Práctica Bloque I.1
Análisis y Visualización Básica de una Red Social con Gephi

Marta Verona Almeida.
DNI: 54087879-K
E-mail: martaverona@correo.ugr.es

13 de mayo de 2019

Índice

1. Introducción

Esta práctica tiene dos objetivos. En primer lugar, familiarizarse con los procedimientos de análisis de redes y con las medidas que habitualmente se consideradan para ello. En segundo lugar, aprender el manejo de Gephi, una herramienta estándar de análisis y visualización de redes.

Así pues, durante esta práctica se realizará un estudio de la red **Diseasome**, extraída de [1]. Se trata de una red de trastornos y genes de enfermedades vinculados a trastornos conocidos, que indican el origen genético común de muchas enfermedades.

2. Resultados gráficos

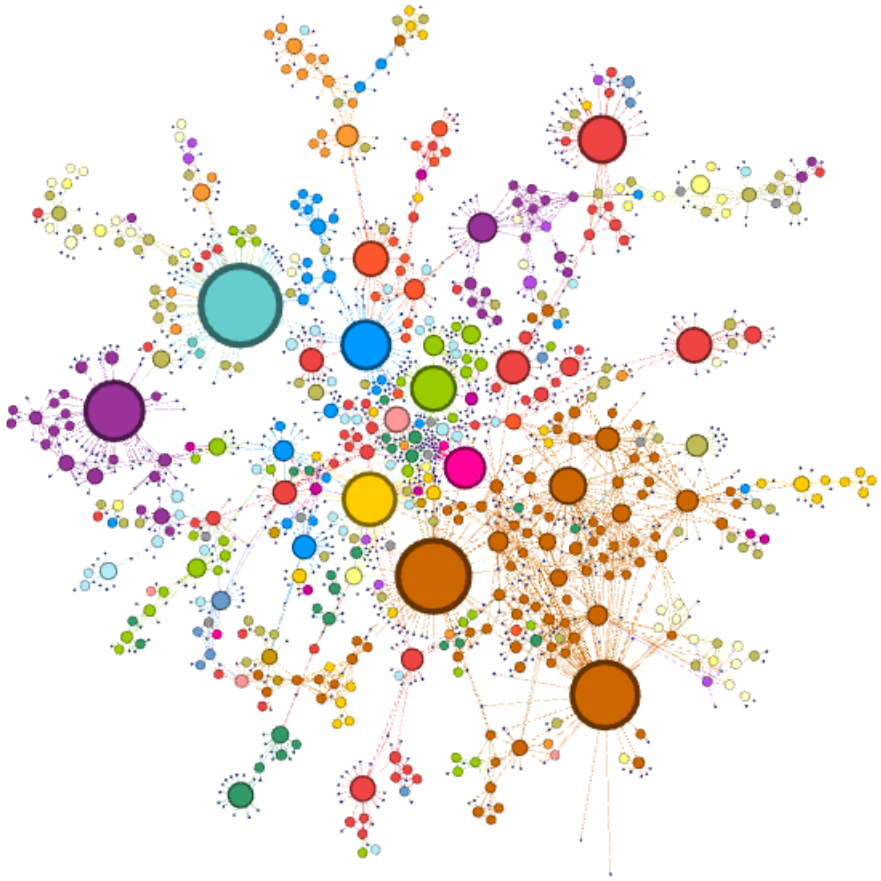

Figura 1: Grafo Completo - Componente Gigante

Medida	Valor
Número de nodos *N*	1419
Número de enlaces *L*	3926
Número máximo de enlaces L_{max}	1963000
Densidad del grafo *L/Lmax*	0,002
Grado medio de entrada (soporte)	2.766737
Grado medio de salida (prestigio)	2.766737
Diámetro d_{max}	15
Distancia media *d*	6,6487
Coeficiente medio de clustering <C>	0,414
Número de componentes conexas	1
Número de nodos componente gigante (y %)	1419 (100%)
Número de aristas componente gigante (y %)	2738 (100%)

Figura 2: Medidas de la red

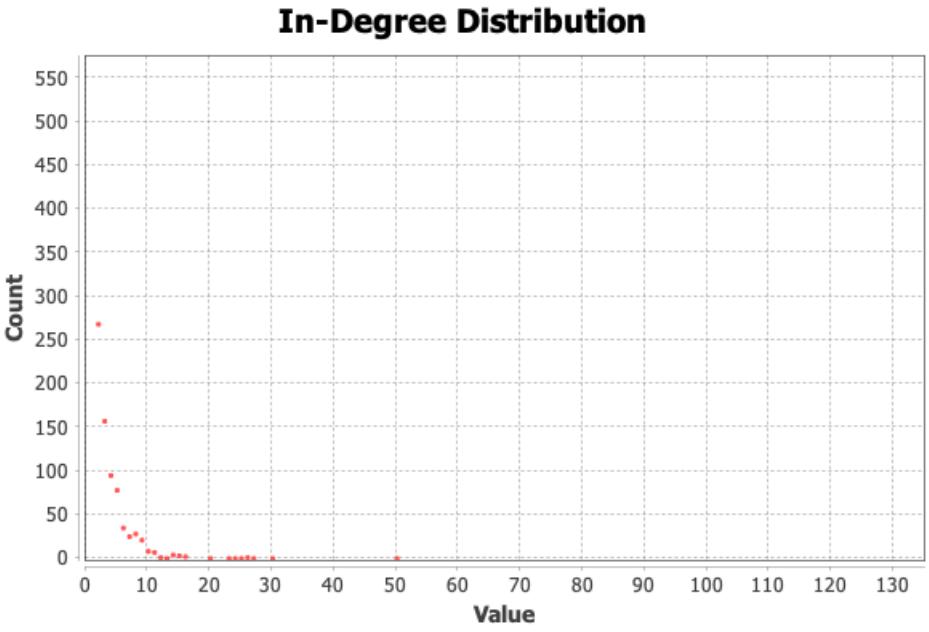

Figura 3: Distribución del grado de entrada (Soporte)

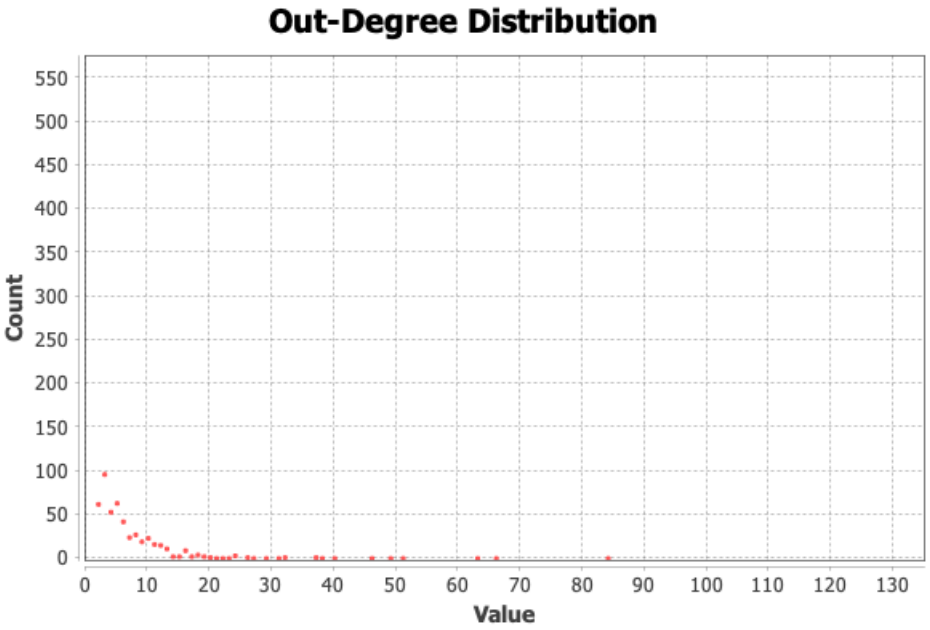

Figura 4: Distribución del grado de salida (Prestigio)

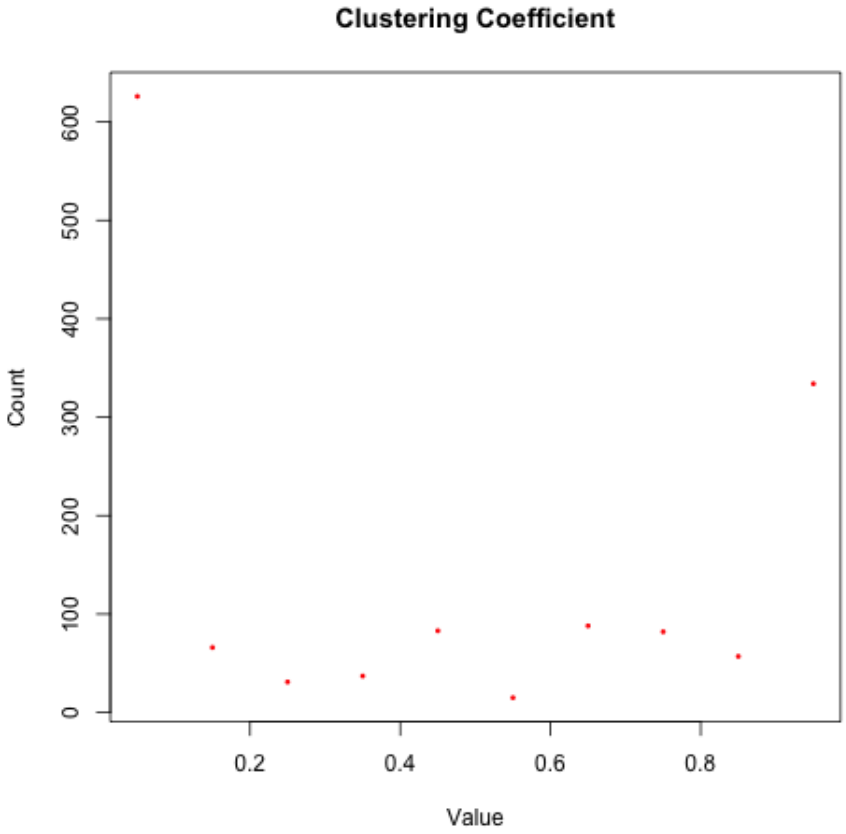

Figura 5: Distribución del coeficiente de clustering

3. Análisis de la red

En primer lugar, en la imagen 2, se puede observar el grafo dirigido que será objeto de estudio a lo largo de la práctica. Para su visualización se ha utilizado como `Force Atlas 2` como algoritmo de layout, con gravedad 15, activando las opciones *"Evitar solapamiento"* y *"Disuadir Hubs"*.

Esta red consta de 1419 nodos y 3926 enlaces, siendo el número máximo de enlaces 1963000, de modo que la densidad del grafo es 0.002. Además, el diámetro de la red es 15, siendo la distancia media 6.6487, de modo que la red se encuentra bastante centrada. Esta información puede verse en la tabla 2.

Por otra parte, dado que el grafo objeto de estudio es un grafo dirigido, analizamos el grado medio de entrada (soporte) y de salida de forma disjunta (prestigio). En 2 se observa que, en ambos casos, el grado medio es 2.767. Eso quiere decir que cada nodo interactúa con unos 3 nodos más. Además, en las imágenes 2 y 2 se muestran las distribuciones asociadas al soporte y al prestigio respectivamente. Observamos que ambas distribuciones son similares, en ambos casos existe una mayoría de nodos con una interacción muy baja, aunque se llegan alcanzar valores altos. Por tanto, se puede afirmar que la red estudiada es una red de libre escala

En cuanto a la conectividad, en la figura 3 se observa que existe una única componente conexa, es decir, el grafo es conexo. Por lo tanto, la componente gigante es el propio grafo, agrupando el 100 % de los nodos. De esta forma, se observa que en esta red no existe fragmentación.

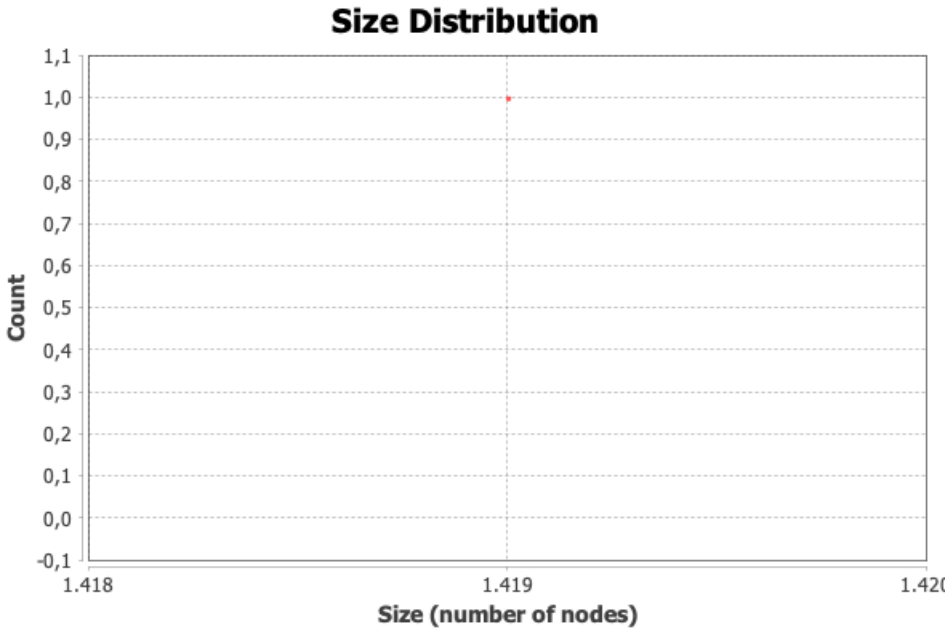

Figura 6: Componentes conexas

El coeficiente medio de clústering es 0.414 lo que es bastante alto, indicando un grado muy significativo de clústering local. Además, en la figura 2 puede observarse la distribución de los coeficientes de clustering. Se puede observar que existe una mayoría de nodos con coeficiente de clusteirng muy bajo, es decir, que no se encuentran conectados con muchos nodos. Sin embargo existen otros con un alto coeficiente de clústering, superior a 0.8, éstos se encuentra altamente conectados.

4. Análisis de la centralidad

En esta sección, se realizará un análisis de la centralidad de los actores de la red. Para ello, se estudiará la intermediación, cercanía, excentricidad y vector propio de la red.

En primer lugar, se analiza la distribución de intermediación. Esta medida de centralidad indica si un nodo se encuentra situado entre los caminos más cortos (geodésicos) de otros dos nodos. Es decir, será más central aquel nodo que se encuentra en muchas caminos más cortos entre otros pares de nodos.

En la figura 4 puede verse la distribución de intermediación en la red estudiada. En ella, se observa que los valores de intermediación son muy bajos en todos los nodos, no llegando al 0.2 en ningún caso.

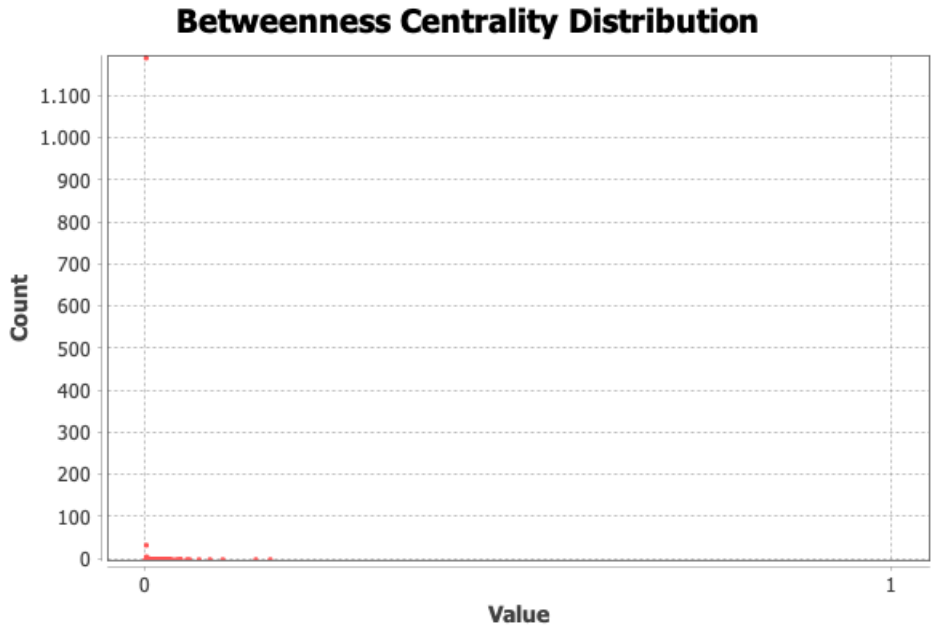

Figura 7: Intermediación

En segundo lugar se estudia la cercanía, que mide la centralidad en función de si un nodo se encuentra cerca del centro, aunque no se encuentre conectado con muchos nodos. La distribución de la cercanía puede verse en la figura 4. En ella, observamos que 900 nodos tienen cercanía prácticamente cero, por loque son muy periféricos. El resto parecen distribuirse en [0.1,0.2], por lo que no hay un nodo que destaque por su cercanía al centro de la red.

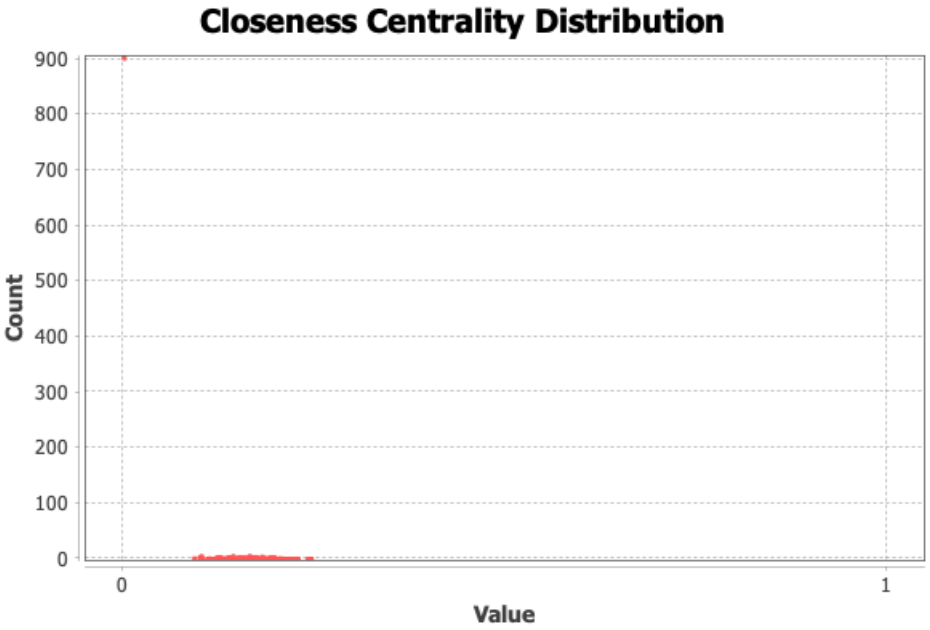

Figura 8: Cercanía

En tercer lugar, la excentricidad mide la máxima distancia de un nodo en la red. Su distribución puede verse en la figura 4. En este caso, se observa que más de 800 nodos tienen una máxima distancia geodésica cercana a cero, lo que indica que, en general, los nodos se encuentran muy cerca entre sí.

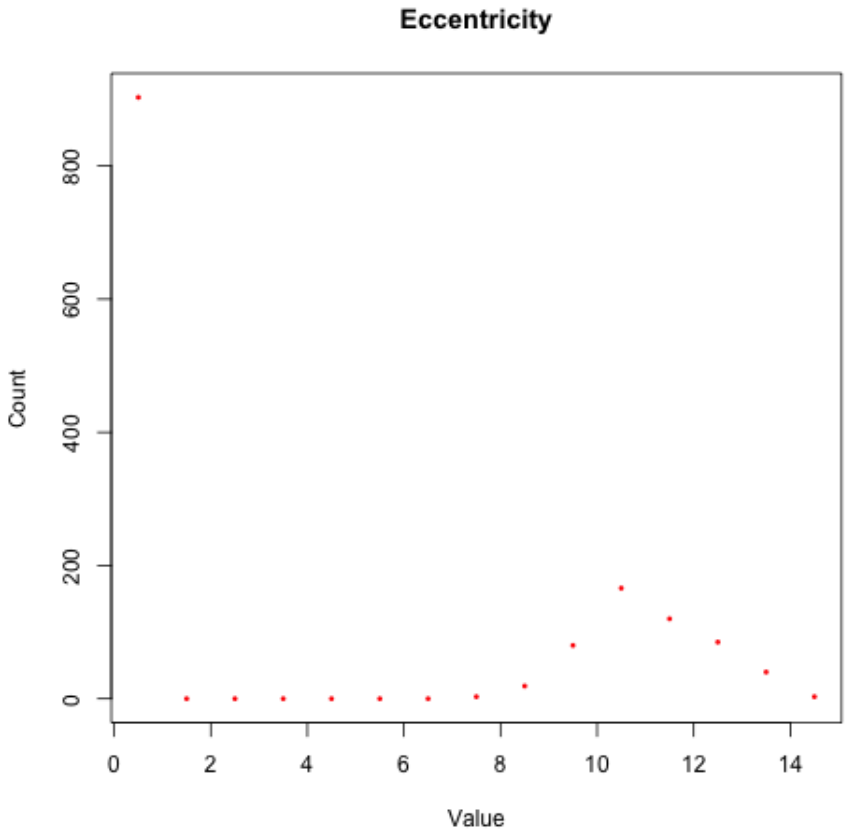

Figura 9: Excentricidad

En cuarto lugar, se estudia la centralidad del vector propio y su distribución puede verse en la figura 4. Esta medida se basa en que la centralidad de un nodo depende de cómo de centrales sean sus nodos vecinos. De esta forma, no sólo se tiene en cuenta la cantidad de conexiones de un nodo, sino su calidad.

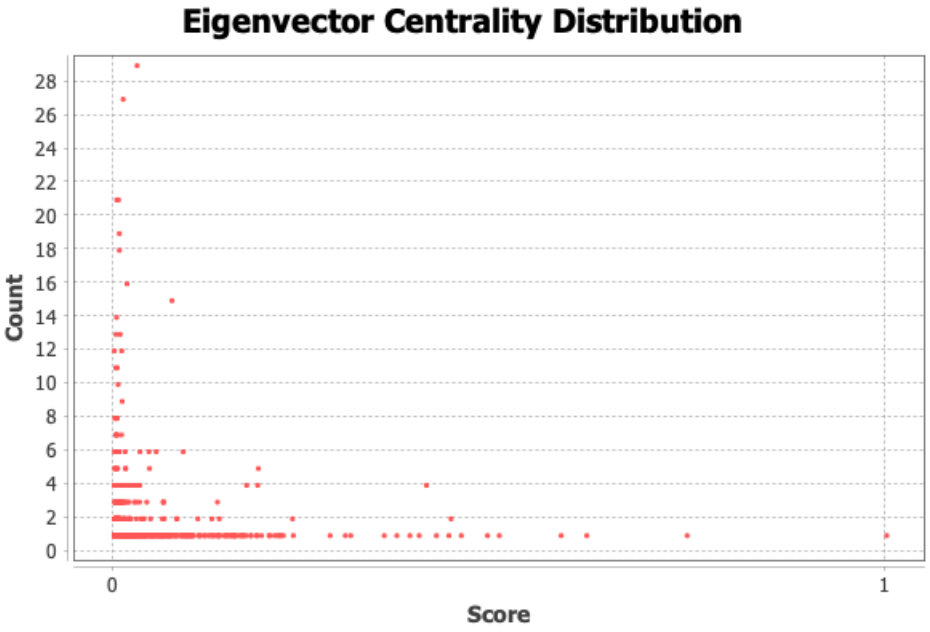

Figura 10: Vector propio

Por último, en la tabla 4, se recogen los nodos con mayor valor para cada una de las cuatro medidas de centralidad expuestas en esta sección.

Grado de entrada	Grado de salida	Intermediación	Cercanía	Vector propio
Colon cancer: 50	Colon cancer: 84	Cardiomyopathy: 0.1662	Lipodystrophy: 0.2454	Colon cancer: 1.0
Breast cancer: 30	Deafness: 66	Lipodystrophy: 0.1470	Diabetes mellitus: 0.2443	Breast cancer: 0.7420
Gastric cancer: 27	Leukemia: 63	Diabetes mellitus: 0.1028	Glioblastoma: 0.2404	Thyroid carcinoma: 0.6120
Leukemia: 26	Diabetes mellitus: 51	Glioblastoma: 0.0855	Obesity: 0.2281	Pancreatic cancer: 0.5789
Thyroid carcinoma: 26	Breast cancer: 49	Deafness: 0.0709	Cardiomyopathy: 0.2268	Gastric cancer: 0.4989

Figura 11: Nodos Centralidad

Como puede verse, existen muchos actores comunes en estas medidas de centralidad. En primer lugar es destacable que cuatro de los cinco actores con máximo valor en intermediación y cercanía coinciden, lo que nos indica que los nodos que se encuentran en más caminos más cortos del resto de pares de nodos, son los que más cerca se encuentran del centro.

Además, existe un nodo que presenta altos valores para la influencia, la intermediación y la cercanía: *Diabetes mellitus*.

Por otra parte, tres de los cinco agentes con mayor soporte tienen también mayor prestigio. Además tres de los cinco agente con mayor soporte también tienen máximo valor de centralidad de vector propio.

En definitiva, se puede observar que existen muchos agentes comunes a todas las medidas de centralidad empleadas, por lo que todas parecen explicar correctamente la centralidad de la red.

5. Estudio de las comunidades

En esta sección se realizará un estudio de las estructuras de comunidades para determinar la estructura modular de la red.

Las comunidades son regiones del grafo que presentan alta concentración de enlaces y una baja concentración de enlaces entre otras regiones.

A la hora de realizar este estudio se ha utilizado el método Lovaina con diferentes valores del parámetro *resolución*, el cual determina el número de comunidades obtenido por el algoritmo. Para determinar qué valor de resolución se ajusta mejor a nuestra red se tienen en cuenta dos factores. En primer lugar, se busca obtener un número de comunidades que nos permita realizar un análisis de las mismas, por lo que se busca que este número sea lo menor posible. En segundo lugar, se debe tener en cuenta la modularidad, que mide la calidad de una partición concreta de una red en comunidades. Su valor se encuentra comprendido en el intervalo $[-1, 1]$ y toma valor máximo cuando la red presenta todos los enlaces dentro de cada comunidad y ninguno entre comunidades. Por lo tanto, se buscará maximizar la modularidad y minizar el número de comunidades.

En la tabla 1 pueden verse los valores obtenidos para la resolución y la modularidad dependiendo de la resolución fijada. Puede verse que, como ya se sabe por el funcionamiento del método Lovaina, cuanto mayor es la resolución menos comunidades se encuentran. Sin embargo, la modularidad oscila algo más, probablemente por la componente aleatoria del método, siendo en todos los casos bastante alta. Teniendo en cuenta la relación buscada entre la modularidad y el número de comunidades, se decide establecer la resolución a 5, de manera que se obtiene una modularidad de 0.825 y 10 comunidaes.

Resolución	Modularidad	Número de Comunidades
0.5	0.868	36
1	0.871	28
1.5	0.873	22
2	0.856	19
2.5	0.830	15
3	0.833	14
3.5	0.835	13
4	0.822	12
5	0.825	10
4	0.780	9

Cuadro 1: Resultados Lovaina

Una vez se ha fijado la resolución para el método Lovaina de detección de comunidades, se realiza un análisis de las comunidades detectadas. En la imagen 5 se observa el tamaño de cada una de las 10 comunidades obtenidas. Puede verse que el abanico de tamaños es bastante amplio, comprendido en el intervalo $(40, 340)$. Se observa una comunidad notablemente más grande que el resto, con unos 330 nodos, mientras que la menor contiene unos 50 nodos.

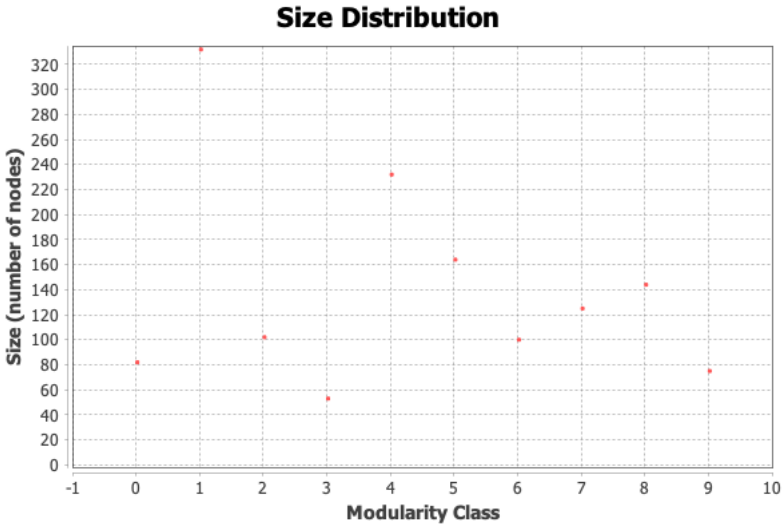

Figura 12: Tamaño de las comunidades

El último paso en este análisis consistirá en visualizar el grafo atendiendo a las comunidades obtenidas. En la imagen 5, se observa el grafo coloreado según la comunidad a la que pertecene. En ella, se puede ver en lila la comunidad de gran tamaño que se resaltó anteriormente. Del mismo modo, se pueden ver las comunidades más pequeñas en distintos tonos de gris.

Figura 13: Modularidad

6. Visualizaciones adicionales

En esta sección se mostrarán algunas gráficas adicionales, que complementan la información sobre la red que ha sido analizada a lo largo de la práctica.

En primer lugar, en la imagen 6, se ha realiza una visualización de la red mostrando la intermediación en el tamaño de los nodos y la centralidad de vector propio en la intensidad del color. De esta forma, los nodos de mayor tamaño son los que presentan mayor intermediación y los que tienen un color más oscuros los que presentan un valor mayor en la centralidad de vector propio. En ella se salta a la vista que los nodos que presentan mayor intermediación no presentan mayor centralidad de vector propio. Esto ya se resaltó al analizar la tabla 4, que contenía los nombres de los agentes con mayor valor en cada medida de centralidad. Por tanto, confirmamos que en esta red, estas medidas no crecen de manera proporcional.

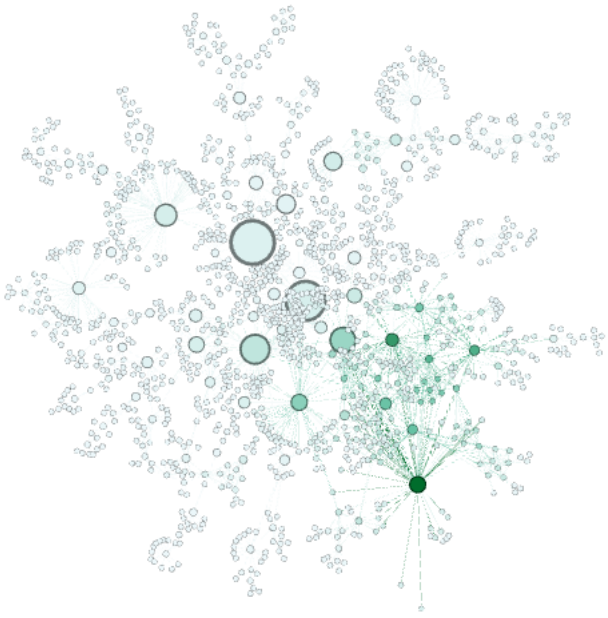

Figura 14: opcional1

En segundo lugar, en la figura 6, enfrentamos los valores de intermediación y centralidad de vector propio para comprobar si existe relación entre ellas. Una vez más comprobamos que no existe una relación clara entre ellas.

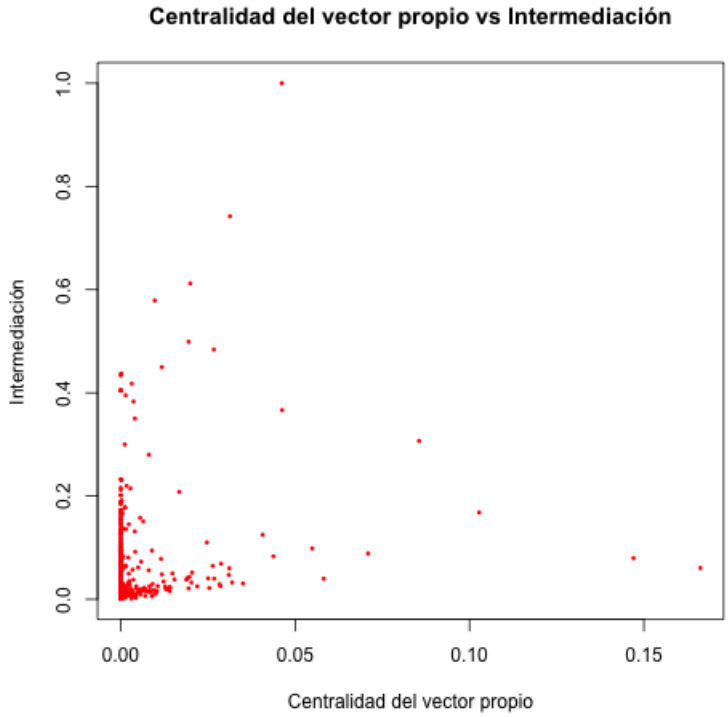

Figura 15: opcional2

Referencias

[1] https://github.com/gephi/gephi/wiki/Datasets