Gestión de Información en la Web
Master en Ingeniería Informática

Práctica 2: Análisis y Evaluación de Redes en Twitter

Luis Alberto Segura Delgado
DNI: 45922174-Y
segura2010@correo.ugr.es

Martes 5 de Abril de 2016

Índice

1 Introducción

El objetivo de esta segunda práctica es formalizar todos los conocimientos adquiridos en el curso aplicándolos a un caso real de análisis de una red social online generada a partir de un medio social. Para ello, se ha seleccionado un medio social concreto (Twitter) y una pregunta de investigación. A partir del medio social elegido, se obtendrá el conjunto de datos y se construirá una red social, que será analizada con objetivo de responder a la pregunta de investigación planteada.

2 Trabajo Realizado

En esta sección se detalla el trabajo realizado en la práctica, indicando en primer lugar el problema concreto que se ha planteado y el conjunto de datos y la forma de obtenerlos para resolver dicho problema. A continuación se explicará el análisis realizado sobre los datos y la red social obtenida y finalmente las conclusiones obtenidas del estudio.

2.1 Descripción del Problema

El problema a estudiar es detectar cuales son los usuarios más relevantes en la discusión de Twitter sobre la emisión en **Periscope**[1] que tuvo lugar el día 25 de marzo, organizada por Gerard Piqué[2].
Para abordar el problema, se han recopilado tweets publicados durante la emisión en los que se mencionaba a Piqué (@3gerardpique) y se incluía la palabra "Periscope". Y como la obtención de los datos se realizó unos días después, se han limitado la búsqueda a los tweets que se publicaron el día 25 de Marzo, día de la emisión[3]. Para obtener los tweets, se ha utilizado la herramienta NodeXL.

De cara a evaluar la red correctamente, se ha decidido eliminar el nodo de Piqué de la red, pues todos los tweets lo mencionan, por tanto se conecta con todos los usuarios, y esto dificulta el análisis de la red y su visualización al mismo tiempo que no resulta interesante.

2.2 Cálculo de los valores de las medidas de análisis

Para el análisis de la red se ha utilizado la herramienta **Gephi**.
Nuestra red social tiene los siguiente valores para las medidas de análisis:

- **Número de Nodos (N)**: 1763
- **Número de Enlaces (L)**: 1464
- **Densidad (D)**: 0.001
- **Grado Medio** ($< k >$): 1.661
- **Diámetro** (d_{max}): 2
- **Distancia Media** ($< d >$): 1.013
- **Distancia Media para la red aleatoria equivalente** ($< d_{aleatoria} >= \frac{log(N)}{log(<k>)}$): 14.73
- **Coeficiente de Clustering Medio** ($< C >$): 0.05

[1] https://www.periscope.tv
[2] http://as.com/videos/2016/03/25/portada/1458916408_738738.html
[3] Búsqueda avanzada de Twitter: @3gerardpique periscope since:2016-03-25 until:2016-03-26 (https://twitter.com/search?vertical=default&q=%403gerardpique%20periscope%20since%3A2016-03-25%20until%3A2016-03-26&src=typd)

- **Coeficiente de Clustering Medio para la red aleatoria equivalente** ($< C_{aleatoria} >= \frac{<k>}{N}$): 0.0009

El número de componentes conexas es de 929, mientras que 883 de los nodos no están conectados con ningún otro, ya que los usuario mencionan principalmente a Piqué (eliminado de la red) y a Iker Casillas. En general los usuarios no se mencionan entre si, salvo excepciones. La componente gigante de nuestra red es Iker Casillas (@casillasworld), ya que recibe la mayor parte de menciones de los usuarios. Tiene un grado de 434 (grado de entrada=434 ; grado de salida=0), por tanto 434 aristas de las 1464 totales son dirigidas a Casillas (un 29.64%). Como vemos, Casillas es, principalmente, el centro de la red. Cosa que tiene sentido, pues es conocido y es el protagonista de la emisión junto a Piqué. Sin embargo, como veremos más adelante, hay otros usuarios importantes en nuestra red, que son mencionados por los espectadores a la hora de comentar el evento.

Los nodos que hemos visto, cuyo grado es 0, no están conectados a ningún otro nodo de la red, pues mencionaban únicamente a Piqué (que ha sido eliminado de la red por ser el centro de la misma). Estos usuarios, no son de gran interés y en general son usuarios aislados que en algún momento han comentado la emisión. El resto de usuario que si están conectados a otros usuarios nos ayudarán a detectar comunidades y actores relevantes de nuestra red. En cuanto a la detección de comunidades, veremos más adelante exactamente que nos cuentan, pero a priori las comunidades de nuestra red deberían representar conversaciones entre diferentes usuarios sobre el evento.

2.3 Propiedades de la red

2.3.1 Distribuciones de Grados

En las figuras 1 y 2 podemos ver las distribuciones de grados de entrada y salida respectivamente, mientras que en la figura 3 podemos ver la distribución global de grados.

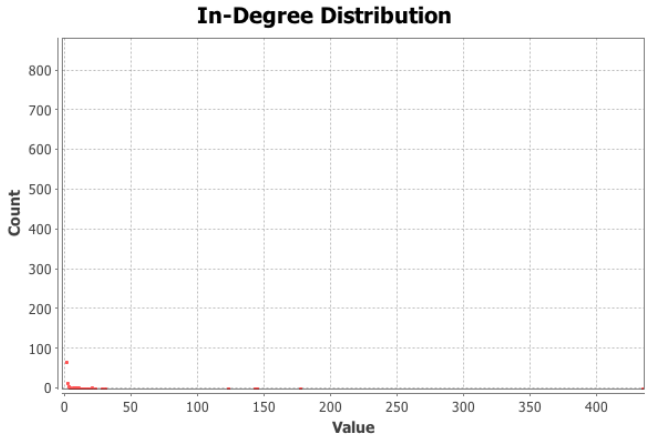

Figura 1: Distribución de grados de entrada

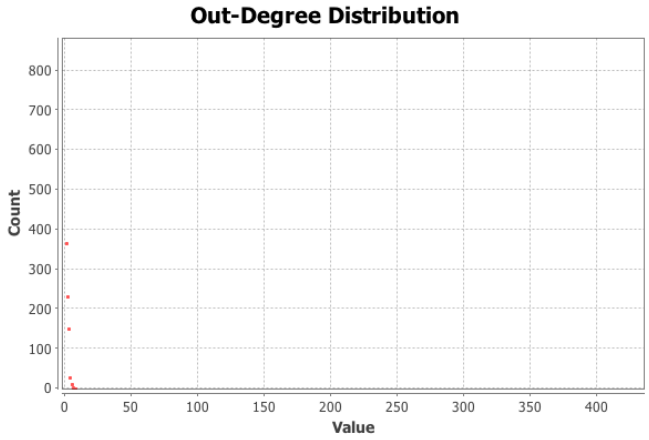

Figura 2: Distribución de grados de salida

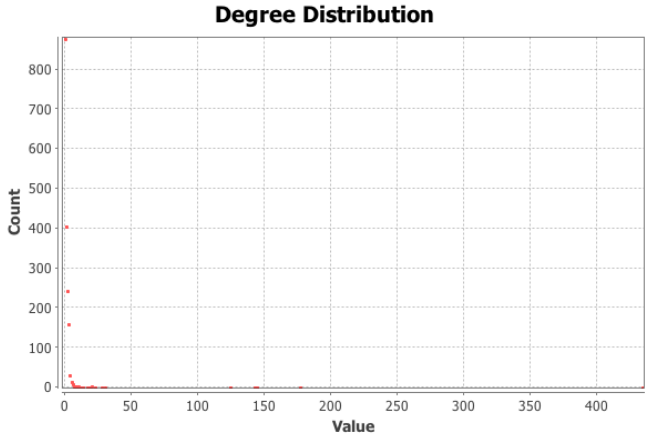

Figura 3: Distribución de grados

A partir de estas distribuciones de probabilidad de los grados de entrada y salida de los nodos de nuestra red social, podemos determinar que nuestra red es libre de escala, y por tanto, que sigue la **Ley de la Potencia**. Las gráficas nos muestran una distribución de larga estela, como ya se ha mencionado en teoría, por ello, **podemos deducir que la red es libre de escala**.

2.3.2 Distribuciones de Distancias

En nuestra red pequeña, nos encontramos ante un caso de *mundo ultra-pequeño*, ya que como vimos anteriormente la distancia media ($< d >$) tenía un valor de 1.01, mientras que la distancia media para una red aleatoria equivalente era de 14.73. Para un mundo ultra-pequeño, la distancia media debe ser aún menor que $\frac{log(N)}{log(log(N))}$, que en este caso obtiene un valor de 3.715, por lo que la distancia media de nuestra red tiene un valor más bajo incluso que la distancia media para un mundo ultra-pequeño. Por tanto, podemos concluir que **nuestra red social es un mundo ultra-pequeño**.

2.3.3 Coeficiente de Clustering

El coeficiente de Clustering nos permite conocer la densidad local de la red, o dicho en otras palabras, la proporción de los vecinos de cada nodo que están conectados. En nuestra red, el coeficiente de clustering

medio es de 0.05. Tenemos un coeficiente de clustering muy bajo, cosa que es lógica si nos fijamos en la gran cantidad de nodos que no están conectados con ningún otro nodo de la red. Como ya vimos antes, al eliminar a Piqué de la red, gran cantidad de nodos quedan totalmente desconectado y no tienen ninguna arista que los una a otros.

2.4 Calculo de los valores de las medidas de análisis de redes sociales

Ya hemos visto el valor medio de los grados, ahora vamos a ver para cada usuario el grado concreto y a tratar de analizar, a partir de esta y otras de las medidas de centralidad, cuales son los usuarios más importantes de nuestra red social. Para empezar, vamos a analizar los actores más interesantes de nuestra red en base al grado. Al trabajar con una red dirigida, trabajaremos con dos grados, el de entrada (que nos indica el prestigio de un usuario a la hora de ser citado/mencionado) y el de salida (que nos indica el alcance de la influencia de un usuario). En nuestro problema, nos interesa más el grado de entrada, pues nos indica los usuarios que han sido más mencionados, y por tanto, los más conocidos y/o interesantes para el resto de usuarios. Como no podría ser de otra manera, el usuario más mencionado (con mayor grado de entrada) es Iker Casillas. Esto ya lo habíamos visto antes, y como decíamos, es lógico al ser el protagonista de la emisión. Veamos que otros usuarios son también mencionados de forma frecuente.

2.4.1 Grado

Usuario	Grado Entrada	Grado Salida
casillasworld	434	0
as_tomasroncero	177	0
elchirincirco	144	0
jpedrerol	143	0
elsimiolopez	123	1
juanmacastano	30	0
hoyendeportes4	28	0
barcastuff	23	0
chuycorona25	20	0
miguel_layun	20	0
mundodeportivo	18	1
abc_deportes	17	0
eukarolyi	15	3
miseleccionmx	14	0
txikiforero	13	0
sefutbol	12	0
sergioramos	10	0
sientelaroja	10	0
jordialba	9	0
marcbartra	9	0

Tabla 1: Usuarios ordenados por grado de entrada

En la tabla 1 podemos ver la lista de usuarios más mencionados. Como podemos ver, a parte de Casillas, el usuario más importante de nuestra red es Tomás Roncero (@as_tomasroncero), seguido por @elchirincirco y @jpedrerol. Estos resultados me parecen interesantes y curiosos, pues muestran algunos detalles interesantes en nuestro problema. Es curioso que los usuarios más mencionados, y por tanto más importantes de nuestra

red desde el punto de vista del grado de entrada, no sean jugadores de fútbol que se encontrasen convocados con la selección española, ya que en el momento de la emisión los jugadores estaban convocados y concentrados en el hotel para afrontar un serie de partidos amistosos. De hecho, me resulta curioso que haya usuarios mas mencionados que Cesc Fabregas, que aparecía junto a Piqué y Casillas en la emisión (aunque con mucho menos protagonismo). Pero de estas curiosidades que nos muestra la red, hablaremos más tarde en la sección dedicada a conclusiones.

2.4.2 Cercanía

Figura 4: Visualización de la parte central de la red. El color de los nodos indica, de menos rojo a más rojo, el valor de la medida de centralidad, de menor a mayor respectivamente. El tamaño de los nodos representa el grado

En la figura 4 podemos ver como queda la visualización de la parte central de la red, en la que el color de los nodos nos indica el valor de la medida de centralidad (Harmonic Closeness Centrality en Gephi). Si nos fijamos, esta medida no nos da demasiada información en nuestra red, ya que al estar gran parte de los nodos conectados a 'hubs', casi todos los nodos (usuarios) son importantes desde el punto de vista de esta medida. Como sabemos, el planteamiento de esta medida es diferente a la de intermediación (que veremos a continuación) y al grado, la centralidad trata de evaluar la cercanía de un nodo al centro de la red, a otros nodos centrales, a los hubs. Y como hemos visto en la imagen de nuestra red de la práctica, se puede ver facilmente esto, como los nodos cercanos y conectados a los nodos centrales de la red obtienen un valor mayor de centralidad que el resto. Y por como es nuestra red, en la que muchos están conectados directamente a los hubs (Casillas, Roncero, ...), muchos de los nodos tienen una centralidad cuyo valor es 1. En la figura 4 no se ha incluido toda la red, sino una parte central para que se vea bien la medida de centralidad. Sin embargo, hay mucho nodos que han quedado totalmente desconectados al eliminar a Piqué de la red, como ya se ha comentado anteriormente. Con lo cual, estos nodos tienen un valor de centralidad igual a 0, cosa que, con lo que sabemos, es lógico. Por tanto, está medida, al igual que el grado, nos permite ver cuales son los hubs, con respecto a lo centrales que son. Y al igual que el grado, los nodos centrales (con un valor menor de centralidad, pero que permiten que los que se conecta a ellos tengan un valor alto), son prácticamente los

mismo que veíamos según el grado.

2.4.3 Intermediación

Usuario	Intermediación
eukarolyi	6
pepvergeli	6
jorditenerife95	3
franarteaga	2
culealcalaino	1
damosasa	1
real_oscarinn	1

Tabla 2: Usuarios ordenados por Intermediación

Como vemos en la tabla 2, tenemos usuarios menos conocidos que son los que obtienen los mejores valores para la medida de intermediación. Como sabemos, la medida de intermediación trata de reflejar en que medida un nodo se encuentra en el centro, en la zona intermedia entre diferentes conjuntos de nodos, de forma que la única forma de que la información fluya de un grupo a otro sea a través de él, permitiéndole controlar los que pasa de un lado a otro. Estos nodos conectan grupos de nodos de forma que pueden controlar la información que se intercambia entre los grupos, por ello, desde el punto de vista de esta medida, estos son los usuarios más importantes.

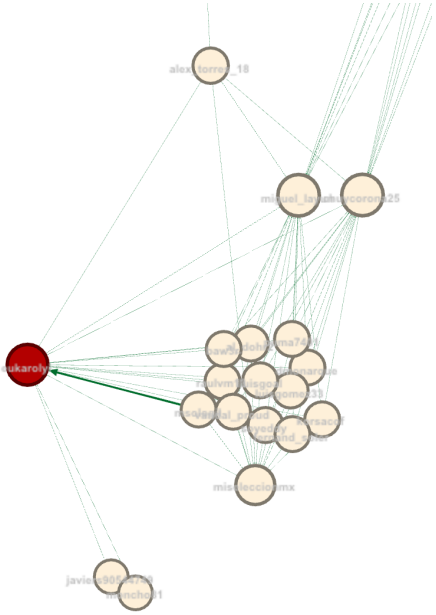

Figura 5: Visualización de la parte de la red unida por el nodo con mayor intermediación. El color de los nodos indica, de menos rojo a más rojo, el valor de la medida de intermediación, de menor a mayor respectivamente. El tamaño de los nodos representa el grado

En la figura 5 podemos ver lo que hemos visto anteriormente. El usuario eukarolyi, es el encargado de conectar los nodos de ese grupo (a simple vista esta es una especie de comunidad cuyo centro y el que une

a todos los nodos es este usuario, veremos más tarde que nos dice el algoritmo de comunidades de Lovaina).

2.4.4 Vector Propio

Usuario	Vector Propio
casillasworld	1
as_tomasroncero	0.467376
elchirincirco	0.327179
jpedrerol	0.32491
elsimiolopez	0.279027
juanmacastano	0.068055
hoyendeportes4	0.063518

Tabla 3: Usuarios ordenados por Vector Propio

La centralidad de Vector Propio es una versión más elaborada de la centralidad de grado, por ello, los resultados son prácticamente iguales, y como podemos ver en la tabla 3, los usuarios más importantes de nuestra red desde el punto de vista de esta medida, son los mismos que nos indicaba la centralidad de grado. De hecho, si nos fijamos en la figura 6, podemos ver como también los colores de los nodos y el tamaño (vector propio y grado) coinciden al representar a los nodos más importantes.

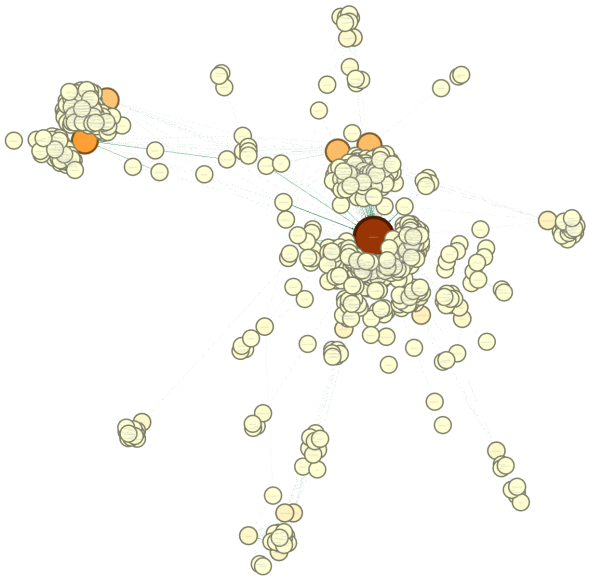

Figura 6: Visualización de la parte central de la red. El color de los nodos indica, de menos rojo a más rojo, el valor de la medida de Vector Propio, de menor a mayor respectivamente. El tamaño de los nodos representa el grado

2.4.5 PageRank

El PageRank es una forma diferente de calcular el vector propio, basandose en colocarse de forma aleatoria en un nodo e ir recorriendo recursivamente los enlaces de un nodo a otro, y contabilizando cuantas veces llegamos a los nodos a través de otros. De esta forma se reajusta el PageRank para cada nodo en función de

las veces que llegamos a él a través de otros y el peso que tengan en ese momento los nodos que conectan con él. Personalmente es de las medidas que más me gusta, ya que su planteamiento me parece muy interesante. En nuestro problema, los actores más importantes siguen siendo los casi los mismos que nos encontrabamos con el grado y con el vector propio de las subsecciones anteriores, ya que nuestro problema es muy concreto y esta muy claro quienes son los actores importantes.

Usuario	PageRank
casillasworld	0.095674
as_tomasroncero	0.054631
elsimiolopez	0.020393
jpedrerol	0.016179
elchirincirco	0.016122
hoyendeportes4	0.008424
abc_deportes	0.006155
juanmacastano	0.005478

Tabla 4: Usuarios ordenados por PageRank

Con el PageRank, como podemos ver en la tabla 4, aparecen usuarios que estaban posicionados peor según otras medidas. Por ejemplo, *@elsimiolopez* es más importante según el PageRank, de lo que lo era según el grado o según el vector propio. Aunque básicamente, todos los usuarios destacados por el PageRank, también eran importantes (algo más o algo menos) según el resto de medidas.

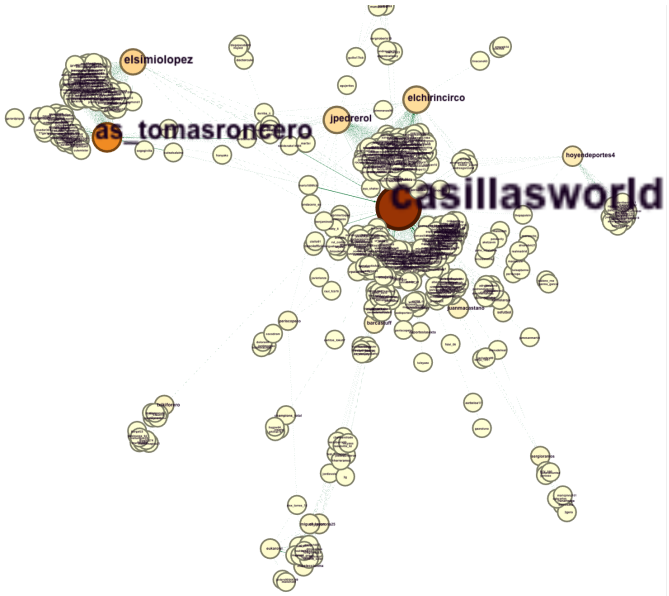

Figura 7: Visualización de la parte central de la red. El color de los nodos indica, de menos rojo a más rojo, el valor de la medida de PageRank, de menor a mayor respectivamente. El tamaño de los nodos representa el grado

En la figura 7 podemos ver una representación en la que los nodos representan el PageRank. Como vemos, es muy parecida a la visualización que obteníamos con el grado y con el vector propio.

2.4.6 Medidas más interesantes

En esta sección hemos estudiado las medidas de centralidad en nuestra red, pero no todas son igual de interesantes de cara a obtener información en nuestro problema concreto. Entre todas las medidas que hemos visto, bajo mi punto de vista y en vista de los resultados obtenidos por cada una de ellas de cara a responder la pregunta que nos hemos planteado, las más interesantes son **PageRank, Vector Propio y grado**. La intermediación puede ser interesante, pero para nuestro estudio no lo es, ya que solo nos permite descubrir usuario muy concretos que en pequeños grupos aparecen como intermediarios. Estos pequeños grupos no son de mucho interés en nuestro estudio. Tampoco es de gran interés la cercanía, aunque esta claro que de cara a utilizarla en la visualización para detectar los hubs/nodos centrales, puede ser interesante; aunque para ello ya tenemos las medidas que hemos dicho que si son interesante, como el PageRank o el vector propio.

2.5 Descubrimiento de comunidades

A continuación, vamos a tratar de detectar comunidades en nuestra red. Anteriormente ya nos habíamos imaginado, a partir de la visualización de la red, que podrían existir algunas comunidades, y ahora es el momento de comprobarlo utilizando un algoritmo especifico para tal efecto. Debido a las limitaciones de la última versión de Gephi por las cuales no están disponibles todos los algoritmos de detección de comunidades, solamente se ha utilizado el algoritmo de Lovaina.
En un primer intento, utilizando como valor para el parámetro de resolución 1.0, la separación en comunidades, desde el punto de vista de la medida de Modularidad, si hay comunidades, ya que se obtiene un valor Q=0,647. El problema es que, aunque ese valor es muy bueno, implica que se detecten 942 comunidades, que vienen principalmente de los nodos inconexos de nuestra red (a causa de eliminar a Piqué). Sin embargo, como podemos ver en la figura 8, las comunidades que ha detectado parecen correctas y muy acertadas.

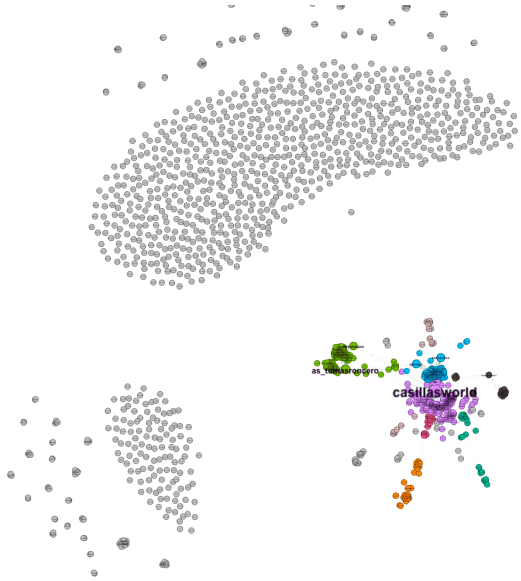

Figura 8: Comunidades de la red detectadas por el algoritmo de Lovaina

En la figura 8 podemos ver que las 942 comunidades, son realmente comunidades de nodos individuales inconexos. Realmente tenemos 8 comunidades importantes, como podemos ver en la figura 9, mientras que el resto son comunidades que ha detectado el algoritmo correspondientes a nodos aislados, y que no nos interesan.

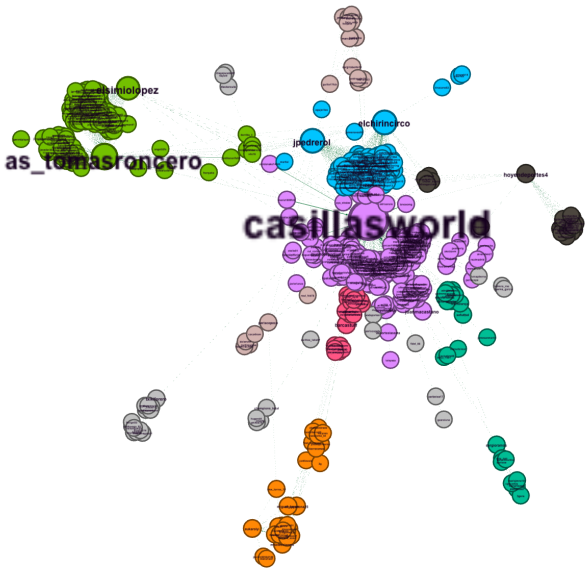

Figura 9: Comunidades interesantes de la red detectadas por el algoritmo de Lovaina

De las 8 comunidades que si que son interesantes, podemos ver las siguientes:

- casillasworld: Comunidad formada, principalmente, por los usuarios que han mencionado a Casillas. No todos los usuarios que han mencionado a Casillas forman parte de esta red, hay algunos otros que al haber mencionado en el mismo tweet a otros usuarios de otras comunidades han quedado asignados a esas otras comunidades (Ver figura 10).
- as_tomasroncero y elsimiolopez: Comunidad formada, principalmente, por los usuarios que han mencionado a Roncero y a elsimiolopez. Aparentemente buscando hacer comentarios graciosos sobre Roncero y mencionando a elsimiolopez, que es una cuenta de humor. Básicamente podríamos decir que vemos a los alumnos enseñándole lo aprendido a su profesor de humor, aunque con los tweets recogidos no les hace ningún caso.
- elchirincirco y jpedrerol: Comunidad formada, principalmente, por los usuarios que han mencionado a elchirincirco y a jpedrerol. Esta comunidad refleja exactamente lo mismo que la anterior, usuarios mencionando a jpedrerol para hacer la gracia y que elchirincirco las lea y les retwitee. Esto es algo muy común en Twitter.
- hoyendeportes4: Comunidad formada, principalmente, por los usuarios que han mencionado a hoyendeportes4. La intención de estos usuarios parece ser avisar a la cuenta de Deportes Cuatro que Piqué está haciendo una emisión.
- sefutbol y sergioramos: Comunidad formada, principalmente, por los usuarios que han mencionado a sefutbol y a sergioramos (Ver figura 11).
- barcastuff: Comunidad formada, principalmente, por los usuarios que han mencionado a barcastuff.
- miguel_layun y chuycorona25: Comunidad formada, principalmente, por los usuarios que han mencionado a miguel_layun y a chuycorona25 (Ver figura 12).

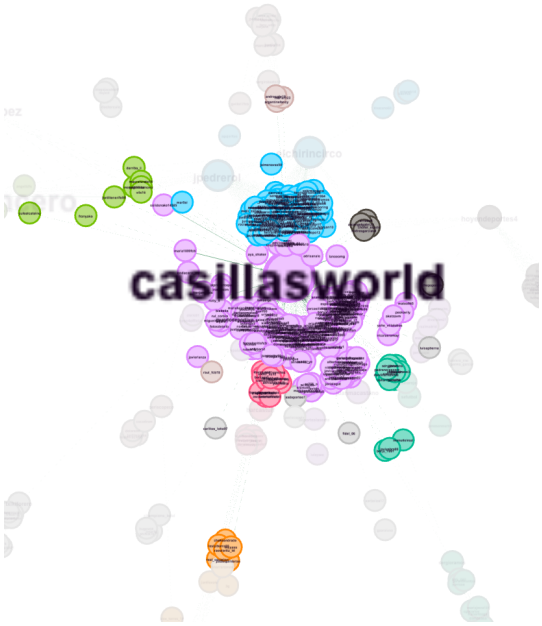

Figura 10: Comunidad Casillas

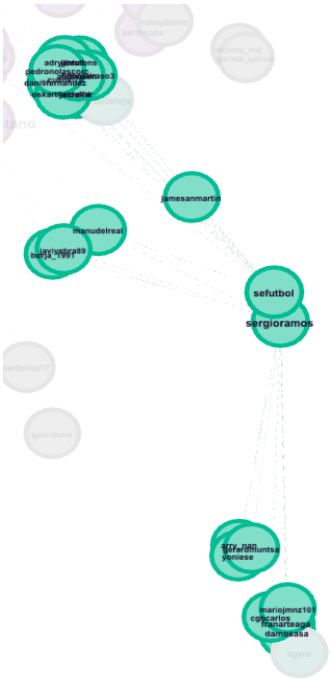

Figura 11: Comunidad sefutbol y sergioramos

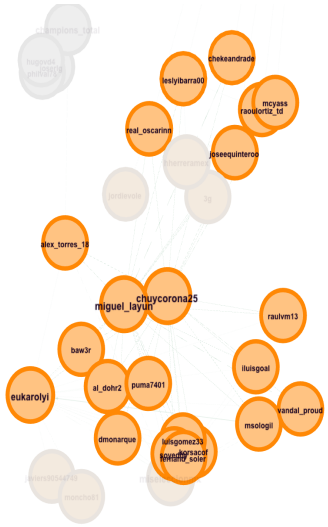

Figura 12: Comunidad de miguel_layun y chuycorona25

2.6 Visualización de la red social

2.7 Discusión de los resultados y Conclusiones