Gestión de Información en la Web

UNIVERSIDAD DE GRANADA

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

Autor
Juan Manuel Castillo Nievas

Máster Profesional en Ingeniería Informática 2020/2021

—

Granada, 28 de mayo de 2021

Índice

1. Introducción

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 Twitter. El análisis se ha hecho usando la herramienta **Gephi** [1].

Para la realización de esta práctica y aprovechando la celebración del concurso de Eurovisión y el gran impacto mediático que tiene (sobre todo en Twitter), se han planteado las siguientes preguntas de investigación:

- ¿Qué usuarios de Twitter tuvieron un número mayor de interacciones durante la segunda semifinal de Eurovisión 2021?
- ¿Cuáles son las cuentas que se deben seguir si alguien está empezando a introducirse en el mundo de Eurovisión?

2. Construcción de la red social a analizar y visualizar

El conjunto de datos se ha obtenido utilizando la herramienta **Twitter Streaming Importer** [2] que es un plugin adicional de Gephi. La búsqueda que se ha hecho para obtener el conjunto de datos es muy sencilla y se compone exclusivamente de dos hashtags: #Eurovision y #Eurovision2021 (ver Figura 1).

Esta búsqueda tuvo lugar durante el desarrollo de la segunda semifinal de Eurovisión el día **20 de mayo** desde las **21:00 hasta las 23:00**, aproximadamente. Cuando finalizó el programa, se habían recolectado un total de **4852 nodos** y **8783 aristas**. En la Figura 2 se muestra una primera visualización de la red completa.

El grafo obtenido tiene las siguientes características:

- **Nodos:** representan los usuarios de Twitter que comentan con los hashtags de Eurovisión.
- **Aristas:** representan las conexiones entre los usuarios de Twitter, bien a través de respuestas, retweets, o las recientes citas de Twitter.
- Es un grafo **dirigido**.

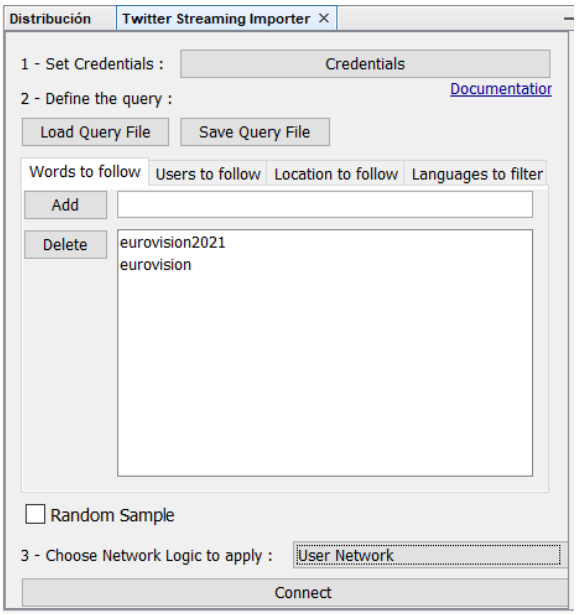


Figura 1: Obtención del conjunto de datos con Twitter Streaming Importer

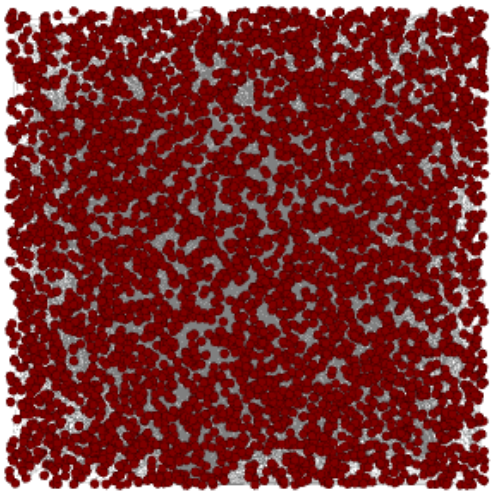


Figura 2: Visualización de la red

2.1. Visualización

Para una mejor visualización de la red se ha aplicado el algoritmo **Force Atlas 2**, que es una versión mejorada del clásico **Force Atlas**. Este algoritmo está basado en el **Kamada Kawai**, en el peso de los enlaces. No todos los enlaces van a tener la misma longitud, sino que consideran los caminos mínimos entre cada par de nodos.

En la Figura 3 se muestra la visualización obtenida. Se ha aumentado la gravedad para que algunos nodos no se fueran tan lejos del centro.

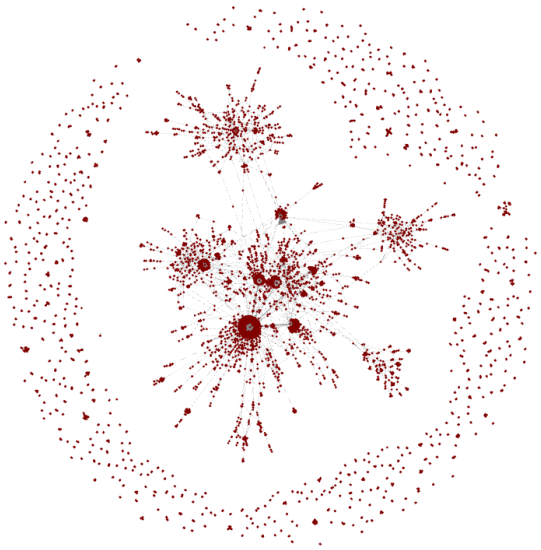


Figura 3: Visualización de la red aplicando **Force Atlas 2**

2.2. Poda

Para reducir el número de nodos y aristas y teniendo en cuenta la visualización que se ha obtenido usando Force Atlas 2, se han aplicado los siguientes criterios para hacer una poda sin perder la información más relevante:

2.2.1. Componente gigante

Primero se ha obtenido la componente gigante de la red, es decir, la componente que conecta el mayor número de nodos. En la Figura 4 se muestra la red obtenida. Ahora hay **3166 nodos** y **6999 aristas**.

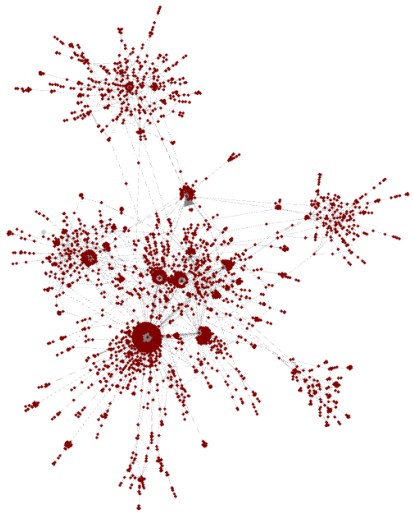
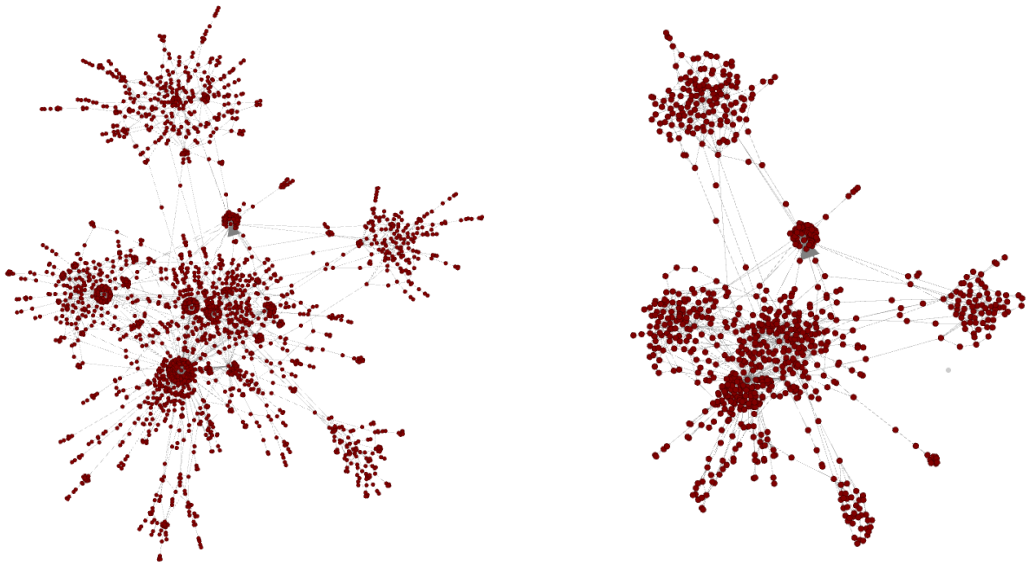


Figura 4: Componente gigante de la red

2.2.2. K-core

El filtro **K-core** se usa para el filtrado de subredes. Para formar parte de la nueva red, cualquier nodo que pertenezca a la red debe tener mínimo un grado k. En la Figura 5 se muestra una comparación de aplicar este algoritmo con los valores **k=2** y **k=3** para ver cuál es la poda que mejor se obtiene.

Se ha decidido aplicar finalmente un K-core de **k=3** porque se eliminan muchos más nodos y aristas, ya que con k=2 se queda con el 53.94 % de la red y con k=3 se queda con un **17.89 %** de la red.



(a) K-core con k=2

(b) K-core con k=3

Figura 5: Algoritmo K-core

3. Cálculo de los valores de las medidas de análisis

En la Tabla 1 se muestran los valores obtenidos para la red inicial.

Medida	Valor
Número de nodos N	4.852
Número de enlaces L	8.783
Densidad D	0
Grado medio k	1.81
Diámetro d_max	5
Distancia media d	1.43
Distancia media para la red aleatoria	$ln(4852)/ln(1{,}81) = 14{,}3$
Coeficiente medio de clustering C	0,025
Número de componentes conexas	665
Número de nodos componente gigante (y %)	3.166 (65.25 %)
Número de aristas componente gigante (y %)	6.999 (79.69 %)

Tabla 1: Valores de la red inicial

De estos valores se puede sacar el siguiente análisis:

- El **grado medio de la red es de 1.81**, que indica que cada cada usuario de Twitter interactúa con casi otros 2 usuarios.
- El **coeficiente medio de clustering es de 0.025**, que indica que hay muy poca probabilidad de que los vecinos de un nodo sean también sean vecinos entre sí. Es un valor muy bajo.
- El **diámetro máximo es de 5** y la **distancia media es de 1.43**. Estamos ante un **mundo muy pequeño**.
- La **componente conexa gigante** abarca un **65.25 % de nodos** y un **79.69 % de aristas** de la red.
- La **densidad de grafo es 0**. Un grafo denso es un grafo en el que el número de aristas es cercano al número máximo de aristas. En este caso, estamos ante un **grafo totalmente disperso**.

En la Tabla 2 se muestran los valores obtenidos después de haber hecho la poda.

Medida	Valor
Número de nodos N	868
Número de enlaces L	2.978
Densidad D	0.004
Grado medio k	3.43
Diámetro d_max	5
Distancia media d	1.39
Distancia media para la red aleatoria	$ln(868)/ln(3{,}43) = 5{,}49$
Coeficiente medio de clustering C	0
Número de componentes conexas	1
Porcentaje de nodos	17.89 %
Porcentaje de aristas	33.91 %

Tabla 2: Valores de la red después de la poda

De estos valores podemos sacar el siguiente análisis:

- Se ha quedado el **17.89 % de nodos** y el **33.91 % de aristas**.
- La **densidad de grafo sube a 0.004**. No es mucho, pero algo sube.
- El **grado medio sube a 3.43**. Quiere decir que ahora cada usuario interactúa con una media de entre 3 y 4 usuarios.
- La **distancia media es de 1.39**, muy parecida a la anterior, por lo que no cambia mucho.
- El **coeficiente medio de clustering es 0**. No es habitual encontrar un coefiente medio de clustering con este valor.
- El **diámetro es 5**, con lo cual **no cambia**.

En conclusión, a pesar de haber reducido un porcentaje bastante importante de nodos y aristas, el análisis que queda no varía tan drásticamente, con lo cual se ha hecho una buena poda.

4. Determinación de las propiedades de la red

4.1. Distribución de grados

En la Figura 6 se muestra la distribución de grados de la red. Se puede ver que se da la **propiedad de libre escala** porque hay muy pocos nodos que estén altamente conectados, mientras que la mayoría de nodos presentan muy pocas conexiones. Sólo hay dos nodos que tengan un grado mayor a 150, y la mayoría de nodos poseen un grado menor a 40.

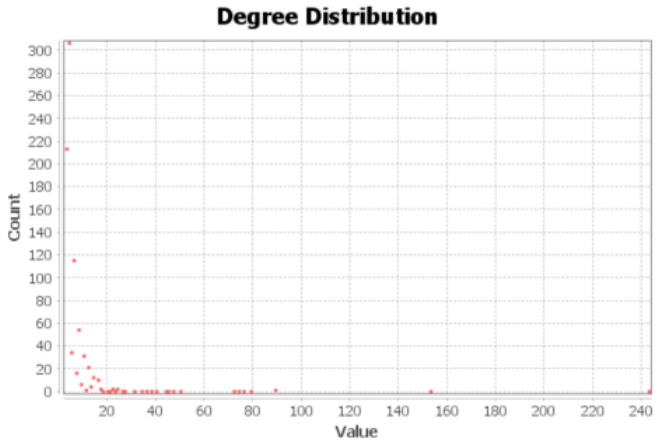
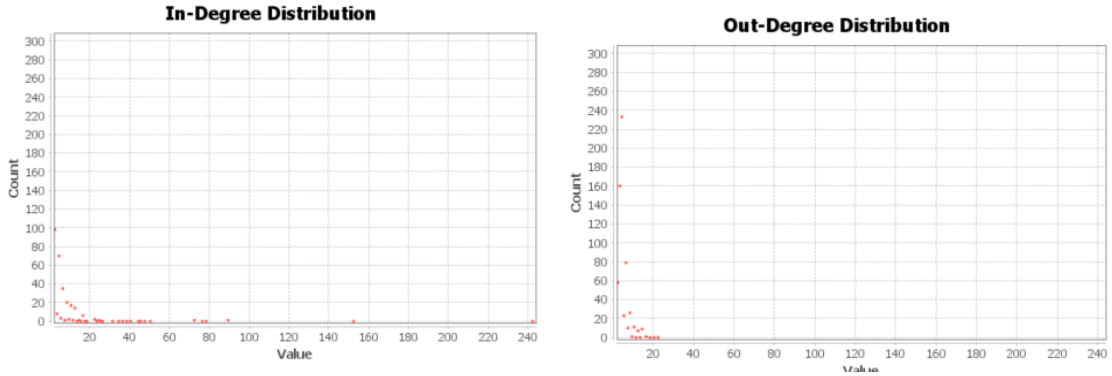


Figura 6: Distribución de grados

Al ser un **grafo dirigido**, en la Figura 7 se muestra la distribución de grados de entrada y salida. La **distribución de grados de entrada** indica el número de conexiones hacia un nodo en concreto, y se puede observar que algunos nodos presentan grandes conexiones y que, a pesar de que algunos tienen pocas conexiones, el número de estos nodos es muy alto. Sin embargo, la **distribución de grados de salida** indica el número total de conexiones que salen de un nodo en concreto, es decir, no reciben muchas interacciones. En este caso, hay varias cuentas que reciben pocas interacciones, pero hay que tener en cuenta que hay muchísimas más cuentas que sí reciben muchas interacciones (grados de entrada).



(a) Distribución de grados de entrada (b) Distribución de grados de salida

Figura 7: Distribución de grados de entrada y salida

4.2. Distribución de distancias

Para empezar esta sección se va a probar la **propiedad de mundos pequeños**. Para ello, la distancia media de la red debe tener una escala logarítmica (o incluso menor) con respecto al tamaño de la red. En este caso, **la distancia media de la red es 1.39**, y **la distancia media para la red aleatoria es 5.49**. La distancia media es mucho menor que la distancia media para la red aleatoria, por lo tanto **se cumple la propiedad de mundos pequeños**.

4.2.1. Intermediación de la red

En la Figura 8 se muestra la **intermediación de la red**. Esto es la frecuencia de aparición de nodos que conectan caminos entre otros nodos de la red (son intermediarios, como bien dice la palabra).

4.2.2. Distribución de cercanía de la red

En la Figura 9 se muestra la **distribución de cercanía de la red**, que es la cercanía de que presentan los nodos en una determinada zona de la red. Se pueden ver las zonas de la red que presentan un número mayor de nodos (más interacción entre los usuarios de Twitter).

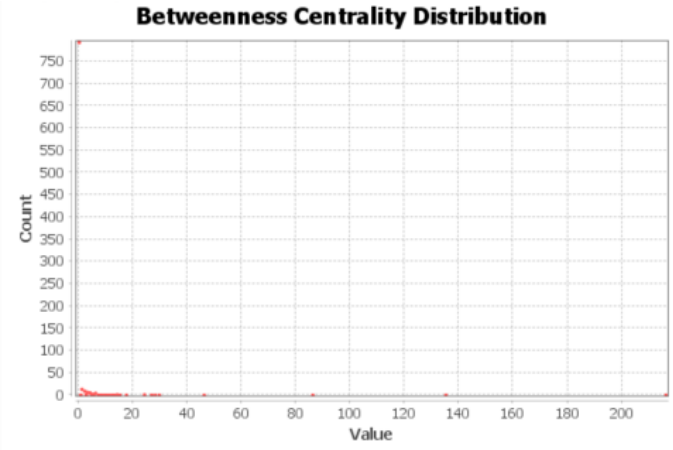


Figura 8: Intermediación de la red

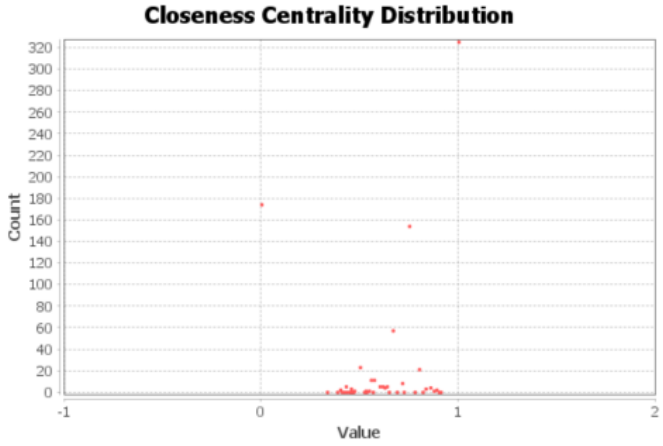


Figura 9: Distribución de cercanía de la red

4.2.3. Excentricidad de la red

En la Figura 10 se muestra la **excentricidad de la red**. La excentricidad mide la distancia que hay entre un nodo y el nodo más alejado de este. Cuando se ven nodos con un valor de 0, lo que significa realmente es que es imposible llegar a esos nodos.

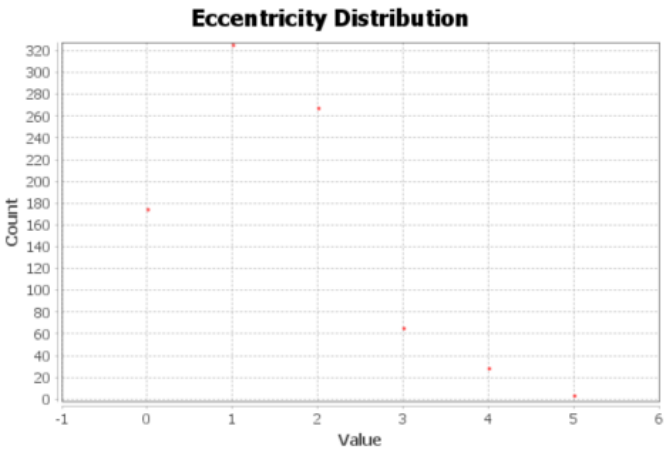


Figura 10: Excentricidad de la red

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

A lo largo de esta sección se van a estudiar los usuarios principales durante la emisión de la segunda semifinal de Eurovisión 2021 teniendo en cuenta las medidas de grado, intermediación, cercanía y vector propio.

5.1. Grado medio

En la Figura 11 se muestran los **15 primeros usuarios** que tienen un **mayor grado medio**.

Label	followers_count	Grado
@eurovision	489750	243
@thisismaneskin	96231	153
@itsmalbert	124169	89
@commonescgirl	17968	89
@stevievanzandt	244143	79
@ilcuoredilara	4010	76
@dadimakesmusic	29220	74
@bbceurovision	219911	72
@scottygb	57944	50
@yleniaindenial	27557	47
@alber_ab_94	385	45
@sergiano_93	475	44
@huertdeauteuil	20666	40
@euromovidas	4396	38
@renyrenner	7675	36

Figura 11: Usuarios principales con la medida de grado medio

5.2. Grado de entrada

En la Figura 12 se muestran los **15 primeros usuarios** que tienen un **mayor grado de entrada**.

Label	followers_count	Grado de entrada
@eurovision	489750	242
@thisismaneskin	96231	152
@itsmalbert	124169	89
@commonescgirl	17968	89
@stevievanzandt	244143	78
@ilcuoredilara	4010	76
@bbceurovision	219911	72
@dadimakesmusic	29220	72
@scottygb	57944	50
@yleniaindenial	27557	47
@alber_ab_94	385	45
@sergiano_93	475	44
@huertdeauteuil	20666	40
@euromovidas	4396	38
@renyrenner	7675	36

Figura 12: Usuarios principales con la medida de grado de entrada

5.3. Grado de salida

En la Figura 13 se muestran los **15 primeros usuarios** que tienen un **mayor grado de salida**.

Label	followers_count	Grado de salida
@littlemixftlodo	9263	22
@mar91rock	688	20
@escgreece	5232	18
@ryuuneedscoffee	233	16
@yvesaur1	58	16
@icouldbemorex	545	14
@kissed__byfire	1697	14
@narancial0vebot	1391	14
@yuthepooh4a	1075	14
@nacho_uncharted	266	14
@grdvrna	2911	14
@tops_uvillapun	67	14
@mfjenks	1306	14
@park_soomin5	298	14
@elyrian_xiii	60	14

Figura 13: Usuarios principales con la medida de grado de salida

5.4. Intermediación

En la Figura 14 se muestran los **15 primeros usuarios** que tienen una **mayor intermediación**.

Label	followers_count	Betweenness Centrality
@eurovision	489750	216.0
@dadimakesmusic	29220	135.0
@stevievanzandt	244143	86.0
@eskaplzm	53	46.0
@alemosiaa	109	29.5
@blascanto	227456	28.0
@eurovision_tve	94781	26.666667
@thisismaneskin	96231	24.0
@twidade_iucks	23921	24.0
@escgreece	5232	17.333333
@helenaaaaadead	1909	15.0
@cristhianblue99	1677	14.0
@piekielneskoki	213	14.0
@mecagoenlapxta	4751	13.0
@namuplanet	1730	12.666667

Figura 14: Usuarios principales con la medida de intermediación

5.5. Cercanía

En la Figura 15 se muestran los **15 primeros usuarios** que tienen una **mayor cercanía**.

Label	followers_count	Closeness Centrality
@eurovision	489750	1.0
@stevievanzandt	244143	1.0
@alemosiaa	109	1.0
@mecagoenlapxta	4751	1.0
@lalachus3	55961	1.0
@hyi099	3420	1.0
@limel87	151	1.0
@renkounofficial	485	1.0
@arianatorfrompl	430	1.0
@snjegurochka	567	1.0
@lazarevjungkook	4479	1.0
@pavlova2828	4033	1.0
@eurovisn_turkey	12320	1.0
@mauros_ranger	703	1.0
@rtve	1202020	1.0

Figura 15: Usuarios principales con la medida de cercanía

5.6. Centralidad de vector propio

En la Figura 16 se muestran los **15 primeros usuarios** que tienen una mayor **centralidad de vector propio**.

Label	followers_count	Eigenvector Centrality
@eurovision	489750	1.0
@thisismaneskin	96231	0.482645
@stevievanzandt	244143	0.424305
@senhitofficial	6262	0.327004
@dadimakesmusic	29220	0.312391
@itsmalbert	124169	0.292152
@commonescgirl	17968	0.288889
@ilcuoredilara	4010	0.240837
@bbceurovision	219911	0.239497
@alber_ab_94	385	0.154669
@scottygb	57944	0.152806
@yleniaindenial	27557	0.151364
@vika71rus	13342	0.135396
@sergiano_93	475	0.134469
@renyrenner	7675	0.13145

Figura 16: Usuarios principales con la medida de centralidad de vector propio

5.7. Análisis de las medidas

Para contestar a las preguntas planteadas, se obtienen las siguientes conclusiones acerca de estas medidas:

- La **medida de centralidad** no sirve en ningún caso porque hay muchísimos usuarios con una centralidad de 1, con lo cual no nos permite distinguir entre tipos de usuarios.
- La **medida de grado de salida** no sirve en ningún caso porque un grado mayor de salida quiere decir que ese usuario de Twitter es el que hace retweets, menciona y cita otras cuentas, pero no implica que su cuenta sea de interés para otros usuarios.
- La **medida de grado medio y de entrada** nos sirven para identificar cuáles son las cuentas que más retweets, respuestas y citas reciben por parte de los usuarios. Esta medida sí es de gran utilidad.
- La **medida de intermediación** es muy interesante para identificar las cuentas que deberían seguir las personas que estén interesadas en descubrir el festival de Eurovisión, ya que representan a los usuarios que conectan más grupos de nodos.
- La **medida de centralidad de vector propio** permite valorar la calidad de las aristas antes que la cantidad. Puede resultar más interesante que la medida de grado medio y de entrada.

6. Descubrimiento de comunidades en la red

6.1. Método de Lovaina

Este método es el que viene por defecto en Gephi. Como valor del parámetro de **resolución** se ha puesto un **5**, y con ello se ha obtenido una **modularidad de 0.661** y un total de **7 comunidades**, las cuales se van a presentar ahora.

6.1.1. Comunidad 1

En la Figura 17 se muestra la primera comunidad. Esta comunidad está compuesta por los *eurofanes españoles*, ya que las cuentas con un mayor grado son las de **@itsmalbert**, **@alber_ab_94**, **@sergio_93** y **@euromovidas**. Estas cuentas son muy famosas en el mundo eurovisivo español.

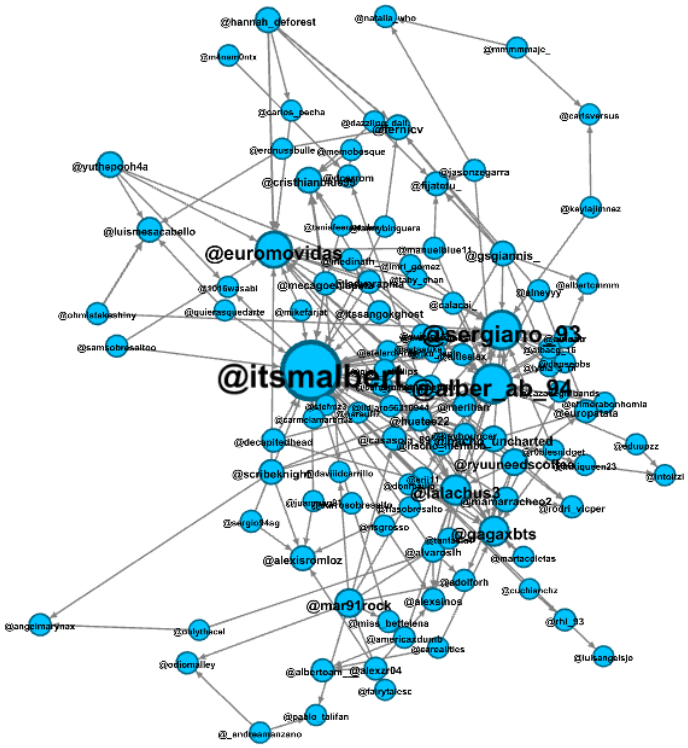


Figura 17: Comunidad 1: eurofanes españoles

6.1.2. Comunidad 2

En la Figura 18 se muestra la segunda comunidad. Esta comunidad se puede decir que está compuesta por los medios de comunicación y difusión de Eurovisión. La cuenta más destacada es la de **@eurovision**, que es la cuenta oficial del festival. También podemos encontrar la cuenta de **@bbceurovision** que es la cadena organizadora de Reino Unido en el festival de Eurovisión. La cuenta de **@commonescgirl** es una cuenta que normalmente hace difusión de todas las cosas que pasan en el festival.

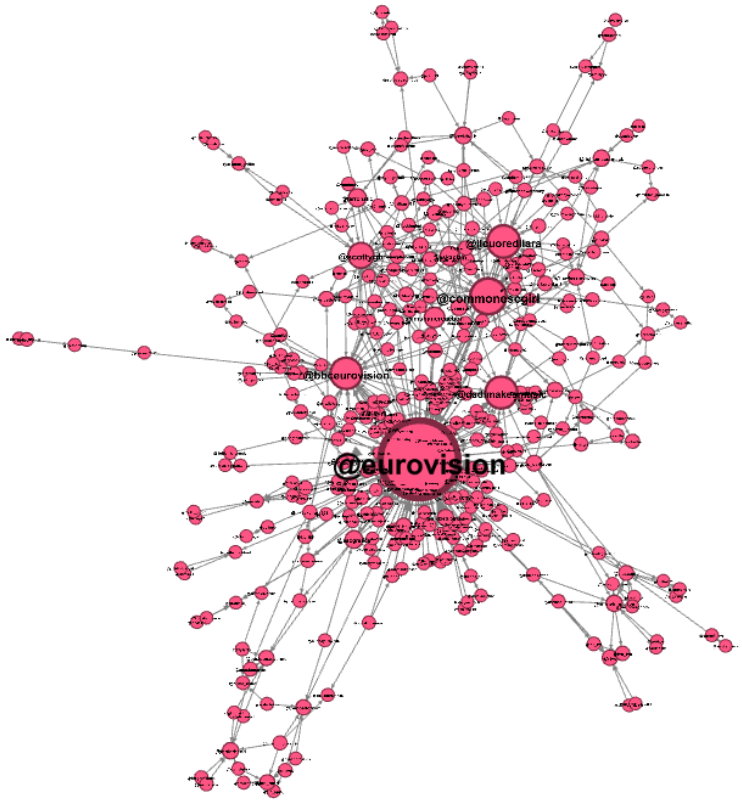


Figura 18: Comunidad 2: medios de comunicación nacionales de Eurovisión

6.1.3. Comunidad 3

En la Figura 19 se muestra la tercera comunidad. La cuenta principal que más destaca por encima de todos es la de **thisismaneskin**, que es la cuenta de los representantes de Italia en el festival y que se han proclamado como **ganadores del festival**, aunque en el momento de esta captura del conjunto de datos aún no se sabía ese dato. También se encuentra la cuenta de **@stevievanzandt**, que es un músico estadounidense que ha apoyado al país italiano abiertamente.

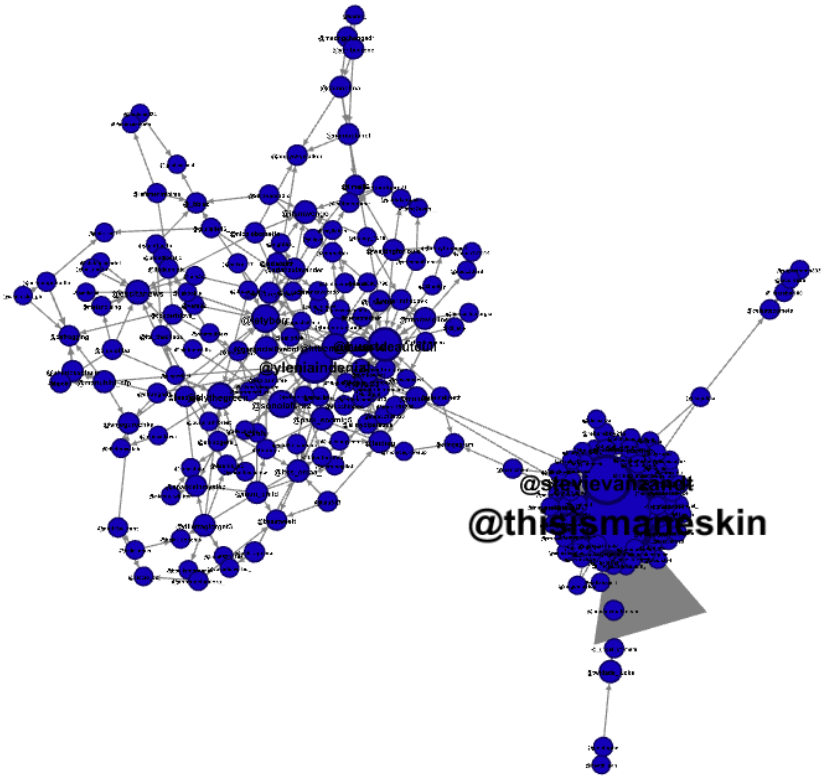


Figura 19: Comunidad 3

6.1.4. Comunidad 4

En la Figura 20 se muestra la cuarta comunidad. No hay cuentas destacables en esta comunidad, por lo que se podría decir que es una comunidad formada por gente a la que le gusta Eurovisión pero que no han tenido mucha relevancia en las interacciones.

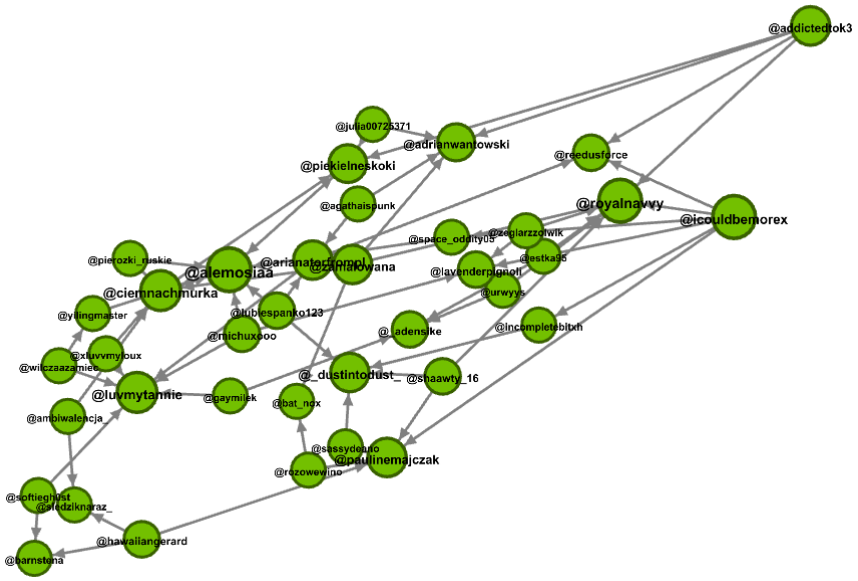


Figura 20: Comunidad 4

6.1.5. Comunidad 5 y 6

En la Figura 21 se muestran las comunidades 5 y 6. Se han agrupado estas comunidades porque no son muy relevantes debido a los pocos usuarios que las forman. Realmente debe percibirse como "errores" en el algoritmo de la modularidad, ya que están muy aislados.

(a) Comunidad 5

(b) Comunidad 6

Figura 21: Comunidades 5 y 6

6.1.6. Comunidad 7

En la Figura 22 se muestra la séptima y última comunidad. Los usuarios más destacados de esta comunidad son **@renyrenner** y **@vika71rus**, que son usuarios procedentes de Rusia. Esto quiere decir que probablemente esta comunidad esté formada por los eurofanes de dicho país y sus alrededores.

Figura 22: Comunidad 7

6.2. Método de Girvan-Newman Clustering

Se ha probado a utilizar este método para obtener la modularidad y comunidades. Después de un tiempo de ejecución de casi 5 minutos, resultado mostrado es el que se ve en la Figura 23. Se detectan **695 comunidades** y una modularidad de **0.88**. Este número de comunidades es demasiado grande, por lo cual se achaca a un error del algoritmo.

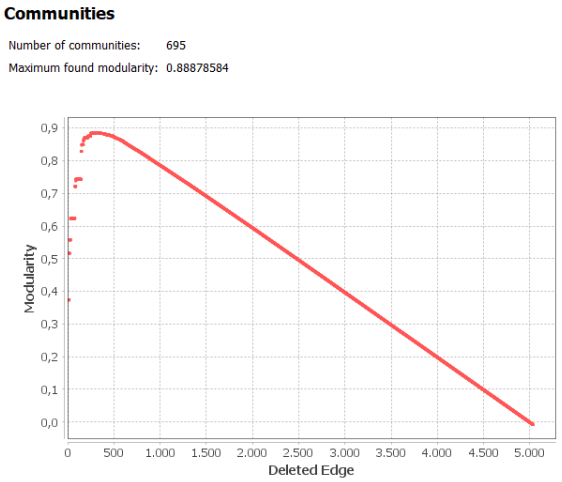


Figura 23: Detección de comunidades con Girvan-Newman Clustering

6.3. Método de Leiden

Se ha intentado ejecutar el **algoritmo de Leiden** para la detección de comunidades pero cada vez que se intetaba ejecutar, la aplicación de Gephi quedaba en un **estado de bloqueo**. Tras varios intentos, se ha decidido dejar esta opción.

7. Visualización de la red social

En la Figura 24 se muestra la visualización final de la red. En la sección 2.1 ya se explicó la visualización realizada. En esta imagen se ha añadido un ranking en los nodos más destacados. Además, se pueden intuir las comunidades que se han presentado en el apartado anterior.

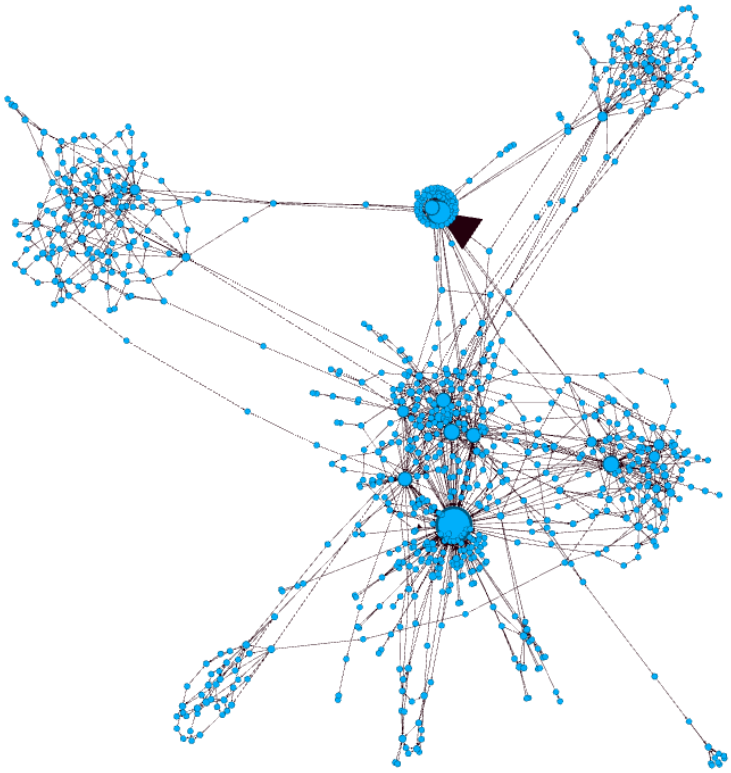


Figura 24: Visualización con Force Atlas 2

8. Discusión de los resultados obtenidos

Se va a dar contestación a las dos preguntas realizadas al principio de este documento.

8.1. ¿Qué usuarios de Twitter tuvieron un número mayor de interacciones durante la segunda semifinal de Eurovisión 2021?

Esta pregunta se ha respondido gracias a la **medida de centralidad de vector propio** y **medida de grado medio y de entrada**. Los usuarios que tenían los valores más grandes de estas medidas eran aquellas cuentas de Twitter que más interacciones habían tenido en cuanto a retweets, respuestas y citaciones.

De acuerdo a la tabla que se obtuvo con los usuarios más relevantes, se puede decir que esta medida **sí sirve para responder a la pregunta**, ya que se encuentran las siguientes cuentas:

- **@eurovision**: es la **cuenta oficial de Eurovisión**.
- **@thisismaneskin**: son los **representantes de Italia en Eurovisión 2021**.
- **@itsmalbert**: es un conocido **youtuber** español que comenta las galas de Eurovisión todos los años.
- **@commonescgirl**: es una cuenta muy conocida entre el público eurofan, ya que ofrece mucha información sobre Eurovisión todo los días del año.

Estas cuentas son muy importantes y está claro que son las que más interacciones reciben.

8.2. ¿Cuáles son las cuentas que se deben seguir si alguien está empezando a introducirse en el mundo de Eurovisión?

Esta pregunta se puede responder de dos maneras: las medidas y la detección de comunidades.

La **medida de intermediación** muestra los usuarios más intermediarios durante la segunda semifinal de Eurovisión. Estos usuarios son a los que la gente que está interesada en introducirse en el mundo de Eurovisión deberían seguir. También puede aceptarse gente que esté interesada en estar al tanto de todas las noticias que pasan a lo largo de la celebración de la gala.

El problema de la **medida de intermediación** es que puede haber usuarios muy intermediarios pero que procedan de otros países cuyo lenguaje no entendamos (por ejemplo, cuentas de Rusia). En la **detección de comunidades** ya se ha visto que la **comunidad 1** está formada por el público eurovisivo españo, con lo cual puede servir también como guía para seguir a cuentas relevantes y que además sean españolas.

En cuanto a las cuentas destacadas de acuerdo a la medida de intermediación se destacan dos cuentas españolas:

- **@blascanto**: es la cuenta de Twitter del **representante de España en Eurovisión 2021**, con lo cual es una muy buena cuenta para aquella gente que quiera introducirse en el mundo de Eurovisión este año.
- **@eurovision_tve**: es la cuenta de Twitter de la **organización española del festival**, también muy acertada para aquellos usuarios que tengan este fin.

9. Bibliografía

[1] Gephi 0.9.2. https://gephi.org/, 2017.

[2] Twitter streaming importer. https://github.com/seinecle/gephi-tutorials/blob/master/src/main/asciidoc/en/plugins/twitter-streaming-importer-en.adoc.