UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

A Study of Virality on Social Networks

Sergio Marino Congosto
Tutor: Simone Santini

Junio 2018

A Study of Virality on Social Networks

AUTOR: Sergio Marino Congosto
TUTOR: Simone Santini

Dpto. Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio de 2018

Resumen (castellano)

Este Trabajo de Fin de Grado se basa en la realización de un estudio del fenómeno denominado ‘viralidad’ usando como caso de estudio unos tweets pertenecientes a nodos/usuarios de la red social Twitter.

En el contexto de nuestro estudio un tweet es “viral” si obtiene muchos retweets en poco tiempo. En este trabajo se intenta averiguar la presencia (o ausencia) de una relación entre la viralidad de un tweet y varias características locales del grafo de las relaciones entre usuarios en que se origina: número de vecinos, coeficiente de clustering, miembros de la comunidad y comunidades adyacentes; todas ellas respecto a un nodo. Entendiéndose como nodo a cada uno de los participantes de la red social en la que se ha basado el estudio.

Para la realización de este estudio ha sido necesario conocer la cantidad de retweets correspondientes a cada nodo. Después se ha relacionado con la cantidad de vecinos a los que ese nodo sigue. Acto seguido se ha calculado el coeficiente de clustering de cada nodo respecto a los nodos vecinos. Luego se ha recurrido al algoritmo japonés Kamada Kawai para poder dibujar, separar y clasificar a los nodos en comunidades. Cada nodo pertenece a una comunidad, de lo que ahí se ha podido deducir el número de miembros de cada comunidad y se ha podido contar el número de comunidades distintas a los que los vecinos de ese nodo pertenecen.

Una vez se han calculado esas variables se ha pasado a la normalización de estos resultados para su posterior comparación con los retweets. A partir de ahí se ha llegado a la conclusión de si esas variables tienen una relación directa, inversa o simplemente no existe ninguna relación entre esas variables y la viralidad.

Abstract (English)

This Bachelor Thesis is based on the carrying out of a study about the famous phenomenom called ‘virality’ using as testbet some tweets belonged to nodes/users in the social network Twitter.

The virality concept is understood in a context where a tweet gets a bunch of retweets in a short period of time. In order to do that a small research with respect to the virality related with several variables: number of neighbors, clustering coefficient, members of the community and adjacent communities; all of them with respect to one node. Understanding a node is each of the participants in the social network in which the study has been based on.

The amount of retweets was one of the indispensable and essential data to carry out this research. After it got related to the number of neighbors that the node in particular follows. Immediately thereafter the coefficient of clustering has been calculated for each node with respect to all its neighbors. Later the japonese algorithm Kamada Kawai was used to draw, divide and classify all the nodes into communities. Each node belongs to one community, from which the number of participants of each community could be calculated and so the number of different communities that the neighbors of the node belongs to.

Once that variables have been calculated it could be possible to restore this outcomes for its later comparison with the retweets. From that point it got reached the conclusion of

which variables either have or not a direct, inverse or just simply not relation between that variables and the virality phenomenon.

Palabras clave (castellano)

Viralidad, nodos, grafo, tuit, retuit, algoritmo, nodo, aristas, vecinos, comunidades, coeficiente de agrupamiento, tfg, librería.

Keywords (inglés)

Tweet, retweet, clustering coefficient, Python, Pearson, Twitter.

Agradecimientos

Me gustaría pararme un momento en este apartado para poder dar las gracias a mucha gente y agradecerle todo lo que han hecho por mí.

Empezando de una manera más personal me gustaría decir gracias a mis padres, abuelos y hermano por haberme inculcado el hábito de estudiar y esforzarse en la vida. De ahí pasando a un plano económico les doy las gracias por haberme pagado la carrera, que no es algo barato ni que todo el mundo se pueda permitir.

Me gustaría también dar gracias a todos los profesores que pasaron por mí. Desde el colegio, instituto y por supuesto Universidad. Tantos profesores, tanto de prácticas como de teoría, que aportan su granito de arena y han hecho que hoy en día haya adquirido los conocimientos que tengo y he podido desarrollar.

Gracias también al Estado y Gobierno de España por permitirme formarme y tener un futuro aquí en España.

Este tfg no hubiera podido ser realizado sin la ayuda de mi tutor que me ha ido guiando, planificando y resolviendo dudas que iban surgiendo por el camino, así que gracias por todo Simone.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

1 Introducción

1.1 Motivación

La viralidad es uno de los fenómenos más evidentes en el panorama de las redes sociales. No se trata de un fenómeno nuevo (su nombre se basa en una metáfora de la difusión de enfermedades, un fenómeno claramente muy antiguo). Tampoco la viralidad es algo nuevo en el campo de la información (Douglas Adams escribió: Nada viaja más rápido que la luz, con la posible excepción de las malas noticias), pero las redes sociales han dado al fenómeno una nueva relevancia y un peso en la difusión de información que antes no tenía. La viralidad se podría definir, si es que hay una definición correcta, como algo que crece, se propaga y duplica de manera rápida en un período de tiempo reducido. Se podría comparar a un virus, el cual crece y se va propagando; siendo su misión duplicarse y crecer cada vez más para hacerse mayor.

Twitter es una red social que puede ser representada como un grafo compuesto por nodos. Estos nodos tienen la capacidad de publicar tweets, que pueden ser entendidos como textos planos, que los nodos tienen la posibilidad de compartir con otros nodos que les siguen, denominándose este hecho de poder compartir tweets entre sí: 'retweet'. Twitter es diferente a otras redes sociales, Twitter permite que un nodo pueda seguir a otro sin que éste le siga de vuelta o viceversa o ambos nodos se pueden seguir. Tiene un concepto de seguimiento que es totalmente distinto a otras redes como Facebook en los que ambos usuarios tienen que ser amigos y aceptar esa amistad. Además, Twitter tiene su especial interacción de retweet. En otras redes sociales los usuarios pueden compartir contenido y comentar, gustar etc. Twitter, aparte de aportar ya esas características previas, permite a los nodos hacer retweet, permite que los nodos que les siguen puedan visualizar el contenido que ese nodo está compartiendo/ 'retuiteando' con ellos. Esto no es posible en redes como Facebook, aunque de una manera posterior Facebook haya incluido un botón de 'Share' para hacer competencia a Twitter.

En este TFG se llevará a cabo un estudio de la estructura de las redes sociales para intentar entender si existe una relación entre la estructura local de red alrededor de un nodo (usuario) y la probabilidad que su información se transforme en viral. Para ello se buscarán ciertas características que cumplen estos nodos que los haga especiales y haga que sus tweets se conviertan en virales.

Durante este estudio se medirán ciertas características de las redes sociales y del entorno de un nodo para determinar si tienen o no relación con la viralidad de los mensajes producidos por esos nodos. La novedad de este TFG es que este estudio no ha sido hecho de manera previa, se trata de una comparación de variables que se mencionarán más delante con la cantidad de retweets que un tweet contenga.

El fenómeno de viralidad tiene que ser entendido como algo muy novedoso y no estrictamente definido. El principal objetivo de este estudio es entender a los nodos que originan ese contenido en forma de tweet, entender qué es lo que hace que esos nodos tengan el poder y la capacidad de llegar a tanta gente, que muchos nodos quieran compartir el contenido de ese tweet.

Los nodos que originan ese contenido inicial generan la cualidad que sus tweets reciban muchos retweets en poco tiempo. Hay varios factores que hacen que tengan tanto éxito. El cometido de este estudio es encontrar esos factores y hacer una comparación con la cantidad de retweets que poseen esos tweets.

1.2 Objetivos

Cada nodo es cada componente del grafo que se puede dibujar si se pintara la red social de Twitter. Cada nodo dispone de ciertas características que lo hacen único y distinto al resto de los nodos. En particular este estudio se ha centrado en las siguientes características: el número de nodos de cada nodo, el coeficiente de clustering de cada nodo, el número de miembros participantes a la que el nodo pertenece y el número de comunidades distintas a las que los vecinos de ese nodo pertenecen.

A la hora de dividir a los nodos en comunidades se ha utilizado el algoritmo Kamada Kawai basado en el principio de que cuantas más conexiones en común tienen dos nodos, más cerca y más probabilidad tienen esos nodos de pertenecer a la misma comunidad.

Por otro lado, se ha calculado otra variable, la cantidad de retweets de todos los tweets que se tienen como muestra para este estudio. En el momento en que un tweet obtenga muchos retweets en poco tiempo, ese tweet pasará a convertirse en viral.

En el momento que ya se han calculado esas variables la idea principal y el principal objetivo es poder relacionarlas mediante correlación. En el estudio se determinará si hay correlación directa, inversa o no hay correlación (es decir, hay independencia estadística) entre el número de vecinos de cada nodo, su coeficiente de agrupamiento, su número de participantes en esa comunidad, y su número de comunidades adyacentes con la cantidad de retweets de los tweets pertenecientes a ese nodo.

1.3 Organización de la memoria

La memoria de este TFG está compuesta por los siguientes capítulos:

- **Introducción**: se trata la motivación, los objetivos y la organización de la memoria.
- **Estado del Arte**: se explican las teorías de los distintos modelos de grafos ER, BA y el algoritmo de agrupación de comunidades Kamada Kawai.
- **Características de Twitter**: se hace referencia a las principales características de la red social Twitter.
- **Medidas**: en este apartado se abordan las distintas medidas que se van a obtener durante este estudio. Consta de la relación que existe entre los nodos y los vecinos, los nodos y los seguidores, los nodos y el coeficiente de agrupamiento, y los nodos y sus respectivas comunidades.
- **Resultados**: se muestran los resultados obtenidos de las medidas previamente tomadas.
- **Conclusiones y Trabajo Futuro**: se desarrollan las conclusiones que se han obtenido durante el estudio y se habla acerca del trabajo que se pudiera realizar en el futuro respecto a este estudio.
- **Referencias:** se citan las fuentes que se han usado para este TFG.
- **Glosario**: se muestran las palabras relevantes y no tan triviales con su respectiva definición.

2 Estado del arte

2.1 Introducción

Ya en los años noventa del siglo XX, en el momento en que la web salió de las universidades para transformarse en un fenómeno social, se observó que, además de su contenido, la estructura de sus enlaces podía proporcionar mucha información. Uno de los primeros resultados de esta observación es el conocido algoritmo PageRank, [4] (algoritmo que utiliza Google como motor de búsqueda creado por Larry Page y Sergey Brin) que usa el estado de equilibrio de un camino aleatorio en el grafo que representa la red para determinar la relevancia de las páginas web.

Desde entonces el interés para estudiar la estructura de grafo de la web primero y de las redes sociales después ha ido aumentando. Dos aspectos de esta investigación nos interesan destacar: por un lado, se han desarrollado medidas sobre grafos que intentan expresar de forma sintética las características de interés para el estudio de las redes sociales; por el otro, se han estudiado modelos de grafos aleatorios que reprodujeron, en la medida de lo posible, estas características, con el fin de disponer de un 'data set' significativo y parametrizado.

En este capítulo se presentarán brevemente algunos de estos trabajos, los que se consideran más relevantes para el trabajo que se desarrolla aquí.

Se considerarán primero varias medidas sobre grafos: la distribución del número de vecinos, el coeficiente de clustering y la detección de comunidades.

Luego, se considerarán los modelos más comunes de grafos aleatorios, y la manera en que sus características (determinadas por las medidas que ya se habrán presentado) reflejan las de las redes sociales reales.

2.2 Medidas

Las medidas sobre los grafos relevantes para este trabajo son las siguientes:

- **Distribución de vecinos**. Se miden las conexiones de los distintos nodos y se calcula un histograma que nos da la probabilidad que un nodo tenga 'n' vecinos.

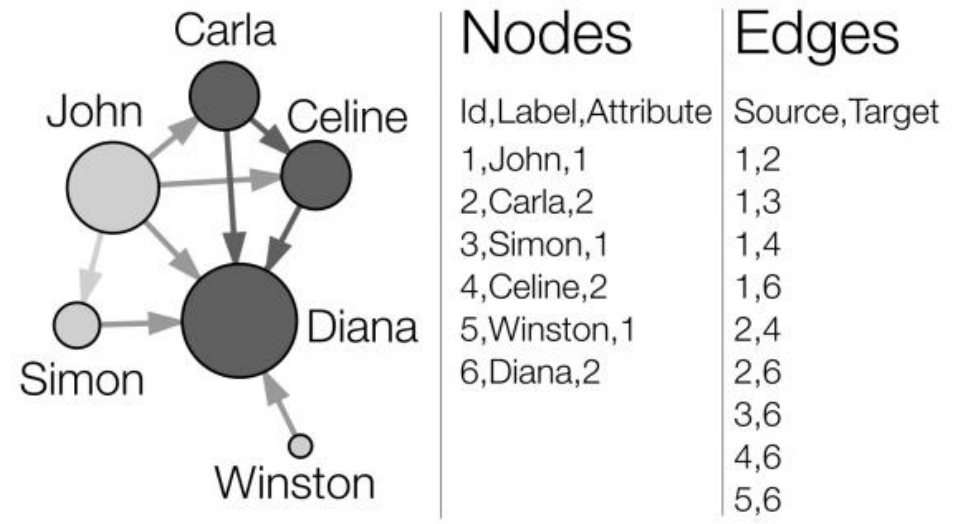

FIGURA 1: Distribución de vecinos

- **Coeficiente de clustering**. Se trata de una medida que evalúa en los nodos el grado en que éstos tienden a agruparse, lo que en teoría de grafos se conoce como coeficiente de agrupamiento (clustering coefficient).

  En las redes de la vida cotidiana y más aún en las redes sociales, que es en dónde se ha realizado este estudio, los nodos tienden a unirse y agruparse formando distintos grupos. Un ejemplo podría ser los distintos grupos políticos que se crean.

  Este coeficiente es calculado de manera semejante, aunque no exactamente igual, para los grafos dirigidos[1] y no dirigidos [7].

  Para el caso de los grafos no dirigidos, el coeficiente de agrupamiento es calculado, para un nodo n:

  $C(n) = |\{(n, u), (u, 2), (2, n) \notin E^3\}| / |\{n, u\} \in E|$

  para el grafo $C = \sum_{n \in V} C(n) * 1 / |V|$

A continuación, se muestra un ejemplo explicativo:

[1] El coeficiente de agrupamiento para los grafos no dirigidos no aplica dado que no ha sido usado en este estudio, por lo que no se detalla.

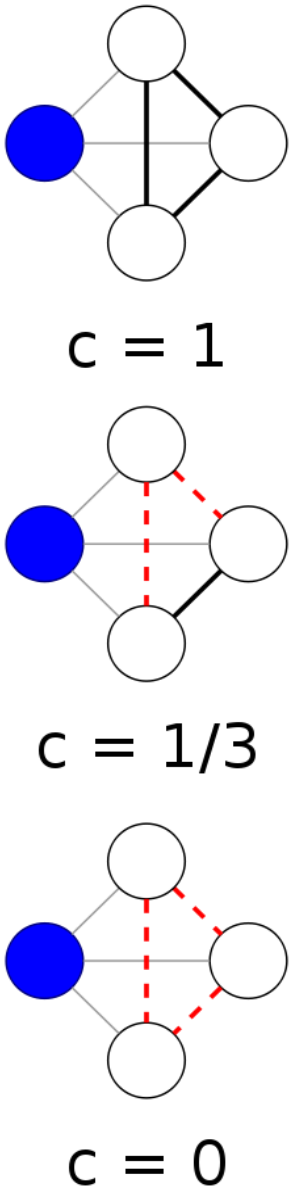

FIGURA 2: Coeficiente de agrupamiento

En este primer caso el nodo azul tiene tres vecinos y se están formando tres triángulos cerrados: triángulo formado por el nodo de la izquierda, nodo superior y nodo derecho; triángulo formado por el nodo de la izquierda, nodo inferior y nodo derecho; y por último el triángulo formado por el nodo de la izquierda, nodo superior y nodo inferior. Por lo que dividimos 3/3 = 1.

En el segundo caso hay también tres vecinos y solo se forma un triángulo debido al nodo izquierdo con el nodo derecho y el nodo inferior. Por lo tanto 1/3 = 0.33.

En el tercer caso sigue habiendo tres vecinos, pero no existe ningún triángulo cerrado, por lo que 0/3 = 0.

- **Comunidad**. Una comunidad es la agrupación de varios nodos caracterizada por tener algo en común. Hay distintos algoritmos que agrupan a los nodos en comunidades dependiendo de distintos criterios [8].

Los miembros o nodos participantes en una red tienden a organizarse y agruparse en comunidades. Por lo que las comunidades son grupos de miembros, ya sea por gustos musicales, aficiones, ideas políticas etc.

El siguiente paso de nuestro estudio fue intentar dividir el grupo de nodos en diversas comunidades.

Un algoritmo bastante conocido para la división en comunidades es el K-means [10]. K-means va colocando 'centroids' que se van moviendo a lo largo del algoritmo hasta estar cada vez más en el centro. K-means es un algoritmo basado en distancias, siendo una de las más populares la 'distancia Euclidea'. En nuestro caso, al no poder calcular distancias entre los nodos, este algoritmo es inútil dado que no puede ser aplicado.

El algoritmo Kamada-Kawai [13] es un algoritmo que posiciona los nodos en un espacio bidimensional ó tridimensional asignándoles la misma longitud a cada arista. Después, a medida que avanza el algoritmo basado en la Ley de Hooke, los nodos que menos enlaces tienen con otros nodos tienden a separarse y los nodos que más conectados están tienden a juntarse. Esta separación o acercamiento de nodos viene dada por unas fuerzas que el algoritmo asigna a las aristas, en función como ya se ha dicho antes, de su menor o mayor conexión con más nodos.

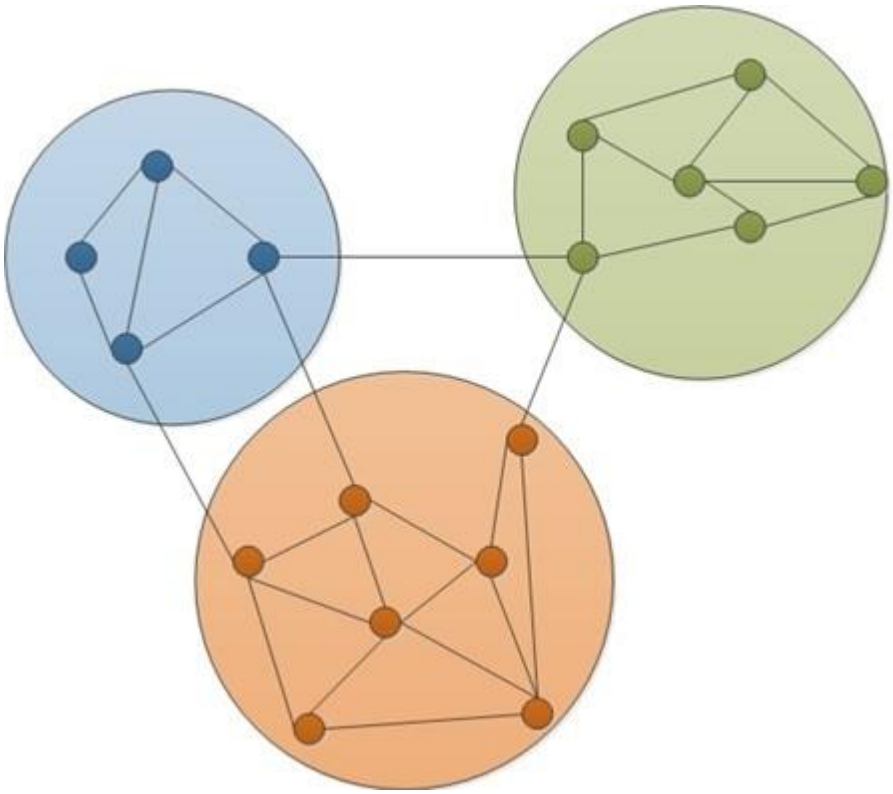

FIGURA 3: Comunidades

2.3 Grafos aleatorios

Un grafo aleatorio es un conjunto de nodos cuyas conexiones se han generado de manera aleatoria, según cierta distribución de probabilidad que determina las características del grafo.

Un grafo aleatorio es una representación de una red compleja, compuesta de diversos participantes. Los nodos del grafo representan los elementos de la red (por ejemplo, páginas web o personas) y los arcos representan conexiones cuya naturaleza depende del tipo de red (por ejemplo, enlaces entre páginas web, relaciones de amistad o de seguimiento para redes sociales, etc.).

Las diferentes redes resultan en modelos diferentes. Por ejemplo, los enlaces web y la red de Twitter se modelan como grafos dirigidos, mientras que Facebook se modela mediante un grafo no dirigido, dado que las relaciones de amistad son simétricas.

2.4 Modelo de grafos ER

El modelo de Erdős-Rényi [14] es un conjunto de nodos (n) a los que se van sumando una serie de aristas (a). Las aristas son producidas con una probabilidad dada.

El grafo aleatorio de Erdős-Rényi, se construye mediante un proceso que empieza con n vértices desconectados a los que se van añadiendo vértices. En realidad existen dos opciones muy relacionadas con el modelo gráfico de Erdős-Rényi (ER):

- Grafo uniformemente al azar con n nodos y m aristas G(n,m).
- Grafo creado con nodos aleatoriamente donde cada gráfico la misma probabilidad siempre y cuando contengan el mismo número de nodos (n) y aristas (a).

En 1960 se describió el comportamiento de un grafo G(n,a) para distintos posibles valores de a aristas, obteniendo diversos resultados tales como:

- Si na < 1, no habrá en un grafo nodos conectados entre sí mayores que O (log (n)).
- Si na = 1, habrá un grafo con un nodo de tamaño del orden $n^{2/3}$.
- Si na → k, donde k es una constante, habrá un gráfico con un componente gigante único con una fracción positiva de los vértices.
- Si $a < \frac{(1-\in)\ln n}{n}$ habrá un gráfico que tendrá vértices aislados.
- Si $a > \frac{(1-\in)\ln n}{n}$ habrá un gráfico conexo.

2.5 Modelo de grafos BA

El algoritmo de Erdős-Rényi fue uno de los primeros que se estudiaron y constituye el punto de partida de una prolífica serie de trabajos.

A pesar de su importancia, no constituye un buen modelo de las redes sociales, sobre todo a causa de su distribución del número de vecinos de un nodo.

Sea $n \in \mathbb{N}$ un número posible de vecinos, y β la fracción de nodos que tiene ese número de vecinos. En el caso de las redes sociales, se ha verificado experimentalmente que estas dos magnitudes siguen una relación del tipo **power law**: $\beta = n^r$ (para $r < 0$). Esta distribución, también conocida como 'scale-free' o libre de escala, supone para una red social que pocos usuarios tienen muchos seguidores y que existe una 'long tail' o cola larga de usuarios, cada vez en mayor número que tiene progresivamente menos seguidores. El parámetro r, cuyo valor depende de la red que se está considerando, regula este descenso.

Un modelo de grafo que exhibe este comportamiento fue propuesto por Barabási-Albert [15]. Este algoritmo permite crear redes aleatorias de 'scale-free' (donde existen nodos muy conectados, mientras que la mayoría de nodos tiene un grado de conexión bajo) utilizando un mecanismo llamado 'preferential attachment'.

El algoritmo funciona de la siguiente manera. Se comienza con una red de m_0 nodos conectados, a la que se van añadiendo nodos. Cada nodo nuevo es conectado preferiblemente por nodos que tienen un mayor grado de conexión. La probabilidad p_i con la que un nuevo nodo se conecte con un nodo 'i' es:

$p_i = k_i / \sum k_j$ (donde k_i es el grado del nodo i).

Es decir, se ve que los nodos se conectarán con más probabilidad con nodos que tienen un alto grado de conexión. Este fenómeno se conoce como 'preferential attachment'.

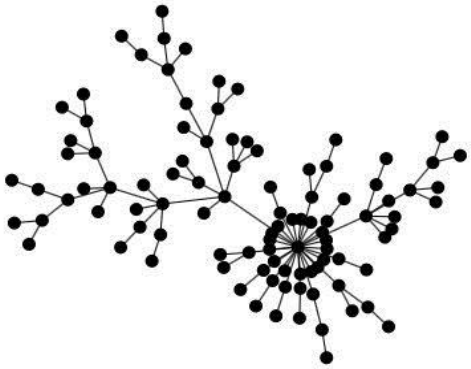

FIGURA 4: Grafo conectado con preferential attachment

2.6 Algoritmo Kamada Kawai

El algoritmo Kamada-Kawai [13] no es, estrictamente hablando, un algoritmo de generación de grafos. Se trata de un algoritmo que, dado un grafo, dispone sus nodos en una espacio de manera tal que queden evidenciadas las características del mismo.

El algoritmo Kamada-Kawai es una variante de los algoritmos llamados "dirigidos por fuerza", es decir, el objetivo es definir la función objetivo a minimizar de tal forma que asigne a cada plano del grafo un número $x \in \mathbb{R}^+$ que representa la energía.

Kamada-Kawai introduce un sistema dinámico, en el que n partículas $p_1, p_2, \ldots, p_n$; (n = |V|) están conectadas en un plano. Cada partícula representa un vértice de los conjuntos de vértices del grafo $v_1, v_2, \ldots, v_n$; (n = |V|). El sistema dinámico actúa de tal forma que los vértices que tienen una arista entre dos están conectados mediante muelles con constantes elásticas.

Sea $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ las coordenadas de las partículas en el plano. La energía del sistema elástico se puede representar mediante la siguiente fórmula:

$$E = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{2} k_{ij} \{ (x_i - x_j)^2 + (y_i - y_j)^2 + l_{ij}^2 - 2 l_{ij} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \}$$

Donde:

$k_{ij} = \frac{K}{d_{ij}^2}$

$d_{ij} =$ Longitud del camino más corto entre $v_i \, y \, v_j$

$K =$ Constante elasticidad

$l_{ij} = L * d_{ij}$

$L =$ Longitud deseable de una arista

El propósito es hallar los valores de x_i, y_i que minimicen la función $E(x_1, .., x_n, y_1, \dots, y_n)$, que se corresponde con la condición de que las derivadas parciales $\frac{\partial E}{\partial x_i}, \frac{\partial E}{\partial y_i}$ sean iguales a cero. Dado que es muy complicado hallar dicho mínimo, la solución puede ser hallada mediante métodos numéricos tal como Newton-Raphson [16].

Este algoritmo presenta las siguientes características:

- **Flexibilidad**. Los algoritmos de fuerza dirigida se adaptan y extienden fácilmente cumpliendo así los criterios adicionales de estética.
- **Interactividad**. Si se dibujan los grafos que van resultando en estados intermedios se puede ver cómo evoluciona el grafo, lo que ayuda a su comprensión y funcionamiento.
- **Intuitivo**. Debido a su afinidad física a los muelles, el algoritmo es fácil de predecir y entender.
- **Simplicidad**. Los algoritmos de fuerza dirigida son fáciles de implementar, teniendo que escribir pocas líneas de código. En este tfg se utilizó una librería de Python, por lo que no fue excesivamente complicado.
- **Fundamentos teóricos.** La convergencia monótona, la cual dicta que el coste del diseño se reduce en cada iteración, garantiza que al final se llegará a un mínimo local y el algoritmo concluirá.

3 Características de Twitter

El agrupamiento de individuos ya sea por diversas aficiones, gustos en común, trabajo, círculos de amigos, ideas políticas etc. en comunidades en Internet es el fenómeno denominado redes sociales. Las redes sociales permiten la interacción entre sus usuarios: éstos pueden publicar fotos, realizar comentarios, marcar las fotos con "me gusta", intercambiar mensajes, etc.

A grandes rasgos las redes sociales se pueden clasificar en tres bloques:

- Genéricas: se trata de las redes sociales sin ninguna particularidad (son las más masivas y numerosas ya que se dirigen a todo tipo de personas). Ejemplos de este tipo de redes sociales son Facebook, Twitter, Instagram etc.
- Profesionales: este tipo de redes crea contactos entre personas basados en intereses profesionales comunes. Un ejemplo de estas redes es Linkedin.
- Temáticas: son aquellas que se caracterizan por tener alguna temática, hobbie, deporte, afición en particular. Un ejemplo es la conocida Flickr, cuya temática es la fotografía.

Las redes sociales surgieron en 1995 con la creación por parte de Randy Conrads de classmates.com [17], una red que buscaba reunir a excompañeros del colegio o trabajo.
Facebook se sitúa en el ranking mundial como la red más usada y con mayor número de usuarios. Esta red social fue creada por un grupo de estudiantes, capitaneados por Mark Zuckerberg, en 2004 [18] cuya finalidad era mantener el contacto entre los estudiantes de Harvard.

Otra red social muy conocida es Twitter. Twitter es una red social creada en 2006 por Jack Dorsey [19]. La red fue creada por Obvious, una compañía de San Francisco, como proyecto de investigación.

Twitter surge con la idea de dejar comunicarse a los usuarios en tan solo 140 caracteres. Un 'tweet', por lo tanto, es un mensaje limitado a 140 caracteres que los usuarios pueden mandar.

Esta idea de limitar los mensajes a 140 caracteres tiene su origen en los SMS. Los SMS están limitados a 160 caracteres debido al antiguo ancho de banda disponible. De estos 160 caracteres, Twitter usa 20 para el nombre de usuario, por lo que solo quedan 140 disponibles para escribir el mensaje.

Twitter es la red social que se usará como referencia en este trabajo. La elección de Twitter se basa en dos consideraciones. La primera es la multiplicidad de posibles relaciones entre personas: follow, retweet y like. Esto nos permite crear varios modelos basados en la misma red. La segunda consideración es la posibilidad de recoger datos: Twitter proporciona una API pública con la información necesaria para crear nuestros modelos.

Las características más relevantes de Twitter son:

I. Asimétrica. Twitter permite que un usuario siga a otro sin que esta relación tenga que ser necesariamente recíproca. Un usuario sigue a otro y este otro puede seguirle o no.
II. Breve y limitado. Los tweets están limitados a 140 caracteres.
III. Descentralizada. La arquitectura de la red social varía dependiendo de las decisiones de los usuarios.
IV. Global. Twitter está en varios países en múltiples idiomas.
V. Hipertextual. Cada tweet puede contener varios enlaces con los símbolos '@' y '#'.
VI. Intuitivo y simple. Twitter posee una interfaz gráfica muy fácil y sencilla, siendo muy intuitiva y no necesitando ningún tipo de conocimiento previo para usarlo.
VII. Multiplataforma. Esta aplicación es soportada por diferentes dispositivos, navegadores etc.
VIII. Viralidad. Permite la rápida circulación de los mensajes y la multiplicación de éstos.

Además, Twitter permite a sus usuarios realizar las siguientes acciones:

a) Mencionar a usuarios. Mediante el carácter '@' los usuarios pueden mencionar a otros. En ese caso los usuarios mencionados reciben una notificación y pueden ver el mensaje en el cual fueron mencionados.
b) Retweet. Una de las características más importantes. Permite compartir el mensaje retuiteado con todos los seguidores de ese usuario.
c) Citar. Permite realizar un comentario sobre el tweet citado en cuestión.
d) Like. Los usuarios pueden dar al botón 'like' para indicar que les gusta ese tweet.
e) Tweets fijos. Hace que se guarden los tweets seleccionados para no perderlos. De cierta manera es como marcar esos tweets como favoritos para no perderlos y tenerlos accesibles.
f) Mensajes privados. Permite a los usuarios enviar mensajes privados entre ellos, permitiendo a los usuarios la comunicación entre ellos.
g) Listas. Los usuarios pueden ser agrupados por listas dependiendo de temas o intereses.
h) News. Las noticias también son visibles, se pueden buscar por hashtags (#).
i) Bloquear usuarios. Los usuarios tienen la capacidad de bloquear usuarios para que esas personas no puedan comunicarse, mencionar ni recibir notificaciones de esos usuarios.
j) Follow. Los usuarios pueden seguir a otros usuarios para de esa manera ver los tweets publicados por ese usuario y toda su información.

4 Medidas

Se ha utilizado los datos proporcionados por Twitter sobre 9832 usuarios y 1 378 158 tweets para construir tres histogramas que representan las relaciones principales de la red social: *follow, retweet, like*. En estos grafos, y especialmente en el grafo de 'follow', se han efectuado las medidas introducidas en el capítulo 2 para determinar sus características.

4.1 Relación nodos – vecinos

Una de las principales características de este grafo, de la figura 5, es la cantidad de vecinos que posee cada nodo. Se adjunta un histograma donde se refleja el número de nodos que siguen al mismo número de nodos:

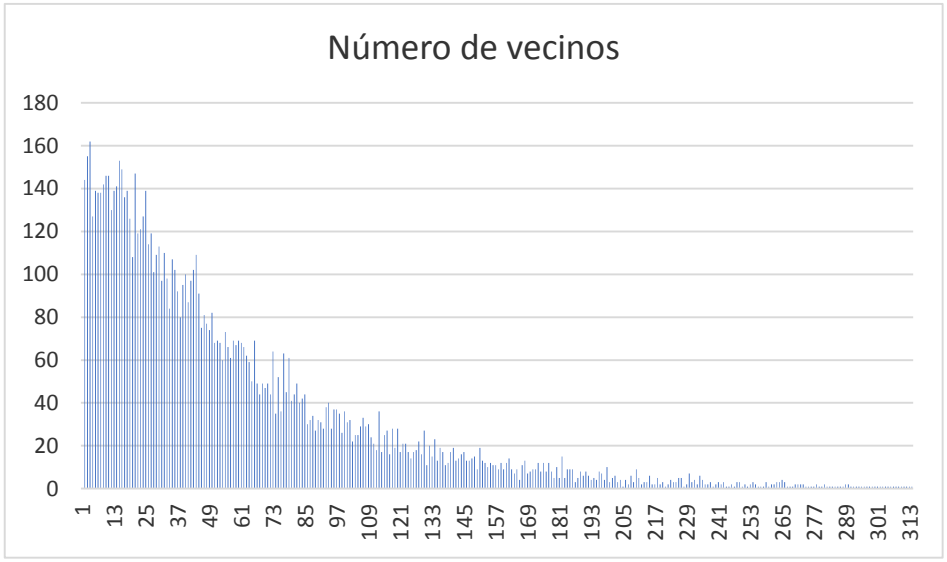

FIGURA 5: Relación nodos - vecinos

En la figura 5 se puede ver cómo abunda el número de nodos que no sigue a muchos usuarios, la mayoría de los nodos siguen a 1 – 100 nodos. Al mismo tiempo se puede ver que hay muy pocos nodos que sigan a muchos nodos.

El nodo que a más vecinos sigue está siguiendo a 313 nodos, pero es solo un nodo, por lo que no es lo común.

Por el contrario, hay 144 nodos que solo siguen a un nodo. Por lo que se ve que la tendencia de estos nodos es seguir a poca gente.

Al mismo tiempo llama la atención que no hay ningún nodo que no esté siguiendo a alguien; todos los usuarios siguen a algún nodo.

4.2 Histograma relación nodos – seguidores

Ahora se muestra la misma medida, pero vista de una manera opuesta, mostrándose así el número de nodos que son seguidos por un número dado de nodos:

FIGURA 6: Relación nodos - seguidores

En la figura 6 se muestra la gran cantidad de nodos que son seguidos muy pocas veces mientras que hay muy pocos nodos que son seguidos muchas veces.

Además hay 670 nodos que solo son seguidos una vez; siendo la mayoría de los nodos seguidos hasta 50 veces, es decir, la mayoría de los nodos tienen 50 o menos seguidores.

Al mismo tiempo como ya se había comentado en el primer gráfico, no hay ningún nodo que no contenga seguidores. Todos los nodos son seguidos al menos una vez, es decir, hay algún nodo que les sigue, es decir, al menos tienen un seguidor.

4.3 Coeficiente de Agrupamiento

En esta sección se estudia el comportamiento estadístico del coeficiente de agrupamiento C(n) para los nodos n del grafo de 'follow'.

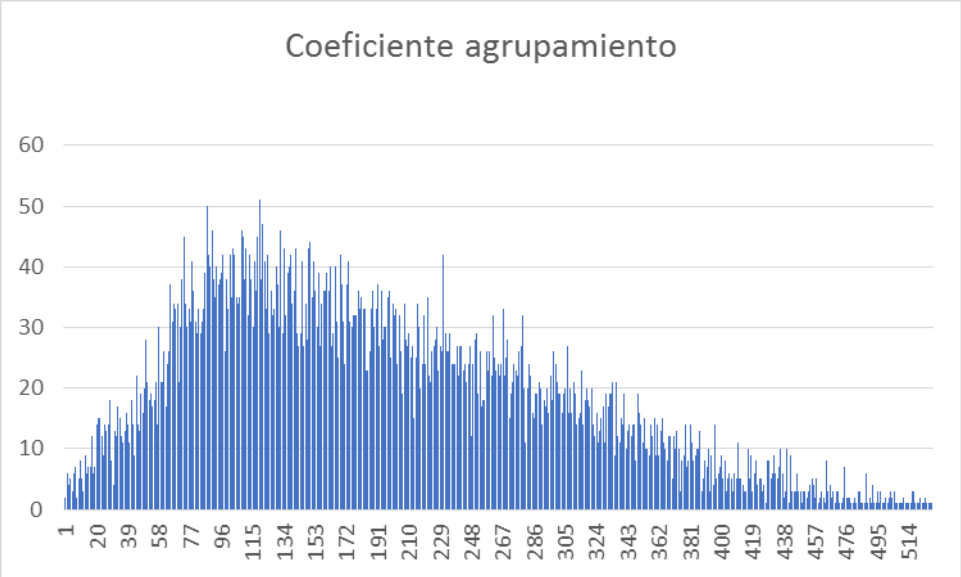

FIGURA 7: Relación nodos - coeficiente agrupamiento

En el gráfico adjunto del coeficiente de agrupamiento se ve como la mayoría de los nodos tienen un coeficiente menor que 300. De cierta manera el coeficiente de clustering se distribuye de una manera lateral, sin tener grandes picos ni bajadas.

El coeficiente de clustering es mayor o igual a 1 en casi todos los casos, solo hay dos nodos que tienen un coeficiente de clustering 0. Esto nos muestra que todos los nodos están muy agrupados entre sí, teniendo un grado alto de adherencia a otros nodos respecto a su número de vecinos.

El nodo que posee un mayor número de coeficiente de clustering es un nodo con un coeficiente de clustering de 529, en cambio en contrapartida hay dos nodos con un coeficiente de clustering con valor 0.

4.4 Comunidades

En este apartado se presenta un gráfico de las comunidades resultantes debido a la aplicación del algoritmo Kamada-Kawai.

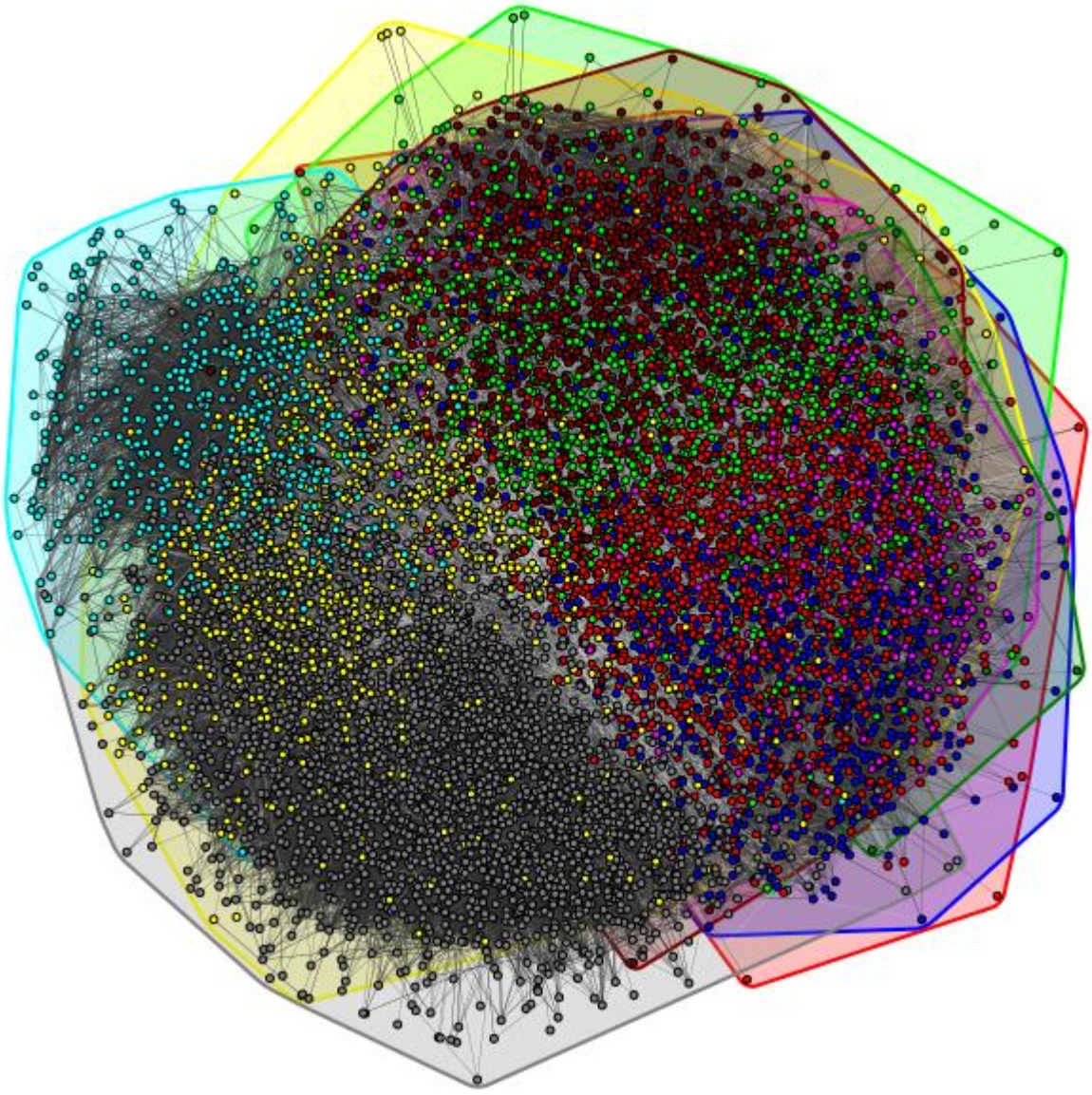

FIGURA 8: División del grafo en comunidades

Como se observa en la figura 8 cada nodo es representado por un círculo. Cada círculo está incluido en alguna comunidad. Las comunidades están representadas por distintos colores para facilitar su comprensión visual.

5 Resultados

El objetivo final de este estudio consiste en la búsqueda de alguna relación entre las características de los nodos proporcionados respecto a la viralidad de sus tweets, entendiendo la viralidad como el número de retweets de ese tweet en cuestión. La búsqueda de esa relación se buscó a través de la Correlación de Pearson.

Las variables con las que se buscó la relación respecto a la viralidad son las siguientes:

- **Número de vecinos**, número de nodos a los que sigue un nodo. En relación con los puntos 4.1 y 4.2.
- **Coeficiente de agrupamiento**, explicado en el punto 4.3.
- **Miembros de la comunidad**. Cada nodo pertenece a una comunidad, como se explicó en el punto 4.4. Se cuentan cuantos participantes hay en cada comunidad.
- **Número de comunidades distintas a las que los vecinos de un nodo pertenecen.** Quizás esta variable pueda parecer un poco compleja de entender a primera vista, por lo que será explicada mejor con un ejemplo a continuación:

  Sea un nodo A el cual sigue a cuatro nodos: {B, C, D, E}. B ∈ Comunidad 1.

  C ∈ Comunidad 2. {C, E} ∈ Comunidad 3. => Entonces el número de comunidades distintas son 3 = {Comunidad 1, 2 y 3}.

Una vez se ha medido esta relación los resultados obtenidos son los siguientes como se puede ver en la figura 9:

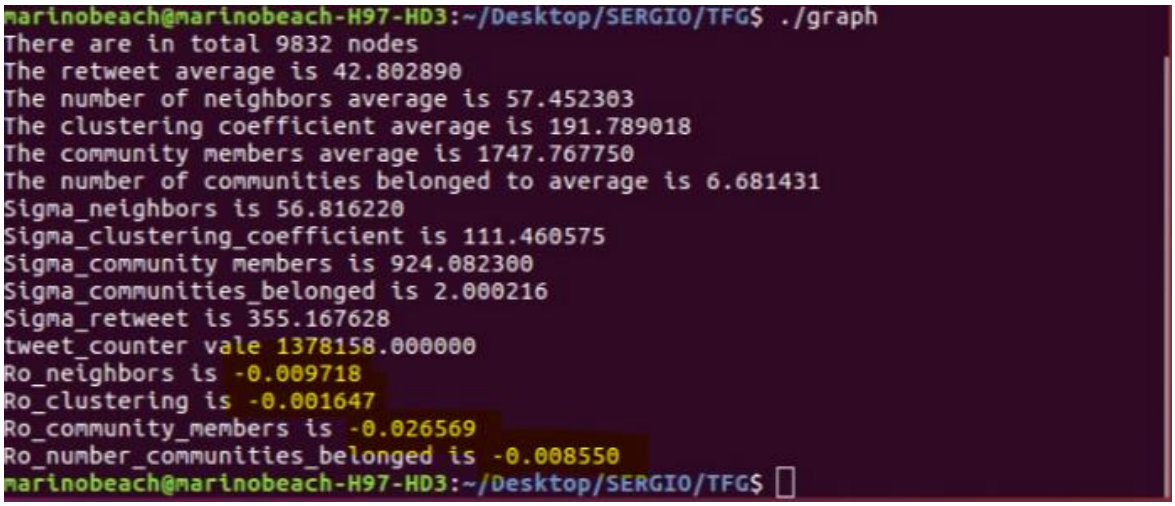

FIGURA 9: Ejecución resultados finales

Los resultados obtenidos se han hecho respecto a 1 378 158 tweets. Cada tweet posee un campo que se hace corresponder con un identificador de un nodo. En total el estudio se ha realizado sobre 9832 nodos. Se pasa ahora a analizar los resultados obtenidos.

El primer resultado importante es la relación de Pearson entre el número de vecinos y la viralidad de los tweets. La relación Pearson r obtenida para este caso es de -0.009718. Este es un valor muy cercano a 0, por lo que significa que los valores obtenidos se sitúan de forma aleatoria, por lo tanto no existe relación lineal entre las variables. Se adjunta una imagen relacionando el número de vecinos y retweets.

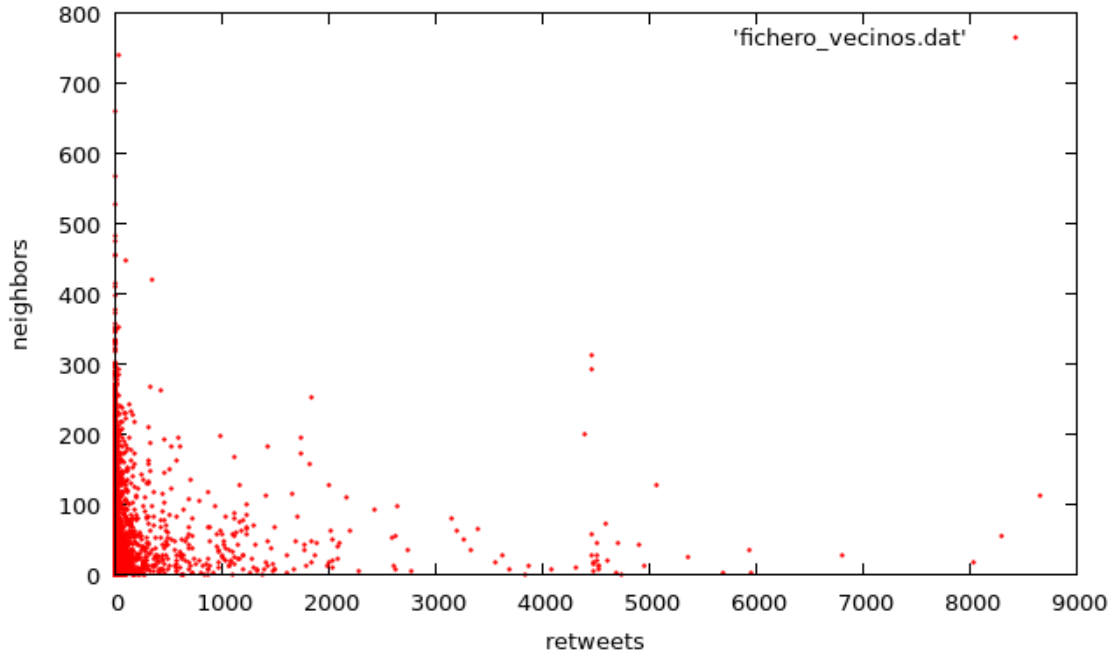

FIGURA 10: Relación neighbors - retweets

El segundo resultado importante es la relación de Pearson entre el coeficiente de agrupamiento/clustering coefficient y la viralidad de los tweets. La relación Pearson r obtenida para este caso es de -0.001647. Este es un valor muy cercano a 0, por lo que significa que los valores obtenidos se sitúan de forma aleatoria, por lo tanto no existe relación lineal entre las variables. Se adjunta una imagen relacionando el coeficiente de agrupamiento y los retweets.

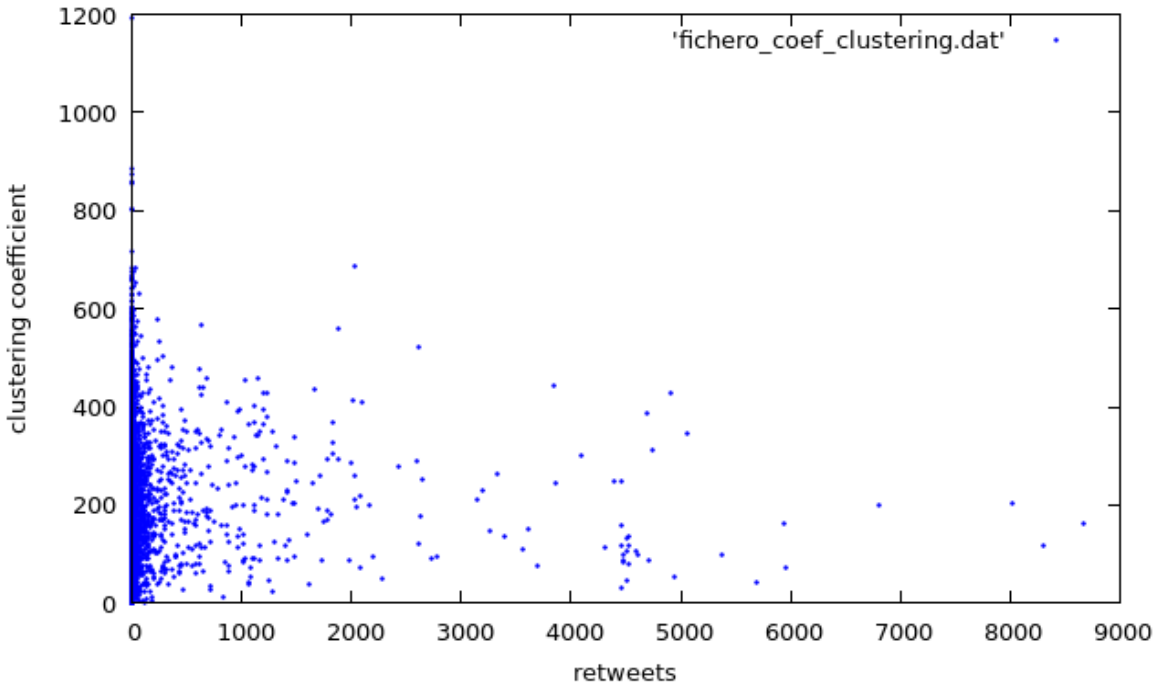

FIGURA 11: Relación coeficiente agrupamiento - retweets

El tercer resultado importante es la relación de Pearson entre los miembros de la comunidad y la viralidad de los tweets. La relación Pearson r obtenida para este caso es de -0.026569. Este es un valor muy cercano a 0, por lo que significa que los valores obtenidos se sitúan de forma aleatoria, por lo tanto no existe relación lineal entre las variables. Se adjunta una imagen relacionando los miembros de la comunidad y los retweets.

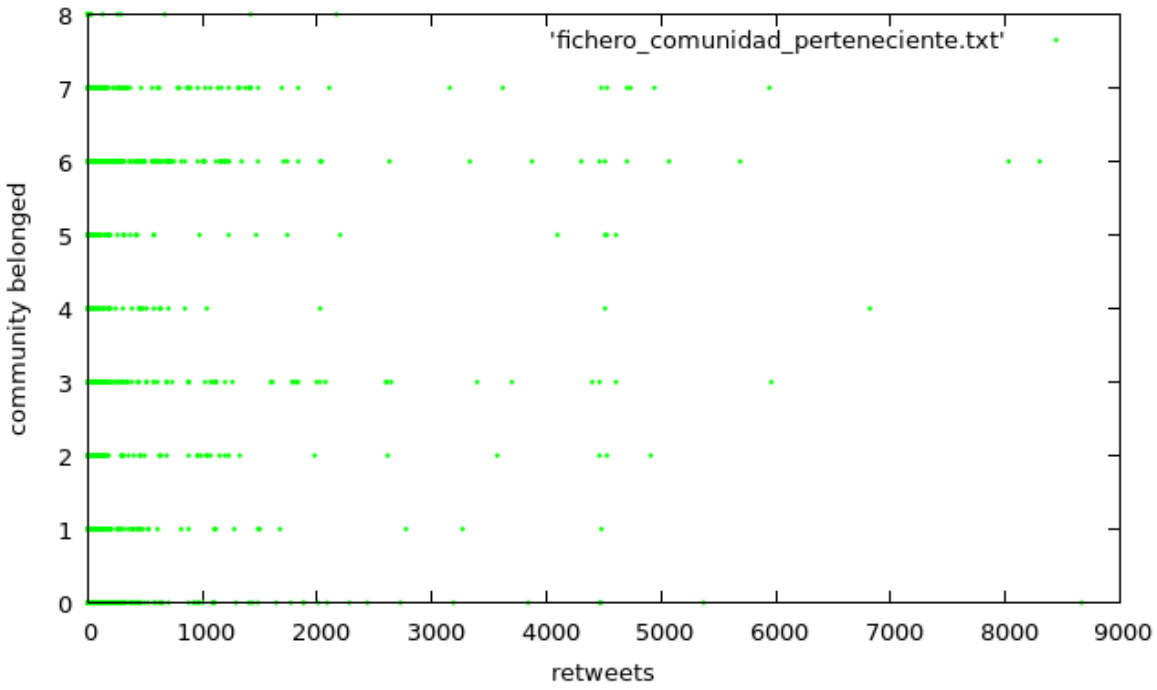

FIGURA 12: Relación miembros comunidad - retweets

El cuarto y último resultado importante es la relación de Pearson entre el número de comunidades distintas a las que los vecinos de un nodo pertenecen y la viralidad de los tweets. La relación Pearson r obtenida para este caso es de -0.008550. Este es un valor muy cercano a 0, por lo que significa que los valores obtenidos se sitúan de forma aleatoria no existiendo relación lineal entre las variables. Se adjunta una imagen que muestra la relación entre el número de comunidades distintas a las que los vecinos de un nodo pertenecen y los retweets.

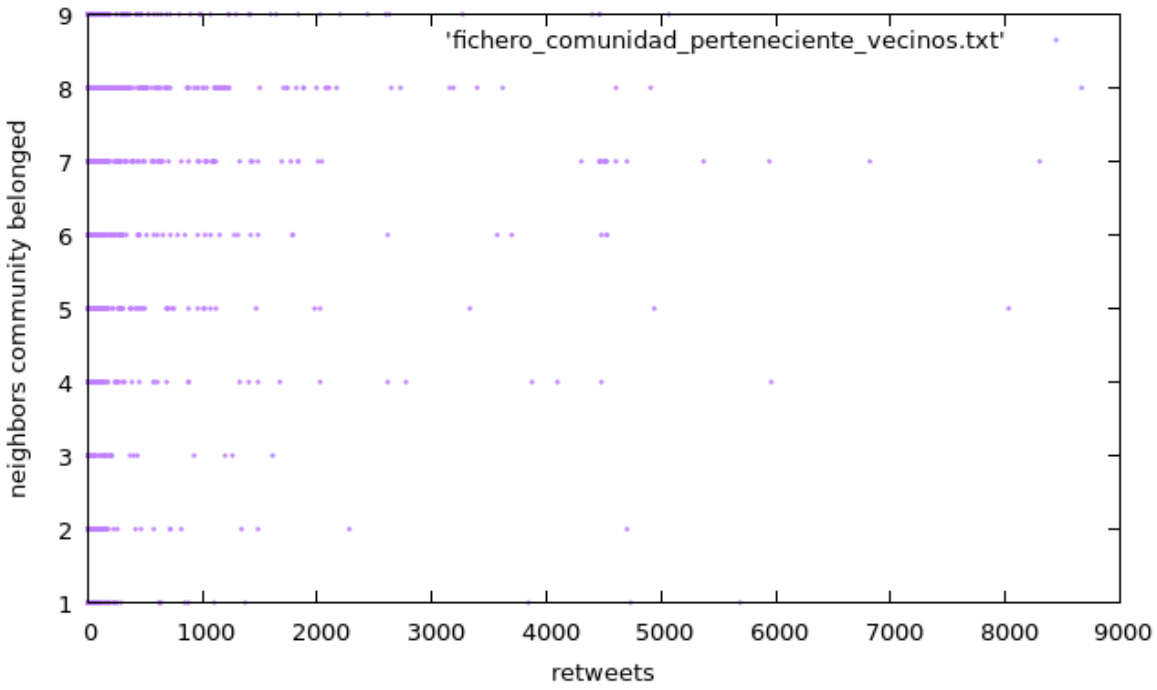

FIGURA 13: Relación número comunidades distintas - retweets

6 Conclusiones y trabajo futuro

Este apartado pretende abordar las distintas conclusiones a las que se ha llegado en este proyecto y un trabajo que se podría realizar en el futuro para ampliar e incluir en este estudio que se ha realizado.

6.1 Conclusiones

En este tfg se ha realizado un estudio acerca de la viralidad de los tweets. Se han ido viendo los datos en los que se inició el proyecto, las diferentes técnicas, procedimientos y librerías que se han utilizado para la obtención de datos con algún sentido y orden.

Se ha demostrado, mediante la relación de Pearson, que finalmente no existe ninguna relación entre el número de vecinos, los miembros de las comunidades, los números de comunidades distintas a las que los vecinos de un nodo pertenecen y la viralidad de los tweets.

6.2 Trabajo futuro

Como se ha ido viendo anteriormente en este tfg se han estudiado ciertos datos y características de los nodos y tweets que componen la red social Twitter. Sin embargo, no se ha podido sacar ninguna relación que explique el porqué de la viralidad de los tweets.

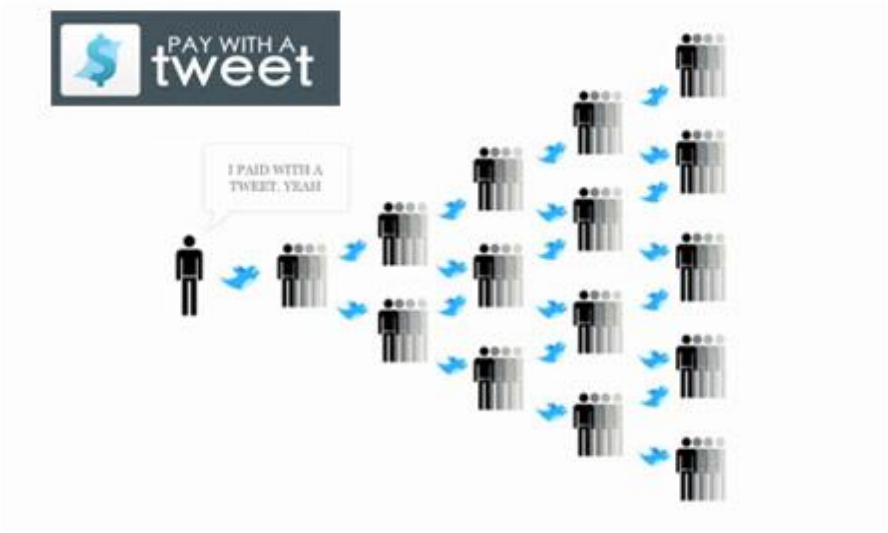

FIGURA 14: Concepto viralidad

Como trabajo futuro quedaría investigar el porqué de esta viralidad. Para ello se podría estudiar el contenido de los tweets en sí con algún analizador morfológico, buscar alguna relación con la fecha en la que esos tweets que se hicieron virales o estudiar de manera personal quién era o qué influencias tiene la persona que publicó esos tweets.

Otro factor que estudiar podría ser el tipo de interacción del tweet: mención, retweet, tweet etc. Y ver qué tipo de peso/weight lleva el tweet en cuestión. También se podría medir el idioma, la longitud del tweet, el tipo de media que lleva asociado etc.

En definitiva, hay muchas variables, factores y características que pueden influir en la viralidad del tweet. En un futuro se tendrá que averiguar cuáles son y cómo influyen.

Referencias

[1] Comunicación online para todos los públicos, “Viralidad: concepto y ejemplos”, 01.01.2012.
[2] Dean Romero, “Viralidad: ¿Qué es en el marketing y cómo sacarle provecho?”, 11.04.2017.
[3] Félix Ruiz, “PageRank: El algoritmo de Google”, 17.11.2013.
[4] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." *Computer networks and ISDN systems* 30.1-7 (1998): 107-117.
[5] Wikipedia, “Clustering Coefficient”, 28.12.2017.
[6] Tavish Srivastava, “Introduction to k-nearest neighbors: Simplified”, 10.10.2014.
[7] Zhang, Peng, Jinliang Wang, Xiaojia Li, Menghui Li, Zengru Di, and Ying Fan. "Clustering coefficient and community structure of bipartite networks." *Physica A: Statistical Mechanics and its Applications* 387, no. 27 (2008): 6869-6875.
[8] Fortunato, Santo. "Community detection in graphs." *Physics reports* 486.3-5 (2010): 75-174.
[9] Andrea Trevino, “Introduction to K-means Clustering”, 12.06.2016.
[10] Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." *IEEE transactions on pattern analysis and machine intelligence* 24.7 (2002): 881-892.
[11] Stephen G. Kobourov, “Force-Directed Drawing Algorithm”.
[12] The igraph core team, “The Kamada Kawai layout algorithm”, 2015.
[13] Kamada, Tomihisa, and Satoru Kawai. "An algorithm for drawing general undirected graphs." *Information processing letters* 31.1 (1989): 7-15.
[14] ERDdS, P., and A. R&WI. "On random graphs i." *Publ. Math. Debrecen* 6 (1959): 290-297.
[15] Wikipedia, “Barabási-Albert model”, 07.03.2018.
[16] Newton-Raphson, Métodos Numéricos.
[17] GorBrit, “Las Redes Sociales: Origen y evolución”, 24.06.2014.
[18] lainformacion.com, “Cómo es la historia de Facebook”, 03.02.2014.
[19] Miguel Jorge, “Historia de Twitter”, 21/03/2011.
[20] José Luis Orihuela, “10 características comunicativas de Twitter”, 20.09.2013.
[21] La Social Media, “Funciones de Twitter”, 25.09.2016.
[22] “Coeficiente de Correlación Lineal de Pearson”.
[23] Minitab, “Interpretar los resultados clave para Correlación”.

Glosario

Arista	Línea que resulta de la intersección de dos nodos en un esquema de representación en forma de árbol.
API	Application Programming Interface.
Centroid/Centroide	Punto que define el centro geométrico de un objeto.
Grafo	Conjunto de objetos denominados vértices o nodos unidos por unos enlaces llamados aristas.
Histograma	Representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados.
Internet	Unión de todas las redes. Red global en la que se juntan todas las redes que utilizan protocolos.
Librería	Conjunto de implementaciones funcionales, codificadas en un lenguaje de programación, que ofrece una interfaz bien definida para la funcionalidad que se invoca.
Mención	Capacidad de Twitter en la que se cita a una persona en un tuit.
Nodo	En un esquema de representación gráfica en forma de árbol, cada uno de los componentes de origen de las diferentes ramificaciones.
Retuitear/Retweet	Acción que permite compartir un tuit/tweet con todos tus seguidores.
Script	Archivo de órdenes. Es un programa usualmente simple almacenado en un archivo de texto plano.
Seguidor	Nodo que te sigue. Este nodo recibirá actualizaciones de estado que publiques.
Tuit	Texto que se publica en una red social con un máximo de 140 caracteres.
Twitter	Una red social que permite comunicarse con tuits.
Vecino	Nodo que es seguido por el nodo en cuestión.
Viralidad	Aquel contenido que se comparte y difunde altamente alcanzando una alta cantidad de visitas en un corto espacio de tiempo.