UNIVERSIDAD DE GRANADA

Gestión de Información en la Web

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

2019-2020

Máster en Ingeniería Informática

Lidia Sánchez Mérida

lidiasm96@correo.ugr.es

Índice de contenidos

Red social

Para realizar esta primera práctica se ha escogido un gráfico web procedente de uno de los repositorios propuestos en el guión. En particular he seleccionado una denominada EPA, la cual recopila las referencias que contienen las páginas web asociadas a la página americana EPA *(United States Environmental Protection Agency)*. Esta es la web oficial de la agencia federal gubernamental de la protección y salud de los seres humanos.

Al descargarnos los datos podemos observar que la información se encuentra en un fichero con extensión *.edge*, el cual dispone de hasta dos columnas en las que se establecen los enlaces entre las distintas páginas. De este modo, cada sitio web dispone de las páginas a las que referencia y de aquellas que lo referencia a él. Como casi cualquier red de webs, es un **grafo dirigido**, por lo que conocemos el orden en el que se realizan las interacciones entre las distintas webs participantes.

Análisis de la red social

En esta sección se van a detallar las medidas y características tanto de la red al completo como de su componente gigante. A continuación se adjunta la tabla con sus valores.

Medida	Valor
Número de nodos N	4271
Número de enlaces L	8965
Número máximo de enlaces Lmax	80362260
Densidad del grafo L/Lmax	0,000491578
Grado medio <k>	2,099
Diámetro dmax	9
Distancia media d	2,629
Coeficiente medio de clustering <C>	0,036
Número de componentes conexas	8
Número de nodos componente gigante (y %)	4253 (99,58%)
Número de aristas componente gigante (y %)	8953 (99,87%)

Tabla 1. Medidas de la red.

Como podemos observar, la red global dispone de un total de **4.271 nodos**, que se corresponden con las páginas web que hacen referencia a la web americana EPA, mientras que por otro lado tiene hasta **8.965 aristas**, que se corresponden con los hiperenlaces que las conectan entre sí. No obstante, teniendo en cuenta que es un grafo dirigido el número **máximo de aristas** que podría tener es de **80.362.260**, calculado mediante esta fórmula [1]:

$$(N * (N - 1))$$

siendo N el número de nodos de la red.

Al calcular la **densidad** de la red con Gephi el resultado fue 0,000. Una densidad de 0 supondría que todos los nodos están aislados, y sin embargo observando la red esto no es así. Simplemente este valor es tan súmamente bajo que creo que Gephi no es capaz de calcularlo. Por ello, lo he calculado manualmente siguiendo esta fórmula específica para grafos dirigidos [2]:

$$\frac{|E|}{|V| \cdot (|V| - 1)}$$

siendo *E* el número de enlaces que tiene la red mientras que *V* es el número de nodos. De este modo he logrado obtener una densidad más concreta de **0,000491578**. Como podemos apreciar es bastante ínfima y por lo pronto, nos indica que la red dispone de una conectividad bastante baja, por lo que hay muchos nodos aislados. Como es una medida bastante teórica comprobaremos esta conjetura con medidas como el grado y la distancia. En el primer caso, el valor del **grado medio** también es bastante bajo, indicando que cada nodo, de media, dispone de un máximo de dos conexiones. Con lo cual cada página se encuentra asociada de media a otras dos webs.

Si observamos el histograma acerca de la distribución de los grados de **entrada** podemos observar que la mayoría de las páginas son muy poco referenciadas, puesto que es en el lateral izquierdo donde se concentran la gran mayoría de ellas. No obstante, existen algunas excepciones que sí disponen de un número alto de referencias, el cual llega hasta aproximadamente 120 si observamos el último punto situado en el lateral derecho. Estos son los denominados ***hubs***, que en nuestro caso son páginas webs que disponen de un alto número de referencias por parte de otros sitios.

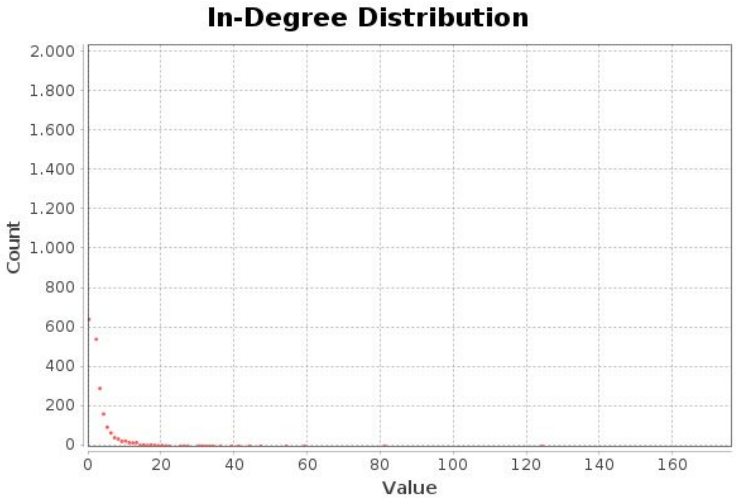

Figura 1. Histograma de la distribución de grados de entrada.

En relación a la distribución de los grados de **salida** podemos apreciar una tendencia similar, puesto que la mayoría de sitios webs, en este caso, no referencian a una gran cantidad de webs. Sin embargo, a diferencia del caso anterior podemos notar que hay una mayor variedad de páginas que referencian a un mayor número de páginas comparado con la media. Incluso, los ***hubs*** aumentan su coeficiente puesto que existen páginas que referencian a un gran número de otros sitios webs. La posible explicación a este hecho podría fundamentarse en que algunas de estas páginas son *blogs* o artículos de investigación que tienen una amplia bibliografía y por lo tanto su grado medio de salida es más alto. Sin embargo, en caso de que no sean muy conocidos puede suceder que estas páginas no sean referenciadas y de ahí que su grado de entrada sea inferior. No obstante, todas teorías son conjeturas puesto que no disponemos de más información acerca de la naturaleza de las páginas que representan los nodos en esta red.

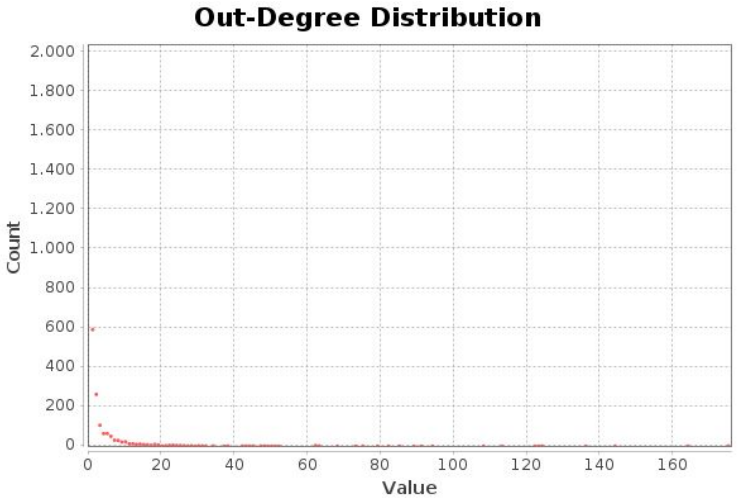

Figura 2. Histograma de la distribución de grados de salida.

En relación a las medidas de longitud, en primer lugar disponemos de un **diámetro** que indica que solo hacen nueve diez nodos para poder alcanzar el nodo más lejano, lo cual podemos determinar que es una distancia bastante corta dada la cantidad de nodos en la red. Asimismo, la **distancia media** de 2,63 nodos nos indica el camino más corto para llegar de una página a otra. Al ser sendos valores bastante menores con respecto al número de nodos y conexiones, podemos determinar que en esta red se podría producir un contagio muy rápido puesto que es bastante sencillo alcanzar todas las páginas de la red.

Si observamos el gráfico de la distribución de las distancias podemos apreciar que la gran mayoría de páginas tienen una distancia de 0. Con una alta probabilidad, estos sitios webs serán los que formen parte de la componente gigante, que como hemos podido visualizar anteriormente, abarca casi la totalidad de la red. Posteriormente podemos observar que un menor número de páginas se encuentran a 1 nodo de distancia, las cuales podemos intuir que están cerca de dicha componente gigante.

En este caso, si bien la densidad indicaba una conectividad baja en la red, las distancias tanto para recorrer el camino mínimo como para alcanzar el nodo más lejano no son considerablemente altas. En el primer caso basta con desplazarse un máximo de 2 nodos mientras que para el camino más largo basta con navegar unas 9 páginas. Sin embargo, la densidad es una medida un poco engañosa puesto que el número posible de enlaces puede ser muy alto mientras que el número real de páginas conectadas puede ser bajo, es decir, depende de las referencias que se efectúen entre sí. Por lo tanto, en base a estos resultados, mi teoría reside en que la red tiene una conectividad muy baja, por lo que cada página referencia y/o es referenciada pocas veces. Sin embargo, como veremos posteriormente, no existen demasiados nodos aislados.

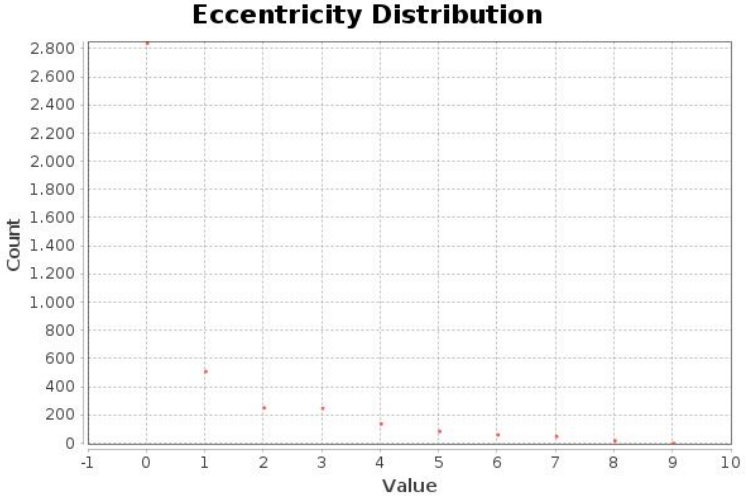

Figura 3. Histograma de distancias.

Por otro lado disponemos del **coeficiente de clustering** cuyo valor es considerablemente bajo. Este indicador refuerza la teoría de que la mayor parte de las páginas son vagamente referenciadas y/o disponen de muy pocas menciones por parte de otras webs. Si observamos la gráfica conseguida a través de *Google Calcs* [3] podemos apreciar de forma gráfica lo explicado anteriormente, puesto que la gran mayoría de páginas tienen un coeficiente de clustering cercano a 0. Esto nos indica que la conectividad a nivel local es

muy baja, ya que como hemos podido comprobar con el grado medio la gran mayoría de nodos solo están conectados con otros dos de media. Asimismo, algunos nodos se alejan bastante del valor medio situándose en 0,5 de coeficiente de clustering. Estos se corresponden con aquellas páginas que están conectadas con sus vecinas, pero que no actúan como nodos *hubs*, de modo que son pequeños grupos interconectados que se forman a lo largo de la red.

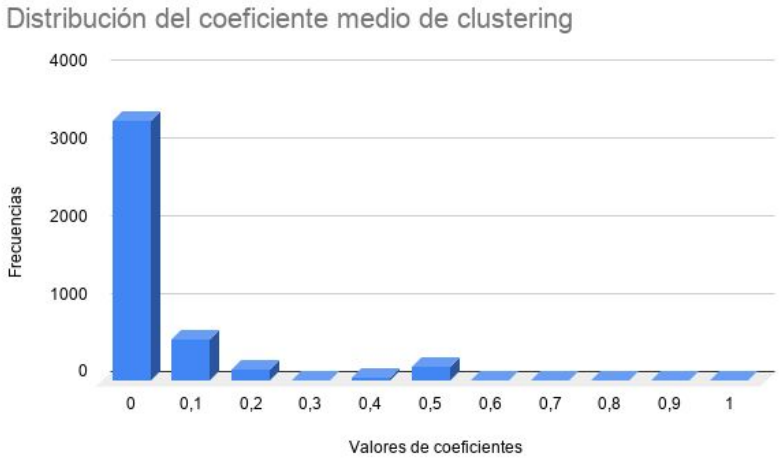

Figura 4. Distribución de los nodos según el coeficiente de clustering.

En cuanto a la conectividad de la red podemos observar en la tabla que existen hasta **8 componentes conexas**, lo cual nos indica que esta red se puede dividir en ocho subgrafos. El mayor de ellos, conocido como **componente gigante**, abarca el 99,58% de los nodos y el 99,87% del total de las aristas. Esto provoca dos consecuencias: por un lado podemos determinar que las siete componentes restantes estarán formadas por nodos aislados, mientras que por otro podemos confirmar que la red global y la red de la componente gigante son sumamente similares. Este hecho lo podemos visualizar en las dos capturas adjuntas a continuación de, en primer lugar la red global, y en segundo lugar la componente gigante.

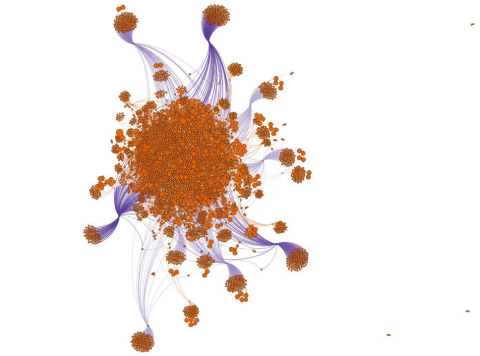

Figura 5. Red completa.

En esta primera imagen podemos observar la red la forma que tiene la **red completa**. El **color** de los nodos se debe al **grado medio** de los mismos, siendo los naranjas aquellos con menor valor, los cuales prácticamente engloban la totalidad de la red, y los morados aquellos con un mayor número de conexiones. Si hacemos zoom sobre la imagen, podemos visualizar hasta cinco nodos de este último color situados como nexo entre el centro de la red y las subredes que se sitúan alrededor. Estos son los nodos *hubs* cuyo objetivo consiste en conectar los nodos contenidos en el centro con los que se localizan en la periferia. Estas páginas son las que sirven de enlace entre el grupo mayoritario y los subgrupos de alrededor.
El **tamaño** de los nodos viene determinado por el **coeficiente de clustering**, por lo que aquellos nodos más grandes son los que tienen una mayor probabilidad de conectar con sus vecinos. De nuevo, si hacemos zoom sobre la imagen podemos observar que la mayoría cuenta con el mismo tamaño, exceptuando algunos que se sitúan en el centro de la red pues disponen de un mayor número de vecinos al que poder estar conectados. Mientras que aquellos con un tamaño menor se suelen localizar en la periferia de la red puesto que están más aislados. Estos se corresponden con aquellas páginas que no hacen referencia a muchas páginas y que tampoco son referenciadas por otros sitios webs.

En la siguiente imagen se puede visualizar la **componente gigante** desde otra perspectiva para destacar, aún más, los nodos *hubs* que conectan las páginas situadas en el centro del grafo con las que se encuentran más alejadas. Estas las podemos visualizar en colores rojizos, mientras que el resto de nodos azules son aquellos que tienen un grado menor, y que por tanto, no disponen de un número considerable de conexiones.
Para el tamaño he vuelto a escoger el **coeficiente de clustering** para destacar a aquellas páginas que tienen una conectividad local más alta. Como hemos mencionado anteriormente, esta cualidad no está asociada a los *hubs* puesto que su trabajo es conectar páginas distantes entre sí, pero su conectividad local es baja.

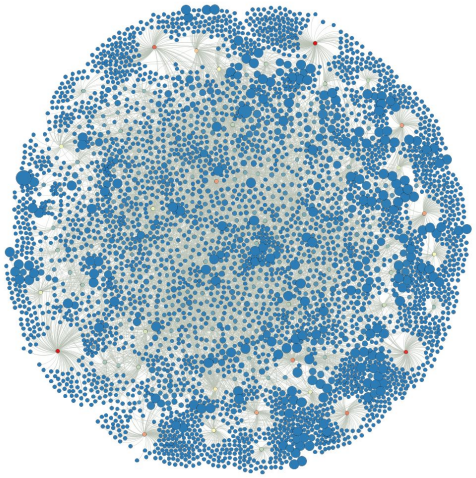

Figura 6. Red de la componente gigante circular.

No obstante, como ya adelantamos anteriormente la componente gigante es muy parecida a la red total puesto que abarca más del 99% de los nodos y aristas. Por lo tanto, en la siguiente captura en la que se representa esta componente gigante con la misma configuración de color y tamaño que dispuse para la red total, es bastante difícil encontrar algunas diferencias. Sin embargo, si hacemos zoom sobre la imagen podemos observar que los **nodos aislados** que se encontraban en la periferia de la red total no aparecen en la componente gigante. Este hecho nos indica que los restantes siete subgrafos en los que se puede dividir la red solo disponen de nodos aislados.

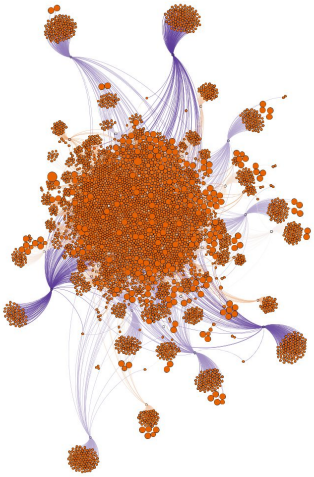

Figura 7. Componente gigante con la misma configuración de color y tamaño que la red total.

Conclusiones

El análisis de esta red particular nos ha desvelado el comportamiento de las páginas asociadas a la EPA en relación a las referencias que realizan y que les son realizadas. Si bien la mayoría dispone de un número bajo de ellas, al menos tienen una que las liga a la componente gigante puesto que esta abarca casi la totalidad de la red global. Este indicio significa que la mayoría de las páginas vinculadas con la EPA se conocen entre sí y por lo tanto todas se encuentran conectadas. Sin embargo, algunas de ellas se encuentran bastante aisladas siendo necesaria la navegación de hasta diez páginas para poder acceder a ellas. Para este suceso se puede plantear la teoría de que o bien son sitios webs nuevos que aún no son conocidos por el resto de páginas o no disponen de interés suficiente para que el resto desee conectar con ellos.

Bibliografía

1. Stackoverflow, *What is the maximum number of edges in a directed graph with n nodes?*,

https://stackoverflow.com/questions/5058406/what-is-the-maximum-number-of-edges-in-a-directed-graph-with-n-nodes

2. Stackoverflow, *What is the definition of the density of a graph?*, https://math.stackexchange.com/questions/1526372/what-is-the-definition-of-the-density-of-a-graph
3. Documentación sobre la función *FRECUENCIA,* https://support.google.com/docs/answer/3094286?hl=es