Posibilidades del uso de las TIC's para el estudio histórico y su difusión

Carlos G. Figuerola, Tamar Groves, Francisco J. Rodríguez Jiménez

Universidad de Salamanca

ÍNDICE

- Retos del historiador
- Digital Humanities y la historia
- El caso de la prensa digital y digitalizada
- Ejemplos de herramientas digitales:
  - Web Crawler
  - Automatic retrivel
  - OPEN CALAIS
  - Clustering
  - Gephi
- Resultados y conclusiones: Conocimiento histórico y difusión de resultados a través de las TICs

RETOS DEL HISTORIADOR

- "La investigación ha sido laboriosa porque los testigos no han dado las mismas versiones de los mismos hechos, sino según las simpatías por unos o por otros, o según la memoria de cada uno." (Tucídides, La Guerra del Peloponeso)
- "A memorandum is written not to inform the reader but to protect the writer." (Dean Acheson)

- PALABRA DE INFORMÁTICO. . . .PALABRA DE HISTORIADOR
  - "La gente quiere creer verdades, pero a la vez quiere creer aquello que confirma sus presupuestos previos y por eso busca los temas o los libros o las novelas que le ratifican su COSMOVISIÓN" (Pérez Garzón, 2012: 256)

RETOS DEL HISTORIADOR

sobre todo del contemporaneísta

- LA DISTANCIA
  - OCUPACIÓN ROMANA península ibérica // GUERRA CIVIL
  - EVITAR "Tentación de servirse de una información 'orientada' o 'selectiva', acorde con sus pre-juicios o preferencias" (VIÑAS, 2013, xiii-xv)
- OBJETIVIDAD
  - El historiador no es insensible al espíritu de su época
  - Zeitgeist (Historiografía alemana)
  - Contexto (BLOCH) Men resamble his...
    $\rightarrow$ ESFUERZO consciente de búsqueda de la objetividad...

RETOS DEL HISTORIADOR

- PRETENSIÓN CIENTÍFICA
  - No mera ELUCUBRACIONES, No meros ensayos apriorísticos o superficiales
- SOBREINFORMACIÓN
  - Ingente masa de información disponible
  - EL ABANICO DE PRUEBAS NO ES ESTÁTICO
  - APARECEN NUEVAS FUENTES / OTRAS PIERDEN RELEVANCIA
  - ¿Cómo separar el grano de la paja? ¿CÓMO ORGANIZAR TAL CANTIDAD?

El mundo de "digital humanities"

- Creando y conservando registros
- Estándares globales de registros de data
- Grandes corpora
- La representación digital de artefactos y la digitalización de textos
- Difusión digital
- Lo digital como un objeto de investigación

Humanidades y lo digital

- El examen de los testimonios de lo que nos define como seres humanos.
- Estrategias interpretativas
- Herramientas que facilitan la exploración de los artefactos / textos

La esencia de los "digital humanities"

- Un estudio critico que incluye la aplicación de algoritmos que facilitan la búsqueda, la recuperación y el analisis de la información.
- La articulación de las preguntas emana del trabajo de reflexión humanística.

El historiador y lo digital

- "Almost all important questions are important precisely because they are not susceptible to quantitative answers" (Swierenga 1970: 33).

La historia y los digital humanities

- Un análisis apoyado por el ordenador de una cantidad enorme de materiales proporciona oportunidades de percibir y analizar patrones, conjunciones, conexiones y ausencias que un ser humano, sin este apoyo, no tendría mucha posibilidad de encontrar


america
analisis
aplicadas
archivos
base
com
consultado
datos
digital
equipo
espana
formacion
fuentes
genero
herramientas
historia
http
humanities
internacional
investigacion
investigadora
investigadores
latina
madrid
mayo
mesa
metodologia
metodologica
miembros
mujeres
nuevas
oral
orales
org
participacion
principal
proyecto
realizado
salamanca
seminario
septiembre
siglo
symposium
tema
transversalidad
universidad
vaciado
www
xx
xxi

Las opciones practicas

- Programas gratis o por pago legos:
  `http://www-958.ibm.com/software/analytics/manyeyes/`
- Programas gratis o por pago que requieren conocimiento básico de informática
  `http://mallet.cs.umass.edu/`
- Programar o adaptar programas

El método de trabajo

- El proceso de desarrollar, aplicar y computar conceptos relacionados con el conocimiento obliga un proceso riguroso de reflexión y definición.
- El dialogo entre el historiador y el informático
- La experiencia de la investigación de cultura científica

Prensa digital y digitalizada

- Productos que nacen en papel
- Productos que nacen digitales

La prensa: el uso de versiones digitales y digitalizadas

Ventajas	Inconvenientes
Fáciles de procesar automáticamente	It's a dynamic source that changes continuously
Permiten manejar una gran cantidad de data	Due to the structure of the WebPages some parts are less accessible than others
Hay una tendencia cada vez más grande de consultar la prensa digital	It is difficult to identify and isolate single articles

Recuperando prensa digital

- la descarga masiva no suele estar prevista en los medios digitales
- en muchos casos las noticias antiguas ya no están en la red
- en muchos casos, los periódicos han cambiado de servidores, aplicaciones web, etc. y no saben qué es lo que ellos mismos tienen
- legalidad dudosa de las descargas masivas

Recuperando prensa digital

La recuperación masiva de ntocias de un medio puede hacerse mediante un *crawler*

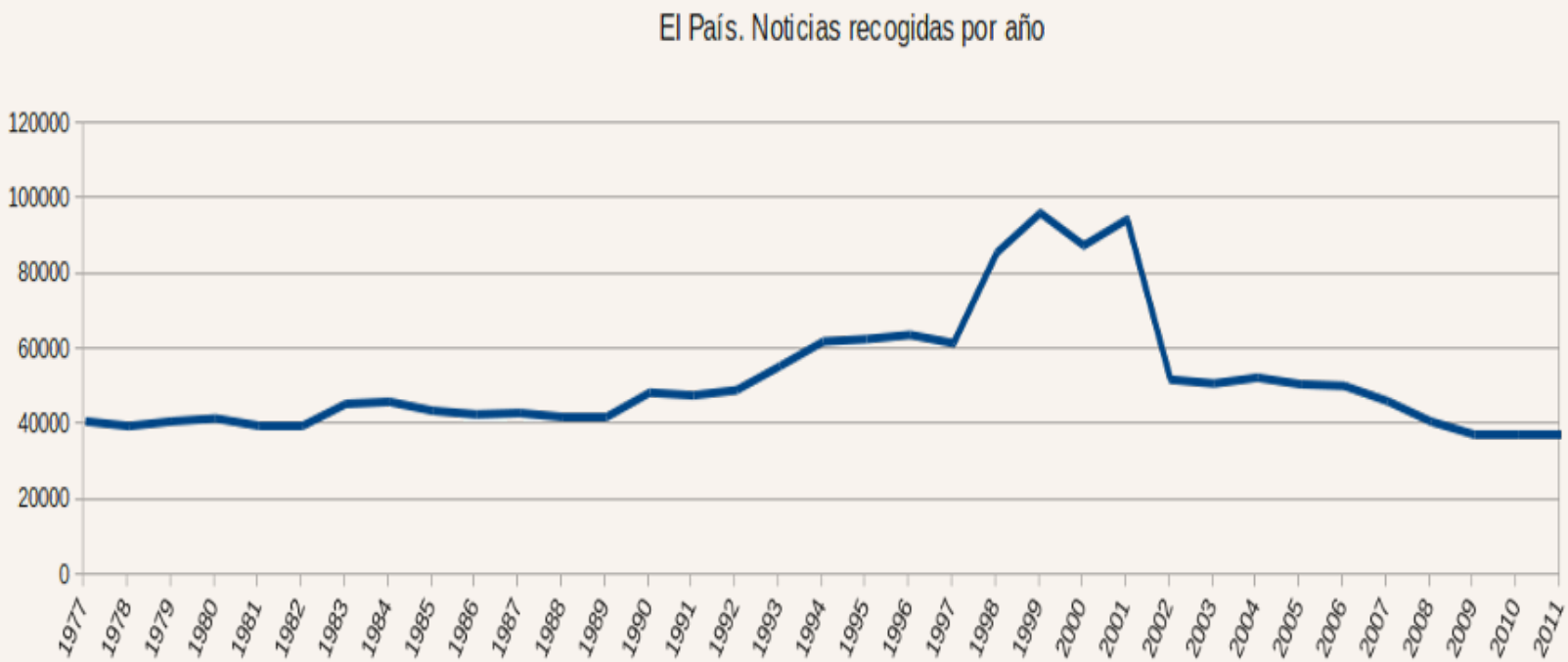
- un programa que navega autónomamente por la red
- su navegación puede ser guiada
- requieren conocer bien la estructura web del periódico a descargar

Pero ...

- las estructuras de los periódicos cambian con el tiempo
- podemos hacer descargas muy exhaustivas, pero no podemos estar seguros de que hemos obtenido el 100 % de las noticias

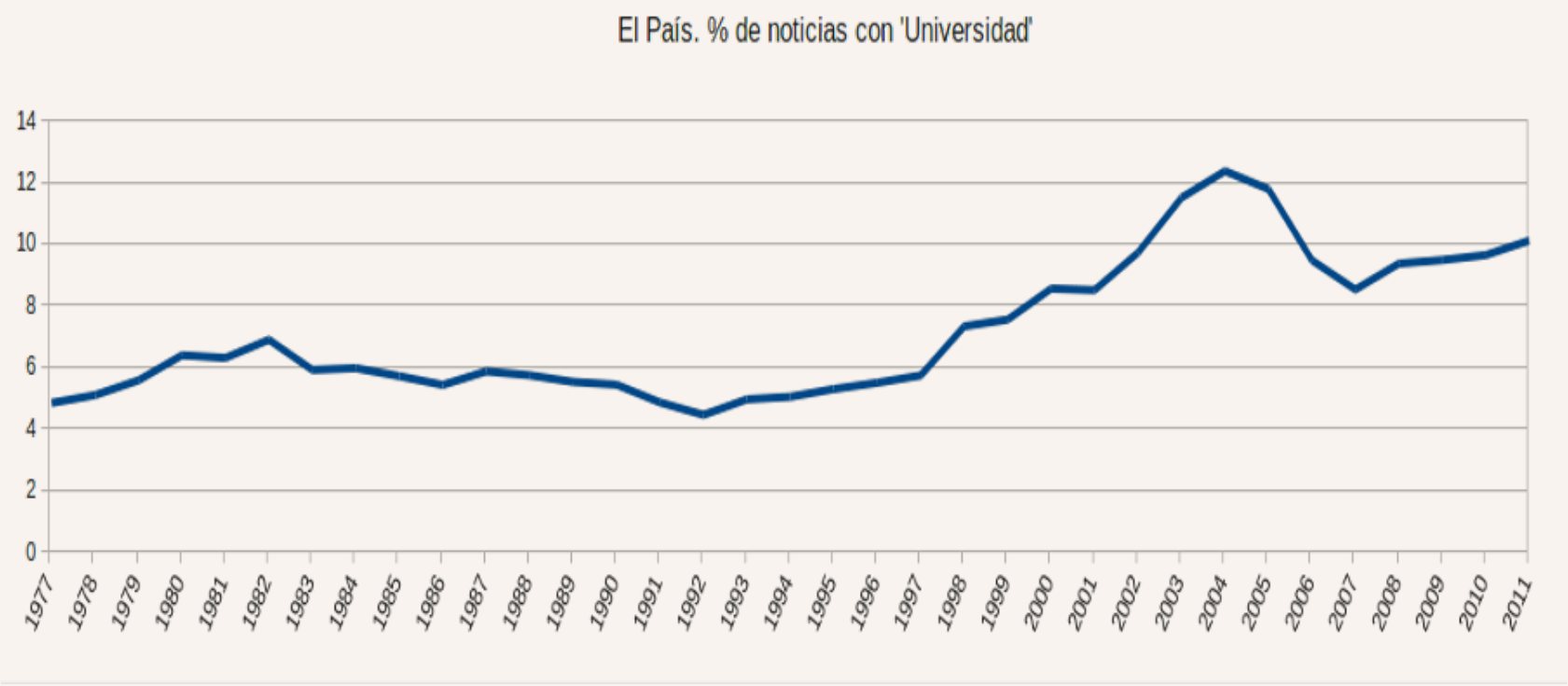
Resultados

- se han descargado las noticias de 1977 a 2011 de todas las secciones
- esto hace un total de **1.836.451** noticias



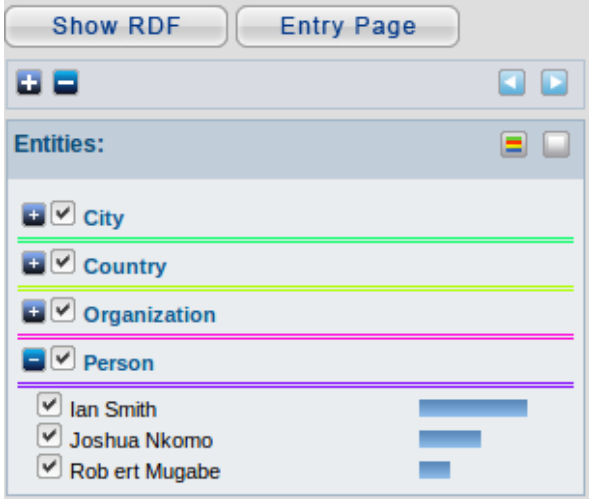
Resultados

- se han seleccionado las noticias que contienen **universidad**
- es un enfoque simplista



Reconocimiento de entidades

- entidades son personas, insituciones, empresas, paises, etc. que aparecen en un texto
- Reconocimiento de Entidades (Named Entities Recognition -NER) es un campo de la lingüística computacional que pretende detectar, normalizar y conectar con otros conocimientos las entidades que aparencen en un documento
- está estrechamente relacionada con elprocesamiento de lenguaje natural y la web semántica
- hay diversos programas que abordan esta tarea
  - ofrecen aciertos elevados, pero no al 100 x %
  - son dependientes del idioma
- uno de los más conocidos es *OpenCalais*



Las conversaciones de Rodesia, al borde de la ruptura

Las conversaciones que sostienen en Salisbury el primer ministro rodesiano y los líderes de los movimientos nacionalistas negros mod erados parecen situadas al borde de la ruptura, después que las organizaciones, dirigidas por el obispo Muzorewa y el reverendo Sith ole, hayan rechazado la nueva fórmula propuesta ayer por Ian Smith para garantizar una representación blanca de un tercio en un futu ro Parlamento independiente de 120 escaños.

Muzorewa y Sithole, cuyas divergencias se habían acentuado en los últimos días, respondieron en común a la iniciativa de **Smith** -elec ciones primarias para el tercio blanco mediante listas separadas para blancos y negros- con la contrapropuesta de reducir a un quint o los escaños de la minoría blanca y aceptar las listas separadas.La semana pasada, la organización de Muzorewa, el **Congreso Naciona** l Africano Unidos, había aceptado una representación parlamentaria blanca de un tercio. La conversación del obispo a la tesis de la fracción del **Congreso Nacional**, que dirige Sithole -sólo veinticuatro escaños para los blancos-, parece obedecer al temor de que un bloque de un tercio en manos del Frente Nacional, partido del primer ministro, **Ian Smith**, podría utilizar las evidentes divisiones e ntre los nacionalistas negros para seguir detentando el poder.

En **Londres** se considera probable que Ian **Smith** acuda a la undécima sesión de las conversaciones, fijada para el martes, con una prop uesta de reducir al 25 %, treinta escaños, la representación parlamentaria blanca. Para el primer ministro rodesiano es vital evitar la ruptura de unas negociaciones que, iniciadas hace ya casi un mes, son verosímilmente su última oportunidad para encontrar un com promiso que salve los derechos de la minoría blanca. **Washington** y **Londres** presionan sobre Salisbury en favor de cualquier acuerdo qu e impida la riada guerrillera de los líderes izquierdistas **Mugabe** y **Nkomo**. Las recientes incursiones navideñas de las fuerzas de **Rob ert Mugabe**, fuertemente apoyado por **Mozambique**, han debilitado aún más las posiciones de **Ian Smith** y contribuido, sin duda, a desacr editar a los ojos de la mayoría negra a los nacionalistas internos que dialogan con el primer ministro.

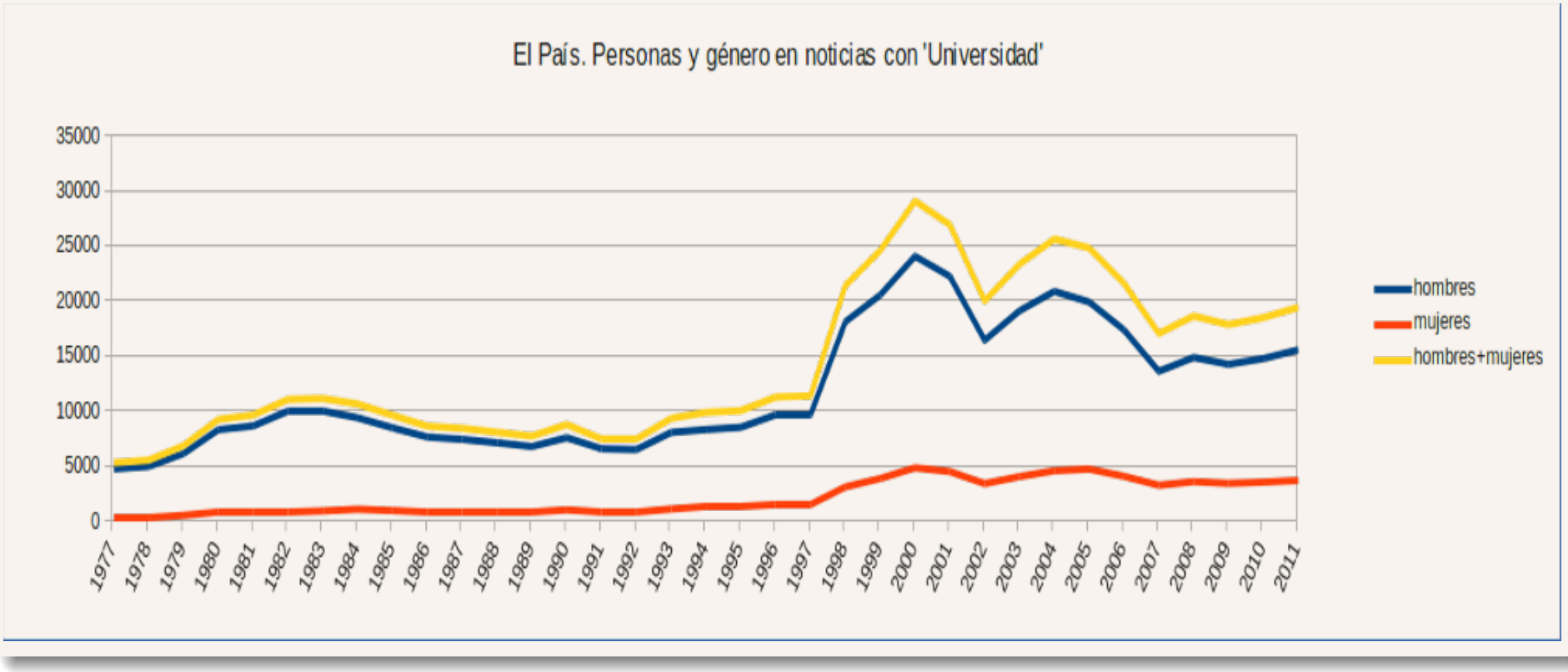
Aunque ninguno de los dos jefes del Frente Patriótico asisten a las negociaciones de Salisbury, **Smith** ha solicitado reiteradamente l a presencia de **Joshua Nkomo**, y fuentes guerrilleras no descartan absolutamente su participación.

Nkomo, apoyado por **Zambia**, y con cerca de 20.000 hombres bien armados y entrenados a su disposición, es ahora mismo el árbitro de la situación rodesiana. Su eventual incorporación a las conversaciones de Salisbury, si es que siguen adelante, sería un vuelco políti

Reconocimiento de entidades

- podemos obtener las peersonas que aparecne en las noticias, pero no saber si son hombres o mujeres
- podemos aproximar el sexo de las personas basándonos en el nombre de pila
- no es necesario revisar todas las personas
  - hay menos nombres de pila que personas
  - unos pocos nombres de pila son mucho más frecuentes
  - la revisión manual de 5.000 nombres de pila cubre el 91 % de todas las personas detectadas en todas las noticias

Resultados


El País. Personas y género en noticias con 'Universidad'
35000
30000
25000
20000
15000
10000
5000
0
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
hombres
mujeres
hombres+mujeres

Técnicas de Análisis de Redes Sociales?

- en realidad, estamos hablando de Teoría de Grafos
- la Teoría de Grafos nos permite representar entidades y sus relaciones
- estas técnicas son aplicadas por los sociólogos (pero también en muchas otras disciplinas)

SNA en breve

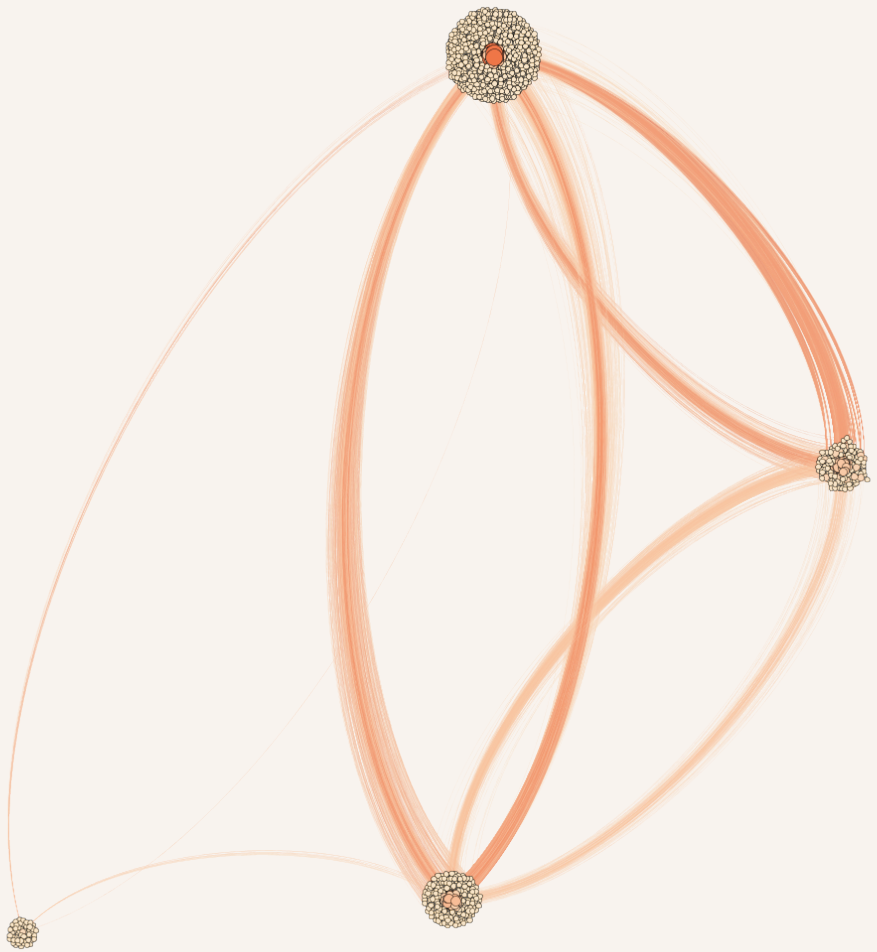
- las entidades pueden representarse como nodos de una red
- las relaciones entre entidades pueden representarse como enlaces entre nodos
- los nodos pueden tener características o atributos (tamaño, color, género, ...)
- los enlaces pueden ser dirigidos o no dirigidos
- los enlaces también pueden tener atributos
- un atributo interesante es el peso: la intensidad de ese enlace

SNA en breve

- la Teoría de Grafos nos proporciona diversas herramientas para:
  - evaluar la importancia de los nodos
  - caracterizar cada uno de los nodos
  - caracterizar toda la red
  - descubrir camninos entre los nodos
  - descubrir comunidades de nodos
  - muchas cosas más ...
- entidades y enlaces pueden usarse para representar cosas muy diversas

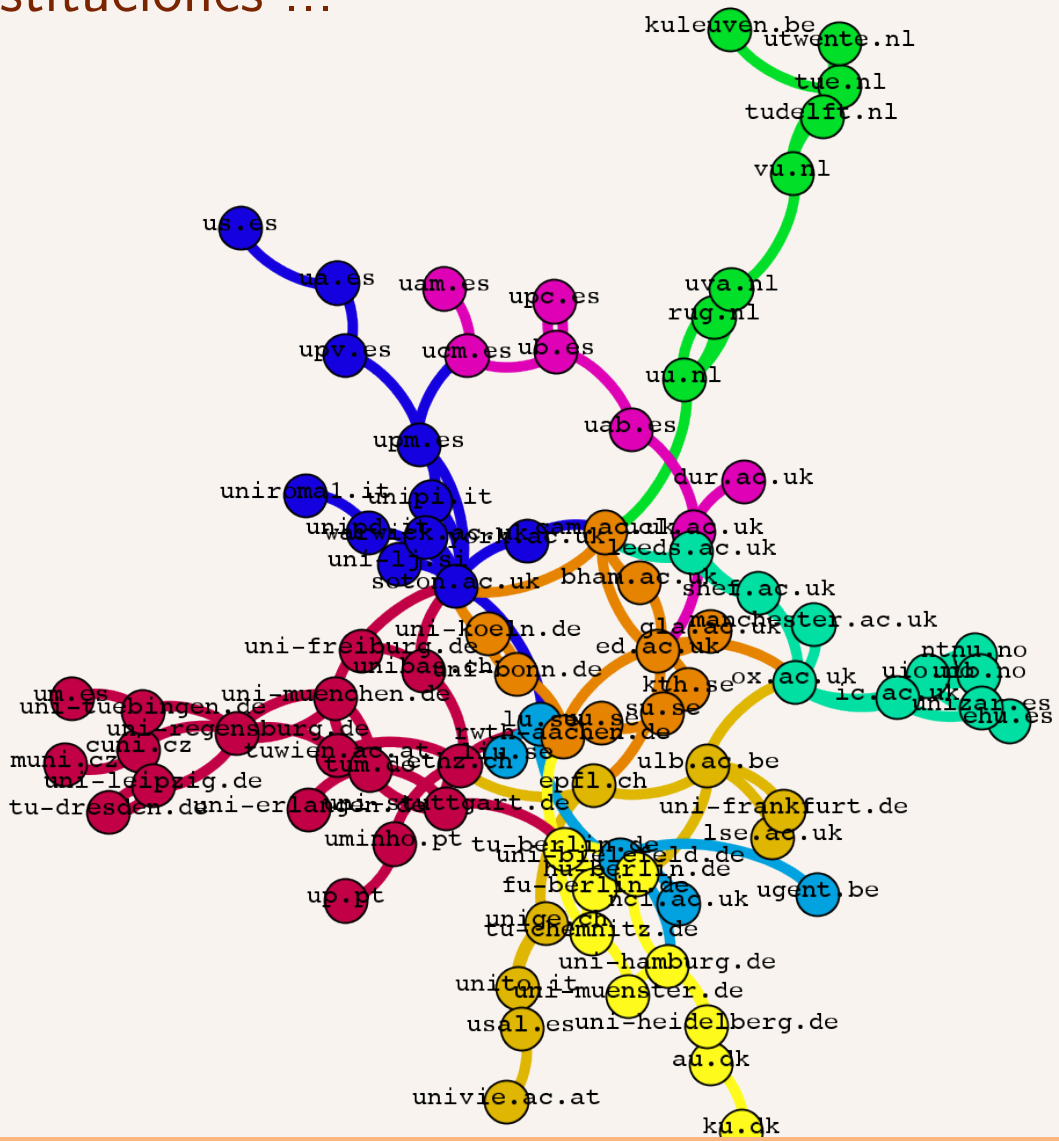
SNA en la práctica

Páginas de portales web...



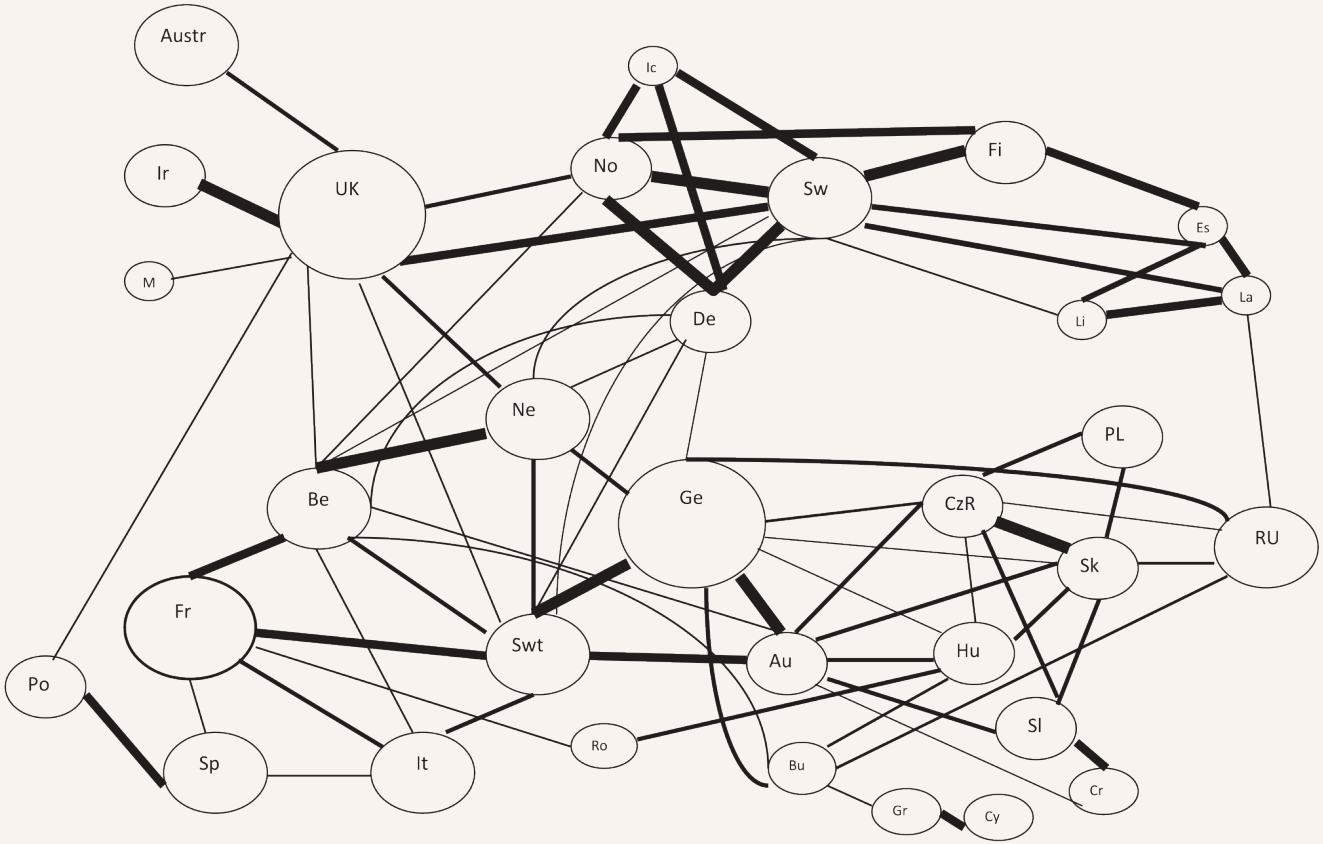
SNA en la práctica

Sedes web de instituciones ...



SNA en la práctica

Redes de investigación científica ...



SNA en la práctica

Artículos de la Wikipedia ...

SNA en la práctica

Entidades y enlaces pueden usarse para representar **relaciones entre personas**

- sabemos qué mujeres aparecen en las noticias de prensa
- podemos presuponer algún tipo de relación entre dos mujeres que aparecen en la misma noticia
- podemos construir una red en la que las mujeres son los nodos y su coocurrencia los enlaces
- de forma simple, el peso de un enlace puede ser el número de coocurrencias de cada par de mujeres

SNA en la práctica

Ejemplo: mujeres en las noticias de prensa sobre universidad en **1977-1982**

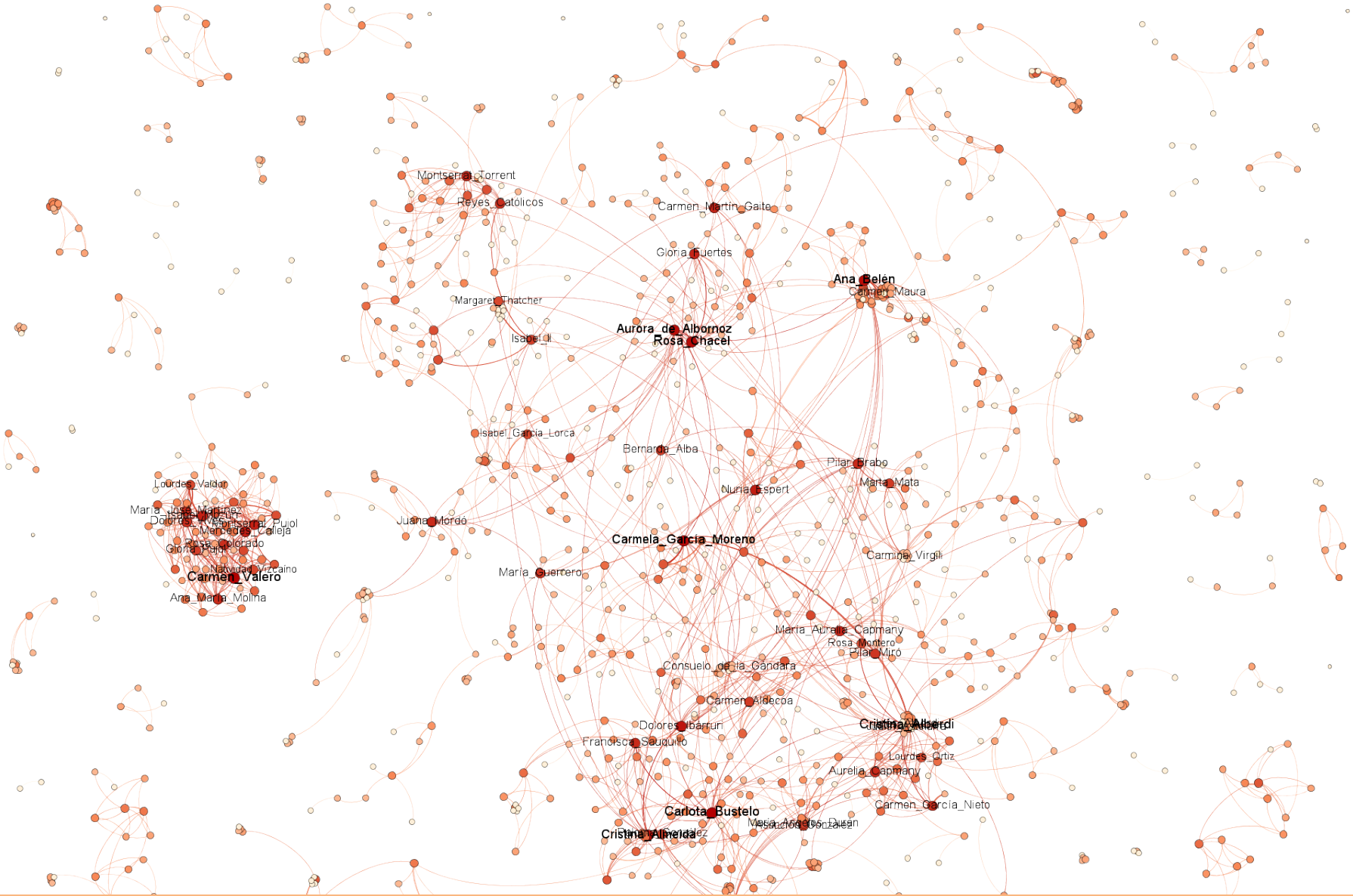
- diversas medidas pueden decirnos:
  - las mujeres más influyentes
  - los grupos de mujeres en diversos aspectos

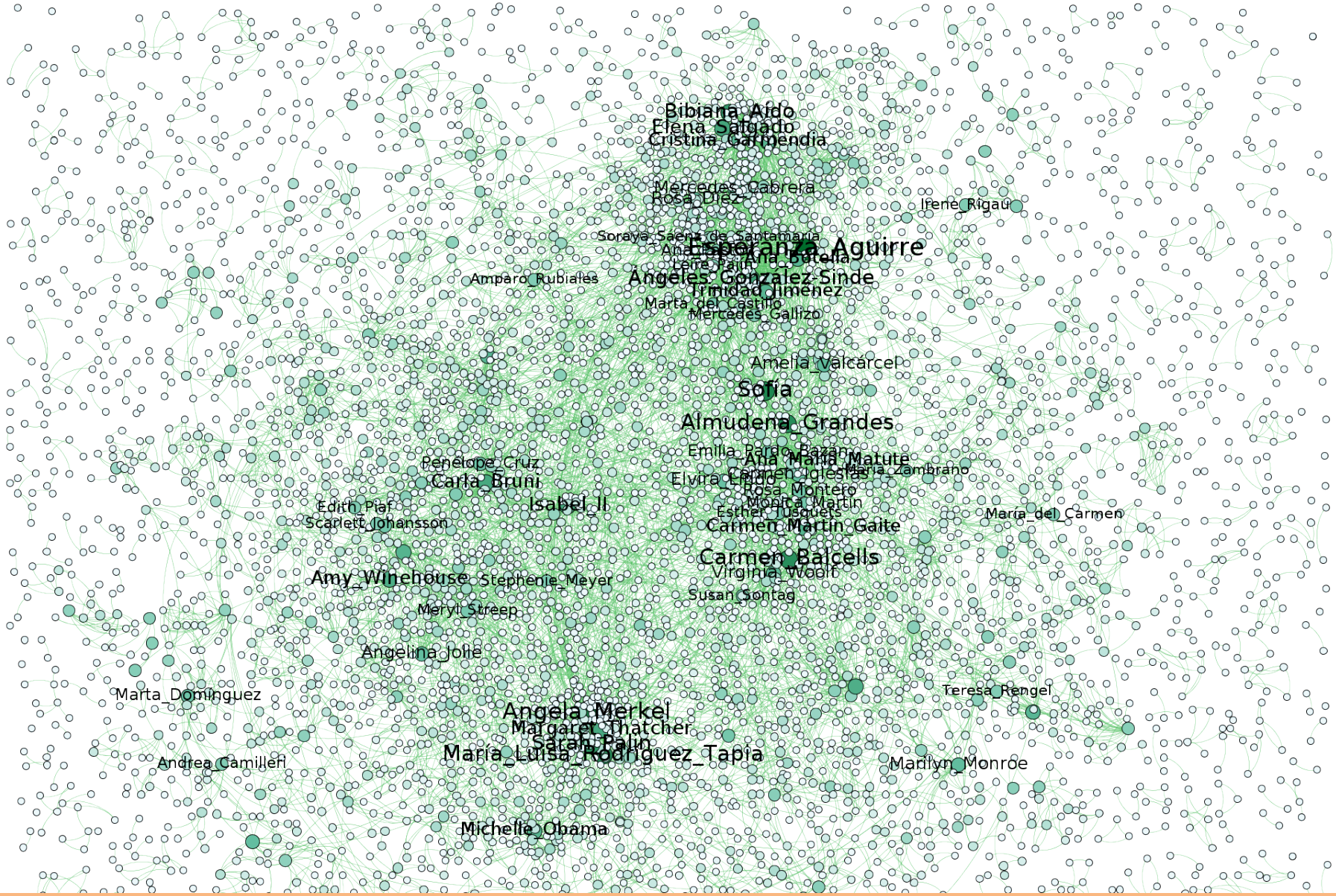
1977-1982

Id	Grado	Betweenness Centrality	PageRank
Cristina_Alberdi	33	0,0279610037	0,0033283736
Carmela_García_Moreno	25	0,018089944	0,0032463605
Carlota_Bustelo	32	0,020221993	0,0031233409
Rosa_Chacel	25	0,0293480212	0,0030726997
Ana_Belén	24	0,0186191111	0,0028006474
Aurora_de_Albornoz	23	0,0225590024	0,002764529
Carmen_Valero	34	4,43E-004	0,0024553748
Cristina_Almeida	23	0,0222441517	0,002264734
Gloria_Fuertes	17	0,012574569	0,0022208303
Dolores_Ibárruri	21	0,0085509555	0,0021862631
Consuelo_de_la_Gándara	21	0,0104848361	0,0021783937
Isabel_II	12	0,0084152943	0,0021467098
Nuria_Espert	19	0,0088234244	0,0021198721
Reyes_Católicos	15	0,0073478804	0,0021100397
Marta_Mata	13	0,0039183493	0,002036819
Juana_Mordó	16	0,0122100696	0,0020066295
Carmen_Martín_Gaite	17	0,0118921143	0,0019856645
Montserrat_Torrent	16	0,0091849321	0,0019416251
Pilar_Brabo	16	0,0085469801	0,0019047582
Carmina_Virgili	12	0,0041315558	0,001836938
María_Aurelia_Capmany	17	0,0115819133	0,0018330409
Margaret_Thatcher	11	0,0063293742	0,0018178536
Carmen_Llorca	19	0,0051503607	0,001741037
Francisca_Sauquillo	16	0,0071569292	0,0017285909
Pilar_Miró	14	0,0095841646	0,0016310046
María_Guerrero	14	0,011393758	0,001614452
Alicia_Urreta	7	1,18E-005	0,0015543884

2008-2011

Id	Grado	Betweenness Centrality	PageRank
Esperanza_Aguirre	142	0,0849896499	0,0049358854
Angela_Merkel	98	0,0450882493	0,0035658419
Sofía	85	0,0382839913	0,0023836925
Almudena_Grandes	64	0,0330072956	0,001707216
María_Luisa_Rodríguez_Tapia	92	0,0320753753	0,0038867664
Carmen_Balcells	57	0,0316760176	0,001497173
Sarah_Palin	51	0,0277308441	0,0017982835
Bibiana_Aído	56	0,0239869329	0,0015588673
Ángeles_González-Sinde	45	0,0218513629	0,0012597488
Elena_Salgado	54	0,019042578	0,0017862311
Isabel_II	57	0,0189153248	0,0018498177
Carla_Bruni	36	0,0175290326	9,94E-004
Margaret_Thatcher	40	0,0174877021	0,0012845486
Amy_Winehouse	55	0,0159092929	0,0015494064
Carmen_Martín_Gaite	56	0,0134217389	0,0014735435
Michelle_Obama	46	0,0133827102	0,0014348169
Ana_María_Matute	41	0,0133345001	0,0012192942
Cristina_Garmendia	41	0,013183799	0,0016481139
Trinidad_Jiménez	46	0,0129034095	0,0014378019
Ana_Botella	39	0,0125215646	0,001120722
Mercedes_Cabrera	46	0,011753169	0,0014533335
Virginia_Woolf	34	0,0115657595	8,67E-004
Rosa_Díez	38	0,0109827831	0,0012587642
Amelia_Valcárcel	37	0,0108070817	0,0011006807
Marilyn_Monroe	30	0,010306948	8,93E-004
Angelina_Jolie	29	0,0098127722	8,19E-004
Carmen_Iglesias	26	0,0092409589	7,07E-004


Montserrat_Torrent
Reyes_Católicos
Carmen_Martín_Gaite
Gloria_Fuertes
Ana_Belén
Carmen_Maura
Margaret_Thatcher
Isabel_II
Aurora_de_Albornoz
Rosa_Chacel
Isabel_García_Lorca
Bernarda_Alba
Pilar_Brabo
Marta_Mata
Nuria_Espert
Lourdes_Valdor
Juana_Mordó
Carmela_García_Moreno
Carmina_Virgili
María_Guerrero
María_Aurelia_Capmany
Rosa_Montero
Pilar_Miró
Consuelo_de_la_Gándara
Carmen_Aldecoa
Dolores_Ibarruri
Francisca_Sauquillo
Cristina_Almeida
Lourdes_Ortiz
Aurelia_Capmany
Carlota_Bustelo
Carmen_García_Nieto
Rosa_Colorado
Natividad_Vizcaíno
Carmen_Valero
Ana_María_Molina


Bibiana_Aido
Elena_Salgado
Cristina_Garmendia
Mercedes_Cabrera
Rosa_Diez
Irene_Rigau
Soraya_Saenz_de_Santamaria
Esperanza_Aguirre
Ana_Botella
Leire_Pajin
Amparo_Rubiales
Angeles_Gonzalez-Sinde
Trinidad_Jimenez
Marta_del_Castillo
Mercedes_Gallizo
Amelia_Valcarcel
Sofia
Almudena_Grandes
Emilia_Pardo_Bazan
Ana_Maria_Matute
Carmen_Iglesias
Maria_Zambrano
Penelope_Cruz
Carla_Bruni
Elvira_Lindo
Rosa_Montero
Monica_Martin
Esther_Tusquets
Edith_Piaf
Isabel_II
Scarlett_Johansson
Carmen_Martin_Gaite
Maria_del_Carmen
Carmen_Balcells
Amy_Winehouse
Stephenie_Meyer
Virginia_Woolf
Susan_Sontag
Meryl_Streep
Angelina_Jolie
Marta_Dominguez
Teresa_Rengel
Angela_Merkel
Margaret_Thatcher
Sarah_Palin
Maria_Luisa_Rodriguez_Tapia
Andrea_Camilleri
Marilyn_Monroe
Michelle_Obama

Detección de temas (Topic Detection)

- en un conjunto de documentos, identificar los principales temas tratados en esos documentos
- los documentos pueden ser noticias de prensa, pero también:
  - mensajes de correo electrónico
  - entradas o comentarios en blogs
  - tweets

La vía clásica de abordar la detección de temas es a través de la clasificación automática de oso documentos (clustering)

Detección de temas

Tenemos un corpus de noticias sobre universidad. También sabemos en qué noticias intervienen mujeres. Nuestro objetivo es:

- detectar los principales temas tratados en las noticias
- conocer en qué temas participan las mujeres, y con qué intensidad

Podemos intentar detectar temas aplicando técnicas de Análisis de Redes Sociales:

- construyendo una red o grafo de noticias
- aplicando a esa red técnicas de descubrimiento de comunidades

La red de noticias

- cada noticia es un nodo en esa red
- podemos enlazar una noticia con otra si ambas son semánticamente parecidas
- el peso de ese enlace puede ser el grado de parecido semántico entre ambas noticias

La distancia semántica entre documentos

- existen muchos métodos para medir la distancia semántica entre documentos
- la Recuperación de Información clásica aplica tales métodos para encontrar documentos adeciados a las búsquedas de los usuarios
- la mayor parte de los motores de búsqueda más conocidos utilizan estas técnicas
- podemos aplicar estos métodos para estimar el parecido semántico entre dos noticias

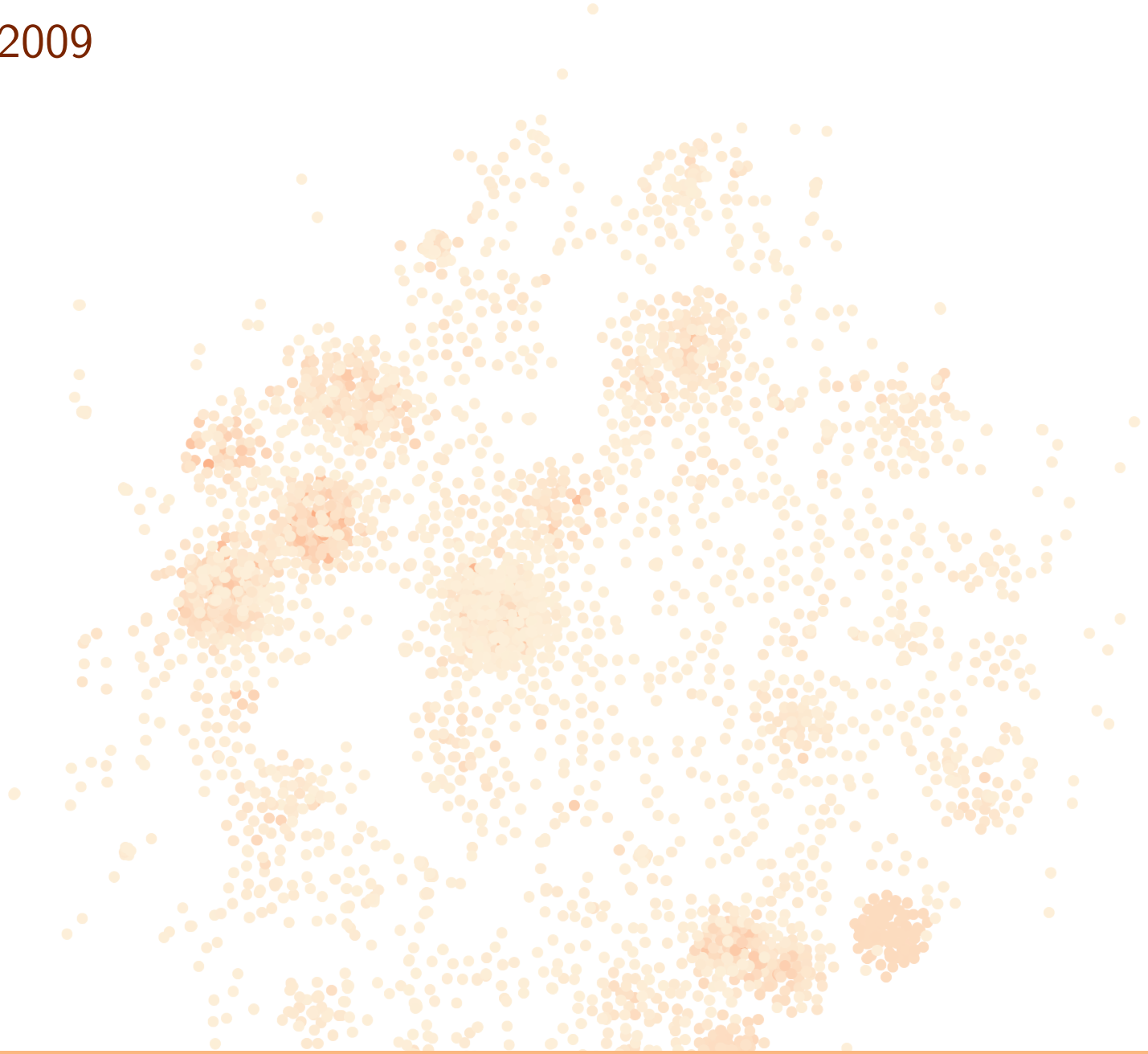
La distancia semántica entre documentos

- podemos ver un documento (una noticia de prensa) como una lista de palabras; de hecho, podemos usar una clase de lista que los matemáticos llaman *vector*
- las palabras en esa listas pueden ser normalizadas de alguna manera; y también pueden tener pesos. Por ejemplo: un número que indique su representatividad sobre el contenido semántico de la noticia
- el peso de una palabra dentro de un documento se puede calcular automáticamente, basándose en su frecuencia, el lugar del documeto en que aparece, etc.
- las matemáticas nos proporcionan métodos para estimar el parecido (distancia o cercanía) entre dos listas (vectores)

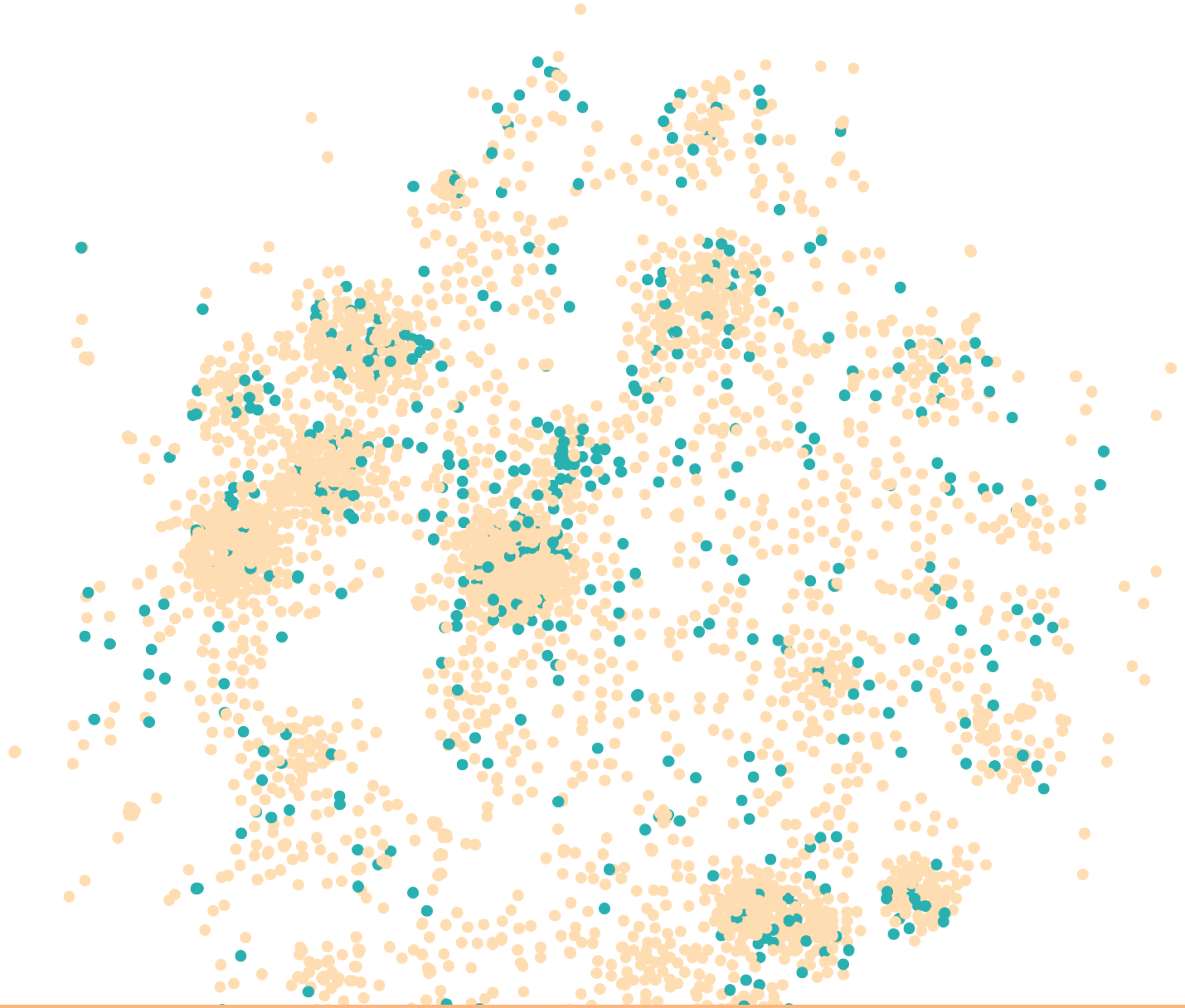
Construyendo la red

- como se ha dicho antes, cada noticia es un nodo y enlazamos dos nodos si son semánticamente parecidos
- el peso del enlace es el grado de parecido
- las noticias que tratan sobre los mismos temas aparecerán más inter-enlazadas

Noticias 2009



Noticias 2009 (noticias con mujeres en color azul)



Descubrimiento de comunidades

- las representaciones gráficas son inexactas
- hay procedimientos no gráficos de *Descubrimiento de comunidades*
  - una *comunidad* es un conjunto de nodos que enlazan fuertemente entre sí y débilmente con los no pertenecientes a ese conjunto
  - en nuestro caso, una comunidad es un conjunto de noticias que tratan (más o menos) del mismo asunto
- estos procedimientos producen un listado de las comunidades halladas y de los nodos miembros de cada una

```
1,1,0.00456697,pais19770226-177,"Profesores no numerarios: sigue la huelga"
1,2,0.00422373,pais19770123-135,"La huelga de profesores continuará esta semana"
1,3,0.00417448,pais19770127-135,"Se acentúa el conflicto de la enseñanza"
1,4,0.00399125,pais19770222-153,"El Ministerio amenaza con sanciones económicas a los PNN de institu
1,5,0.00385055,pais19770204-132,"Los enseñantes piden estabilidad laboral y control democrático de l
...
...
2,1,0.00358621,pais19771008-136,"Los partidos mayoritarios y la Federación Católica de Padres se pro
2,2,0.00348011,pais19770614-099,"Ultimas intervenciones de los líderes en Televisión"
2,3,0.00339085,pais19770216-177,"Sobre los PNN de Universidad"
2,4,0.00274552,pais19770518-061,"Al día siguiente"
2,5,0.00257922,pais19771222-117,"Los problemas concretos y la autocrítica"
2,6,0.00250335,pais19770423-129,"Entrevista con Arias Navarro"
2,7,0.0024123,pais19770720-067,"El voto por la libertad y la democracia"
...
...
13,1,0.00180532,pais19770330-009,"Los estudiantes de Educación Física se manifiestan ante el ministe
13,2,0.00177176,pais19770212-013,"Ministerio de Educación y DND tratarán el lunes los problemas del
13,3,0.00137942,pais19770326-020,"Los estudiantes de Educación Física reivindican un rango universit
13,4,0.00134001,pais19770218-016,"La educación física, un problema académico, laboral y político"
13,5,0.000893123,pais19771227-021,"La figura del profesor, centro de atención"
13,6,0.000752091,pais19771112-149,"Las instalaciones deportivas oficiales, abiertas al público"
13,7,0.000743549,pais19770430-026,"Los educadores físicos, pilares fundamentales del deporte alemán"
...
...
```

Algunos recursos interesantes

- Crawling
  - **wget** `http://www.gnu.org/s/wget/`, `http://gnuwin32.sourceforge.net/packages/wget.htm`
- SNA
  - Pajek `http://pajek.imfm.si/doku.php`
  - Gephi `http://gephi.github.io/`

Muchas gracias por su atención !!

Información de contacto

- Carlos G. Figuerola *figue@usal.es*
- Tamar Groves *tamargroves@usal.es*
- Francisco J. Rodríguez Jiménez *fjrodriguezjimen@gmail.com*