

enero, 2024

Informe Trabajo Cibermetría

Alberto García Martín



Máster en Sistemas Inteligentes
Universidad de Salamanca

Índice

1. Introducción	1
2. Análisis de datos de la revista EPI	1
2.1. Análisis de coautorías	1
2.2. Análisis de coincidencia de palabras clave	6
3. Análisis cibermétrico de la web de Pfizer	11
3.1. Recogida de enlaces	11
3.2. Análisis de redes	12
4. Conclusiones	17

1. Introducción

En este informe se describe el trabajo realizado para la asignatura de Cibermetría, perteneciente al máster en Sistemas Inteligentes. Este informe está dividido en dos secciones principales, la primera, dedicada a la parte de R y netCoin, donde se analizan datos de la revista “El Profesional de la Información”, centrándose en un análisis de redes de coincidencia entre las diferentes características de los registros. En la segunda parte, correspondiente a Nutch y Gephi, se realiza un análisis cibermétrico de la web de Pfizer.

2. Análisis de datos de la revista EPI

El análisis de datos de la revista “El Profesional de la Información” se ha realizado en dos fases. En primer lugar, se ha realizado un análisis de coautorías, para encontrar las relaciones entre los autores de los artículos. Posteriormente, se ha realizado un análisis de coincidencia entre las palabras clave de los distintos artículos.

2.1. Análisis de coautorías

Para realizar el análisis de coautorías, primero se ha realizado un preprocesamiento del campo de los autores. En este preprocesamiento se han sustituido los guiones por espacios, y se han dividido los autores tanto por el carácter “;” como por el carácter “,”, ya que en algunos casos se utilizaba uno u otro para separar los autores.

Una vez divididos los autores de cada artículo, se ha creado una lista de aristas, donde cada arista representa una colaboración entre dos autores, y cada vértice representa un autor. A la hora de crear esta lista, solo se han tenido en cuenta los registros donde había alguna coautoría, es decir, se han descartado los registros donde solo había un autor.

Cada posible combinación de dos autores se ha añadido a la lista de aristas, y se ha añadido toda la información sobre el artículo en el que colaboraron a la arista correspondiente. De esta forma, cada arista contiene como fuente y destino los dos autores que colaboraron, y como atributos el título del artículo, el tema, la sección, el número total de autores, el año de publicación, y el número y volumen de la revista. Todas las aristas serán no dirigidas y tendrán un peso unitario. Si hay dos autores con múltiples colaboraciones, se añadirán tantas aristas como colaboraciones haya entre ellos.

A cada vértice se le añadirá como atributo el nombre del autor, y los artículos en los que ha participado, así como el número total de artículos en los que aparece como autor. También se han calculado medidas de centralidad para cada vértice. El grafo resultante se ha convertido a un objeto de `igraph` y se ha representado con netCoin, como se puede ver en la figura 1.

Trabajo Cibermetría

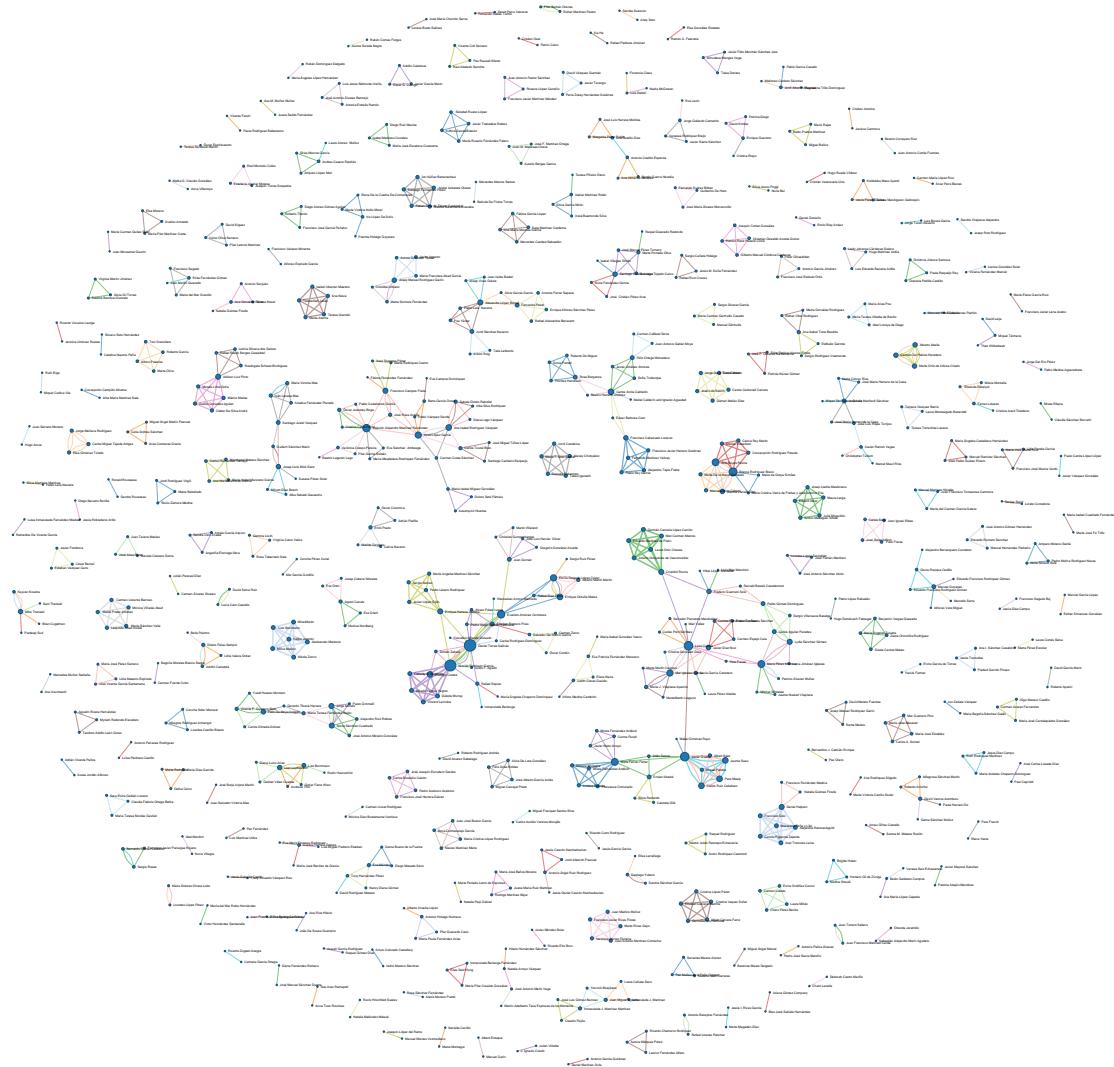


Figura 1: Grafo de coautorías

Las métricas generales del grafo son las siguientes, al tratarse de un grafo que no está totalmente conectado, algunas métricas como el diámetro o el camino medio se han omitido:

- **Número de vértices:** 797. El número total de autores que han colaborado en algún artículo.
- **Número de aristas:** 1003. El número total de colaboraciones entre autores.
- **Densidad:** 0.003. El número de aristas entre el número total de posibles aristas. Esta métrica muestra cuán conectado está el grafo, en este caso el valor es muy bajo, lo que indica que el grafo está poco conectado. Esto coincide con lo que se ve en la red, cada autor tiende a colaborar con un número limitado de otros autores, habiendo subgrupos que no están fuertemente interconectados con otros grupos dentro de la red.
- **Grado medio:** 2.52. El número medio de aristas que inciden en cada vértice. Este valor indica que de media, los autores cuando colaboran con otros autores, lo hacen con dos o tres autores más. La distribución de grados se puede ver en la figura 2, muy pocos autores tienen un grado superior a 6, lo que indica que la mayoría de autores colaboran con un número reducido de otros autores.

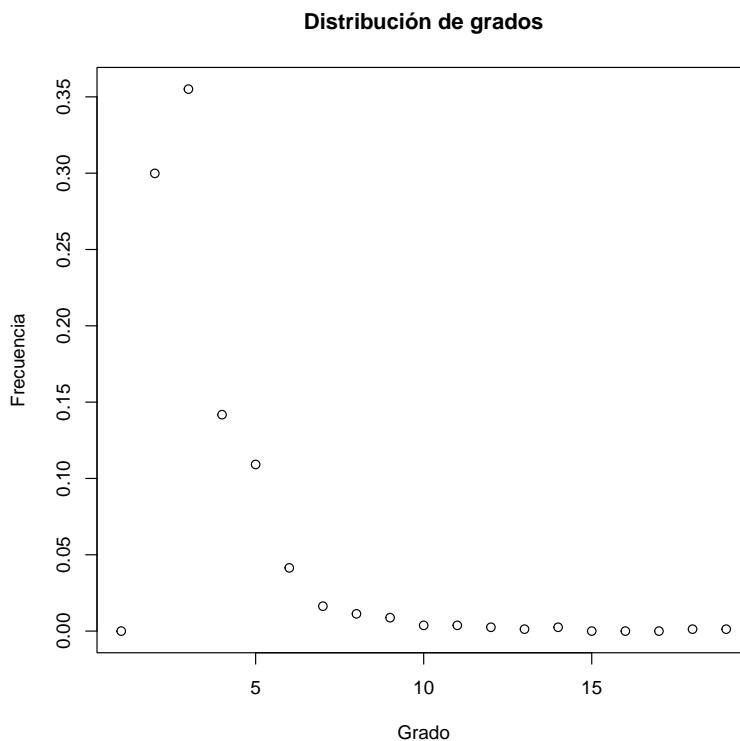


Figura 2: Distribución de grados

Para realizar la representación del grafo de la figura 1, el tamaño de los vértices se ha calculado en función del grado. La anchura de las aristas se ha determinado

Trabajo Cibermetría

según el número de autores que han colaborado en el artículo, y el color en función del tema del artículo.

El grafo de coautorías se caracteriza principalmente por estar formado por varios subgrafos, donde cada subgrafo está formado por autores que han colaborado entre sí, pero que no han colaborado con autores de otros subgrafos. Aplicando un algoritmo de detección de comunidades, como el de Louvain, cada comunidad detectada se corresponde a uno de estos subgrafos, este análisis es trivial sin aplicar ningún algoritmo, ya que se pueden ver claramente los subgrafos en la figura 1.

Para analizar más detalladamente una comunidad en concreto, se ha seleccionado la comunidad más grande, que está formada por un total de 54 autores. En la figura 3 se puede ver este grafo, donde además de los indicadores del grafo principal se ha añadido una agrupación por artículos, para identificar fácilmente los autores que han trabajado en exactamente los mismos artículos. Además el color de los vértices se ha determinado en función de la centralidad de intermediación. Esta métrica mide el grado en que un vértice actúa como un puente en las conexiones entre otros vértices. Los vértices con un color más oscuro desempeñan un papel más importante al conectar grupos que de otro modo estarían menos conectados.

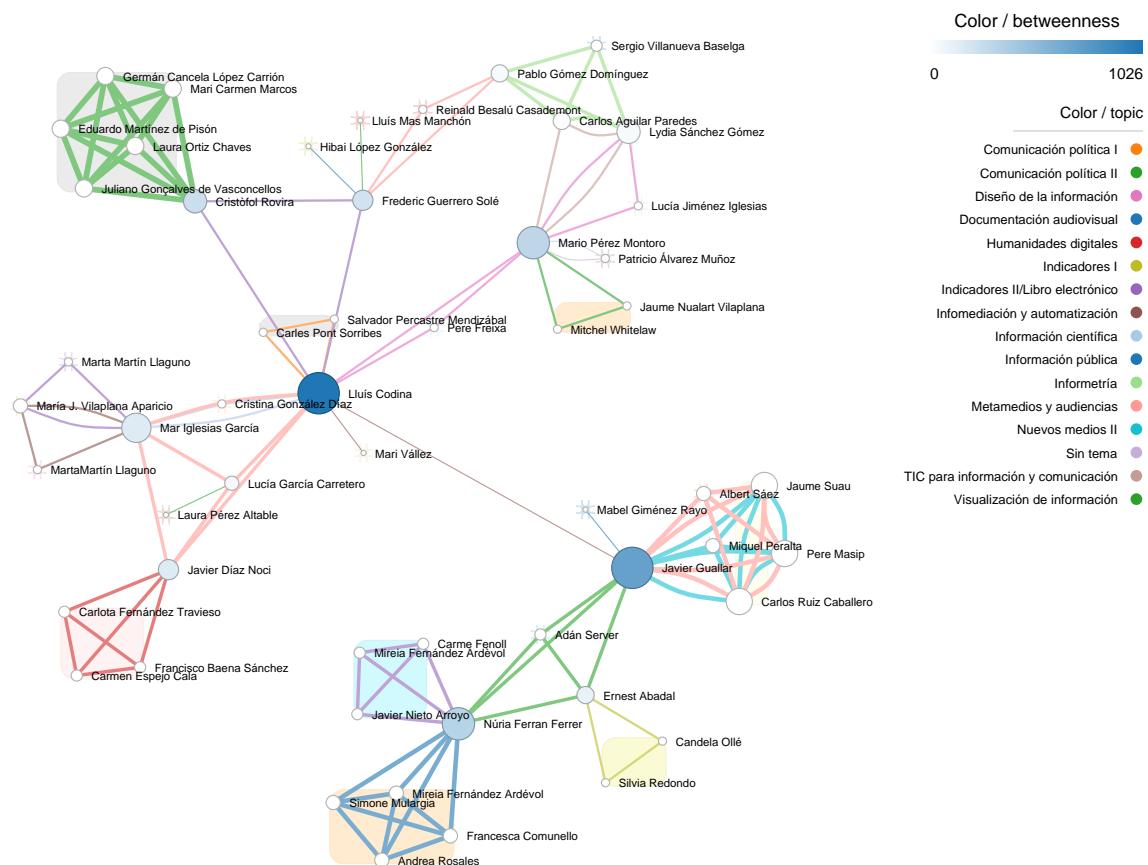


Figura 3: Comunidad más grande del grafo de coautorías

Se podría pensar que una comunidad de autores que colaboran entre sí suelen tratar los mismos temas en sus artículos, pero en este caso no es así, ya que se puede ver que hay autores que han colaborado en los mismos artículos, pero que tratan

temas muy diferentes.

Entre todos los autores destaca “Lluís Codina”, que es el autor que más artículos ha escrito y que más colaboraciones ha tenido. Además es el que tiene una mayor centralidad de intermediación, y de cercanía, lo que indica que es el autor con una mayor cercanía con el resto de autores y el que tiene una mayor influencia en esta comunidad. Su alta centralidad de intermediación se debe en gran parte a su colaboración con “Javier Guallar”, que tiene un número igual de colaboraciones que él, pero menos artículos en total, y una centralidad de intermediación menor.

También se ha analizado la segunda comunidad más grande, que está formada por 37 autores. En la figura 4 se puede ver este grafo, donde se aplican los mismos indicadores que en el grafo anterior, pero en este caso el color de los nodos se determina en función de la centralidad de vector propio. Esta métrica mide la importancia de un vértice en función de la importancia de sus vecinos, cuantificando directamente la influencia de un vértice en su comunidad. Los vértices con un color más oscuro son los que tienen una mayor influencia en la comunidad.

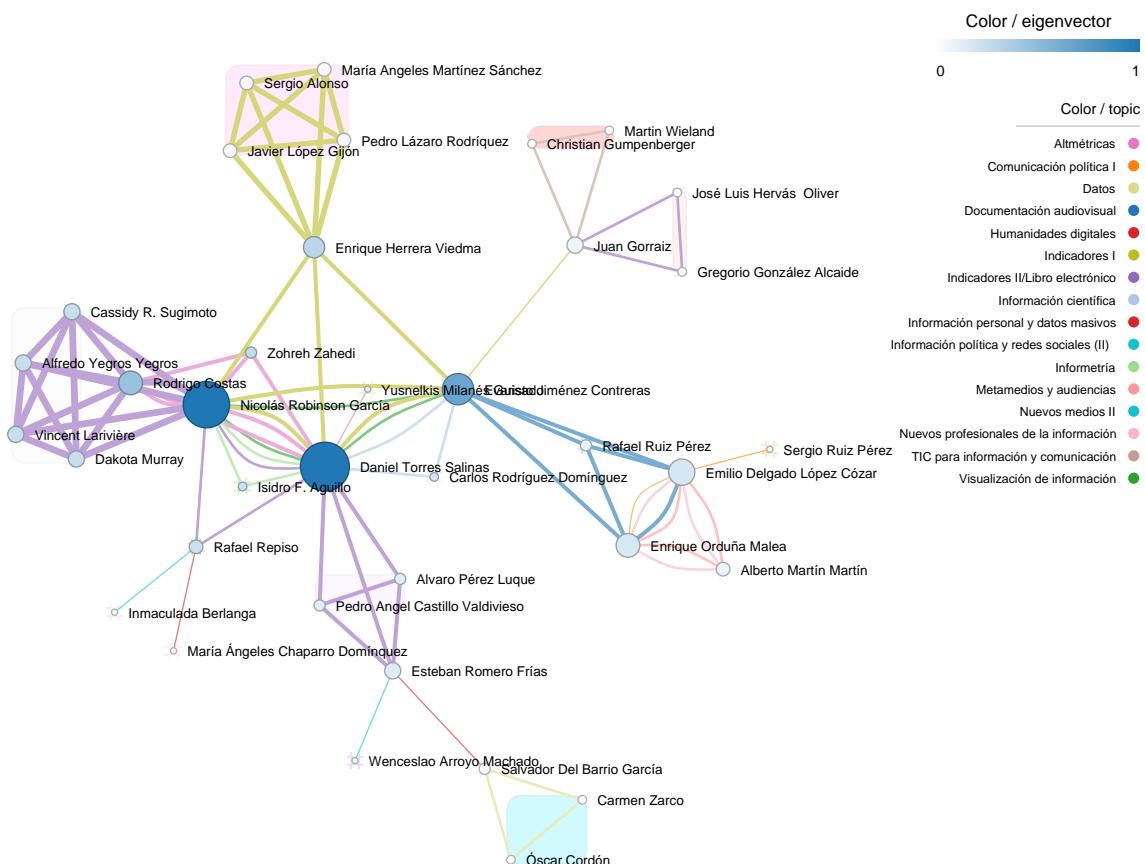


Figura 4: Segunda comunidad más grande del grafo de coautorías

A diferencia de la comunidad anterior, en esta sí que se puede identificar dos temas principales en los artículos de colaboración, correspondientes a los temas “Indicadores I” e “Indicadores II”. En este caso, el autor con mayor centralidad de vector propio es “Daniel Torres Salinas”, que es el que más artículos ha escrito de esta comunidad, y el que mayor grado tiene, no solo en esta comunidad sino en todo

el grafo. También es el que mayor *pagerank* tiene, que es una métrica que mide la importancia de un vértice en función de la importancia de sus vecinos, pero que tiene en cuenta la importancia de los vecinos de los vecinos, y así sucesivamente. Además destaca “Nicolás Robinson García”, que es el segundo autor con mayor centralidad de vector propio, aunque con un menor número de artículos.

2.2. Análisis de coincidencia de palabras clave

En el caso del análisis de coincidencia de palabras clave, se ha realizado un preprocesamiento de los campos de palabras clave, pasando todos los caracteres a minúsculas y dividiendo las palabras clave por los caracteres “;” y “,”. Una vez divididas las palabras clave, se han eliminado los espacios en blanco al principio y al final de cada palabra clave, y se han eliminado los signos de puntuación, normalizando de esta forma las palabras clave.

Las aristas del grafo se han creado de forma similar al caso anterior, añadiendo una arista por cada combinación de dos palabras clave que aparecen en el mismo artículo, y añadiendo los mismos atributos, añadiendo además el número total de palabras clave que aparecen en el artículo. De forma similar al caso de los autores, cada vértice tendrá como atributo el nombre de la palabra clave, y los artículos en los que aparece, así como el número total de artículos y las medidas de centralidad.

El grafo resultante se puede ver en la figura 5, donde se han aplicado los mismos indicadores que en el caso anterior. En este caso se puede ver claramente que el grafo es mayor y está mucho más interconectado y que tan solo hay unas pocas comunidades aisladas del resto.

Las métricas generales del grafo son las siguientes:

- **Número de vértices:** 2393.
- **Número de aristas:** 16289.
- **Densidad:** 0.0056. La densidad es relativamente baja a pesar de tener un alto número de aristas, esto puede deberse a que tan solo hay un número reducido de palabras clave que aparecen en muchos artículos, y actúan como puente entre los grupos de palabras clave que aparecen en uno o dos artículos.
- **Diámetro:** 7. El diámetro es la distancia más larga posible entre dos nodos en la red principal. En este caso, el diámetro es 7, lo que indica que se necesita un máximo de siete pasos para viajar de un nodo a cualquier otro nodo en la red. Este valor corresponde a una red más extendida, con áreas que pueden estar relativamente aisladas unas de otras.
- **Longitud media del camino más corto:** 3.37. Esta métrica indica cuántos pasos, de media, son necesarios para conectar cualquier par de nodos en la red principal. Un valor de 3.37 sugiere que se pueden conectar dos palabras clave a través de aproximadamente tres pasos.

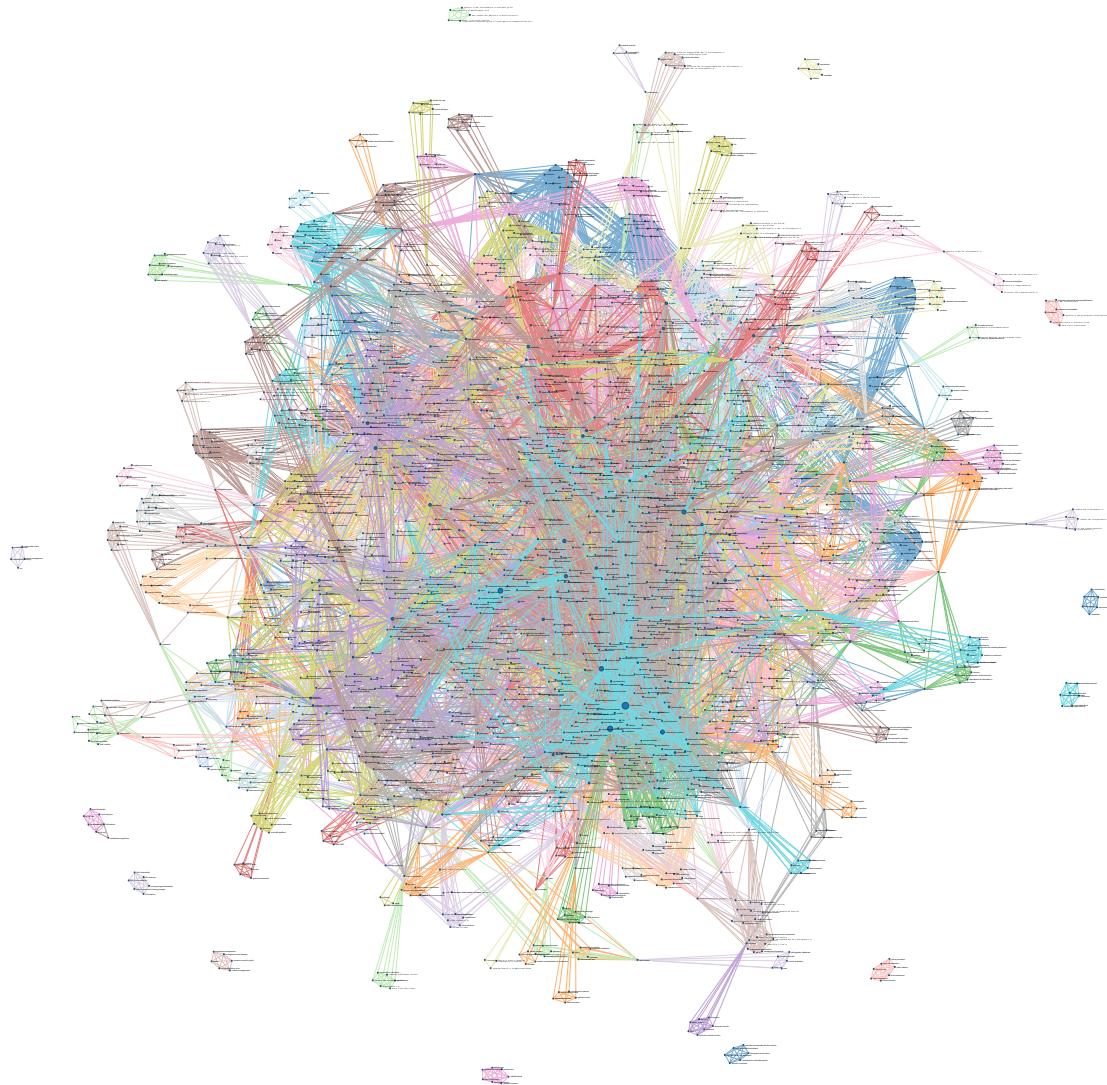


Figura 5: Grafo de palabras clave

- **Grado medio:** 13.61. Este valor significa que cada palabra clave, de media, está directamente relacionada con alrededor de 14 otras palabras clave. La distribución de grados se puede ver en la figura 6, donde se puede ver que hay un número reducido de palabras clave que tienen un grado mayor de 20, pero estos casos atípicos pueden llegar a alcanzar un grado superior a 400.

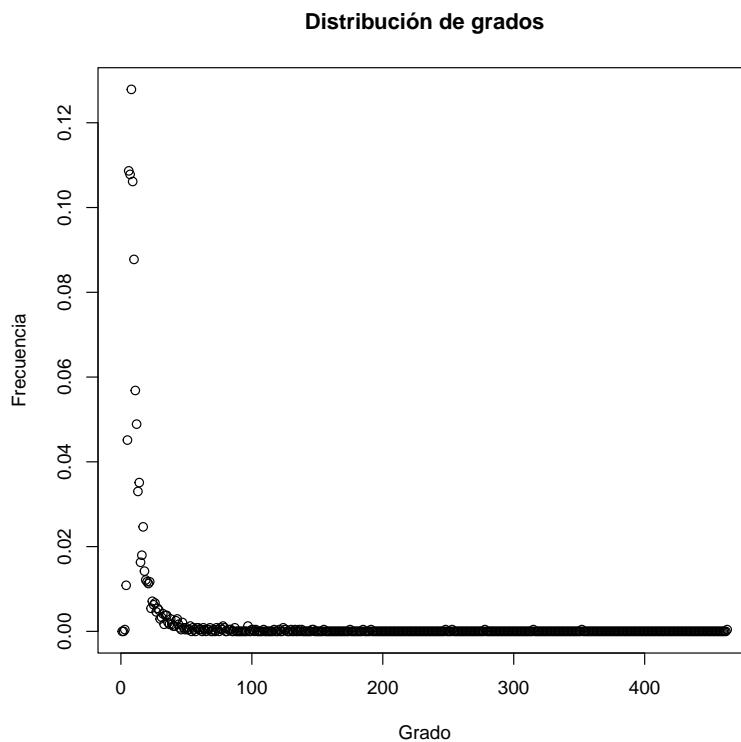


Figura 6: Distribución de grados

Dado el tamaño del grafo, es muy difícil analizarlo en su totalidad, por lo que se ha decidido filtrar el grafo en temas para estudiar cuáles son las palabras clave más importantes en los temas seleccionados. Para ello, primero se ha aplicado un filtro para quedarse con las palabras claves correspondientes a artículos del tema “Información política y redes sociales”, eliminando las palabras claves que se quedan sin aristas al aplicar el filtro. En la figura 7 se puede ver el grafo resultante, en este caso el color de los nodos se determina en función del *pagerank*, indicando que una palabra clave es más importante cuanto más oscuro sea su color.

Se puede ver que hay dos comunidades principales que no están conectadas entre sí, la más pequeña, tiene como nodo central y palabra más importante “españa”, mientras que la comunidad más grande tiene como palabra más importante “redes sociales”. Esta última palabra es la que mayor grado tiene en general en el grafo completo, y la que mayor *pagerank* tiene en esta comunidad. También destaca la palabra “twitter”, que es la segunda palabra más importante en esta comunidad en cuanto a *pagerank*. Otras palabras importantes son “comunicación política”, “medios sociales” e “información política”.

Es llamativo que haya muchas palabras clave que están conectadas solo con su propio grupo de palabras clave perteneciente al mismo artículo, y que no están conectadas con el resto de palabras clave. Esto puede significar que o las palabras clave son demasiado específicas del tema sobre el que trata el artículo, o que el conjunto filtrado no es lo suficientemente grande. Aun así, se puede ver que hay palabras clave que actúan como puentes conectando a grupos de palabras clave que de otra forma estarían aislados, como es el caso de “medios digitales” o “series de televisión” en la parte superior del grafo. El uso de estos términos en vez de otros más concretos y similares entre sí como “podcasts” o “podcasting” puede ser beneficioso a la hora de encontrar artículos relacionados con estos temas.

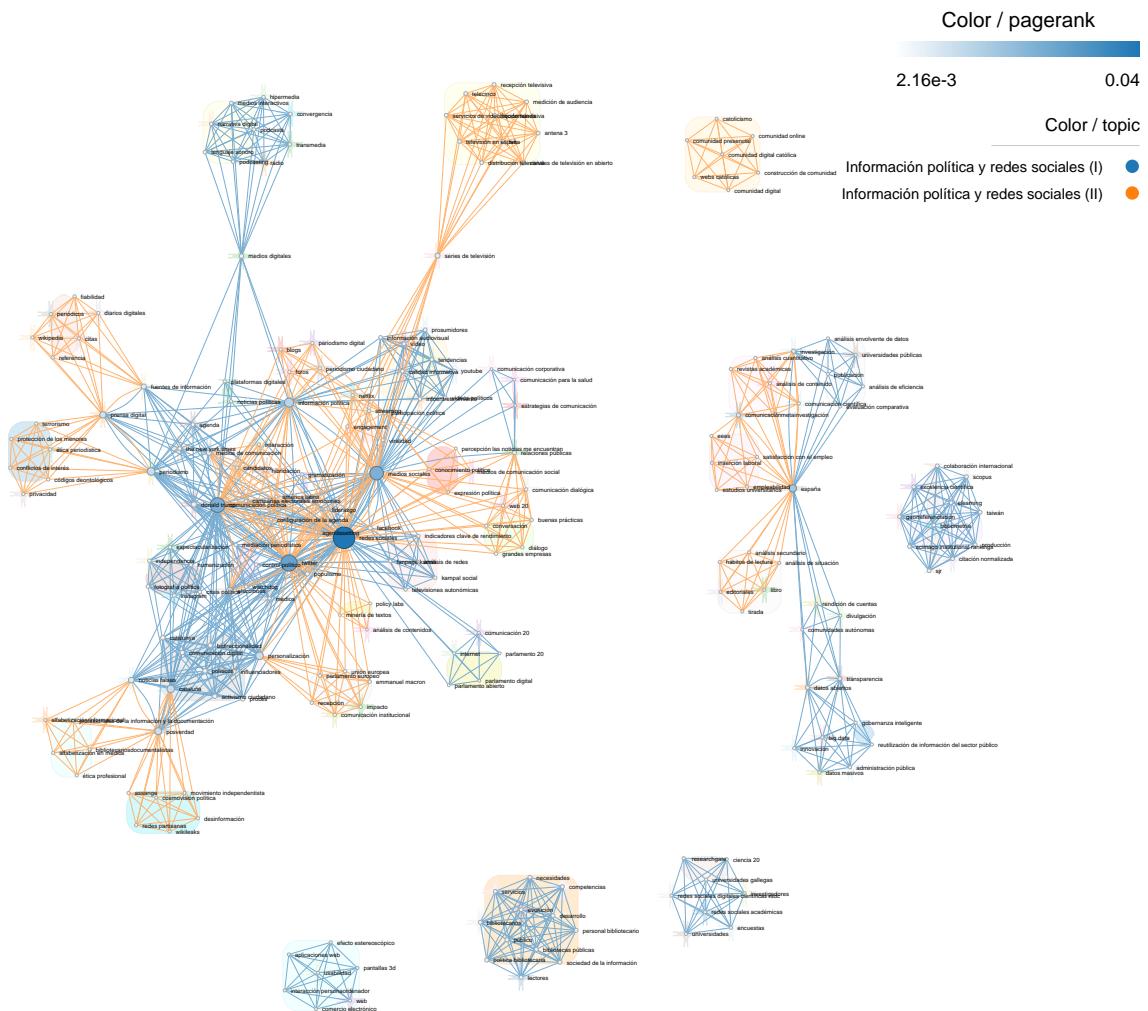


Figura 7: Grafo de palabras clave filtrado por tema

A continuación, se ha vuelto a aplicar un filtro sobre el grafo principal, pero esta vez sobre los artículos pertenecientes al tema “Indicadores”. En la figura 8 se puede ver el grafo resultante. Esta vez, exceptuando un grupo de palabras aislado, todas las palabras clave se encuentran interconectadas en una única comunidad.

Como era de esperar dado el tema de los artículos, la palabra clave más importante es “indicadores”, con el mayor *pagerank* y el mayor grado. En segundo lugar, vuelve

Trabajo Cibermetría

a aparecer la palabra “redes sociales”, que era la de mayor relevancia en el tema anterior. Otras palabras clave importantes son “almétricas”, “bibliotecas públicas”, “indicadores bibliométricos” y “bibliometría”. Como se puede observar, el vocabulario utilizado en los artículos de este tema es mucho más específico que en el tema anterior, y las palabras clave están más interconectadas entre sí.

En esta comunidad sigue habiendo palabras clave que están conectadas solo con su propio grupo de palabras clave perteneciente al mismo artículo, pero en este el número es menor. Aun así, sigue habiendo palabras clave que actúan como puentes, es el caso de “calidad”, “lectura digital” o “libros electrónicos”. Se vuelve a dar el caso en el que hay palabras clave que son muy específicas del tema sobre el que trata el artículo y que podrían ser sustituidas por otras más generales, como es el caso de “libros-app” o “libros-app infantiles”.

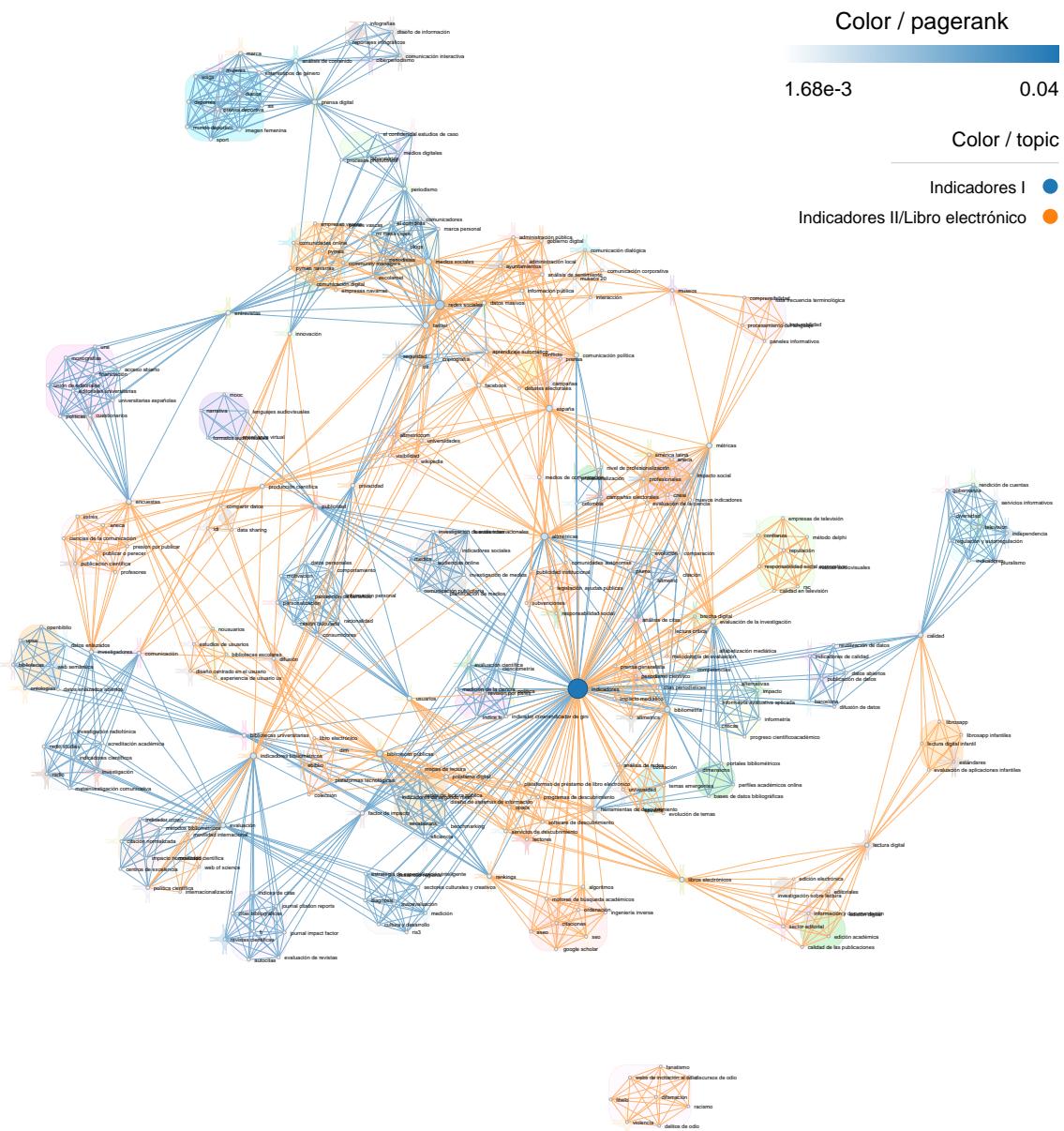


Figura 8: Grafo de palabras clave filtrado por tema

3. Análisis cibermétrico de la web de Pfizer

Para elaborar el informe cibermétrico de la web de Pfizer, primero se han recogido todos los enlaces de la web utilizando la versión 1.19 del *crawler* Apache Nutch. Posteriormente, se ha realizado un análisis de los enlaces recogidos utilizando el programa Gephi en su versión 0.10.1.

3.1. Recogida de enlaces

La recogida de enlaces requiere establecer una configuración adecuada de Nutch para que recopile los enlaces de la web de forma correcta. En primer lugar, se ha creado un fichero con la lista de URL semilla, que son las URL desde las que se empieza a recorrer la web. En este caso, se ha utilizado la URL principal de la web internacional de Pfizer, <https://www.pfizer.com>.

En segundo lugar, se ha editado el fichero `regex-urlfilter.txt` para indicarle a Nutch que solo recopile las URL que pertenezcan al dominio de Pfizer, o a un subdominio de Pfizer. De esta forma, se evita que Nutch recopile enlaces de otras webs que no pertenezcan a Pfizer. También se indica que ignore URL que empiecen por los protocolos `ftp`, `mailto` y `file`, ya que no son enlaces a páginas web. Para evitar posibles bucles se ignoran las URL con longitud superior a 2048 caracteres y las que tengan más de 3 barras hacia delante. También se evitan las URL con caracteres típicos de peticiones HTTP o que correspondan a ciertos tipos de archivos multimedia.

En cuanto a la configuración del propio *crawler* en `nutch-site.xml`, tan solo se han modificado dos parámetros de su configuración por defecto. El parámetro `http.agent.name` se ha cambiado para indicar que el agente que se va a utilizar es `Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/120.0.0.0 Safari/537.3`, este agente se corresponde a la versión más popular de Chrome en Windows 10 en el momento de realizar el análisis, se ha elegido este agente para que la web no detecte que se está utilizando un *crawler* y bloquee la recogida de enlaces. El otro parámetro que se ha modificado es `fetcher.server.delay`, que se ha reducido del valor por defecto de 5 segundos a 0.5 segundos, para que el proceso de recogida de enlaces sea más rápido, pero sin eliminar del todo el tiempo de espera entre peticiones para no saturar el servidor.

Una vez realizada la configuración de Nutch, se ha ejecutado el comando `crawl` para que recopile los enlaces de la web. Al desconocer la profundidad total de la página web y el número de enlaces que se iban a recopilar, primero se comenzó con una profundidad de 4, y se fue aumentando iterativamente hasta llegar a una profundidad de 16, con la que se consiguió recopilar todas las URL de la web. Este proceso en total llevó poco más de una hora.

Para extraer los links se fusionaron los segmentos obtenidos con el comando `nutch mergesegs`, y posteriormente se trajeron los enlaces con el comando `nutch`

readseg. El archivo `dump` obtenido se formateó para que tuviera el formato de un archivo CSV. Se dividió el archivo `dump` en registros según se encontraba el patrón `Recno::`. Para cada registro se ha creado una lista de aristas donde la URL de origen se correspondía con la URL del registro, y la URL de destino se correspondía con cada una de las URL de salida que aparecían en el registro. De esta forma, cada registro contiene una lista de aristas que representan los enlaces de la URL de origen a las URL de destino. Cualquier arista paralela, es decir, que conecte la misma URL de origen con la misma URL de destino, se ha eliminado, ya que se corresponden con enlaces con distintos puntos de anclaje que no aportan información adicional.

Durante el análisis del grafo se detectó que había nodos que apuntaban a enlaces HTTP, los cuales redirigen automáticamente a enlaces HTTPS con la misma dirección. De la misma forma se detectaron enlaces que apuntaban a `pfizer.com` sin ningún subdominio, que redirigen automáticamente a `www.pfizer.com`. Para evitar inconsistencias en el grafo se ha decidido normalizar todos los enlaces a su versión HTTPS y con el subdominio `www` en caso de que no tuvieran ningún subdominio.

Una vez realizado este proceso, ya se disponía de un archivo CSV con todos los enlaces de la web de Pfizer, con un total de 39476 líneas, cada una de ellas correspondiente a una arista, con una URL de origen y una URL de destino. Este archivo se ha utilizado para realizar el análisis de redes con Gephi.

3.2. Análisis de redes

Una vez importado el archivo CSV en Gephi, se ha creado un grafo dirigido, donde cada nodo representa una URL, y cada arista representa un enlace de una URL a otra. Se ha usado el método de *Fruchterman Reingold* para distribuir los nodos en el espacio, y posteriormente se ha utilizado el método *Noverlap* de Gephi para evitar que los nodos se solapen entre sí. El color de los nodos se ha determinado en función de su grado de salida, a mayor número de enlaces salientes, más oscuro es el color del nodo. El tamaño de los nodos se ha determinado en función de su grado de entrada, a mayor número de enlaces entrantes, mayor es el tamaño del nodo. En la figura 9 se puede ver el grafo resultante. La curvatura de las aristas indica la dirección del enlace, al recorrer una arista en el sentido de las agujas del reloj se llega a la URL de destino. Los subdominios identificados en la web de Pfizer son:

- **www.pfizer.com:** Es el subdominio principal de la web de Pfizer, y contiene las páginas que actúan como portada de la web, así como las páginas de las distintas secciones de la web.
- **investors.pfizer.com:** Contiene información relevante a inversores, más relacionada con la parte financiera de Pfizer.
- **insights.pfizer.com:** Contiene información relevante a inversores, pero más relacionada con la parte de investigación y desarrollo de Pfizer.
- **cdn.pfizer.com:** Contiene recursos estáticos de la web, como elementos multimedia o de gran tamaño.

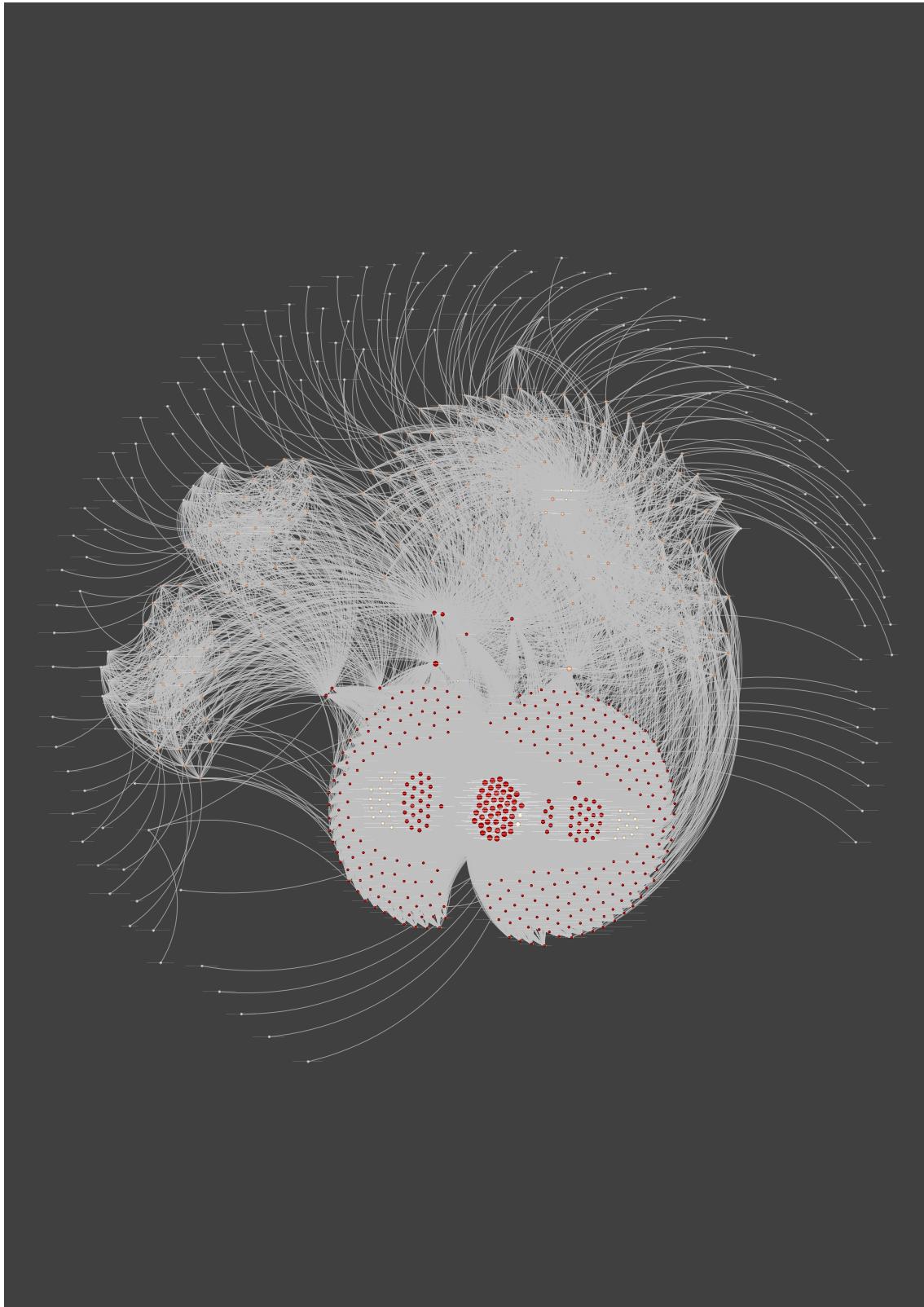


Figura 9: Grafo de enlaces de la web de Pfizer

El grafo final tiene las siguientes métricas, que pintan un cuadro de un sitio web grande y complejo, con distintas secciones o temas que están relativamente bien definidos y conectados internamente, pero con menos conexiones entre estos grupos:

- **Número de nodos:** 775. Este es el número total de páginas en el sitio web.
- **Número de aristas:** 39476. El número total de enlaces entre páginas.
- **Grado medio:** 50.937. Este valor indica que cada URL tiene de media 51 enlaces, salientes y entrantes. En la figura 10 se puede ver la distribución de grados, donde se ve que hay casos atípicos que pueden llegar a alcanzar un grado superior a 450. En la figura 11 se puede ver la distribución de grados de entrada, y en la figura 12 la distribución de grados de salida. Hay más enlaces entrantes que salientes, lo que puede indicar que la web de Pfizer tiene algunas páginas dentro del dominio que actúan como centros de información o que tiene una estructura de árbol.
- **Diámetro:** 32. Este valor indica que la distancia máxima entre dos páginas en el sitio es relativamente grande, lo que sugiere un sitio extenso con posiblemente una gran variedad de contenido.
- **Longitud media del camino más corto:** 7.659. Este valor refuerza la idea de que estamos ante un sitio extenso, mostrando que, en promedio, se requieren varios pasos para navegar de una página a otra arbitraria.
- **Densidad:** 0.066. La densidad es baja, lo que implica que no todas las páginas están densamente interconectadas; esto podría deberse a que el sitio contiene secciones claramente diferenciadas.
- **Modularidad:** 0.288. Apunta a una estructura de comunidad moderadamente fuerte, indicando que el sitio está dividido en subgrupos que probablemente representen diferentes temas o categorías. Estos grupos puede que se correspondan con los subdominios identificados anteriormente.

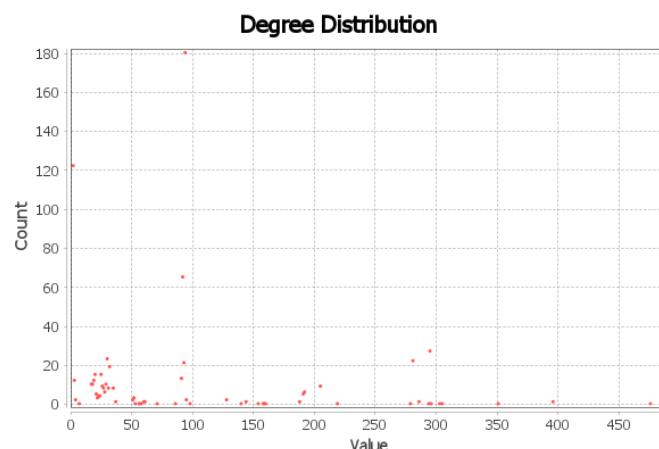


Figura 10: Distribución de grados

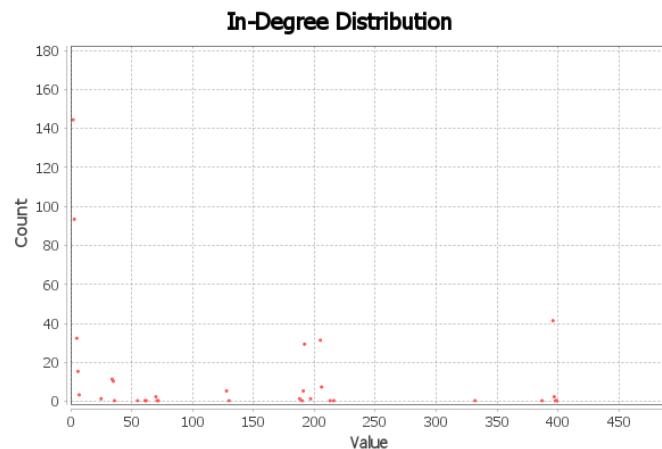


Figura 11: Distribución de grados de entrada

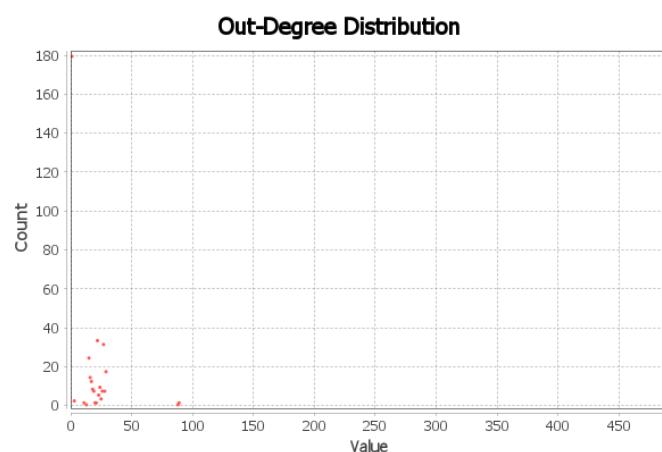


Figura 12: Distribución de grados de salida

En la parte inferior del grafo se puede identificar claramente un grupo de nodos que están muy interconectados entre sí, y que tienen un grado de salida muy alto. En la figura 13 se ha seleccionado la parte de la derecha de este clúster, que se corresponde principalmente a páginas con información para inversores, pertenecientes al subdominio `investors.pfizer.com`. En este gráfico los enlaces de salida de los nodos seleccionados se representan en rojo, los de entrada en azul y cuando se dan ambos casos en amarillo.

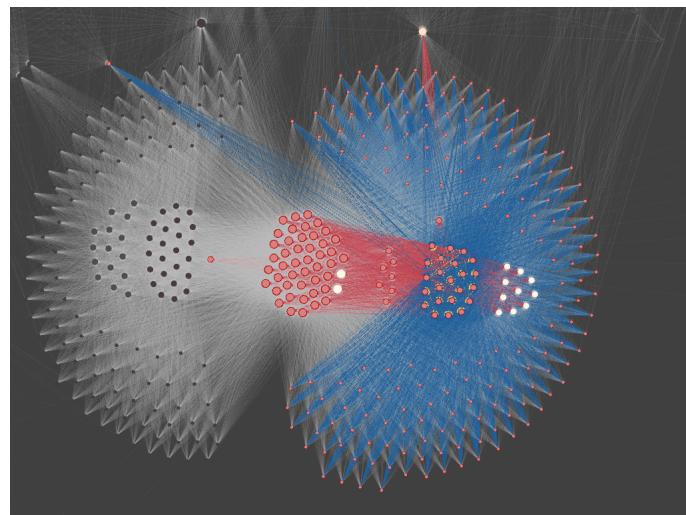


Figura 13: Subgrafo de páginas para inversores

De la misma forma se ha identificado el clúster correspondiente a las páginas de la sección de investigación y desarrollo, pertenecientes al subdominio `insights.pfizer.com`. En la figura 14 se puede ver este grupo. Estas páginas están compuestas principalmente por noticias y secciones paginadas, con un portal central que actúa como portada de la sección.

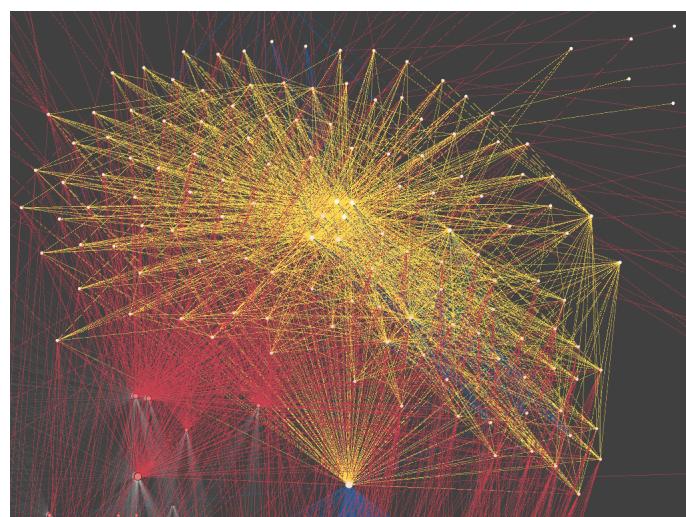


Figura 14: Subgrafo de páginas de investigación y desarrollo

4. Conclusiones

En este informe se ha realizado un análisis de redes de dos conjuntos de datos distintos. En el primer caso se ha realizado un análisis de coautorías y de coincidencia de palabras clave de la revista “El Profesional de la Información”. En el segundo caso se ha realizado un análisis de redes de la web de Pfizer.

En el caso de la revista “El Profesional de la Información”, se ha visto que el grafo de coautorías está formado por varios subgrafos, donde cada subgrafo está formado por autores que han colaborado entre sí, pero que no han colaborado con autores de otros subgrafos. Por lo que se puede determinar que lo más común es que los autores colaboren con un número reducido de otros autores. En el caso del grafo de coincidencia de palabras clave, se ha realizado un estudio sobre las palabras que aparecen en los artículos de dos temas distintos. Se ha visto que en ambos casos hay palabras clave que actúan como puentes conectando a grupos de palabras clave que solo se utilizan para un artículo. Lo que podría indicar que estas palabras clave podrían ser sustituidas por otras más generales que relacionen mejor los artículos.

En el caso de la web de Pfizer, se ha realizado un análisis cibermétrico de la web, donde se ha visto que la web está formada por varios subdominios, que se corresponden con distintas secciones de la web. Se ha visto que estos subdominios se corresponden con comunidades en el grafo. También se han analizado las distintas medidas de la red y se ha visto que la web de Pfizer es un sitio extenso, con distintas secciones o temas que están relativamente bien definidos y conectados internamente, pero con menos conexiones entre estos grupos.