Máster Universitario en Ingeniería Informática

GESTIÓN DE INFORMACIÓN EN LA WEB

Práctica 4:
Caso Práctico de Análisis y Evaluación de Redes en Twitter

UNIVERSIDAD DE GRANADA

Carlos Morales Aguilera
75925767-F
carlos7ma@correo.ugr.es

Curso Académico 2020-2021

Índice

1. Selección de un medio social, definición de una pregunta de investigación y obtención de un conjunto de datos asociado

Para la realización de esta práctica es necesario realizar un análisis de una red social existente en una red social. Para ello, se ha escogido obtener la red de **Twitter**, utilizando la API de la misma y la herramienta **Gephi** tanto para realizar las tareas de scrapping como de análisis de la información de la red.

Uno de los temas más actuales y con mayor influencia en **Twitter** es el estado actual de una de las competiciones deportivas más importantes a nivel nacional, como es la liga española de fútbol. Actualmente existen 4 equipos candidatos al título, y a falta de unas pocas jornadas son muchos los comentarios que se pueden ver de diferentes aficiones en dicha red social.

Con el objetivo de observar que aficiones están más cohesionadas se proponen dos preguntas de cara a la realización de esta práctica:

- **¿Qué club de los líderes interactúa más con su afición?**
- **¿Existe una diferencia mediática en redes de los clubes respecto a los medios?**

Para poder obtener dicha información, se han considerado los siguientes términos, que si bien son pocos, darán lugar a un gran conjunto de información: *sevilla fc*, *sevilla*, *la liga*, *laliga*, *barsa*, *barcelona*, *madrid*, *real*, *real madrid*, *atlético de madrid* y *atleti*.

Por otro lado, se ha obtenido información adicional de las siguientes cuentas de usuario principales: **fcbarcelona_es**, **realmadrid**, **laliga**, **atleti** y **sevillafc**.

2. Construcción de la red social on-line a analizar y visualizar.

Para la obtención de la red, he seguido los siguientes pasos:

1. Crear una cuenta como desarrollador en Twitter, para poder obtener los tokens de acceso a la API.
2. Utilizar el plugin de Gephi **Twitter Streaming Importer**.

Una vez introducida la información deseada, se ha conectado el plugin y se ha obtenido la información (la cual manualmente ha sido finalizada ya que la cantidad existente de interacciones es exageradamente grande).

Por lo tanto cabe destacar dos aclaraciones:

- **Nodo**: Usuarios de Twitter.
- **Aristas**: Interacciones, tales como menciones, respuestas o retweets. Estas son entre usuarios, por lo que el grafo es **bidireccional**. Por otro lado tienen un **peso** en base a la cantidad de interacciones entre usuarios.

A continuación se muestra la red inicial que se obtiene, la cual no posee ni filtros ni algoritmos de visualización, por lo que es normal que se vea como un cuadrado lleno de puntos:

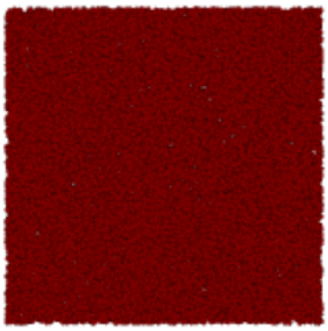
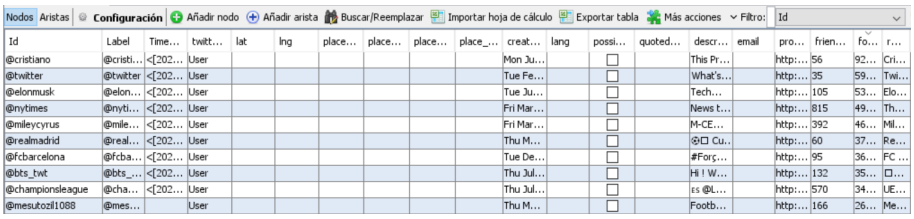


Imagen 1: Visualización inicial de la red

En cuanto a los nodos, se visualiza según número de seguidores, y como es evidente **Cristiano** es una de las cuentas con más seguidores en la actualidad, seguida de grandes cuentas como la propia de **Twitter** o algunos clubes:

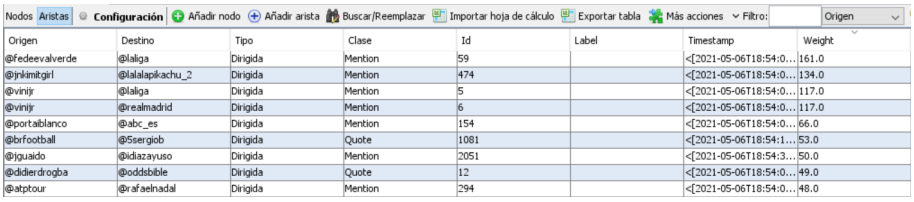
Por último, cabe destacar que las aristas tienen un **peso**, el cual indica la asiduidad que interaccionan esas dos cuentas, en este caso es lógico que un jugador del **Real Madrid**, que juega en **La Liga** como es **Fede Valverde**, sea uno de los que más interactúa con la cuenta oficial de **La Liga**:

Analizando la posibilidad de utilizar diferentes filtros, he buscado información sobre los mismos y estudiado un poco más el contexto de la red, decidiendo finalmente aplicar:



Id	Label	Time…	twitt…	lat	lng	place…	place…	place…	place_…	creat…	lang	possi…	quoted…	descr…	email	pro…	frien…	fo…	r…
@cristiano	@cristi…	<[202…	User							Mon Ju…		☐		This Pr…		http:…	56	92…	Cri…
@twitter	@twitter	<[202…	User							Tue Fe…		☐		What's…		http:…	35	59…	Twi…
@elonmusk	@elon…	<[202…	User							Tue Ju…		☐		Tech…		http:…	105	53…	Elo…
@nytimes	@nyti…	<[202…	User							Fri Mar…		☐		News t…		http:…	815	49…	Th…
@mileycyrus	@mile…	<[202…	User							Fri Mar…		☐		M-CE…		http:…	392	46…	Mil…
@realmadrid	@real…	<[202…	User							Thu M…		☐		⚽☐ Cu…		http:…	60	37…	Re…
@fcbarcelona	@fcba…	<[202…	User							Tue De…		☐		#Forç…		http:…	95	36…	FC …
@bts_twt	@bts_…	<[202…	User							Thu Jul…		☐		Hi ! W…		http:…	132	35…	☐…
@championsleague	@cha…	<[202…	User							Thu Jul…		☐		ES @L…		http:…	570	34…	UE…
@mesutozil1088	@mes…		User							Thu M…		☐		Footb…		http:…	166	26…	Me…

Imagen 2: Ejemplo nodos de la red



Origen	Destino	Tipo	Clase	Id	Label	Timestamp	Weight
@fedeevalverde	@laliga	Dirigida	Mention	59		<[2021-05-06T18:54:0…	161.0
@jnkimitgirl	@lalalapikachu_2	Dirigida	Mention	474		<[2021-05-06T18:54:0…	134.0
@vinijr	@laliga	Dirigida	Mention	5		<[2021-05-06T18:54:0…	117.0
@vinijr	@realmadrid	Dirigida	Mention	6		<[2021-05-06T18:54:0…	117.0
@portaiblanco	@abc_es	Dirigida	Mention	154		<[2021-05-06T18:54:0…	66.0
@brfootball	@Ssergiob	Dirigida	Quote	1081		<[2021-05-06T18:54:1…	53.0
@jguaido	@didiazayuso	Dirigida	Mention	2051		<[2021-05-06T18:54:3…	50.0
@didierdrogba	@oddsbible	Dirigida	Quote	12		<[2021-05-06T18:54:0…	49.0
@atptour	@rafaelnadal	Dirigida	Mention	294		<[2021-05-06T18:54:0…	48.0

Imagen 3: Ejemplo aristas de la red

1. **Componente gigante**: Nos quedaremos con la componente con mayor número de nodos conectados. Nos quedamos con un 26 % de los nodos, puede parecer mucho pero considerando la cantidad de gente que habla sobre el tema de forma global, es coherente.

2. **k-core**: Nos permite restringir a un subgrafo en el que los nodos tienen un valor mínimo **k** de grado. Se ha tomado un valor **k** igual a 4.

3. **Rango de grado**: Elimina todos los nodos que no se encuentren en un rango determinado prefijado. Se ha tomado el rango $[5 - \infty)$.

Tras realizar estas modificaciones, se ha reducido el número de nodos a un 1,38 % y el número de aristas al 3,57 % de la red original. A continuación actualizamos las características de la red.

3. Cálculo de los valores de las medidas de análisis

A continuación se ven una serie de características de la red inicial (antes de filtrar):

Característica	**Valor**
Número nodos	21349
Número aristas	30027
Densidad grafo	0
Grado medio	1,406
Grado medio con pesos	1,615
Diámetro	7
Coeficiente de clustering medio	0,073
Componentes fuertemente conexas	21107
Componentes debilmente conexas	4525
Longitud media de camino	1,683

Se puede observar claramente que la **densidad** es 0, ya que el grafo es dirigido y evidentemente no estará completo (no se relacionan todos con todos). Por otro lado el **grado medio** indica que un nodo tiene de media conexiones con 1,4 nodos, lo cual considerando la cantidad de gente dentro de **Twitter** hablando sobre **La Liga** es bastante lógico.

Si observamos el **diámetro** de la red, podemos comprender que habría que pasar por 7 nodos para llegar de un nodo a cualquier nodo (lógico ya que es fácil llegar de cualquier usuario a un equipo oficial, jugador u organismo).

Por último, se remarca la importancia del **coeficiente medio de clustering**, el cual nos indica la probabilidad de que dos nodos cualquiera estén conectados directamente. Siendo 0,073 es un valor lógico considerando lo anteriormente descrito sobre el ámbito de la red.

A continuación, tras el filtrado se observa:

Característica	**Valor**
Número nodos	294
Número aristas	1071
Densidad grafo	0,012
Grado medio	3,643
Grado medio con pesos	7,735
Diámetro	6
Coeficiente de clustering medio	0
Componentes fuertemente conexas	282
Componentes debilmente conexas	1
Longitud media de camino	1.52

Se pueden apreciar una serie de cambios que se describen a continuación:

- La **densidad** aumenta, aunque mínimamente, lo cual indica que este grafo se encuentra más completo.
- Los **grados medios** han aumento, por lo que la red está más conectada.
- El **coeficiente de clustering medio** sale 0, lo cual creo que carece de sentido basándonos en la propia definición de dicho parámetro. Personalmente lo asocio a un error.
- La **longitud media de camino** se reduce, aunque no es un gran cambio, es positivo.

4. Determinación de las propiedades de la red

Se puede observar en la **distribución de los grados** que la mayoría de los nodos filtrados posee un grado bajo, aunque existen determinados casos en los que se posee un grado elevado, esto se debe a cuentas con bastante interacción.

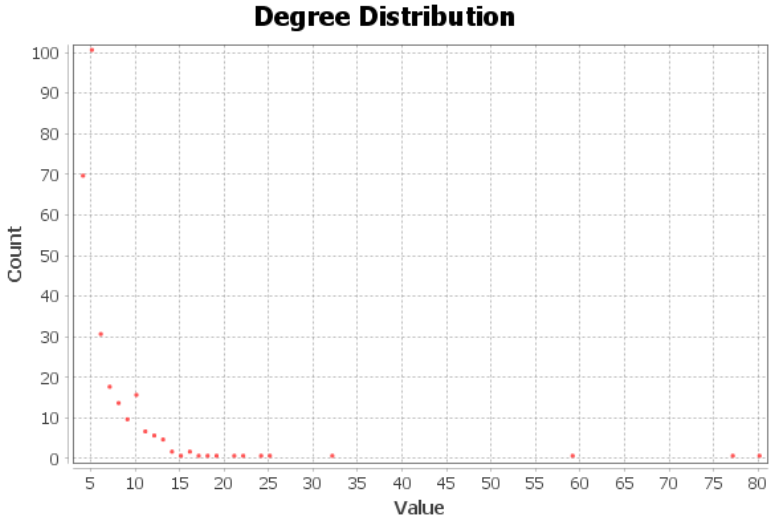


Imagen 4: Distribución de grados

En referencia a los grados de entrada, se puede observar que existe una distribución similar, y que salvo que casos muy aislados, no poseen gran interacción de entrada.

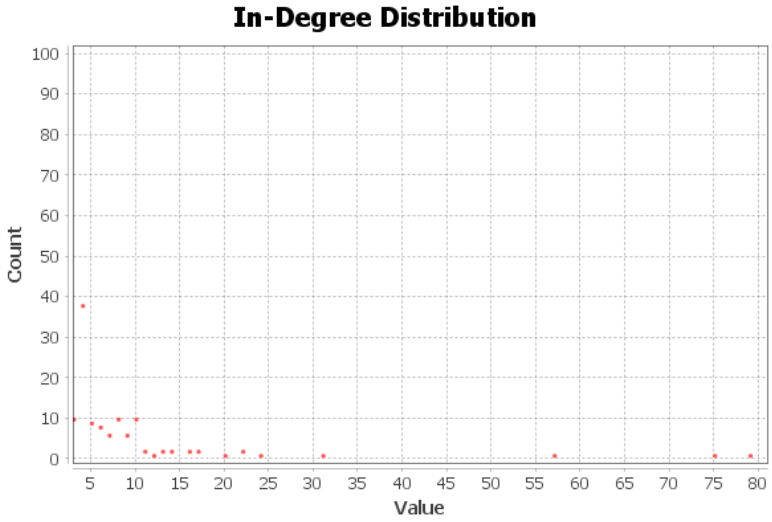


Imagen 5: Distribución de grados de entrada

En el caso de los grados de salida, si se ve, como es lógico, que existen cuentas que reciben pocas interacciones, salvo casos en los que se espera que sean cuentas oficiales donde se reciben más interacciones, pero sin ser tan numerosas como las de entrada.

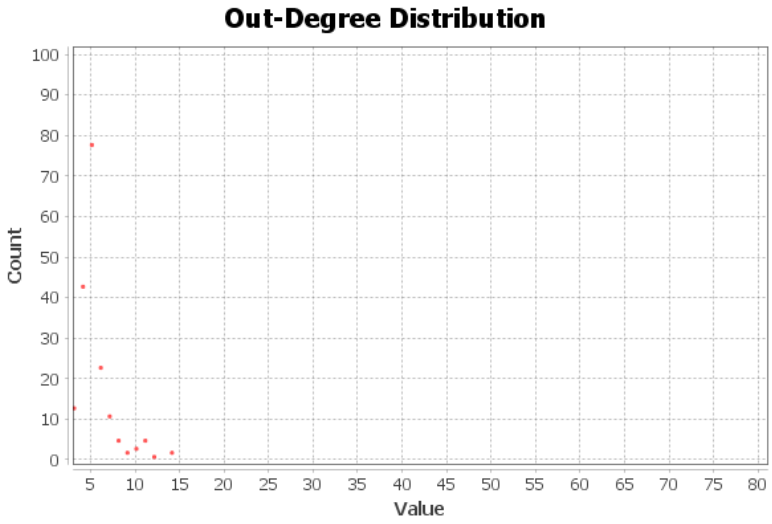


Imagen 6: Distribución de grados de salida

Lo que sí es destacable y curioso es que en la distribución de distancias, existe una distancia media en su mayoría de 2 nodos, lo cual implica que las conexiones entre los usuarios de la red se encuentran muy conectados y cabe esperar comunidades muy cercanas.

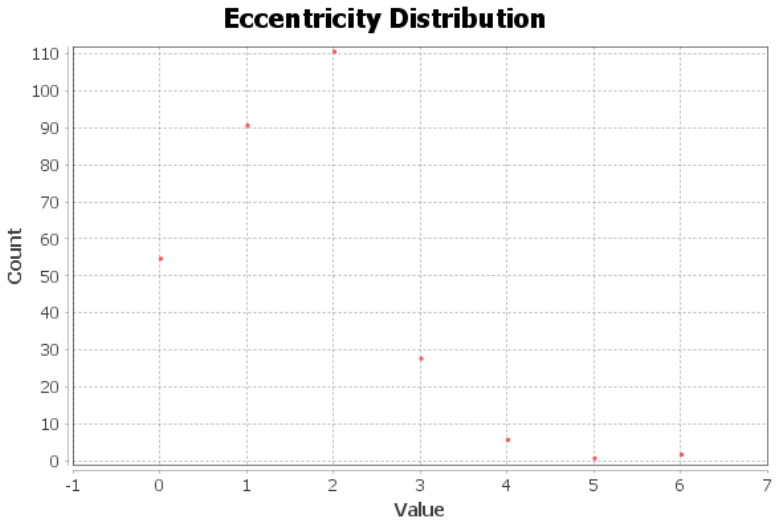


Imagen 7: Distribución de distancias

Por otro lado, si hablamos de la centralidad y cercanía de la red, nos fijamos en la distribución de la cercanía de la red, se puede apreciar que dentro de que existen nodos con *lejanía*, por lo que el centro de la red se encuentra bastante disperso.

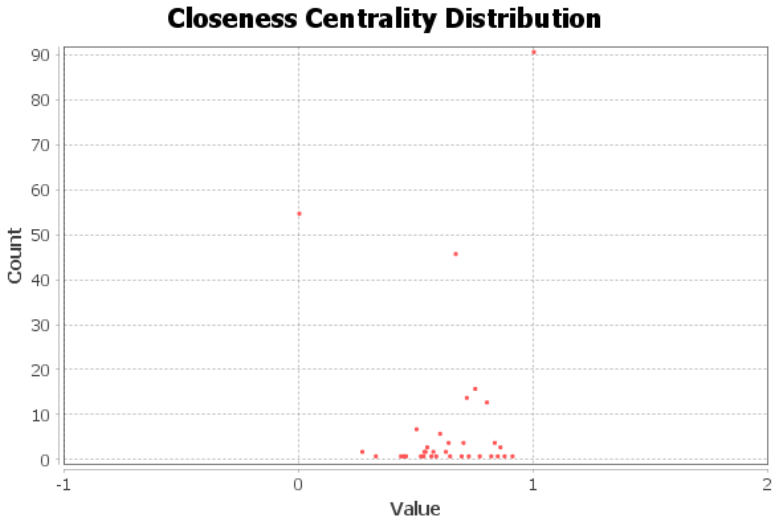


Imagen 8: Distribución de cercanía de la red

Finalmente hablando un poco de la intermediación de los nodos, se aprecia que la mayoría de los nodos tienen una baja intermediación, por lo que se puede observar que no existe una gran cantidad de intermediarios que conecte las redes.

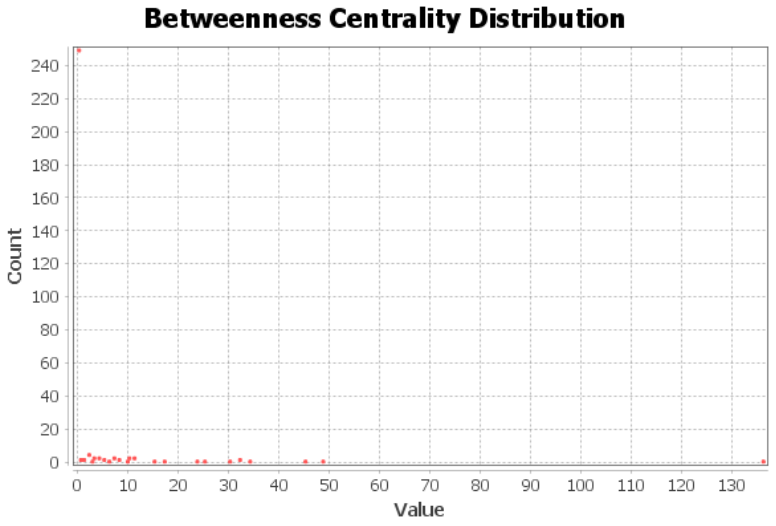


Imagen 9: Distribución de distancias

5. Calculo de los valores de las medidas de análisis de redes sociales

Si hablamos de las cuentas más populares desde el punto de vista de Twitter, podemos mirar por seguidores, siendo lógico que las cuentas más *famosas* sean los propios clubes, jugadores, competiciones y periódicos:

Usuario	**Followers**
@realmadrid	37089802
@championsleague	34298065
@sergioramos	18448180
@fcbarcelona_es	16682558
@tonikroos	9036482
@el_pais	8051410
@laliga	6161755
@marca	5513644
@nachofi1990	2825487
@brfootball	2810498

A continuación, observamos la centralidad de los usuarios en la red definida, donde se aprecia que existen como ejes cuentas similares a las analizadas previamente y donde se çuelanüsuarios de otros ámbitos como la política o el tenis:

Usuario	**Centralidad**
@sergioramos	1
@laliga	1
@nachofi1990	1
@brfootball	1
@didierdrogba	1
@elchiringuitotv	1
@atptour	1
@podemos	1
@eldiarioes	1
@realbetis	1

A continuación observamos la intermediación, donde destaca claramente el usuario *@velascotb*, el cual se trata de un periodista deportivo, y es normal que mantenga una gran interacción con cuentas periodísticas. Se repiten patrones vistos previamente con la centralidad:

Usuario	Intermediación
@velascotb	358.0
@eldiarioes	208.0
@laliga	136.0
@ciudadanoscs	60.0
@publico_es	60.0
@madridxtra	48.5
@absolutechelsea	45.0
@mutuamadridopen	42.3
@begoavila	41.0
@rosadiezglez	36.0

A continuación observamos el grado de entrada, donde como es lógico se aprecian varias celebridades y cuentas conocidas como pueden ser en el ámbito futbolístico futbolistas como **Fede Valverde** o **Vinicius Jr**, y en otros ámbitos como la política **Isabel Díaz Ayuso** o **Juan Guaidó**:

Usuario	Grado de entrada
@fedeevalverde	94
@vinijr	92
@laliga	81
@atptour	64
@idiazayuso	55
@jguaido	51
@realmadrid	37
@rafaelnadal	37
@alexzverev	32
@elchiringuitotv	31

A continuación observamos la centralidad del vector propio, donde se aprecia claramente que en las diferentes cuentas los vectores propios no poseen un alto grado de centralidad, sino que poseen una gran variedad, salvo cuentas como la liga que son el claro centro de forma definida (o en aspecto político Juan Guaidó):

Usuario	Centralidad vector propio
@laliga	1.0
@jguaido	0.679
@realmadrid	0.56
@idiazayuso	0.546
@fcbarcelona_es	55
@5sergiob	0.492
@valenciacf	0.476
@fedeevalverde	0.448
@vinijr	0.432
@rafaelnadal	0.395

6. Descubrimiento de comunidades en la red

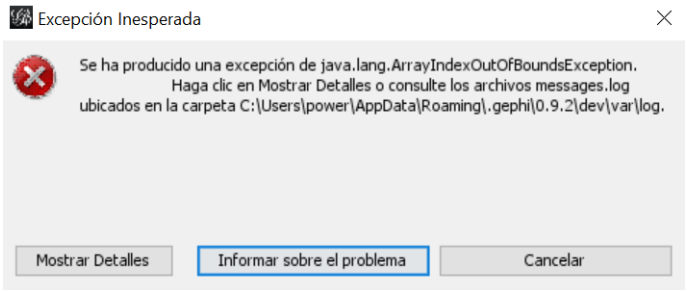
Louvain

Con el objetivo de definir las comunidades que estamos investigando, se ha decidido utilizar la clase de modularidad para poder recoger estas comunidades en un número que consdideremos aceptable teniendo en cuenta la gran comunidad existente en el fútbol.

Para la parametrización se ha utilizado un grado de modularidad de 4.0, obteniendo de esta forma 8 comunidades, las cuales analizaremos posteriormente.

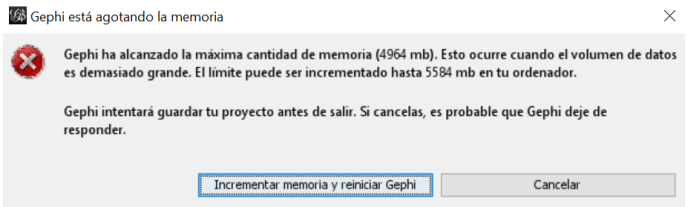
Leiden

Se ha descargado el plugin y ejecutado con semilla 0, pero al lanzarlo siempre produce el mismo error sin posibilidad de saber por qué.



Girvan-Newman

Se ha instalado también el plugin y ejecutado, y tras varios intentos donde constantemente muestra la siguiente ventana y termina dando error, no me ha sido posible ejecutarlo.



7. Visualización de la red social

Se han evaluado diferentes algoritmos de visualización, pero en cuanto a la red, se ha decidido finalmente utilizar el algoritmo de visualización que ya conocemos **Force Atlas 2**, el cual distingue bastante entre las diferentes comunidades que analizaremos posteriormente. La red quedaría de la siguiente forma:

Imagen 10: Red con Force Atlas 2

A continuación, para poder separar las comunidades por colores de los nodos, se ha utilizado una partición por colores utilizando la modularidad de clase (la cual he tenido que trasladar a otra columna por problemas con Gephi).

A continuación se muestra la red con sus comunidades de la siguiente forma:

Imagen 11: Red con Force Atlas 2 y colores

A continuación con el objetivo de visualizar las diferentes comunidades se muestran:

Comunidad 1

En esta primera comunidad no se aprecia ninguna característica, son usuarios desconocidos que simplemente interactúan bastante con algunas de las cuentas de otras comunidades y entre ellas, conformando una comunidad reducida y poco reconocible.

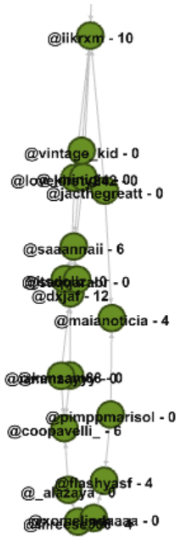


Imagen 12: Comunidad 1: Usuarios desconocidos

Comunidad 2

En esta comunidad se encuentran cuentas con un mismo grado, pero que a su vez es evidente que si nos fijamos existe un patrón. Si conocemos un poco el escenario, y actualidad sobre el mismo, y nos fijamos detenidamente se pueden apreciar usuarios como *@fcbarcelona_es*, *@cristobalsoria* o *5sergiob*. Por lo que se trata de usuarios cercanos al **FC Barcelona**.

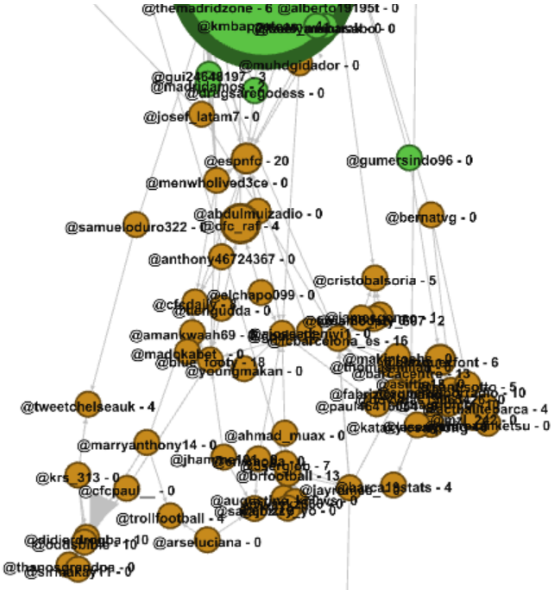


Imagen 13: Comunidad 2: Usuarios cercanos al FC Barcelona

Comunidad 3

Pese a no ser uno de los equipos buscados inicialmente, se aprecia como una de las aficiones más activas, ya que el **Valencia CF** posee su propia comunidad (pequeña ya que evidentemente no era el objetivo) pero es curioso encontrar un equipo histórico no buscado a propósito.

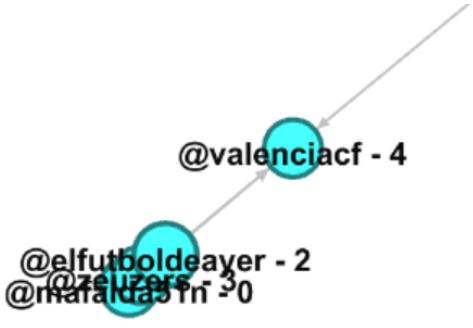


Imagen 14: Comunidad 3: Usuarios cercanos al Valencia CF

Comunidad 4

Otra de las comunidades más llamativas es la formada por empresas y entidades como el **Banco Santander**, **Mutua Madrid** o algunos famosos deportistas como **Rafael Nadal** o **Feliciano López**. Además se encuentran claramente asociados con otra comunidad que se examinará posteriormente (Comunidad 8).

Imagen 15: Comunidad 4: Patrocinadores de la liga o fútbol

Comunidad 5

Esta es una de las comunidades más curiosas, y que quizás se representa pero con el algoritmo utilizado, pero se examinará por partes para poder comprender dicha comunidad. Si se puede apreciar a simple vista algo claro, es una comunidad de periodismo.

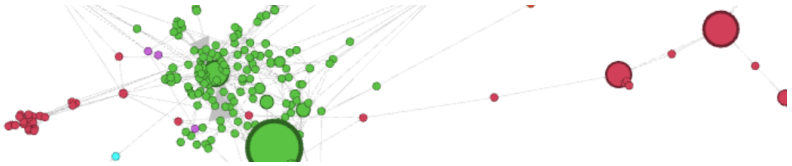


Imagen 16: Comunidad 5: Periodismo

Se visualiza la primera parte, la cual se puede asociar al periodismo tradicional de grandes diarios, entre el cual se destaca claramente **@eldiarioes**, **@iescolar** y **@publico_es**. Se puede asociar a una comunidad más tradicional de periodismo o más genérica.

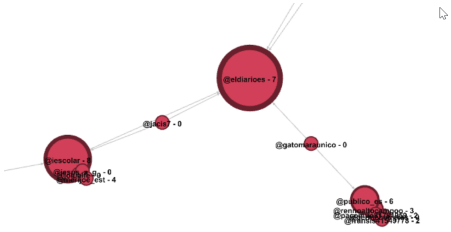


Imagen 17: Comunidad 5.1: Periodismo tradicional

Por otro lado, si observamos la otra parte de la comunidad se ve claramente que se tratan de cuentas de diferentes medios de comunicación (especialmente radio). Lo que sí resulta curioso es que aparezca **Eden Hazard** entre ellos, pero se encuentra claramente cerca a otra comunidad que se explicará posteriormente lo cual es coherente (Comunidad 8).

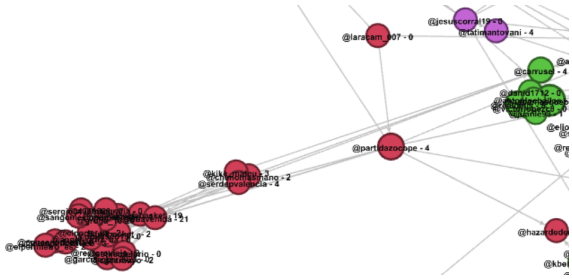


Imagen 18: Comunidad 5.2: Periodismo de radio

Comunidades 6 y 7

Estas comunidades se mencionan de forma conjunta ya que ambas se encuentran muy separadas del resto de las comunidades, salvo pequeños enlaces, y sin embargo se encuentran muy unidas entre sí. Tiene una clara explicación, son comunidades de política.

Se puede apreciar claramente que la comunidad de naranja (6) se asocia más a partidos como **PP** o **Ciudadanos**, mientras que la comunidad rosa (7) se asocia más a partidos como **VOX**, **PSOE** o **Podemos**.

Aunque no es objeto de este trabajo, cabe destacar que es curioso ver como afecta el impacto de las recientes elecciones de Madrid, y como se asocian movimientos similares u opuestos por las constantes interacciones entre ellos (ya sean positivas o negativas).

Esta comunidad evidentemente surge de la relación con ciertos equipos de la capital y principalmente de la relación con los periódicos que se han mencionado previamente.

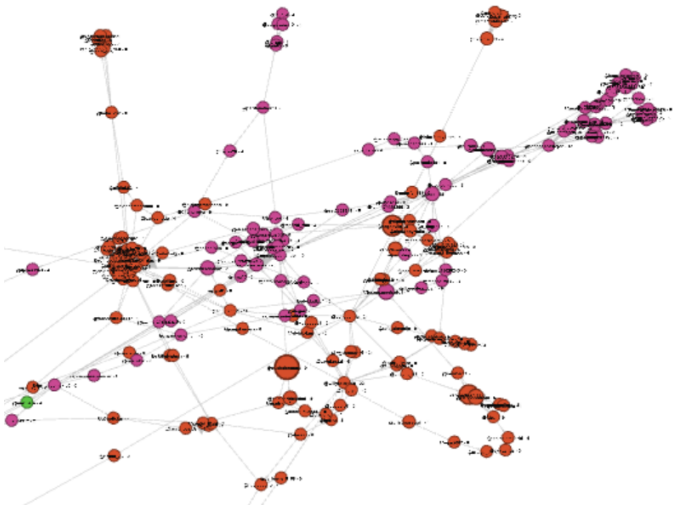


Imagen 19: Comunidades 6 y 7: Política

Comunidad 8

Esta es claramente la comunidad más extensa e influyente, donde se aprecian además los usuarios más dominantes. Aunque considerando el usuario *@velascotb* es un periodista y *@laliga* es un usuario de la competición, si nos centramos en mayor detalle se aprecia un clarísimo patrón, siendo este el de usuarios que giran en torno al **Real Madrid**.

Es lógico que siendo el club más seguido en redes sociales, con más jugadores activos en la misma y más influyente, sea el que más interacciones recibe tanto de organismos oficiales, periodistas, otros clubes o aficionados.

Es de lejos la comunidad más grande y más céntrica, por lo que se comprenden ciertos aspectos que se han mencionado previamente como que **Eden Hazard** pese a pertenecer a otra comunidad se encuentre tan cercano.

También que organismos de la comunidad 4 y sobre todo deportistas reconocidos que son abiertamente aficionados del **Real Madrid**, se encontrarán claramente más asociados a esta comunidad que a la del **FC Barcelona** u otro club.

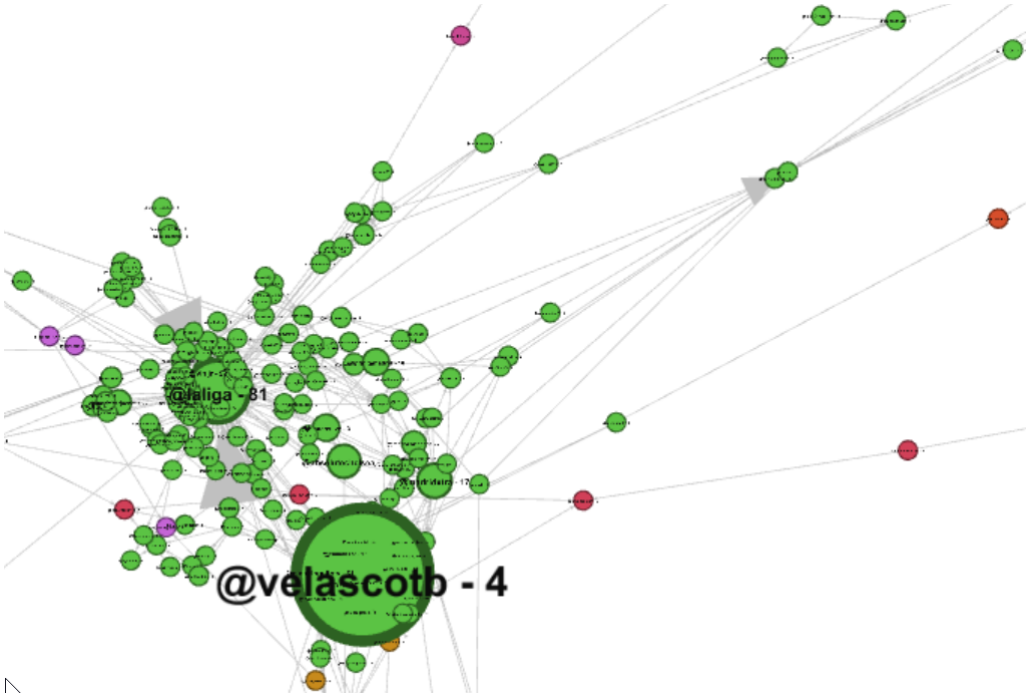


Imagen 20: Comunidad 8: Usuarios cercanos al Real Madrid CF

8. Discusión de los resultados obtenidos

Tras ver las diferentes conclusiones a las que se han llegado en cada uno de los apartados del trabajo, y seguir una seria de ideas y conclusiones, finalmente se procede a anotar una seria de conclusiones a modo de resumen sobre la red estudiada.

- El mundo del fútbol en España tiene una clara influencia sobre el mundo del periodismo y empresas o entidades relevantes en el país.
- Existe una estrecha relación con la comunidad política, ya que no solo a través de los diarios se unen, sino que muchas veces se asocia a la misma, quizás por ideología de las diferentes aficiones de los equipos (Real Madrid se encuentra más cerca que el FC Barcelona, lo cual es coherente dada la realidad social).
- Los periodistasa tienen una gran influencia en el mundo del fútbol, de ahí que *@velascotb* sea el nodo con mayor relevancia de toda la red (aunque obteniendo datos de otros días podría variar).
- El mundo del fútbol se encuentra estrechamente unido, de ahí la aparición de nodos de clubes no explorados inicialmente como el **Valencia CF** o el **Elche CF**.
- Los grandes equipos históricos de la competición como son **Real Madrid** y **FC Barcelona** destacan de lejos frente al resto de clubes dentro de la liga.

Finalmente, tras estas anotaciones, se procede a responder las dos preguntas planteadas inicialmente en el trabajo.

¿Qué club de los líderes interactúa más con su afición?

Ante esta pregunta hay una clara respuesta, y no solo con su afición, sino en general con el panorama futbolístico y a nivel global, el **Real Madrid** es el club que más interactúa o se relaciona con sus aficionados en **Twitter**.

¿Existe una diferencia mediática en redes de los clubes respecto a los medios?

Sorprendentemente, la afirmación en este caso es que **sí** existe una diferencia mediática. Se observa claramente en las diferentes comunidades que el **Real Madrid** y el **FC Barcelona** se encuentran claramente más unidos a la comunidad periodística que otros clubes como son los otros dos aspirantes **Sevilla** y **Atlético de Madrid**, que apenas poseen presencia.

Por último, tras investigar en profundidad, cabe destacar que la escasa presencia de estos últimos dos clubes en redes no implica que sus aficiones no interactuen con ellos. Como es conocido **Real Madrid** y **FC Barcelona** son dos de los principales clubes globales, y es lógico que posean un mayor grado de interacción.

Si eliminasemos dichas interacciones con estos clubes, se obtendrían otros resultados, pero no es objeto de este proyecto.

9. Referencias

[1] Twitter API.

[2] Material asignatura GIW - Máster Profesional en Ingeniería Informática en la UGR.

[3] Gephi. Página oficial.

[4] Forum Gephi. Enlace.