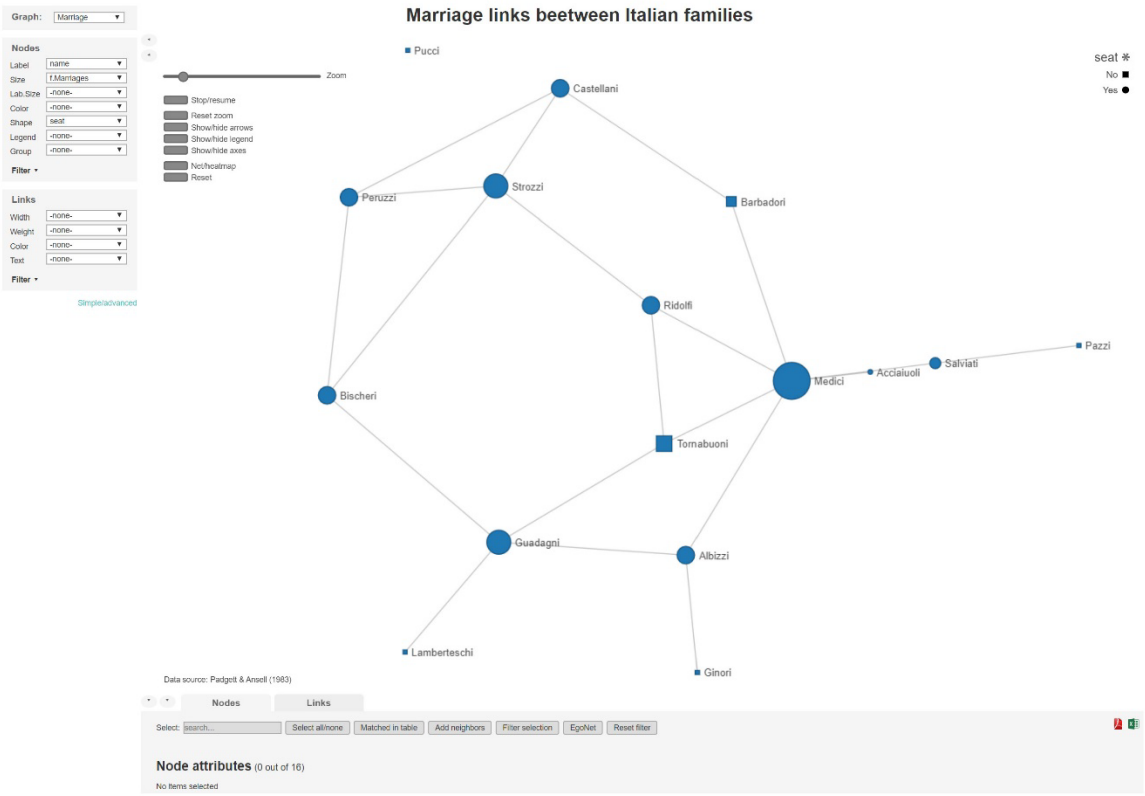


Figure 2: Graph of links between Italian families.

```
R> G

Title: Marriage links between Italian families

Nodes(16):
       name f.Marriages f.Business wealth priorates seat
 Acciaiuoli           1          0     10         53  Yes
    Albizzi           3          0     36         65  Yes
  Barbadori           2          4     55          0   No
   Bischeri           3          3     44         12  Yes
 Castellani           3          3     20         22  Yes
     Ginori           1          2     32          0   No
...

Links(20):
     source     target frequencies
    Albizzi   Guadagni           1
    Albizzi     Medici           1
    Albizzi     Ginori           1
 Acciaiuoli     Medici           1
  Barbadori Castellani           1
```

```
  Barbadori      Medici            1
...

Data source: Padgett & Ansell (1983)
```

The 'netCoin' object G (as well as the non-shown H) is composed of two data frames. In the first (nodes) there are the families' attributes: frequency of marriage links (f.Marriages), frequency of business links (f.Business), a wealth index (wealth), number of priories held (priorates) and holding of at least one priorate (seat). In every row of the links data frame there are two families with a column indicating the existence of a link (coincidence) between them.

Once the two networks are ready, the function multigraphCreate() generates both graphs in the specified directory (see Figure 2).

```
R> multigraphCreate(Marriage = G, Business = H, dir = "italian")
```

Sanderson's analysis of species co-occurrences

This section uses one of the most renowned data examples in ecology. Charles Darwin compiled data about 13 species of finches and 17 of the Galápagos Islands (Sanderson 2000) on which they could be found.

We prepare the nodes' attributes (finches) and their incidences in the islands (Galapagos). Afterwards, we have to add the images in a specific directory in order to refer to them in the allNet() function.

```
R> data("Galapagos", package = "netCoin")
R> data("finches", package = "netCoin")
R> finches$species <- system.file("extdata", finches$species,
+    package = "netCoin")
```

Here, a few extra features are added to the graph shown in Figure 3:

- criteria = "hyp": The statistical criteria to be used for the strength of the edges.
- maxL = 0.05: Maximum value of the statistic to include the edge in the list.
- lwidth = "Haberman": Name of the vector with width variable in the links data frame.
- lweight = "Haberman": Name of the vector with weight variable in the links data frame.
- image = "file": Name of the vector with image files in the nodes data frame.
- layout = "mds": The algorithm selected for the network topology.

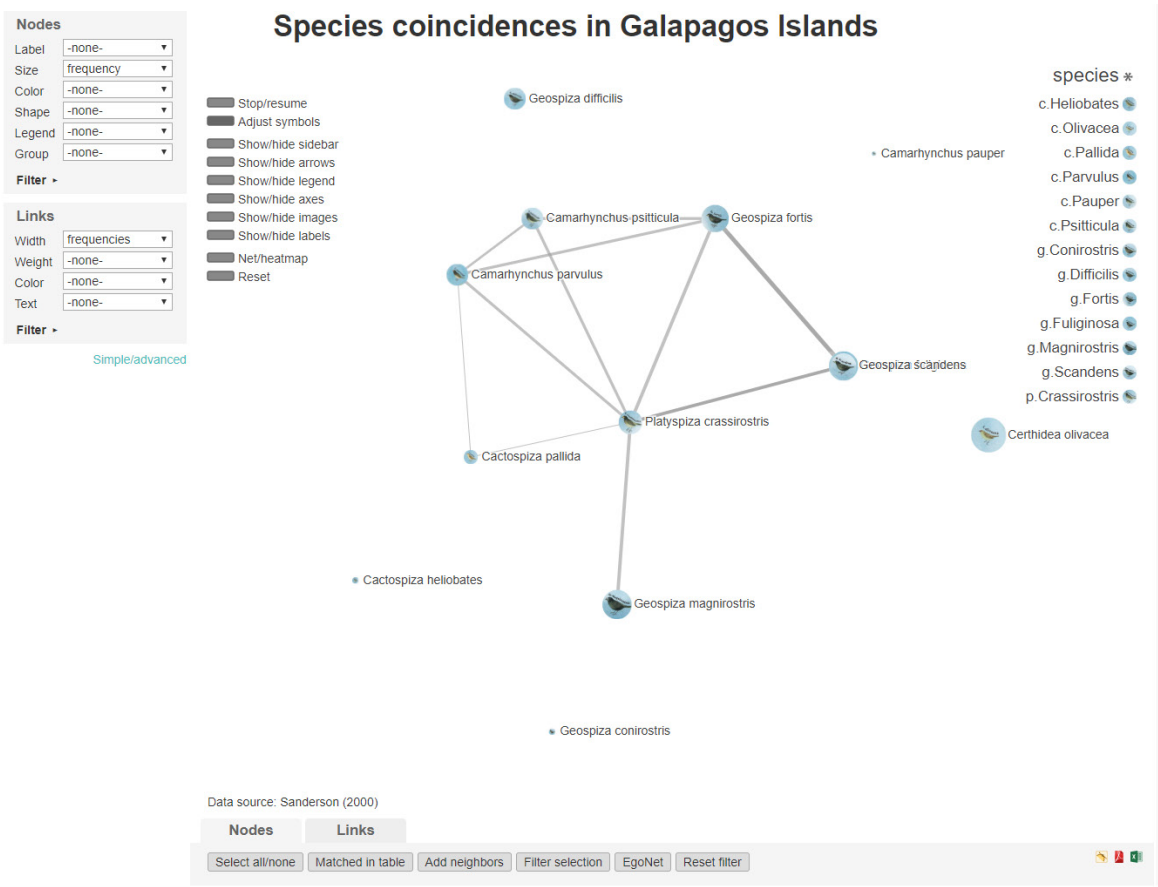


Figure 3: Graph of finches coincidences in Galápagos Islands.

```
R> Net <- allNet(Galapagos,
+     frequency = TRUE, procedures = "frequencies", criteria = "hyp",
+     layout = "mds", nodes = finches, maxL = 0.05, size = "frequency",
+     image = "species", lwidth = "frequencies", cex = 1.35, controls = 2:4,
+     main = "Species coincidences in Galapagos Islands",
+     note = "Data source: Sanderson (2000)")
R> Net

Title: Species coincidences in Galapagos Islands

Nodes(13):
                   name frequency        %     type
 Geospiza magnirostris        14 82.35294 Geospiza
       Geospiza fortis        13 76.47059 Geospiza
   Geospiza fuliginosa        14 82.35294 Geospiza
   Geospiza difficilis        10 58.82353 Geospiza
     Geospiza scandens        12 70.58824 Geospiza
  Geospiza conirostris         2 11.76471 Geospiza
...
```

```
Links(14):
                Source                    Target frequencies   p(Fisher)
 Geospiza magnirostris Platyspiza crassirostris          11 0.029411765
       Geospiza fortis      Geospiza fuliginosa          13 0.005882353
       Geospiza fortis        Geospiza scandens          12 0.002100840
       Geospiza fortis  Camarhynchus psitticula          10 0.014705882
       Geospiza fortis    Camarhynchus parvulus          10 0.014705882
       Geospiza fortis Platyspiza crassirostris          11 0.006302521
...

Data source: Sanderson (2000)
```

In this example, the only attributes of nodes are `frequency`, percentage (%) and `type`. The column `specs` has been suppressed because it is used to create the images from the images file names. More importantly, the links attributes are 1) `frequencies`, for example the number of coincidences of source and target finches, and 2) `p(Fisher)`, which is the error probability of rejecting the one-side alternative hypothesis, in case that it is true that two species are not coincident on each island (scenario).

Once the '`netCoin`' object is ready, the function `plot()` generates its graphical representation in a temporary directory (see Figure 3), or in the directory specified in the `dir` argument. In this way, all the necessary files to be deposited in a web server are saved so that anyone can view them and interact with them using a browser.

```
R> plot(Net)
```

Graphical comparison of two networks

netCoin can also be used to graphically compare networks of co-occurrences. For instance, the previous graph of the Galápagos Islands finches (`Net`) can be compared with a random null model obtained from the same data with the function `cooc_null_model()` of the **EcoSimR** package (Gotelli, Hart, and Ellison 2015). Among the possibilities offered by this program, we opted for the nullity of co-occurrences and the Sim9 algorithm, which is a sequential swap (Gotelli 2000; Strona, Nappo, Boccacci, Fattorini, and San-Miguel-Ayanz 2014).

Once the theoretical or null model is randomly obtained (`nullData`), it could be analyzed and represented with the command `allNet()` assessing the significance of its co-occurrence links. Previously, in order to better compare the empirical data obtained by Darwin with the random null model data, the positions of the nodes of the null model are set using those of the empirical model. After using the hypergeometric distribution (`criteria = "hyp"`) and a level of significance of 0.05 (`maxL = 0.05`), the new graph (`NullNet`) only has two co-occurrences out of the possible 78 (paired combinations of 13 fiches).

To represent these two or more networks at the same time, the function `multigraphCreate()` is used with the parallel argument assigned as true. It can be observed (Figure 4) that the species are located in the same place and have the same size, proportional to their presence in the islands, but the number of links is much smaller, because they have been randomized and a filter of significance in the argument of the `allNet()` function has been set.

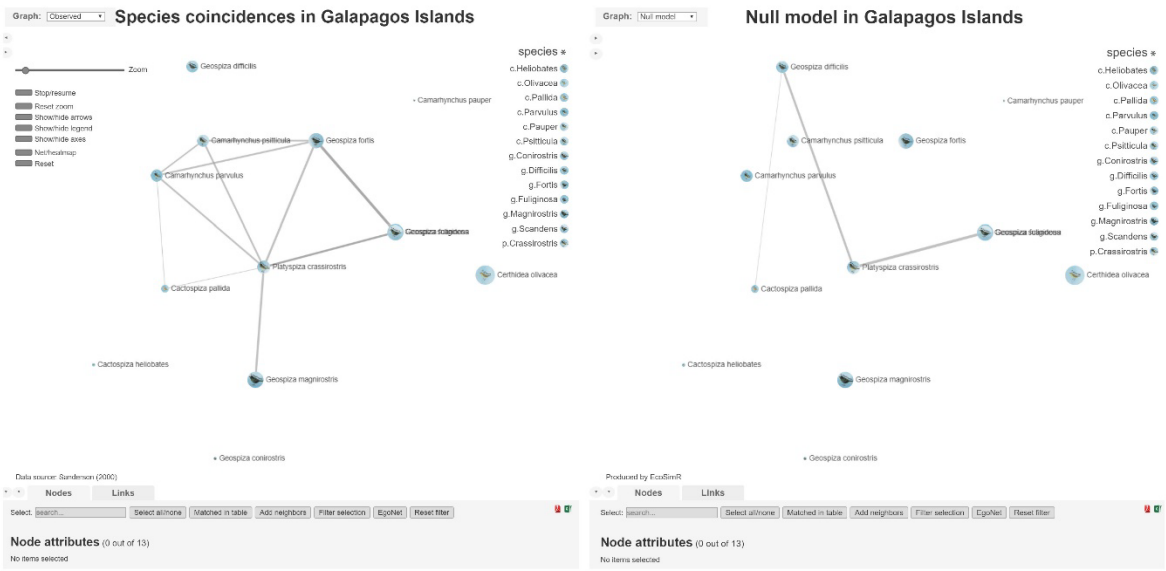


Figure 4: Graph of finches coincidences in Galápagos Islands.

```
R> library("EcoSimR")
R> layout.Net <- cbind(Net$nodes$fx, Net$nodes$fy)
R> set.seed(2016)
R> nullModel <- cooc_null_model(t(Galapagos), nReps = 1000, burn_in = 500,
+    algo = "sim9", metric = "checker")
R> nullData <- t(nullModel$Randomized.Data)
R> colnames(nullData) <- colnames(Galapagos)
R> NullNet <- allNet(nullData, frequency = TRUE, procedures = "frequencies",
+    criteria = "hyp", maxL = 0.05, layout = layout.Net, nodes = finches,
+    size = "frequency", image = "species", lwidth = "frequencies",
+    cex = 1.4, controls = 2:3, main = "Null model in Galapagos Islands",
+    note = "Produced by EcoSimR")
R> multigraphCreate("Observed" = Net, "Null model" = NullNet,
+    mode = "parallel")
```

Survey analysis

Another interesting use for **netCoin** is that of survey analysis applied to explore relationships between variables including those from multiple choice questions. The straightforward analysis shown below uses the package **haven** (Dusa and Thiem 2020) to read a SPSS (IBM Corporation 2017) survey demo file. Three variables are selected for the analysis: `gender`, `inccat` (income category in thousands) and `carcat` (primary vehicle price category).

The `plot()` function is applied to the result of the `surCoin()` function with those three variables as inputs. This produces the graph in Figure 5 where the male node is connected to the lowest and highest incomes as well as the economy and luxury vehicle categories. On the other hand, the female node is linked to income categories in the middle range and either the standard or the luxury vehicle price category.

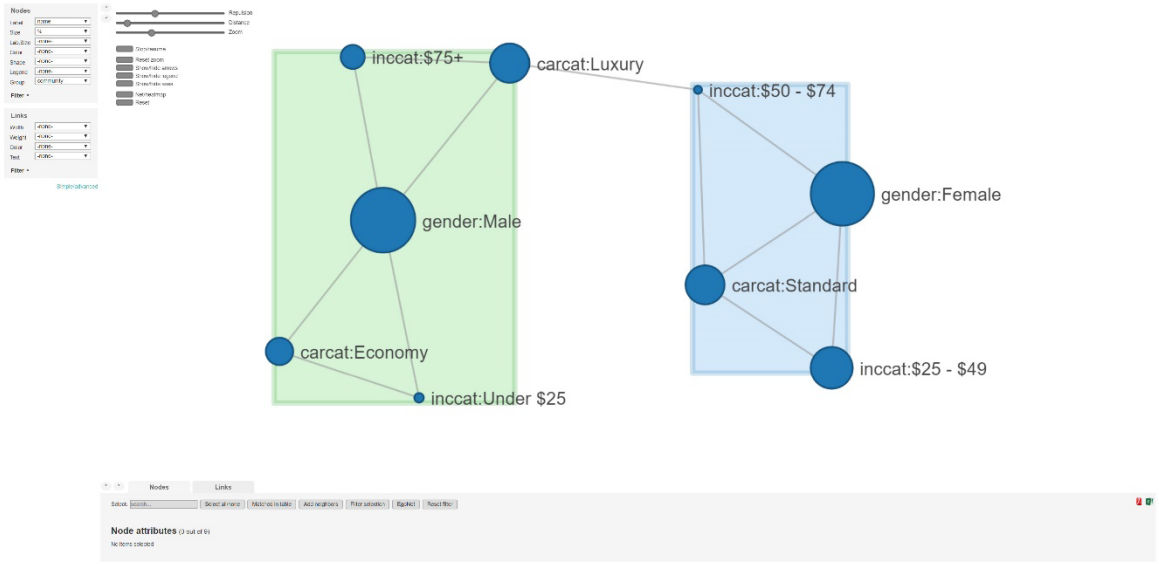


Figure 5: Graph of survey data multiresponse question.

```
R> library("haven")
R> survey <- read_spss(file = "demo.sav")
R> variables <- c("gender", "inccat", "carcat")
R> plot(surCoin(survey, variables, communities = "Louvain"))
```

4.4. Performance

To test the **netCoin** performance, several random datasets were generated with a different number of cases (1,000 and 50,000) and events (10, 50, 100). Tests were performed with six datasets: M(1,000× 10), M(1,000×50), M(1,000×100), M(50,000×10), M(50,000×50), M(50,000 ×100). Calculations for Jaccard were compared using **netCoin** and **parallelDist**. The results show faster times for **parallelDist** when the number of cases or events is smaller. But when the number of cells (cases times events) grows, then **netCoin** offers better results as shown by Figure 6. As time grows exponentially with the number of cells, time is represented by its logarithmic values in this figure.

The package produces interactive graphs that work well with up to 1500 edges. Using more than 1500 edges makes the interaction with the graph slow due to browser memory limitations.

5. Concluding comments

The **netCoin** package offers an opportunity for the interactive analysis and visualization of data sets composed of every kind of data insofar as variables are dichotomized. It contains a large variety of similarity measures to connect the events that co-occur in the same scenarios.

In order to select the relevant coincidences, **netCoin** incorporates two models of probability: the normal distribution through the Haberman residuals for a large number of scenarios, and the hypergeometric model for small data collections. Its main aim is to represent coincidences through a graph, which is particularly useful when many events are to be analyzed.

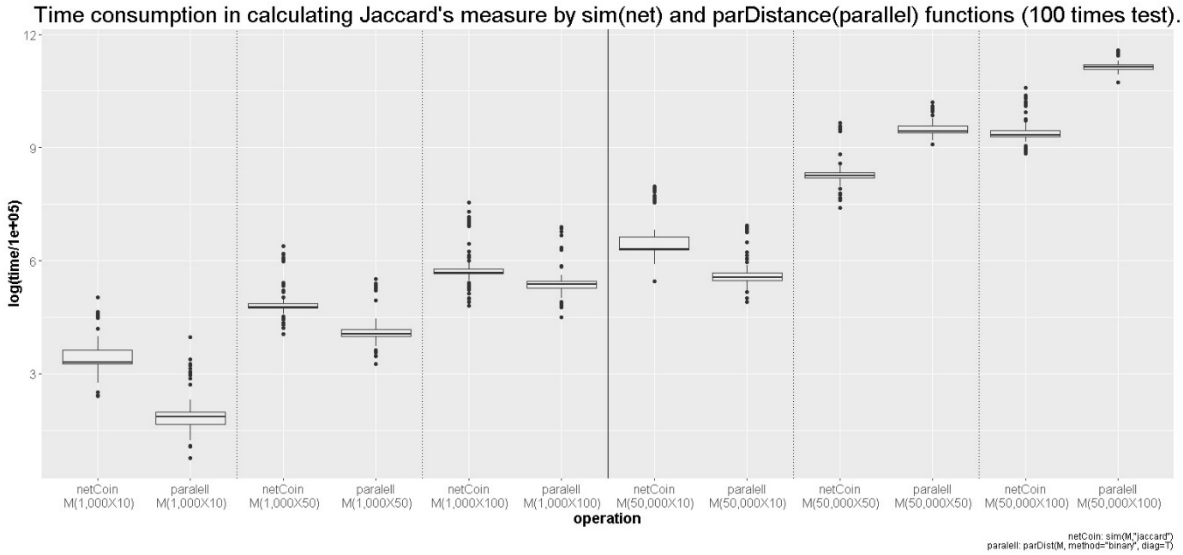


Figure 6: Performance of **netCoin** compared with **parallelDist**.

By means of routines from **igraph**, **netCoin** can reproduce different types of layouts and obtain communities with various algorithms, which facilitate the analysis and interpretation of coincidences. Data are then converted into D3 interactive graphs with controls enabling an interactive event analysis that can be shared with users online.

Acknowledgments

The work reported in this paper was supported by two grants to Modesto Escobar (CS-2013-49278-EXP and CSO2015-65094-P) from the State Program for R&D&i, whose funds comes from FEDER (European Union).

We are also grateful to the anonymous reviewers for comments on successive drafts of the paper, and to Carlos Prieto and David Barrios for their collaboration on **netCoin**.

References

Allaire JJ, Grandrud C, Rusell K, Yetman CJ (2015). ***networkD3: D3*** *JavaScript Network Graphs from R*. R package version 0.4, URL https://CRAN.R-project.org/package=networkD3.

Almende BV, Benoit T, Titouan R (2019). ***visNetwork****: Network Visualization Using* ***vis.js*** *Library*. R package version 2.0.9, URL https://CRAN.R-project.org/package=visNetwork.

Anderberg MR (1973). *Cluster Analysis for Applications*. Academic Press, New York.

Bastian M, Heymann S, Jacomy M (2009). "**Gephi**: An Open Source Software for Exploring and Manipulating Networks." In *International AAAI Conference on Weblogs and Social Media*.

Batagelj V, Mrvar A (1998). "**Pajek**-Program for Large Network Analysis." *Connections*, **21**(2), 47–58.

Bates D, Mächler M (2019). ***Matrix**: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-18, URL https://CRAN.R-project.org/package=Matrix.

Baumgartner M (2009). "Inferring Causal Complexity." *Sociological Methods & Research*, **38**(1), 71–101. doi:10.1177/0049124109339369.

Baumgartner M, Thiem A (2015). "Identifying Complex Causal Dependencies in Configurational Data with Coincidence Analysis." *The R Journal*, **7**(1), 176–184. doi:10.32614/rj-2015-014.

Blei DM, Ng A, Jordan M (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, **3**, 993–1022.

Blondel VD, Guillaume JL, Lefebvre E (2008). "Fast Unfolding of Communities In Large Networks." *Journal of Statistical Mechanisms: Theory and Experiment*, **2008**, P10008. doi:10.1088/1742-5468/2008/10/p10008.

Borgelt C (2012). "Frequent Item Set Mining." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2**(6), 437–456. doi:10.1002/widm.1074.

Bostock M, Ogievetsky V, Heer J (2011). "D^3 Data-Driven Documents." *IEEE Transactions on Visualization and Computer Graphics*, **17**(12), 2301–2309. doi:10.1109/TVCG.2011.185.

Butts CT (2008). "**network**: A Package for Managing Relational Data in R." *Journal of Statistical Software*, **24**(2), 1–36. doi:10.18637/jss.v024.i02.

Butts CT (2019). ***network**: Classes for Relational Data*. R package version 1.16.0, URL https://CRAN.R-project.org/package=network.

Carley K (1993). "Coding Choices for Textual Analysis; A Comparison of Content Analysis and Map Analysis." *Sociological Methodology*, **23**, 75–126. doi:10.2307/271007.

Chang J (2015). ***lda**: Collapsed Gibbs Sampling Methods for Topic Models*. R package version 1.4.2, URL http://CRAN.R-project.org/package=lda.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J (2020). ***shiny**: Web Application Framework for R*. R package version 1.4.0.2, URL https://CRAN.R-project.org/package=shiny.

Connor EF, Simberloff D (1983). "Interspecific Competition and Species Co-Occurrence Patterns on Islands: Null Models and the Evaluation of Evidence." *Oikos*, **41**(3), 455–465. doi:10.2307/3544105.

Corman SR, Kuhn T, McPhee R, Dooley K (2002). "Studying Complex Discursive Systems: Centering Resonance Analysis of Communication." *Human Communication*, **28**(2), 157–206. doi:10.1111/j.1468-2958.2002.tb00802.x.

Cox TF, Cox MA (2001). *Multidimensional Scaling*. Chapman & Hall/CRC, Boca Raton.

Csardi G (2020). ***igraph*** *– The Network Analysis Package*. R package version 1.2.5, URL https://CRAN.R-project.org/package=igraph.

Csardi G, Nepusz T (2006). "The **igraph** Software Package for Complex Network Research." *InterJournal Complex Systems*, **1695**, 1–9.

Diaconis M, Mosteller F (1989). "Methods for Studying Coincidences." *Journal of the American Statistical Association*, **84**(408), 853–861. doi:10.1080/01621459.1989.10478847.

Diamond JM, Gilpin ME (1982). "Examination of the "Null" Model of Connor and Simberloff for Species Co-Occurrences on Islands." *Oecologia*, **52**(1), 64–74. doi:10.1007/bf00349013.

Dusa A, Thiem A (2020). ***QCA****: Qualitative Comparative Analysis*. R package version 3.7, URL https://CRAN.R-project.org/package=QCA.

Eckert A (2018). ***parallelDist****: Parallel Distance Matrix Computation Using Multiple Threads*. R package version 0.2.4, URL https://CRAN.R-project.org/package=parallelDist.

Epskamp S, Costantini G, Haslbeck J, Isvoranu A (2020). ***qgraph****: Graph Plotting Methods, Psychometric Data Visualization and Graphical Model Estimation*. R package version 1.6.5, URL https://CRAN.R-project.org/package=qgraph.

Epskamp S, Cramer AO, Waldorp LJ, Schmittmann VD, Borsboom D (2012). "**qgraph**: Network Visualization of Relationships in Psychometric Data." *Journal of Statistical Software*, **48**(4), 1–18. doi:10.18637/jss.v048.i04.

Escobar M (2015). "Studying Coincidences with Network Analysis and Other Multivariate Tools." *The Stata Journal*, **15**(4), 1118–1156. doi:10.1177/1536867x1501500410.

Escobar M, Barrios D, Prieto C, Martinez-Uribe L (2020). ***netCoin****: Interactive Analytic Networks*. R package version 1.1.25, URL https://CRAN.R-project.org/package=netCoin.

Everitt BS (2003). *The Cambridge Dictionary of Statistics*. Cambridge University Press, Cambridge.

Feinerer I (2019). *Introduction to the* **tm** *Package Text Mining in R*. R package version 0.7-7, URL https://CRAN.R-project.org/web/packages/tm/vignettes/tm.pdf.

Feinerer I, Hornik K (2019). ***tm****: Text Mining Package*. R package version 0.7-7, URL https://CRAN.R-project.org/package=tm.

Feinerer I, Hornik K, Meyer D (2008). "Text Mining Infraestructure in R." *Journal of Statistical Software*, **25**(5), 1–52. doi:10.18637/jss.v025.i05.

Finney DJ (1948). "The Fisher-Yates Test of Significance in 2×2 Contingency Tables." *Biometrika*, **35**(1–2), 145–156. doi:10.1093/biomet/35.1-2.145.

Fisher RA (1935). "The Logic of Inductive Inference." *Journal of the Royal Statistical Society*, **98**(1), 39–82. doi:10.2307/2342435.

Frick A, Ludwing A, Mehldau H (1995). "A Fast Adaptative Layout Algorithm for Undirected Graphs." In R Tamassia, IG Tollis (eds.), *Lecture Notes in Computer Science*, pp. 388–403. Springer-Verlag, Berlin. doi:10.1007/3-540-58950-3_393.

Fruchterman TMJ, Reingold EM (1991). "Graph Drawing by Force-Directed Placement." *Software: Practice and Experience*, **21**(11), 1129–1164. doi:10.1002/spe.4380211102.

Girvan M, Newman MEJ (2002). "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences of the United States of America*, **99**(12), 7821–7828. doi:10.1073/pnas.122653799.

Good BH, de Montjove YA, Clauset A (2009). "The Performance of Modularity Maximization in Practical Contexts." *Physical Review E*, **81**, 046106. doi:10.1103/physreve.81.046106.

Gotelli NJ (2000). "Null Model Analysis of Species Co-Ocurrence Patterns." *Biology*, **81**(9), 2606–2621. doi:10.1890/0012-9658(2000)081[2606:nmaosc]2.0.co;2.

Gotelli NJ, Hart E, Ellison A (2015). ***EcoSimR**: Null Model Analysis for Ecological Data*. R package version 0.1.0, URL https://CRAN.R-project.org/package=EcoSimR.

Gower JC (1985). "Measures of Similarity, Dissimilarity, and Distance." In *Encyclopedia of Statistical Sciences*, volume 5. John Wiley & Sons, New York.

Grandrud C, Allaire JJ, Rusell K (2016). "**D3** JavaScript Network Graphs from R." URL http://christophergandrud.github.io/networkD3/.

Griffith DM, Veech JA, Marsh CJ (2016a). "**cooccur**: Probabilistic Species Co-Occurrence Analysis in R." *Journal of Statistical Software*, **69**(2), 1–17. doi:10.18637/jss.v069.c02.

Griffith DM, Veech JA, Marsh CJ (2016b). ***cooccur**: Probabilistic Species Co-Occurrence Analysis in R*. R package version 1.3, URL https://CRAN.R-project.org/package=cooccur.

Grimmer J, Stewart BM (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis*, **21**(3), 267–297. doi:10.1093/pan/mps028.

Grosjean P (2014). ***tcltk2**: Tcl/Tk Additions*. R package version 1.2-11, URL https://CRAN.R-project.org/package=tcltk2.

Haberman SJ (1973). "The Analysis of Residuals in Cross-Classified Tables." *Biometrics*, **29**(1), 205–220. doi:10.2307/2529686.

Hahsler M, Grün B, Hornik K (2005). "**arules** – A Computational Environment for Mining Association Rules and Frequent Item Sets." *Journal of Statistical Software*, **14**(15), 1–25. doi:10.18637/jss.v014.i15.

Hamann U (1961). "Merkmalsbestand und Verwandtschaftsbeziehungen der Farinosae. Ein Beitrag zum System der Monokotyledonen." *Willdenowia*, **2**, 639–768.

Hubálek Z (1982). "Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: An Evaluation." *Biological Reviews*, **57**(4), 669–689. doi:10.1111/j.1469-185x.1982.tb00376.x.

IBM Corporation (2017). *IBM SPSS Statistics 25*. IBM Corporation, Armonk. URL http://www.ibm.com/software/analytics/spss/.

Jaccard P (1901). "Distribution de la Flore Alpine dans le Bassin des Dranses et dans Quelques Règions Voisines." *Bulletin de la Societe Vaudoise des Sciences Naturelles*, **37**, 241–272. doi:10.5169/seals-266440.

Jurka T, Collingwood L, Boydstun AE, Grossman E, van Atteveldt W (2014). ***RTextTools**: Automatic Text Classification via Supervised Learning*. R package version 1.4.2, URL https://CRAN.R-project.org/src/contrib/Archive/RTextTools.

Kamada T, Kawai S (1989). "An Algorithm for Drawing General Undirected Graphs." *Information Processing Letters*, **31**(1), 7–15. doi:10.1016/0020-0190(89)90102-6.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO (2010). "**Picante**: R Tools for Integrating Phylogenies and Ecology." *Bioinformatics*, **26**(11), 1463–1464. doi:10.1093/bioinformatics/btq166.

Kulczynski S (1927). "Die Pflanzenassoziationen der Pieninen." *Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles, B*, **Suppl. II**, 57–203.

Loiseau S, Vaudor L, Decorde M, Heiden S (2015). ***textometry**: Textual Data Analysis Package Used by the **TXM** Software*. R package version 0.1.4, URL https://CRAN.R-project.org/package=textometry.

Lucas C, Nielsen R, Roberts M, Stewart B, Storer A, Tingley D (2015). "Computer Assisted Text Analysis for Comparative Politics." *Political Analysis*, **23**(2), 254–277. doi:10.1093/pan/mpu019.

Lund K, Burgess C (1996). "Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence." *Behavior Research Methods, Instruments, & Computers*, **28**(2), 203–208. doi:10.3758/bf03204766.

Martin S, Brown WM, Klavans R, Boyack KW (2008). "DRL: Distributed Recursive (Graph) Layout." *Technical report*, Sandia National Laboratories.

Matsuo Y, Ishizuka M (2004). "Keyword Extraction from a Document Using Word Co-Occurrence Statistical Information." *International Journal of Artificial Intelligence Tools*, **13**(1), 157–169. doi:10.1142/s0218213004001466.

Meyer D, Buchta C (2019). ***proxy**: Distance and Similarity Measures*. R package version 0.4.23, URL https://CRAN.R-project.org/package=proxy.

Newman MEJ (2006). "Finding Community Structure in Networks Using the Eigenvectors of Matrices." *Physical Review E*, **74**, 1–22. doi:10.1103/physreve.74.036104.

Ochiai A (1957). "Zoogeographic Studies on the Soleoid Fishes Found in Japan and Its Neighbouring Regions." *Bulletin of the Japanese Society of Scientific Fisheries*, **22**(9), 526–530. doi:10.2331/suisan.22.526.

Oksanen J, Blenchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2019). ***vegan***: *Community Ecology Package*. R package version 2.5-6, URL https://CRAN.R-project.org/package=vegan.

Padgett JF, Ansell CK (1993). "Robust Action and the Rise of the Medici, 1400–1434." *American Journal of Sociology*, **98**(6), 1259–1319. doi:10.1086/230190.

Pearson K (1900). "Mathematical Contributions to the Theory of Evolution. – VII. On the Correlation of Characters Not Quantitatively Measureable." *Philosophical Transactions of the Royal Society of London A*, **195**, 1–47. doi:10.1098/rsta.1900.0022.

Pons P, Latapy M (2006). "Computing Communities in Large Networks Using Random Walks." *Journal of Graph Algorithms and Applications*, 10(2), 191–218.

Popping R (2000). *Computer-Assisted Text Analysis*. Sage Publications, London.

Popping R (2003). "Knowledge Graphs and Network Text Analysis." *Social Science Information*, **42**(1), 91–106. doi:10.1177/0539018403042001798.

Prieto C, Barrios D (2017). ***RJSplot***: *Interactive Graphs with R*. R package version 2.5, URL https://CRAN.R-project.org/package=RJSplot.

Raghavan UN, Albert R, Kumara S (2007). "Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks." *Physical Review E*, **76**(3), 036106. doi:10.1103/physreve.76.036106.

Ragin C (1987). "The Comparative Method: Moving beyond Qualitative and Quantitative Methods." University of California, Berkeley.

Ragin CC (2000). *Fuzzy-Set Social Science*. University of Chicago Press, Chicago.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reichardt J, Bornholdt S (2006). "Statistical Mechanics of Community Detection." *Physical Review E*, **74**, 016110. doi:10.1103/physreve.74.016110.

Reingold EM, Tilford JS (1981). "Tidier Drawings of Trees." *IEEE Transactions on Software Engineering*, **7**(2), 223–228. doi:10.1109/tse.1981.234519.

Roberts ME, Stewart BM, Tingley D (2019a). "**stm**: An R Package for Structural Topic Models." *Journal of Statistical Software*, **91**(2), 1–40. doi:10.18637/jss.v091.i02.

Roberts ME, Stewart BM, Tingley D (2019b). ***stm***: *Estimation of the Structural Topic Model*. R package version 1.3.5, URL https://CRAN.R-project.org/package=stm.

Roberts ME, Tingley D, Lucas C, Leder-Luis J, Gadarian S, Alberrtson B, Rand D (2014). "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science*, **58**(4), 1064–1082. doi:10.1111/ajps.12103.

Robinson D, Silge J (2020). ***tidytext**: Text Mining Using **dplyr**, **ggplot2**, and Other Tidy Tools*. R package version 0.2.3, URL https://CRAN.R-project.org/package=tidytext.

Rogers DJ, Tanimoto TT (1960). "A Computer Program for Classifying Plants." *Science*, **132**(3434), 1115–1118. doi:10.1126/science.132.3434.1115.

Russell PF, Rao TR (1940). "On Habitat and Association of Species of Anopheline Larvae in South-Eastern Madras." *Journal of the Malaria Institute of India*, **3**, 153–178.

Sanderson J (2000). "Testing Ecological Patterns A Well-Known Algorithm from Computer Science Aids the Evaluation of Species Distributions." *American Scientist*, **88**(4), 332–339.

Schult DA, Swart P (2008). "Exploring Network Structure, Dynamics, and Function Using **NetworkX**." In *Proceedings of the 7th Python in Science Conference*.

Sneath PHA, Sokal RR (1962). "Numerical Taxonomy." *Nature*, **193**, 855–860. doi:10.1038/193855a0.

StataCorp (2019). *Stata Statistical Software: Release 16*. StataCorp LLC, College Station. URL http://www.stata.com/.

Strona G, Nappo D, Boccacci F, Fattorini S, San-Miguel-Ayanz J (2014). "A Fast and Unbiased Procedure to Randomize Ecological Binary Matrices with Fixed Row and Column Totals." *Nature Communications*, **5**(4114). doi:10.1038/ncomms5114.

Sugiyama K, Tagawa S, Mitsuhiko T (1981). "Methods for Visual Understanding of Hierarchical Systems Structure." *IEEE Transactions on Systems Man and Cybernetics*, **11**(2), 109–125. doi:10.1109/tsmc.1981.4308636.

Van Attenveld W (2008). *Semantic Network Analysis. Techiques for Extracting, Representing, and Quering Media Content*. Routledge, London.

Van Rossum G, *et al.* (2011). *Python Programming Language*. URL https://www.python.org/.

Veech JA (2013). "A Probabilistic Model for Analysing Species Co-Occurrence." *Global Ecology and Biogeography*, **22**(2), 252–260. doi:10.1111/j.1466-8238.2012.00789.x.

Wakita K, Tsurumi T (2007). "Finding Community Structure in Mega-Scale Social Networks." arXiv:cs/0702048 [cs.CY], URL https://arxiv.org/abs/cs/0702048.

Wasserman S, Faust K (1994). *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press, Cambridge.

Wickham H, Miller E (2019). ***haven**: Import and Export SPSS, Stata and SAS Files*. R package version 2.2.0, URL https://CRAN.R-project.org/package=haven.

Young L, Soroka S (2012). "Affective News: The Automated Coding of Sentiment In Political Texts." *Political Communication*, **29**(2), 205–231. doi:10.1080/10584609.2012.671234.

Yule GU (1900). "On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society." *Philosophical Transactions of the Royal Society of London A*, **194**, 257–319. doi:10.1098/rsta.1900.0019.

Zhang J (2016). ***spaa***: *SPecies Association Analysis*. R package version 0.2.2, URL https://CRAN.R-project.org/package=spaa.

Affiliation:

Modesto Escobar
Department of Sociology and Communication
Faculty of Social Sciences
University of Salamanca
37071 Salamanca, Spain
E-mail: modesto@usal.es
URL: http://sociocav.usal.es/web/en/miembros/sociologia/escobar-mercado/

Luis Martinez-Uribe
Doctoral Program in Social Sciences
University of Salamanca

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
May 2020, Volume 93, Issue 11 *Submitted:* 2017-06-15
doi:10.18637/jss.v093.i11 *Accepted:* 2019-12-04

MATHEMATICAL POPULATION STUDIES
https://doi.org/10.1080/08898480.2017.1418116

Digital archives as Big data

Luis Martinez-Uribe

Departamento de Sociología y Comunicación, Universidad de Salamanca, Salamanca, Spain; DataLab, Biblioteca de la Fundación Juan March, Madrid, Spain

ABSTRACT
Digital archives contribute to Big data. Combining social network analysis, coincidence analysis, data reduction, and visual analytics leads to better characterize topics over time, publishers' main themes and best authors of all times, according to the British newspaper *The Guardian* and from the 3 million records of the British National Bibliography.

KEYWORDS
Big data; coincidence analysis; social network analysis; open data

1. Introduction

Latour's (2007: 2) quote: *"It is as if the inner workings of private worlds have been pried open because their inputs and outputs have become thoroughly traceable"* summarizes Big data.

Big data comprise not only government databases, social media, transactions such as credit cards and online clicks, and global positioning systems or accelerometers, but also digitized documents from libraries and archives (Martinez-Uribe and Fernandez, 2015).

In bibliometrics, clustering, time series, and network analysis have already been used to rate citations and identify production and co-authorship (Kejžar, Černe, and Batagelj, 2010; Battisti and Salini, 2012; Ferrara and Salini, 2012). Library resources are of high quality (Topçu et al., 2014), be it in literature (Moretti, 2005), history (Cohen, 2006), musicology (Tuppen et al., 2016), or sociology (Escobar, 2009; Escobar and Isla, 2015).

We show what kind of insight bring graphics and visualization (Healy and Moody, 2014; Cook et al., 2016) by combining coincidence analysis, data reduction, social network, and visual analytics. We treat the case of the bibliography of the British National Library.

2. Data and method

2.1. *The British National Bibliography*

National bibliographies are devised to include every publication in the country (Evans, 2005). The British Library is the legal depository of the

CONTACT Luis Martinez-Uribe lmartinez@march.es

United Kingdom and Ireland since 1662. The British National Bibliography contains mentions of all books published since 1950.

In 2001, the British Library opened access to the British National Bibliography (Deliot, 2014). It converted its documents from the library bibliographic format "MARC21" to the linked data format of the Resource Description Framework. This dataset has links to other open library datasets such as the Virtual International Authority File, Geonames, or the Library of Congress Subject Headings.

Each of the over-three million British National Bibliography records informs about the title, the author, the date and place of publication, and the subjects.

2.2. *Network coincidence analysis*

The purpose of network coincidence analysis (Fisher, 1924; Fisher, 1928; Diaconis and Mosteller, 1989; Escobar, 2015) is to detect which people, subjects, objects, attributes, or events appear simultaneously in different spaces, which are called scenarios.

M events X_j, $j = 1, \ldots, M$, are random variables recorded in each of the N scenarios. $X_j = 1$ if the j-th event occurs, $X_j = 0$ otherwise. Two events are said to be "coincident" if they occur in the same scenario.

In the "incidence" matrix $X = (x_{ij})$, the rows $i = 1, \ldots, N$ stand for scenarios and the columns $j = 1, \ldots, M$ for events. This matrix is binary, with elements x_{ij} equal to 0 or 1 indicating if the event X_j occurs or not in the i-th scenario:

$$X = x_{ij} \quad i = 1, ..., N; \; j = 1, ..., M. \tag{1}$$

The coincidence matrix $C = (c_{ij})_{i,j=1,\ldots,M}$ is the symmetric $M \times M$ matrix

$$C := X^{\mathrm{T}}X, \qquad \text{with} \quad c_{ij} := \sum_{k=1}^{N} x_{ki}x_{kj} = c_{ji}, \tag{2}$$

where X^{T} denotes the transposed matrix of X. Because only scenarios k in which both events X_i and X_j occur ($x_{ki} = x_{kj} = 1$) contribute to c_{ij}, the element c_{ij} represents the total number of joint ocurrences of the events X_i and X_j. The total number of scenarios in which X_j occurs is c_{jj}.

By definition, the two events X_i and X_j are independent of each other when the conditional probability $P(X_i|X_j) = P(X_i)$. Then the probability of recording both events X_i and X_j is $P(X_i \cap X_j) = P(X_i)P(X_j)$. Two events X_i and X_j coincide in probability if:

$$c_{ij} > \frac{c_{ii}c_{jj}}{N}, \tag{3}$$

making X_i and X_j dependent of each other, or $P(X_i \cap X_j) > P(X_i)P(X_j)$.

We normalize the data using the statistical residuals e_{ij} between the recorded and the expected values through the Pearson residual:

$$e_{ij} := \frac{c_{ij} - \frac{c_{ii}c_{jj}}{N}}{\left(\frac{c_{ii}c_{jj}}{N}\right)^{\frac{1}{2}}}. \tag{4}$$

When non null, these residuals represent independent and coincident events. Haberman (1973) further divides e_{ij} by the standard deviation of all residuals:

$$d_{ij} := \frac{e_{ij}}{\left((1 - \frac{c_{ii}}{N})(1 - \frac{c_{jj}}{N})\right)^{\frac{1}{2}}}, \quad i \neq j. \tag{5}$$

The adjusted residuals d_{ij} are normally distributed with mean zero and standard deviation one. This allows us to test $d_{ij} = 0$ against $d_{ij} \neq 0$. With the entire population, it is no longer necessary to calculate probabilities. Thus we build the $M \times M$ adjacency matrix $A = (a_{ij})_{i,j=1,\ldots,M}$ using the Haberman residuals d_{ij} from all scenarios, following the rule:

$$a_{ii} = 0; \quad a_{ij} = \begin{cases} 1 \text{ if } d_{ij} > 0 \\ \\ 0 \text{ if } d_{ij} \leq 0 \end{cases} \quad i \neq j. \tag{6}$$

With sample data, we compute the adjacency matrix with the probability that the adjusted residual d_{ij} is non-negative.

Here the scenarios are the books mentioned in the British National Bibliography. The events include subjects, authors, and publishers. In the incidence matrix X, the rows correspond to books and the columns to subjects, authors, or publishers. The coincidence matrix C comprises the frequencies of those events in the diagonal and the frequencies of coincidences of two events elsewhere in the matrix. The adjacency matrix A determines which events (authors, subjects, or publishers) coincide in the set of scenarios (books), with the Haberman residual d_{ij} indicating the strength of that coincidence.

2.3. *Visual analytics*

Visualization is interactive: its purpose it to help make out patterns (Keim et al., 2008). Network graphs represent coincidences between events. A network graph $G = (V, E)$ represents a system made up of nodes $V = \{v_1, v_2, \ldots, v_m\}$ connected by edges $E = \{e_1, e_2, \ldots, e_l\}$ (Wasserman and Faust, 1994). We represent events as nodes and their coincidences as edges. The strengths of edges between connected nodes is given by Haberman residuals. The size of each node represents the frequency c_{ii} of the event in the set of scenarios.

The spatial distribution of the nodes in a network depends on the method. Graph drawing algorithms are multidimensional scaling (Kruskal and Wish, 1978; Kamada and Kawai, 1989; Fruchterman and Reingold, 1991).

2.4. *Software*

The R statistical software contains network coincidence analysis and the associated D3.js javascript visualization library in the netCoin R package (Escobar et al., 2017) available in the Comprehensive R Archive Network at https://cran.r-project.org/package=netCoin.

NetCoin generates a Web page with network graphs of coincident events. Users load this Web page with a Web browser and interacts with the network through a control panel. They can customize the network for the location, color, shape, size of the nodes, and width and color of the edges. They can zoom and move the network.

3. Case study: The British National Bibliography

After getting rid of irrelevant data, we retain frequent enough events. We calculate the strengths of edges between events, generate the interactive network graph, and remove edges having too low connection strengths.

3.1. *Topics of the British National Bibliography over time*

Figure 1 shows that the total number of books catalogued per year in the British Library has increased since 1950.

We filter the scenarios by relevance. After discarding books with no date of publication or published before 1960, the dataset contains 2,816,615 books.

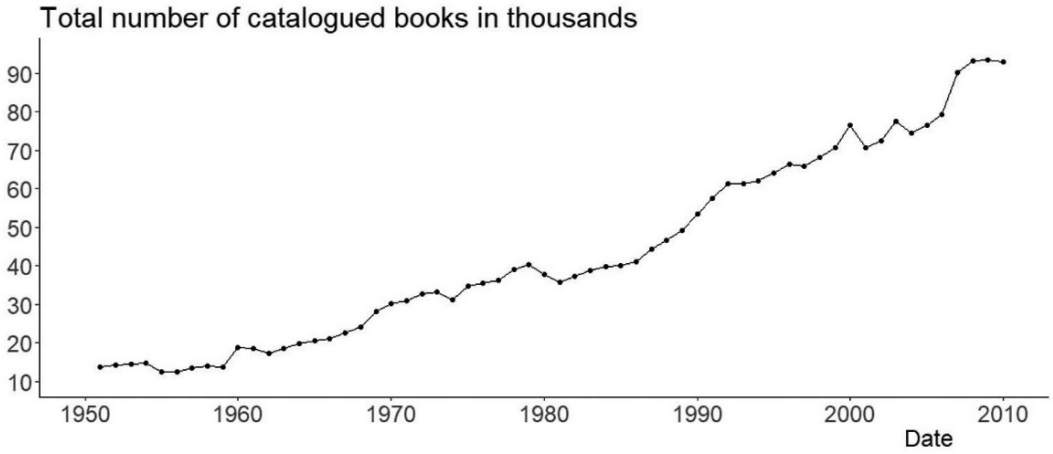


Figure 1. Total number of books catalogued in the British National Bibliography by year of publication.

Discarding the books with no subject leads to 2,279,781 books. We classify all books published by decade of publication. The total set of scenarios comprises 287,233 subjects. We consider the most frequent 160 ones. We then produce the network graph in Figure 2, where decades are indicated with a cross and the size of the node represents the frequency of the event. The graph contrasts the 60s, 70s, and 80s on the left-hand side of Figure 2 to the 90s, 2000s, and 2010s on the right-hand side.

Figure 3 presents the main subjects from 1960 to 1980 and Figure 4 from 1980 to 2010. The size of the nodes indicates the frequency of the events in the 2 million scenarios. These network graphs show that, although topics such as "fiction in English" keep a constant proportion across decades, others change. From the 60s to the 80s, the main subjects have ceased to be on Great Britain only, and, have gradually involved information technology, business,

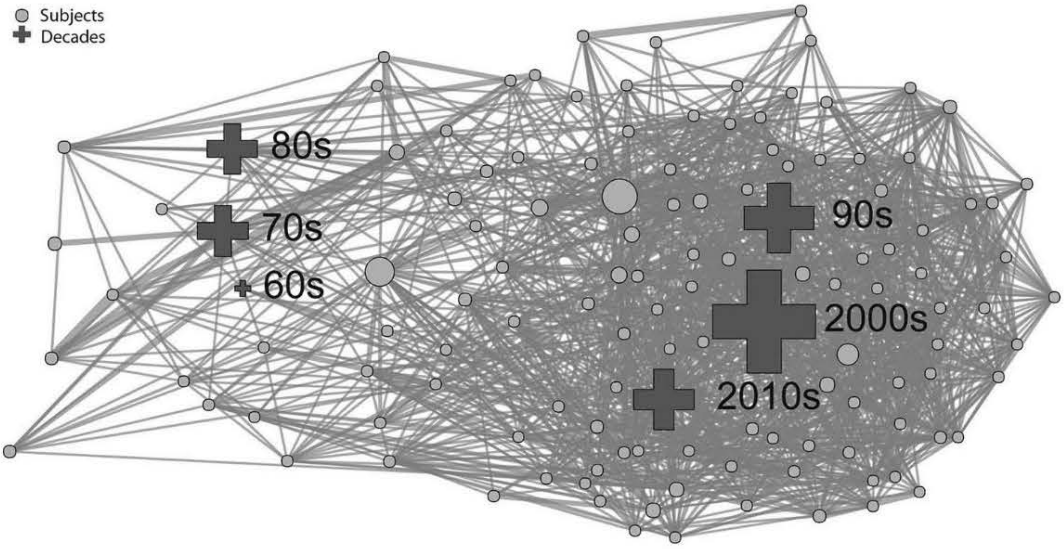


Figure 2. Network of the main subjects in the British National Bibliography by decade.

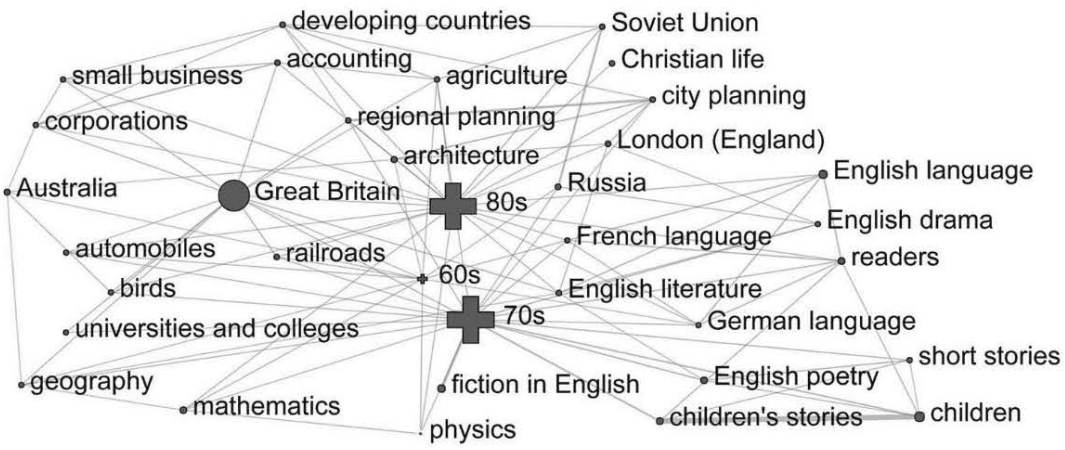


Figure 3. Network of the subjects in the British National Bibliography from 1960 to 1980.

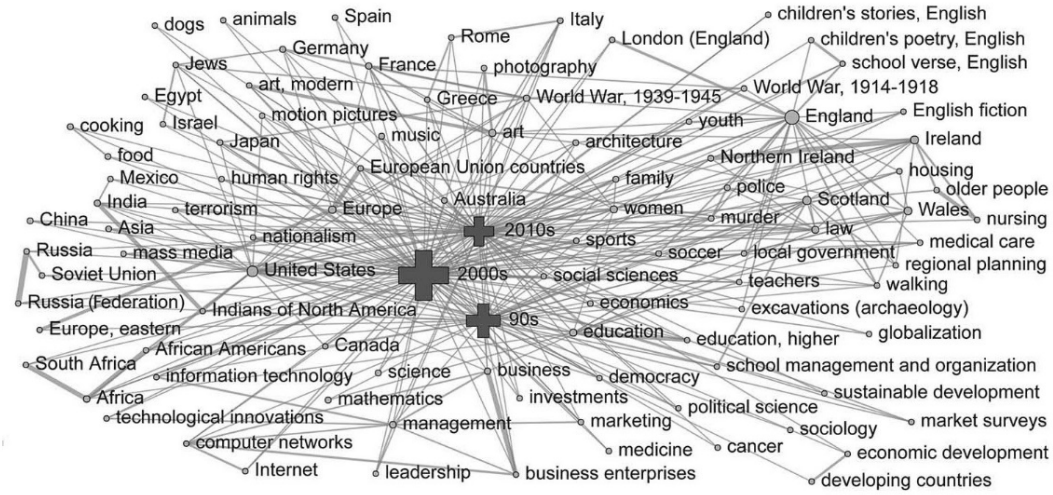


Figure 4. Network of the main subjects in the British National Bibliography from 1990 to 2016.

management, and social sciences, beside mathematics, physics, and geography.

Our data reduction has produced around 2 million scenarios and 300,000 events. The network graphs of Figures 2–4 highlight the most prevalent and the most important connections.

3.2. *Published themes*

The British National Bibliography comprises 214,131 publishers. Figure 5 shows the publishers with the highest total number of books. How similar are their published themes?

The data reduction starts with the selection of relevant scenarios, among the four most prolific publishers: Oxford University Press, Cambridge

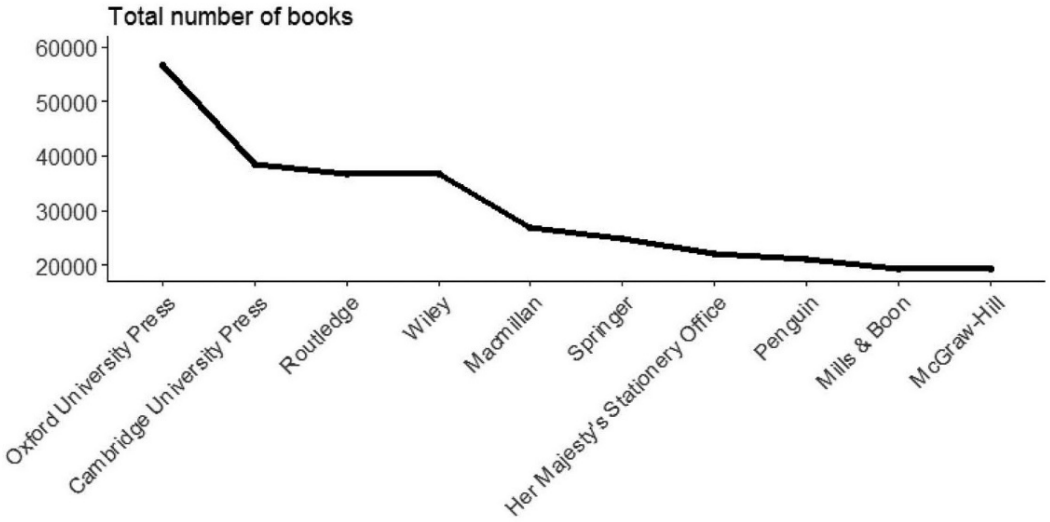


Figure 5. Total number of books by publisher in the British National Bibliography.

University Press, Routledge, and Wiley. This reduced set of scenarios comprises 168,294 books.

We retain the four publishers and their 20% most frequent topics. For Oxford University Press, we retain 40 topics out of 20,126; for Cambridge University Press, 49 out of 17,972; for Wiley 52 out of 11,668; and for Routledge 80 out of 14,694. Because of common topics, the sum amounts to 157 topics, to which we add the four publishers considered as events to obtain 161 events. Figure 6 represents these topics together with the publishers and their coincidences in a network graph using crosses for publisher nodes. Wiley is located away from the other publishers, which indicates that Wiley publishes relatively uncommon topics. The proximity of Oxford University Press, Routledge, and Cambridge University Press to one another indicates how much they share topics. Figures 7 and 8 highlight Wiley amidst its published topics, as an example.

3.3. *The 100 best novels' authors of all times in the British National Bibliography according to* The Guardian

In October 2013, the newspaper *The Guardian* published a list of what its editors considered to be the 100 best novels of all times (Mccrum, 2013). How do their authors situate themselves in the British National Bibliography?

The 100 authors are the "events". We reduce the data by selecting the books whose authors are mentioned in the 100 best novels by *The Guardian*. They amount to 13,216 books from which we retain 6,613 books having clear topics, the "scenarios". These books comprise 2,008 distinct subjects. We retain the 116

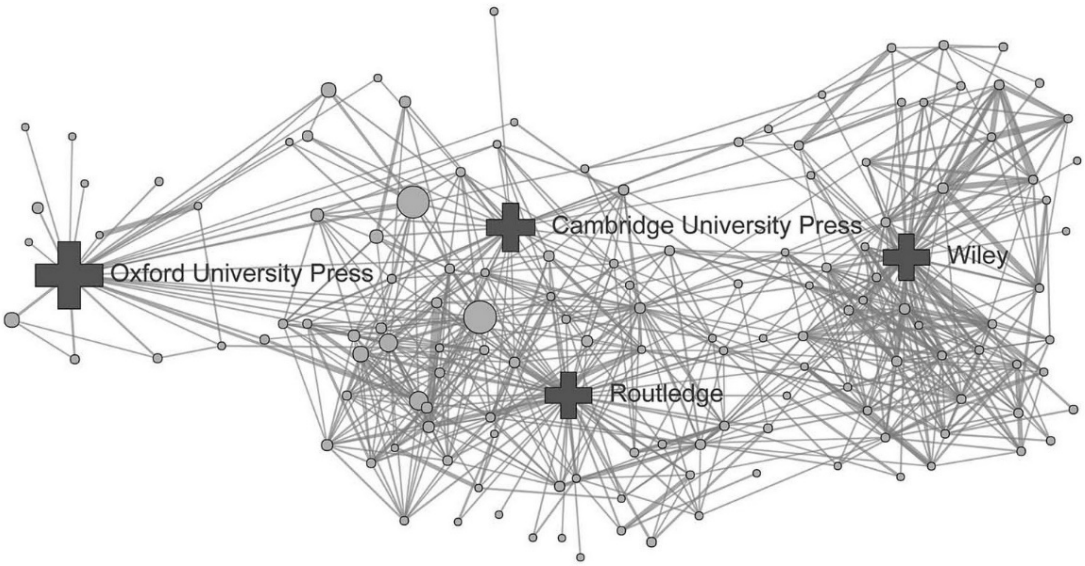


Figure 6. Publishers and their topics.

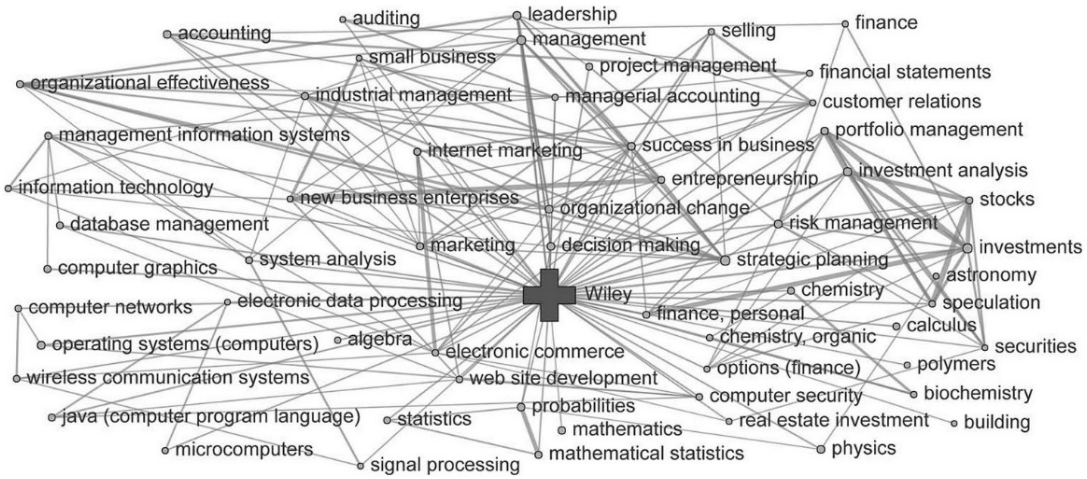


Figure 7. Wiley and its topics.

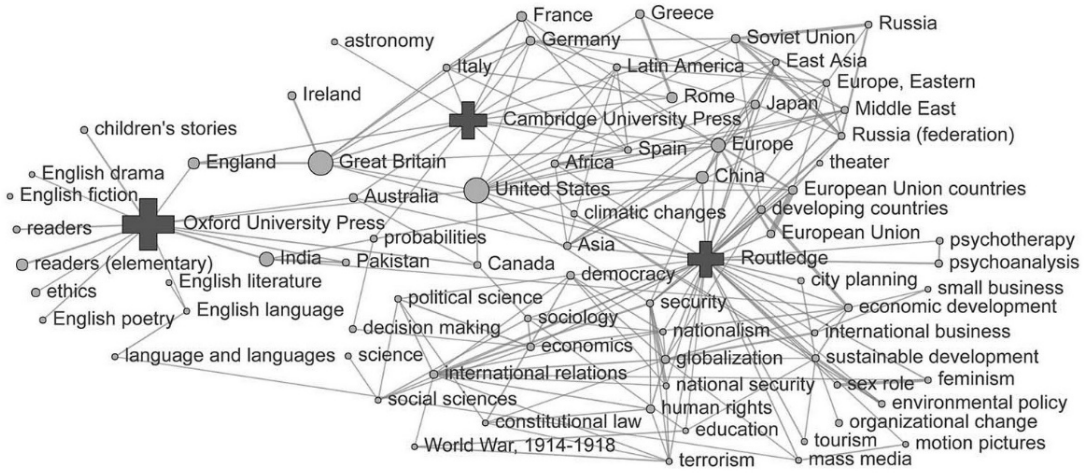


Figure 8. Oxford University Press, Cambridge University Press, Routledge, and their topics.

most frequent ones, amounting to more than half of all occurrences. For each author, we add images and geographical and topical information.

The resulting network in Figure 9 highlights a focus on English literature, with "England" as the most frequent event.

In Figure 10 authors around the node of "England" include mostly romantic and realistic British authors such as Charles Dickens, Jane Austen, and Anthony Trollope, and treat fiction, family and friendships, and social life in England.

4. Conclusion

We showed that collections prepared by libraries can contribute to Big data. Network coincidence analysis combines statistical methods and social network analysis to reduce the total number of events, measure relationships, and delineate trends. Our treatment of a bibliography has shown how to

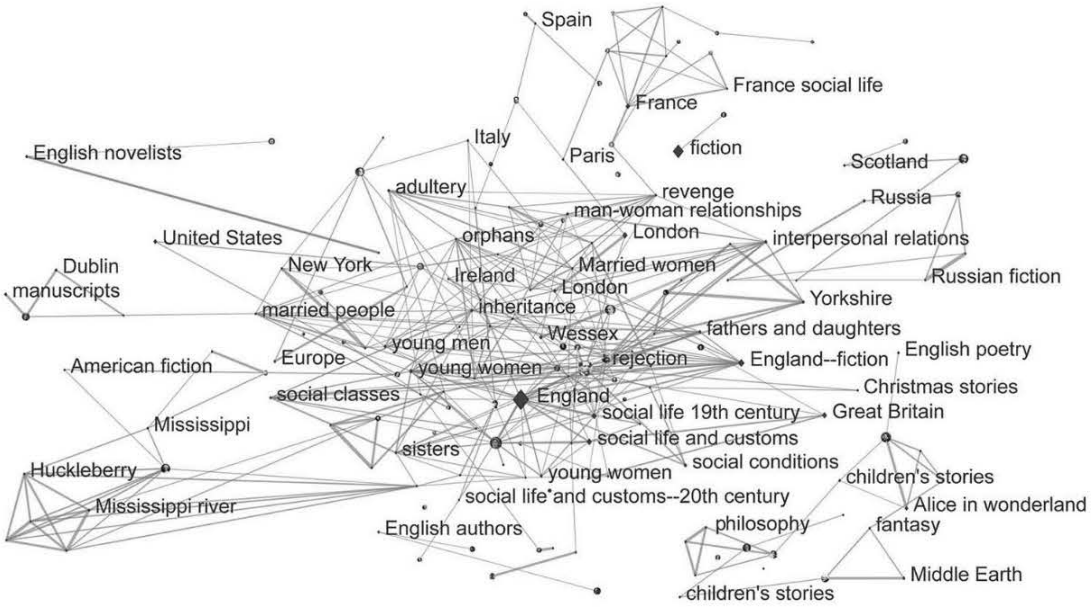


Figure 9. The 100 best novels' authors according to *The Guardian* in the British National Bibliography with their main topics.

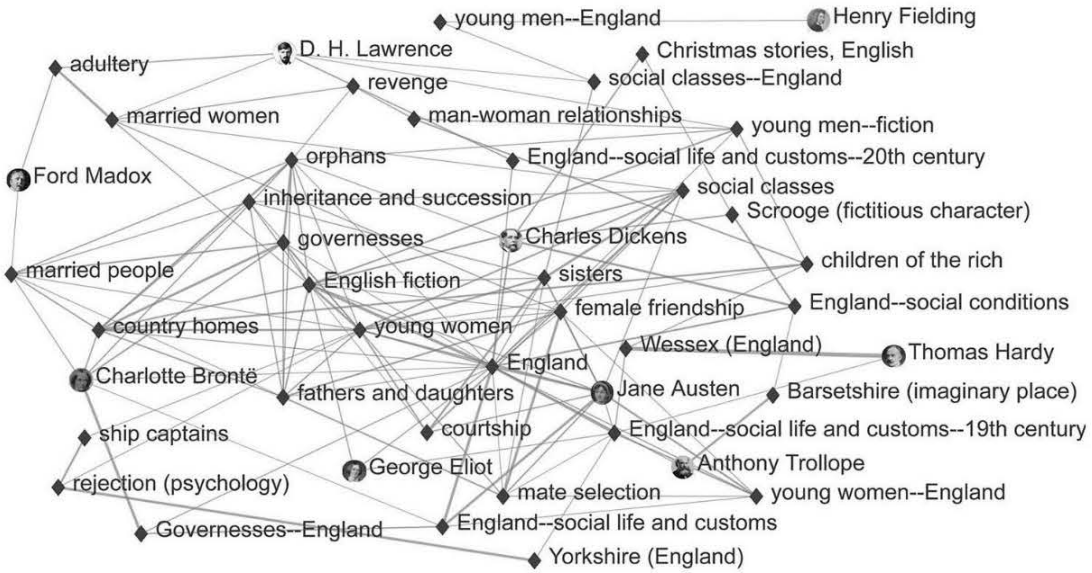


Figure 10. The 100 best novels' authors in the British National Bibliography surrounding the "England" node.

extract insights from more than 3 million records, 800,000 person names, and 300,000 subject headings.

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (FEDER) (research project number CSO2013-49278- 180 EXP).

ORCID

Luis Martinez-Uribe http://orcid.org/0000-0002-7795-3972

References

Battisti, F. (de) and Salini, S. (2012). Bibliographic data: a different analysis perspective. *Electronic Journal of Applied Statistical Analysis*, *5*(3): 353–359.

Cohen, D. J. (2006). From babel to knowledge: Data mining large digital collections. *Dlib Magazine*, *12*(3). [Online]. Retrieved from http://www.dlib.org/dlib/march06/cohen/03cohen.html

Cook, D., Eun-Kyung, L., and Mahbulul, M. (2016). Data visualization and statistical graphics in big data analysis. *Annual Review of Statistics and Its Application*, *3*(1): 133–159.

Deliot, C. (2014). Publishing the British national bibliography as linked open data. *Catalogue and Index*, *174*(1): 13–18.

Diaconis, P. and Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association*, *84*(408): 605–622.

Escobar, M. (2009). Redes semánticas en textos periodísticos: propuestas técnicas para su representación. *Empiria: revista de metodología de ciencias sociales*, *17*(1): 13–39.

Escobar, M. (2015). Studying coincidences with network analysis and other multivariate tools. *Stata Journal*, *15*(4): 1118–1156.

Escobar, M. and Isla, J. G.. (2015). La expresión de la identidad a través de la imagen: los archivos fotográficos de Miguel de Unamuno y Joaquín Turina. *Revista española de investigaciones sociológicas (REIS)*, *152*(1): 23–34.

Escobar, M., Prieto, C., Barrios, D., et al. (2017). netCoin: INTERACTIVE networks with R. Retrieved from https://cran.r-project.org/web/packages/netCoin/index.html

Evans, G. E. (2005). *Developing Library and Information Center Collections*. Toledo, OH, U.S. A.: Libraries Unlimited.

Ferrara, A. and Salini, S. (2012). Ten challenges in modeling bibliographic data for bibliometric analysis. *Scientometrics*, *93*(3): 765–785.

Fisher, R. A. (1924). A method of scoring coincidences in tests with playing cards. *Proceedings of the Society for Psychical Research*, *34*. Glasgow: University Press Glasgow. 181–185.

Fisher, R. A. (1928). The effect of psychological card preferences. *Proceedings of the Society for Psychical Research*, *38*. Glasgow, University Press Glasgow. 269–271.

Fruchterman, T. M. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11): 1129–1164.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, *29*(1): 205–220.

Healy, K. and Moody, J. (2014). Data visualization in sociology. *Annual Review of Sociology*, *40*(1): 105–128.

Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, *31*(1): 7–15.

Keim, D., Andrienko, G., Fekete, J. D., et al. (2008). Visual analytics: definition, process, and challenges. In A. Kerren, J. T. Stasko, J. D. Fekete, C. North (Eds.). *Information visualization. Lecture notes in computer science, 4950*. Berlin: Springer.

Kejžar, N., Černe, S. K., and Batagelj, V. (2010). Network analysis of works on clustering and classification from web of science. In H. Locarek-Junge, and C. Weihs (Eds.), *Classification as a Tool for Research. Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin: Springer.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. London: Sage.

4. Artículo III. La sociología a través de sus publicaciones en revistas de impacto mediante el uso de big data

La sociología a través de sus publicaciones en revistas de impacto mediante el uso de big data

Sociology through its publications in impact factor journals using big data

Luis Martínez-Uribe
Fundación Juan March
Universidad de Salamanca
lmartinez@march.es (ESPAÑA)
ORCID https://orcid.org/0000-0002-7795-3972

Recibido: 14.09. 2020
Aceptado: 23.11.2021

RESUMEN

Al igual que las demás disciplinas científicas, la sociología se puede concebir como un sistema social compuesto de complejas relaciones entre actores que incluyen a investigadores, instituciones, revistas y editoriales. Actualmente, el fenómeno del big data ofrece la posibilidad de usar grandes colecciones de datos que permiten la representación de los vínculos sociales que se dan en la ciencia. En concreto, a través de las grandes fuentes de datos bibliométricas, big scholarly data, la sociología tiene a su alcance ingentes cantidades de datos para describir y estudiar con detalle la evolución de las disciplinas científicas.

En este artículo describimos la sociología de los últimos años a través de las publicaciones en las revistas de impacto. Para hacer esto, se emplean datos de revistas de sociología del Journal Citation Reports ampliados con la información de los artículos del Microsoft Academic Graph. Realizamos un análisis descriptivo de las revistas, sus países de origen, lenguas, editoriales y décadas de aparición e impacto. A continuación, evaluamos la evolución temporal del número de artículos y citas, así como la coautoría y el género de los autores. Tras esto, establecemos cuatro grupos de tipos de revistas y estudiamos sus diferencias en las dimensiones anteriores mediante contrastes de hipótesis. Finalmente, representamos las relaciones entre autores y revistas usando una red de afiliación que nos permite detectar grupos de revistas que forman interesantes comunidades temáticas y geográficas.

La novedad del trabajo consiste en haber utilizado una fuente de datos de las denominadas big scholarly data con más de 300 millones de publicaciones y

EMPIRIA. Revista de Metodología de Ciencias Sociales. N.º 53 enero-abril, 2022, pp. 53-89.
ISSN: 1139-5737, DOI/ empiria.53.2022.32612

forma de red. Los resultados corroboran análisis de estudios previos y presentan la sociología como una disciplina dominada por revistas generalistas anglosajonas que abarca una amplia variedad de temáticas y con enfoques metodológicos diversos que dependen del ámbito geográfico en el que se desarrollan. Unas pocas revistas dominan las citas, mientras que los artículos de revistas metodológicas tienen un grado mayor de coautoría y las revistas temáticas un menor sesgo de género.

PALABRAS CLAVE

Sociología del conocimiento, sociología, grandes datos, datos masivos, redes de afiliación, bibliometría, métodos de investigación, Microsoft Academic Graph, análisis de coincidencias.

ABSTRACT

Like other scientific disciplines, sociology can be observed as a social system made up of researchers, institutions, journals and publishers. These relationships are established via conceptual communications which form networks that establish the way in which disciplined are organized. At present, the big data phenomena offers the capacity to use large data collections to analyse social processes. Big scholarly data sources offer sociology immense quantities of data useful to describe and study the evolution of scientific disciplines in detail.

In this article we characterised the last thirty years of sociology through its publications in impact factor journals. To do this, we use data about the sociology journals from Journal Citation reports augmented with article information from Microsoft Academic Graph. The analysis starts by describing the journals, countries of origin, languages, publishers, the decades in which they appeared and their impact factor. After this, we evaluate the evolution of numbers of articles and citations as well as co-authorship and gender proportion. Subsequently, we establish four groups of journal types and study their differences in the previous dimensions using hypothesis tests. Finally, we represent the relationships between authors and journals using an affiliation network that allows us to detect groups of journals that form interesting thematic and geographic communities.

The novelty of the work consists in having used a data source of the so-called big scholarly data with more than 300 million publications. The paper also provides several strategies to select the data of interest among the millions of publications to reduce their dimensionality in order to represent them in the form of a network. The results show a discipline dominated by Anglo-Saxon countries and large publishing conglomerates. The most prominent journals dominate citations whilst methodological journals have a higher degree of co-authorship and thematic journals have the lowest gender bias. The affiliation network between authors and journals contains two large groups, one formed by the pioneering

American journals together with quantitative methodological journals and another one made up of English and qualitative methodological journals.

KEYWORDS

Sociology of knowledge, sociology, big data, affiliation networks, bibliometrics, research methods, Microsoft Academic Graph, coincidence analysis.

1. INTRODUCCIÓN

La sociología como disciplina científica surge en distintas comunidades nacionales en distintos momentos en el tiempo y ha evolucionado en una variedad amplia de temáticas y métodos para resolver el extenso espectro de problemas que se plantea. La disciplina ha sido acusada de estar dominada por los países anglosajones, fragmentada en temáticas volátiles, con falta de cohesión metodológica y sesgo de género. Estudios bibliométricos en los últimos años se han encargado de mapear y analizar la evolución de las disciplinas a partir de la información de sus publicaciones. En sociología estos estudios se han centrado en medir la producción y el impacto de ciertas comunidades y subdisciplinas, pero no de la disciplina en su conjunto. La reciente aparición del fenómeno *big data* ha generado nuevas fuentes de datos que aportan nuevas oportunidades y retos metodológicos. En el terreno de bibliometría este es el caso que nos plantean los datos del Microsoft Academic Graph, una fuente de datos reciente, denominada *big scholarly data*, que cuenta con 230 millones de publicaciones con información enriquecida mediante técnicas de inteligencia artificial. ¿Hasta qué punto estos datos corroboran algunos de estos aspectos de la sociología? ¿Cuáles son los retos metodológicos que nos plantea su utilización?

El objetivo de este artículo consiste en caracterizar la disciplina sociológica a través de sus publicaciones en revistas de impacto utilizando *big scholarly data*, nuevos métodos y estrategias que nos permiten reducir la dimensionalidad para explorar grandes cantidades de datos.

En este artículo se concibe la sociología como una disciplina en la que se entretejen complejos sistemas de relaciones sociales entre investigadores, instituciones, asociaciones y editoriales. Partiendo de la premisa de que esas relaciones y su evolución en el tiempo se conforman en forma de redes, se da cuenta de la sociología de los últimos años a través de sus revistas de impacto. Para ello, se emplean datos de revistas de sociología del *Journal Citation Reports* ampliados con la información de los artículos del *Microsoft Academic Graph*.

El artículo se estructura de la siguiente forma: comienza con una breve introducción de la sociología como disciplina científica resaltando el corpus teórico que la conceptualiza como un sistema social dominado por sus comunicaciones. Se presenta a continuación el fenómeno *big bata* con sus expectativas y retos para después enfocarse en los grandes datos bibliométricos con su capacidad

para analizar las disciplinas científicas. Las secciones de datos y metodologías describen las fuentes de datos que se emplean y los métodos estadísticos que se aplican. Tras esto, se exploran los datos obtenidos a través de análisis descriptivos, test de hipótesis y análisis de redes sociales.

2. LA SOCIOLOGÍA Y SU CONCEPTUALIZACIÓN COMO DISCIPLINA

La definición de la sociología a través de la identificación de sus dominios y métodos ha sido un tema recurrente desde sus inicios. Emile Durkheim (1982) la definía como la ciencia de los hechos sociales y de esta forma la diferenciaba del resto de disciplinas científicas. Por otro lado, Albion W. Small (1906), fundador del primer departamento de sociología en Estados Unidos, leía un artículo ante el *Sociology Club* de la Universidad de Chicago donde consideraba una pérdida de tiempo intentar definir la disciplina.

La sociología como disciplina ha sido acusada de fragmentarse con nuevas áreas de interés que surgen constantemente y que amenazan su estatus y unidad (Moody y Light, 2006, O'Reilly, 2009). Esta misma preocupación la recoge Smelser al destacar la falta de cohesión en el nivel conceptual:

> "...sociology, by comparison with some other sciences, lacks a single, accepted conceptual framework. The field is difficult to distinguish from other because it contains a diversity of frameworks, some of which it shares with other fields such as psychology and social anthropology. If anything, then, sociology is too comprehensive, diffuse, soft in the center, and fuzzy around the edges." (Smelser, 2014)

Otras perspectivas la describen como una disciplina que se origina de modo diverso en distintas comunidades nacionales, que con el tiempo pasan a formar redes supranacionales que terminan siendo globales (Vanderstraeten, 2010). Distintos estudios revelan el dominio de prominentes revistas generalistas anglosajonas (Moody y Light, 2006), una división marcada entre la tradición americana más cuantitativa y la británica más teórica (Zougiris, 2018), una coautoría más habitual en los trabajos cuantitativos (Moody, 2004) y un cierto sesgo de género entre los autores que publican (Grant y Ward, 1991).

3. MARCO TEÓRICO: LAS DISCIPLINAS COMO SISTEMAS SOCIALES CONSTITUIDOS POR REDES DE PUBLICACIONES

Este artículo se encuadra en un marco teórico que concibe la producción científica desde el estudio de los procesos de interacción social, de su entorno y de su evolución en el tiempo. La imaginación social (Mills, 1959) cobra entonces un papel esencial para apreciar el escenario en el que estos procesos actúan.

Estudios en la sociología del conocimiento sugieren que el conjunto de ideas que uno considera verdaderas depende en gran medida del grupo al que pertenezcas. Así científicos que pertenecen a redes de colaboración comparten ideas, utilizan metodologías similares y se influencian unos a otros (Moody, 2004).

Las disciplinas científicas pueden observarse como sistemas sociales cuya comunicación es fundamental para entender los mecanismos que las conforman. Estos dominios científicos son dependientes de comunicaciones conceptuales observadas por terceros (Stichweh, 2008). Los artículos académicos son un buen ejemplo de comunicaciones, suponen prestigio, establecen reglas de recompensa y reclaman autoridad (Stinchcombe, 1984). De manera discontinua, estas disciplinas se forman y evolucionan en el tiempo a través de un compendio compartido de teorías, métodos y problemas que han de ser solucionados (Khun, 1962:11). La ciencia es pues un ámbito social donde están presentes el interés, las relaciones de poder, cuando no la ostentación de prestigio (Bourdieu, 2004: 29).

La actual comunicación científica se engloba en la sociedad red definida por Castells como "*la nueva estructura social de la Era de la Información, basada en redes de producción, poder y experiencia*" (Castells, 1998:350). Estas redes se convierten en interesantes laboratorios para entender e interpretar el proceso de producción científica (Latour y Woolgar, 1987). De este modo, las disciplinas pueden estudiarse como redes de publicaciones en las que se van construyendo temas y procedimientos aceptados por la comunidad científica (Luhmann, 1995).

Moody (2004) utiliza la idea de estructura de red y propone tres formas posibles de redes de colaboración en sociología. Una primera estructura es la que se ve afectada por no disponer de una teoría unificada de la disciplina, esto hace que la red esté compuesta por múltiples especialidades desconectadas y con subredes altamente agrupadas. La segunda estructura está determinada por la idea de que la producción científica depende de unos pocos científicos, *scientific stars,* cuyo trabajo determina el curso de la disciplina y genera redes en forma de estrella. Finalmente, las colaboraciones con límites teóricos más permeables generan redes de gran alcance y cohesionadas estructuralmente.

4. *BIG SCHOLARLY DATA*, UNA FUENTE PARA EL ANÁLISIS DE DISCIPLINAS

El primer uso documentado del término *big data* aparece en un artículo de científicos de la NASA en 1997 describiendo un problema de visualización de datos debido a conjuntos de datos tan grandes que ponen a prueba la capacidad de la memoria principal, el disco local e incluso el disco remoto. A esto lo llamaron el problema del *big data* (Press, 2014). El *big data* se puede concebir como una combinación de nuevas fuentes de datos de gran tamaño con las infraestructuras tecnológicas y los métodos para su tratamiento. El fenómeno del *big data* trae consigo beneficios para la investigación social como la capacidad de usar datos sobre fenómenos sociales no disponibles anteriormente, poseer

información sobre poblaciones completas, recabar datos de modo inconsciente para la persona observada (*unobtrusive data*) y analizar información de procesos sociales en tiempo real (Boyd y Crawford, 2012, Espeland y Stevens, 2008, Manovich, 2015, Martinho, 2018, McFarland et al., 2016, Moretti, 2000, Tinati et al., 2014). Como contrapeso al entusiasmo suscitado por estos nuevos recursos, no faltan los análisis críticos que reflexionan sobre sus limitaciones y reabren el debate ya acontecido en otros dominios científicos sobre la investigación centrada en datos versus la investigación basada en hipótesis (Carroll, 2009, Neresini, 2017). Ya en los años 80 los estudios interpretativos sobre la construcción social de la tecnología consideran los datos y la tecnología como el resultado de un proceso de construcción social y advertían del peligro de las perspectivas positivistas donde los datos son considerados objetivos (Pinch y Bijker, 1984). La literatura más reciente nos alerta de los riesgos de los métodos y teoría que acompañan al Big Data ya que en ocasiones los datos solo capturan información de ciertas actividades, pueden ser incorrectos, parciales o no contar con información sobre su procedencia (Giardullo, 2016, Halford y Savage, 2017, McFarland y McFarland, 2015).

En las ciencias sociales nuevas subdisciplinas como *social data science o computational social science* están asumiendo este reto, abordándolo de manera interdisciplinar junto con matemáticos e ingenieros (Blok et al., 2017, Burrows y Savage, 2014, Lazer et al., 2009, Savage y Burrows, 2007). En este contexto, los sociólogos han de contribuir con su conocimiento teórico a interpretar cómo se estructura lo social (Tubaro, 2014). Los datos por si mismos no son suficientes (Grimmer, 2015) y el trabajo con grandes datos debe seguir siendo una operación teórica, pues la interpretación es crucial para analizar la realidad social (Boyd y Crawford, 2012).

La sociología ha utilizado métodos cuantitativos desde sus inicios aportando rigor a la disciplina. Asimismo, en muchos casos la sociología ha contribuido de manera importante al desarrollo de métodos estadísticos (Clogg, 1992). La revisión de Raftery (2000) del uso de estadísticas en sociología presenta tres generaciones de métodos estadísticos en sociología: una primera generación, a partir de la Segunda Guerra Mundial, centrada en las tabulaciones cruzadas de encuestas y censos de pocas variables; una segunda generación desde 1960 que trata con datos de encuesta con muchas variables, y una tercera generación que comienza en los años 80 donde los formatos varían e incluyen datos textuales, redes sociales o datos espaciales. Actualmente, también se utilizan técnicas como la visualización de datos (Healy y Moody, 2014), metodologías de procesamiento del lenguaje natural (PLN) (Evans y Aceves, 2016), minería de datos o modelos de aprendizaje automático (*machine learning ML)* (Amaturo y Punziano, 2017, Frank et al., 2019).

Por otro lado, la bibliometría ha acuñado el término *big scholarly data* (Xia et al., 2017), para referirse a aquellas fuentes académicas de datos que han sufrido un crecimiento exponencial en los últimos años. Bases de datos, algunas de pago como *Web of Knowledge* y *Scopus* o gratuitas como *Google Scholar* y *Microsoft Academic Graph*, acumulan una inmensa cantidad de información

académica sobre autores, citas, artículos de revistas, actas de congresos, tesis y libros. Estos datos proveen indicadores que se analizan generalmente en investigaciones bibliométricas, para medir el impacto y la productividad científica o evaluar las redes académicas de comunicación e investigación.

Además, en los últimos años, según describen Su & Lee (2010), existe una multitud de estudios que mapean y estudian la evolución de dominios científicos a partir de los datos y metadatos de sus publicaciones científicas. Existe una amplia tradición de estudio de disciplinas científicas en la sociología de la ciencia. Estos estudios entienden la comunicación científica a través de revistas especializadas como una forma de organización y control de las disciplinas constituidas como construcciones sociales (Vanderstraeten, 2010). Sin embargo, Gupta & Battacharya (2004) argumentan que un nuevo enfoque de estudio surge cuando comienzan los estudios cuantitativos con información de publicaciones, ya que estas no solo revelan la estrategia científica de sus autores, sino que también proporcionan información sobre las dinámicas compartidas por la disciplina a la que pertenecen.

Los estudios bibliométricos de la sociología se han centrado hasta el momento en la producción sociológica y en los patrones de coautoría de países concretos como Australia, Países Bajos, Italia, Francia o España (De Haan, 1997, Gantman y Dabós, 2018, Phelan, 2000, Riviera, 2015, Vanderstraeten, 2010), en subdisciplinas como la sociología médica o la desigualdad de riqueza, (Korom, 2019, Seale, 2008), en las élites académicas en sociología (Korom, 2020), en la división de escuelas metodológicas (Oromaner, 1981, Schwemmer y Wieczorek, 2019), en la diferencia de citas entre los artículos teóricos y metodológicos (Peritz, 1983), así como en los efectos de los rankings, que favorecen ciertos tipos de culturas de investigación en los departamentos de sociología (Moksony et al., 2014).

Sin embargo, tales perspectivas dejan aún sin describir o explicar cuestiones tales como las diferencias entre diversos países, o entre diversas subdisciplinas temáticas y metodologías, ni abordan la evolución de los patrones de género y colaboración a través de la coautoría, temas que serán objeto de revisión en las próximas páginas.

5. DATOS: JOURNAL CITATION REPORTS Y MICROSOFT ACADEMIC GRAPH

Para disponer de datos que permitan evaluar la producción científica en sociología se han seleccionado las revistas de impacto utilizando la base de datos de *Web of Science Journal Citation Reports* (JCR) (2020)[1], fuente comúnmente

[1] JCR incluye datos de citas extraídos de aproximadamente 12,000 revistas académicas y técnicas y actas de congresos de más de 3,300 editoriales en más de 60 países. Cubre casi todas las especialidades en ciencia, tecnología y ciencias sociales y permite la evaluación y comparación de revistas para identificar las revistas mejor clasificadas y de mayor impacto en un campo en

reconocida para el establecimiento de rankings de las revistas académicas en función de su impacto en la producción científica. En la fecha de consulta de los datos, JCR contenía información acerca de 165 revistas en la categoría de sociología entre 1997 a 2018. Esta información se enriquece añadiendo las editoriales, los idiomas de publicación, los países de origen y la década en la que aparecen cada una de las revistas. Jacobs (2016) identifica limitaciones en los datos JCR de sociología al no ser del todo exhaustivos. En la categoría de sociología se echan en falta importantes revistas como *American Sociologist, Context* o *Work, Family and Community*. Además, algunas de las revistas presentes, en particular *Annals of Tourism Research* y *Cornell Hospitality Quarterly)*, cuya temática es de dudosa inclusión en esta categoría. Estas dos últimas las eliminamos de nuestra selección. Además, los resultados de analizar estos datos pueden verse condicionados por el sesgo anglosajón en la cobertura de las revistas JCR enfocadas en publicaciones americanas, inglesas y de países bajos debido a la fuerte presencia de importantes editoriales comerciales (Rodriguez-Yunta, 2009).

La información de los artículos y autores de estas revistas se incorporaron desde los datos de Microsoft *Academic Graph* (MAG) *(Sinha et al., 2015)*, fuente que se distribuye libremente con una licencia de datos abierta y contiene información de 230 millones de publicaciones incluyendo artículos de revistas y actas de congresos, casi 240 millones de autores, 50.000 revistas, 4.500 conferencias y 25.500 instituciones. La base de datos de MAG es el resultado de procesos de captura de datos que mezclan información indizada por el buscador *Bing* junto con las fuentes de sindicación (rss) de los editores. Estos datos son enriquecidos mediante procesos automáticos que utilizan inteligencia artificial, en concreto de procesamiento del lenguaje natural, que ayudan a detectar y desambiguar entidades y sus relaciones (autores, afiliaciones, revistas) además de identificar conceptos que definen cada una de las publicaciones y que posteriormente se organizan en una taxonomía (Wang et al., 2019). Estos datos tienen limitaciones al no incluir otros tipos de comunicación científica. La sociología, al igual que otras disciplinas de ciencias sociales, utiliza otros canales de difusión además de los artículos de revistas como son los libros o seminarios (Clemens et al., 1995).

Los datos de JCR se extrajeron de su plataforma online. Para el acceso y la consulta de los datos de MAG se desplegó una infraestructura de *Big Data* en la nube de Microsoft Azure. Los componentes necesarios incluían un almacenamiento de ficheros de texto plano con los datos y un motor de analítica con el que definir el esquema de los ficheros y realizar consultas en lenguaje U-SQL para obtener la selección de datos necesaria[2]. Además, los datos de los autores de las publicaciones se enriquecieron añadiendo el género a través del nombre del autor utilizando Gender-API[3], un servicio online que utiliza diversas fuentes

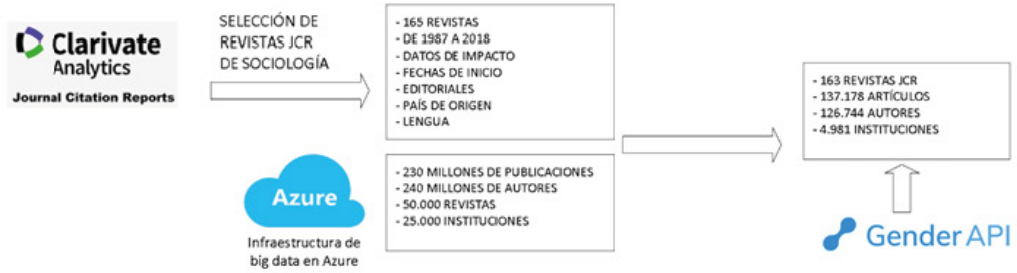
particular.

[2] Instrucciones para desplegar la infraestructura necesaria en la nube están disponibles en la siguiente dirección: https://docs.microsoft.com/en-us/academic-services/graph/get-started-set-up-provisioning

[3] Gender API es una plataforma web capaz de determinar el género a partir del nombre https://gender-api.com/

y clasifica los nombres con un grado por género con un grado de acierto alto (Santamaría y Mihaljević, 2018). La figura 1 muestra esquemáticamente todo el proceso anterior de captura, selección y enriquecimiento de datos.

Figura 1. Proceso de captura y enriquecimiento de datos



6. METODOLOGÍA

Toda la manipulación y análisis de datos para este artículo se ha realizado utilizando el lenguaje de programación estadística R. En la siguiente sección se presentará un análisis descriptivo de las revistas JCR de sociología por países, idiomas de publicación, editoriales, década de aparición e impacto. Posteriormente, se mostrarán las revistas a través de sus artículos describiendo la evolución anual del número de artículos y citas. Estos análisis nos permiten empezar a identificar patrones en las tipologías de las revistas. Tras esto, se seleccionarán algunas de las revistas y se organizarán en cuatro grupos para los que se estudiarán sus diferencias en las dimensiones de citas, coautoría y proporción de género aplicando contrastes de hipótesis.

En la última sección de análisis representamos las relaciones entre revistas y autores a través de una red de afiliación. Las redes de afiliación son redes bipartitas o bimodales, es decir, aquellas que tienen dos tipos de nodos distintos y las conexiones solo se producen entre elementos que pertenecen a tipos diferentes:

Una red de afiliación $G=(U,V,E)$
donde $\forall\, u_1,u_2 \in U, \forall\, v_1, v_2 \in V$ no existen aristas e tales que $e=(u_1, u_2)$ ni $e=(v_1,v_2)$

Las redes bimodales pueden proyectarse convirtiéndose en redes de modo uno. Esta reducción de dimensionalidad permite centrarse en las relaciones de un tipo de nodo. Las redes de afiliación son especialmente útiles para establecer la relación de pertenencia entre actores y grupos. En nuestro caso, la red de afiliación tiene por actores a los autores de las publicaciones y los grupos son las revistas JCR donde publican. A través de esta red podremos detectar aquellos grupos de revistas con comunidades comunes de autores que publican en ellas.

El marco estadístico metodológico que se emplea para generar la red de afiliación es el análisis de coincidencias (Escobar, 2015, Escobar y Tejero, 2018) que es aplicado usando el paquete de R netCoin (Escobar y Martinez-Uribe,

2020). Este marco tiene por objetivo principal detectar el tipo de personas, eventos, atributos, etc. que tienden a aparecer de manera simultánea en un número limitado de espacios.

Partimos de N espacios delimitados denominados escenarios, en cada escenario hay un conjunto de tamaño M de variables aleatorias denominadas eventos $X_j, j=1,...,M$. $X_j = 1$ si el evento j ocurre y $X_j = 0$ si no ocurre. Dos eventos son coincidentes si ocurren en el mismo escenario. A partir de esta misma información, también podría decirse que dos escenarios son semejantes si concurren en ellos los mismos eventos de modo no aleatorio.

El conjunto de los escenarios y los eventos forman una matriz binaria de incidencias $X=(x_{ij})$ de dimensiones NxM con los escenarios en las filas y los eventos en las columnas. Esta matriz es binaria con sus elementos x_{ij} iguales a 0 o 1 indicando si el evento X_j ocurre en el escenario i - ésimo.

Con la matriz de incidencias puede obtenerse la matriz simétrica de coincidencias C de tamaño MxM a través de la operación $\boldsymbol{C= X^T X}$, donde $\boldsymbol{X^T}$ es la matriz transpuesta de X. Cada elemento c_{ij} representa el número de escenarios en los que X_i y X_j tienen el valor 1, es decir, coinciden. Esta matriz de incidencias nos permite obtener las siguiente tres métricas probabilísticas:

La primera es la probabilidad de que suceda un evento X_i que se obtiene dividiendo el número de veces que sucede el evento entre el número de escenarios:

$$P(X_i) = \frac{c_{ii}}{N}$$

También disponemos de la probabilidad conjunta de dos eventos X_i y X_j, denotada por $P(X_i \cap X_j)$, que viene dada por la frecuencia de ocurrencia de los dos eventos en el mismo escenario dividido entre el número de escenarios:

$$P(X_i \cap X_j) = \frac{c_{ij}}{N}$$

Finalmente, tenemos la probabilidad condicional de que sucedan dos eventos X_i y X_j, denotada por $P(X_i \mid X_j)$,, que expresa la probabilidad de que ocurra un evento cuando un segundo evento ya ha ocurrido y se obtiene dividiendo las probabilidades de cada evento:

$$P(X_i|X_j) = \frac{P(X_i \cap X_j)}{P(X_j)} = \frac{c_{ij}}{c_{jj}}$$

De esta forma podemos hablar de que dos eventos coinciden en probabilidad basándonos en el concepto de eventos independientes. Dos eventos X_i y X_j, son independientes cuando:

$$P(X_i|X_j) = P(X_i) \leftrightarrow \frac{c_{ij}}{c_{jj}} = \frac{c_{ii}}{N}$$

Lo cual se cumple si la frecuencia c_{ij} correspondiente a la probabilidad conjunta de X_i y X_j viene dada por:

$$c_{ij}^{*} = \frac{c_{ii}c_{jj}}{N} \; donde \; c_{ij}^{*} \; es \; la \; frecuencia \; esperada$$

Así diremos que dos eventos X_i y X_j son coincidentes en probabilidad (tienen grado de dependencia) si:

$$c_{ij} > \frac{c_{ii}c_{jj}}{N} = c_{ij}^{*}$$

La diferencia entre c_{ij} y c^{*}_{ij} toma una distribución normal con el siguiente error estándar (Haberman, 1973):

$$\sqrt{((\left(1-\frac{c_{ii}}{N}\right)-\left(1-\frac{c_{jj}}{N}\right))}$$

Este error puede utilizarse para estandarizar la diferencia entre el valor empírico de eventos coincidentes c_{ij} y la frecuencia esperada c^{*}_{ij} suponiendo que son independientes y obtener así el obtener el residuo de Haberman ($r_{jk} \sim N(0,1)$).

$$r_{ij} = \frac{c_{ij} - {c_{ij}}^{*}}{((\left(1-\frac{c_{ii}}{N}\right)-\left(1-\frac{c_{jj}}{N}\right))^{\frac{1}{2}}}$$

Así se conforma la matriz de adyacencias A de dimensiones MxM donde dos eventos X_i y X_j se dice que son coincidentes si cumplen la siguiente norma:

$$\mathbf{A}[i,j] = 1 \leftrightarrow P(r_{ij} \leq 0) < c \; \wedge \; i \neq j \; \; donde \; c \; es \; el \; nivel \; de \; significación$$

El valor de c debe ser la probabilidad de concluir que una cierta hipótesis es falsa cuando resulta que es cierta (Error de tipo I) con c supuesto pequeño. Por ello la hipótesis debe ser que $(r_{ij} > 0)$, es decir la hipotesis es que $c_{ij} > c^{*}_{ij}$ lo cual equivale a que X_i y X_j son coincidentes en probabilidad. Así $A[i,j]=1$ cuando la probabilidad de que r_{ij} sea menor o igual a cero es pequeña. Es decir que lo muy probable es que $c_{ij} > c^{*}_{ij}$

Alternativamente, se podría obtener la matriz S de similitudes de escenarios mediante la fórmula $\mathbf{S= XX^{T}}$, en cuyo caso aparecería los elementos s_{pq}, que indicaría cuántos eventos iguales comparten los escenarios p y q. De modo análogo, podrían calcularse el residuo de Haberman y otra matriz de adyacencias A de dimensiones NxN a fin de detectar cuando hay una similitud entre escenarios.

A partir de la matriz de adyacencias A se elabora una red en la que los eventos son los nodos y sus vínculos los valores de los residuos de Haberman. En este artículo, los escenarios son las revistas JCR de sociología y los autores

que publican los artículos en ellas sería los eventos. De esta manera la matriz de adyacencias ***A*** indica qué revistas son semejantes en los autores que en ella publican usando el residuo de Haberman para indicar la fuerza de la relación.

7. ANÁLISIS DESCRIPTIVO DE LAS REVISTAS JCR DE SOCIOLOGÍA

Las revistas JCR de sociología conforman un conjunto de publicaciones especializadas controladas por editoriales en múltiples idiomas. Las tablas a continuación (tabla 1) muestran la distribución de las revistas por países, idiomas y editoriales. De las 163 revistas 67 son de Estados Unidos y 54 de Reino Unido. El resto de las revistas provienen de otros 21 países, la mayor parte del continente europeo, aunque también de Asia, América y Oceanía.

El 90% de las revistas publican en inglés, aunque también las hay con artículos en distintos idiomas simultáneamente. A nivel editorial, dominan los grandes conglomerados editoriales internacionales como SAGE, Taylor & Francis, Elsevier, Blackwell o Routledge. A estos los acompañan, con menor representación, editoriales universitarias como las de Oxford, Cambridge y Chicago.

Tabla 1. Distribución de revistas por países de origen, idioma y por editoriales

País	*Revistas*	%
Estados Unidos	67	41.1%
Reino Unido	54	33.1%
Países bajos	10	6.13%
Alemania	7	4.29%
Canadá	3	1.84%
Otros	22	13.5%

Idioma	*Revistas*	%
Inglés	147	90.2%
Alemán	5	3.07%
Frances	3	1.84%
Español	3	1.84%
Otros	5	3.07%

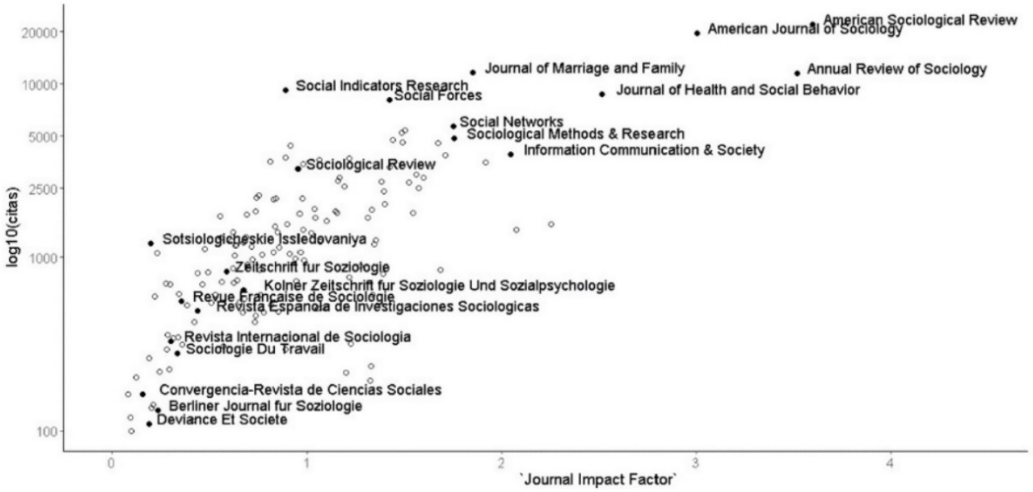
Editoriales	Revistas	%
SAGE	40	24.5%
Taylor & Francis	13	7.98%
Blackwell Publishing Inc	12	7.36%
Elsevier Ltd.	12	7.36%
Routledge	6	3.68%
Wiley	6	3.68%
Oxford University Press	5	3.07%
Cambridge University Press	4	2.45%
Kluwer Academic Press	4	2.45%
Springer Verlag	4	2.45%
University of Chicago Press	4	2.45%
Otras	53	32.5%

Khun (1962:19) asocia la aparición de revistas especializadas en una disciplina a momentos de cambio de paradigma. En la sociología las publicaciones pioneras son *American Journal of Sociology*, creada en 1895 por Albion W. Small en la Universidad de Chicago y *Sociological Review* fundada en Reino Unido en 1908 por Leonard T. Hobhouse. Tras estas publicaciones aparecen otras tres americanas, *Social Forces* en 1922, *American Sociological Review* en 1936 y el *Journal of Marriage and Family* en 1939. La figura 2 muestra las revistas por década de aparición. En la década de los 60 hay un aumento importante de aparición de revistas, este aumento alcanza su máximo en la década de los años 70 con más de 40 revistas. En esta década encontramos *Annual Review of Sociology*, *Sociological Methods and Research* y *Social Networks.* Un 40% de las revistas aparecen en las tres últimas décadas, aunque en la década más reciente apenas empiezan su andadura 7 revistas.

Figura 2. Distribución de revistas por década de aparición

Las publicaciones quedan clasificadas de acuerdo con su influencia en la disciplina a través de métricas de impacto. La métrica más sencilla es el número total de citas, JCR lo calcula teniendo en cuenta las citas entre las revistas disponibles en su base de datos. Sin embargo, la métrica por excelencia para medir el impacto es el *Journal Impact Factor (JIF)* que se calcula anualmente al dividir las citas de la revista ese año por el número de artículos en los dos años anteriores. En la figura 3 representamos las revistas con su impacto en el eje horizontal y las citas en el eje vertical. Aparece para cada revista el promedio de los JIF anuales y el total de citas más reciente. Solo han quedado representadas aquellas revistas con más de 100 citas con un eje vertical expresado en escala logarítmica en base 10 para ajustar las importantes diferencias en el total de citas entre las revistas.

Figura 3. Revistas organizadas por factor de impacto y número de citas



En la parte superior derecha del gráfico aparecen las revistas pioneras y generalistas. *American Sociological Review* y *American Journal of Sociology* ocupan los más lugares destacados. Junto a estas dos aparece *Annual Review of Sociology*, uno de los denominados *review journals*, que por este motivo obtienen un número elevado de citas (Moed, 2005). Tras estas publicaciones aparecen revistas temáticas tales como *Journal of Marriage and Family* y *Journal of Health and Social Behaviour que* incorporan investigación de interés actual sobre el género, la familia la salud y la medicina. Más centrada en el gráfico aparece *Social Forces*, una de las revistas prominentes en el campo y junto a *American Sociological Review* y *American Journal of Sociology* perteneciente a la denominada "Triple Corona". Su diferencia de citas e impacto con las otras dos posiblemente sea debido a la capacidad de las otras dos de distanciarse de las demás (Jacobs, 2016). Alrededor de Social Forces se encuentra *Information Communication & Society,* revista que comienza en 2001y centrada en temáticas de creciente interés como son los estudios sobre la sociedad de la comunicación y el impacto de las nuevas tecnologías. Las otras tres revistas de este grupo, todas de los años 70, presentan una aproximación empírica y metodológica. Son *Sociological Methods and Research, Social Networks* y *Social Indicators Research.* Un poco más abajo aparece *Sociological Review,* otra de las pioneras pero que no llega a tener los indicadores de influencia de ASR y AJS. Las publicaciones en otros idiomas como el alemán, francés o castellano aparecen en la parte inferior izquierda.

Explorar la procedencia y el impacto de las revistas es un buen punto de partida. Sin embargo, es necesario incluir en el análisis a los autores e instituciones de las revistas. Es pues preciso bajar un nivel de profundidad y estudiar los artículos. Para ello se emplean los datos de MAG filtrando los artículos de las 163 revistas JCR de sociología de 1997 a 2018. Hay cinco revistas para las

que no hay artículos en MAG en estos años: *Ethology and Sociobiology, Innovation, Australian and New Zealand Journal of Sociology* y *Studies in Symbolic Interaction.* En total se cuenta con 164.036 artículos que han sido revisados para eliminar las recensiones de libros eliminando aquellos títulos que contienen las palabras *"book review"*. Al final de este proceso se obtuvieron 139.452 artículos. Finalmente, tras descartar los artículos repetidos se consiguió la cifra final de 137.178 artículos de estas 158 revistas entre 1997 y 2018.

En las tablas de abajo (tabla 2) se presentan los listados de las diez revistas con más artículos y citas por artículo. La revista *Contemporary Sociology* cuenta con 6.648 representando casi el 5% del total. *American Sociological Review* es la revista con la media más alta de citas por artículo con 119, además sus citas representan un 4.7% del total de citas en esta base de datos.

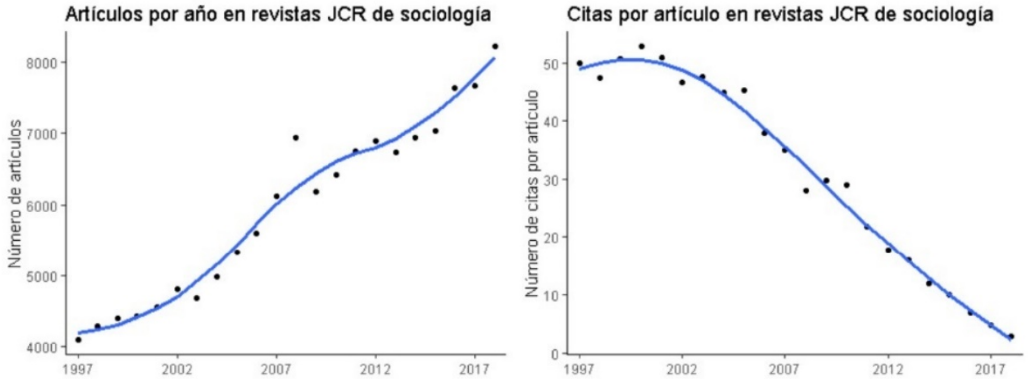
Tabla 2. Arriba las diez revistas con más artículos en los datos y abajo la diez revistas con la media de citas por artículos más alta en los datos

Revistas	Número de artículos en MAG	Proporción con respecto al total
Contemporary Sociology	6.648	4.85%
Social Indicators Research	3.904	2.85%
American Journal of Sociology	3.482	2.54%
Ethnic and Racial Studies	2.901	2.11%
Journal of Marriage and Family	2.372	1.73%
Society	2.028	1.48%
Human Ecology	1.956	1.43%
Social Forces	1.906	1.39%
Social Science Quarterly	1.782	1.3%
The Sociological Review	1.779	1.3%

Revistas	***Número de citas en MAG***	***Media de citas por artículo***	***Proporción de citas con respecto al total***
American Sociological Review	128.115	119	4.69%
Journal of Health and Social Behavior	60.218	91.5	2.20%
Sociology Of Education	31.323	68.1	1.15%
Journal of Marriage and Family	134.822	56.8	4.93%
Gender & Society	46.993	53.8	1.73%
Sociological Methodology	16.599	53.0	0.61%
Social Networks	44.956	50.9	1.65%
Social Problems	34.936	49.9	1.28%
Evolution and Human Behavior	54.333	46.6	2.00%
Population and Development Review	42.699	45.8	1.57%

De aquí en adelante, y con el fin de intuir mejor las tendencias temporales, representaremos las series temporales con un método de curva suavizada de ajuste de regresión polinómica. En los gráficos siguientes (figura 4) se representa la evolución anual del número de artículos y citas de las 158 revistas de sociología. En 20 años el número de artículos se duplica y las medias más altas de citas por artículo se acumulan en los primeros años.

Figura 4. Evolución anual del número de artículos y de citas. Las curvas utilizan el método de suavizado de ajuste de regresión polinómica.



Al fijarnos en los artículos con más citas (tabla 3), resalta que hay 3 de ellos del *American Journal of Sociology,* que la mitad son de 1997 o 1998 y que presentan una diversidad de enfoques y temáticas desde temas de salud, a elementos metodológicos y teóricos.

Tabla 3. Los diez artículos con más citas en MAG

Citas acumuladas	Referencias
5.901	Idler, Ellen L., and Yael Benyamini. "Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies." Journal of Health and Social Behavior 38, no. 1 (1997): 21-37.
4.539	Burnham, K. P., & Anderson, D. R. (2004). "Multimodel Inference: Understanding AIC and BIC in Model Selection". *Sociological Methods & Research*, *33*(2), 261–304.
3.700	Connell, R. W., & Messerschmidt, J. W. (2005). "Hegemonic Masculinity: Rethinking the Concept". Gender & Society, 19(6), 829–859.
3.325	Woolcock, Michael. "Social Capital and Economic Development: Toward a Theoretical Synthesis and Policy Framework." *Theory and Society* 27, no. 2 (1998): 151-208.
2.878	Inglehart, Ronald, and Wayne E. Baker. "Modernization, Cultural Change, and the Persistence of Traditional Values." American Sociological Review 65, no. 1 (2000): 19-51

Citas acumuladas	Referencias
2.730	Burt, Ronald S. "Structural holes and good ideas." *American Journal of Sociology* 110, no. 2 (2004)
2.647	Heckathorn, Douglas D. "Respondent-driven Sampling: a New Approach to the Study of Hidden Populations." *Social Problems* 44, no. 2 (1997): 174-199.
2.478	Meyer, John W., John Boli, George M. Thomas, and Francisco O. Ramirez. "World society and the nation-state." American Journal of sociology 103, no. 1 (1997): 144-181.
2.463	Reckwitz, Andreas. "Toward a theory of social practices: A development in culturalist theorizing." European Journal of Social Theory 5, no. 2 (2002): 243-263.
2.359	Emirbayer, Mustafá, and Ann Mische. "What is Agency?". (1998) American Journal of Sociology. Vol 13, no. 4, Jan 1998. 962-1023.

Los datos de MAG proporcionan información de los 126.744 autores presentes en la selección de artículos. Además, hay 4.981 instituciones relacionadas con estos autores. Esta información aparece en un 37% de las relaciones entre autor y artículo. Las tablas 4 y 5 presentan los autores e instituciones con más citas. Paul R. Amato, de la Universidad de Pennsylvania, es el autor con más citas, seguido de Robert J. Sampson de la Universidad de Harvard, Alejandro Portes de la Universidad de Princeton, Ed Diener de la Universidad de Illinois, Tom A. B. Snijders de la Universidad de Oxford y Ellen L. Idler de la Universidad de Emory. Entre las instituciones con más citas encontramos un predominio de universidades americanas de gran prestigio como Michigan, Pennsylvania, Cornell, Ohio, Texas, Harvard, California (Berkeley y UCLA) o University of Wisconsin-Madisson.

Tabla 4. Autores con más citas

Autor	***Número de artículos***	***Número de citas***	**Revistas distintas**	Media de citas por artículo	***Artículo con más citas de cada autor***
Paul R. Amato	48	8.163	7	170.0	Amato, Paul R. "The consequences of divorce for adults and children." Journal of Marriage and Family 62, no. 4 (2000): 1269-1287.
Robert J. Sampson	23	7.306	11	317.6	Sampson, Robert J., and Stephen W. Raudenbush. "Systematic social observation of public spaces: A new look at disorder in urban neighborhoods." American Journal of Sociology 105, no. 3 (1999)

Autor	***Número de artículos***	***Número de citas***	**Revistas distintas**	Media de citas por artículo	***Artículo con más citas de cada autor***
Alejandro Portes	43	6.322	16	147.0	Portes, Alejandro, Luis E. Guarnizo, and Patricia Landolt. "The study of transnationalism: pitfalls and promise of an emergent research field." Ethnic and Racial Studies 22, no. 2 (1999): 217-237.
Ed Diener	23	6.314	2	274.5	Diener, Ed, and Eunkook Suh. "Measuring quality of life: Economic, social, and subjective indicators." Social Indicators Research 40, (1997)
Tom A. B. Snijders	43	6.239	6	145.0	Snijders, Tom AB, Gerhard G. Van de Bunt, and Christian EG Steglich. "Introduction to stochastic actor-based models for network dynamics." Social networks 32, no. 1 (2010): 44-60.
Ellen L. Idler	7	6.078	3	868.2	Idler, Ellen L., and Yael Benyamini. "Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies." Journal of Health and Social Behavior 38, no. 1 (1997): 21-37.
Stephen P. Borgatti	22	5.994	3	272.4	Borgatti, Stephen P. "Centrality and network flow." Social Networks 27, no. 1 (2005): 55-71.
Douglas D. Heckathorn	10	5.723	5	572.3	Heckathorn, Douglas D. "Respondent-driven sampling: a new approach to the study of hidden populations." Social Problems 44, no. 2 (1997)
Douglas S. Massey	73	5.506	19	75.4	Massey, Douglas S., and Kristin E. Espinosa. "What's driving Mexico-US migration? A theoretical, empirical, and policy analysis." American Journal of Sociology 102, no. 4 (1997): 939-999.
Michel Callon	17	5.169	5	304.0	Callon, Michel. "Introduction: the embeddedness of economic markets in economics." The Sociological Review 46, no. 1_suppl (1998): 1-57.

Tabla 5. Instituciones con más citas

Institución	*Número de artículos*	*Número de citas*	*Número de autores*	*Media de citas por artículo*
University of Michigan	1346	72.135	722	53.59
Pennsylvania State University	1537	65.473	645	42.60
Cornell University	1148	52.972	488	46.14
Ohio State University	1034	52.611	440	50.88
University of Texas at Austin	1138	47.811	502	42.01
University of Wisconsin-Madison	1050	43.810	572	41.72
University of California, Los Angeles (UCLA)	971	43.753	520	45.06
Harvard University	924	43.558	534	47.14
University of North Carolina at Chapel Hill	857	42.833	424	49.98
University of California, Berkeley	836	38.743	497	46.34

7.1. Análisis de citas, coautoría y proporción de género por tipos de revistas

En los anteriores análisis descriptivos hemos visto que hay diferentes tipos de revistas: las revistas más antiguas, las temáticas, las metodológicas o las de lengua no inglesa. ¿Tienen estos tipos de revistas distintos patrones de citas, coautoría o proporción de género? Para averiguarlo organizamos algunas de las revistas en los cuatro tipos que indicamos en la tabla abajo (tabla 6). Incluimos las tres revistas de la triple corona, todas las revistas de habla no inglesa y todas las revistas puramente metodológicas. Para las revistas temáticas seleccionamos las trece primeras en cuanto a su factor de impacto y número de citas en JCR.

Tabla 6. Agrupación de algunas revistas en cuatro grupos

Revistas de la triple corona
American Sociological Review
American Journal of Sociology
Social Forces
Revistas de lengua no inglesa
Archives Européennes de Sociologie
Berliner Journal fur Soziologie
Convergencia-Revista de Ciencias Sociales
Deviance Et Societe
Drustvena Istrazivanja
Filosofija-Sociologija

Revistas temáticas
Ethnic and Racial Studies
Gender & Society
Information Communication & Society
International Journal of Intercultural Relations
International Political Sociology
Journal for The Scientific Study of Religion
Journal of Consumer Culture
Journal of Health and Social Behavior
Journal of Marriage and Family
Society & Natural Resources

Kolner Zeitschrift fur Soziologie Und Sozialpsychologie	Socio-Economic Review
Revista Española de Investigaciones Sociológicas	Sociology of Education
Revista Internacional de Sociología	Sociology of Health & Illness
Revue Française de Sociologie	**Revistas de métodos**
Sociologicky Casopis	Journal of Mathematical Sociology
Sociologie Du Travail	Qualitative Research
Sociologija I Prostor	Qualitative Sociology
Sociologisk Forskning	Social Indicators Research
Sotsiologicheskie Issledovaniya	Social Networks
Soziale Welt-Zeitschrift fur Sozialwissenschaftliche Forschung Und Praxis	Sociological Methodology
Zeitschrift fur Soziologie	Sociological Methods & Research
	Journal of Mathematical Sociology

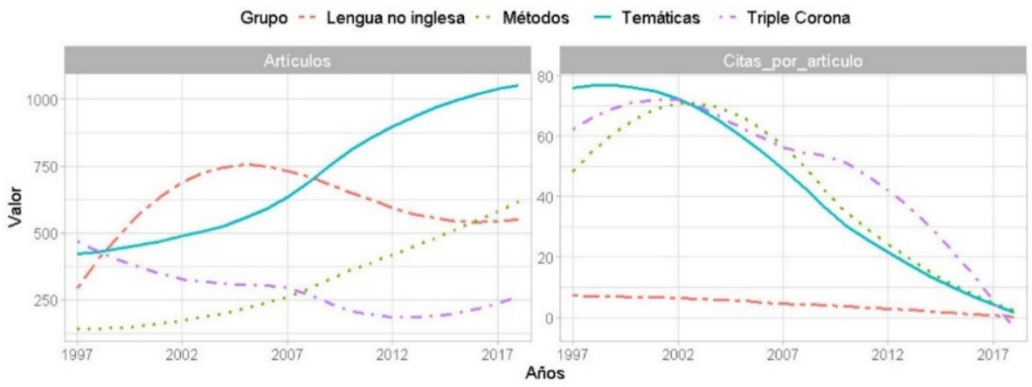
Para cada uno de estos grupos se obtuvieron el número de artículos y la media de citas por artículo (tabla 7). Las revistas de la triple corona son las que poseen la media más alta de citas por artículo, casi 55, seguidas de las temáticas y las de métodos. Las revistas de lengua no inglesa tienen la media de citas por artículo más baja con tan solo 4,5.

Tabla 7. Artículos, media de citas por artículo y desviación estándar por grupo de revistas

Grupo	Artículos	Media de citas por artículo	Desviación estándar
Triple corona	5.984	54,6	142,0
Temáticas	14.531	37,3	98,9
Métodos	6.294	34,2	108,0
Lengua no inglesa	12.719	4,5	16,4

Al representar anualmente los artículos y la media de citas por artículo (figura 6) puede apreciarse que el número de artículos de las revistas de la triple corona desciende con los años. Este efecto se debe principalmente al descenso de artículos publicados en la *American Journal of Sociology*. Por otro lado, el número de artículos de las revistas temáticas y de métodos aumenta debido al aumento de artículos en revistas como *Social Indicators Research* o *Ethnic and Racial Studies*. Basándose en las citas por artículo, las revistas de los tres grupos tienen tendencias similares a primera vista. Por su lado, las revistas de lengua no inglesa alcanzan su máximo de artículos en torno al año 2005, pero como ya se ha comentado están lejos de las revistas americanas o británicas en citas por artículo.

Figura 6. Evolución del número de artículos y citas por artículo por grupos de revistas. Las curvas utilizan el método de suavizado de ajuste de regresión polinómica.



Para comprobar si las diferencias en la media de citas por artículo entre los cuatro grupos de revistas son significativas se empleó el análisis de varianza (ANOVA) para probar la hipótesis nula de que no hay diferencias entre las medias de los diferentes grupos. Los resultados de ANOVA entre grupos verifican que la diferencia de medias es estadísticamente significativa ($p<.05$) con una $F_{3,39.524}$ de 494,4. La prueba post hoc de Tukey (tabla 8) señala que todos los grupos difieren significativamente ($p<.05$), excepto la media de citas por artículo de las revistas temáticas y las de métodos. Las revistas de la triple corona son las que obtienen la media de citas por artículos más alta y con mayor diferencia que las otras revistas en los otros grupos.

Tabla 8. Resultados de la prueba de Tukey para las medias de citas

Resultados dos a dos	***Diferencia***	***p-valor***
Métodos - Lengua no inglesa	29,73	0
Temáticas - Lengua no inglesa	32,84	0
Triple Corona - Lengua no inglesa	50,05	0
Temáticas – Métodos	3,1	0,11
Triple Corona – Métodos	20,13	0
Triple Corona - Temáticas	17,21	0

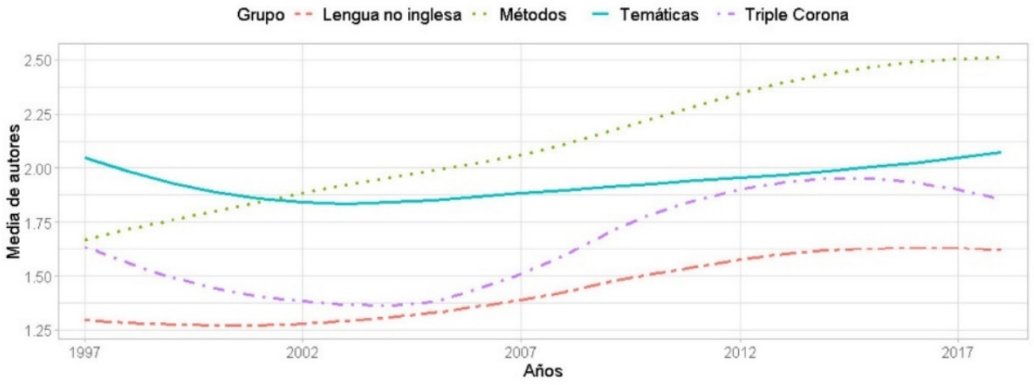
El número de coautores por artículo se ha ido incrementando en todas las disciplinas. En sociología es algo cada vez más común motivado por razones de competitividad y progresión académica, el aumento de las oportunidades de colaboración o la propia naturaleza de la investigación interdisciplinar (Taylor & Francis, 2017). Al calcular la media de autores por artículo y por grupo de revistas (tabla 9), pueden observarse diferencias entre los distintos grupos. Las revistas de métodos con más de 2 autores por artículo superan a las temáticas con 1,9 y a las de la triple corona con 1,6.

Tabla 9. Media de coautores por artículo y desviación estándar por grupo de revistas

Grupo	Media de autores por artículo	***Desviación estándar***
Triple Corona	1,59	0,99
Temáticas	1,94	1,52
Métodos	2,22	1,52
Lengua no inglesa	1,42	0,81

El gráfico de evolución anual (figura 7) muestra un crecimiento en la media de coautores para los todos los grupos a excepción de las revistas de métodos que parten de una media de dos autores y la mantienen en el tiempo. Las revistas temáticas son las que experimentan mayor aumento pasando de 1,6 autores por artículo a casi 2,5.

Figura 7. Media anual de coautores por grupos de revistas



De nuevo, se comprueba que existen diferencias significativas en la media de autores por artículo entre los cuatro grupos de revistas utilizando el análisis de varianza (ANOVA) para probar la hipótesis nula de que no hay diferencias entre las medias de autores de los diferentes grupos. Los resultados de ANOVA entre grupos verifican la diferencia estadísticamente significativa ($p<.05$) de estas medias ($F_{3,39.400}=737$, $p<2e-16$). La prueba post hoc de Tukey (tabla 10) señala que todos los grupos difieren significativamente ($p<.05$). Las revistas de métodos son las que obtienen la media de autores por artículos más alta superando en 0.6 coautores por artículo a las de la triple corona, 0,23 a las temáticas y en 0,74 a las de lengua no inglesa.

Tabla 10. Resultados de la prueba de Tukey para las medias de autores

Resultados dos a dos	*Diferencia*	*P adj*
Métodos - Lengua no inglesa	0,79	0
Temáticas - Lengua no inglesa	0,51	0
Triple Corona - Lengua no inglesa	0,17	0
Temáticas – Métodos	-0,28	0
Triple Corona – Métodos	-0.62	0
Triple Corona - Temáticas	-0.34	0

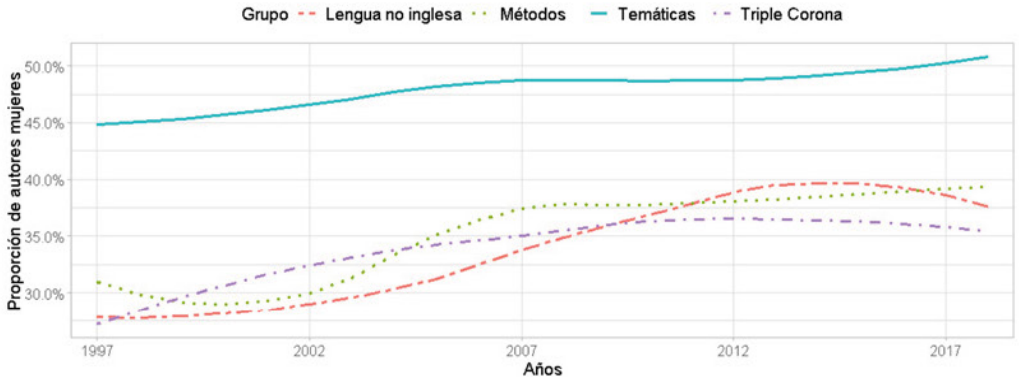
Otra de las cuestiones que puede abordarse es la referente a la existencia de sesgo de género en las publicaciones de estas revistas. A través de los nombres de los autores se puede asignar un género con el fin de estudiar la proporción de mujeres en las publicaciones. Del total de 19.985 nombres únicos de autores se ha dispuesto del género en 18.341 casos, un 92%. Un 44% de los autores de los artículos son mujeres. La proporción de género es distinta en los artículos de los cuatro grupos de revistas (tabla 11) con las revistas temáticas con la proporción de mujeres más alta.

Tabla 11. Proporción media de autores de género femenino por artículo y desviación estándar por grupo de revistas

Grupo	*Proporción media de autoras por artículo*	*Desviación estándar*
Triple Corona	0,329	0,425
Temáticas	0,477	0,436
Métodos	0,356	0,401
Lengua no inglesa	0,321	0,431

En la figura 8 puede observarse el aumento en la proporción de autores de género femenino desde 1997 para los cuatro grupos y la importante diferencia de las revistas temáticas con las de los otros tres grupos. Revistas como *Gender & Society* cuentan con una proporción del 83% de mujeres autores o *Journal of Marriage and Family* con un 60%. Llama también la atención encontrar dentro de las revistas de métodos dos con una proporción de mujeres por encima de las demás en el grupo, se trata de *Qualitative Sociology* y *Qualitative Research.*

Figura 8. Evolución anual de la proporción de autores de género femenino para los cuatro grupos de revistas



El análisis de varianza ANOVA nos permite ver si la diferencia entre las medias de citas y de autores es significativa, pero no nos vale para la diferencia entre proporciones de género. Para este caso usamos una prueba de proporciones con la que comprobamos si la diferencia entre las proporciones de autores de género femenino de los distintos grupos es estadísticamente significativa. Utilizamos la hipótesis nula de que las proporciones son iguales entre los grupos. El resultado de la prueba confirma las diferencias significativas ($p<.05$) entre las proporciones de los cuatro grupos. Las revistas temáticas son las que obtienen la proporción más alta de mujeres por artículo superando en un 15% a las de la triple corona y las de lengua no inglesa y en un 12% a las temáticas.

8. LA RED DE AFILIACIÓN ENTRE AUTORES Y REVISTAS

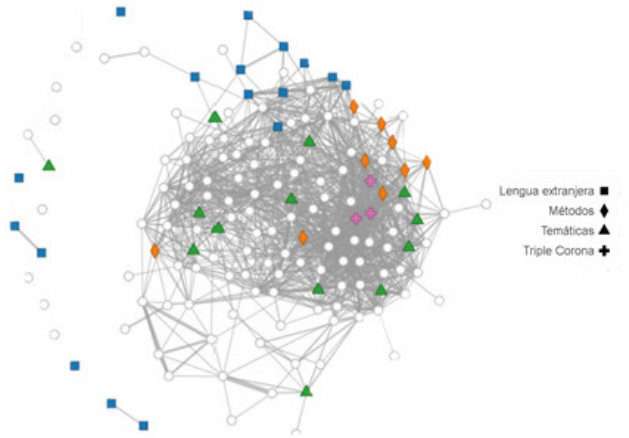
Con el fin de entender mejor la relación entre los autores y las revistas JCR de sociología, se representarán sus relaciones a través de una red de afiliación donde los actores son los autores de las publicaciones y los grupos las revistas donde publican. A través de esta red podrán detectarse aquellos grupos de revistas con comunidades comunes de autores.

Para los cálculos del análisis de coincidencias eliminamos los artículos de *Contemporary Sociology* para evitar el efecto que pueda tener una revista con una proporción de artículos de revisión elevada, pero con un impacto bajo. Así partimos de la matriz de incidencias $\boldsymbol{X}$ que tiene 127.368 filas que representan a los autores y 157 columnas, una por revista. Los valores x_{ij} de esta matriz toman el valor 1 si el autor *i - ésimo* ha publicado alguna vez en la revista *j - ésima* y cero en caso contrario. A partir de esta matriz binaria de incidencias $\boldsymbol{X}$ se aplica el análisis de coincidencias y se calcula la matriz de adyacencias A. La red de afiliación pasa de ser bipartita a modo uno formada por nodos que representan las 158 revistas cuyas relaciones, establecidas mediante el residuo de Haberman,

representan la fuerza de la relación entre las revistas. Cuantos más autores hayan publicado en dos revistas, más fuerte será su relación. Para simplificar la red solo se tienen en cuenta aquellas relaciones con Haberman > 3 asegurando que las relaciones son significativas con un nivel de confianza mayor del 99,8% en pruebas de una sola cola.

La red resultante se muestra en la figura 9. Los nodos representan las revistas JCR y las aristas representan los enlaces que establecen la relación entre las revistas. Utilizamos el algoritmo basado en fuerzas de atracción repulsión de redes de Fruchterman-Reingold (Fruchterman y Reingold, 1991). Esta red consta de un componente central con 139 nodos conectados y otros 18 nodos separados. Las formas de los nodos representan los grupos de la sección anterior. Todas las revistas de la triple corona y todas las de métodos forman parte del componente central mientras que una de las temáticas y varias de las revistas de lengua no inglesa están fuera de él. Las tres revistas más prestigiosas se encuentran próximas a la derecha del componente central. Las revistas de métodos se agrupan arriba a la derecha del componente principal, aunque dos de ellas (las cualitativas) aparecen en otras zonas de este mismo componente. Las revistas temáticas están distribuidas por varias zonas sin una agrupación clara. Finalmente, las revistas de lengua no inglesa están en la parte superior de la red y siete de ellas desconectadas del componente central.

Figura 9. Red de afiliación de revistas utilizando la forma del nodo para representar los grupos de revistas



Se calcularon dos medidas de centralidad de redes: el grado con pesos y la centralidad de intermediación. En la red de la figura 10 los colores de los nodos representan el grado ponderado en una escala de color. Las revistas de la triple corona tienen todas valores altos, siendo *Social Forces* la revista con mayor número de enlaces con otras revistas. Al fijarnos en la tabla 12 con las quince revistas con mayor grado con peso identificamos revistas americanas y británicas

Figura 10. Redes de afiliación con escalas de color para representar el grado con pesos de los distintos nodos. La red de la derecha es la misma que la de la izquierda, pero con centrada en los nodos con mayor grado con pesos junto de sus nombres.

Tabla 12. Las quince revistas con mayor grado con pesos de la red de afiliación

Revista	Grado con pesos
Social Forces	7.177
Social Science Research	6.079
American Sociological Review	4.694
American Journal of Sociology	3.447
Sociological Forum	3.284
British Journal of Sociology	3.142
Journal of Marriage and Family	3.132
Sociological Quarterly	3.014
Social Problems	2.960
Sociology	2.856
European Sociological Review	2.850
Sociological Perspectives	2.813
The Sociological Review	2.546
Sociology Compass	2.157
Ethnic and Racial Studies	2.066

En la figura 11 los colores de los nodos denotan el grado de intermediación y la tabla 13 presenta los datos de las primeras 20 revistas. Los nodos con mayor

grado de intermediación son *British Journal of Sociology, European Societies* y *Sociological Theory*. Estos nodos tienen un papel importante en la red ya que suelen actuar como controladores de flujo de información al ser puentes entre grupos.

Figura 11. Redes de afiliación con escalas de color para representar la intermediación de los distintos nodos. La red de la derecha es la misma que la de la izquierda, pero centrada en los nodos con mayor grado de intermediación junto con sus nombres.

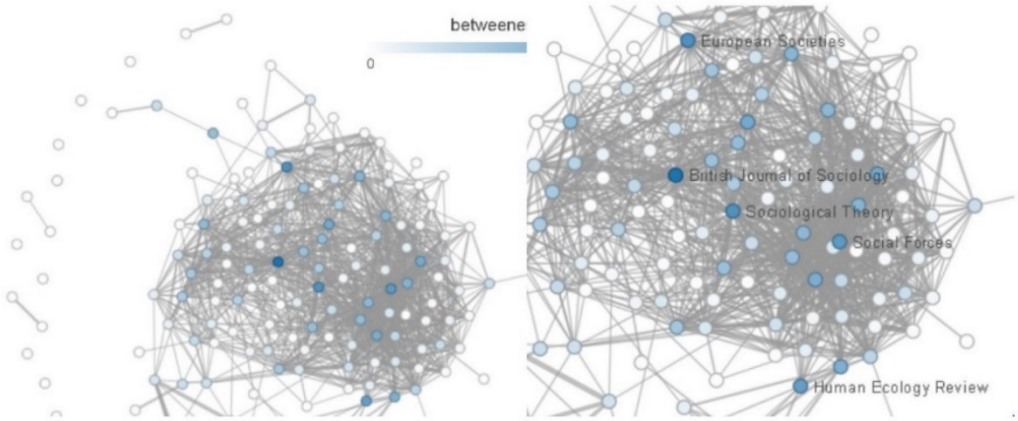
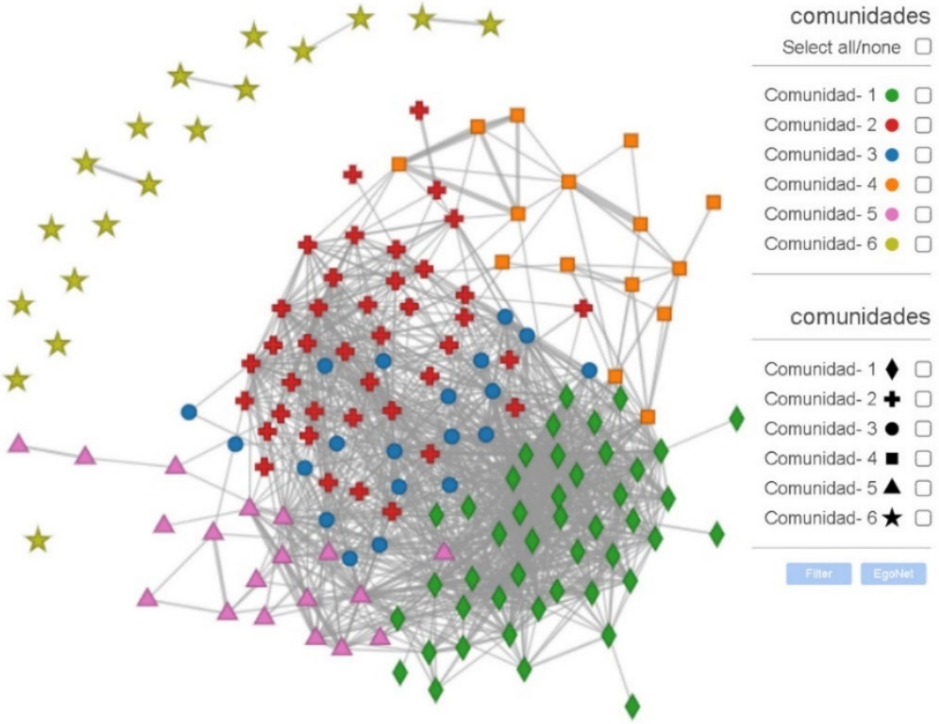


Tabla 13. Las quince revistas con mayor grado de intermediación de la red de afiliación

Revista	Centralidad de intermediación
British Journal of Sociology	611,02
European Societies	478,88
Sociological Theory	474,16
Social Forces	440,56
Human Ecology Review	420,08
International Sociology	370,58
Social Science Research	367,06
Sociological Perspectives	354,68
Rural Sociology	346,39
American Sociological Review	325,63
European Sociological Review	320,33
Sociological Quarterly	306,29
Research in Social Stratification and Mobility	305,68
European Journal of Social Theory	289,02
Sociological Forum	281,95

Tras estos análisis, se aplicó el algoritmo de Louvain para el cálculo de comunidades (Blondel et al., 2008) y se obtuvieron 19 comunidades en la red de afiliación. Como 5 de las comunidades acumulan el 90% de los nodos las restantes se agruparon en una única comunidad (Comunidad-6). La figura 12 presenta la disposición de estas seis comunidades.

Figura 12. La red de afiliación con la forma y color de los nodos identificando las comunidades obtenidas por el algoritmo de Louvain.



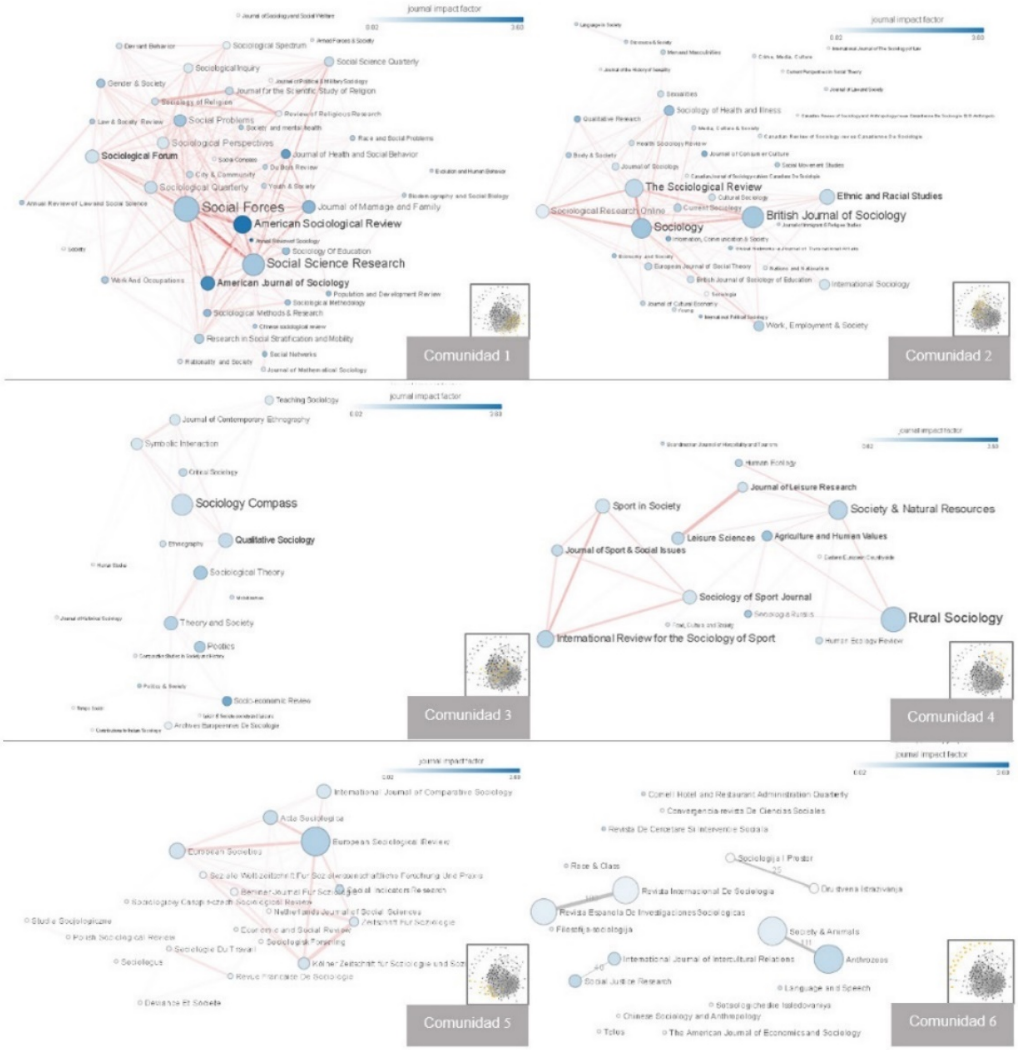
A continuación, se describirán cada una de las comunidades. Para ello se representarán en la figura 13 las subredes mediante el grado como tamaño de los nodos y un degradado de color para el factor de impacto de los nodos. Los enlaces también utilizan un degradado de color que representa el número de autores coincidentes entre las revistas.

La comunidad 1, situada a la derecha de la red, es la más extensa con 45 nodos. Son los nodos con mayor factor de impacto, con predominancia de revistas americanas, especialmente las más tradicionales. Las revistas de la triple corona ocupan un lugar central en esta comunidad que incluso contiene la mayoría de las revistas de métodos cuantitativos. Además, contiene revistas temáticas enfocadas a la religión, la sociología militar, el género, la familia, la juventud y la salud. La comunidad 2, es la siguiente en tamaño con 40 nodos y aparece a

la izquierda del componente principal de la red. Esta comunidad está dominada por cuatro revistas del Reino Unido: *British Journal of Sociology*, *Sociology, Sociological Review* y *Sociological Research Online*. Cuenta con una revista de métodos cualitativa *Qualitative Research* e incluye dos de las revistas temáticas de mayor impacto *Information, Communication & Society* y *Journal of Consumer Culture*. Además, contiene revistas que cubren temáticas como la sociología del lenguaje, salud, media y cultura, inmigración, estudios raciales, nacionalismo y estudios internacionales.

La comunidad situada entre las dos anteriores es la comunidad 3 y consta de 20 nodos. Contiene revistas teóricas como *Sociological Theory* y *Theory and Society,* además de revistas de corte cualitativo como *Qualitative Sociology, Ethnography* y *Journal of Contemporary Ethnography*. En la parte superior de la red está la comunidad 4, está comunidad está compuesta por revistas temáticas que abordan materias como la sociología rural, la agricultura, los recursos naturales, el ocio o el deporte. La comunidad que aparece en la parte inferior de la red es la comunidad 5 e incluye gran cantidad de revistas de lengua no inglesa enlazadas con el componente principal a través de revistas de temática internacional como *European Sociological Review, International Journal of Comparative Sociology* y *European Societies*. Finalmente, la comunidad 6 agrega 18 revistas desconectadas del componente principal donde encontramos enlaces entre revistas de los mismos países como las dos revistas españolas *Revista Española de Investigaciones* y la *Revista Internacional de Sociología* o las croatas *Sociologija I Prostor* y *Drustvena Istrazivanja*. También hay enlaces entre revistas de temática común como *Society & Animals* y *Anthrozoos*.

Figura 13. Las seis comunidades representadas como subgrafos usando el grado para el tamaño de los nodos y el factor de impacto para el color



9. DISCUSIÓN DE RESULTADOS

Las revistas JCR de sociología y sus artículos conforman una rica fuente para conocer el conjunto de temáticas, sistema de comunicación compuesto por investigadores, instituciones y editoriales que complementa sostienen el conocimiento conjunto de la sociología como disciplina. Como expresó operativamente Inkeles (1964, 8): la sociología puede ser abordada como "lo que hacen los

sociólogos" y en la actualidad la mayor parte de los sociólogos académicos se dedican a publicar artículos en revistas de impacto.

Los análisis realizados muestran una sociología dominada por grupos editoriales internacionales y en menor medida por unas pocas editoriales universitarias. Como es de común conocimiento, el idioma predominante es el inglés y dos de cada tres revistas tienen su origen en Estados Unidos y Reino Unido. Las universidades de ambos países también aportan los autores con más artículos y citas.

Destacan en la disciplina las revistas de la llamada triple corona: *American Journal of Sociology, American Sociological Review y Social Forces*. En el listado de las incluidas en el listado JCR, son las primeras en aparecer y ejercen un dominio absoluto en citas e impacto. Por otro lado, estas revistas poseen un grado de coautoría bajo y la mayor desproporción de género entre sus autores. Tras ellas, en los años setenta aparecen gran cantidad de revistas temáticas y metodológicas. Las revistas metodológicas tienen la media más alta de coautores con más de dos autores por artículo aportando similares resultados a los obtenidos por Moody (2004). Por su parte, Grant y Ward (1991) también señalaron un mayor equilibrio de género entre los autores que publican en las revistas temáticas y, en especial, en aquellas con enfoques de género, familia o salud, que obtienen proporciones más altas de autoría de género femenino (Grant y Ward, 1991

Las revistas se agrupan en comunidades a través de las redes de afiliación que forman con los autores. La red resultante corresponde claramente con el primer tipo de red de colaboración que propone Moody (2004). Se trata de una red compuesta por múltiples especialidades desconectadas y con subredes altamente agrupadas reflejando la falta de una teoría unificada de la disciplina. Las tres revistas americanas más importantes lideran la comunidad con mayor número de miembros como ya anticipaban (Moody y Light, 2006). Esta subred está también compuesta por un grupo de revistas metodológicas más recientes y otras sobre temas de religión, sociología militar, juventud y salud. La siguiente comunidad en tamaño la dominan las revistas del Reino Unido donde hay una sola revista metodológica de corte cualitativo y otras temáticas de inmigración, estudios raciales o nacionalismo. Esta división entre Estados Unidos y Reino Unido coincide con el análisis de Zougiris (2018). Además, nos encontramos con otros dos tipos de comunidades, las temáticas y las de lengua no inglesa. Las temáticas abordan temas como la sociología rural, la agricultura, los recursos naturales, las relaciones humanos-animales, el ocio y el deporte. Las de lengua no inglesa, responden a las comunidades nacionales de Vanderstraeten (2010) e incluyen revistas alemanas, francesas, españolas o del este europeo con una conexión fuerte con las revistas de enfoque europeo.

10. CONCLUSIONES

En este artículo se ha caracterizado el conjunto de la disciplina sociológica a partir de las publicaciones en revistas especializadas de impacto en los últimos años. La novedad del trabajo consiste en haber utilizado una fuente de datos de las denominadas *big scholarly data*. En concreto se ha utilizado el Microsoft Academic Graph que cuenta con más de 300 millones de publicaciones. El artículo aporta varias estrategias que permiten seleccionar los datos de interés entre los millones de publicaciones y reducir su dimensionalidad para representarlos en forma de red. La selección inicial se realiza a través de las revistas conforme al ranking de revistas JCR de sociología. Tras esto se vuelve a realizar una selección de revistas y se reúnen en tres grupos para los que se comparan sus citas, coautoría y sesgo de género. Finalmente, se utiliza el análisis de coincidencias para establecer una red de afiliación entre autores y revistas que se proyecta a una red de modo uno formada por solo revistas. Esta red permite localizar agrupaciones de revistas enlazadas por autores comunes.

Los resultados corroboran análisis de otros estudios presentando la sociología como una disciplina dominada por revistas generalistas anglosajonas. Es una disciplina que abarca una amplia variedad de temáticas y con enfoques metodológicos diversos que dependen del ámbito geográfico en el que se desarrollan. Unas pocas revistas dominan las citas, mientras que los artículos de revistas metodológicas tienen un grado mayor de coautoría y las revistas temáticas un menor sesgo de género.

Estos resultados pueden verse condicionados por el sesgo anglosajón en la cobertura de las revistas JCR y por la utilización únicamente de artículos de revistas y no de libros o comunicaciones en congresos.

Una posible investigación para estudios futuros consistiría en clasificar temática y metodológicamente los artículos de las revistas seleccionadas y poder evaluar cuestiones tales como la relación entre el número de citas y las distintas temáticas o metodologías, la evolución de las tendencias en los últimos años en relación con la distribución geográfica de las revistas o la correspondencia entre el género, las materias y los métodos que se utilizan.

11. BIBLIOGRAFÍA

AMATURO, E. y PUNZIANO, G. (2017): "Blurry Boundaries: Internet, Big-New Data, and Mixed-Method Approach", en Data Science and Social Research, Cham, Springer International Publishing, pp. 35-55

BLOK, A., CARLSEN, H. B., JØRGENSEN, T. B., MADSEN, M. M., RALUND, S. y PEDERSEN, M. A. (2017): "Stitching Together the Heterogeneous Party: A Complementary Social Data Science Experiment", Big Data & Society, 4(2), DOI: 10.1177/2053951717736337

BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R. y LEFEBVRE, E. (2008): "Fast Unfolding of Communities in Large Networks", Journal of Statistical Mechanics: Theory and Experiment, 2008(10), DOI: 10.1088/1742-5468/2008/10/p10008

BOURDIEU, P. (2004): Science of Science and Reflexivity, Cambridge, Polity Press

BOYD, D. y CRAWFORD, K. (2012): "Critical Questions for Big Data", Information, communication & society, 15(5), pp. 662-79. DOI: 10.1080/1369118X.2012.678878

BURROWS, R. y SAVAGE, M. (2014): "After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology", Big Data & Society, 1(1), DOI: 10.1177/2053951714540280

CARROLL, S. (2009): "Defining the Scientific Method", Nature Methods, 6(4), pp. 237-37. DOI: 10.1038/nmeth0409-237

CASTELLS, M. (1998): The Rise of the Network Society, Oxford, Blackwell Publishers

CLARIVATE ANALYTICS (2020): Journal Citation Reports, disponible en https://jcr.clarivate.com/ [consulta: Agosto 2019]

CLEMENS, E. S., POWELL, W. W., MCILWAINE, K. y OKAMOTO, D. (1995): "Careers in Print: Books, Journals, and Scholarly Reputations", American Journal of Sociology, 101(2), pp. 433-94. DOI: 10.1086/230730

CLOGG, C. C. (1992): "The Impact of Sociological Methodology on Statistical Methodology", Statistical Science, 7(2), pp. 183-96.

DE HAAN, J. (1997): "Authorship Patterns in Dutch Sociology", Scientometrics, 39(2), pp. 197-208. DOI: 10.1007/BF02457448

DURKHEIM, E. (1982): "Sociology and the Social Sciences (1903)", en The Rules of Sociological Method: And Selected Texts on Sociology and Its Method, London, Macmillan Education UK, pp. 175-208

ESCOBAR, M. (2015): "Studying Coincidences with Network Analysis and Other Multivariate Tools", Stata Journal, 15(4), pp. 1118-56.

ESCOBAR, M. y TEJERO, C. (2018): "El Análisis Reticular De Coincidencias", Empiria. Revista de metodología de ciencias sociales, 39(2018) DOI:10.5944/empiria.39.2018.20879

ESCOBAR, M. y MARTINEZ-URIBE, L. (2020): "Network Coincidence Analysis: The Netcoin R Package", Journal of Statistical Software, 93(11), pp. 1-32, DOI: 10.18637/jss.v093.i11

ESPELAND, W. N. y STEVENS, M. L. (2008): "A Sociology of Quantification", European Journal of Sociology, 49(3), pp. 401-36. DOI: 10.1017/S0003975609000150

EVANS, J. A. y ACEVES, P. (2016): "Machine Translation: Mining Text for Social Theory", Annual Review of Sociology, 42(1), pp. 21-50. DOI: 10.1146/annurev-soc-081715-074206

FRANK, M. R., WANG, D., CEBRIAN, M. y RAHWAN, I. (2019): "The Evolution of Citation Graphs in Artificial Intelligence Research", Nature Machine Intelligence, 1(2), pp. 79-85. DOI: 10.1038/s42256-019-0024-5

FRUCHTERMAN, T. M. J. y REINGOLD, E. M. (1991): "Graph Drawing by Force-Directed Placement", Software: Practice and Experience, 21(11), pp. 1129-64. DOI: 10.1002/spe.4380211102

GANTMAN, E. R. y DABÓS, M. P. (2018): "Research Output and Impact of the Fields of Management, Economics, and Sociology in Spain and France: An Analysis Using Google Scholar and Scopus", Journal of the Association for Information Science and Technology, 69(8), pp. 1054-66. DOI: 10.1002/asi.24020

GIARDULLO, P. (2016): "Does 'Bigger' Mean 'Better'? Pitfalls and Shortcuts Associated with Big Data for Social Research", Quality & Quantity, 50(2), pp. 529-47. DOI: 10.1007/s11135-015-0162-8
GRANT, L. y WARD, K. B. (1991): "Gender and Publishing in Sociology", Gender and Society, 5(2), pp. 207-223
GRIMMER, J. (2015): "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together", Political Science & Politics, 48(1), pp. 80-83. DOI: 10.1017/S1049096514001784
GUPTA, B. y BHATTACHARYA, S. (2004): "Bibliometric Approach Towards Mapping the Dynamics of Science and Technology", DESIDOC Journal of Library & Information Technology, 24(1)
HABERMAN, S.J. (1973): "The Analysis of Residuals in Cross-Classified Tables", Biometrics, 29(1), pp. 1-25. DOI: 10.18637/jss.v014.i15
HALFORD, S. y SAVAGE, M. (2017): "Speaking Sociologically with Big Data: Symphonic Social Science and the Future for Big Data Research", Sociology, 51(6), pp. 1132-48. DOI: 10.1177/0038038517698639
HEALY, K. y MOODY, J. (2014): "Data Visualization in Sociology", Annual Review of Sociology, 40(1), pp. 105-28. DOI: 10.1146/annurev-soc-071312-145551
JACOBS, J. A. (2016): "Journal Rankings in Sociology: Using the H Index with Google Scholar", The American Sociologist, 47(2), pp. 192-224. DOI: 10.1007/s12108-015-9292-7
KHUN, T. S. (1962): The Structure of Scientific Revolutions, United States, The University of Chicago Press
KOROM, P. (2019): "A Bibliometric Visualization of the Economics and Sociology of Wealth Inequality: A World Apart?", Scientometrics, 118(3), pp. 849-68. DOI: 10.1007/s11192-018-03000-z
KOROM, P. (2020): "The Prestige Elite in Sociology: Toward a Collective Biography of the Most Cited Scholars (1970-2010)", The Sociological Quarterly, 61(1), pp. 128-63. DOI: 10.1080/00380253.2019.1581037
LATOUR, B. y WOOLGAR, S. (1987): Laboratory life, New Jersey, Princeton University Press
LAZER, D., PENTLAND, A., ADAMIC, L., ARAL, S., BARABÁSI, A.-L., BREWER, D., CHRISTAKIS, N., CONTRACTOR, N., FOWLER, J., GUTMANN, M., JEBARA, T., KING, G., MACY, M., ROY, D. y VAN ALSTYNE, M. (2009): "Computational Social Science", Science, 323(5915), pp. 721. DOI: 10.1126/science.1167742
MANOVICH, L. (2015): "Data Science and Digital Art History", International Journal for Digital Art History, 0(1), DOI: 10.11588/dah.2015.1.21631
MARTINHO, D. T. (2018): "Researching Culture through Big Data: Computational Engineering and the Human and Social Sciences", Social Sciences, 7(12), DOI: 10.3390/socsci7120264
MCFARLAND, D. A. y MCFARLAND, H. R. (2015): "Big Data and the Danger of Being Precisely Inaccurate", Big Data & Society, 2(2), DOI: 10.1177/2053951715602495
MCFARLAND, D. A., LEWIS, K. y GOLDBERG, A. (2016): "Sociology in the Era of Big Data: The Ascent of Forensic Social Science", The American Sociologist, 47(1), pp. 12-35. DOI: 10.1007/s12108-015-9291-8
MILLS, C. W. (1959): The Sociological Imagination, New York, Oxford University Press

MOED, H. F. (2005): "Citation Analysis of Scientific Journals and Journal Impact Measures", Current Science, 89(12), pp. 1990-96.
MOODY, J. (2004): "The Structure of Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999", American Sociological Review, 69(2), pp. 213-238. DOI: 10.1177/000312240406900204
MOODY, J. y LIGHT, R. (2006): "A view from Above: The Evolving Sociological Landscape", The American Sociologist, 37(2), pp. 67-86
MOKSONY, F., HEGED□S, R. y CSÁSZÁR, M. (2014): "Rankings, Research Styles, and Publication Cultures: A Study of American Sociology Departments", Scientometrics, 101(3), pp. 1715-29. DOI: 10.1007/s11192-013-1218-y
MORETTI, F. (2000): "Conjectures on World Literature", New left review, pp. 54-68.
NERESINI, F. (2017): "On Data, Big Data and Social Research. Is It a Real Revolution?" en On Data, Big Data and Social Research. Is It a Real Revolution?, Cham, Springer International Publishing, pp. 9-16
O'REILLY, K. (2009): "For Interdisciplinarity and a Disciplined, Professional Sociology", Innovation: The European Journal of Social Science Research, 22(2), pp. 219-32. DOI: 10.1080/13511610903075761
OROMANER, M. (1981): "Cognitive Consensus in Recent Mainstream American Sociology: An Empirical Analysis", Scientometrics, 3(2), pp. 73-84. DOI: 10.1007/BF02025631
PERITZ, B. C. (1983): "Are Methodological Papers More Cited Than Theoretical or Empirical Ones? The Case of Sociology", Scientometrics, 5(4), pp. 211-18. DOI: 10.1007/BF02019738
PHELAN, T. J. (2000): "Bibliometrics and the Evaluation of Australian Sociology", Journal of Sociology, 36(3), pp. 345-63. DOI: 10.1177/144078330003600305
PINCH, T. J. y BIJKER, W. E. (1984): "The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other", Social Studies of Science, 14(3), pp. 399-441. DOI: 10.1177/030631284014003004
PRESS, G. (2014):"12 Big Data Definitions: What´s Yours?" Forbes. Disponible en la página web: https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/ [consulta: Septiembre 2021]
RAFTERY, A. E. (2000): "Statistics in Sociology, 1950–2000", Journal of the American Statistical Association, 95(450), pp. 654-61. DOI: 10.1080/01621459.2000.10474245
RIVIERA, E. (2015): "Testing the Strength of the Normative Approach in Citation Theory through Relational Bibliometrics: The Case of Italian Sociology", Journal of the Association for Information Science and Technology, 66(6), pp. 1178-88. DOI: 10.1002/asi.23248
RODRIGUEZ-YUNTA, L. (2009): "Revistas españolas en WoS", Anuario ThinkEPI, 2010(4), pp. 250-253 web: https://recyt.fecyt.es/index.php/ThinkEPI/article/view/31268 [consulta: Septiembre 2021]
SANTAMARÍA, L. y MIHALJEVI□, H. (2018): "Comparison and Benchmark of Name-to-Gender Inference Services", PeerJ Computer Science, 4, pp. e156.
SAVAGE, M. y BURROWS, R. (2007): "The Coming Crisis of Empirical Sociology", Sociology, 41(5), pp. 885-99. DOI: 10.1177/0038038507080443
SCHWEMMER, C. y WIECZOREK, O. (2019): "The Methodological Divide of Sociology: Evidence from Two Decades of Journal Publications", Sociology, 54(1), pp. 3-21. DOI: 10.1177/0038038519853146

SEALE, C. (2008): "Mapping the Field of Medical Sociology: A Comparative Analysis of Journals", Sociology of health & illness, 30(5), pp. 677-95. DOI: 10.1111/j.1467-9566.2008.01090.x

SINHA, A., SHEN, Z., SONG, Y., MA, H., EIDE, D., HSU, B.-J. y WANG, K. (2015): "An Overview of Microsoft Academic Service (Mas) and Applications." en An Overview of Microsoft Academic Service (Mas) and Applications, pp. 243-46

SMALL, A. W. (1906): "The Relation between Sociology and Other Sciences", American Journal of Sociology, 12(1), pp. 11-31.

SMELSER, N. J. (2014): "The Optimum Scope of Sociology (1969)." en Getting Sociology Right, University of California Press, pp. 15-34

STICHWEH, R. (2008): "The Sociology of Scientific Disciplines: On the Genesis and Stability of the Disciplinary Structure of Modern Science", Science in Context, 5(1), pp. 3-15. DOI: 10.1017/S0269889700001071

STINCHCOMBE, A. L. (1984): "The Origins of Sociology as a Discipline", Acta Sociologica, 27(1), pp. 51-61. DOI: 10.1177/000169938402700104

SU, H.-N. y LEE, P.-C. (2010): "Mapping Knowledge Structure by Keyword Co-Occurrence: A First Look at Journal Papers in Technology Foresight", Scientometrics, 85(1), pp. 65-79. DOI: 10.1007/s11192-010-0259-8

TAYLOR & FRANCIS (2017): Co-Authorship in the Humanities and Social Sciences, disponible en página web https://authorservices.taylorandfrancis.com/wp-content/uploads/2017/09/Coauthorship-white-paper.pdf [consulta: Agosto 2019]

TINATI, R., HALFORD, S., CARR, L. y POPE, C. (2014): "Big Data: Methodological Challenges and Approaches for Sociological Analysis", Sociology, 48(4), pp. 663-81. DOI: 10.1177/0038038513511561

TUBARO, P. (2014): "Sociology and Social Networks." en Sociology and Social Networks, SAGE Publications Sage UK: London, England, pp.

VANDERSTRAETEN, R. (2010): "Scientific Communication: Sociology Journals and Publication Practices", Sociology, 44(3), pp. 559-76. DOI: 10.1177/0038038510362477

WANG, K., SHEN, Z., HUANG, C., WU, C.-H., EIDE, D., DONG, Y., QIAN, J., KANAKIA, A., CHEN, A. y ROGAHN, R. (2019): "A Review of Microsoft Academic Services for Science of Science Studies", Frontiers in Big Data, 2, pp. 45.

XIA, F., WANG, W., BEKELE, T. M. y LIU, H. (2017): "Big Scholarly Data: A Survey", IEEE Transactions on Big Data, 3(1), pp. 18-35. DOI: 10.1109/TBDATA.2016.2641460

ZOUGIRIS, K. (2018):"Detecting Topical Divides and Topical Bridges Across National Sociologies". The American Sociologist, 50, pp. 63-84. DOI: 10.1007/s12108-018-9392-2

5. Conclusiones

Esta sección presenta las conclusiones de la tesis doctoral, sus limitaciones e implicaciones para investigaciones futuras. La tesis, compuesta de tres artículos, aspira a realizar una contribución metodológica y se enmarca en la rama de la sociología computacional.

Tras la aparición de una amplia variedad de fuentes de datos con el fenómeno del *big data*, se están desarrollando multitud de metodologías de análisis y tratamiento de grandes cantidades de información que combinan métodos tradicionales con novedosas metodologías computacionales. Inicialmente fueron ingenieros, matemáticos y tecnólogos los que comenzaron a analizar estas fuentes para estudios sociológicos en el ámbito privado gracias a sus capacidades técnicas. Sin embargo,las nueves fuentes *big data* traen consigo retos metodológicos importantes y bien conocidos por los sociólogos. Es pues fundamental su participación en los estudios sociológicos que exploten estas nuevas fuentes de datos para interpretar debidamente la estructura social. Así, uno de los retos mencionados recurrentemente en la literatura, que aborda la aparición de la sociología computacional, es la necesidad de formar a los sociólogos computacionales. Es crucial para la disciplina que los sociólogos estén capacitados en las habilidades necesarias para trabajar con *big data* y tengan a su alcance tecnologías, infraestructura, metodologías y estrategias para acceder, manipular y analizar grandes cantidades de información.

El análisis de coincidencias aporta un método de detección de patrones, aplicable a grandes y pequeños conjuntos de datos, para así poder comprenderlos mejor

Con el objetivo de contribuir a las metodologías y estrategias para el tratamiento de estas nuevas fuentes de datos se presenta la herramienta netCoin en el primer artículo de este trabajo. Se trata de una herramienta desarrollada en el lenguaje de programación R que implementa el análisis de coincidencias. Este análisis aporta una estrategia para la búsqueda de patrones en los datos multivariables arrojando información sobre las relaciones intrínsecas entre las distintas dimensiones presentes. Lo hace a través de un marco estadístico que aplica una variedad de medidas de distancia para establecer la probabilidad de que dos sucesos aparezcan de manera conjunta en los datos de los que se dispone. De esta forma, el análisis de coincidencias aporta una valiosa estrategia de análisis a sociólogos que la pueden aplicar tanto a los datos de encuesta tradicionales como a aquellas nuevas fuentes de datos. En los artículos presentados en los capítulos 3 y 4 se demuestra cómo esta capacidad de detectar patrones en grandes cantidades de datos se convierte en un elemento fundamental para extraer conocimiento de datos brutos.

El uso de análisis de redes interactivas permite realizar análisis exploratorios de grandes cantidades de información y la presentación de resultados finales

La herramienta netCoin además incorpora la capacidad de generar gráficos de redes para representar las coincidencias entre los sucesos seleccionados. Estas redes se presentan a través de un interfaz interactivo en el que es posible hacer filtros precisos o customizar la forma, el color y el tamaño de los nodos y las aristas. Estas redes interactivas cuyo funcionamiento es explicado en el primer artículo han sido útiles en el segundo y tercer artículo para la realización de análisis exploratorios iniciales y para la presentación de resultados finales. Para la preparación de los artículos de los capítulos 3 y 4 ha sido necesario realizar multitud de exploraciones de los datos para entender mejor su contenido y estructura, así como para identificar posibles análisis más profundos a realizar. Por otro lado, en ambos artículos se han preparado varias redes que han sido presentadas para poder explicar los resultados obtenidos. Así, el estudio aporta una estrategia que permite a los sociólogos computacionales el uso del análisis de redes tanto la exploración de los datos con los que estén trabajando como la presentación de resultados en sus trabajos de investigación.

Estrategias como el cruce con datos externos, la representación de diversas dimensiones en redes con nodos diferenciados y la proyección de redes bimodales posibilitan reducir la dimensionalidad de los datos para una gestión y análisis más sencillos y simples

Igualmente, al trabajar con grandes cantidades de información es importante contar con metodologías que permitan reducir la dimensionalidad y filtrar la información de interés. En los artículos de los capítulos 3 y 4 hemos mostrado la utilidad de cruzar los datos principales con los que trabajamos con otras fuentes de datos externas permitiéndonos hacer selecciones que filtran la información de las fuentes de *big data.* No solo eso, el análisis de coincidencias proporciona varias estrategias de reducción de dimensionalidad convenientes. Una de ellas es la posibilidad de representar los resultados del análisis de coincidencias como redes en las que los nodos dispuestos con colores y formas diversos representen variables o categorías. Esta estrategia permite mostrar relaciones entre las distintas dimensiones de datos multivariantes. Cabe no olvidar que el análisis de coincidencias parte de una matriz de eventos que suceden en un número finito de espacios. Esta matriz representa una red bimodal que al proyectarse a una red de modo uno representa solamente las relaciones en los datos en una de las dimensiones y así reduce la complejidad de la información. Estas técnicas posibilitan partir de fuentes de *big data* y transformarlas en datos más fáciles de manejar y analizar.

Los datos bibliográficos y de bibliotecas proporcionan una fuente de calidad y de fácil acceso para los sociólogos computacionales

Al revisar las fuentes de datos utilizadas en publicaciones en el campo de la sociología computacional hemos detectado que las fuentes habituales incluyen los medios sociales, algunas bases de datos bibliográficas de pago, experimentos online, datos de telefonía y dispositivos móviles, mensajería instantánea, servicios de Internet o de fuentes de información mantenidas por la comunidad como Wikipedia. La mayoría de estos datos son propiedad de empresas privadas y su acceso presenta dificultades importantes para la mayoría de los investigadores. Asimismo, hay una importante omisión de estudios que empleen datos gestionados por bibliotecas y archivos.

Ante las barreras a nivel de acceso que pueden presentar las fuentes de datos de big *data* al ser de propiedad privada, los datos bibliográficos y de bibliotecas y archivos se alzan como una opción de interés para los sociólogos computacionales. Principalmente, esto es debido a la calidad de los datos gestionados por las bibliotecas, así como por la sencillez para su acceso y utilización. Las bibliotecas y archivos cuentan con una importante tradición de catalogación y enriquecimiento de sus datos utilizando tesauros y vocabularios de normalización que garantizan la calidad de los datos. En los últimos años están además participando de movimientos de ciencia abierta preparando sus colecciones para un acceso abierto y computacional que permita su explotación. De esta forma los datos de bibliotecas, por la heterogeneidad de contenidos y formatos, por su calidad, por ser datos abiertos y accesibles a nivel computacional pueden aportar a los sociólogos computacionales nuevas perspectivas de investigación o complementar las ya existentes. En este trabajo lo hemos mostrado a través de dos ejemplos en los artículos de los capítulos 3 y 4, uno con los datos del *British National Bibliography* de la *British Library* y otro con los datos abiertos del *Microsoft Academic Graph,* como puede fácilmente utilizarse

Es preciso mencionar varias limitaciones de esta tesis. Por un lado, está la selección de los conjuntos de datos utilizados en los ejemplos. Esta selección no cubre todos los tipos importantes de datos de bibliotecas y archivos. En particular, habría sido especialmente interesante haber podido completar los ejemplos presentados con otros referidos a datos de colecciones digitales o archivos. Estos datos son diferentes ya que pueden incluir la información que minuciosamente describen los documentos y los objetos digitales tales como imágenes, grabaciones sonoras o

vídeos. Esta información requiere un tratamiento también diferente para su análisis que seguramente pueda abordarse combinando los análisis tradicionales de contenido con nuevas aproximaciones computacionales.

Tampoco se ha profundizado en exceso en algunos de los elementos más tecnológicos relacionados con la preparación y configuración de las infraestructuras en las que se han podido almacenar y manipular los datos. Quizás es más adecuado que esta información aparezca en revistas de investigación especializadas en dominios más técnicos. Sin embargo, el montaje de estas infraestructuras es inevitable cuando se trabaja con grandes cantidades de información y los sociólogos computacionales tendrán que realizar este tipo de instalaciones irremediablemente. Aun así, no existe una única forma de preparar estas infraestructuras que además dependen de las fuentes de datos que se utilicen.

En cuanto a las implicaciones para investigaciones futuras, existe un área en particular en la que se podría continuar extendiendo las aportaciones realizadas en esta tesis. En los ejemplos presentados a través de los artículos hemos trabajado con grandes redes que han sido imposible de explorar visualmente debido a las limitaciones de capacidad de procesamiento de los equipos utilizados y las capacidades humanas para comprender grandes cantidades de información presentadas gráficamente. Por ello, en nuestros ejemplos hemos utilizado varias técnicas de reducción de dimensionalidad optando por filtrar los nodos y aristas de la red total de una manera determinista para quedarnos con aquellas partes de la red que nos interesaban. En matemáticas y ciencias de la computación existe un corpus de literatura científica dedicada al estudio de las técnicas de agrupamiento y unión de nodos para reducir la densidad de las redes conservando sus propiedades esenciales conocido como g*raph coarsening.* Estas técnicas podrían incorporarse a netCoin permitiendo la representación y exploración de grandes redes y posibilitando el estudio de los procesos sociales a nivel micro y a nivel macro.

6. Referencias

Adams, Jimi and Ryan Light. 2015. "Scientific consensus, the law, and same sex parenting outcomes." *Social Science Research* 53:300-10. doi: 10.1016/j.ssresearch.2015.06.008.

Aragona, Biagio. 2022. "Tipos de big data y análisis sociológico: usos, críticas y problemas éticos." *Empiria. Revista de metodología de ciencias sociales* (53):15-30. doi: 10.5944/empiria.53.2022.32610.

Askin, Noah and Michael Mauskapf. 2017. "What makes popular culture popular? Product features and optimal differentiation in music." *American Sociological Review* 82(5):910-44. doi: 10.1177/0003122417728662.

Ayris, Paul and Tiberius Ignat. 2018. "Defining the role of libraries in the open science landscape: a reflection on current European practice." *Open Information Science* 2(1):1-22. doi: doi:10.1515/opis-2018-0001.

Bail, Christopher A. 2014. "The cultural environment: measuring culture with big data." *Theory and Society* 43(3):465-82. doi: 10.1007/s11186-014-9216-5.

Bail, Christopher A. 2016. "Cultural carrying capacity: Organ donation advocacy, discursive framing, and social media engagement." *Social Science & Medicine* 165:280-88. doi: 10.1016/j.socscimed.2016.01.049.

Bail, Christopher A., Taylor W. Brown and Marcus Mann. 2017. "Channeling hearts and minds: Advocacy organizations, cognitive-emotional currents, and public conversation." *American Sociological Review* 82(6):1188-213. doi: 10.1177/0003122417733673.

Bail, Christopher A., Taylor W. Brown and Andreas Wimmer. 2019. "Prestige, proximity, and prejudice: How Google search terms diffuse across the world." *American Journal of Sociology* 124(5):1496-548. doi: 10.1086/702007.

Battisti, Francesca de and Silvia Salini. 2012. "Bibliographic data: a different analysis perspective." *Electronic Journal of Applied Statistical Analysis* 5(3):353-59. doi: 10.1285/i20705948v5n3p353.

Bernard, H. Russell. 2012. "The science in social science." *Proceedings of the National Academy of Sciences* 109(51):20796-99. doi: 10.1073/pnas.1218054109.

Berners-Lee, Tim. 2009, "Linked data". url: https://www.w3.org/DesignIssues/LinkedData.html. Consultado en marzo de 2022.

Boyd, Danah and Kate Crawford. 2012. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." *Information, communication & society* 15(5):662-79. doi: 10.1080/1369118X.2012.678878.

Breiman, Leo. 2001. "Statistical modeling: The two cultures." *Statistical Science* 16(3):199-231.

Burrows, Roger and Mike Savage. 2014. "After the crisis? Big data and the methodological challenges of empirical sociology." *Big Data & Society* 1(1):1-6. doi: 10.1177/2053951714540280.

Centola, Damon. 2010. "The spread of behavior in an online social network experiment." *Science (New York, N.Y.)* 329(5996):1194-97. doi: 10.1126/science.1185231.

Cioffi-Revilla, Claudio. 2014. "Computation and social science." Pp. 23-66 in *Introduction to computational social science: Principles and applications*, edited by C. Cioffi-Revilla. London: Springer London.

Deliot, Corine. 2014. "Publishing the British National Bibliography as linked open data." *Catalogue & Index* 174:13-18.

Donoho, David. 2017. "50 years of data science." *Journal of Computational and Graphical Statistics* 26(4):745-66. doi: 10.1080/10618600.2017.1384734.

Driscoll, Kevin. 2012. "From punched cards to "big data": A social history of database populism." *Communicatio* 1:1-33.

Eagle, Nathan, Alex Pentland and David Lazer. 2009. "Inferring friendship network structure by using mobile phone data." *Proceedings of the National Academy of Sciences* 106(36):15274-78. doi: 10.1073/pnas.0900282106.

Edelman, Benjamin G. and Michael Luca. 2014. "Digital discrimination: The case of Airbnb.com." url: http://dx.doi.org/10.2139/ssrn.2377353. Consultado en marzo de 2022.

Edelmann, Achim, Tom Wolff, Danielle Montagne and Christopher A. Bail. 2020. "Computational social science and sociology." *Annual Review of Sociology* 46(1):61-81. doi: 10.1146/annurev-soc-121919-054621.

Escobar, Modesto. 2015. "Studying coincidences with network analysis and other multivariate tools." *The Stata Journal* 15(4):1118-56. doi: 10.1177/1536867X1501500410.

Espeland, Wendy Nelson and Mitchell L. Stevens. 2008. "A sociology of quantification." *European Journal of Sociology* 49(3):401-36. doi: 10.1017/S0003975609000150.

Evans, James and Jacob G. Foster. 2019. "Computation and the sociological imagination." *Contexts* 18(4):10-15. doi: 10.1177/1536504219883850.

Evans, James A and Jacob G Foster. 2011. "Metaknowledge." *Science (New York, N.Y.)* 331(6018):721-25.

Evans, James A. and Pedro Aceves. 2016. "Machine translation: mining text for social theory." *Annual Review of Sociology* 42(1):21-50. doi: 10.1146/annurev-soc-081715-074206.

Farber, Michael, Mike Cameron, Christopher Ellis and Josh Sullivan. 2013. "Massive data analytics and the cloud." url: https://web.archive.org/web/20150616043728/http://www.boozallen.com:80/media/file/MassiveData.pdf. Consultado en marzo 2022.

Fisher, R. A. 1925. *Statistical methods for research workers*. London: Oliver and Boyd.

Flores, René D. 2017. "Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using Twitter data." *American Journal of Sociology* 123(2):333-84. doi: 10.1086/692983.

García Ferrando, Manuel. 1980. *Sobre el método. Problemas de investigación empírica en sociología*, Edited by C. d. I. Sociológicas. Madrid: Centro de Investigaciones Sociológicas.

Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Lieberman Aiden Erez and Li Fei-Fei. 2017. "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States." *Proceedings of the National Academy of Sciences* 114(50):13108-13. doi: 10.1073/pnas.1700035114.

Golder Scott, A. and W. Macy Michael. 2011. "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures." *Science (New York, N.Y.)* 333(6051):1878-81. doi: 10.1126/science.1202775.

González-Alcaide, Gregorio. 2021. "Bibliometric studies outside the information science and library science field: uncontainable or uncontrollable?". *Scientometrics* 126(8):6837-70. doi: 10.1007/s11192-021-04061-3.

Gualda, Estrella. 2022a. "Social big data, sociología y ciencias sociales computacionales." *Empiria. Revista de metodología de ciencias sociales* (53):147-77. doi: 10.5944/empiria.53.2022.32631.

Gualda, Estrella. 2022b. "Altruism, solidarity and responsibility from a committed sociology: contributions to society." *The American Sociologist* 53(1):29-43. doi: 10.1007/s12108-021-09504-1.

Halford, Susan and Mike Savage. 2017. "Speaking sociologically with big data: symphonic social science and the future for big data research." *Sociology* 51(6):1132-48. doi: 10.1177/0038038517698639.

Hallo, María, Sergio Luján-Mora, Alejandro Maté and Juan Trujillo. 2016. "Current state of linked data in digital libraries." *Journal of Information Science* 42(2):117-27. doi: 10.1177/0165551515594729.

Healy, Kieran and James Moody. 2014. "Data visualization in sociology." *Annual Review of Sociology* 40(1):105-28. doi: doi:10.1146/annurev-soc-071312-145551.

Helbing, Dirk, Illés Farkas and Tamás Vicsek. 2000. "Simulating dynamical features of escape panic." *Nature* 407(6803):487-90. doi: 10.1038/35035023.

Hoffman, Mark Anthony. 2019. "The materiality of ideology: cultural consumption and political thought after the american revolution." *American Journal of Sociology* 125(1):1-62. doi: 10.1086/704370.

Horowitz, Irving Louis. 2006. "Big five and little five: Measuring revolutions in social science." *Society* 43(3):9-12. doi: 10.1007/BF02687589.

Hrynaszkiewicz, Iain, Natasha Simons, Azhar Hussain, Rebecca Grant and Simon Goudie. 2020. "Developing a research data policy framework for all journals and publishers." *Data Science Journal* 19(1).

IFLA. 2020. "IFLA statement on libraries and artificial intelligence." url: https://repository.ifla.org/handle/123456789/1646. Consultado en marzo de 2022.

Kaur, Ranjit, Parveen Kumar and Raminder Pal Singh. 2014. "A Journey of digital storage from punch cards to cloud." *IOSR Journal of Engineering* 4(3):36-41.

Khun, Thomas S. 1962. *The structure of scientific revolutions*. United States: The University of Chicago Press.

King, Gary and Nathaniel Persily. 2019. "A new model for industry-academic partnerships." *PS: Political Science and Politics* 53:703-09.

King, Molly M., Carl T. Bergstrom, Shelley J. Correll, Jennifer Jacquet and Jevin D. West. 2017. "Men set their own cites high: gender and self-citation across fields and over time." *Socius* 3:2378023117738903. doi: 10.1177/2378023117738903.

Kitchin, Rob. 2014a. *The data revolution: Big data, open data, data infrastructures and their consequences*: Sage.

Kitchin, Rob. 2014b. "Big data, new epistemologies and paradigm shifts." *Big Data & Society* 1(1). doi: 10.1177/2053951714528481.

Lauriault, Tracey P., Barbara Lazenby Craig, D. R. Fraser Taylor and Peter L. Pulsifer. 1969. "Today's data are part of tomorrow's research: Archival issues in the sciences." *Archivaria* 64:123-79.

Lazer, David and Jason Radford. 2017. "Data ex machina: Introduction to big data." *Annual Review of Sociology* 43(1):19-39. doi: 10.1146/annurev-soc-060116-053457.

Leahey, Erin and James Moody. 2014. "Sociological innovation through subfield integration." *Social Currents* 1(3):228-56. doi: 10.1177/2329496514540131.

Lewis, Kevin, Kurt Gray and Jens Meierhenrich. 2014. "The structure of online activism." *Sociological Science* 1(1):1-9. doi: 10.15195/v1.a1.

MacDonald, Stuart and Luis Martínez-Uribe. 2008. "Libraries in the converging worlds of open data, e-research, and Web 2.0."

Manovich, Lev. 2015. "Data science and digital art history." *International Journal for Digital Art History* 0(1). doi: 10.11588/dah.2015.1.21631.

Martinez-Uribe, Luis and Paz Fernandez. 2015. "Data services: a strategic function of 21st century libraries." *El profesional de la información* 24(2):193-99. doi: 10.3145/epi.2015.mar.13.

Martinho, D. Teresa. 2018. "Researching culture through big data: Computational engineering and the human and social sciences." *Social Sciences* 7(12). doi: 10.3390/socsci7120264.

McFarland, Daniel A, Kevin Lewis and Amir Goldberg. 2016. "Sociology in the era of big data: The ascent of forensic social science." *The American Sociologist* 47(1):12-35. doi: 10.1007/s12108-015-9291-8.

McFarland, Daniel A. and H. Richard McFarland. 2015. "Big data and the danger of being precisely inaccurate." *Big Data & Society* 2(2):2053951715602495. doi: 10.1177/2053951715602495.

Mills, C. Wright. 1959. *The sociological imagination*. New York: Oxford University Press.

Molina, Mario and Filiz Garip. 2019. "Machine learning for sociology." *Annual Review of Sociology* 45(1):27-45. doi: 10.1146/annurev-soc-073117-041106.

Moore, Niamh, Andrea Salter, Liz Stanley and Maria Tamboukou. 2016. *The archive project: Archival research in the social sciences*. London: London: Taylor & Francis Group.

Murray-Rust, Peter. 2008. "Open data in science." *Nature Precedings*. doi: 10.1038/npre.2008.1526.1.

Open Knowledge Foundation. 2022. "Open data handbook." url: https://opendatahandbook.org/. Consultado en marzo de 2022.

Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke and Stewart Varner. 2019. "Always already computational: Collections as data: Final report." url: https://digitalcommons.unl.edu/scholcom/181. Consultado en marzo de 2022.

Park, Jaram, Young Min Baek and Meeyoung Cha. 2014. "Cross-Cultural Comparison of Nonverbal Cues in Emoticons on Twitter: Evidence from Big Data Analysis." *Journal of Communication* 64(2):333-54. doi: 10.1111/jcom.12086.

Potârcă, Gina and Melinda Mills. 2015. "Racial preferences in online dating across European countries." *European Sociological Review* 31(3):326-41. doi: 10.1093/esr/jcu093.

Press, G. 2014. "12 Big data definitions: What´s yours?". *Forbes*. url:https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours. Consultado en marzo de 2022.

Raftery, Adrian E. 2000. "Statistics in sociology, 1950–2000." *Journal of the American Statistical Association* 95(450):654-61. doi: 10.1080/01621459.2000.10474245.

Rzhetsky, Andrey, G. Foster Jacob, T. Foster Ian and A. Evans James. 2015. "Choosing experiments to accelerate collective discovery." *Proceedings of the National Academy of Sciences* 112(47):14569-74. doi: 10.1073/pnas.1509757112.

Saavedra, Serguei, Jordi Duch and Brian Uzzi. 2011. "Tracking traders' understanding of the market using e-communication data." *PLOS ONE* 6(10):e26705. doi: 10.1371/journal.pone.0026705.

Savage, Mike and Roger Burrows. 2007. "The coming crisis of empirical sociology." *Sociology* 41(5):885-99. doi: 10.1177/0038038507080443.

Shirado, Hirokazu and Nicholas A. Christakis. 2017. "Locally noisy autonomous agents improve global human coordination in network experiments." *Nature* 545(7654):370-74. doi: 10.1038/nature22332.

Shirado, Hirokazu, George Iosifidis, Leandros Tassiulas and Nicholas A. Christakis. 2019. "Resource sharing in technologically defined social networks." *Nature Communications* 10(1):1079. doi: 10.1038/s41467-019-08935-2.

Shor, Eran, Arnout van de Rijt, Alex Miltsov, Vivek Kulkarni and Steven Skiena. 2015. "A paper ceiling: Explaining the persistent underrepresentation of women in printed news." *American Sociological Review* 80(5):960-84. doi: 10.1177/0003122415596999.

Su, Hsin-Ning and Pei-Chun Lee. 2010. "Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight." *Scientometrics* 85(1):65-79. doi: 10.1007/s11192-010-0259-8.

Tinati, Ramine, Susan Halford, Leslie Carr and Catherine Pope. 2014. "Big data: methodological challenges and approaches for sociological analysis." *Sociology* 48(4):663-81. doi: 10.1177/0038038513511561.

Toole, Jameson L, Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C González and David Lazer. 2015. "Tracking employment shocks using mobile phone data." *Journal of The Royal Society Interface* 12(107):20150185.

Tubaro, Paola. 2014. "Sociology and social networks." *Sociology* 48(2):410-16. doi: 10.1177/0038038513517319.

Tufekci, Zeynep and Christopher Wilson. 2012. "Social media and the decision to participate in political protest: Observations from Tahrir Square." *Journal of Communication* 62(2):363-79. doi: 10.1111/j.1460-2466.2012.01629.x.

Tukey, John W. 1962. "The future of data analysis." *The Annals of Mathematical Statistics* 33(1):1-67.

Tukey, John W. 1977. *Exploratory data analysis*. Reading: Addison-Wesley.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer and Ben Jones. 2013. "Atypical combinations and scientific impact." *Science (New York, N.Y.)* 342(6157):468-72. doi: 10.1126/science.1240474.

Vanderstraeten, Raf. 2010. "Scientific communication: Sociology journals and publication practices." *Sociology* 44(3):559-76. doi: 10.1177/0038038510362477.

Vasi, Ion Bogdan, Edward T. Walker, John S. Johnson and Hui Fen Tan. 2015. ""No fracking way!" Documentary film, discursive opportunity, and local opposition against hydraulic fracturing in the United States, 2010 to 2013." *American Sociological Review* 80(5):934-59. doi: 10.1177/0003122415598534.

Wagner, Claudia, Eduardo Graells-Garrido, David Garcia and Filippo Menczer. 2016. "Women through the glass ceiling: gender asymmetries in Wikipedia." *EPJ Data Science* 5(1):5. doi: 10.1140/epjds/s13688-016-0066-4.

West, Jevin D, Jennifer Jacquet, Molly M King, Shelley J Correll and Carl T Bergstrom. 2013. "The role of gender in scholarly authorship." *PLOS ONE* 8(7):e66212.

White House. 2009. *Open government directive.* url: https://web.archive.org/web/20100113232617/https://www.whitehouse.gov/sites/default/files/microsites/ogi-directive.pdf. Consultado en marzo de 2022.

Wuchty, Stefan, F. Jones Benjamin and Brian Uzzi. 2007. "The increasing dominance of teams in production of knowledge." *Science (New York, N.Y.)* 316(5827):1036-39. doi: 10.1126/science.1136099.