



Universidad
Internacional
Menéndez Pelayo

Máster Universitario en Investigación en Inteligencia Artificial

Curso 2020-2021

**Recuperación y extracción de información,
grafos y redes sociales**

Análisis y Visualización Básica de una Red Social con Gephi

15 de febrero de 2021

Laura Rodríguez Navas
DNI: 43630508Z
e-mail: rodrigueznavas@posgrado.uimp.es

La Red

La red *Diseasome*[1] seleccionada para realizar esta práctica, es una red no dirigida de trastornos y genes de diferentes enfermedades vinculadas por asociaciones conocidas entre trastornos y genes, que indican el origen genético común de muchas enfermedades. La red está formada por 526 enfermedades y 903 genes, donde los genes asociados con trastornos similares muestran una mayor probabilidad de interacciones físicas entre sus diagnósticos y una mayor similitud de perfiles de expresión para sus tratamientos, lo que respalda la existencia de distintos clústers funcionales específicos de cada enfermedad.

El conjunto de datos de *Diseasome* viene como un archivo .zip, que se puede descargar [aquí](#). Una vez se ha descargado y descomprimido, obtenemos un archivo .gexf, un archivo de grafos. Importamos el archivo de grafos a *Gephi*[2] y empezamos a probar diferentes opciones de visualización.

Después de probar diferentes visualizaciones nos decidimos por el algoritmo de distribución: [Fruchterman Rein-gold](#)[3] (en la ventana *Distribución*). Para evitar que las componentes conexas queden fuera de la vista principal, fijamos el valor del parámetro *Gravedad* a 20 y también marcamos las opciones *Disuadir Hubs* y/o *Evitar el solapamiento*. Esto convertirá nuestra visualización de la red en un círculo y colocará la red alrededor de una misma área (ver Figura 1).

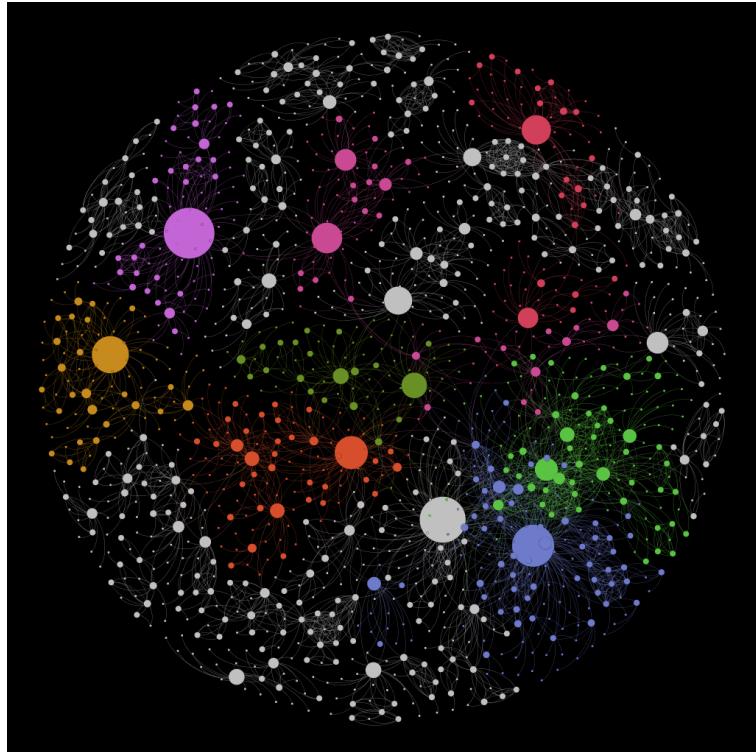


Figura 1: Red completa sobre un fondo negro sin etiquetas.

De una primera visualización pasamos a la detección de comunidades para colorear los clústers de la red. *Gephi* implementa el método de [Louvain](#)[4] para la detección de comunidades (disponible en el panel de *Estadísticas*). Para ello, damos clic en ejecutar *Modularidad* y veremos como el algoritmo de detección de comunidades nos ha creado un nuevo parámetro de particionamiento llamado *Modularity Class*. Si seleccionamos este nuevo parámetro podremos observar las comunidades encontradas y si finalmente pulsamos *Aplicar* colorearemos los nodos según las comunidades encontradas. Esto hace que la visualización sea más colorida y se vea bien donde se encuentra cada comunidad. También añadimos etiquetas a los nodos (ver Figura 2).

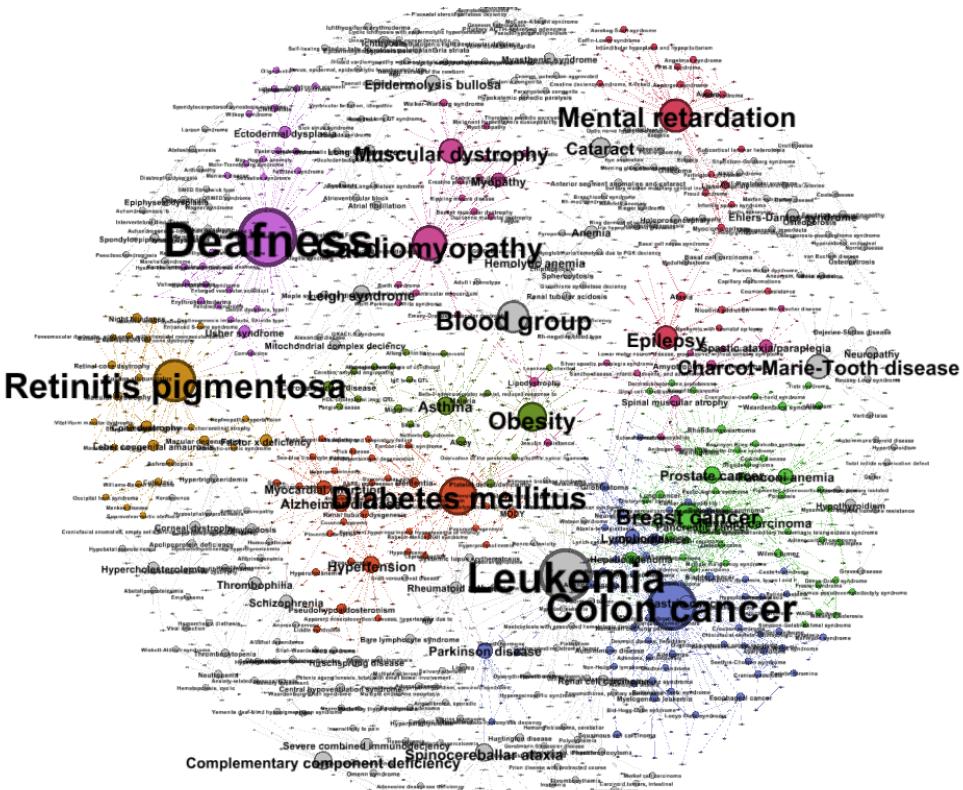


Figura 2: Red completa sobre un fondo blanco con etiquetas.

Como podemos observar en la Figura 2, las diferentes comunidades están agrupadas por colores. El tamaño de las etiquetas depende del tamaño del nodo. Está claro que los cánceres son la enfermedad más dominante de todas, siendo una de las enfermedades más comunes en comparación con otras enfermedades que existen en la actualidad. Un dato curioso es que la sordera es la enfermedad que se lleva la mayor porción. También vemos otros clústers además de los cánceres, como la diabetes, la salud mental, etc. Se da la propiedad libre de escala (*scale-free*), muy común en redes reales, porque muchos nodos de la red poseen un gran número de enlaces a otros nodos.

Análisis Básico de la Red

Como pudimos observar en la sección anterior, la red parece demasiado compleja para analizarla visualmente, así que para los primeros pasos del análisis de la red, comenzamos por anotar los valores de las medidas globales básicas: el número de nodos (N) y el número de enlaces (L), que aparecen directamente en la ventana *Contexto*. El número de nodos de la red es igual a 1419 y el número de enlaces es igual a 3926. Además calculamos manualmente el número máximo de enlaces L_{max} .

$$L_{max} = \frac{N * (N-1)}{2} = \frac{1419 * (1419-1)}{2} = 1006071$$

Posteriormente calculamos otra medida global, el grado medio $\langle k \rangle$, ejecutando la opción correspondiente en la ventana *Estadísticas*. El valor del grado medio $\langle k \rangle$ es igual a 5,533, es decir, que cada trastorno de la red está

conectado con 5 genes en media. Al realizar el cálculo del grado medio $\langle k \rangle$, también obtenemos la distribución de grados de la red completa (ver Figura 3).

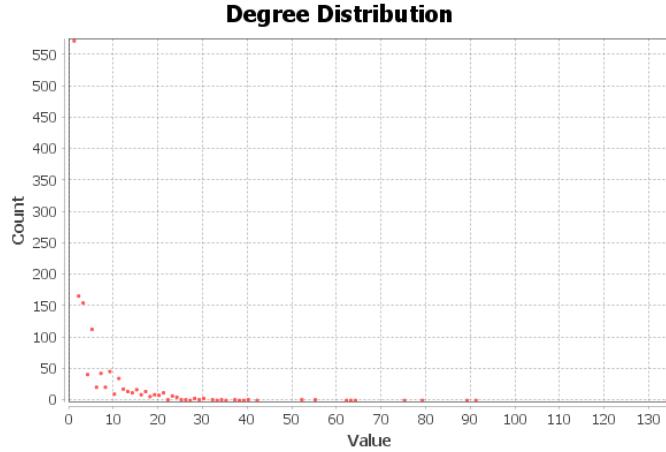


Figura 3: Distribución de grados de la red completa.

Existen algunas enfermedades fuertemente conectadas (*hubs*), la mayor con grado 135. Concretamente 10 enfermedades tienen más de 50 variaciones de genes.

La opción *Densidad* de grafo mide la relación entre el número de enlaces (L) y el número máximo de enlaces (L_{max}). Ejecutamos la opción y vemos que su valor es igual a 0,004 (valor cercano a 0, densidad mínima). Esto nos indica que el grafo es un grafo disperso, el número de enlaces (L) no es cercano al número de máximos de enlaces (L_{max}).

A continuación, ejecutamos la opción *Coeficiente medio de clustering $\langle C \rangle$* para obtener la medida del mismo nombre. El valor del coeficiente medio de clustering es igual a 0,819. Es un valor muy alto que nos indica un grado muy significativo de clustering local. Al realizar el cálculo del coeficiente medio de clustering también obtenemos la distribución de coeficientes de clustering de la red completa (ver Figura 4), donde vemos que el coeficiente de clustering es mucho mayor en los nodos poco conectados que en los nodos más conectados (*hubs*). En este caso, los nodos de grado bajo se sitúan en vecindarios localmente densos y viceversa como consecuencia de la jerarquía de la red.

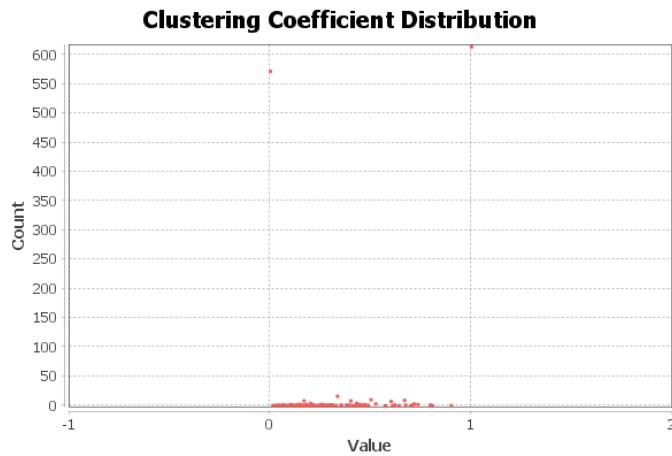


Figura 4: Distribución de coeficientes de clustering de la red completa.

Ahora, pasamos a analizar la conectividad de la red. En primer lugar, obtenemos el número de componentes conexas ejecutando la opción *Componentes conexos* (disponible en el panel de *Estadísticas*). Vemos que el número de componentes conexas es igual a 1. En este caso, como solo tenemos una componente conexa, determinamos que la componente gigante de la red es la red completa actual.

Finalmente, calculamos las medidas globales restantes (diámetro d_{max} y distancia media $\langle d \rangle$) ejecutando la opción correspondiente al *Diámetro de la red* en la ventana *Estadísticas*. El valor del diámetro (d_{max}) es igual a 15. Viendo la red (ver Figura 2), pensaríamos que hay variaciones grandes en las distancias entre los nodos pero la red tiene una distancia media baja ($\langle d \rangle = 6,783$). El cálculo del diámetro también nos proporciona los valores de las tres medidas de centralidad (intermediación, cercanía y excentricidad), que podemos observar en las figuras 5, 6 y 7. En la Figura 5 podemos comprobar que no hay variaciones grandes en las distancias entre los nodos, la mayoría de los nodos están bastante juntos. Solo 4 se encuentran muy alejados del resto.

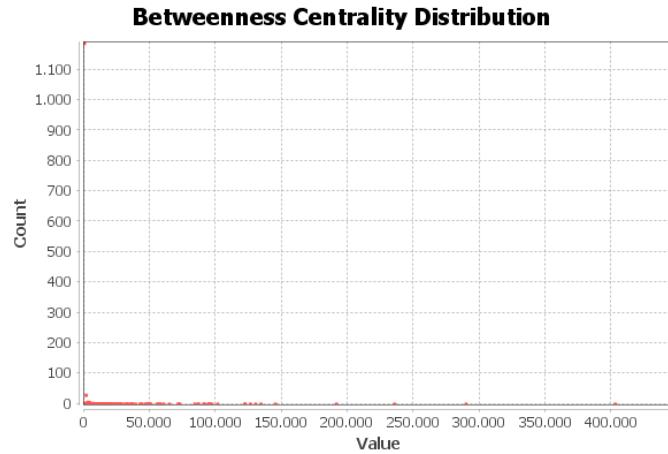


Figura 5: Centralidad de intermediación no normalizada de la red completa.

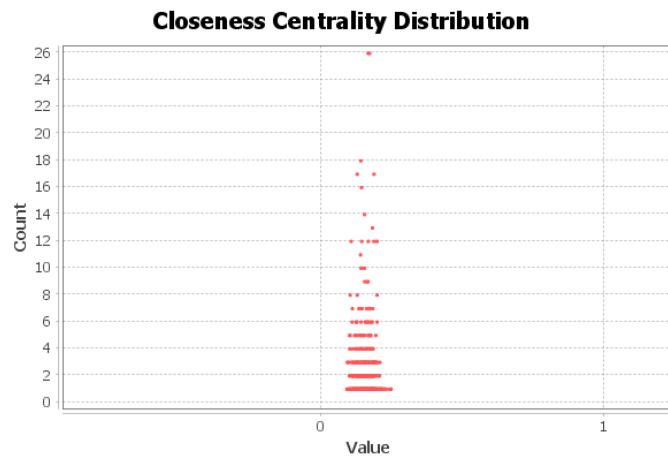


Figura 6: Centralidad de cercanía no normalizada de la red completa.

Observando la Figura 6 parece que se de la propiedad de mundos pequeños (*small-world*). La mayoría de los nodos no son vecinos entre sí, y sin embargo la mayoría de los nodos pueden ser alcanzados desde cualquier nodo origen a través de un número relativamente corto de saltos entre ellos. También observamos que no existen distancias largas en la red y que bastantes nodos no están muy alejados del centro de la red.

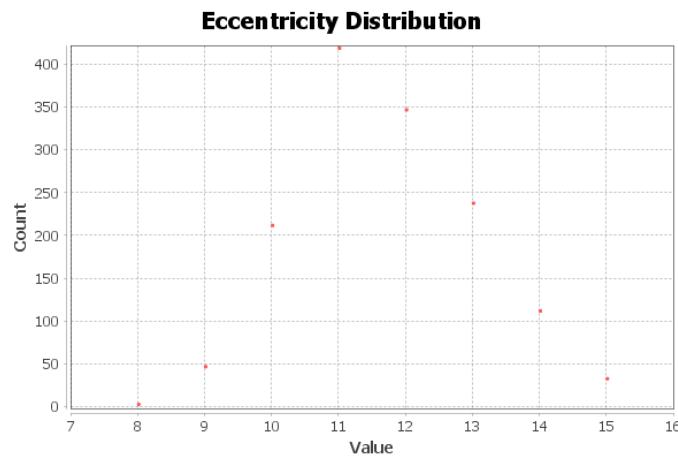


Figura 7: Centralidad de excentricidad no normalizada de la red completa.

Observando la Figura 7 vemos que la excentricidad no es muy constante. Esto nos indica que al no ser constante, no existen ni nodos muy periféricos ni muy centrales en la red.

A continuación, mostramos la tabla que resume los valores de las medidas calculadas anteriormente (ver fichero *MedidasRedesPracticaPartel-1.xlsx* para observar la tabla en formato Excel).

Medida	Valor
Número de nodos N	1419
Número de enlaces L	3926
Número máximo de enlaces Lmax	1006071
Densidad del grafo L/Lmax	0.004
Grado medio $\langle k \rangle$	5.533
Diámetro dmax	15
Distancia media d	6.783
Coeficiente medio de clustering $\langle C \rangle$	0.819
Número de componentes conexas	1
Número de nodos componente gigante	1419
Número de aristas componente gigante	3926

Tabla 1: Tabla con los valores de las medidas estudiadas.

En la siguiente sección de la práctica empleamos la medidas de centralidad calculadas.

Estudio de la Centralidad de los Actores

En esta sección se realiza un pequeño análisis de redes sociales sobre nuestra red basado en las medidas de centralidad. El análisis determina los 5 actores principales de la red mediante las medidas de centralidad de grado, intermediación, cercanía y vector propio.

Los valores de tres de estas medidas (grado, intermediación y cercanía) ya están calculados en los pasos que se han realizado en la sección anterior. La centralidad de grado (no normalizada) se generó al calcular el *Grado medio* $\langle k \rangle$ en la ventana *Estadísticas*. Las medidas de centralidad de intermediación y cercanía (no normalizadas) se generaron con la opción *Diámetro de la red*. En este caso, las volvemos a calcular para obtener las medidas normalizadas con el checkbox *Normalizar centralidades en el rango [0, 1]*.

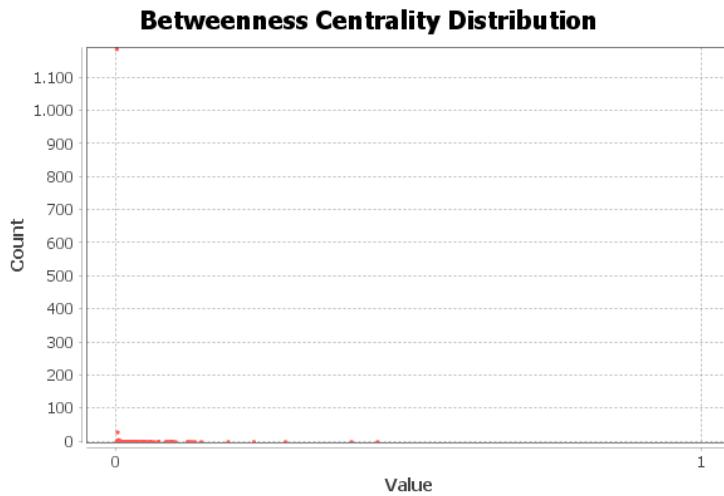


Figura 8: Centralidad de intermediación normalizada de la red completa.

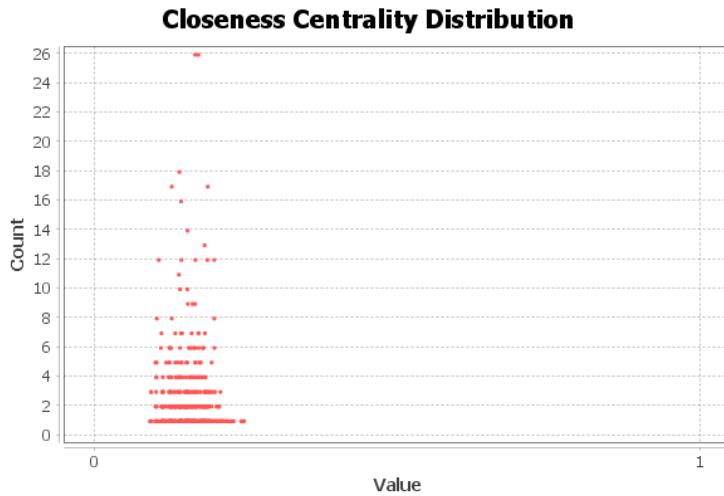


Figura 9: Centralidad de cercanía normalizada de la red completa.

Finalmente, calculamos la centralidad de vector propio que se calcula en la opción del menú *Estadísticas* del mismo nombre (ver Figura 10).

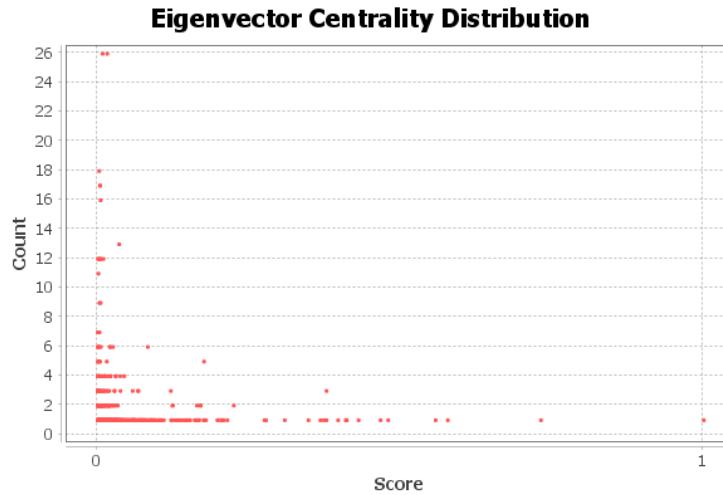


Figura 10: Centralidad de vector propio de la red completa.

Una vez ejecutadas las opciones de menú correspondientes, los valores de centralidad de cada nodo pueden visualizarse en la tabla *Nodos* de la pestaña *Tabla de datos*, junto con el resto de la información asociada a cada nodo. A continuación, anotamos los nombres de los 5 actores con mejor valor para cada una de las cuatro medidas anteriores, así como el valor de la medida en cada caso y los almacenamos en la tabla siguiente:

Centralidad de Grado	Centralidad de Intermediación	Centralidad de Cercanía	Centralidad de Vector propio
Colon Cancer 134	Cardiomyopathy 0.445069	Lipodystrophy 0.245414	Colon cancer 1.000000
Deafness 91	Lipodystrophy 0.400534	Diabetes mellitus 0.244272	Breast cancer 0.731462
Leukemia 89	Diabetes mellitus 0.28795	Glioblastoma 0.240402	Thyroid carcinoma 0.577893
Breast Cancer 79	Glioblastoma 0.234041	Obesity 0.228084	Pancreatic cancer 0.557891
Diabetes mellitus 75	Deafness 0.19061	Cardiomyopathy 0.226771	Gastric cancer 0.479706

Tabla 2: Los 5 actores con mejor valor por medida de centralidad.

Los actores mas importantes de la red desde una perspectiva global en función de los valores de las medidas de centralidad pertenecen a cánceres, diabetes, sordera, cardiopatías, VIH y obesidad (ver Tabla 2).

De estos actores realizamos un pequeño análisis a continuación, sin tener en cuenta la centralidad de grado, que aunque refleja el número de conexiones de cada actor, no tiene en cuenta la estructura global de la red.

Como sabemos, una medida bastante importante es la centralidad de intermediación, que nos indica que actores hacen de puente entre otras regiones de la red. Por lo cual pueden conectar distintas comunidades entre sí. En el caso que nos ocupa (*Diseasome*), podemos observar en la Tabla 2 que las cardiopatías son el actor mayoritario de esta medida de centralidad. Eso es que las cardiopatías son la enfermedad más conectada a otras enfermedades distintas, podemos decir que se diagnostica junto a muchísimas otras enfermedades. Por ejemplo, podemos sufrir cardiopatías con la obesidad, con la diabetes, etc. En la Figura 12 se puede observar como el nodo correspondiente a las cardiopatías tiene la mayor intensidad de color de la red.

La centralidad de cercanía mide cómo de cerca está un actor del centro de la red. En este caso parece ser que el actor más centrado pertenece a la enfermedad lipodistrofia, que se asocia a enfermedades como el virus de la inmunodeficiencia humana (VIH). Pero esta medida no nos serviría de mucho en un análisis más profundo, ya que muchos de los actores tienen un valor parecido de centralidad de cercanía. En la tabla 2 no se puede apreciar bien esta conclusión, pero si observamos la Figura 13 o la Figura 17, se puede observar perfectamente. En la Figura 13 vemos diferentes nodos con la misma intensidad de color, y en la Figura 17 muchas enfermedades con valores parecidos de centralidad de cercanía.

Por último, la medida de centralidad de vector propio es una medida recursiva que asigna importancia a un nodo en función de la importancia de sus vecinos. Es decir, tiene en cuenta la calidad de las conexiones, en lugar de la cantidad. El primer actor, el cáncer de colon, tiene un valor de esta medida de 1, lo cual indica que es el nodo más importante y con el mayor número de conexiones importantes. Es el actor a tener más en cuenta de la red. Tanto en la Figura 14 y Figura 18 se puede observar.

Visualizaciones y Gráficos adicionales

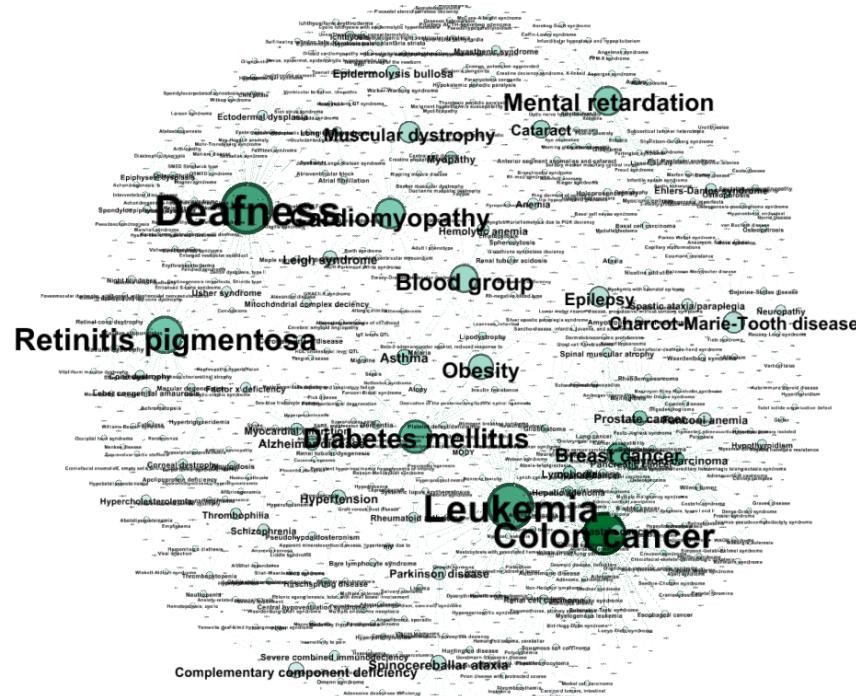


Figura 11: Centralidad de grado de la red completa.

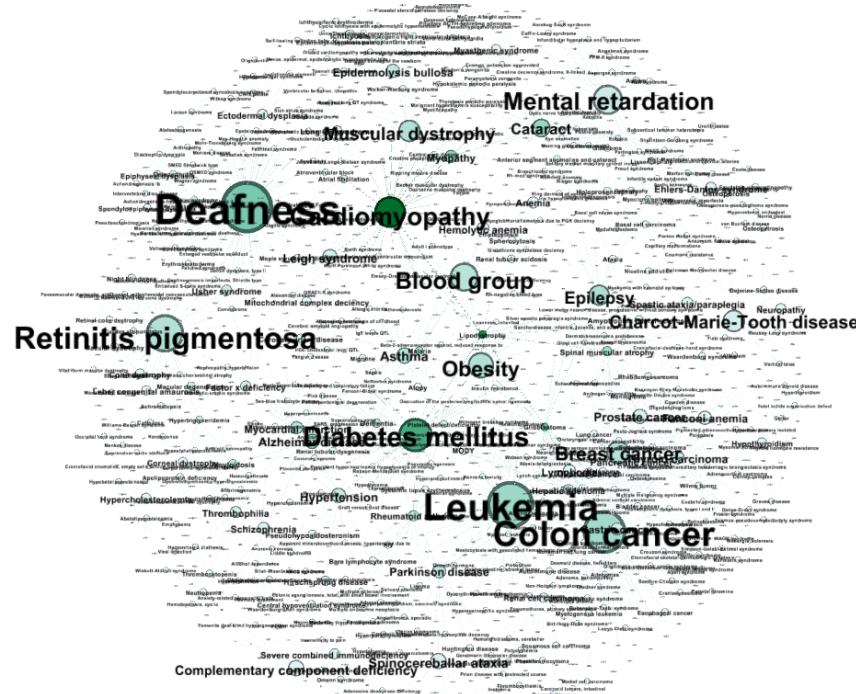


Figura 12: Centralidad de intermediación de la red completa.

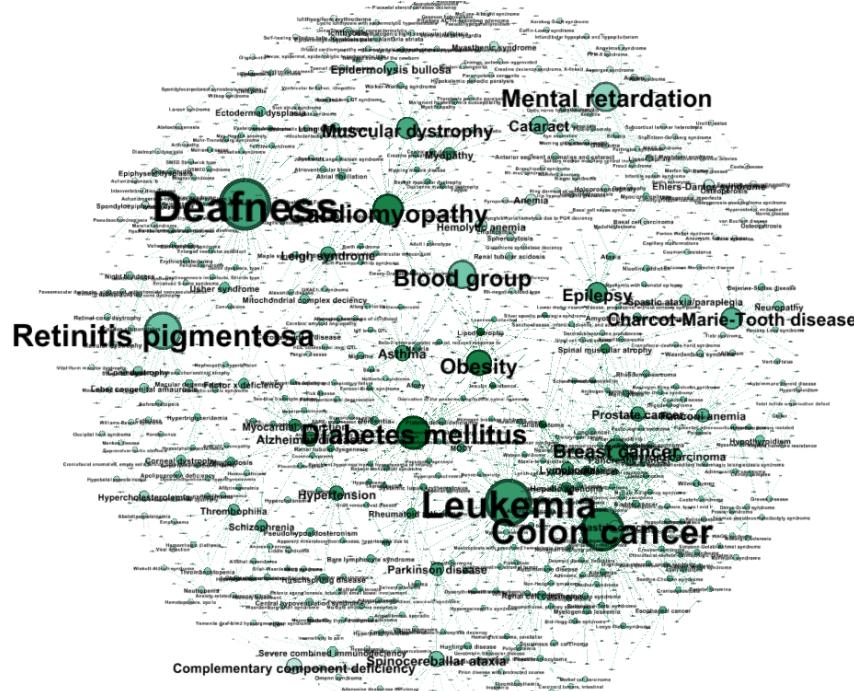


Figura 13: Centralidad de cercanía de la red completa.

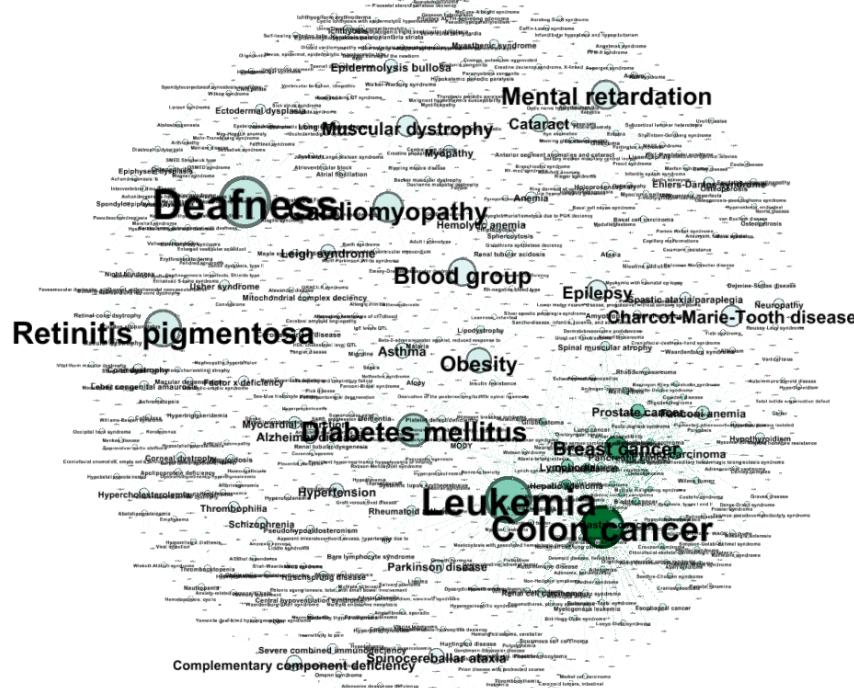


Figura 14: Centralidad de vector propio de la red completa.

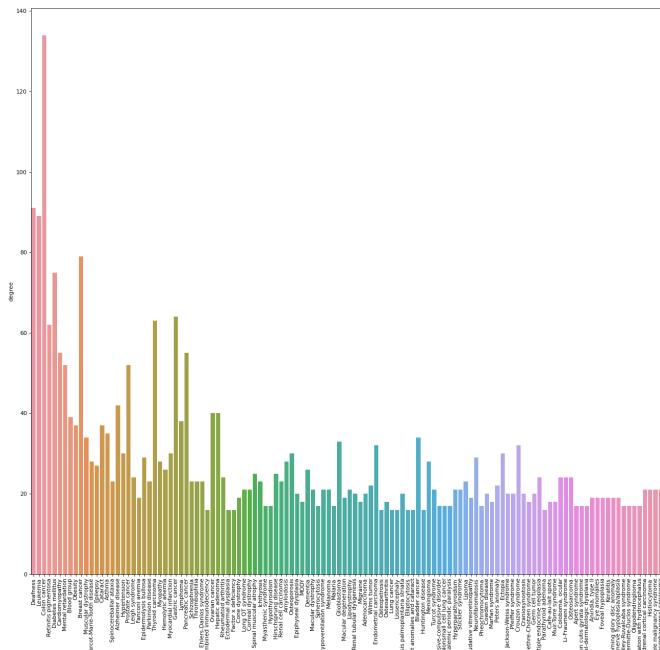


Figura 15: Actores de la medida de centralidad de grado.

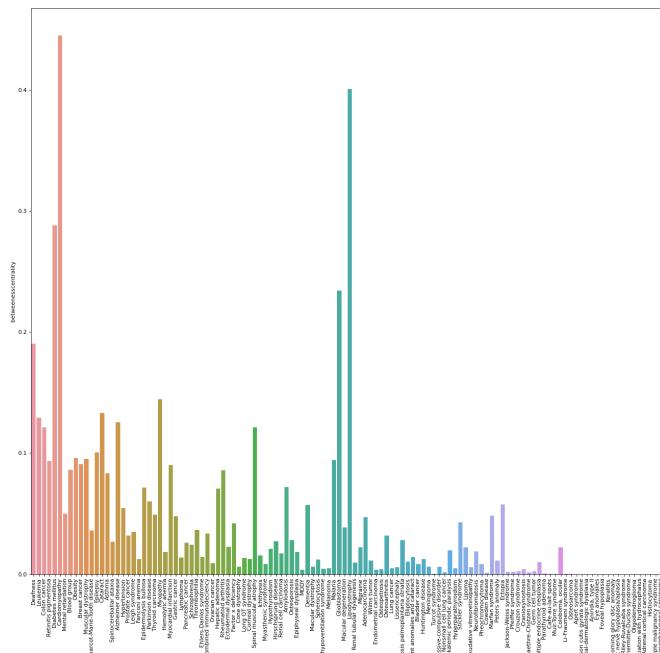


Figura 16: Actores de la medida de centralidad de intermediación.

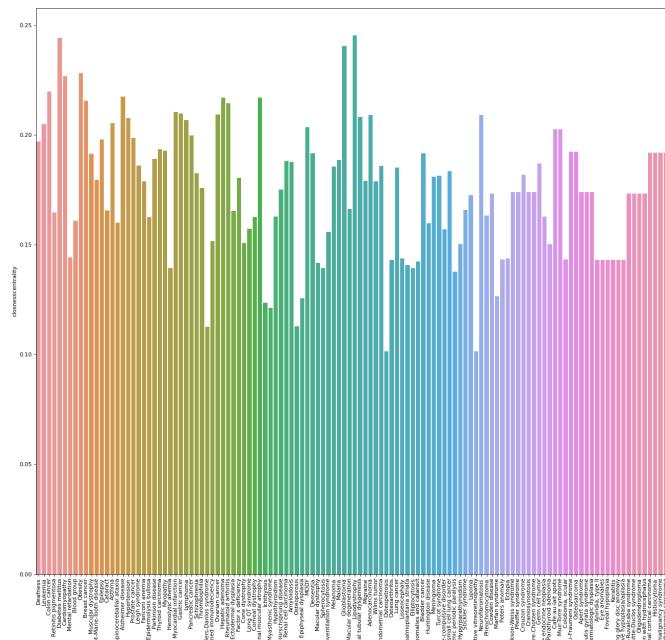


Figura 17: Actores de la medida de centralidad de cercanía.

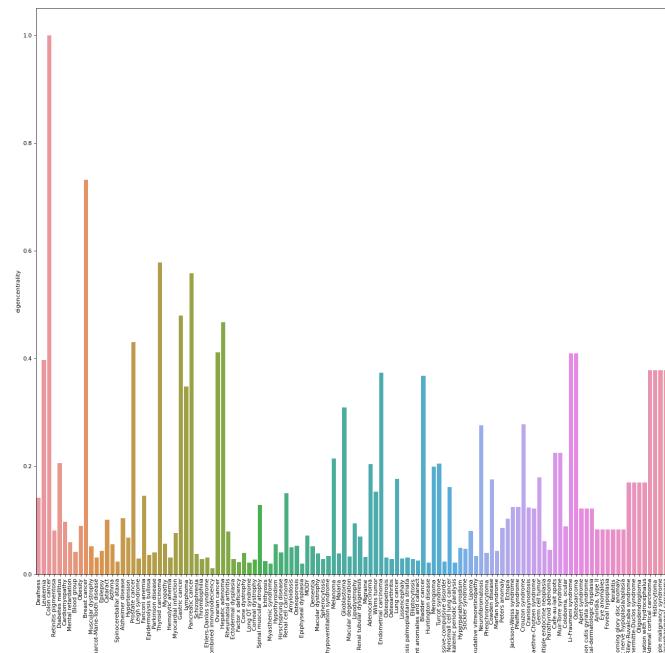


Figura 18: Actores de la medida de centralidad de vector propio.

Para realizar los gráficos anteriores, que representan los valores de las medidas para todos los actores de las enfermedades de la red en ejes de coordenadas, donde se han excluido algunas enfermedades y todos los genes, para poder mejorar las visualizaciones, se ha utilizado un *script* en Python que se incluye en los ficheros de la práctica y que también podemos ver a continuación:

```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

def createGraph(values , col):
    plt.figure(num=None, figsize=(20, 18), dpi=80, facecolor='w' , edgecolor='r')
    g = sns.barplot(x="Label" , y=col , data=values)
    plt.xticks(rotation=90)
    plt.savefig( 'images/{}.png' .format(col))

if __name__ == '__main__':
    df = pd.read_csv("dataset/diseasome.csv" , sep=",")
    df_disease = df[df[ '0' ].str.contains("disease")]
    df_disease = df_disease[df_disease[ "degree" ] > 15]

    colNames = [ "degree" , "betweenesscentrality" , "closenesscentrality" , "eigencentrality" ]
    for colName in colNames:
        createGraph(df_disease , colName)
```

Bibliografía

- [1] Kwang-II Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.
- [2] Mathieu Bastian, Sébastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009.
- [3] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.
- [4] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.