PRÁCTICA 1

Sistemas de información web

Universidad De Granada,
Master en Ingeniería Informática 2016/2017

AUTOR

JUAN PABLO GONZÁLEZ CASADO
PABLO12614@CORREO.UGR.ES / *PABLO12614@GMAIL.COM*

DESCRIPCIÓN

ESTUDIO DE RED SOCIAL SOBRE EL INCIDENTE DE SAN PETERSBURGO 04 DE ABRIL DE 2017

CONTENIDO

INTRODUCCIÓN

A día de hoy las redes sociales pueden ser una gran fuente de información a la hora de querer analizar comunidades de usuarios, tendencias o cualquier otro tipo de aspecto sobre la sociedad.

Hemos visto cómo se puede intentar predecir la viralidad de mensajes, analizar protestas sociales hasta el punto de crear partidos, determinar comunidades de usuarios como medio de promoción o determinar tendencias culturales.
Sobre este último ejemplo es sobre lo que vamos a plantear el problema y tratar de analizar los niveles de generar información relacionada con este tipo de casos.

Es estudio de redes sociales es un ámbito complejo en el que se tiene que jugar con la imaginación y en algunos casos puede resultar engañoso por el tipo de datos generados, la falta de filtrado o por un mal análisis del problema.

INVESTIGACIÓN

La investigación se va a centrar en un tema que a día de hoy parece tomar fuerza, tratando de analizar si se trata de un "buen" uso de las redes sociales de las empresas o de una tendencia entre los usuarios de las redes sociales.

El problema a tratar es el incidente de la explosión de un artefacto en San Petersburgo, el 3 de Abril de 2017.

Actualmente las sociedades parecen estar ante una amenaza constante de ataques terroristas.
En este caso, se va a aprovechar esta tendencia para determinar quiénes son los principales generadores de información a la hora de informar sobre un atentado, pudiendo analizar el apoyo que ciertos medios le dan a este tipo de noticias y dejando a una libre interpretación conclusiones de intereses.

Para analizar este problema se han realizado búsquedas por la tendencia en Twitter 'San Petersburgo', descargando twitts por medio del complemento NodeXL para Excel y tratando de filtrar los datos con el fin de obtener una conclusión.

La estructura de datos se ha dividido en nodos y aristas.

La tabla de datos de nodos tiene la siguiente estructura:

Id	**Etiqueta**	**Seguidos**	**Seguidores**	**Favoritos**	**Tweets**	**Zona horaria**
Texto	*Texto*	*Número*	*Número*	*Número*	*Número*	*Texto*
Pablo126	Pablo126	1500	1499	23	432	Madrid

Podemos ver como en la tabla de nodos se especifican los usuarios que forman la red social y algunas de sus características.

En el caso de las aristas su tabla es algo así:

Origen	**Destino**	**Mensaje**
Texto	*Texto*	*Texto*
Pablo126	Aarbeloa17	RT: ¡Hala Madrid!

Aquí podemos identificar las interacciones entre los usuarios de la red, pudiendo ver como un usuario trata de destacar algo que otro ha escrito, dando a entender que este usuario mantiene una postura parecida al que ha generado el mensaje.

El volumen de datos ha sido de unos 10000 twitts, eliminando repetidos y extrayendo nodos y aristas se ha quedado reducido a 5368 y 7160 respectivamente, aunque posteriormente se volverá a filtrar para omitir información poco útil.

CALCULO DE MEDIDAS DE ANÁLISIS

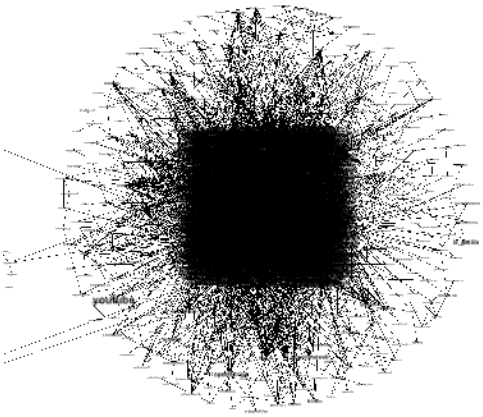
Vamos a calculas las medidas de análisis de la red con toda la información obtenida, aunque posteriormente se realizará un filtrado para hacerla mas manejable.

- **(N)** Número de nodos: 5956
- **(L)** Número de enlaces: 7159
- **(D)** Densidad: 0
- **(k)** Grado medio: 1,202
- **(d_max)** Diámetro: 5
- **(d)** Distancia media: 1,202
- **(d_aleatoria)** Distancia media para la red aleatoria equivalente: 47,24
- **(C)** Media de coeficiente de clustering: 0,024
- **(C_aleatoria)** Media coef. Clustering para la red aleatoria equiv: 2,018

Número de componentes conexas: 1149

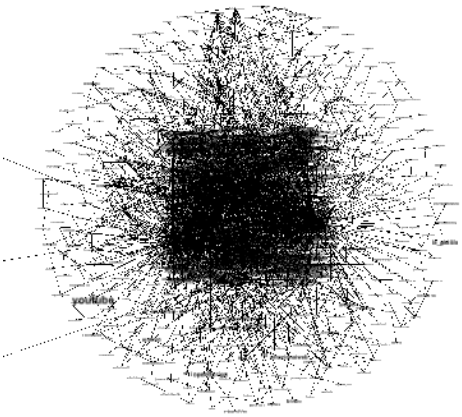
Ahora vamos a aplicar una componente gigante, aunque también aplicarmos un K-core para reducir el número de nodos y aristas “débiles”.

A primera vista, aplicando la componente gigante y visualizando con Fruchterman Reingold, podemos ver la complejidad de la red y la falta de uniformidad:



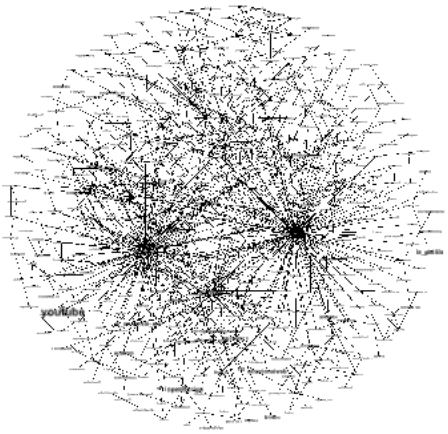
Compontente gigante + K-core(1)

Aumentando el nivel de K-core a 2 obtenemos el siguiente grafo:



Componente gigante + K-core(2)

Vemos como se debilita la zona central aunque siguen sobrando muchos nodos poco útiles.



Componente gigante + K-core(3)

Ahora sí que parece que hemos obtenido una red representativa, limpia de enlaces débiles. Seguramente este sea el filtrado con el que vamos a continuar trabajando, pero para asegurar vamos a probar con un nivel más.

Componente gigante + K-core(4)

Esta red seguramente nos pueda resultar muy útil aunque el filtrado de datos parece ser alto y estamos empezando a perder información, quedando descartada esta combinación para nuestro objetivo.

Una vez seleccionados los filtros que vamos a realizar, los aplicamos y obtenemos los siguientes valores:

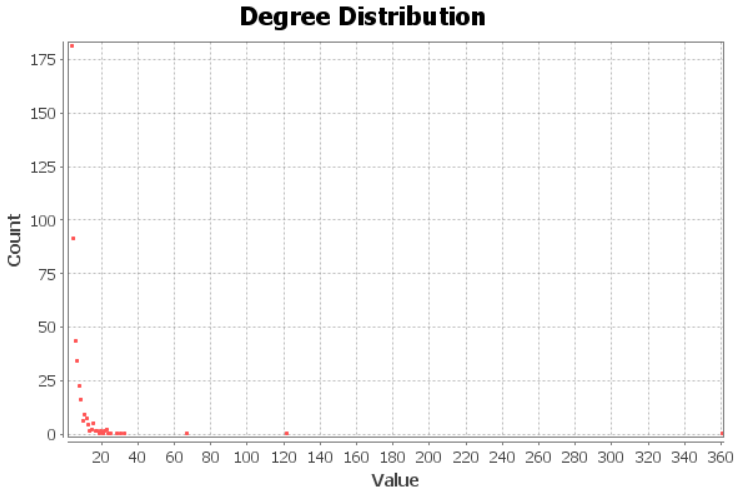
- **(N)** Número de nodos: 455 (7,64%)
- **(L)** Número de enlaces: 522 (21,26%)
- **(D)** Densidad: 0,007
- **(k)** Grado medio: 3,345
- **(d_max)** Diámetro: 5
- **(d)** Distancia media: 3,345
- **(d_aleatoria)** Distancia media para la red aleatoria equivalente: 5,068
- **(C)** Media de coeficiente de clustering: 0,133
- **(C_aleatoria)** Media coef. Clustering para la red aleatoria equiv: 0.0073

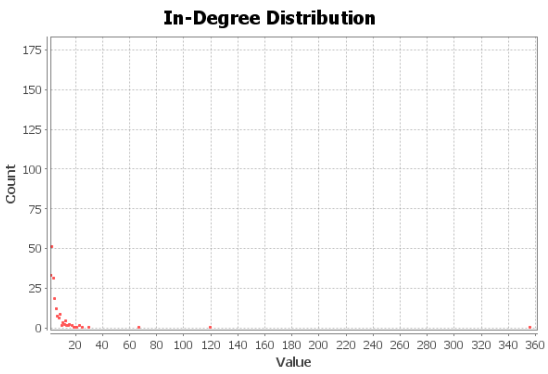
Número de componentes conexas: 4

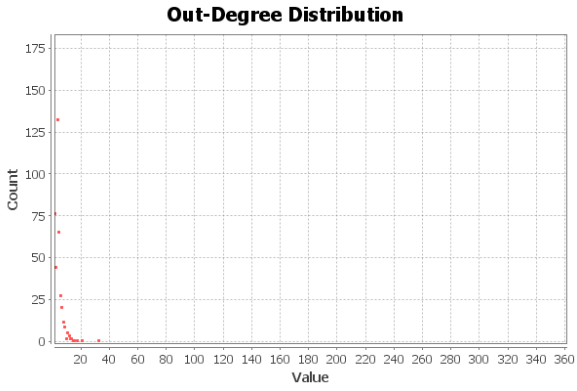
DETERMINACIÓN DE LAS PROPIEDADES DE LA RED

DISTRIBUCIÓN DE GRADOS

A continuación vamos a representar las distribuciones de grados totales, de entrada y de salida.



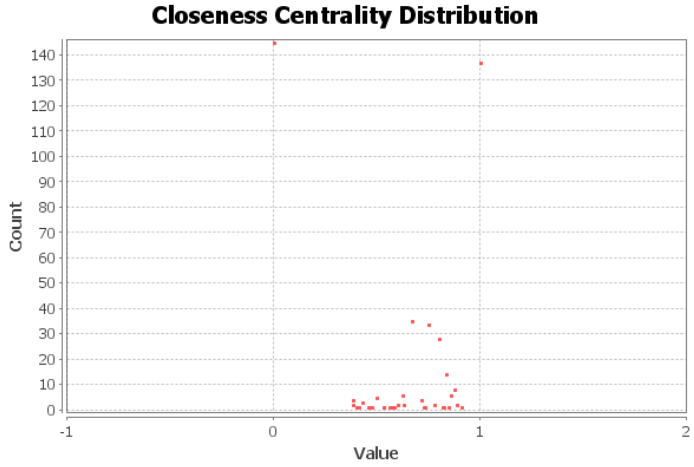


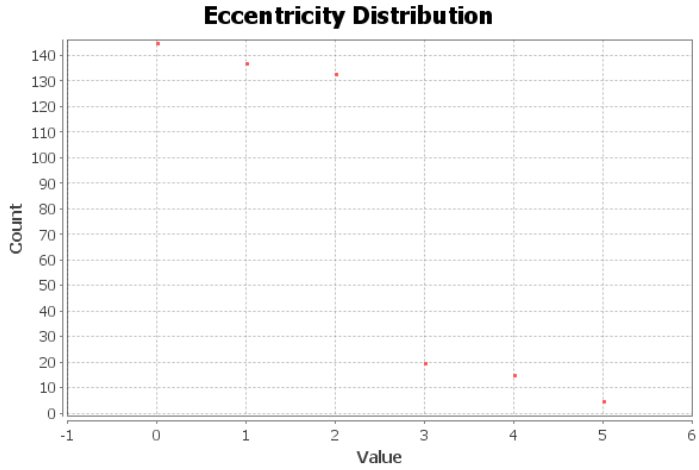


Podemos ver que la distribución es la clásica a la hora de trabajar con una red social.
Existen muchos nodos con pocos enlaces y muy pocos enlaces con muchos nodos. Algo bastante común y con el que se pueden identificar los principales nodos de la red.

DISTRIBUCIÓN DE DISTANCIAS

Vamos a comprobar las distancias máximas de cada nodo al más lejano dentro de la red.





En el primer gráfico podemos ver como la mayoría de los nodos se encuentran entorno al centro, representando así al sector que se dedica a retwittear, teniendo conexión con los centros de la red (1). Por otro lado vemos que hay un gran número de nodos centrales (entre 130 y 140), seguramente se trate de empresas dedicadas a la comunicación.

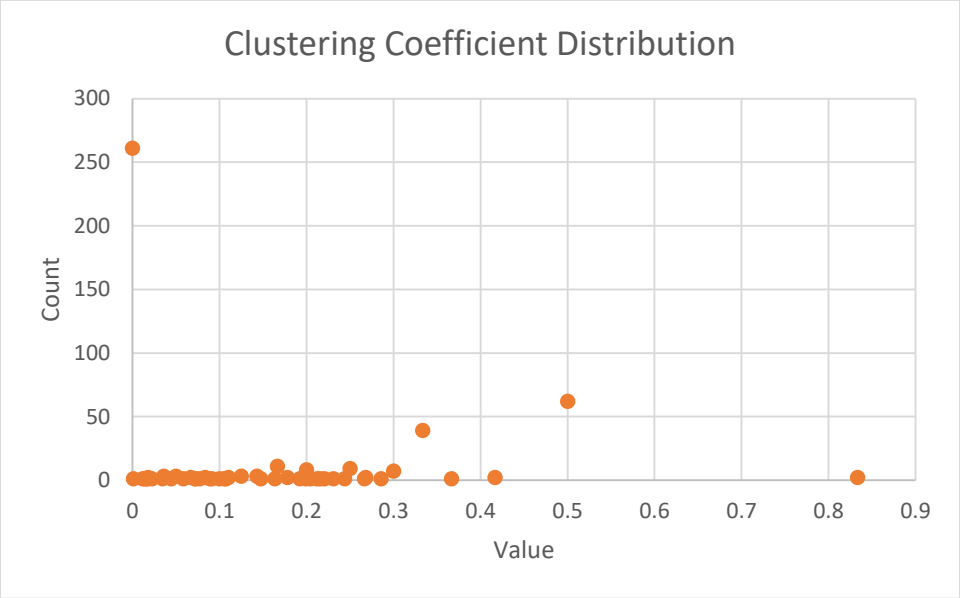
En el segundo grafico se puede ver mejor la agrupación, pudiendo ver que hay nodos inconexos y otros que tienen una mayor conexión aunque se número es más reducido.

Los que tienen un valor 0 son tweets que nombran el acontecimiento pero no mantienen relación con los centros de la red o los generadores de información masiva.

DISTRIBUCIÓN DE COEFICIENTE DE CLUSTERING MEDIO

Obteniendo la distribución de los coeficientes de clustering, podemos ver en la gráfica la acumulación de nodos con un coeficiente de clustering mayores que 0 y menores de 0,3.

Una vez más podemos ver la agrupación de nodos dispersos con una cantidad igual a 0 y los nodos muy conectados con una cantidad cercana al 1.



Tras este análisis podemos determinar que por la distribución de grados y la ley de la potencia, se trata de una red libre de escala.

La diferencia entre la distancia de la red y la distancia aleatoria equivalente (3,345 y 5,068 respectivamente), también hace pensar que estamos trabajando con una red de mundo pequeño.

CALCULO DE LOS VALORES DE LAS MEDIDAS DE ANÁLISIS DE REDES SOCIALES

Si nos preguntamos quienes son los principales generadores de este tipo de información, seguramente se llegue a la conclusión de que se trata de los periódicos digitales.
El problema lo vamos a centrar en la importancia que le dan ciertos medios a un tipo de noticias, por lo que para ello vamos a calcular las siguientes medidas:

Usuario	Grado
actualidadrt	360
el_pais	121
cnnee	66
letradoarmando1	32
a3noticias	30
foro_tv	28
elmundoes	24
sputnikmundo	23
conelmazo_dando	22
ntelevisa_com	22

Usuario	Intermediación
actualidadrt	160
foro_tv	60
betzabezumaya	45
ntelevisa_com	40
pablogrodrgz	25
isabelcatolica2	16
superalitita	8
antonnic1	8
momentses	7
aneshali	7

Usuario	Cercania
actualidadrt	1
supernews2017	1
isabelcatolica2	1
aliciadiego69	1
cristanjc	1
diegokpri050	1
rayala266	1
connie_gdl	1
alvarezp1973	1
momentses	1

Usuario	Vector propio
actualidadrt	1
el_pais	0.22444233
mauricioampuero	0.15904108
cnnee	0.08561992
foro_tv	0.08180957
conelmazo_dando	0.06324931
supernews2017	0.06052104
temuco	0.05606968
a3noticias	0.04568348
isabelcatolica2	0.03815963

Mi investigación se basa en ver los medios de comunicación que le dan más importancia a este tema.

El caso de la intermediación no es determinante ya que se va a analizar medios de un mismo país y estos estarán divididos en comunidades de usuarios que les gusta más un medio u otro.

La cercanía no es nada determinante. Existen muchos nodos con cercanía 1, así que cualquier usuario de estos puede aparecer entre los 10 primeros.

En cuanto al vector propio lo que se trata es de estudiar el grado pero teniendo en cuenta si las conexiones de ese nodo son importantes o no. Por lo tanto me parece relevante para el estudio.

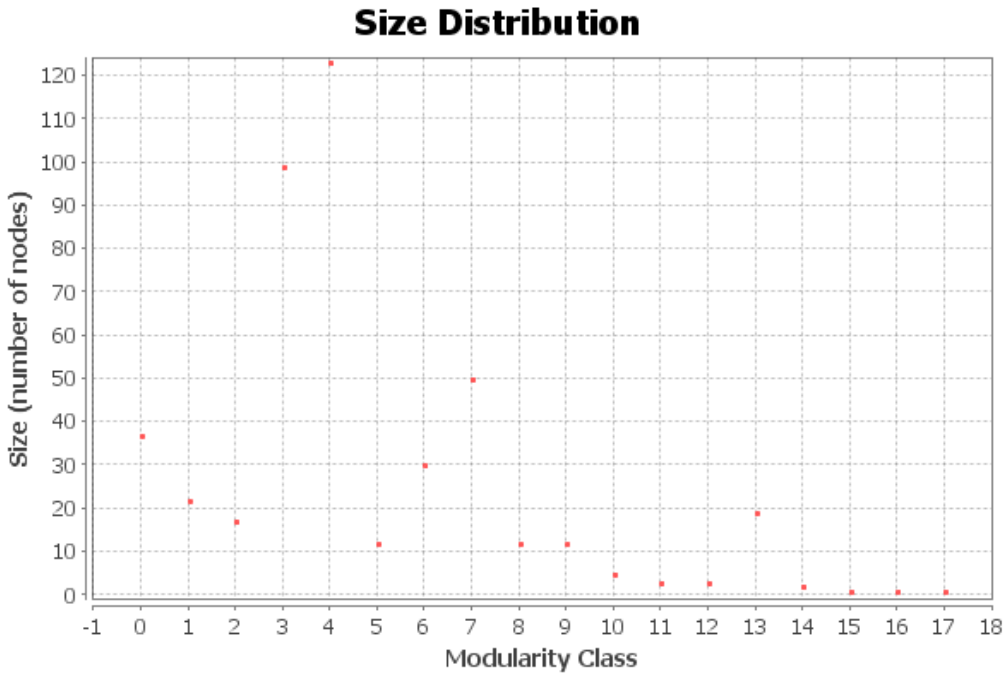
Tanto grado como vector propio me parecían medidas relevantes para el problema a tratar. La manera de decidir una de las dos ha sido fijándome en los usuarios de ambas tablas y viendo como en el caso del grado, aparecen usuarios no asociados a empresas de comunicación, así que teniendo en cuenta el fin de la investigación, seleccionamos vector propio.

DESCUBRIMIENTO DE COMUNIDADES EN LA RED

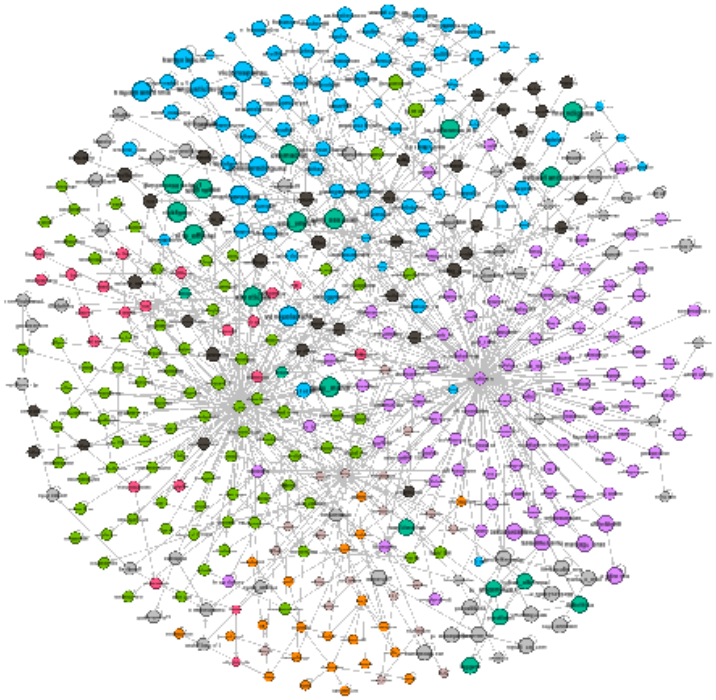
Para descubrir comunidades se ha utilizado el método Louvain.

Primero realizamos unas pruebas calculando modularidad con diferentes resoluciones, obteniendo las resoluciones que dé el valor más alto en modularidad. En mi caso el valor de la resolución esta entre 1 y 1,5 por lo que he elegido 1,3.

Para una resolución de 1,3 se ha obtenido una modularidad de 0,566 y 18 comunidades.



Ahora, observando el grafo y aplicando algunos filtros, se puede obtener bastante información de una manera gráfica y deducir la importancia de muchas comunidades.



Comunidad	Medio/centro
Azul	OkDiario
Morada	RT
Verde	El Pais
Naranja	CNNEE
Verde/azul	La Sexta
Gris	Medios de Mexico y otros paises
Rosa	Antena3

Aislando comunidades se puede llegar a ver que la comunidad de CNNEE está unida con la comunidad de OKDiario y aunque parecen dos comunidades totalmente distintas, parece que una es generadora de otra.

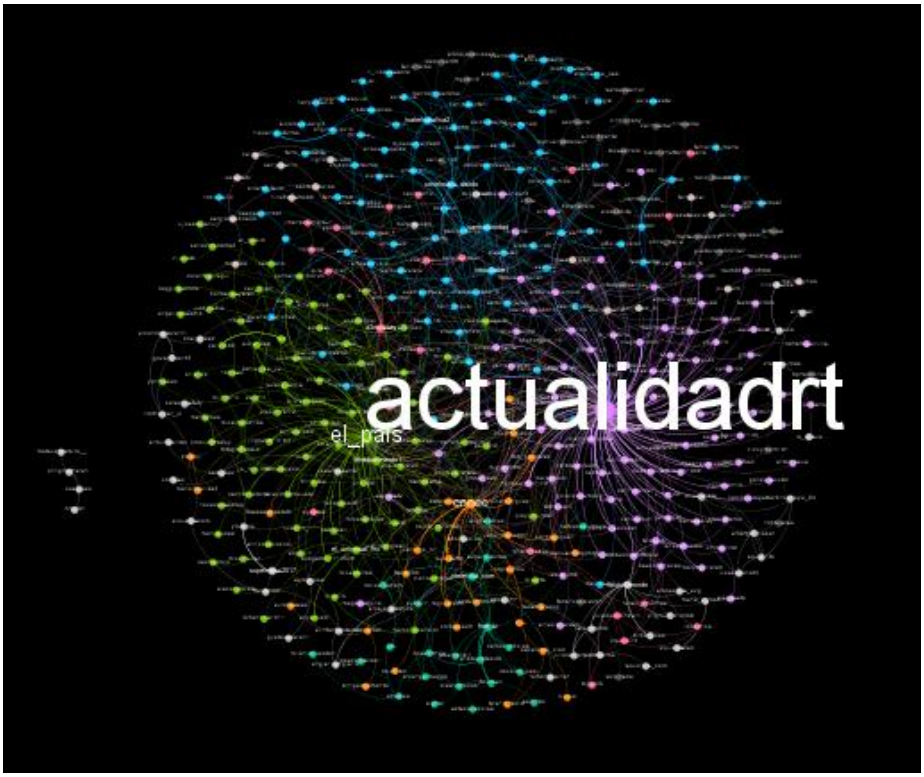
La morada, RT, es un medio de comunicación de Rusia por lo que es la más grande y la más generadora de información. Está prácticamente aislada.

La rosa por ejemplo es Antena3, aunque está demasiado dispersa.

Aunque hemos encontrado 18 comunidades, tras filtrar, se puede ir viendo como hay comunidades poco importantes o que se han ido generando por medio de otras comunidades más pequeñas. También ocurren casos de conversaciones aisladas.

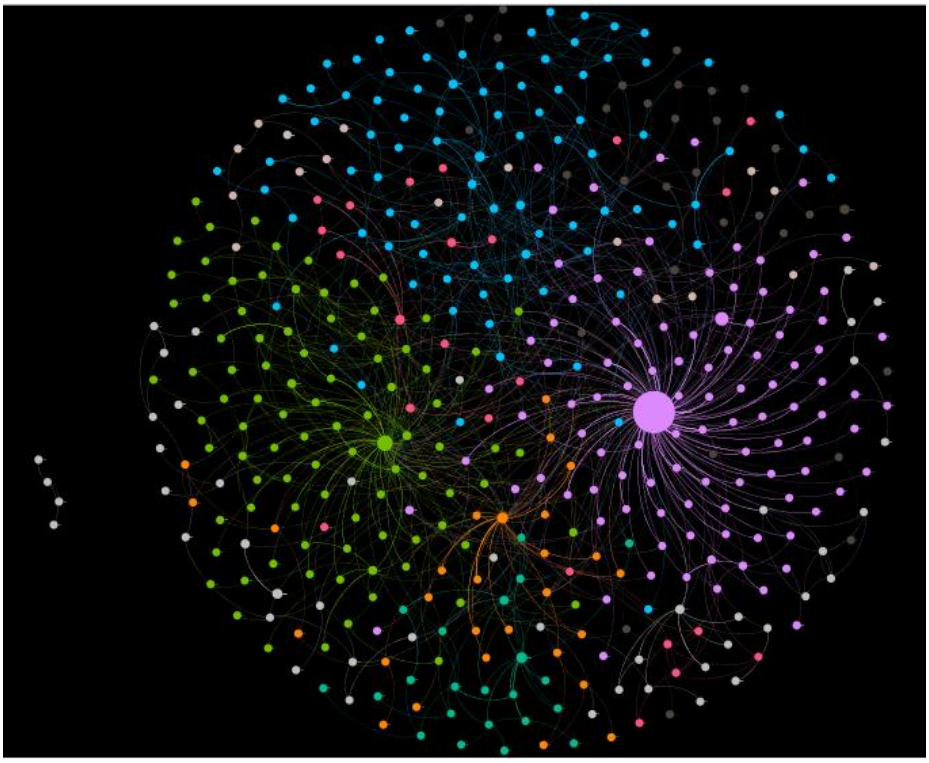
VISUALIZACIÓN DE LA RED SOCIAL

En una primera visualización vamos a ver el grado, un valor importante a la hora obtener los generadores de información.



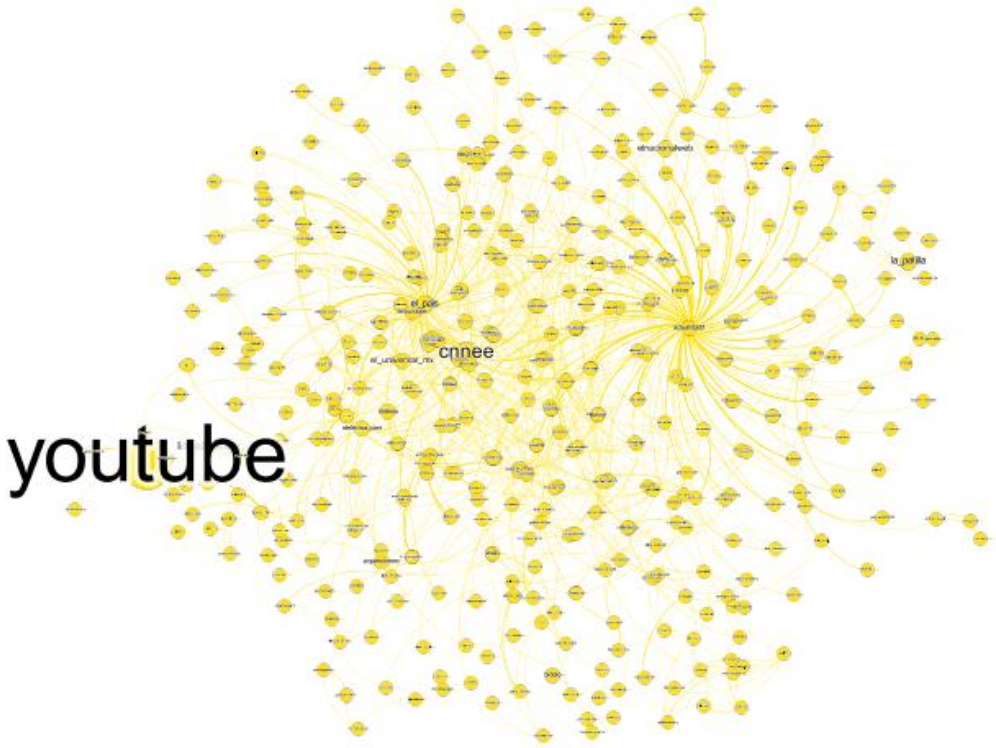
Vemos como actualidadRT es la más grande. También se ve El Pais y destaca en tercer lugar CNNEE.

Aquí tenemos la misma visualización pero esta vez con el vector propio de centralidad.



Prácticamente es igual que la del grado, pero en este caso hemos eliminado las etiquetas. De esta manera podemos visualizar mejor los enlaces, viendo que los que mas tienen son actualidadRT, ElPais y CNNEE. El resto de comunidades tienen aristas muy dispersas, esto quiere decir que no hay una comunidad tan determinada como en el caso de las tres anteriores.

La siguiente visualización formar parte de un análisis de la importancia de los seguidores a la hora de generar información.



En ella podemos ver como youtube es el perfil con más seguidores y aun así nunca se encuentra entre los más importantes a la hora de tratar este tema.
También vemos como actualidadRT no se necuentra ni entre los 5 con más seguidores y aun así es el más influyente.

De aquí se puede deducir que a la hora de necesitar información, los usuarios saben dónde tienen que acudir y el foco de las conversaciones.

DISCUSIÓN DE LOS RESULTADOS OBTENIDOS

Tras los resultados obtenidos, podemos decir que se han detectado claramente comunidades de usuarios, se ha determinado los focos de información y aunque la pregunta de investigación es bastante interpretativa, se puede deducir que los medios de comunicación que hacen eco en Twitter de este tipo de noticias normalmente son los que apoyan las plataformas digitales.

Por ejemplo, LaSexta es una cadena de televisión con bastante interacción en las redes y como hemos visto, se encuentra entre los más importantes. También está el caso de periódicos digitales como ElPais y OKDiario.

Se ha podido visualizar como existen canales de información con una gran cantidad de seguidores pero que a la hora de generar cierta información no son medios valiosos.

El tema de la apuestas de los medios de comunicación por las redes sociales tiene una gran profundidad y aquí solo se han mostrado algunos matices que hacen entrever el poder que puede suponer.

También se ha podido visualizar como unos medios de alimentan de otros como el caso de OKDiario y CNNEE, algo que puede ayudar a la hora de determinar los medios que generan información original y contrastada.

Se ha confirmado que el comportamiento de esta red social sigue el patrón de todas las redes, donde encontramos una gran cantidad de usuarios que generan información pero con poco alcance y luego se encuentran unos cuantos usuarios que generan información y es utilizada por el resto.

Por último se han podido ver apariciones de comunidades aisladas y que tras un aislamiento se han comprobado que pertenecen a otros países. Quizás esto se deba a la forma de obtener la información de Twitter. El rango de descarga de datos es muy determinante, ya que puede estar siendo omitido una región en la que es de noche y la población no genera tanta información.