



UNIVERSIDAD DE GRANADA

Redes y Sistemas Complejos

*Práctica 4:
Caso práctico de análisis y
evaluación de redes*

Javier León Palomares

15 de diciembre de 2017

Índice

1. Introducción y definición de objetivos	3
2. Descripción de la red	3
3. Análisis global inicial	5
4. Propiedades de la red	8
4.1. ¿Es una red libre de escala?	11
4.2. ¿Es una red de mundos pequeños?	11
5. Medidas de redes sociales	12
5.1. Centralidad de grado	12
5.2. Centralidad de grado ponderado	13
5.3. Centralidad de intermediación	14
5.4. Centralidad de cercanía	15
5.5. Centralidad de vector propio	16
6. Análisis de comunidades	18
6.1. Comunidades incluidas en los datos originales	18
6.2. Detección con el método de <i>Louvain</i>	20
7. Conclusiones	25

Índice de figuras

1.	Red original de coautoría dibujada con <i>OpenOrd</i>	4
2.	Red original de coautoría dibujada con <i>Yifan-Hu</i>	4
3.	Distribución de coeficientes de <i>clustering</i>	5
4.	Red original con los nodos de mayor coeficiente de <i>clustering</i> destacados en verde intenso.	6
5.	Red que contiene sólo los nodos de coeficiente de <i>clustering</i> igual a 1	6
6.	Distribución de tamaños de las componentes conexas.	7
7.	Componente gigante de la red.	7
8.	Distribución de grados.	8
9.	Distribución de coeficientes de <i>clustering</i> de la red completa.	9
10.	Distribución de coeficientes de <i>clustering</i> de la componente gigante.	9
11.	Distribución de distancias.	10
12.	Coeficientes de <i>clustering</i> frente a grados.	10
13.	Red original con los nodos diferenciados por su grado.	12
14.	Red original con los nodos diferenciados por su grado ponderado.	13
15.	Red original con los nodos diferenciados por su intermediación.	14
16.	Componente gigante con los nodos diferenciados por su centralidad de cercanía.	15
17.	Sección de la componente gigante con sus etiquetas.	16
18.	Nodos de la componente gigante según grado (tamaño) y valor de centralidad de vector propio (color).	17
19.	<i>Close-up</i> de la componente gigante.	17
20.	Comunidades según la distribución ya incluida en la red.	19
21.	Visualizaciones de las 4 comunidades elegidas.	19
22.	Comunidades según el método de <i>Louvain</i> con Resolución = 1.0.	20
23.	Visualizaciones de las 2 nuevas comunidades.	20
24.	Comunidades según el método de <i>Louvain</i> con Resolución = 1.0.	21
25.	Detalle de la componente gigante en una región de nodos importantes.	21
26.	Comunidades grandes según el método de <i>Louvain</i> con Resolución = 10.0.	22
27.	Las dos comunidades más importantes de la ejecución con Resolución = 10.0.	22
28.	Comunidades grandes según el método de <i>Louvain</i> con Resolución = 0.5.	23
29.	<i>Close-up</i> del centro de la red para comunidades con Resolución = 0.5.	23
30.	Las 5 comunidades escogidas.	24

1. Introducción y definición de objetivos

Con la motivación del estudio de las redes complejas en esta asignatura, surge una pregunta acerca del propio campo de investigación que puede ser resuelta mediante las técnicas aprendidas: ¿cuáles son los autores más relevantes en su bibliografía?

Con esta pregunta en mente, utilizaremos una red de coautoría en la que los nodos serán los autores y los enlaces representarán colaboraciones entre ellos en publicaciones oficiales.

Asimismo, pues esta pregunta puede saber a poco, trataremos de descubrir no sólo la importancia de unas pocas personas sino también autores o grupos de autores que contribuyan a otros campos, ya que la ciencia de redes tiene numerosas aplicaciones. Para hacer esto, destacaremos nodos o comunidades de nodos que llamen la atención y buscaremos información acerca de ellos.

2. Descripción de la red

El conjunto de datos utilizado fue compilado en 2006 por *Mark Newman*, investigador de este campo del conocimiento, a partir de la información disponible en el momento de su creación. Existe una red inmediata asociada a estos datos.

Los distintos elementos y características de dicha red son:

- Nodos: los distintos autores de la bibliografía.
- Enlaces: colaboraciones entre investigadores (por ello, es una red no dirigida y ponderada).
- Pesos: número de colaboraciones entre los investigadores (más peso, mayor colaboración).
- Comunidades: aunque trataremos de detectar comunidades mediante las herramientas a nuestra disposición, el conjunto de datos parece contener ya información acerca de grupos de colaboradores, y los analizaremos también.

Por tanto, los autores que más colaboren entre sí tenderán a estar más cerca en las visualizaciones; incluso podríamos aventurar que la componente gigante tendrá entre sus nodos a los investigadores más influyentes ya que éstos suelen estar muy solicitados a la hora de publicar.

Para concluir este apartado, en la siguiente página veremos la red original representada con los algoritmos *Yifan-Hu* y *OpenOrd* (a mayor tamaño de nodo, mayor grado). Esto nos permite, entre otras cosas, decidir si es necesario realizar algún tipo de filtrado o poda para interpretar mejor la red.

Como podremos observar, en principio tenemos bastantes nodos potencialmente interesantes a juzgar por su grado; basándonos en el razonamiento intuitivo que acabamos de usar, consideraremos descartar autores con pocas colaboraciones porque los grandes investigadores suelen ser *hubs* que reúnen a otros nombres de similar o menor relevancia.

No obstante, esto no significa que vayamos a descartar a todas las personas que no tengan conexiones con las más importantes. Sin ir más lejos, ya aparecen algunos subconjuntos de nodos interesantes que podrían quedar fuera de la componente gigante.

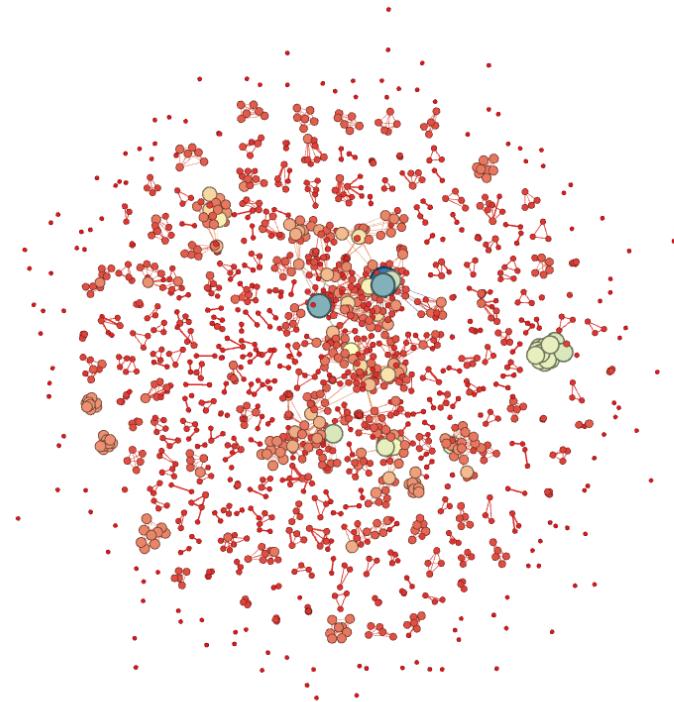


Figura 1: Red original de coautoría dibujada con *OpenOrd*.

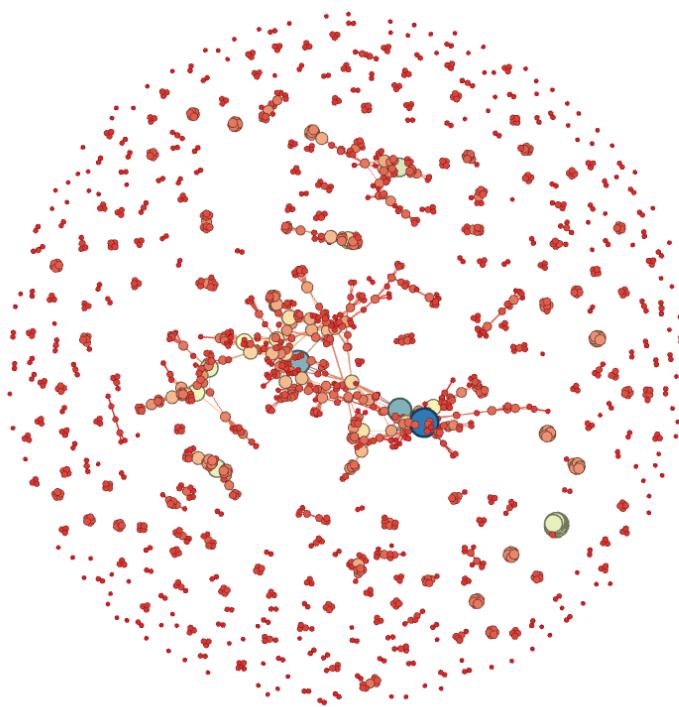


Figura 2: Red original de coautoría dibujada con *Yifan-Hu*.

3. Análisis global inicial

Las primeras medidas que vamos a obtener son las relativas a la red como conjunto. Cabe mencionar que, si bien ya podríamos filtrar nodos, en principio no será necesario hacerlo porque la red tiene un tamaño asumible para hacer cálculos (no así para representarla de forma clara).

A continuación tenemos la tabla con las medidas de nuestra red y las correspondientes, si procede, de la red aleatoria equivalente:

Medida	Valor red original	Valor red aleatoria
N	1589	1589
L	2742	2742
D	0.00217	0.00217
$\langle k \rangle$	3.451	3.451
d_{max}	17	-
$\langle d \rangle$	5.823	5.950
$\langle C \rangle$	0.878	0.00217

Donde:

- $\langle k_{aleatoria} \rangle = \frac{2L}{N}$
- $\langle d_{aleatoria} \rangle = \frac{\log(N)}{\log(\langle k \rangle)}$
- $\langle C_{aleatoria} \rangle = \frac{\langle k_{aleatoria} \rangle}{N}$

Vemos que es una red bastante poco densa pero que tiene una distancia media pequeña; esto puede ser interesante. Pero lo que más nos llama la atención es el coeficiente de *clustering* tan alto. Las redes reales son demasiado dispersas como para que este fenómeno sea consecuencia de una gran densidad, y ésta no es una excepción. Así pues, nos disponemos a encontrar la explicación mediante la distribución de coeficientes de *clustering* y la visualización de la red:

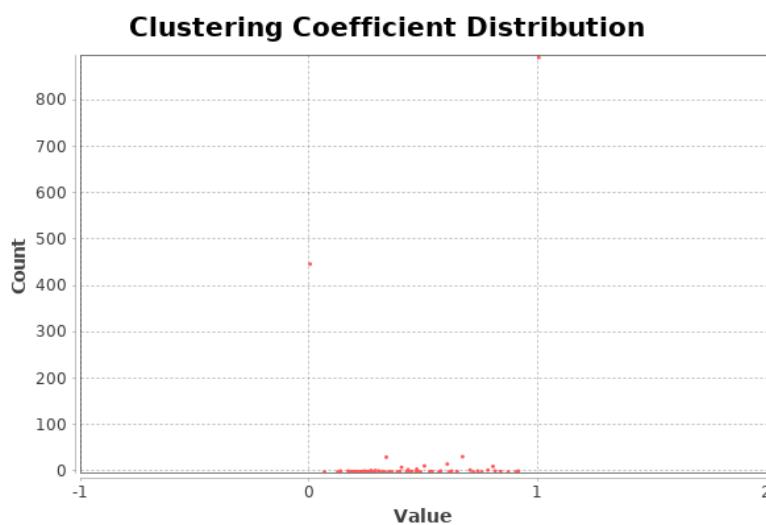


Figura 3: Distribución de coeficientes de *clustering*.

La distribución anterior contiene muchos valores en los extremos, y sobre todo en el 1. Si vemos ahora la red destacando los nodos con valor de coeficiente igual a 1, entenderemos qué está ocurriendo:



Figura 4: Red original con los nodos de mayor coeficiente de *clustering* destacados en verde intenso.

Adicionalmente, dejemos sólo los que tienen coeficiente igual a 1:

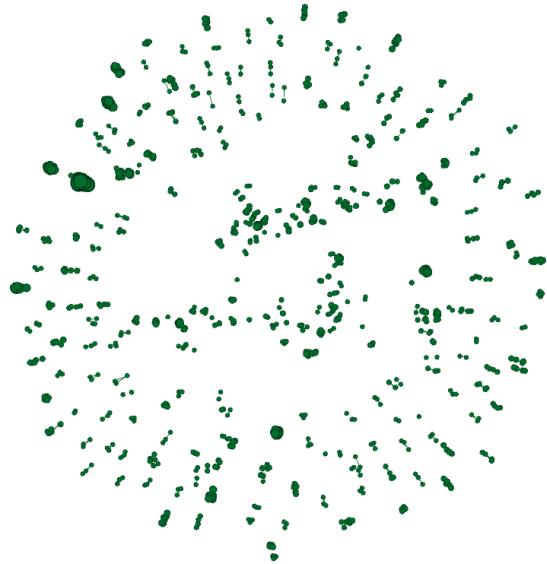


Figura 5: Red que contiene sólo los nodos de coeficiente de *clustering* igual a 1.

Los investigadores que colaboran en un mismo artículo tienen conexiones entre todos ellos en la red de coautoría. Esto significa que para cada uno de esos grupos de personas se crea un **clique**, que siempre tiene un coeficiente de *clustering* local de 1. Si no escriben más artículos o escriben con los mismos autores, esos cliques se mantienen, y esto ocurriendo muchas veces es exactamente lo que causa ese valor medio tan anómalo.

Una vez hecho esto, es el momento de continuar el análisis con la detección de las componentes conexas. La figura que acabamos de mostrar nos dice que la componente gigante no será muy grande. Comprobémoslo mediante la distribución de tamaños y una visualización:

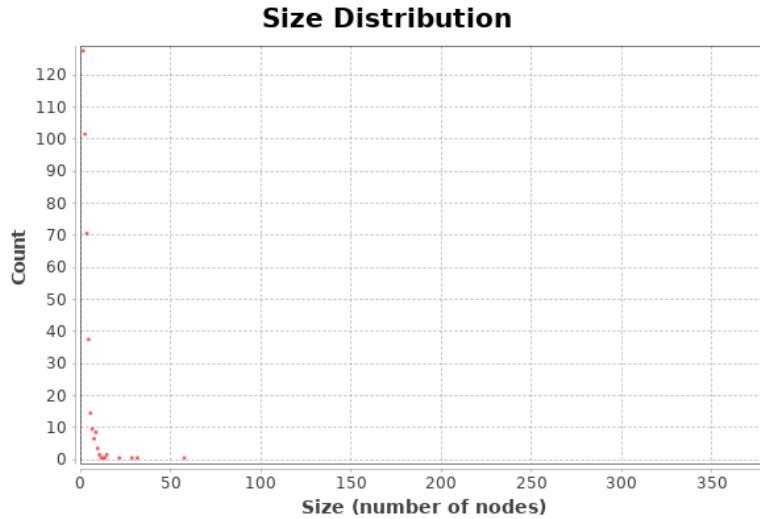


Figura 6: Distribución de tamaños de las componentes conexas.

En efecto, la distribución nos indica que existen muchos nodos aislados y muchos grupos pequeños de autores (estos últimos son los que causan el alto *clustering*). Si avanzamos lo suficiente en el eje x podemos ver el punto que corresponde a la componente gigante, que tiene un tamaño de 379 nodos con 914 enlaces entre ellos; estos números representan, respectivamente, el 23.85% y el 33.33% de los originales. En total hay 396 componentes conexas.

Si representamos la componente gigante mediante el algoritmo *Yifan-Hu*, el resultado es el siguiente:

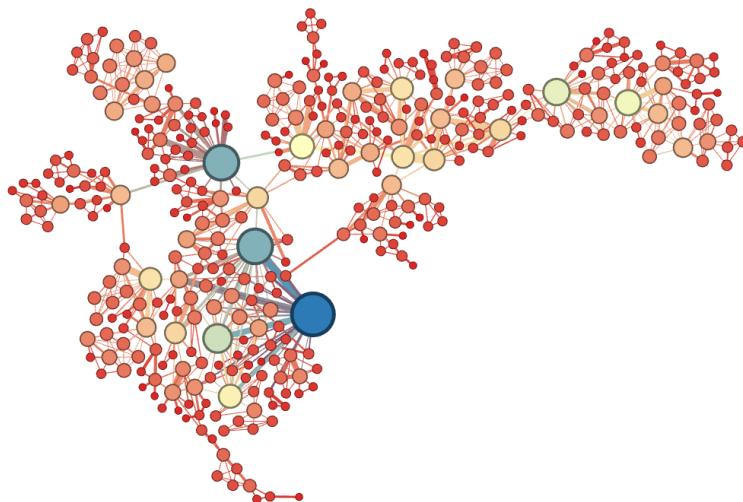


Figura 7: Componente gigante de la red.

Por completitud, estas son las medidas para la componente gigante:

Medida	Valor componente gigante	Valor red aleatoria	Valor red original
N	379	379	1589
L	914	914	2742
D	0.0127	0.0127	0.00217
$\langle k \rangle$	4.823	4.823	3.451
d_{max}	17	-	17
$\langle d \rangle$	6.042	3.773	5.823
$\langle C \rangle$	0.798	0.010	0.878

Al quedarnos sólo con la componente gigante, hemos eliminado muchos de los pequeños grupos autocontenidos. Esto ha hecho que el coeficiente de *clustering* disminuya bastante (aunque no hasta cifras más comunes porque el efecto de los grupos de investigación sigue presente). También ha aumentado un poco el grado medio aunque, como veremos más adelante, esa medida podría ser poco informativa.

4. Propiedades de la red

Empleando medidas básicas y algunas gráficas de apoyo somos capaces de identificar importantes propiedades que subyacen a la red. En particular, podemos saber si cumple las propiedades de las redes libres de escala o de mundos pequeños, descartando en el proceso que se trate de una red de tipo aleatorio.

Comenzamos analizando la distribución de grados de los nodos. Si es una red libre de escala, se caracterizará por seguir una **ley de la potencia**:

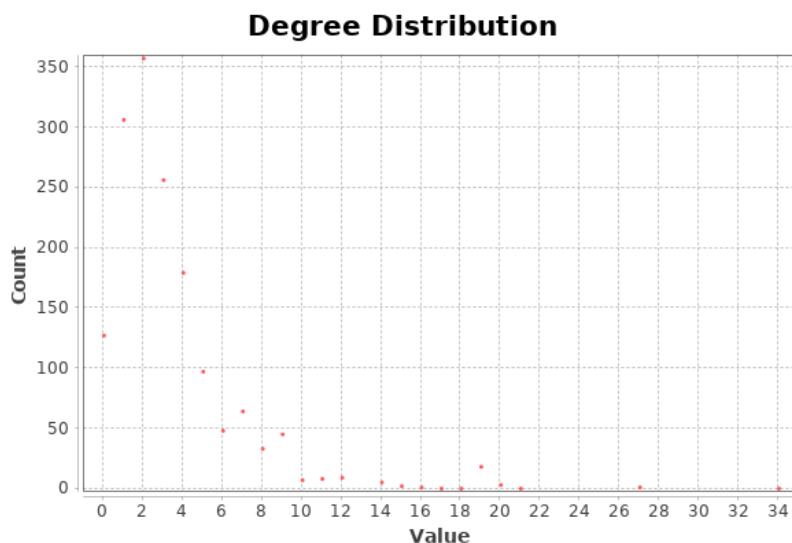


Figura 8: Distribución de grados.

A la vista de esta gráfica la red puede ser libre de escala, ya que en efecto posee una característica curva que muestra muchos nodos de grado bajo y algunos con grado alto, los *hubs*; por contra, en una red aleatoria tendríamos una campana de Gauss que anularía la aparición de *hubs* tan significativos.

La siguiente distribución que veremos es la de coeficientes de *clustering*. Ya la hemos presentado antes para encontrar la causa del alto coeficiente de *clustering*, pero nos conviene tenerla cerca ya que, junto con las distancias, nos va a permitir decidir después si es una red de mundos pequeños.

Así pues, a continuación tenemos las distribuciones tanto de la red original como de la componente gigante. El mostrar ambas se debe a que las redes de mundos pequeños se caracterizan por tener unas distancias pequeñas a pesar de que la mayoría de los nodos no son vecinos de muchos otros; esto ocurre porque hay un alto *clustering* local y muchos *hubs* interconectando unos grupos de nodos con otros. No obstante, en la red original había muchos cliques aislados, lo cual podría hacernos relacionar erróneamente las bajas distancias con el alto *clustering*. De aquí que analicemos también la componente gigante para ver si en una partición conexa de la red siguen teniendo sentido las conclusiones.

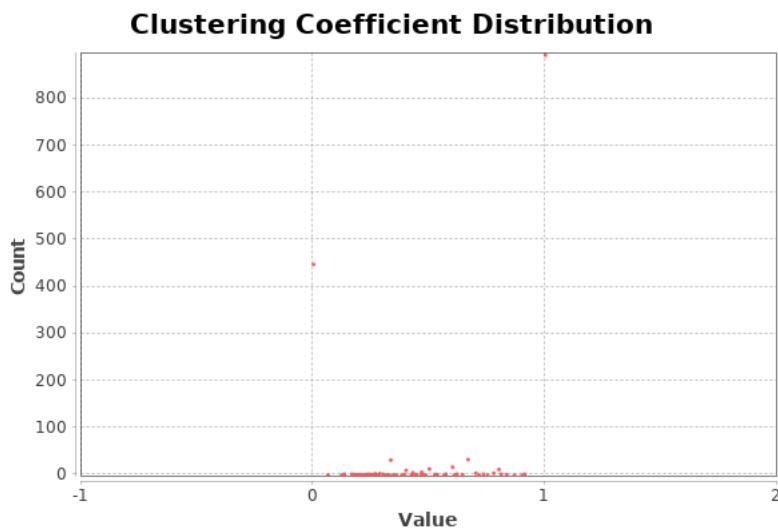


Figura 9: Distribución de coeficientes de *clustering* de la red completa.

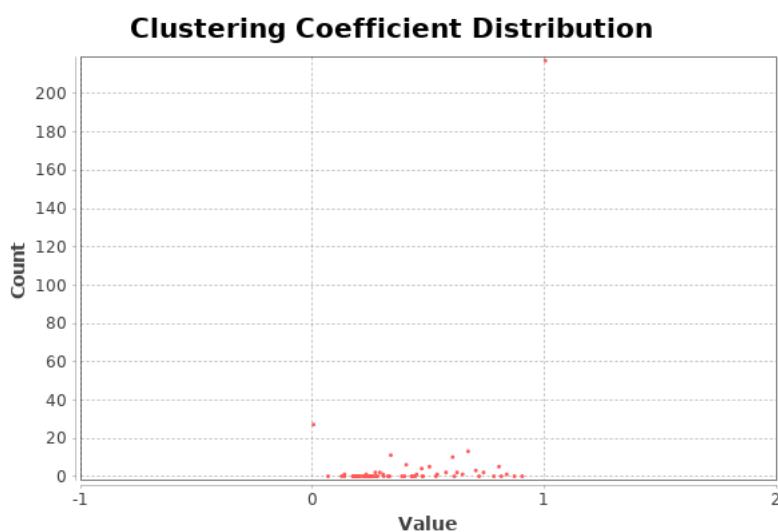


Figura 10: Distribución de coeficientes de *clustering* de la componente gigante.

Vemos que las gráficas mantienen una gran similitud, lo que nos da confianza de cara al análisis que faremos. La diferencia entre ellas es principalmente la desaparición de la mayoría de nodos con coeficiente 0. Los que quedan son autores que sólo han colaborado con otro autor.

Pasamos a ver la última gráfica que se pedía en esta sección, la de distribución de distancias. Debido a que Gephi no ofrece esta funcionalidad, ha sido realizada con una librería externa.

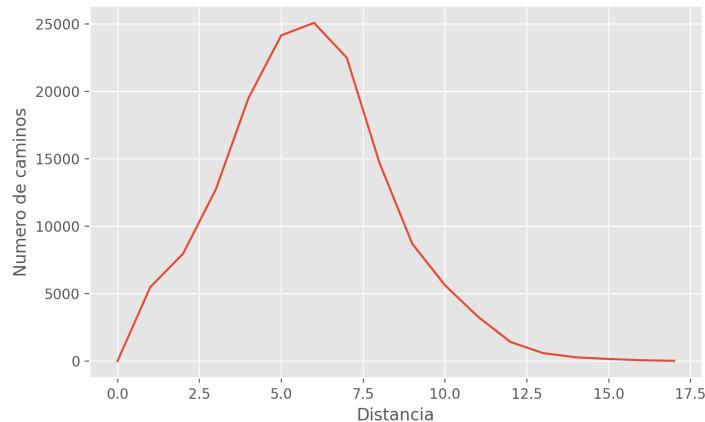


Figura 11: Distribución de distancias.

La tendencia que se observa es la de una mayor proporción de los caminos en la primera parte del eje x, lo que implica menores distancias. Además, mirando su escala nos damos cuenta de que los caminos son en general bastante cortos para las dimensiones y la densidad de la red.

Antes de deducir sus propiedades, vamos a mostrar por completitud una figura extra que muestra la correlación entre los coeficientes de *clustering* y los grados de los nodos:

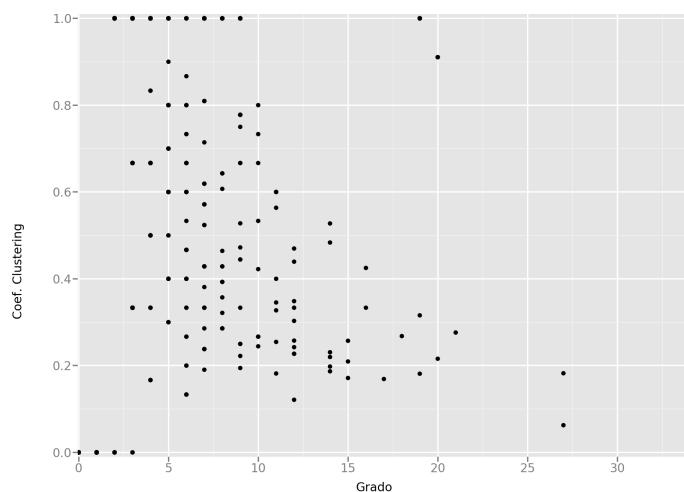


Figura 12: Coeficientes de *clustering* frente a grados.

Esta gráfica nos reafirma desde otro punto de vista dos convicciones: que la red pueda ser de mundos pequeños, ya que la alternativa es que la combinación de alto *clustering* y bajas distancias provenga de una red muy conectada, algo incompatible con un *clustering* decreciente con el grado; y que estemos hablando de una red libre de escala, porque al seguir esta curva otra ley de la potencia, está implicando la existencia de subgrafos densos conectados por grandes *hubs*.

4.1. ¿Es una red libre de escala?

Siguiendo el hilo del último párrafo, la distribución de grados, como ya habíamos dicho, parece encajar con una ley de la potencia. Si fuese aleatoria, tendría un pico de frecuencia en el centro y se dispersaría hacia los lados, cosa que no pasa. Además, si consideramos el grado medio obtenido vemos que no es representativo de lo que de verdad ocurre: se sitúa muy a la izquierda en la gráfica de la distribución, mientras que la sucesión de grados se alarga demasiado en el eje x como para que la media albergue significado. Por tanto, sí, es una red libre de escala que presenta unas características homólogas a otras redes del mismo tipo como las de actores o las de colaboraciones y citas en el ámbito científico general.

Su exponente es $\gamma = 2.032$, calculado de nuevo de forma externa a Gephi. Las implicaciones inmediatas de este valor son dos: en primer lugar, el encontrarse en el intervalo $2 < \gamma < 3$ hace que el comportamiento libre de escala tenga relevancia (no sería una red aleatoria o sería difícil distinguirla de una); en segundo lugar, la teoría nos dice que también exhibe un comportamiento de mundo ultra-pequeño. Con esto último damos paso al próximo apartado.

4.2. ¿Es una red de mundos pequeños?

Para decidirlo de forma aislada nos tenemos que fijar en el *clustering* y en las distancias. Si el primero es alto y las segundas son pequeñas, la red es de mundos pequeños. Ya hemos mostrado mediante las medias y las gráficas que se cumplen ambas condiciones.

Concluimos el apartado anterior sugiriendo que estábamos ante un caso de mundos ultra-pequeños. En los mundos pequeños normales, la distancia media crece logarítmicamente respecto al número de nodos de la red. En los mundos ultra-pequeños, la distancia crece de manera doblemente logarítmica. El intervalo $2 < \gamma < 3$ sugiere un fenómeno como el último, aunque quizás por estar el valor tan cerca del límite inferior los cálculos no se aproximan a la realidad (podría haber necesidad de una constante multiplicativa); en su lugar, ya que la propiedad de mundos pequeños a secas sí se acerca ($\langle d \rangle = 5.823$, $\log(1589) = 7.370$), nos decantamos por ella.

5. Medidas de redes sociales

La red de coautoría se puede considerar una red social porque modela la interacción de investigadores a la hora de publicar. Por este motivo, vamos a hacer un análisis básico para intentar reunir más información sobre la importancia de los autores en este contexto utilizando ya sus nombres.

5.1. Centralidad de grado

Comenzamos viendo formalmente la **centralidad de grado**, aunque ya la habíamos utilizado sin nombrarla al pintar la red por primera vez. Esta medida nos revela la importancia de cada autor según su entorno inmediato. Puesto que la red es ponderada, hablaremos tanto de grados a secas como de grados ponderados.

En primer lugar, los autores más importantes según su grado no ponderado son:

Autor	Centralidad de grado
Albert-László Barabási	34
Mark Newman	27
Hawoong Jeong	27
Zoltán Oltvai	21
Malcolm Young	20

Era de esperar que uno de los mayores nombres en la ciencia de redes en los últimos años ocupase el primer lugar en alguna medida de centralidad. Empezamos, pues, teniendo a *Barabási* como el autor con mayor grado simple. Aunque analizaremos más adelante algunos de estos autores, vamos a ver la red con los nodos de mayor grado destacados en color y tamaño porque hay información que no queda reflejada en la tabla:

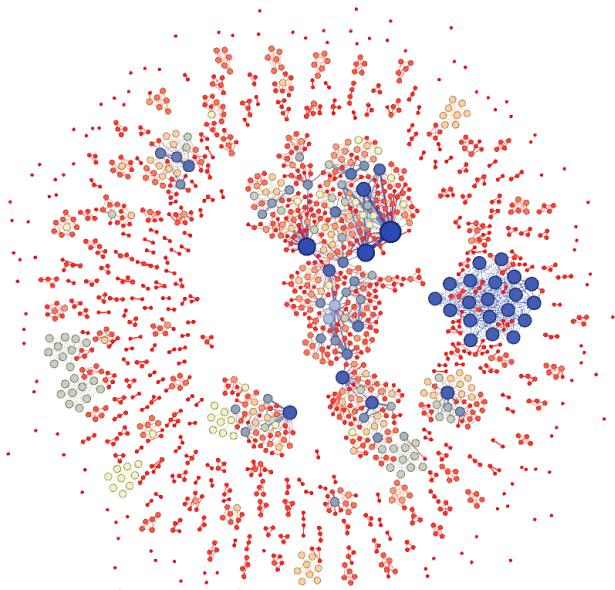


Figura 13: Red original con los nodos diferenciados por su grado.

En general parece no existir nada extraño, pero destacan algunos grupos de nodos cuyo grado parece deberse simplemente a haber publicado algún artículo entre un número inusual

de personas. Estando atentos, nos damos cuenta de que uno de estos grupos (se ve en la mitad derecha de la imagen) distorsiona la percepción de la importancia de los autores según la medida que estamos utilizando: son prácticamente un clique cuyos elementos se refuerzan su valor entre sí pero que **no interactúan con el resto de la red**. Si la tabla hubiese sido más extensa, estos nodos ocuparían muchas de las primeras posiciones porque el grado de todos ellos está en torno a 20; esto nos habría llevado a equivocaciones al tratar de identificar a otras personas destacadas pero no tan prolíficas como *Barabási* y similares.

5.2. Centralidad de grado ponderado

Con el objetivo de ser más fiables se ha propuesto introducir la medida del **grado ponderado**, que analizaremos tras este párrafo. Pensándolo un poco, parece lógico que los autores **relevantes y activos de verdad** no sólo no tengan una única colaboración, sino que incluso se asocien múltiples veces con otros autores con los que hayan logrado buenos resultados. Por tanto, queremos favorecer enlaces fuertes. Veamos la nueva tabla:

Autor	Centralidad de grado ponderado
Albert-László Barabási	30
Mark Newman	23
Hawoong Jeong	18
Romualdo Pastor-Satorras	17
Alessandro Vespignani	15

Un apunte es que los pesos de los enlaces no corresponden exactamente al número de colaboraciones entre dos autores, pero tienen una lógica similar. Dicho esto, vemos que los primeros puestos se mantienen sin cambios. Se puede concluir que son candidatos para más adelante, cuando respondamos a la pregunta original del trabajo. Los otros dos nombres que aparecen también se tendrán en cuenta. Pero lo que más nos interesa ahora mismo es ver si esta modificación en la medida ha tenido el efecto deseado:

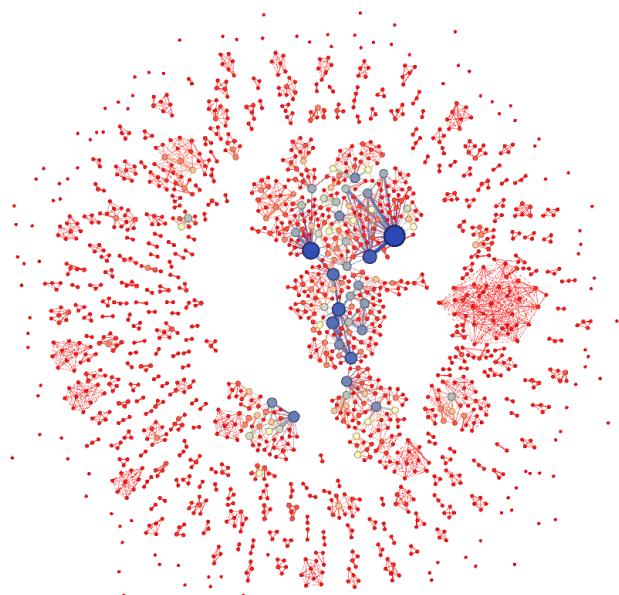


Figura 14: Red original con los nodos diferenciados por su grado ponderado.

Comparando con la imagen anterior, los nodos en cliques han perdido su inflada importancia. Es muy posible que las publicaciones de dichos grupos sean muy citadas como hitos puntuales, pero nuestro objetivo principal no es encontrar artículos famosos. Por consiguiente, esta medida es más adecuada aquí que la de grados simples.

5.3. Centralidad de intermediación

Continuando nuestro análisis de medidas de redes sociales, la siguiente es la **intermediación**. La intermediación favorece a los nodos por los que pasan muchos caminos mínimos entre otros nodos. Aquí están los primeros puestos:

Autor	Centralidad de intermediación
Mark Newman	28300
Romualdo Pastor-Satorras	24592
Yamir Moreno	20379
Ricard Solé	19249
Stefano Boccaletti	18200

Los *hubs* tienen facilidad para obtener valores altos de intermediación al ser eso lo que hacen a alto nivel cuando conectan diversas secciones de las redes. Por tanto, si dibujamos la red de coautoría según este criterio, nos queda algo no muy distinto a lo que acabamos de ver:

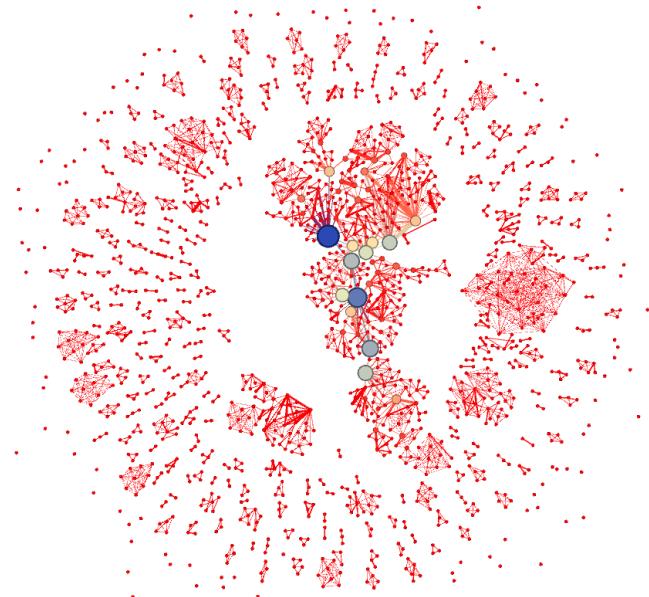


Figura 15: Red original con los nodos diferenciados por su intermediación.

Los mayores valores aparecen en la componente gigante, como es lógico: en las componentes conexas más pequeñas no se da la posibilidad de que muchos caminos mínimos pasen por sus nodos, así que encontramos a autores cercanos a *Barabási* cumpliendo la función de puentes o “peajes” entre comunidades. Según podríamos interpretar, estas personas tendrían una mayor facilidad para controlar el contacto entre diferentes grupos de investigación o a individuos pertenecientes a los mismos; a priori, sin conocer sus ámbitos de trabajo, las personas adecuadas podrían incluso favorecer grupos interdisciplinares.

5.4. Centralidad de cercanía

Seguimos con la penúltima medida de centralidad: la **cercanía**. La cercanía valora la capacidad de estar bien situado en la red, de forma que las distancias a los demás nodos sean cortas. Para este cálculo vamos a tener que utilizar sólo la componente gigante porque Gephi toma en consideración únicamente los nodos alcanzables desde un determinado nodo, y esto resulta en muchos cliques teniendo valores máximos que no nos dicen nada útil.

La tabla con los mejores autores por cercanía (normalizada) es la siguiente:

Autor	Centralidad de cercanía
Mark Newman	0.256
Ricard Solé	0.249
Romualdo Pastor-Satorras	0.247
Petter Holme	0.243
Guido Caldarelli	0.233

Como se observa en la siguiente figura, los nodos con mayor valor de cercanía tienden de forma natural a situarse en el centro de las visualizaciones, sobre todo en las basadas en fuerzas como *Yifan-Hu*:

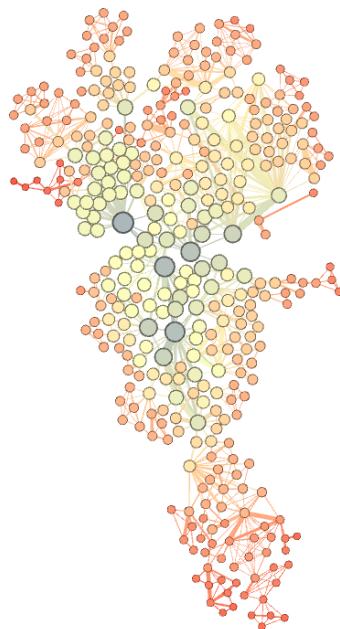


Figura 16: Componente gigante con los nodos diferenciados por su centralidad de cercanía.

Una posible interpretación ahora es que los autores más importantes según este criterio tienen más fácil contactar a otros para solicitar colaboración, ayuda o información, por ejemplo.

Los colores suaves que tienen los nodos ayudan a notar que hay diferencias en los valores pero que no existen grandes contrastes, siendo 0.09 el mínimo. Gracias a los mismos *hubs* que son relativamente centrales en esta medida, el resto de los nodos también pueden tener algo de ese valor.

Para terminar, aprovechando que en esta ocasión no tenemos tantos nodos circundantes, podemos hacer un poco de zoom para ver algunos nombres sin perder mucho contexto:

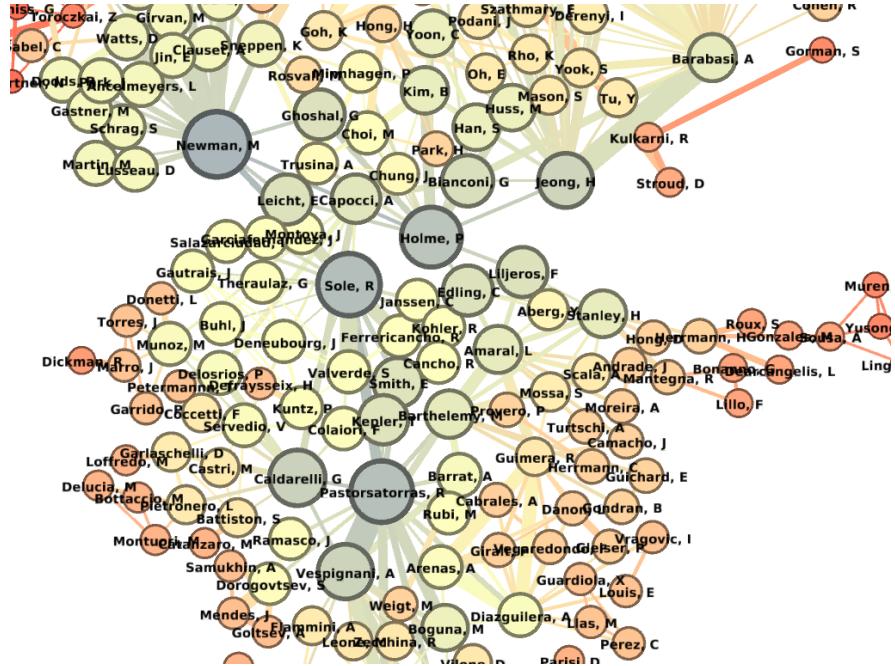


Figura 17: Sección de la componente gigante con sus etiquetas.

Destacamos a *Mark Newman*, que ha aparecido en los primeros puestos de todos los criterios hasta ahora. No sólo ha realizado un alto número de colaboraciones y ha repetido algunas de ellas varias veces, sino que parece tener un papel importante a la hora de vincular conjuntos de autores.

5.5. Centralidad de vector propio

Finalmente, la última medida que calcularemos es la **centralidad (normalizada) de vector propio**, que elabora más el concepto de la centralidad de grado considerando que cada enlace incidente en un nodo varía en importancia según la centralidad del nodo del otro extremo. De nuevo, vamos a utilizar la componente gigante para evitar la interferencia de nodos como los que daban problemas en la centralidad de grado simple. Los 5 mejores autores en este caso son:

Autor	Centralidad de vector propio
Albert-László Barabási	1
Hawoong Jeong	0.913
Zoltán Oltvai	0.817
Tamas Vicsek	0.611
Erzsebet Ravasz	0.550

Antes de comentar los resultados, vamos a mostrar dos imágenes que nos ayuden a comprender los matices de esta medida. La primera es una vista general de la componente

gigante, contraponiendo el tamaño (grado de los nodos) y el valor de centralidad de vector propio (rojo significa menos, azul significa más, colores claros indican valores intermedios):

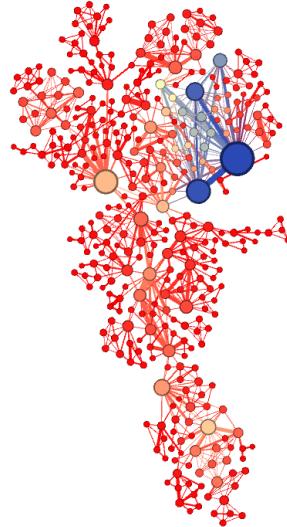


Figura 18: Nodos de la componente gigante según grado (tamaño) y valor de centralidad de vector propio (color).

Comprobamos que no es posible identificar una correlación entre el grado y este criterio de centralidad, ya que a primera vista podemos encontrar nodos con bajo grado que son más centrales que otros nodos con grado mayor. En concreto, vamos a acercarnos a la zona de los nodos azules para poner nombres a algunos ejemplos:

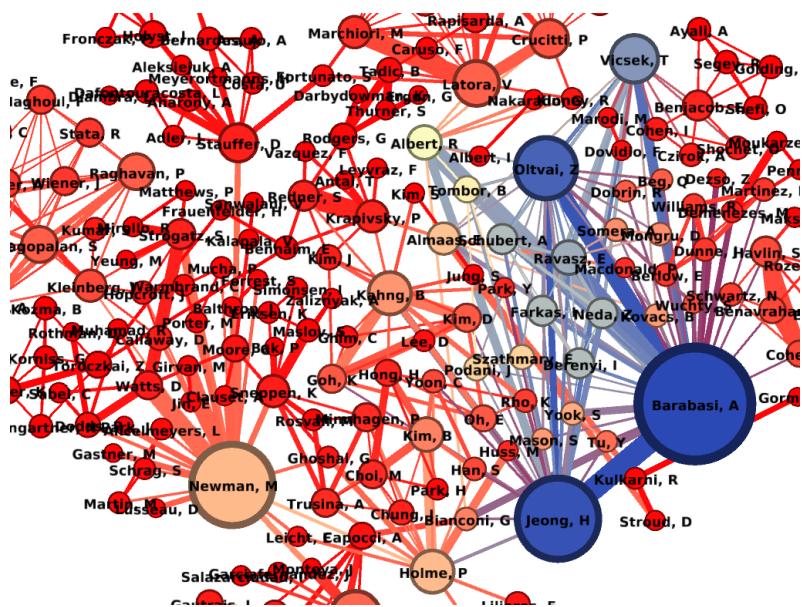


Figura 19: Close-up de la componente gigante.

La relevancia de *Barabási* viene dada no sólo por su gran conectividad, sino también por los enlaces que comparte con otros autores casi tan centrales como él (*Hawoong* y *Oltvai*).

Pero el foco lo vamos a situar sobre *Mark Newman* y *Réka Albert*; el primero es un autor altamente conectado pero poco central, y lo sabemos porque está en un entorno diferente al de los otros investigadores mencionados; la segunda es una autora con menos conexiones pero cuya centralidad es mayor por su implicación directa con *Jeong*, *Oltvai*, y especialmente con *Barabási* en los dos trabajos de mayor trascendencia de la ciencia de redes moderna, como veremos luego.

Utilizando la información que nos han dado estas medidas de centralidad, cuando sea el momento buscaremos a los autores aparentemente más notables de la literatura de la ciencia de redes y veremos si el conjunto de datos empleado nos permitía responder a la pregunta original. Sin embargo, aún queda trabajo por hacer, trabajo al que dedicaremos la siguiente sección: detección de comunidades. Las comunidades quizás nos proporcionen más referencias que buscar, ya que por la inmensa cantidad de autores era necesario un severo recorte en el número de entradas de las tablas.

6. Análisis de comunidades

Hemos tenido la ocasión de constatar en varias oportunidades que muchos autores se agrupan de diversas formas y cumplen diversos roles en la estructura de la red. Vamos a tratar de aprovechar esos patrones para identificar más elementos interesantes. Asimismo, cabe destacar que, por la naturaleza de esta sección, ignoraremos los nodos aislados.

Un detalle a tener en cuenta que ya se comentó en la descripción inicial de la red es que los creadores de este conjunto de datos parecen haber incluido su propia evaluación de las comunidades interesantes (atributo ComponentID). Como podemos considerarlo conocimiento experto, haremos un análisis separado tanto de este conjunto de comunidades como de algunos encontrados por el método de *Louvain* que implementa Gephi, si arrojase resultados con sentido.

Antes de comenzar, de nuevo debemos recordar que el volumen de datos es inmanejable a menos que enfoquemos nuestro trabajo, así que limitaremos nuestra elección final de comunidades a un máximo de 5. Como heurística, daremos prioridad a aquéllas que se repitan aproximadamente en ambos sitios.

6.1. Comunidades incluidas en los datos originales

La partición original consta de 8 comunidades principales representadas con colores distintos, siendo las demás descartadas mediante un gris común a todas ellas. Algo llamativo lo encontramos en la comunidad más grande, que engloba a toda la componente gigante; esto deja a un lado por el momento las subcomunidades que hubiese en ella, y a su vez nos indica que algunos de esos grupos aislados que excluimos en la sección anterior por ser tan molestos para los cálculos de centralidad son comunidades con potencial. Gracias a este hecho quizás podríamos conseguir algunos datos más que analizar.

Dado que a priori no tenemos conocimiento suficiente para determinar qué comunidades son más interesantes, vamos a elegir algunas a mano y vamos a acercar la vista para conocer sus autores y así poder saber si las elige también el método de *Louvain*. Para hacer esto necesitaremos ver la red, así que la tenemos en la página siguiente.

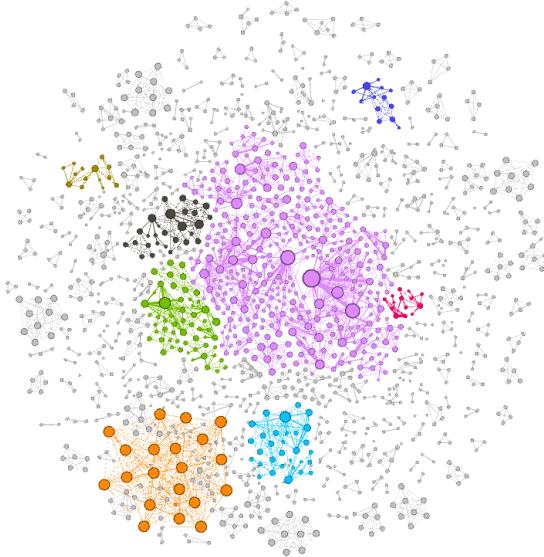


Figura 20: Comunidades según la distribución ya incluida en la red.

Ignorando la componente gigante por ser demasiado extensa y conocer ya autores importantes en ella, nos fijamos en las otras comunidades. Destacan las de color naranja, turquesa, verde y gris oscuro por ser las de mayor tamaño. Pasamos a dibujarlas por separado:

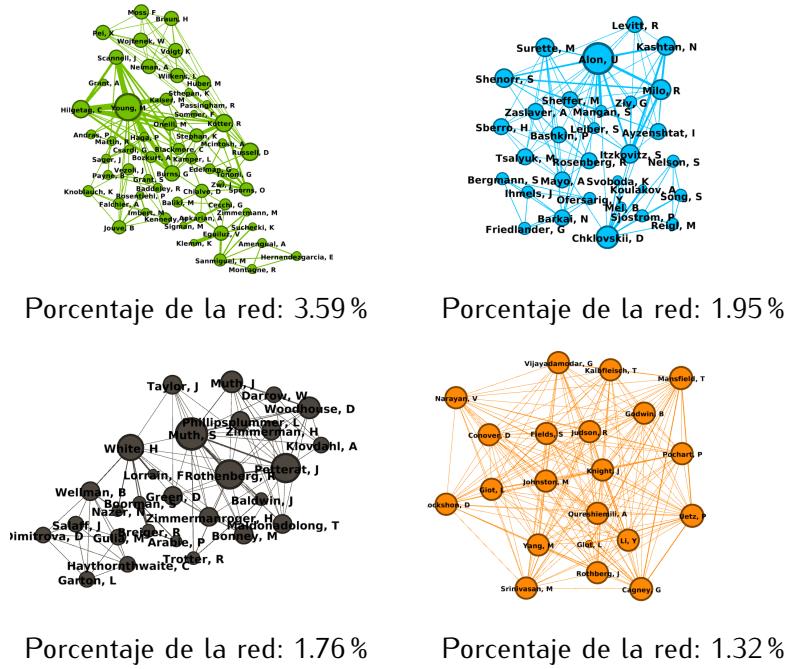


Figura 21: Visualizaciones de las 4 comunidades elegidas.

Observamos que en todas las comunidades parece haber personas más relevantes que concentran la actividad, excepto en la naranja. En esta comunidad todos los autores parecen tener el mismo peso (el nodo más pequeño puede que exista por un error tipográfico), algo que nos anima a elegirla para averiguar por qué.

6.2. Detección con el método de *Louvain*

Ya tenemos una serie de comunidades candidatas sacadas de la información que acompaña a la red. El siguiente paso es ver qué resultados obtenemos si ejecutamos un algoritmo específico para la tarea.

La primera ejecución será con un valor del parámetro *Resolución* igual a 1, el valor por defecto. Esto resulta en una modularidad de 0.955 (405 comunidades), que es una cantidad muy alta e indica una considerable estructura por *clusters* (nada nuevo después de ver la red). Pasemos a la parte interesante, que es ver los grupos más grandes que ha encontrado (17 en total):

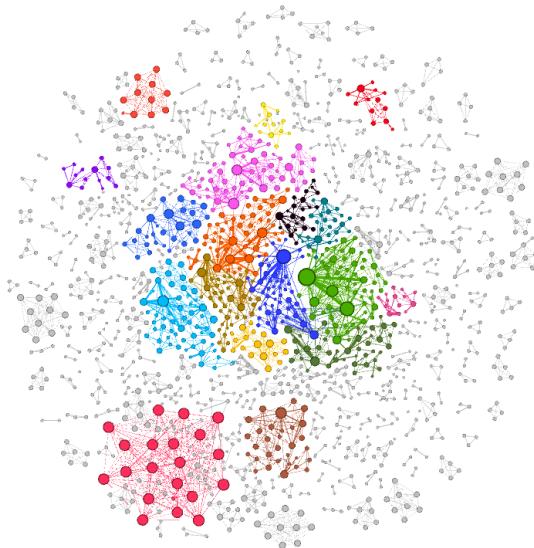
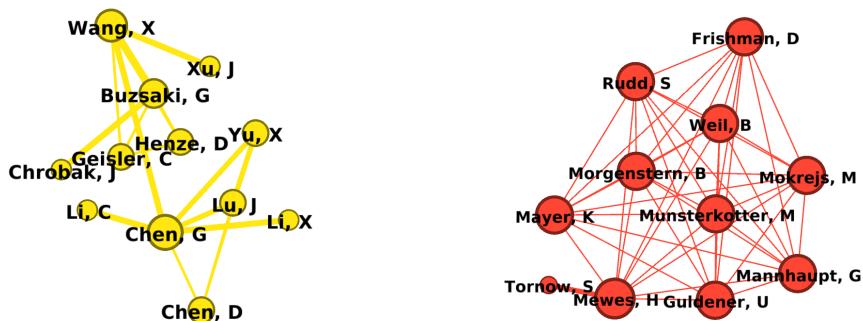


Figura 22: Comunidades según el método de *Louvain* con Resolución = 1.0.

Como es posible comprobar, no sólo ha encontrado todas las comunidades predefinidas (el dibujo es el mismo, por lo que podemos identificarlas a simple vista) sino que ha particionado también la componente gigante en grupos que veremos en breve. También aparecen otras dos agrupaciones aisladas que no teníamos antes, por si fuesen de interés:



Porcentaje de la red: 0.76 %

Porcentaje de la red: 0.69 %

Figura 23: Visualizaciones de las 2 nuevas comunidades.

A modo de cierre, echemos un vistazo a cómo ha dividido la componente gigante:

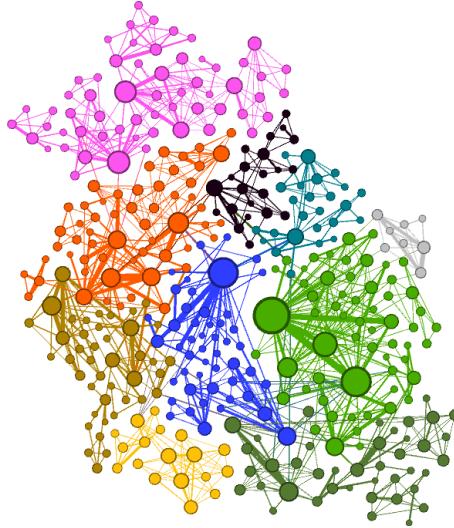


Figura 24: Comunidades según el método de *Louvain* con Resolución = 1.0.

Al tener una imagen más cercana entendemos mejor por qué existe tal grado de modularidad: si nos fijamos, existen pocos enlaces conectando cada grupo de color, aunque al final todos los nodos que hay en la imagen son mutuamente alcanzables.

Si aproximamos la vista a la zona donde están *Barabási*, *Jeong* y *Newman* (tres de los nodos más importantes según las medidas de centralidad), podemos entender mucho mejor dos cosas: la comunidad de investigadores tan amplia que se fundamenta en los dos primeros, y el alto valor de intermediación de *Newman*, que interconecta a su propia comunidad y a las otras dos que vemos.

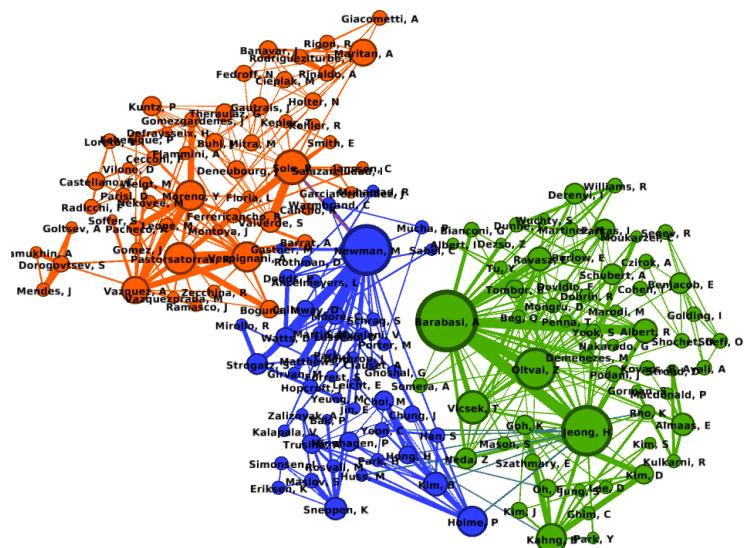


Figura 25: Detalle de la componente gigante en una región de nodos importantes.

Continuando con la detección de comunidades, vamos a ejecutar el algoritmo con un valor mayor que 1 de Resolución y otro menor para favorecer comunidades más grandes y más pequeñas, respectivamente.

El valor mayor que uno elegido es 10, fijado tras varios experimentos en los que el algoritmo no producía resultados diferentes notables. Para un valor de modularidad de 0.921 y 398 comunidades, aquí tenemos el resultado:

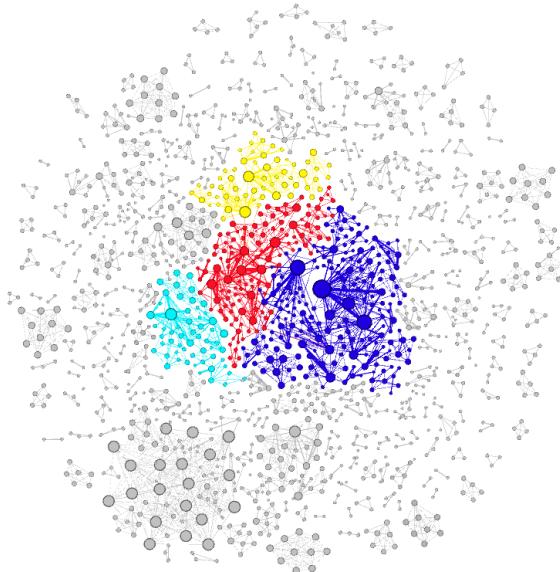


Figura 26: Comunidades grandes según el método de *Louvain* con Resolución = 10.0.

Aun no siendo poco informativa por sí misma, lo único novedoso que nos aporta esta división es que *Newman* y su comunidad pueden ser considerados parte de la de *Barabási* y *Jeong*, que se expande un poco. Veámoslo antes de proseguir:

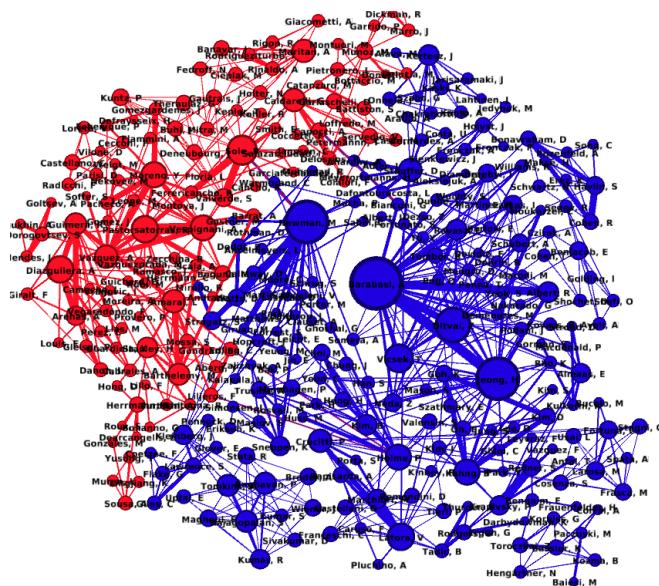


Figura 27: Las dos comunidades más importantes de la ejecución con Resolución = 10.0.

En último lugar, veamos qué pasa si hacemos más pequeño el parámetro del algoritmo de *Louvain*. Su valor es de 0.5 y nos da un total de 414 comunidades con 0.951 de modularidad:

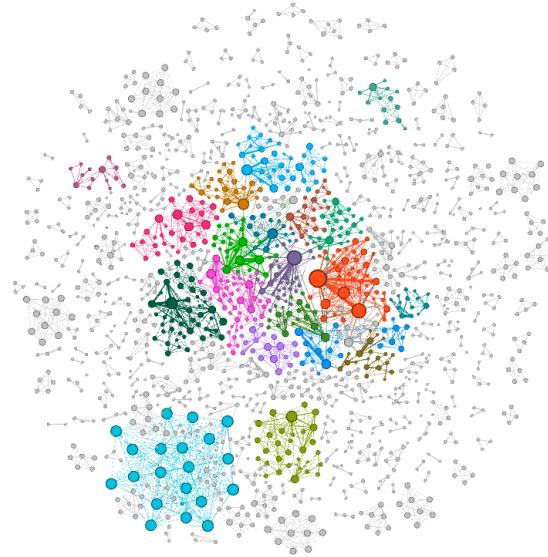


Figura 28: Comunidades grandes según el método de *Louvain* con Resolución = 0.5.

Ahora tenemos un número superior de comunidades destacadas, todas ellas más pequeñas pero más orientadas al entorno cercano de cada autor. Muchos grupos que ya conocíamos vuelven a aparecer, y otros se fragmentan. Si quisiésemos investigar en profundidad la red más estrecha de colaboradores de un determinado autor, estas comunidades nos darían una idea más acertada que las vistas anteriormente.

Por familiaridad con la visualización que estamos usando, podemos intuir cómo se han fragmentado los *clusters* más centrales sin necesidad de enturbiar la imagen con las etiquetas:

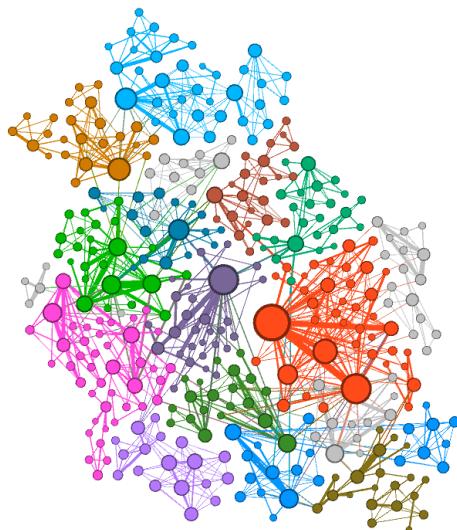


Figura 29: Close-up del centro de la red para comunidades con Resolución = 0.5.

Para terminar, hemos de elegir qué comunidades serán tomadas en consideración para investigar en la sección de conclusiones. Debido a que las medidas de centralidad de redes sociales nos han dado muchos autores situados en la componente gigante, nos centraremos en la periferia.

Como las cuatro comunidades predefinidas que hemos elegido han aparecido con *Louvain*, las aceptamos. La restante será la segunda de las adicionales que encontramos en la ejecución del mismo. Todas ellas se pueden apreciar de nuevo en la siguiente figura:

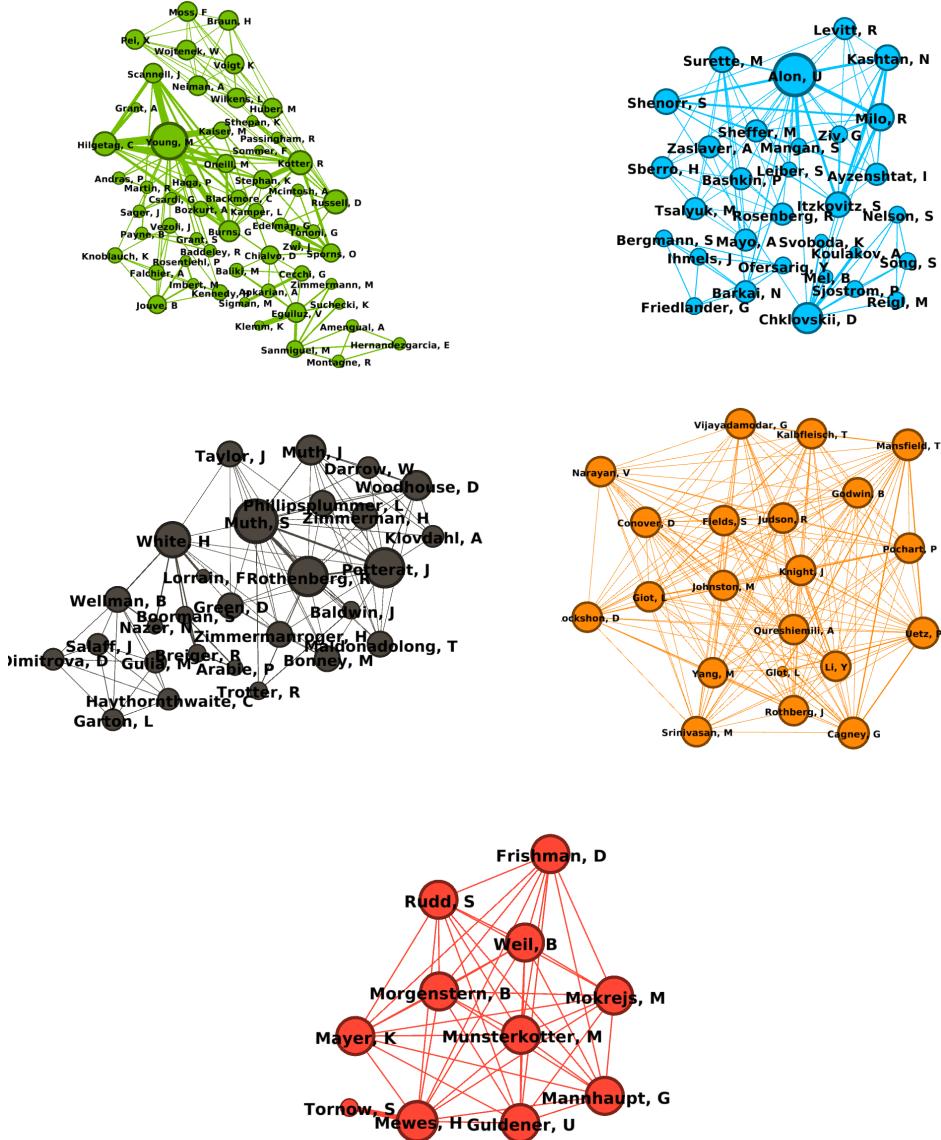


Figura 30: Las 5 comunidades escogidas.

7. Conclusiones

Nuestro objetivo primordial era averiguar cuáles son los autores más relevantes de la bibliografía de la ciencia de redes, o al menos cuáles eran hasta 2006, año de creación del conjunto de datos.

Así pues, vamos a buscar pruebas que confirmen que íbamos bien encaminados con los autores seleccionados por las medidas de centralidad. Salvo que se indique otra cosa, se indagará acerca de unos pocos autores utilizando *Google Scholar*.

Empezamos con *Barabási*. Es autor de dos de las más extensamente citadas publicaciones en la literatura de este campo: *Emergence of scaling in random networks*, 1999, con 29984 citas, y *Statistical mechanics of complex networks*, 2002. Junto con *Réka Albert* escribió estos dos artículos, que sirvieron para dar un potente impulso a la ciencia de redes moderna.

Además, podemos encontrar un libro suyo (*Linked: The New Science of Networks*) citado por 9000. Destacamos también un artículo en colaboración con *Hawoong Jeong* y de nuevo *Albert* (2000), con 7988 citas, y otro con *Zoltán Oltvai* (2004), con 6289 citas. Aparece incluso *Mark Newman*, en *The structure and dynamics of networks* (2006), con 2903 citas.

Como podemos comprobar, no sólo ha hecho valiosas contribuciones (un total de 175903 citas), sino que ha participado en ellas con autores que también habíamos considerado relevantes según nuestro análisis de la red.

Newman, por su parte, también ha aportado publicaciones altamente relevantes por su cuenta: *The structure and function of complex networks* (2003), con 16631 citas, o *Community structure in social and biological networks* (2002), con 10496 citas.

Hawoong Jeong no tiene unas cifras tan impresionantes como los dos anteriores, pero sus frecuentes colaboraciones con *Barabási* le han valido 39470 citas en total.

Para no alargar en exceso esta sección, terminamos con *Zoltán Oltvai*, tercero en centralidad de vector propio. Este puesto lo ha conseguido, entre otros motivos, por sus trabajos con *Barabási* y *Jeong*; véanse *Network biology: understanding the cell's functional organization* (2004), 6289 citas, o *The large-scale organization of metabolic networks* (2000), 5118 citas.

Por el breve repaso a algunas de estas personas que hemos hecho, parece que en efecto la red nos ha permitido encontrar a los autores más destacados.

Para concluir, veamos qué nos pueden ofrecer los cinco grupos elegidos en la parte de detección de comunidades. Nos referiremos a ellos por el color de sus nodos (los artículos se pueden ver pinchando en los nombres):

- Grupo verde: *Anatomical connectivity defines the organization of clusters of cortical areas in the macaque monkey and the cat* (2000), 505 citas. De *Hilgetag, Burns, O'Neill, Scannell* y *Young*, todos ellos presentes en la comunidad.
- Grupo turquesa: *Network motifs in the transcriptional regulation network of *Escherichia coli** (2002), 2675 citas, de *Shen-Orr, Milo, Mangan* y *Alon*; *Network motifs: simple building block of complex networks* (2002), 5304 citas, de *Milo, Shen-Orr, Itzkovitz, Kashtan, Chklovskii* y *Alon*. Todos ellos presentes también.
- Grupo gris: *Social network dynamics and HIV transmission* (1998), 238 citas. De *Rothenberg, Potterat, Woodhouse, Muth, Darrow* y *Klovahl*. Nuevamente, todos están en la comunidad.
- Grupo naranja: *A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae** (2000), 5243 citas. De todos y cada uno de los autores presentes.

Esta comunidad es especialmente llamativa porque se ha formado por una colaboración de 20 personas en un único artículo, lo que les confiere el mismo grado a todos los nodos.

- Grupo rojo: [*MIPS: a database for genomes and protein sequences*](#) (2002), 1064 citas. De todos excepto *Tornow*, el de grado más pequeño. Nuevamente un grado igual resulta en un artículo publicado en conjunto por todos los autores de la comunidad.

A la vista de estos resultados, también ha sido posible encontrar grupos de colaboración y contribuciones a otros campos (basta con ver los nombres de los *papers*).

Con todo esto, que es tan sólo una fracción de lo que se podría haber ahondado en la información disponible, podemos concluir que la red permite de manera efectiva entender la estructura de investigación que existía en la ciencia de redes hasta el año 2006.