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

12 de enero 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 nos indican el origen genético común de muchas enfermedades. La forman 526 enfermedades y 903 genes, donde los genes asociados con trastornos similares muestran una mayor probabilidad de interacciones físicas entre sus productos y una mayor similitud de perfiles de expresión para sus transcripciones, lo que respalda la existencia de distintos módulos funcionales específicos de la enfermedad.

El conjunto de datos de *Diseasome* viene como un archivo *.zip*, que se puede descargar en el siguiente enlace: `http://gephi.org/datasets/diseasome.gexf.zip`. Una vez descargado y descomprimido, obtenemos un archivo *.gexf*, que contiene un archivo de grafos. Importamos el archivo de grafos a *Gephi*[2] y comenzamos a probar diferentes opciones de visualización.

Después de probar diferentes visualizaciones encojemos el algoritmo de distribución: Fruchterman Reingold (en la ventada *Distribución*). Para evitar que las componentes conexas queden fuera de la vista principal, fijamos el valor del parámetro *Gravedad* a 20. También marcamos las opciones *Disuadir Hubs y/o Evitar el solapamiento*. Esto convierte la visualización en un círculo y coloca la red alrededor de la misma área (ver Figura 1).

De aquí pasamos a la detección de comunidades para colorear los clústers de la red. *Gephi* implementa el método Louvain disponible en el panel de *Estadísticas*. Damos clic en ejecutar *Modularidad* y veremos como el algoritmo de detección de comunidades nos ha creado un nuevo parámetro de particionamiento (*Modularity Class*). Si seleccionamos este nuevo parámetro observaremos 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 vean bien donde se encuentra cada comunidad.

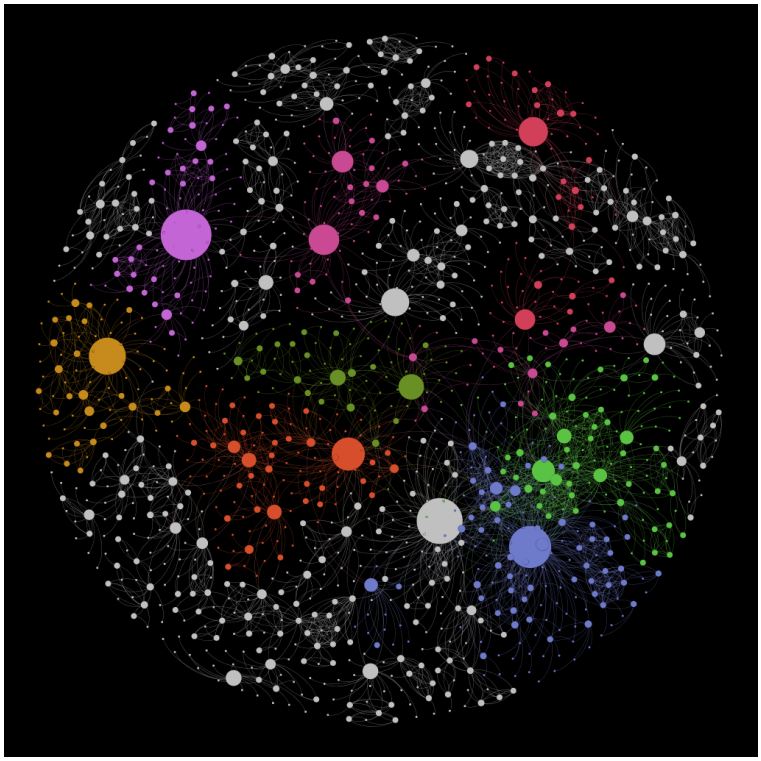

Figura 1: Red completa sobre un fondo negro sin etiquetas.

A continuación, añadiremos las etiquetas a los nodos para ver la red con más a detalle (ver Figura 2).

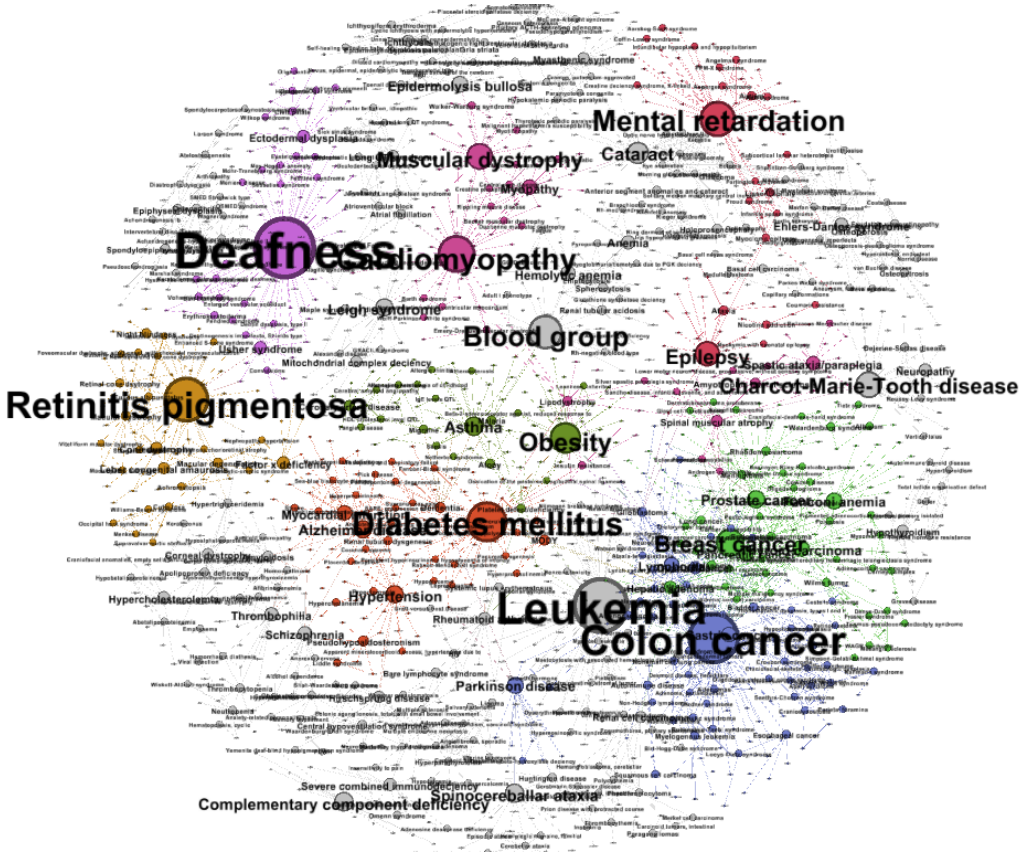

Figura 2: Red completa sobre un fondo blanco con etiquetas.

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, es una de las enfermedades más comunes en comparación con otras que existen en la actualidad. La sordera es la enfermedad que se lleva la mayor porción. También vemos diferentes clústers además de los cánceres, como la diabetes, la salud mental, etc.

Análisis Básico de la Red

Para los primeros pasos del análisis de la red, comenzamos por anotar los valores de las medidas globales básicas: el número de nodos 1419 (N) y el número de enlaces 3926 (L), que aparecen directamente en la ventana Contexto. 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 <k>, ejecutando la opción correspondiente en la ventana *Estadísticas*. El valor del grado medio <k> es 5,533. Al realizar el cálculo del grado medio, también obtenemos la distribución de grados de la red completa (ver Figura 3).

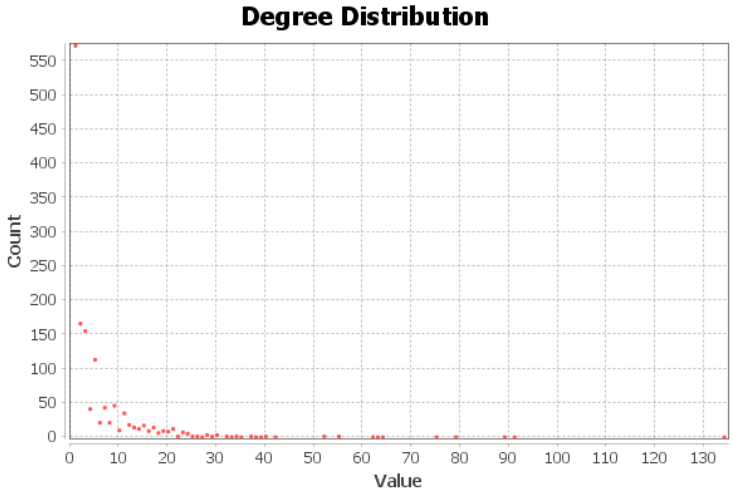

Figura 3: Distribución de grados de la red completa.

La opción *Densidad* de grafo mide la relación entre número de enlaces L y el número máximo de enlaces L_{max}. La ejecutamos y vemos que su valor es 0,004.

A continuación, ejecutamos la opción *Coeficiente medio de clustering* para obtener la medida del mismo nombre, <C>. El valor del coeficiente medio de clustering <C> es 0,819. Al realizar el cálculo del coeficiente medio de clustering, también obtenemos la distribución de coeficientes de clustering de la red completa (ver Figura 4).

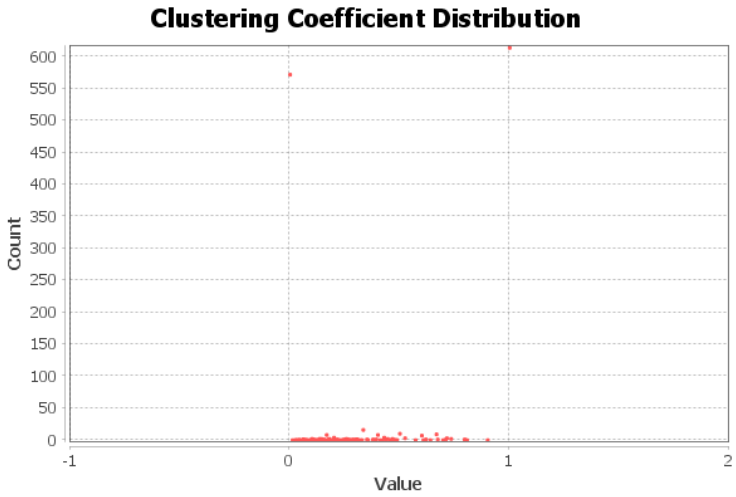

Figura 4: Distribución de coeficientes de clustering de la red completa.

Pasamos a analizar la conectividad de la red. En primer lugar, obtenemos el número de componentes conexas ejecutando la opción *Componentes conexos*. Vemos que el número de componentes conexas es 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 d) ejecutando la opción correspondiente al *Diámetro de la red* en la ventana Estadísticas. El valor del diámetro (d_{max}) es igual a 15. El cálculo del diámetro también nos proporciona el valor de la distancia media (d), 6.783, así como el valor de tres medidas de centralidad (intermediación, cercanía y excentricidad), que podemos observar en la siguientes figuras.

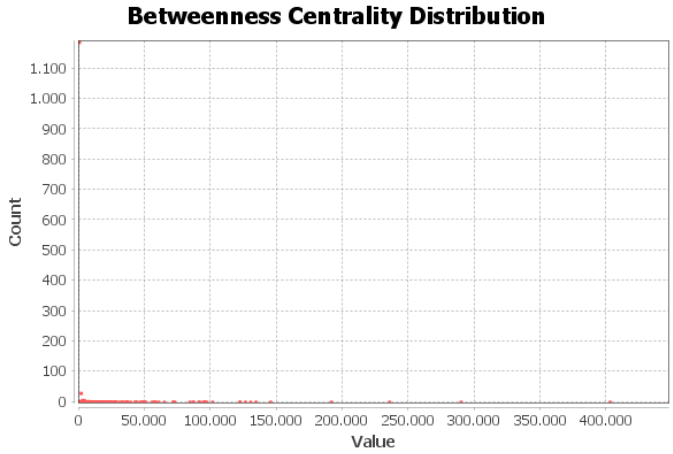

Figura 5: Intermediación de la red completa.

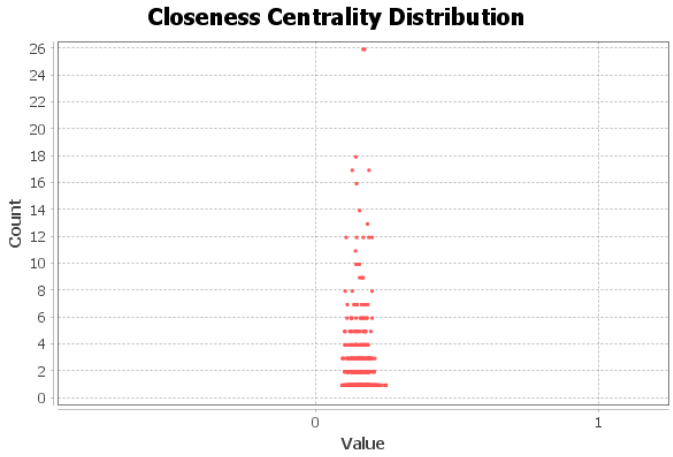

Figura 6: Cercanía de la red completa.

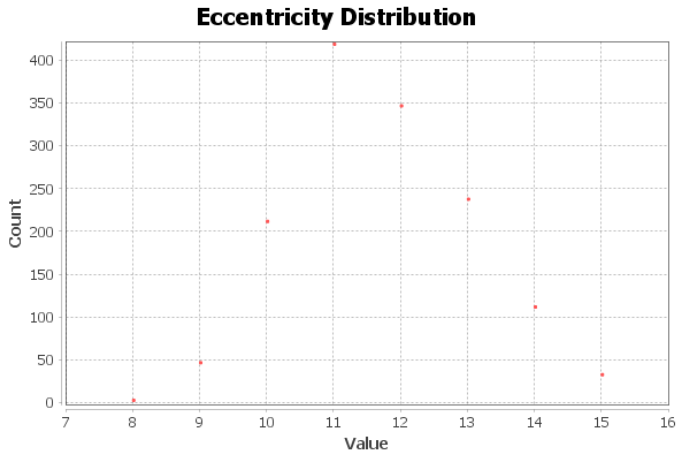

Figura 7: Excentricidad de la red completa.

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 grado, intermediación, cercanía y vector propio.

El valor de tres de estas medidas ya está calculado en los pasos que se han realizado en la sección anterior. La centralidad de grado (no normalizada) se generó al calcular el *Grado medio* en la ventana *Estadísticas*. Las medidas 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]*. El resultado de la repetición de cálculo lo podemos visualizar en las siguientes figuras.

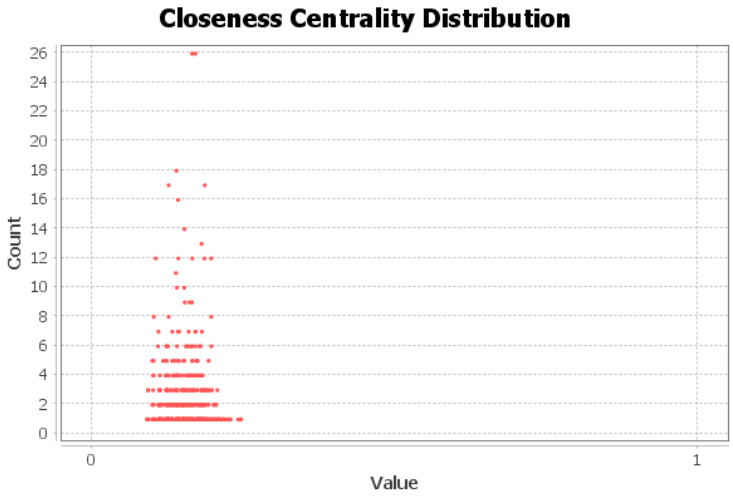

Figura 8: Cercanía normalizada de la red completa.

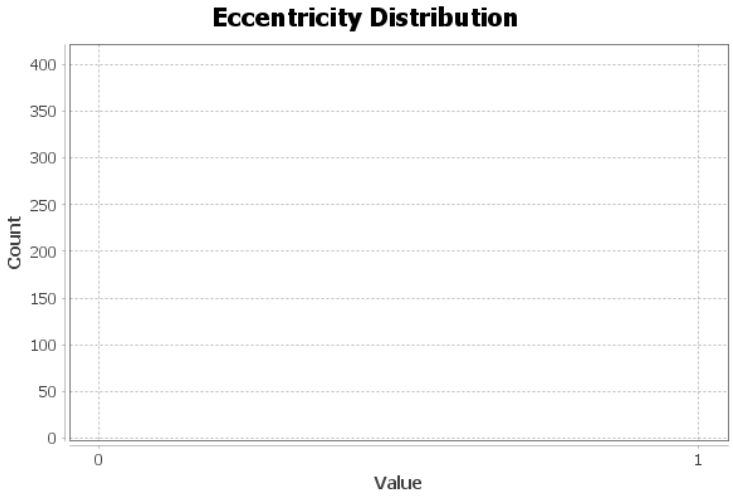

Figura 9: Excentricidad normalizada de la red completa.

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

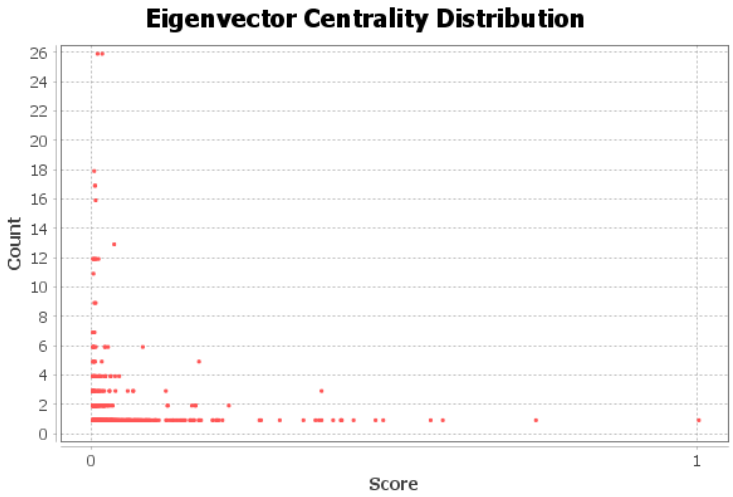

Figura 10: Centralidad de vector propio de la red completa.

Visualizaciones y Gráficos adicionales

Bibliografía

[1] Kwang-Il 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, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009.