Máster en Ciencia de Datos e Ingeniería de Computadores

Minería de Medios Sociales

Bloque I.1: Análisis de una Red Social con Gephi

Curso 2021-2022

Autora
Lidia Sánchez Mérida
76065456-V

Contacto
lidiasm96@correo.ugr.es

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

Granada, Abril de 2022

Análisis básico de la red

La red social seleccionada para realizar esta práctica se denomina ***rt-voteonedirection*** procedente de la web *Network Repository*. Se trata de una **red dirigida** compuesta por un total de **4.273 nodos** correspondientes a usuarios de Twitter y **4.935 enlaces** que representan los *retuits* generados entre todas las cuentas que participaron en el ***hashtag voteonedirection***. Con él se pretendía apoyar a la banda de música *One Direction* en la celebración de los *MTV Video Music Awards* del año 2012, una gala orientada a premiar los mejores videoclips de diferentes categorías musicales en el ámbito internacional.

A continuación, en la Figura 1 se puede observar una representación de la red global. La primera característica destacable es la existencia de un **conjunto de grupos muy voluminosos situados en el centro del gráfico**. Presumiblemente estarán compuestos por un conjunto de usuarios que han interaccionado masivamente con publicaciones de otras cuentas, aumentando así el número de *retuits*. No obstante, también se pueden observar **otros grupos de nodos situados en la periferia** del gráfico de menor tamaño aunque suficiente para ser visibles fácilmente.

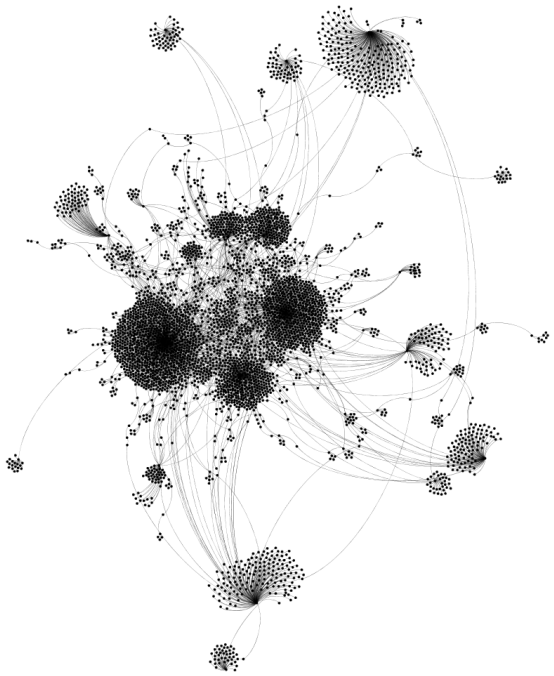

Figura 1. Representación gráfica de la red completa.

En la Tabla 1 se recopilan los valores de las diferentes métricas de estudio para esta red social. Comenzamos describiendo la fórmula para calcular el **número de enlaces máximos de una red dirigida**:

$N * (N - 1)$, siendo *N* el número de nodos existentes.

Comparando el número de enlaces reales con el número de enlaces máximos teóricos podemos observar que la **diferencia es sumamente notoria**, del orden de hasta catorce millones. Como consecuencia, la **densidad de la red es tan considerablemente inferior** que este valor ha sido calculado manualmente puesto que Gephi está limitado a un máximo de tres decimales. Para ello hemos aplicado la siguiente fórmula extraída de la fuente [1]:

$\frac{|E|}{|V|*(|V|-1)}$, donde E es el número de enlaces y V el número de nodos.

Una densidad con un valor prácticamente despreciable indica que la conectividad entre los diferentes nodos de la red es prácticamente nula, lo que teóricamente favorece la existencia de multitud de **nodos aislados**. Esta teoría se apoya en el valor tan ínfimo del grado medio, que representa un total de una conexión de media para cada nodo. Si lo traducimos al ámbito de esta red social, **cada usuario únicamente ha *retuiteado* una publicación** en base a la media. Por otro lado, también disponemos de un coeficiente medio de clustering de cero, lo que simboliza una **conectividad local baja**. Como consecuencia, podemos determinar que apenas existen interacciones entre los usuarios que han hecho uso del *hashtag* monitorizado durante la emisión del programa musical.

Medida	**Valor**
Número de nodos *N*	4273
Número de enlaces *L*	4935
Número máximo de enlaces *Lmax*	18.254.256
Densidad del grafo *L/Lmax*	2,70E-04
Grado medio <*k*>	1,155
Diámetro *dmax*	7
Distancia media *d*	2
Coeficiente medio de clustering <*C*>	0
Número de componentes conexas	1
Número de nodos componente gigante (y %)	4273(100%)
Número de aristas componente gigante (y %)	4935(100%)

Tabla 1. Medidas básicas de una red social.

Observando las medidas del diámetro y la distancia media, podemos determinar que solo hacen falta un total de **siete nodos para alcanzar al usuario más aislado** mediante un camino mínimo compuesto por, aproximadamente, dos nodos. Ambas métricas simbolizan que esta red dispone de un **alto grado de difusión** puesto que, comparado con el número de usuarios total, se necesita una cantidad de ellos muy inferior para propagar un *tweet* por la red completa.

Si bien hemos concluido que la conectividad global y local de la red es especialmente escasa en base a los valores de ciertas métricas, como la densidad, el grado medio y el coeficiente medio de clustering, **la componente gigante coincide con la red completa**. Una consecuencia directa de esta cualidad es que **no existen nodos sin conexiones**, es decir, anteriormente hemos concluido que los usuarios que han utilizado el *hashtag voteonedirection* apenas han generado interacciones entre sí, pero como mínimo han interaccionado con una publicación de otro usuario para aparecer en esta red social.

Distribución de grados

En la Figura 2 podemos observar la **distribución de grados de entrada** que representa la proporción de usuarios cuyos *tweets* han sido *retuiteados* por otros usuarios que han utilizado también el *hashtag voteonedirection*. Tal y como se puede apreciar en esta primera representación, la mayoría de los puntos se concentran en el primer intervalo [0, 50] donde los valores son mínimos. Como consecuencia podemos determinar que la **mayoría de usuarios apenas han recibido *retuits*** por parte de otras cuentas que han utilizado el mismo *hashtag*. También es notable la existencia de unos pocos nodos situados sobre valores más elevados en el eje X, representando a aquellos usuarios o ***nodos hubs*** que han recibido un **mayor número de interaccione**s en comparación con la mayoría. La cola que se genera como consecuencia de este fenómeno representa una cualidad asociada a las redes sociales **libres de escala,** una propiedad que modela la probabilidad de que una minoría de nodos tengan un número de relaciones entrantes considerablemente mayor a la media.

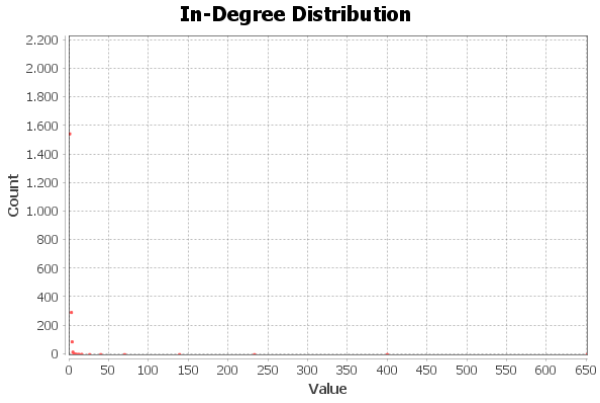

Figura 2. Distribución de grados de entrada.

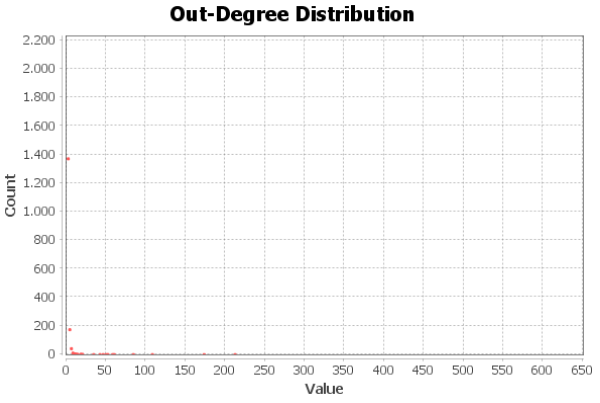

Figura 3. Distribución de grados de salida.

De forma similar, en la Figura 3 se encuentra la **distribución de grados de salida** compuesta por el número de conexiones producidas por cada nodo. Como podemos observar en la gráfica, la mayoría de ellos se encuentran en el primer intervalo de valores mínimos, por lo que podemos confirmar que **la mayoría de usuarios tampoco han realizado apenas algún *retuit*** del contenido publicado por otras cuentas. Por lo tanto, la interacción de la mayoría de participantes del *hashtag voteonedirection* es prácticamente insignificante. Sin embargo, algunos **usuarios registran unos valores notablemente altos** y diferentes de la media de interacciones salientes, y por ende estos ***nodos hubs*** se sitúan en intervalos más elevados del

gráfico. Nuevamente conforman una estela, como ocurre en la distribución de grados de entrada, aunque ligeramente menos acentuada.

Conectividad de la red

Tal y como se indicó en la Tabla 1, en la Figura 4 podemos visualizar que en esta red social existe **una única componente** que se encuentra formada por todos los nodos y enlaces disponibles. Por lo tanto la **componente gigante coincide con la red completa.** Una posible explicación a este fenómeno puede encontrarse en la misma naturaleza de la red. Al estar modelada a partir de los *retuits* producidos entre usuarios que participaban en el *hashtag voteonedirection*, **cada cuenta ha debido recibir o realizar un *retuit*** para aparecer en esta red particular.

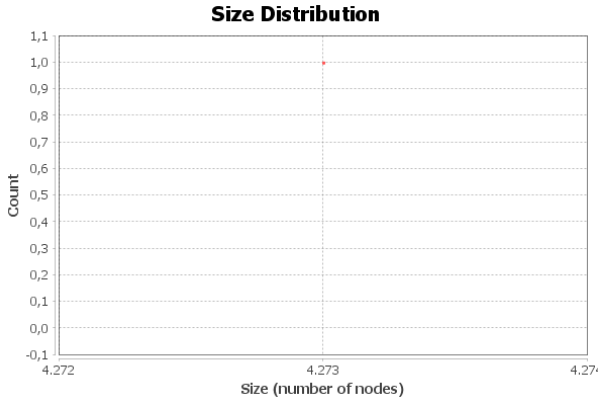

Figura 4. Distribución de componentes conexas.

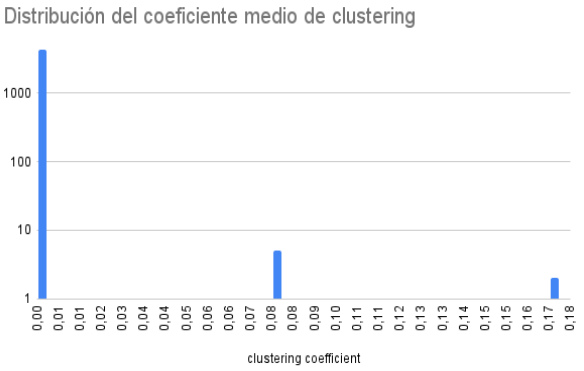

Figura 5. Distribución del coeficiente medio de clustering.

Si observamos la Figura 5 en la que se encuentra representada la **distribución del coeficiente medio de clustering,** con una escala logarítmica para apreciar mejor los resultados y graficada con *Google Calc* ante la imposibilidad de generarla con Gephi, podemos apreciar que se refuerza la teoría de la **escasa participación** en el *hashtag* monitorizado. Existe una elevada concentración de nodos en torno al primer valor representando una **conectividad local escasa**, como se había comentado en el análisis anterior. No obstante, también podemos apreciar la localización de diversos nodos con un coeficiente de clustering mayor al resto. Considerando la naturaleza de la red, se trata de **usuarios que presentan una alta conectividad entre sí**. Este conjunto de nodos **no actúan como *hubs,*** puesto que este tipo de usuarios suele tener un coeficiente de clustering bastante inferior ya que su función consiste en conectar nodos aislados y no suelen disponer de una red de vecinos tan amplia. Probablemente se trate de un grupo de usuarios que se conozcan entre sí y/o que compartan el mismo gusto por esta banda musical para realizar *retuits* al contenido que han publicado bajo el *hashtag voteonedirection*.

Estudio de la Centralidad de los Actores

El objetivo de esta sección consiste en identificar y analizar los roles más relevantes de algunos de los nodos que componen esta red social. Comenzamos con el estudio de la **centralidad de grado** que representa a aquellos **usuarios que han realizado o recibido un mayor número de *retuits***. Al tratarse de una red dirigida, es más interesante dividir esta métrica en los grados de entrada o **soporte**, cuyos valores se encuentran en la columna *Centralidad (GE)*, y los grados de salida o **influencia** localizada en la segunda columna denominada *Centralidad (GS)*. En el primer caso podemos visualizar los identificadores de los **usuarios que más *retuits* han recibido** al utilizar el *hashtag voteonedirection*. Presumiblemente, estos nodos se encuentran vinculados a cuentas populares de Twitter que generan publicaciones de interés sobre este grupo musical. Sin embargo, en la segunda columna se encuentran los **usuarios que más *retuits* han realizado** sobre el contenido de otras cuentas. Por lo tanto, es posible que se trate de seguidores de la banda musical a los que probablemente les interese el contenido que generan las cuentas representadas en la primera columna.

Centralidad (GE)	Centralidad (GS)	Intermediación	Cercanía	Vector propio
1534 : 650	*1370 : 212*	*611 : 0.000676*	*611 : 1.0*	*1534 : 1.0*
310 : 399	*1425 : 173*	*2058 : 0.000199*	*2058 : 1.0*	*310 : 0.596748*
611 : 232	*34 : 108*	*764 : 0.000108*	*1105 : 1.0*	*1346977909 : 0.49950874*
1105 : 138	*1986 : 84*	*272 : 0.000074*	*1652 : 1.0*	*1523 : 0.49680058925892*
2058 : 69	*2092 : 60*	*1105 : 0.000062*	*702 : 1.0*	*1347074532 : 0.49680058*

Tabla 2. Medidas de centralidad mostrando los cinco nodos con mayor valor para cada métrica.

A continuación procedemos a analizar los resultados de la **intermediación**, cuyo objetivo consiste en identificar a aquellos **nodos involucrados en el máximo número de caminos**. De este modo, los usuarios con mayor valor disponen de la habilidad de controlar la información que reciben los demás nodos, pudiendo añadir o suprimir flujos de datos a su antojo. Como consecuencia, los identificadores de la tercera columna se corresponden con **usuarios que han podido acceder a una mayor cantidad de *retuits*** en comparación con el resto de participantes.

En la Figura 6 se representa un gráfico de la red social en el que se combinan dos de las medidas anteriormente explicadas. El color de los nodos representa la intermediación, mientras que el tamaño está asociado al grado de entrada. El objetivo consiste en descubrir si aquellos usuarios que reciben un **mayor número de interacciones también disponen de la capacidad de controlar el flujo de información**. Tal y como podemos apreciar, los nodos que representan a los usuarios más populares se sitúan en el centro de la red con un tamaño

más considerable con respecto al resto. Sin embargo, debido a su tonalidad azulada podemos determinar que disponen de un menor valor de intermediación, por lo que **rechazamos la hipótesis** formulada previamente. El único nodo con un grado de intermediación considerablemente alto se encuentra coloreado de rojo y localizado alrededor de los nodos más populares, con un tamaño más moderado. En base a estas características, por un lado podemos determinar que **este usuario es el único en la red con la capacidad de controlar el flujo de información** al que pueden acceder el resto de cuentas que han participado en este *hashtag*, mientras que se encuentra **dentro de la lista de usuarios más populares** puesto que su tamaño refleja un número de *retuits* recibidos mayor que la media del resto de nodos.

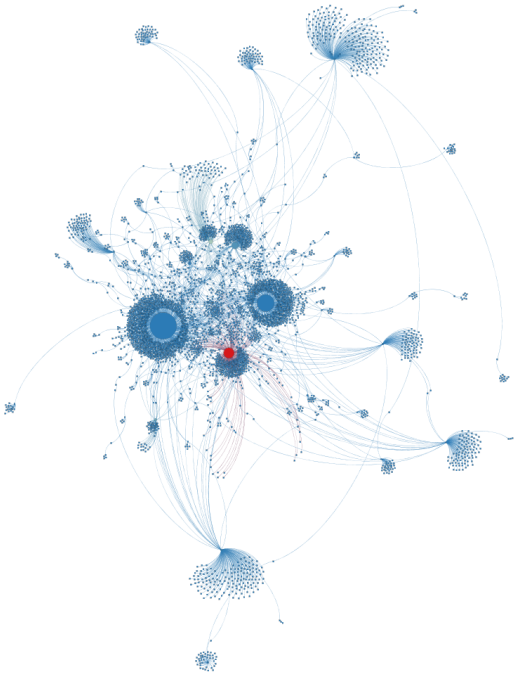

Figura 6. Representación gráfica de la red completa coloreando los nodos con el grado de entrada y estableciendo su tamaño según la intermediación.

Continuamos con el análisis de una métrica conocida como **cercanía**, cuyo objetivo consiste en clasificar como relevantes a aquellos **nodos que se encuentran cerca del centro de la red**, basándose en la hipótesis de que en dicha localización se sitúan los usuarios más importantes de la red social. Observando los valores de esta métrica en la Tabla 2 podemos apreciar que los dos primeros usuarios **coinciden con los mejor valorados según la intermediación**. Como consecuencia podemos intuir que existe una relación que conecta ambas métricas, y en el caso particular de esta red, se ha podido observar en el gráfico anterior que el usuario con mayor control de información se encuentra localizado en el centro de la red, entre los nodos de mayor tamaño y las cuentas de usuarios tradicionales.

Finalmente procedemos a realizar un estudio sobre la **centralidad de vector propio** que intenta identificar los nodos relevantes a partir tanto de **su propia relevancia como la de sus vecinos**. De nuevo, podemos observar una **relación entre esta métrica y el grado de**

entrada puesto que ambas cuentan con los mismos nodos en los dos primeros puestos del *ranking* visualizado en la Tabla 2. A diferencia del caso anterior, esta conexión está fuertemente fundamentada en que el cálculo de la centralidad de vector propio se trata de una **versión mejorada de la centralidad de grado**, a la que añade la importancia de los nodos de alrededor para calcular la centralidad de uno particular. Al tratarse de una red dirigida y haber dividido la centralidad de grado en entrante y saliente, podemos determinar que la centralidad de vector propio de estos dos primeros usuarios parece encontrarse altamente influenciada por la cantidad de *retuits* que han recibido.

En la Figura 7 se representa gráficamente la red social utilizando el algoritmo de visualización *Fruchterman Reingold* para representar los nodos con colores que reflejan la cercanía y fijando sus tamaños según la centralidad de vector propio. Como podemos apreciar existe un número considerable de **usuarios que ocupan posiciones centrales en la red** representados con un color rosáceo. Sin embargo todos ellos se caracterizan por tener un tamaño ínfimo, lo que indica que **no son relevantes en función del vector propio** ya sea por falta de importancia propia o de su vecindario. Este hecho pone de manifiesto que, a diferencia de las métricas anteriores, la **centralidad de cercanía y de vector propio son considerablemente diferentes**. Contrariamente, aquellos nodos con mayor tamaño se encuentran coloreados mediante tonalidades verdosas, por lo que no se encuentran alrededor del centro de la red pero sí son identificados como relevantes gracias a la importancia propia y a la de sus vecindarios. Probablemente se trate de **cuentas populares entre los *fans*** de la banda musical dentro de la red social Twitter. Para confirmar esta teoría se ha generado un segundo gráfico en la Figura 8 con la misma configuración que el de la Figura 7, estableciendo el color en función del grado de entrada. Así podemos apreciar a los dos usuarios destacados en el análisis anterior coloreados con tonalidades rosáceas.

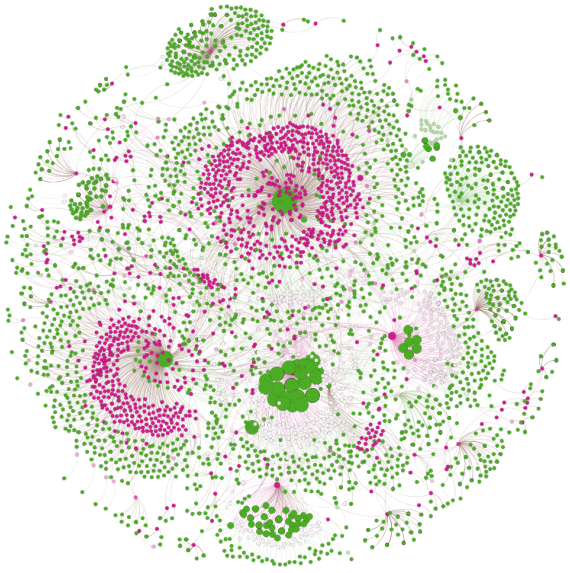

Figura 7. Representación de la red completa

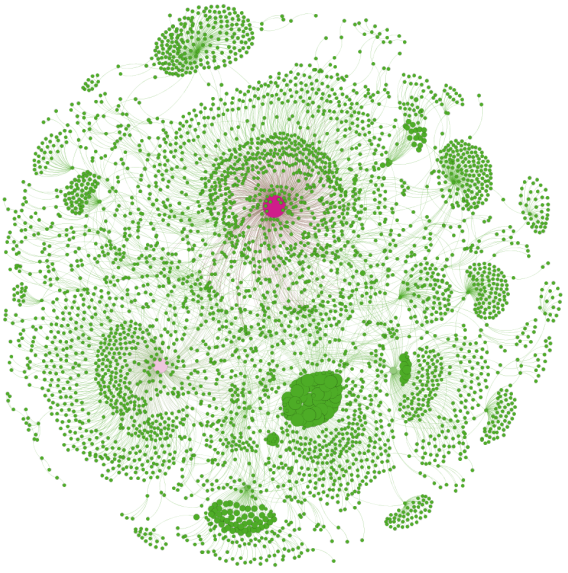

Figura 8. Representación de la red completa

coloreando los nodos con cercanía y su tamaño con vector propio.

coloreando los nodos según el grado de entrada y su tamaño con el vector propio.

Detección de Comunidades con el algoritmo de Lovaina

En la Tabla 3 se encuentran los valores resultantes tras aplicar el **método de Lovaina para realizar un análisis de las comunidades** existentes en la red social completa. Como podemos observar, a menor resolución mayor valor de modularidad aunque también aumenta considerablemente el número de comunidades. Con valores iguales o menores a uno el análisis de estas agrupaciones es prácticamente imposible de realizar. Por lo tanto, para esta red social particular, resulta **beneficioso disminuir el valor de la resolución** para conseguir un número de comunidades razonable, así como una modularidad considerablemente alta para poder realizar un estudio de las agrupaciones generadas con cierta fiabilidad.

Resolución	Modularidad	Número de comunidades
0.4	0.786	278
0.7	0.809	63
1.0	0.809	44
5.0	0.761	11
10.0	0.623	5

Tabla 3. Comunidades obtenidas tras aplicar el método de Lovaina con diferentes resoluciones

A continuación, la Tabla 4 contiene los valores relativos a las comunidades identificadas tras aplicar el método de Lovaina con una resolución igual a diez. Tal y como podemos apreciar, las **dos primeras comunidades reúnen prácticamente al 80% de los nodos** totales. Comparando el número de usuarios de cada agrupación con el número de conexiones, podemos observar que existen numerosas conexiones entre ellos, por lo que podemos determinar que **todas las comunidades son altamente cohesivas.**

Número de comunidad	Número de nodos	Porcentaje de nodos	Número de enlaces	Porcentaje de enlaces
Comunidad 1	1717	40.18%	1997	40.47%
Comunidad 3	1580	36.98%	1726	34.97%
Comunidad 2	402	9.41%	406	8.23%
Comunidad 4	323	7.56%	323	6.55%
Comunidad 0	251	5.87%	250	5.07%

Tabla 4. Composición de las comunidades obtenidas con Lovaina y resolución igual a diez.

En la Figura 9 se representan gráficamente las cinco comunidades cuyos valores han sido mostrados y analizados anteriormente. La comunidad con mayor población se encuentra situada en el lateral izquierdo y marcada de color rojo, mientras que la segunda más popular se localiza a la derecha de color amarillo. Uno de los aspectos más destacables es que **ambas comunidades se encuentran mezcladas** entre sí, por lo que si aumentamos progresivamente el valor de la resolución, probablemente el algoritmo de Lovaina agrupe todos los nodos en una sola comunidad. Adicionalmente, podemos determinar que las dos agrupaciones se posicionan en torno a **dos nodos *hubs*** que se corresponden con los nodos de mayor tamaño en sendas comunidades: los usuarios con identificadores 1534 y 310, respectivamente. Si recordamos el análisis de centralidad de actores realizado en la sección anterior, estos dos usuarios se caracterizan por disponer de los **valores más altos en las medidas de centralidad del grado de entrada y del vector propio**. Como ya se comentó anteriormente, probablemente se trate de dos cuentas orientadas al seguimiento de la banda musical y por ende reciben una gran cantidad de interacciones por parte de los *fans* en Twitter.

Finalmente, a diferencia de las dos comunidades previas, las tres agrupaciones restantes disponen de un menor número de nodos y enlaces, además de caracterizarse por una alta dispersión, puesto que existen **diferentes subcomunidades dentro de cada comunidad** localizadas por la periferia del gráfico. Este comportamiento suele ser característico de **grupos de usuarios que se conocen** e interactúan entre sí, con la capacidad de ampliar su círculo de conexiones en caso de que algunos de ellos se relacionen de forma similar con nodos de otras subcomunidades.

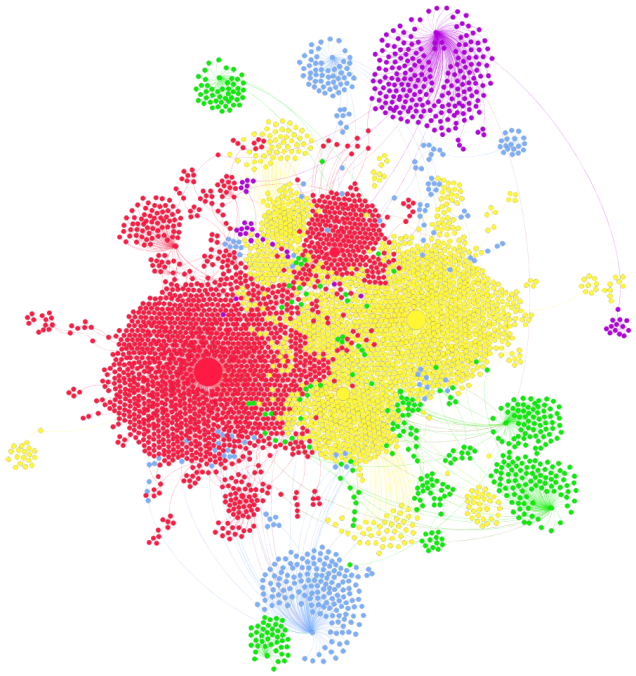

Figura 9. Cinco comunidades generadas mediante el método de Lovaina con resolución igual a diez.

Detección de Comunidades con el algoritmo de Leiden

El segundo algoritmo seleccionado para la detección de comunidades es el **método de Leiden**, cuyo objetivo consiste en componer comunidades en las que todos los nodos se encuentran conectados entre sí. A diferencia del algoritmo de Lovaina, cuanto menor es la resolución, menor número de comunidades de mayor calidad en base al incremento de la modularidad. Sin embargo, **a partir de una resolución particular, el número de agrupaciones no disminuye** hasta que su valor es tan considerablemente inferior que el resultado converge a una sola comunidad. En el caso de esta red, tal y como podemos observar en la Tabla 6, a partir de 0.1 el número de comunidades apenas varía de siete a ocho.

Resolución	Modularidad	Número de comunidades
0.01	0.9900000000000000	1
0.04	0.9630781869434030	7
0.07	0.9401235531944910	7
0.1	0.9236541785805200	8
0.4	0.8734064789655120	21
0.7	0.8398571114603860	38
1.0	0.8123190927843320	51
5.0	0.6279552358137060	626
10.0	0.5380072473062850	884

Tabla 6. Comunidades obtenidas tras aplicar el método de Leiden con diferentes resoluciones.

En la Tabla 7 se representan los valores generados tras aplicar una segunda detección de comunidades con el algoritmo de Leiden y una resolución de 0.04, puesto que es con este valor con el que se alcanza un menor número de agrupaciones maximizando la modularidad. Como podemos apreciar, la **primera comunidad contiene más del 88% de los nodos** totales de la red. Antes de visualizar el resultado gráficamente, podemos intuir que este método probablemente ha **agrupado las dos comunidades de mayor población identificadas con Lovaina**. En mi opinión, parece que la fusión de sendos grupos es lógica ya que, tal y como se observó en la Figura 9, los nodos de sendas comunidades se encuentran entremezclados. Sin embargo, al aumentar considerablemente el número de participantes en la primera comunidad y generar una mayor cantidad de agrupaciones, la mayoría de comunidades son prácticamente insignificantes en referencia al porcentaje de nodos y enlaces que contienen. No obstante, el **valor de la modularidad es significativamente más elevado** que el conseguido por el método de Lovaina, lo que indica que la cohesión entre las siete comunidades generadas es mayor que la relativa a las cinco comunidades analizadas en la sección anterior.

Número de comunidad	Número de nodos	Porcentaje de nodos	Número de enlaces	Porcentaje de enlaces
Comunidad 0	3765	88.11%	4410	89.36%
Comunidad 1	276	6.46%	275	5.57%
Comunidad 2	98	2.29%	97	1.97%
Comunidad 3	54	1.26%	53	1.07%
Comunidad 4	43	1.01%	42	0.85%
Comunidad 5	19	0.44%	18	0.36%
Comunidad 6	18	0.42%	17	0.34%

Tabla 7. Composición de las comunidades obtenidas con Leiden y resolución igual a 0.04.

Finalmente en la Figura 10 se representan las siete comunidades generadas por el algoritmo de Leiden estableciendo como resolución el valor 0.04. La primera característica más destacable es la confirmación de la teoría previa relacionada con la **fusión de las dos comunidades más populares** obtenidas con el algoritmo de Lovaina. En este caso se trata de la *comunidad 0* representada en color rojo, la cual se sitúa en torno a los dos usuarios con mayor centralidad de grado de entrada y de vector propio mencionados en el estudio de la sección anterior.

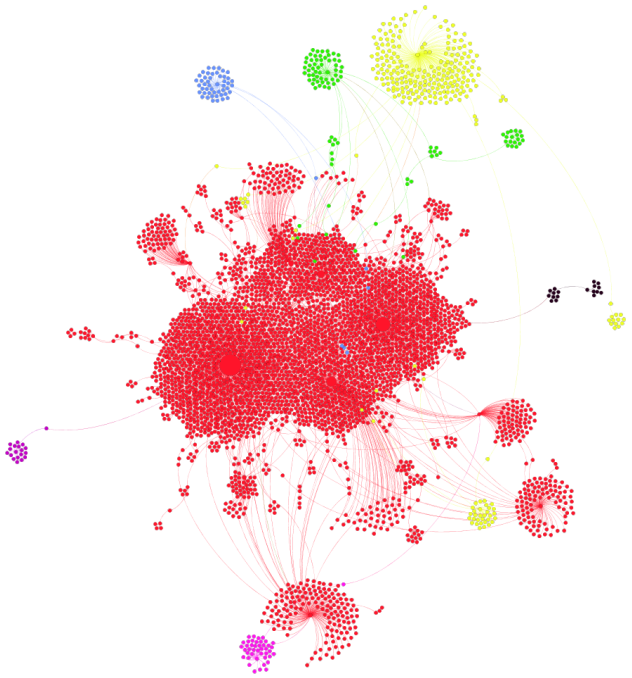

Figura 10. Siete comunidades generadas mediante el método de Leiden con resolución igual a 0.04.

Si comparamos las agrupaciones resultantes de la aplicación de sendos algoritmos, podemos apreciar que la mayoría de las **comunidades representadas con tonalidades verdosas y azuladas también han sido fusionadas,** por el algoritmo de Leiden, dentro del grupo mayoritario coloreado en rojo. Una posible explicación a este fenómeno reside en la existencia de multitud de nodos pertenecientes a estas comunidades entremezclados en el grupo de mayor población. Por lo tanto, la generación de estas agrupaciones puede ser representativo del error asociado al algoritmo de Lovaina, consistente en introducir nodos en una misma comunidad aún no estando conectados entre sí. Sin embargo, el método de Leiden al aplicar otra metodología para agrupar los diferentes nodos, es capaz de solventar este fenómeno garantizando que todos los nodos incluidos en una comunidad se encuentren conectados entre sí.

Referencias

1. Stackoverflow, What is the definition of the density of a graph?, https://math.stackexchange.com/questions/1526372/what-is-the-definition-of-the-density-of-a-graph