Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

Máster Profesional en Ingeniería Informática

Curso 2020/2021

PRÁCTICA 4

Gestión de Información en la Web

Breve descripción

Evaluación de Redes en Twitter

Autor

Álvaro de la Flor Bonilla (alvdebon@correo.ugr.es) 15408846-L

RESUMEN

El objetivo de esta práctica es formalizar todos los conocimientos adquiridos en el curso aplicándolos a un caso real de análisis de una red social on-line generada a partir de un medio social. Para ello, el alumno seleccionará un medio social concreto, planteará una pregunta de investigación, obtendrá un conjunto de datos del medio en cuestión, construirá una red social on-line adecuada y la analizará para responder a la pregunta planteada. En principio, nos plantearemos el uso de redes de Twitter, pero el alumno puede considerar cualquier red obtenida de cualquier otro medio social (Facebook, Instagram, Pinterest, Youtube, Flickr, Wikipedia, etc.).

ÍNDICE DEL PROYECTO

ÍNDICE DE ILUSTRACIONES

1 INTRODUCCIÓN

1.1 Temática de investigación

El lunes día 24/05/2021 el seleccionador nacional Luis Enrique hizo pública la lista de los jugadores elegidos para disputar finalmente la Eurocopa, campeonato de futbol de carácter europeo que se disputan 55 países distribuidos en 10 grupos.

De los 26 posibles jugadores elegibles solo fueron convocados 24. En concreto fueron:

- ***Porteros***: Unai Simón, David de Gea y Robert Sánchez.
- ***Defensas***: José Luis Gayá, Jordi Alba, Pau Torres, Aymeric Laporte, Eric García, Diego Llorente, César Azpilicueta y Marcos Llorente.
- ***Centrocampistas***: Sergio Busquets, Rodri Hernández, Pedri, Thiago Alcántara, Koke Resurrección y Fabián Ruiz.
- ***Delanteros***: Dani Olmo, Mikel Oyarzabal, Álvaro Morata, Gerard Moreno, Ferran Torres, Adama Traoré y Pablo Sarabia.

Fruto de la lista anterior, se originó un gran conflicto en Twitter respecto a la elección, con mayor o menor crítica, en función de si el jugador favorito del equipo del usuario estaba en ella. Como ejemplo a destacar, podemos señar el caso del gran centrocampista Sergio Canales Madrazo, actual jugador del Real Betis Balompié, que pesé a su gran temporada y calidad de juego mostrada este año (además de su gran actuación en su último partido representando la selección) ha sido descartado para disputar la competición europea.

La comunidad futbolera y aficionada a la selección española de futbol es internacional, sin embargo, la crítica a la no convocatoria de Sergio Canales es más local (a pesar de la internacionalidad de la afición bética). La pregunta que planteo es: ¿la conectividad de la comunidad crítica de la no convocatoria de Sergio Canales es en parte igual de reducida que en la realidad? Por otro lado, también nos gustaría medir el grado nivel de influencia que son capaces de ciertos usuarios ¿Serán béticos o fanes del futbol en general?

1.2 Medio social y conjunto de datos

Para dar respuesta a las preguntas anteriores, y como ya adelantamos, vamos a usar como medio social Twitter. Además, para la extracción de los datos ya que estamos usando *"Gephi"* utilizaremos el plugin *"Twitter Streaming Importer"* pudiéndolo añadir directamente desde la interfaz del programa.

Para hacerlo, hemos tenido que crear una cuenta y crear una aplicación en la sección de desarrolladores de *Twitter*. Después de este paso seremos capaces de hacer uso de la API de Twitter directamente desde la interfaz de *"Gephi"*.

2 CONSTRUCCIÓN DE LA RED

Haciendo uso de las herramientas mencionadas anteriormente y una vez concluidos todos los pasos anteriores hemos añadido los hashtags #EURO2020, #SergioCanales y nombrar a *"sergio canales"*.

La red descargada es del tipo *"Full Twitter Network"*, por tanto, esta red contiene los *"tweets"* de todos los usuarios, los *"hashtags"*, enlaces y los contenidos multimedia. El proceso de recolección de los datos se realizó el día 24 de mayo de 2021, máximo exponente de crítica de la lista de seleccionados configurada.

Es una red dirigida, compuesta por un total de 7520 nodos y 15169 aristas (o enlaces). De estos nodos, como bien lo hemos indicado antes se representan los usuarios, los *"tweets"*, los enlaces, contenido multimedia y los *"hashtag"*.

Sin embargo, la definición de enlace y no la hemos visto hasta ahora, y puede representar las siguientes acciones:

- Un *"tweet"* se conecta con los enlaces, contenido multimedia y los *"hashtag*" que contiene.
- Un *"tweet"* conecta con un usuario que lo menciona o hacer un *"retweet"* de tipo mención.
- Un usuario se conecta con otro al que menciona o hace *"retweet*".
- Un usuario se conecta con el *"tweet"* que el mismo publica, hace *"retweet"* o cita.

Ilustración 1 – Muestra de la red

La distribución de colores que puede ver en la figura de arriba corresponde al siguiente código de colores que puede ver en la imagen:

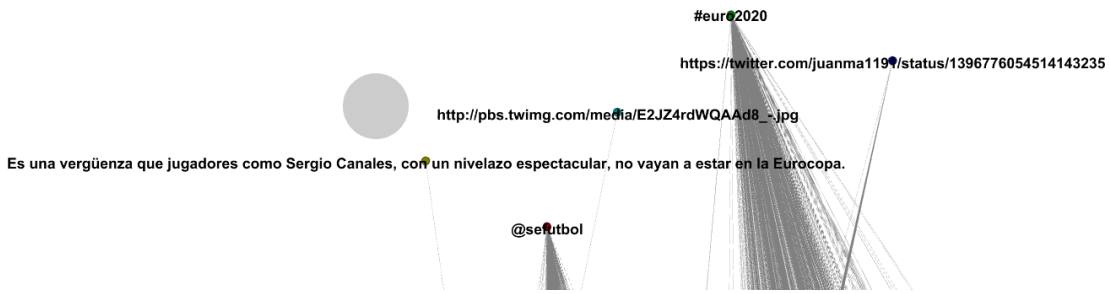

Ilustración 2 – Código de colores

El código de colores es el siguiente:

- El color rojo corresponde con usuarios.
- El verde claro con el contenido del *"tweet".*
- El verde oscuro con los *"hashtag"*.
- El azul con el contenido multimedia.
- El morado con referencias a otro *"tweet"*.

3 ANÁLISIS

3.1 Características estructurales

Como podemos ver en la imagen superior (Figura 1), se señalan claramente tres agrupaciones de nodos. Dos claramente señaladas y muy reconocible y otra algo más dispersa, pero de gran tamaño, aun así.

Todas estas agrupaciones siguen una misma estructura claramente señalada, compuesta de agrupaciones en las que en el centro se encuentran una pareja de nodos correspondientes con un *"tweet"* y el autor de este. Alrededor de esta dupla se encuentran los usuarios que han interactuado con él.

Dándole un primer vistazo a nuestra red podemos presuponer que, como elementos más importantes podemos señalar:

- La mayoría de las interacciones hacen uso del *"hashtag"* #Euro2020

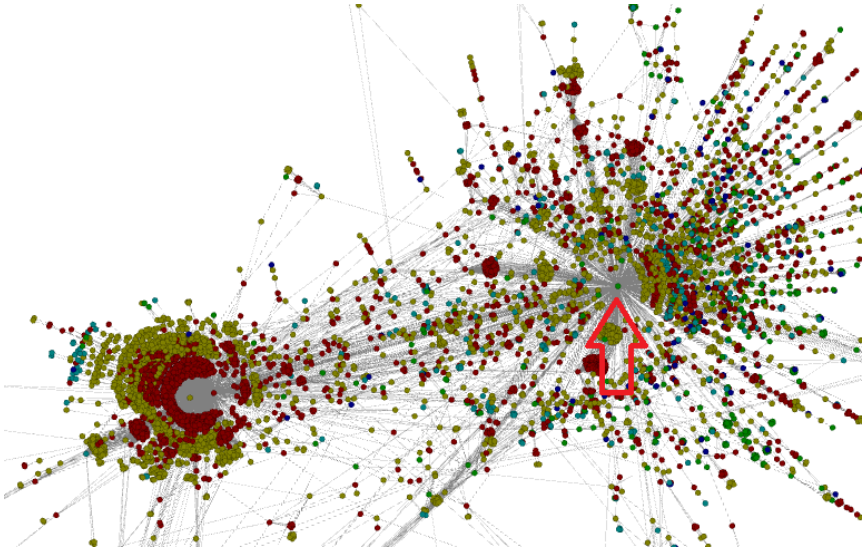

Ilustración 3 – Nodo #EURO2020

- Podemos observar una gran comunidad que se genera a partir del *"tweet"* que realizó la cuenta oficial de la selección (@sefutbol):

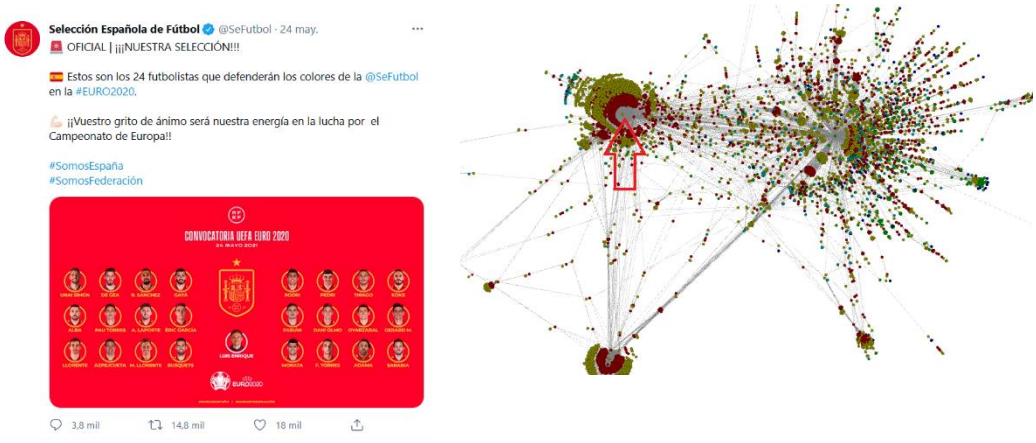

Ilustración 4 – Segunda gran comunidad

- Existe otra gran comunidad basada en la interacción entre un usuario y la respuesta de la compañía *"Aliexpress"*.

Ilustración 5 – Tercera comunidad

Aún así, realizaremos un análisis de las distintas métricas para realizar un estudio aún más profundo ya que por ejemplo aún nos queda por analizar comunidades más pequeñas y aisladas de las que no hemos comentado nada.

3.2 Métricas globales

La interfaz de *"Gephi"* nos proporciona el siguiente calculo de alguna de las medidas de la red, las cuales son:

MEDIDA	VALOR
Número de nodos *N*	7520
Número de enlaces *L*	15169
Densidad del grafo L / L_{MAX}	0.001
Grado medio $<K>$	2.017
Diámetro d_{MAX}	8
Distancia media $<d>$	1,9869549374468602
Distancia media para la red aleatoriamente equivalente $<d_{aleatoria}>$	12.72117758
Coeficiente medio de *clustering* $<C>$	0.257
Coeficiente medio de *clustering* para la red aleatoria equivalente $<C_{aleatoria}>$	0.00026821808
Número de componentes conexas	36
Número de nodos de la componente gigante *(y %)*	7378 (98.11 %)
Número de enlaces de la componente gigante *(y %)*	15039 (99.14 %)

3.3 Propiedades de la red

El siguiente paso que vamos a realizar es, haciendo uso de las métricas de la red, realizar una caracterización de esta con el objetivo de determinar las propiedades más características de la red.

3.3.1 Distribución de grados

Lo primero que vamos a estudiar es la **distribución de grados de entrada**, que en nuestro caso podemos observar que la mayoría de *"tweets"* han recibido muy pocas interacciones y por tanto se concentran en los valores más pequeños de la primera gráfica de la ilustración 6.

Sin embargo, existen algunas excepciones como las que podemos apreciar en los tramos 400-500 y 1100-1200. Estas interacciones son los llamados *"hubs"* que no es otra cosa que los *"tweets"* de los usuarios que si que han recibido más menciones o *"retweets"* ya sea por ser grandes influyentes o éxito de su opinión.

En cuanto a la **distribución de grados de salida**, en este caso se observa que la mayoría de los usuarios no interaccionan con otros, simplemente dan su opinión. Me reafirmo en esta opinión ya que hay muy pocos *"hubs*" como comentamos anteriormente, pero existe una gran concentración en los primeros valores del intervalo.

¿Qué esta ocurriendo entonces? Pues como he comentado anteriormente la mayoría de los usuarios simplemente están dando su opinión propia, sin reaccionar a otras decisiones de los usuarios.

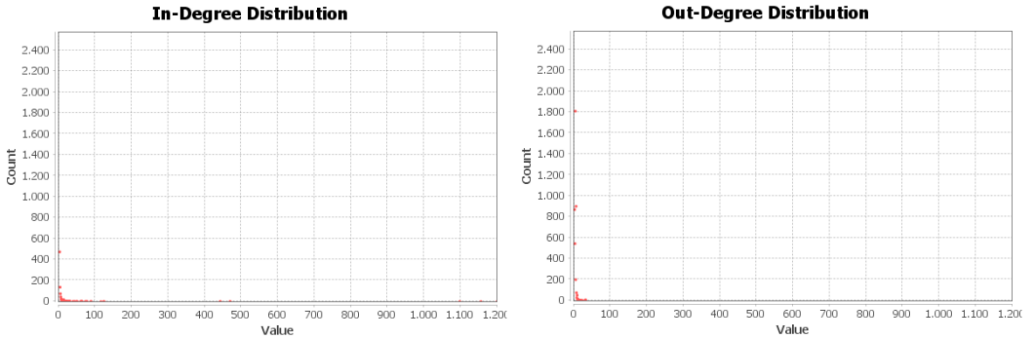

Ilustración 6 – Distribución de grados

Observando ambas gráficas, podemos ver como las distribuciones siguen la ley de la potencia, por lo que podemos afirmar que la red que estamos estudiando es libre de escala. Damos esta afirmación ya que se pueden apreciar un conjunto de nodos que tienen un mayor número de interacciones en comparación con el resto.

Como resultado, hay una gran posibilidad de que existan nodos que tengan un mayor número de conexiones de entrada que el resto.

3.3.2 Análisis de la distancia media

La distancia media real que se ha obtenido mediante *"Gephi"* es 1,99 mientras que la que hemos calculado teóricamente es 12,72. Es evidente que la distancia real ha resultado ser bastante inferior a la calculada.

Es común en redes sociales que sean redes de mundo pequeño, redes en las que, aunque la mayoría de los nodos son vecinos entre sí, puede llegarse de un nodo a otro usando un número relativamente pequeño de saltos.

Para comprobar que una red sea de mundo ultra-pequeño, utilizaremos la siguiente ecuación:

$$< d > = \frac{\ln(N)}{\ln(\ln(< N >))}$$

Como resultado del cálculo anterior obtenemos el valor 4.078. Como este valor es inferior a la distancia media de nuestra red, no se trata de una mundo ultra-pequeño, sino que es de un mundo pequeño.

3.3.3 Coeficiente de clustering medio

En la siguiente imagen puede apreciarse como en la mayoría de los casos los usuarios no reciben muchas interacciones a sus *"tweets"*. En relación con ello podemos estudiar la transitividad de la red.

En las redes sociales, la transitividad lleva a grafos más densos, es decir más cercanos a un grafo completo. Por tanto, esta transitividad puede ser medida fijándonos en como de cerca está nuestro grafo de ser completo, utilizando para ello el coeficiente de clustering de nodos.

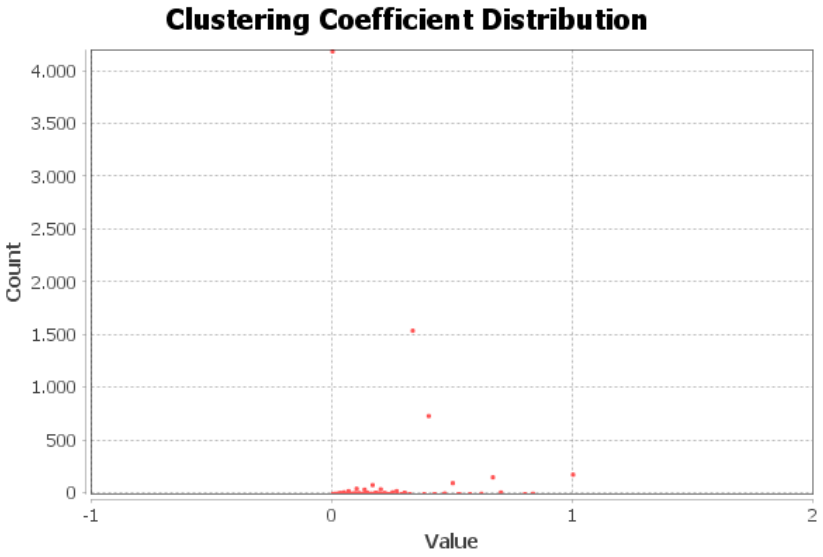

Ilustración 7 – Distribución del coeficiente de clustering medio

Es habitual que en redes sociales se considera alto un valor medio de coeficiente clustering de entorno al 0.6, en ese caso indicaría que la transitividad es alta. Nuestra red tiene un coeficiente de 0.257 lo cual indica una transitividad que comienza a ser bastante baja.

3.4 Medidas de análisis de redes sociales

3.4.1 Intermediación

En este apartado analizaremos los nodos que son nexo de unión y para ello utilizaremos los valores de intermediación de los usuarios de la red.

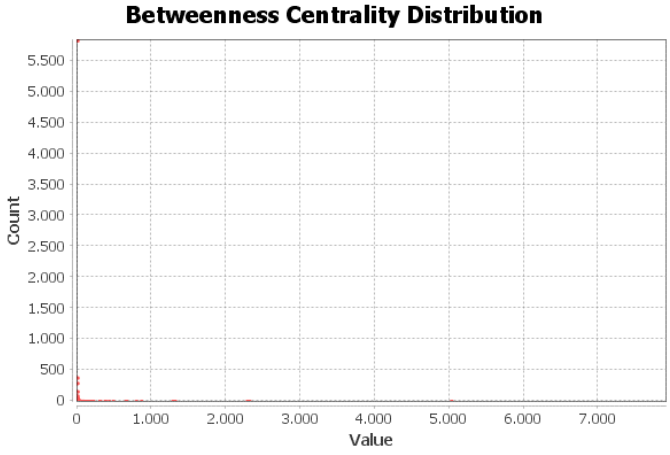

Ilustración 8 – Distribución de la centralidad de intermediación

La mayoría de los resultados se centran en el primer intervalo de la gráfica donde los valores son más bajos, lo cual significa que la mayoría de los usuarios no son fundamentales en los caminos en los que forman parte.

Por otro lado, si que es cierto que existen usuarios que si juegan un papel crucial, ya que conectan a un mayor número de nodos debido a las interacciones que han recibido.

3.4.2 Cercanía

Este bloque analizará cuántos de los usuarios de la red que estamos estudiando se encuentran más cercanos a ella. En otras palabras, lo que queremos conocer es cuáles son los que tienen que dar el menor número de saltos necesarios para alcanzar un *"tweet"*.

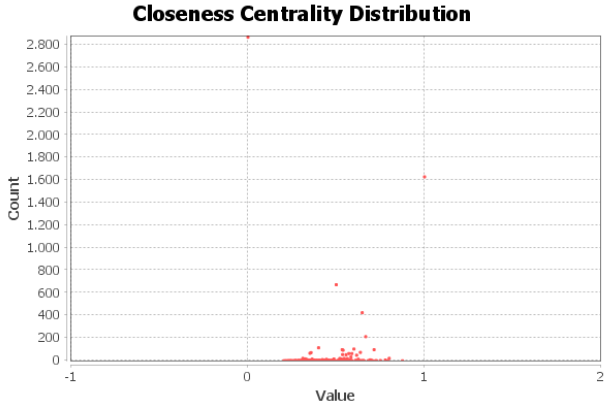

Ilustración 9 – Distribución de la centralidad de cercanía

En la gráfica que podemos ver justo arriba, se observa como hay un denso grupo de usuarios con cercanía 0 siendo por tanto nodos muy alejados de la red. Por otro lado, existe un pequeño grupo de usuarios con cercanía 1, indicando así que se encuentra en la parte central de la red.

Finalmente, existe un tercer grupo que alberga los valores intermedios, siendo nodos céntricos, pero a la vez demasiado alejados.

3.4.3 Vector propio

En esta ocasión estudiaremos como de importante son los usuarios en función de la relevancia de sus vecinos visualizando así los usuarios que son más influyentes por el hecho de estar conectados a los usuarios más famosos.

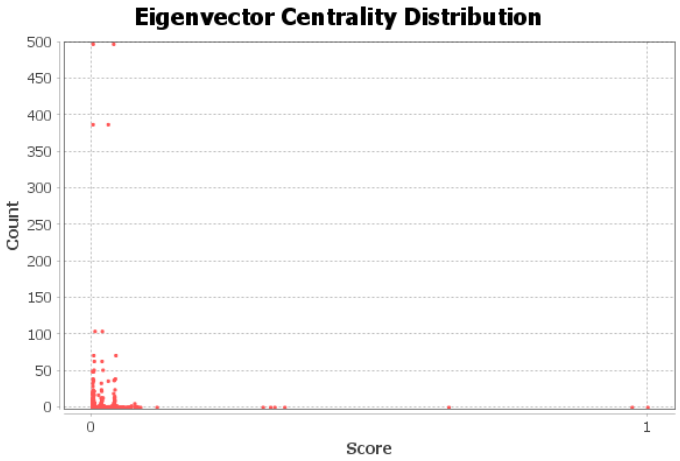

Ilustración 10 – Distribución de la centralidad de vector propio

La inmensa mayoría de los nodos tienen una baja importancia, ya que se encuentran conectados con usuarios que no son muy relevantes. En la sección de visualización observaremos este estudio de forma gráfica.

3.5 Comunidades

Uno de los principales usos de las redes complejas es su uso para la detección de comunidades, entendiendo estas como regiones de la red en la que hay una alta concentración de enlaces, mientras que esa concentración es baja entre ellas.

Utilizando la herramienta que nos ofrece *"Gephi"* por defecto obtenemos:

- Modularidad 0.780 y 116 comunidades con resolución 1.0

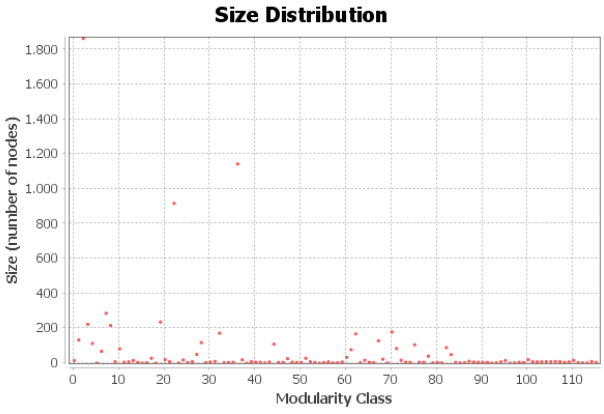

Ilustración 11 – Modularidad sin tratar

- Modularidad 0.017 y 36 comunidades con resolución 38.0

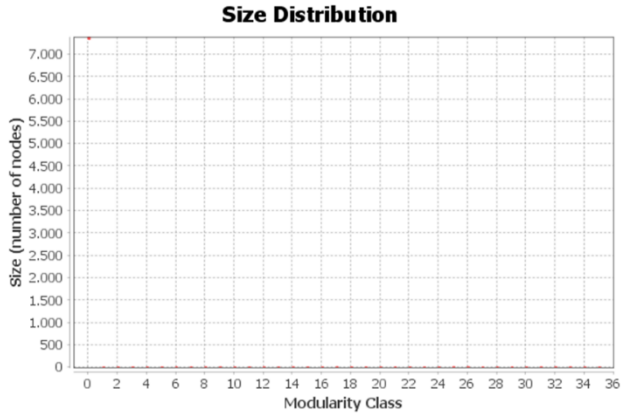

Ilustración 12 – Modularidad tratada

En el apartado de visualización extenderemos aún más este contenido de forma visual.

Por otro lado, si utilizamos el método *"Girvan-Newman"* se obtienen los siguientes resultados:

Girvan-Newman Report

Parameters:

Respect edge type for shortest path betweeness: no
Respect parallel edges for shortest path betweeness: no

Respect edge type for modularity computation: no
Respect parallel edges for modularity computation: no

Processed Graph Data

Nodes: 7520
Edges 15070

Processing time: 232.249 sec.

Communities

Number of communities: 36
Maximum found modularity: 0.017175207

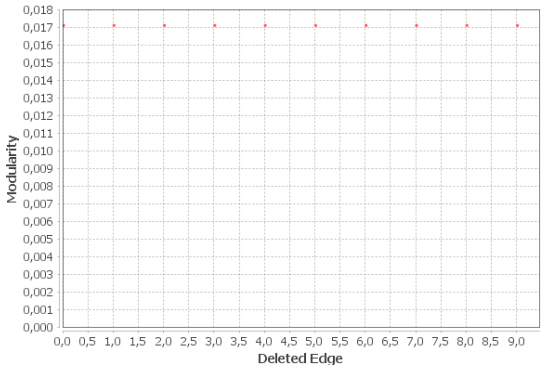

Ilustración 13 – Detección de comunidades usando *"Girvan-Newman"*

Resultados iguales a los dados por la función utilizada por *"Gephi"* de manera estándar en la propiedad modularidad utilizando una resolución 38.0 como valor.

4 VISUALIZACIÓN

4.1 Centralidad intermedia

En esta primera visualización coloreamos los nodos en función al valor de la centralidad intermedia.

Ilustración 14 – Centralidad de intermediación

Cuanto más cercano al violeta sea el rojo del nodo, son los que disponen de menor valor, por lo tanto, apenas intervienen en algunos caminos. Son los usuarios que menos capacidad de difundir tienen.

Por otro lado, los nodos que tienen colores cercanos al azul o vivos una influencia muchísimo mayor. Su valor de centralidad provoca que si se eliminan quedarían desconectados muchísimos caminos

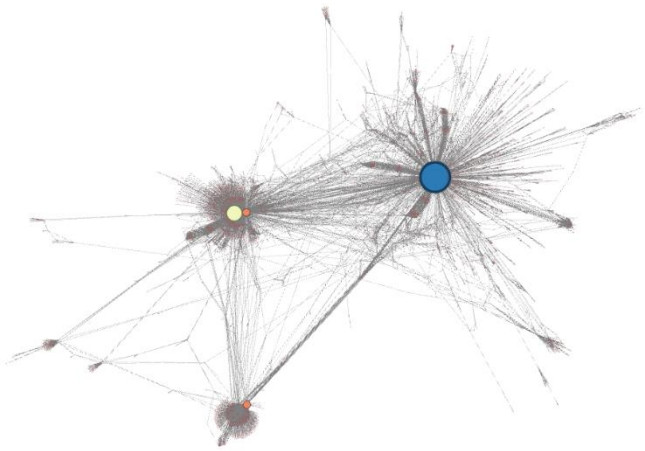

Ilustración 15 – Nodos más importantes

En la ilustración anterior podemos ver como, sobre el resto, quedan tres nodos señalados principalmente. El más grande de todos, el azul corresponde al *"hashtag"*

#EURO2020, el nodo amarillo al "tweet" que señalamos en el inicio de nuestro análisis realizado por la cuenta @sefutbol, que es el nodo naranja y finalmente el otro nodo naranja corresponde al *"tweet" que compara la convocatoria con "Aliexpress"*.

4.2 Centralidad de cercanía

Esta segunda medida, como bien explicamos antes lo que intenta mostrar es que nodos se encuentra más cercanos al centro de nuestra red. Normalmente son aquellos que sean más céntricos ya que cuentan con caminos más cortos en la mayoría de las ocasiones para llegar a cualquier nodo.

Ilustración 16 – Gráfica de centralidad de cercanía

En esta ocasión, los nodos que tienen los colores de tonos azules tienen que recorrer menor distancia que el resto de los nodos para alcanzar el centro de la red. Por otro lado, los nodos con tonos del amarillo al rojizo (azul más cerca, rojo más lejano) sí que tienen que recorrer una distancia mayor, en comparativa de sus vecinos.

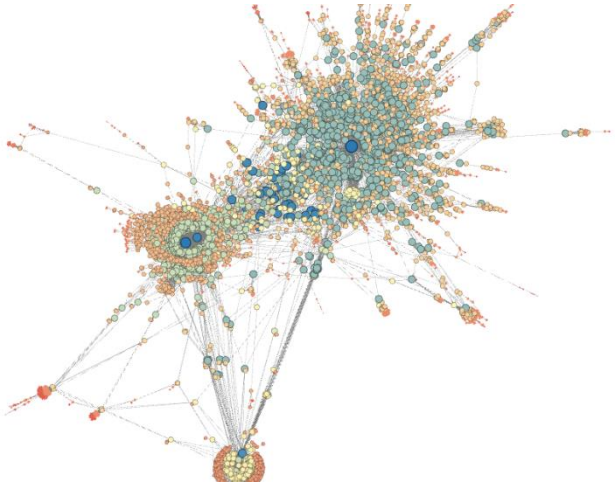

Ilustración 17 – Nodos más influyentes según centralidad

El gráfico de justo arriba muestra los usuarios más influyentes destacando en tamaño usando los mismos códigos de colores.

Enfocándonos un poco más en esto datos:

Label	twitter_type	description	friends_co...	followers...	real_name	Closeness...
La lista de Luis Enrique me ha dejado "descolocado". Que no va Sergio Canales....?	Tweet					1.0
@rbezanilla11	User	Español, C...	2549	1197	ricardo beza...	1.0
Els 3 interiors convocats per Luis Enrique Koke: 1 gol i 2 assistFabian Ruiz:3 gols i 1 assistThiago: 1 gol i 0 assistTotal: 5 gols i 3 assit...	Tweet					1.0
@llado_7	User	Estudiant ...	9	9	Pere Lladó ...	1.0
@bola149	User	Sentidiño.	285	287	bOla149	1.0
Sergio Ramos este como este debe ir a la eurocopa es el capitan y el lider y va aportar seguro... Q no vayan navas, nacho, canales y...	Tweet					1.0
@albertillo_10	User		459	467	Alberto C.T	1.0
Temporadón Don Sergio Canales... https://t.co/nSQLxhFa6s	Tweet					1.0
@Jvor010 @DonMANUELZUAR @jotajordi13 Y pablo sarabia que pinta en esta lista??? Jugó poco y nada en el PSG, enserio esta por e...	Tweet					1.0
Bajo mi punto de vista, y pensando en frio, esto pienso de la convocatoria de Luis Enrique:Jugadores que sobran: Unai Simón, Eric G...	Tweet					1.0
@yerayonne	User	Un chico n...	366	66	Yerayonne YT	1.0
@jesulileon17 Por supuesto, que Jesús Navas, Sergio Canales e Iago Aspas no hayan sido convocados ha sido una injusticia gorda gord...	Tweet					1.0
Jesús Navas, Sergio Canales, Nacho Fernández y Iago Aspas Luis enrique tu padre es Amunike	Tweet					1.0
@jesusgamez30520	User	Jesus_Ga...	211	76		1.0
@MundoMaldini El problema es el nivel del entrenador que es cortita por muchos que nos vendanJugadores como Sergio Ramos-Can...	Tweet					1.0
@albertoowono 📌 PORTERO: Edgar Badia📌 DEFENSA: Kounde📌 MEDIO: Sergio Canales📌 DELANTERO: Benzema 📌 MVP: Marcos Llo...	Tweet					1.0
@chicadelpiano	User		210	7	La chica del...	1.0
@akalj_	User		191	32	Aka Lj	1.0
@josemcteskone	User	chemack ...	1168	303	Chemack	1.0
@MTxabi No es por eso, es porque Luis Enrique no ha convocado a jugadores que han hecho temporadas muy buenas y se lo merecí...	Tweet					1.0
@Gerardmb99 Mis 26 jugadoresGk:De Gea/unai/pachecoDFC: Sergio Ramos/pau torres/iñigo martinez y nachoLTI: Jordi y gayaLTD: ...	Tweet					1.0
@RadioFCB Yo no rabio por el Madrid...rabio porque jugadores como Iago Aspas,Sergio Asenjo,Jesús Navas,Nacho Fernández,Mario ...	Tweet					1.0
@FCamposoficial Sergio Canales pra mim tbm merecia uma chance na lista final da Euro, fez boa temporada no Bétis.	Tweet					1.0
@sergiopatron10	User	Que Dios ...	317	211	Sergio Patr...	1.0
La lista de Luis Enrique es indigna, como no llevas a Sergio Canales	Tweet					1.0
@blandonrmcf	User	Me suspe...	308	79	Juan	1.0
@aleaguilargalan	User	No tenem...	239	260	alejandro a...	1.0
Es una pena lo de Sergio Ramos, porque él solo es mejor que todos los centrales que hemos convocado, pero apenas ha competido ...	Tweet					1.0
@alexborre07	User	Que le de...	1081	323	Alex B.	1.0
@AbuAbuGamer no, ramos yo no lo hubiese llevado. nacho con los ojos cerrado. sergio canales, aspa, hermoso, pero son desiciones...	Tweet					1.0
¡Qué lujos se da España y Luis Enrique! mira que no llevar a la Eurocopa a Sergio Ramos, Aspas y Canales ya dice mucho, pero borr...	Tweet					1.0
@RafaelEscrig @MisterFantasyES Me faltan Alex remiro, Pacheco, Sergio Herrera, Jesús Navas, Nacho, Gabriel Paulista o Mario Herm...	Tweet					1.0
@Papa__Gueye Sergio Canales et Jesús Navas non plus	Tweet					1.0
@MrKinoccio A mi me parece necesaria cierta regeneración..pero me faltan varios elementos en la lista. Me faltan: NAVAS, NACHO, I...	Tweet					0.8
@joseeimp	User	Comuni...	393	110	Jose Ignacio	0.8

Ilustración 18 – Uso de parámetro *"Closeness Centrality"* en laboratorio de datos

En esta ocasión podemos observar como muchos de los *"tweets"* se refieren entre otros, a la no convocatoria de Sergio Canales. Sin embargo, estos nodos son los que se encuentran visualmente en el exterior de la red, rompiendo con el orden que realiza *"ForceAtlas2"*.

4.2.1 Vector propio

En este caso el usuario más influyente ha sido @sefutbol, como es evidente con la declaración de Luis Enrique de la convocatoria oficial. En la imagen inferior azul representa la información más relevante y rojo la que menos.

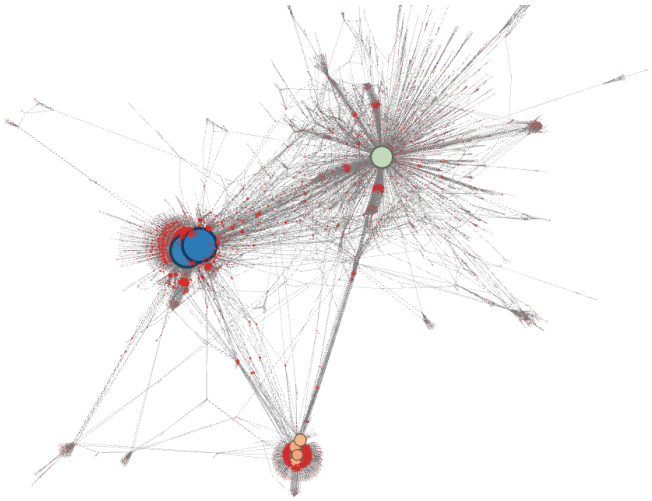

Ilustración 19 – Autores más influyentes

4.2.2 Comunidades y subredes

En la siguiente imagen se muestra el conjunto de comunidades que es capaz de detectar *"Gephi"* de forma automática. Lo más curioso que he descubierto analizando las comunidades, es que los comentarios relacionados a *"Sergio Canales"* representa solo la comunidad de abajo a la izquierda, prácticamente nada respecto al resto.

Ilustración 20 – Comunidades de usuarios

5 CONCLUSIONES

Una vez hemos concluido el análisis de la red, hemos podido identificar a los usuarios más relevantes, que claramente han sido @sefutbol, @juanma1191 y @aliexpress.

Ya, en menor medida, los usuarios más influyentes han sido @albertotegaes1 en base a quien se ha originado toda la comunidad crítica a la no convocatoria de Sergio Canales.

También he podido ver que esta red cumple con las propiedades más comunes vistas en clase.

Por otro lado, he descubierto la escasa participación respecto a la red general que ha tenido la crítica de la no convocatoria de Sergio, ha sido más comentada la no llamada a Iago Aspas.

Uno de los nodos más influyentes ha sido el *"hashtag"* #EURO2020 que ha sido utilizado por multitud de usuarios cada vez que realizaban su interacción.