PRÁCTICA 2: ANÁLISIS Y EVALUACIÓN DE REDES EN TWITTER, GIW

Manuel Blanco Rienda

MASTER EN INGENIERIA INFORMÁTICA, UGR

Informe de la práctica 2 por Manuel Blanco Rienda

(Dataset de Twitter del día del comienzo del proceso de Brexit de UK)

Descripción de la red a analizar

La red a analizar ha sido obtenida a través del uso de NodeXL como herramienta de descarga y almacenamiento de los datos de Twitter. Estos datos han sido almacenados en ficheros csv a partir de la diferenciación entre vértices y aristas que hace la plantilla de NodeXL asociada a la red. En cuanto al contenido del grupo de datos a analizar, proviene del contexto del día 29 de Marzo de 2017, día marcado por el comienzo del proceso de separación del Reino Unido de la Unión Europea. Es por ello que toda la información que se tratará a lo largo de esta práctica, debe ser contemplada desde el mencionado contexto. Teniendo en cuenta este contexto, la pregunta me cuestiono y que es la base de este trabajo es: ¿Cuáles son los usuarios más relevantes en este tema?

Representación de la red (Frutcherman-Reingold)

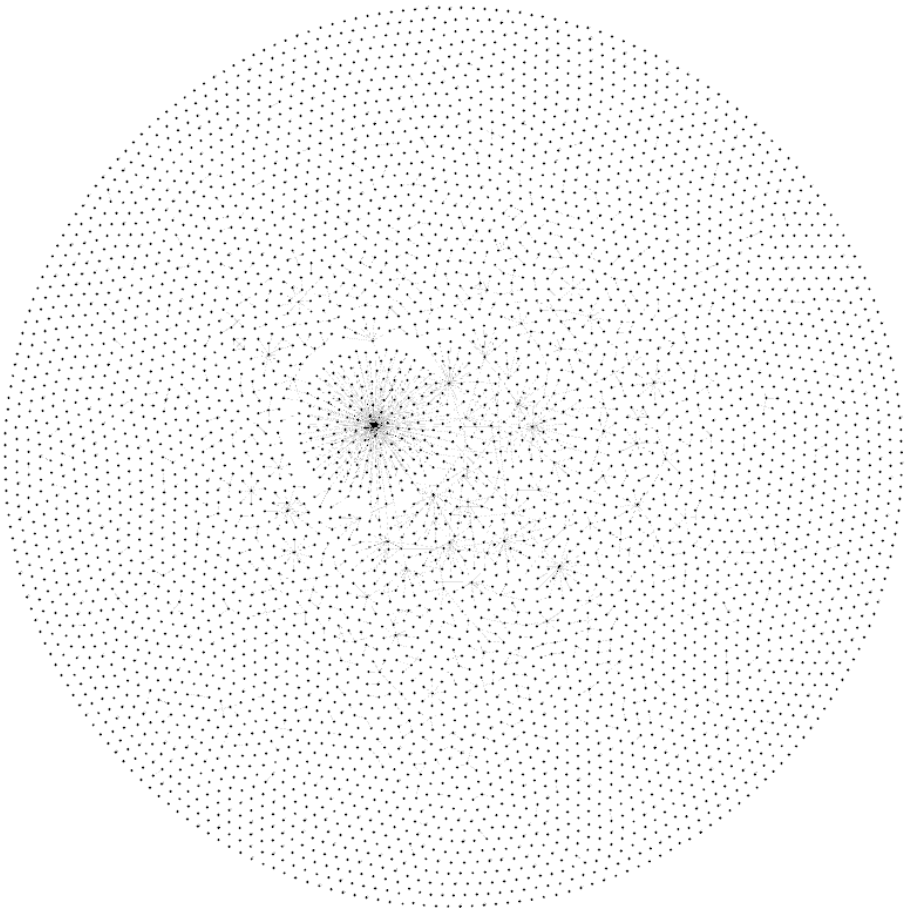

Figura 1: Representación de la red original según Frutcherman-Reingold

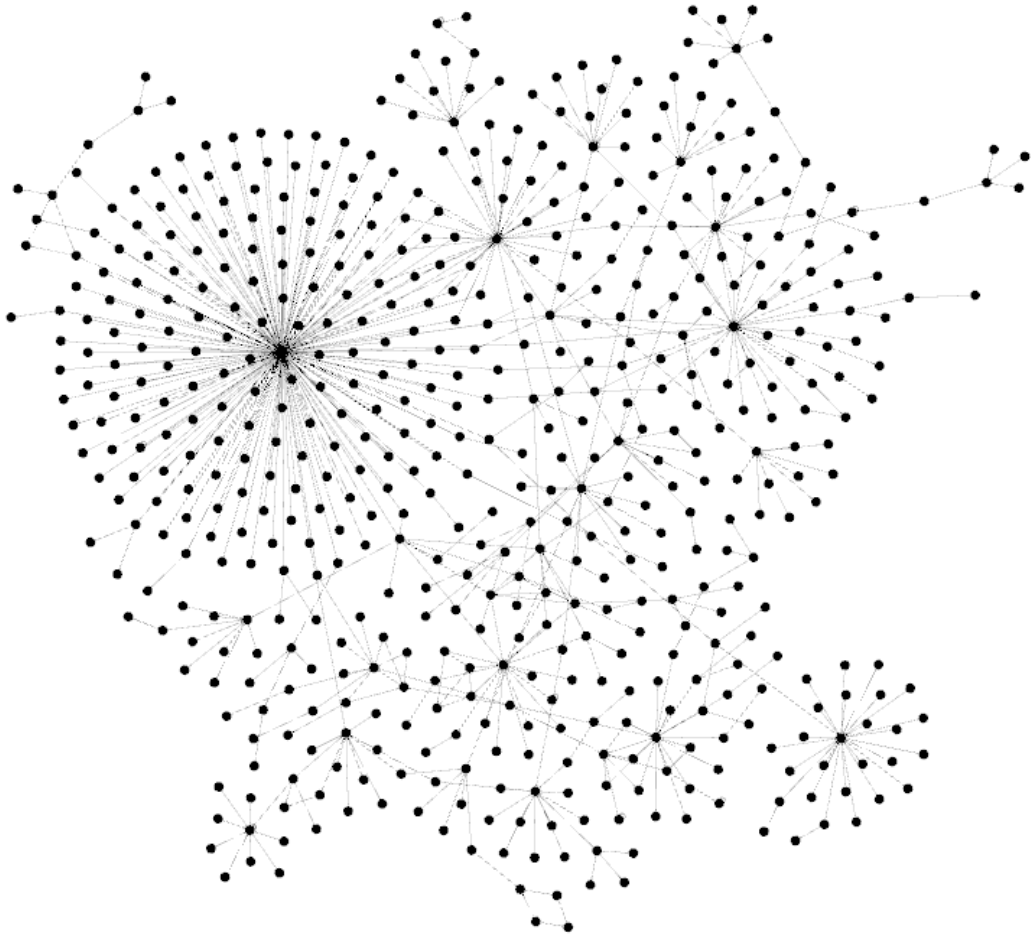

Figura 2: Representación de la componente gigante según Frutcherman-Reingold

Tal y como puede verse, tras realizar la representación de la componente gigante, vemos que ésta solo supone un 11,57% en nodos y un 29,51% en aristas. De este hecho podemos concluir que la mayor parte de la red que estudiamos se encuentra aislada del cluster principal de nodos. Ello se debe a que muchas de las interacciones con el hagstag “#Brexit” son individuales y no relacionales entre usuarios: Se trata de twits que sólo hacen referencia a temas relacionados con el brexit pero que no tienen conexión con los temas que habla el resto del mundo.

A pesar de que con la componente gigante obtenemos los nodos más relevantes de la red, no podemos podar casi un 90% de la misma porque perderíamos mucha información. Para ello, paso a aplicar otra serie de filtrados sobre el grupo, que poden la red, pero que no por ello la dejen incompleta de datos.

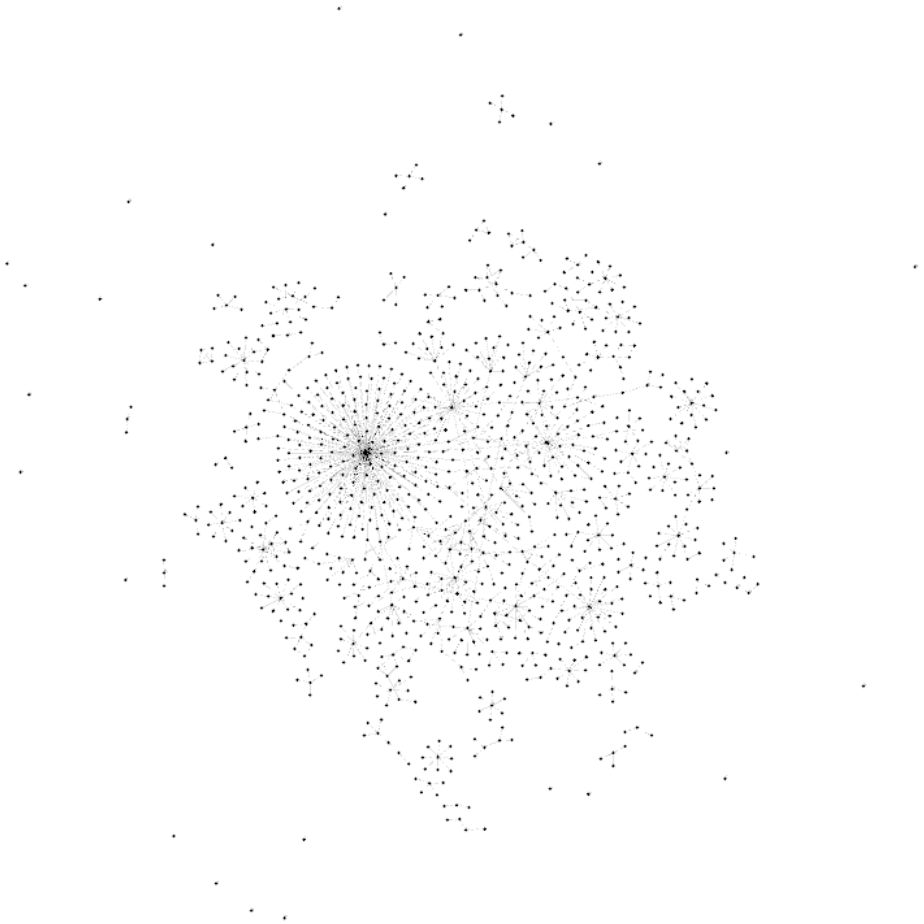

Figura 3: Representación de la red podada

Tras realizar varias pruebas de podado, se ha llegado a un modelo final para realizarlo: un doble filtrado (ya que no se trata de una técnica de podado al uso, sino que opto por filtrar los datos que ya tengo). Por un lado nos hemos quedado con los nodos más importantes de la red (aquellos que tienen un grado superior a 4, dado que la mayor parte de los nodos de la red son inconexos del resto o bien sólo se conectan con ellos mismos) y por otro, hemos añadido a la red resultante aquellos nodos que están conectados a una distancia de 3 o menos pasos. La red resultante después de la poda tiene 1036 nodos y 1187 aristas, representando un 21% y un 51% respectivamente, del total. A partir de ahora utilizaré esta red, de cara evitar la sobrecarga del programa Gephi al tratar con gran cantidad de los datos de la red.

Después de tener la red podada, paso a mostrar las etiquetas de los nodos en una proporción adecuada para que las principales sean visibles y vuelvo a aplicar Frutcherman-Reingold, para agrupar los nodos podados en una representación más cohesiva.

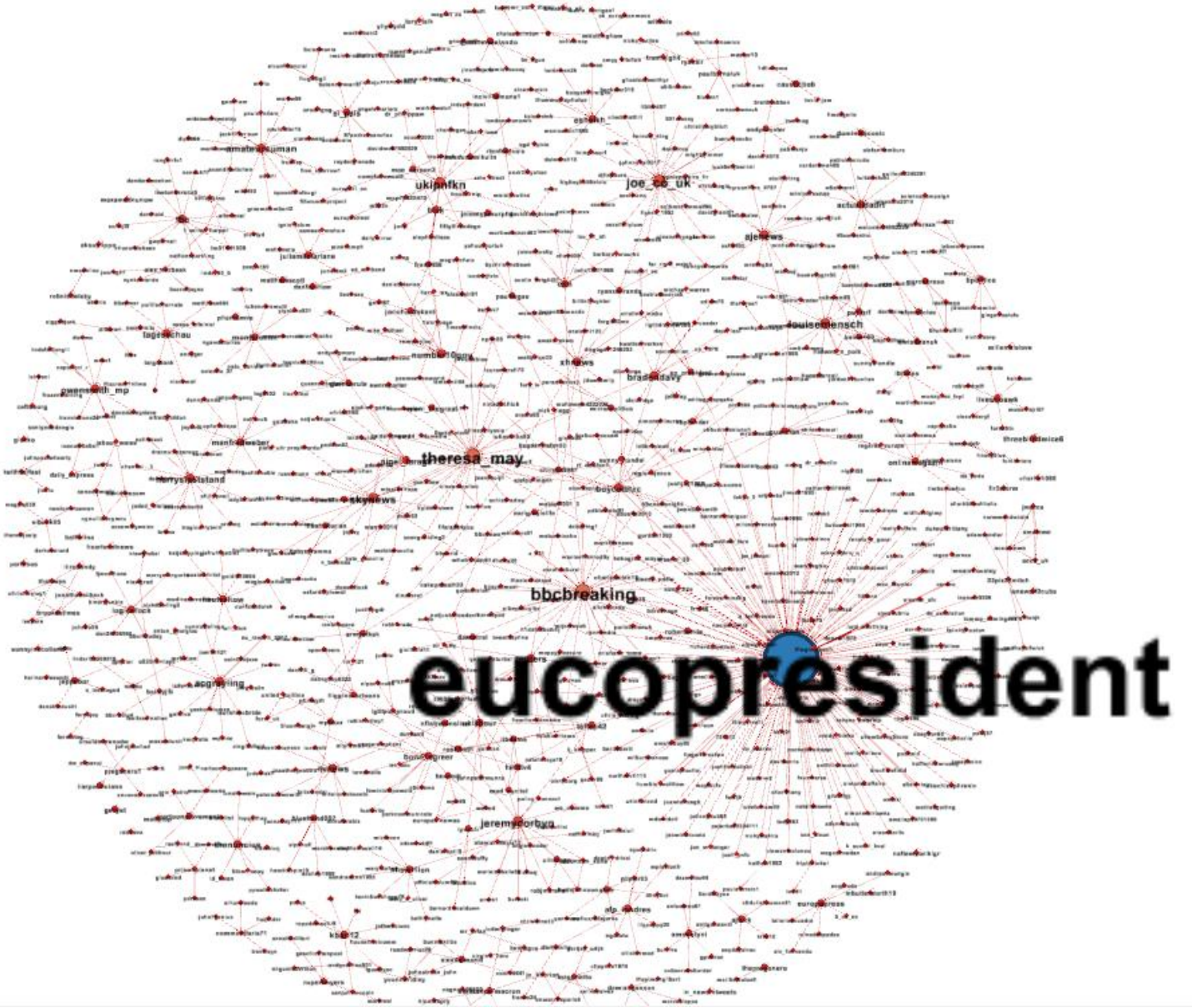

Figura 4: Representación de la red podada con etiquetas según Frutcherman-Reingold

<u>Datos Básicos (Red Original)</u>

Número de nodos: 4910

Aristas: 2301

Se trata de un grafo dirigido.

Cálculo de los valores de las medidas de análisis (Red Podada)

Número de nodos (N): 1036 (21%)

Aristas (L): 1187 (51%)

Se trata de un grafo dirigido.

Densidad (D): 0,001

Número máximo de conexiones posibles (Lmax): L/D => 1.187.000 conexiones máximas posibles.

Grado medio (K): 2,292

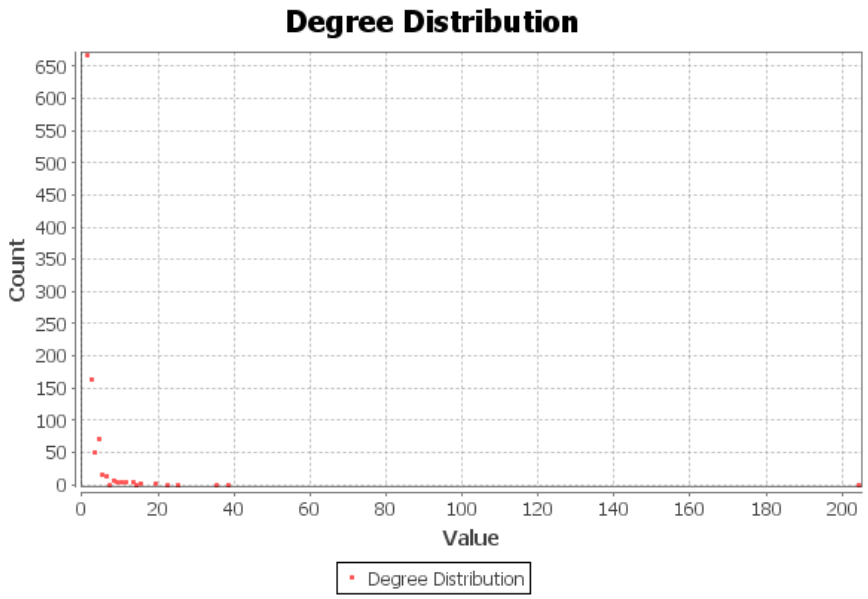

Figura 5: Distribución de grados de la red

Coeficiente de clustering: 0,027

Diámetro (dmax): 3

Radio: 0

Distancia media (d): 1,079

Distancia media para la red aleatoria equivalente (daleatoria): (ln N)/(ln K) => 8,37

Coeficiente de clustering medio para la red aleatoria equivalente (Caleatorio): (K/N) => 0,00221

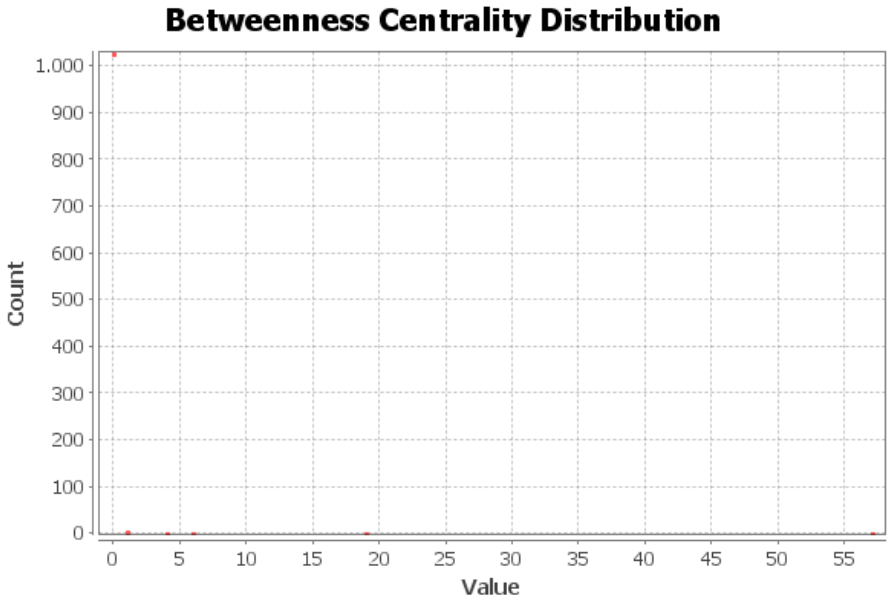

Figura 6: Distribución de distancias de la red

Análisis de la conectividad de la red original (sin podar)

Componentes conexas: 3160

Componente gigante

Número de nodos (N): 568 (11,57% visible)

Número de aristas (L): 679 (29,51% visible)

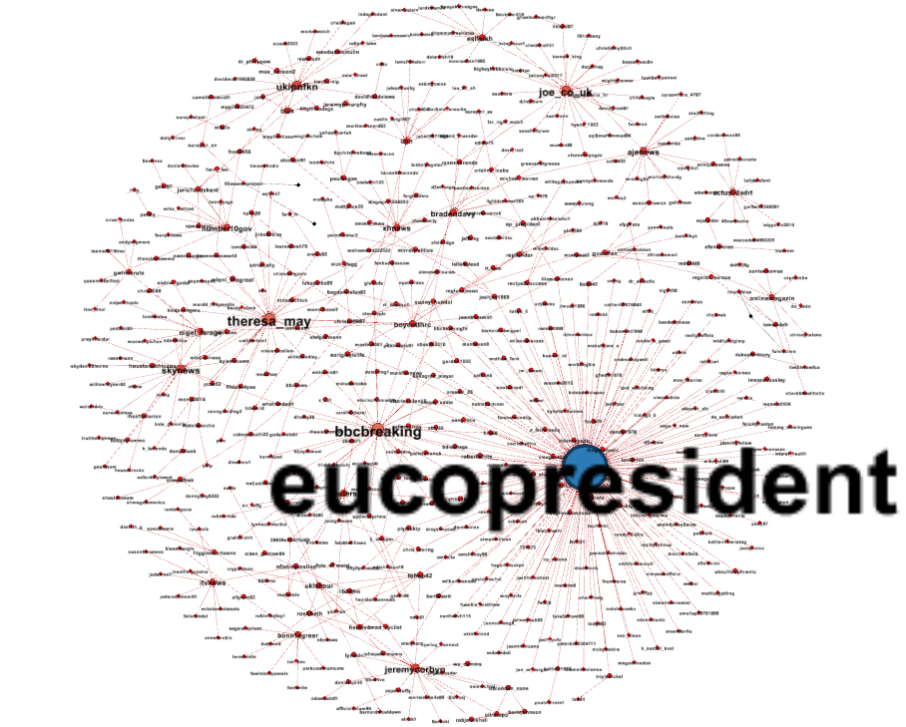

Figura 7: Representación de la componente gigante de la red original según Frutcherman-Reingold con etiqueta

Tabla resumen de las medidas realizadas

Medida	Valor
Número de nodos *N (red podada)*	1036
Número de enlaces *L (red podada)*	1187
Número máximo de enlaces L_{max} *(red podada)*	1.187.000
Densidad del grafo *L/Lmax (red podada)*	0,001
Grado medio *<k> (red podada)*	2,292
Diámetro d_{max} *(red podada)*	3
Distancia media *d (red podada)*	1,079
Distancia media para red aleatoria equivalente *(red podada)*	8,37
Coeficiente medio de clustering *<C> (red podada)*	0.027
C para red aleatoria equivalente *(red podada)*	0,00221
Número de componentes conexas (red original)	3160
Número de nodos componente gigante (y %) *(red original)*	568 (0.1157)
Número de aristas componente gigante (y %) *(red original)*	679 (0.2951)

Tabla 1: Resumen de las medidas realizadas a la red

Determinación de las propiedades de la red (Red Original)

Grado medio (K): 0,937

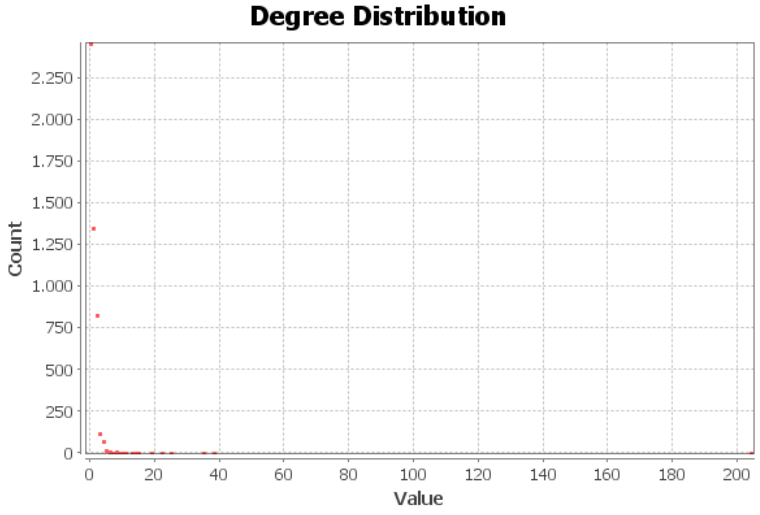

Figura 8: Distribución de grados de media de la red original

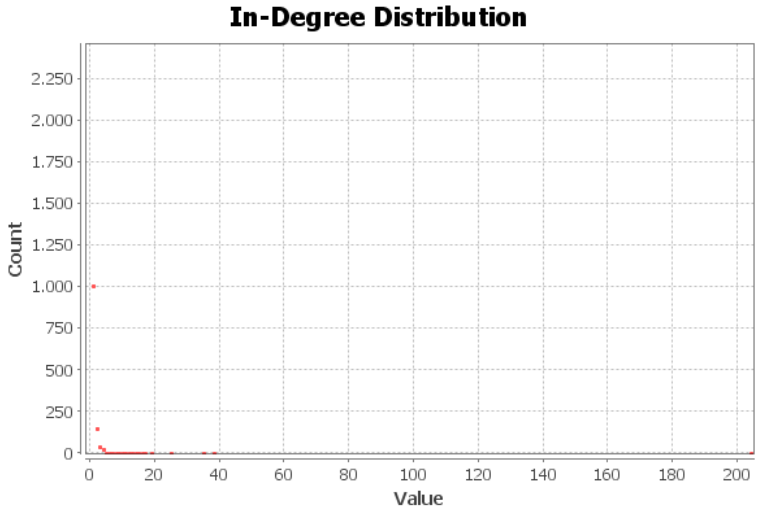

Figura 9: Distribución de grados de entrada de la red original

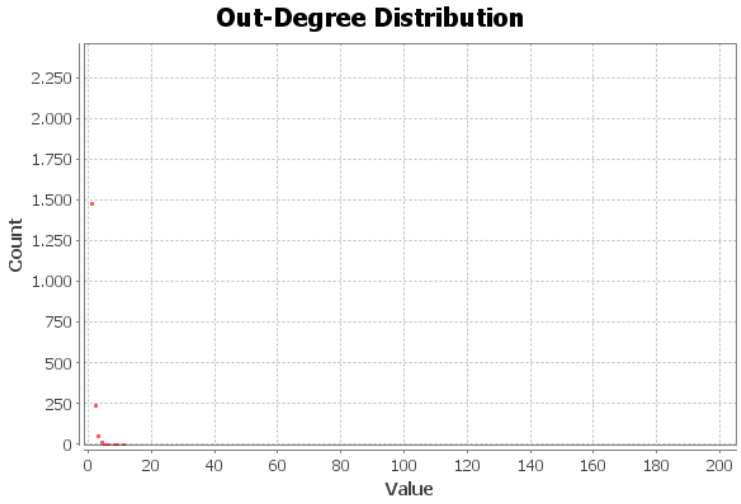

Figura 9: Distribución de grados de salida de la red original

Distancia media (d): 1,053

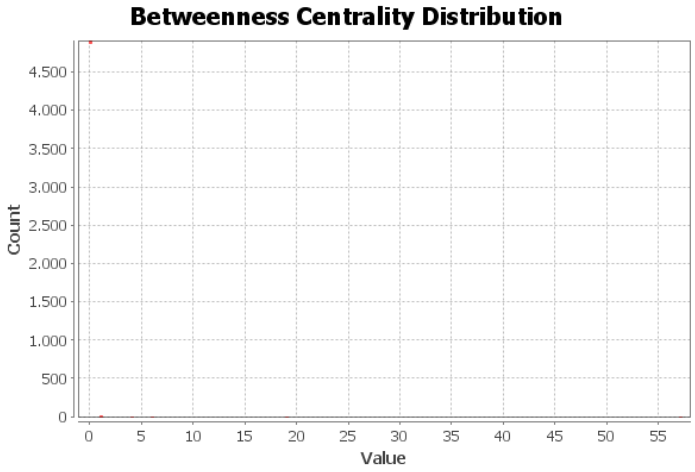

Figura 10: Distribución de distancias de la red original

Coeficiente de clustering: 0,01 (No se visualiza bien la gráfica)

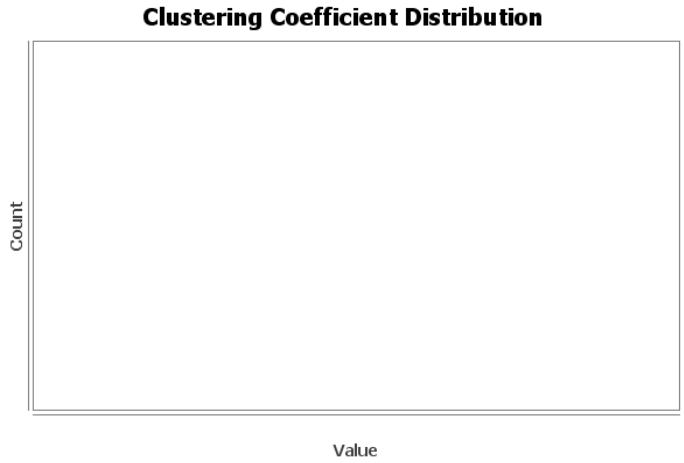

Figura 11: Distribución del coeficiente de clustering de la red original

¿Red libre de escala?

Una red libre de escala es aquella en la que algunos pocos nodos tienen un gran número de conexiones con otros nodos, pero en las que la mayor parte de los elementos que la componen tienen muy pocas o ningunas conexiones. Para demostrar que la red que estudio es de este tipo utilizo la ley de la potencia: $P(k) \sim k^{(-\gamma)}$, siendo γ un valor no constante que sin embargo suele tomar valores entre 2 y 3. La ley de la potencia establece que dado un cierto valor de γ, llega el momento que la curva de la ley de la potencia está por encima de la distribución de grados. Dado que es una red dirigida, realizamos esta comprobación para la distribución de grados de entrada y de salida:

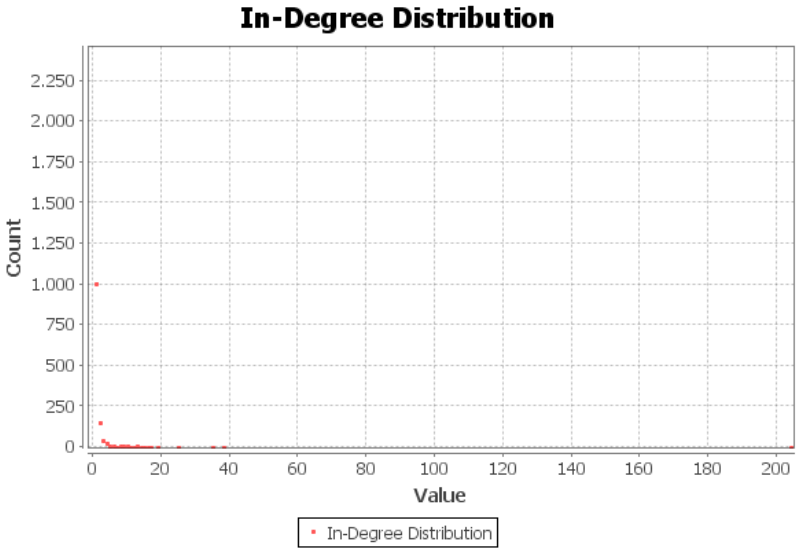

Figura 12: Distribución de grados de entrada de la red original

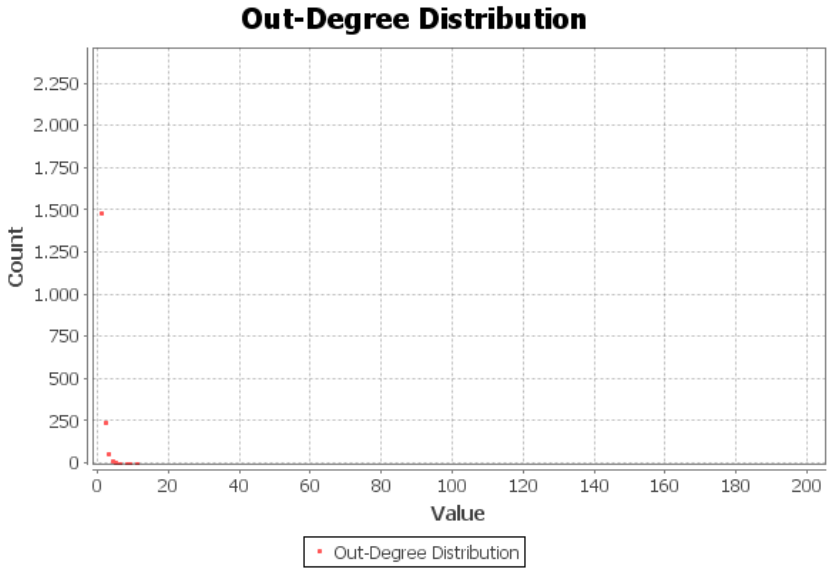

Figura 13: Distribución de grados de salida de la red original

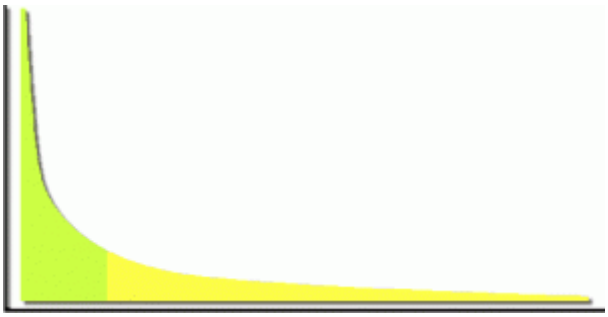

Figura 14: Gráfica de la ley de la potencia k^(-y)

Comparando ambas gráficas de distribución de grados con la que corresponde a la ley de la potencia, podemos ver claramente que se cumple la ley y que por tanto la red es libre de escala. Además, esto puede verse refutado en las representaciones gráficas de la propia red, en la que vemos unos pocos nodos con muchas conexiones mientras que la gran mayoría de ellos tienen muy pocas relaciones o ninguna (especialmente claro a la vista con Frutcherman-Reingold).

<u>¿Es un mundo pequeño?</u>

Para comprobar si mi red se trata de un mundo pequeño hay que tener en cuenta que la distancia media de la red tenga una escala logarítmica con respecto a la cantidad de nodos que tiene, o incluso menor:

- ln(N)/ln(<k>) = 10,24

Tal y como vemos, la distancia media (1,053) entra dentro del tipo de calificación de: mundo ultra-pequeño: comparado con el tamaño de la red, los nodos

se encuentran a distancias tan cercanas que hacen que el “mundo” que representa este conjunto de tweets sea muy pequeño.

Coeficiente de clustering medio

A partir del valor obtenido del coeficiente de clustering medio (0,01) es el momento de observar si estamos tratando con una red regular, aleatoria o de libre grado.

- ¿Red regular? Para ello, el coeficiente de clustering debería ser constante al tamaño de la red. Sin embargo, tras realizar varias pruebas con Gephi, en las que vamos podando de diferente forma la red original vemos que el coeficiente va variando, por lo que podemos decir que ésta no se trata de una red regular.

- ¿Red aleatoria? Para ello, el coeficiente de clustering debe estar alrededor de <k>/N => 0,937/4910 = 0,00019, es un valor bastante inferior al del coeficiente, por lo que no podemos decir que esté en su umbral y por tanto, la red no puede considerarse aleatoria.

- ¿Red de libre grado? Tal y como hemos visto antes, la red es de libre escala, ya que tiene unos pocos nodos con muchas conexiones con otros nodos y la mayor parte de los elementos que tienen muy pocas o nulas relaciones.

Cálculo de los valores de las medidas de análisis de redes sociales

Dado que la pregunta de investigación se centra en el descubrimiento de los principales actores dentro de la red social creada en torno al hagstag “Brexit”, es necesario analizar el grupo como red social.

Si prestamos atención a las medidas de grado, vemos que los principales actores de la red social (aquellos que destacan por nivel de grado por encima de todos) son: eucopresident (), theresa_may (), bbcbreaking (), joe_co_uk () y ukipnfkn() en ese orden (ya que pueden verse en las anteriores representaciones de la red, dibujada en función del grado de los nodos). Si por otro lado vemos la intermediación, se observa que la correspondencia con la mencionada técnica no es equivalente, ya que tenemos nodos que pese a tener un gran número de conexiones dichas relaciones no son las más importantes. Este es el caso de Louise Mensch, una periodista independiente inglesa con mucha reputación en Reino Unido y que fue miembro del grupo parlamentario conservador de Corbyn entre 2010 y 2012. Aquí podemos ver que aunque Donald Tusk (eucopresident), es el principal actor en la red en cuanto a conexiones con el resto de nodos que la forman, dichas conexiones no son tan relevantes como las que puede tener una periodista como la anteriormente mencionada.

0.0	(75,3%)
1.0	(24,03%)
0.6	(0,22%)
0.6666666666666666	(0,1%)
0.75	(0,1%)
0.5555555555555556	(0,08%)
0.625	(0,08%)
0.5	(0,02%)
0.6363636363636364	(0,02%)
0.7	(0,02%)
0.8333333333333334	(0,02%)

Figura 15: Valores de centralidad de los nodos

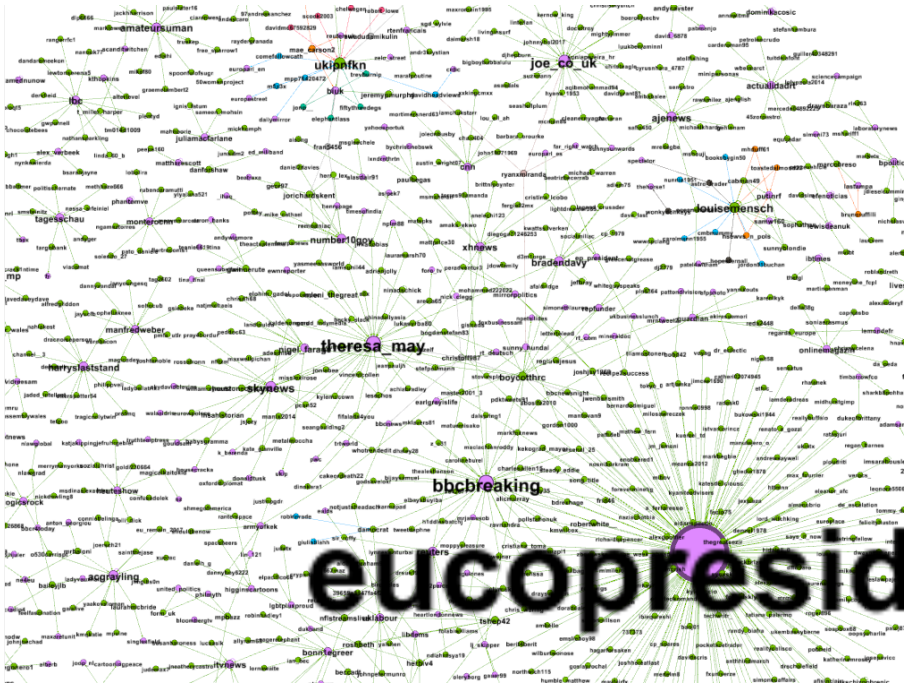

Figura 16: Visualización de la red con los colores de centralidad anteriores

Si por último hacemos un análisis de la centralidad de cercanía de nodos observamos que la mayor parte de los nodos de la red tienen un valor elevado para esta medida, dado que tal y como hemos demostrado a lo largo de esta práctica, la red está formada de pequeños grupos de gente aislada que aportan entre sí dentro de comunidades de amigos cercanas. Es por ello que los principales actores en cuanto a grado tienen valores muy pequeños para esta medida.

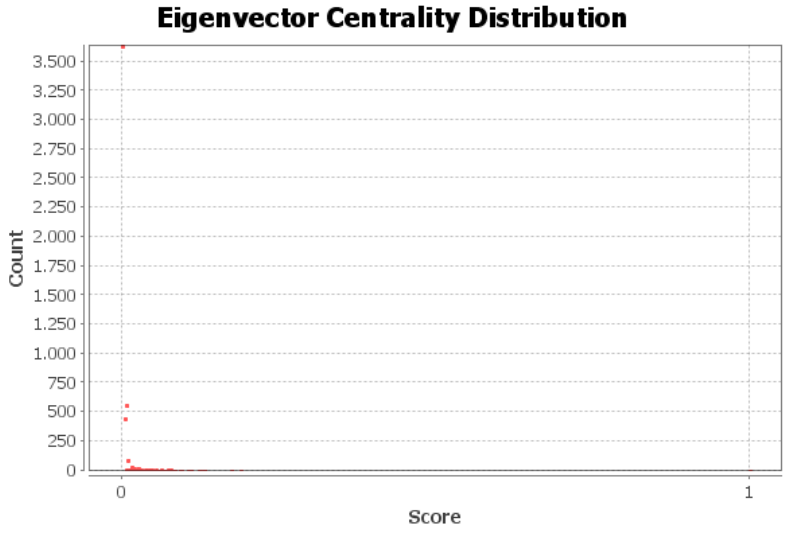

Figura 17: Distribución del vector de centralidad propio

Analizando las conclusiones a las que hemos llegado al aplicar las tres medidas de centralidad, podemos resumir que en este tipo de red, lo mejor para detectar los principales actores es usar en conjunción la intermediación y el grado, ya que son las dos herramientas que mejores resultados pueden arrojar en un contexto como es en el que se mueve este grupo.

Detección de comunidades

A través del uso de la utilidad “modularidad” de Gephi (la cuál es una implementación del algoritmo de Lovaina para detección de comunidades) obtengo la conclusión de que la red con la que estamos trabajando no tiene ninguna comunidad aparente (tanto la original como la podada). Si observamos la representación de la red, se puede ver que “casi” hay una comunidad diferenciada del resto de nodos que conforman la red: la que está formada por los principales políticos que participan en la ejecución del Brexit. Sin embargo, hay nodos que sirven de enlace con esta “casi comunidad” que evitan que finalmente lo sea. El nodo más llamativo de estos “usuarios puente” es el que corresponde con la cadena informativa del Reino Unido, la BBC, lo cual es perfectamente comprensible, ya que los enviados de la propia cadena se encontraban retransmitiendo el debate en el parlamento, desde dentro. Eran los conectores entre la cámara parlamentaria y el exterior.

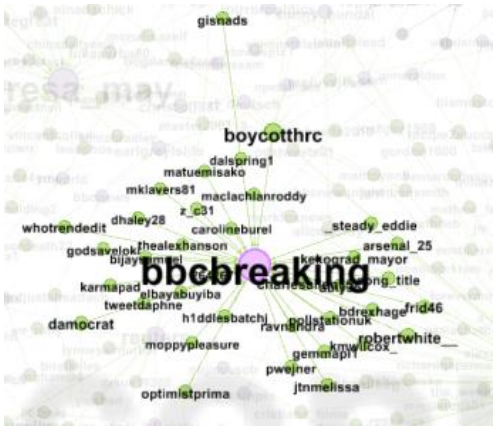

Figura 17: Nodo puente entre la “casi comunidad” que forman los políticos del brexit y el exterior

Discusión de los resultados obtenidos

Teniendo en cuenta los resultados obtenidos durante la realización del análisis de la red podada, podemos sacar las siguientes conclusiones:

- Cada usuario de la red tiene conexión de media con otro usuario de la misma. Dicho valor queda tan menguado a pesar de comprobar que hay varios nodos con muchas relaciones a causa de que la mayor parte de la red está conformada por usuarios que son inconexos con el resto: son personas que dan su opinión sobre el tema del que se habla: El “Brexit”, pero que están dentro de grupos sociales en los que no se habla de ese tema.

- Prestando atención a la distribución de grados, podemos ver que la mayor parte de los nodos que forman la red tienen muy pocas relaciones con los demás (1-3 conexiones), mientras que a parte se puede observar un hub muy grande: La cuenta de Donald Tusk, el actual presidente del Consejo Europeo, y varios más pequeños (la cadena BBC, Theresa May la principal impulsora del brexit, etc). Este principal Hub aglomera la mayor parte de conexiones, que pueden ser identificadas como menciones dentro de los tweets de los usuarios que centran su opinión del proceso de separación del Reino Unido, sobre el presidente del Consejo Europeo.

- Se han hallado una gran cantidad de componentes inconexas con el grupo principal de la red, lo que evidencia la no cohesividad de sus miembros. En unas circunstancias como las que he analizado: un hagstag relacionado con una noticia pública es un hecho que se ve claramente refutado, ya que la mayor parte de los nodos conforman opiniones de usuarios aislados que apuntan hacia los actores principales que forman parte de este proceso de separación del Reino Unido de Europa.

- El coeficiente de clustering (0,027 para la red podada y 0,01 para la red original) evidencia el hecho de que no hay grandes vecindarios de nodos que formen parte de la red: la mayor parte son nodos aislados entre sí que apuntan en una cierta dirección (ya que se trata de una red dirigida).

- La red en su conjunto representa un mundo muy pequeño, ya que los nodos conectados están a una distancia muy pequeña entre sí, por lo que todo está muy aglomerado. Si nos fijamos en la

distribución de distancias, queda refutado este hecho ya que se puede ver que hay muchos nodos que sólo tienen conexión consigo mismos: han dado su opinión sobre el tema del Brexit sin más (distancia 0) y otros que nombran a los actores principales del proceso de separación y que se encuentran a distancias muy cercanas: de 1 a 2 pasos. Además es una red libre de escala: unos pocos nodos con muchas conexiones y la mayoría con pocas.

- Para terminar, podemos responder a la pregunta inicial a causa de la cual se ha producido todo este trabajo: Los principales actores (aquellos más destacados descubiertos en el proceso de cálculo de las medidas de análisis de redes sociales) de la red son: políticos por un lado como: Donald Tusk y Theresa May y periodistas de opinión reputada como: La cadena BBC y Louise Mensch, que en el momento de producirse el debate en el parlamento servían de puentes de comunicación con el exterior, consiguiendo más centralidad, más protagonismo desde fuera que la que podían ofrecer los propios políticos, aun teniendo mayor grado de conexiones.

BIBLIOGRAFÍA

1 – *Práctica de redes sociales con GEPHI*, José Luis Molina, Universidad Autonómica de Barcelona. Obtenido de la dirección: http://pagines.uab.cat/joseluismolina/sites/pagines.uab.cat.joseluismolina/files/PR%C3%81CTICA%20DE%20REDES%20SOCIALES%20CON%20GEPHIversionEspa%C3%B1ol.pdf

2 – *Artículo de Louise Mensch,* de Wikipedia, obtenido de la dirección: https://en.wikipedia.org/wiki/Louise_Mensch

3 – *Artículo de la Red libre de escala,* de Wikipedia, obtenido de la dirección: https://es.wikipedia.org/wiki/Red_libre_de_escala

4 – *Tutorial: Network Analysis of a Twitter hashtag using Gephi and NodeXL,* Clara Guibourg, Periodista Digital, obtenido de la dirección: https://cguibourg.wordpress.com/2015/05/04/tutorial-network-analysis-of-a-twitter-hashtag-using-gephi-and-nodexl/