# Análisis del grafo de Twitter

Jorge Merlino

15 de Diciembre de 2015

**Resumen**

Para este trabajo se obtuvo un subgrafo con alrededor del 1 % de los nodos de Twitter a partir de la página web www.twitter.com. Se realizaron varios análisis de la estructura de este grafo basados en la disciplina de Social Network Analysis que permitieron obtener algunos datos interesantes principalmente por estudiar un subgrafo sesgado con nodos cercanos a nuestro país.

## 1. Introducción

Para este proyecto se obtuvo un subgrafo del grafo completo de Twitter y se realizó un análisis del mismo con el fin de obtener algunas medidas estadísticas de interés.

Estas medidas están basadas en Social Network Analysis (SNA). SNA es una disciplina que estudia las relaciones e intercambios entre personas, organizaciones, cosas o cualesquiera otras entidades interconectadas [35]. SNA utiliza como fundamento las teorías de redes [31] y grafos [10] y es parte de una corriente de sociología moderna liderada por Georg Simmel [34] y Émile Durkheim [19] que realza la importancia de las interconexiones entre actores, más allá de los actores en sí mismos.

SNA es aplicable a todo conjunto de entidades en donde sus interacciones pueden modelarse como un grafo. Por ejemplo ha mostrado su utilidad en sociología [21], estudio de redes de comunicaciones [2, 3], biología [8, 6], teoría de la información [37, 17], etc. SNA ayuda a visualizar y modelar las redes y sus participantes. Permite identificar los actores más importantes de una red, a detectar comunidades, a rastrear cómo se difunde la información y cómo se generan las opiniones, a predecir la dinámica futura de una red, etc.

Además, se compararon las medidas obtenidas con otros estudios del mismo tipo en la bibliografía para ver similitudes y diferencias entre los resultados.

## 2. Obtención del grafo

Se define un grafo dirigido en el que los nodos son los usuarios de Twitter y las aristas corresponden a la acción de seguir. Es decir que si el usuario $A$ sigue al usuario $B$ se tiene una arista dirigida desde $A$ hasta $B$.

Esta probablemente no sea la elección más natural para el sentido de las aristas dado que la información fluye en el sentido inverso, es decir desde el usuario seguido hacia sus seguidores. La razón para usar este sentido fue para simplificar la implementación dado que, si bien los usuarios promedio siguen a más usuarios de los que los siguen a ellos, en los casos extremos la cantidad de usuarios seguidos se espera que sea considerablemente menor en comparación al número de seguidores. Como ejemplo hay celebridades o figuras políticas que son seguidas por decenas de millones de usuarios mientras que en nuestro grafo el caso más extremo de usuarios seguidos llega a dos millones y está muy lejos del segundo mayor.

Como se verá más adelante no hay forma de realizar una selección aleatoria no sesgada de los seguidores de un usuario por lo que no hay más alternativa que recuperarlos a todos. Por esta razón, llegar al nodo de una celebridad podría dejarnos muchos días procesando sus usuarios seguidos (por ejemplo recolectar 50 millones de usuarios nos llevaría cerca de 150 días de la forma en que funciona el algoritmo que se describe más adelante en esta sección)

Para obtener la lista de usuarios que sigue un usuario determinado se intentó usar las operaciones del API de Twitter pero esto no resultó práctico dado que para esto se debe crear una aplicación y solo se pueden obtener los datos de los usuarios que siguen a dicha aplicación.

Como alternativa se recurrió al *scraping* de la página web de Twitter (www.twitter.com) dado que estos datos si son accesibles libremente desde la misma. Esto está expresamente prohibido en las condiciones de servicio de Twitter que dicen:

> Crawling the Services is permissible if done in accordance with the provisions of the robots.txt file, however, scraping the Services without the prior consent of Twitter is expressly prohibited.

Por esta razón se creó un usuario de prueba para obtener el grafo de forma de evitar el bloqueo de un usuario real en caso de ser detectados. Con el fin de evitar la detección el script de scraping solo realiza un pedido cada 5 segundos. Con esta configuración nunca se tuvieron problemas de bloqueos. El usuario de prueba solamente sigue a mi usuario dado que solo cumple la función de intermediario.

En la página de Twitter los usuarios que sigue una persona se presentan como una lista con *scroll* infinito. Es decir que a medida que el usuario se acerca al final de la lista se realiza un nuevo pedido asincrónico al servidor que agranda dinámicamente la lista que se está recorriendo.

Se investigó cual era el método invocado para obtener la lista inicial, y cual era el método invocado asincrónicamente para agrandar la lista. Para esto se usaron las funciones de debug de $Firefox$ que permiten obtener los parámetros del comando $cURL$ para simular los pedidos realizados por el navegador. Al procesar un usuario primero se obtiene la página con la lista de usuarios inicial y luego se llama cada 5 segundos a la función asincrónica simulando un scroll manual sobre la misma.

Cada una de las llamadas a estos métodos retornan a lo sumo 20 usuarios en formato *JSON*. Los usuarios retornan ordenados según lo que Twitter estima que son los usuarios más interesantes y no se pueden saltear páginas. Esto es así pues en cada pedido se recibe un código que se debe usar para pedir la siguiente página. Esto es un número de alrededor de 20 dígitos que no es creciente por lo que es muy difícil predecir cual va a ser el número asociado a la página *n-esima* de la lista. Por estas razones es que se indicó que no era posible realizar un muestreo aleatorio de las aristas del grafo.

Se desarrolló un algoritmo que recorre el grafo obteniendo los usuarios que son seguidos por cada nodo. La idea es ir procesando los nodos en orden *BFS* recorriendo primero todos los de distancia uno, para luego mirar los de distancia dos, etc. Se intentó buscar un grafo con todos los nodos con distancia hasta cuatro del usuario original. Esto es equivalente a todos los usuarios con distancia menor o igual a tres de mi usuario. El tiempo no fue suficiente para obtener el grafo completo así que se cortó antes de obtener todos los usuarios a distancia cuatro.

El algoritmo fue implementado en *Python* usando una base de datos *SQLite* para guardar la información del grafo. Se usa el programa *cURL* para efectuar los pedidos a la página web de Twitter. Se guarda una lista de adyacencia en una tabla mientras que otra indica los usuarios que ya fueron procesados para no procesarlos más de una vez. Esto puede ocurrir cuando distintos usuarios siguen a la misma persona. Además se tuvieron en cuenta algunos otros detalles como que a veces es posible que Twitter retorne al mismo usuario más de una vez en la lista de usuarios seguidos. Esto puede ocurrir sobre todo para los usuarios que siguen a mucha gente y en los que el algoritmo puede estar varias horas para obtener la lista completa de usuarios.

# 3. Estructura del grafo

El grafo completo obtenido tiene 7.804.481 nodos y 13.593.164 aristas en total. Se consideraba que había alrededor de 650 millones de usuarios en Twitter en setiembre de 2015 [38] por lo que este subgrafo contiene algo más del 1 % de los nodos del grafo total. No hay información publicada sobre la cantidad de aristas del grafo.

Si nos limitamos a los nodos a distancia dos el grafo tiene 29 nodos y 143 aristas. Podemos ver una representación del grafo en la figura 1. Si consideramos los nodos de hasta distancia tres tenemos 9.323 nodos y 1.213.808 aristas. Se puede ver una representación de este grafo en la figura 2.

En estas figuras el tamaño de los nodos está dado por su valor de centralidad de cercanía (ver sección 4.4.1) y el color esta determinado por el grado de salida. Son más azules los de menor grado y más rojos los de grado mayor. El usuario inicial está marcado con dos asteriscos (**) y mi usuario con uno (*). En la figura 2 no son visibles dado que quedan cubiertos por otros nodos.

Para visualizar los grafos se usó la biblioteca *gephi* [41].

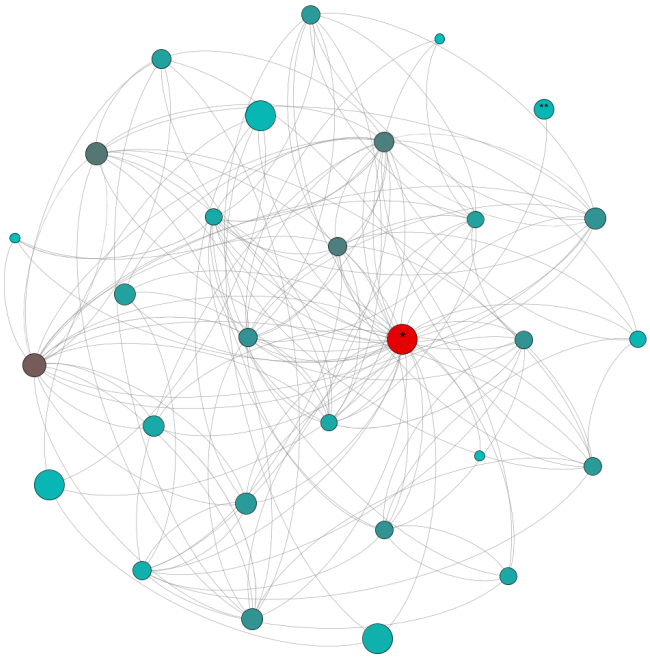

Figura 1: Grafo hasta distancia dos

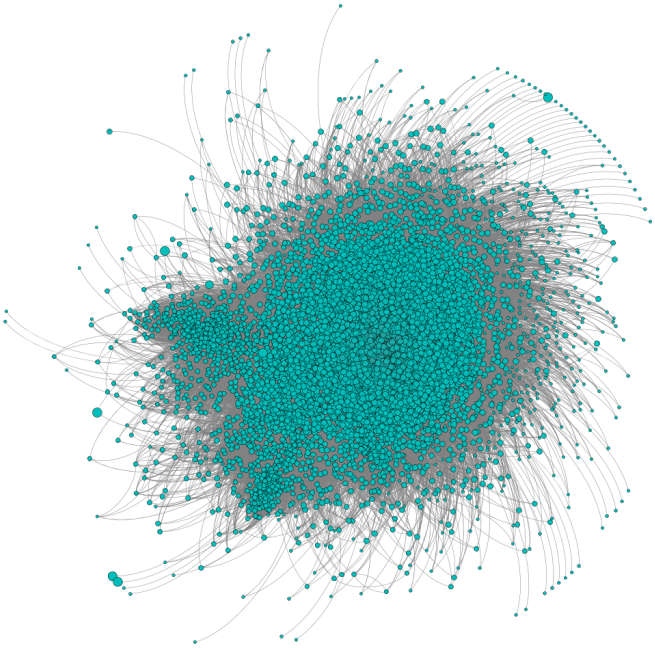

Figura 2: Grafo hasta distancia tres

Mirando por nivel se tienen 29 nodos a distancia dos, 9.294 a distancia tres y 7.795.158 a distancia cuatro considerando que estos últimos no son todos porque no fue suficiente el tiempo para obtenerlos todos.

# 4. Estadísticas calculadas

Se determinaron varias estadísticas interesantes para calcular en el grafo principalmente a partir de lo planteado en [27] y [14]. Para ello se utilizó el programa *R* [39] con la biblioteca *igraph* [40]

## 4.1. Diámetro

El diámetro de un grafo se define como el máximo camino más corto entre cualquier par de nodos. Para grafos grandes es muy difícil de calcular dado que su tiempo de ejecución es $O(n^3)$ siendo $n$ la cantidad de nodos.

En este caso el diámetro es cuatro por la forma en que se construye el grafo. Considerando que, como ya se comentó, se estima que hay unos 650 millones de usuarios de Twitter y la relación que hay entre los niveles se podría suponer que se conseguiría obtener el grafo completo (o al menos su componente conexa principal) siguiendo solamente hasta distancia cuatro (ignorando el nodo inicial). Por esto podríamos asumir que el grafo completo no debe tener distancia mayor a cuatro o cinco.

Según la bibliografía es esperable que el diámetro de este grafo sea pequeño dado que lo mismo se observa en grafos de conectividad de Internet, links de páginas web y también en redes sociales [5, 29, 4, 11, 15, 18, 36]. Es lo que se conoce en la literatura como el fenómeno de *mundo pequeño* (small world).

## 4.2. Grado de entrada y salida

Para medir los grados de entrada y salida no se consideraron los nodos con grado de salida en cero. Es decir los nodos de distancia cuatro y los de distancia tres que no llegaron a ser procesados. Sí se cuentan las aristas que llegan a estos nodos para el grado de salida de los nodos de origen.

En primera instancia es interesante ver cuales son los nodos con mayor grado de entrada y salida. Esto se corresponde con los usuarios que siguen a más gente y con los usuarios que son seguido por más usuarios respectivamente. Los cinco mayores grados de entrada se pueden ver en la tabla 1 y los cinco mayores grados de salida están en la tabla 2.

| Posición | Usuario | Grado |
|---|---|---|
| 1 | @Tiranos_Temblad | 4.641 |
| 2 | @ObservadorUY | 3.702 |
| 3 | @LuisSuarez9 | 3.587 |
| 4 | @DiegoForlan7 | 3.412 |
| 5 | @NoToquenNada | 3.350 |

Tabla 1: Mayores grados de entrada

| Posición | Usuario | Grado |
|---|---|---|
| 1 | @alispagnola | 2.013.910 |
| 2 | @BarackObama | 639.562 |
| 3 | @leopoldolopez | 334.252 |
| 4 | @Nexofin | 331.011 |
| 5 | @Variety | 299.207 |

Tabla 2: Mayores grados de salida

El promedio de grado de entrada es de 136 aunque la desviación estándar llega a 256 por lo que se ve que es bastante alta. Por el lado del grado de salida el promedio es mayor con 1.549 aunque la varianza es mucho mayor que antes valiendo 24.424. Se ve que naturalmente los usuarios en promedio

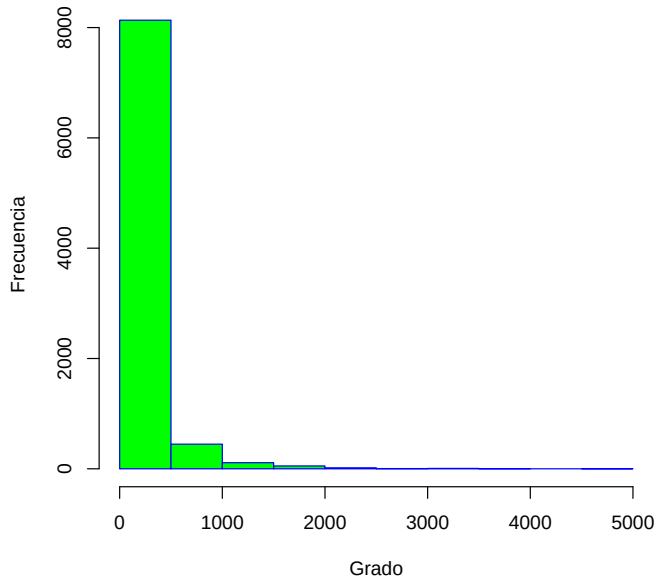


Figura 3: Distribución de grado de entrada

siguen a más gente de la que los sigue a ellos y que el número de seguidores tiene mucha menos varianza que el número de usuarios seguidos.

Considerando lo comentado antes es interesante analizar la distribución de grado de entrada y de salida. Se muestra un histograma para el grado de entrada en la figura 3 y para el grado de salida en la figura 4. Para el grado de salida se usaron 1000 particiones de las cuales no se muestran todas en la figura.

Podemos ver que la distribución de grado responde a la llamada *ley potencial* (power law) dado que se cumple aproximadamente que la cantidad de nodos $N_d$ de tamaño $d$ está dada por $d^{-\lambda}$ donde $\lambda$ es el exponente de la ley potencial.

Este tipo de distribuciones de grado se observa en varios tipos diferentes de grafos como ser llamadas telefónicas [1], Internet [20], páginas web [24, 15, 9, 23, 25], grafos de citas bibliográficas [33] y redes sociales [16]

## 4.3. Coeficiente de clustering y reciprocidad

Para esta estadística se consideró solamente el grafo completo con los nodos hasta distancia 3 dado que las aristas a nodos no procesados afectan negativamente estos valores.

El coeficiente de clustering $C_v$ de un vértice $v$ se define de esta forma: si $v$ tiene $d$ vecinos entonces a lo sumo puede haber $d(d-1)/2$ aristas entre ellos; entonces $C_v$ es la fracción de estas aristas que realmente existen. Dicho de otra forma $C_v$ es la cantidad de aristas entre un nodo y sus vecinos dividido la cantidad total posible de aristas que pueden existir. El coeficiente de clustering para este grafo es de 0,188. Este dato parece consistente con la bibliografía. En [22] se dice que el coeficiente de clistering de Flickr [30] es 0,108, el de LiveJournal [7] es 0,118 y el de una red de referencias bibliográficas [28] es 0,187.

Otra medida relacionada es la reciprocidad que es la probabilidad de que si existe una arista en un sentido también exista una en el sentido opuesto. La reciprocidad del grafo es de 0,334. Este valor parece ser mayor en nuestro grafo que las estimaciones que se encuentran en la bibliografía donde se considera a Twitter como una red con baja reciprocidad. Por ejemplo en [26] se habla del 22 %.

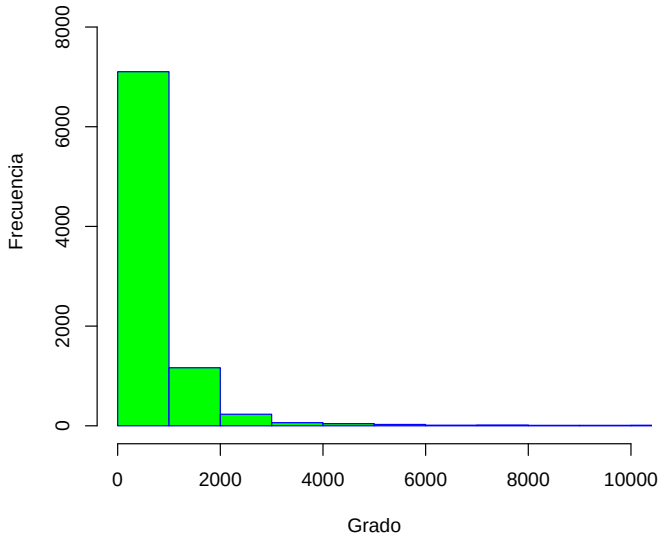


Figura 4: Distribución de grado de salida

## 4.4. Centralidad

Informalmente la centralidad de un nodo intenta asignarle un valor al mismo indicando que tan importante es el lugar que ocupa en el grafo. Hay varias formas de medir esto, pero según [14] para un grafo del tipo de Twitter las mas relevantes serian las llamadas centralidades de cercanía y de grado.

Para estas estadísticas y las siguientes no se consideran los nodos con grado de salida cero y las aristas hacia ellos. Es decir no se considera a los usuarios que no fueron procesados.

### 4.4.1. Centralidad de cercanía

Se define como:

$$C_c(i,g) = \frac{n-1}{\sum_{i \neq j} d(i,j)}$$

donde $i$ y $j$ son nodos, $g$ es un grafo, $d(i,j)$ es la distancia mínima entre $i$ y $j$ y $n$ es la cantidad de nodos de $g$ [21]. Intuitivamente representa que tan cerca está el nodo con respecto al resto de los nodos del grafo.

En un grafo dirigido, podemos pensar en la centralidad de cercanía entrante como la cantidad promedio de pasos que se deben dar para llegar a un nodo dado desde todos los otros nodos alcanzables en el grafo. La centralidad de cercanía saliente mide lo mismo en el otro sentido.

En el contexto de flujo de información podemos interpretar la centralidad de cercanía como el tiempo esperado hasta la llegada de un flujo en la red [13]. Los nodos con centralidad alta tienden a recibir los flujos antes asumiendo que se originan desde cualquiera de los otros nodos con igual probabilidad y también asumiendo que recorren los caminos más cortos. En el caso de Twitter como normalmente la información recorre todos los caminos al hacerse en *broadcast* en particular usará el más corto.

Para mostrar otros ejemplos, las organizaciones con alta centralidad en una red de transferencia de tecnología de I&D serán capaces de desarrollar nuevos productos antes que otros. Por otro lado, individuos con alta centralidad en una red sexual tienen más probabilidad de contraer enfermedades en forma temprana, posiblemente antes de que haya tratamientos disponibles en el caso de enfermedades nuevas.

Considerando que la información viaja en el sentido inverso a la dirección de las aristas de nuestro grafo, los nodos con alta centralidad de cercanía entrante son los que son más escuchados por los otros nodos (la información que originan llega más rápido) mientras que los de alta centralidad de cercanía saliente son los que reciben más rápidamente la información desde el resto de la red.

El nodo con mayor centralidad de cercanía entrante en nuestro grafo es @BBCBreaking seguido por @BBCWorld y @RollingStones. @BBCBreaking tiene casi 20 millones de seguidores en total y 666 en nuestro grafo. Tiene sentido que sean sitios de noticias los de mayor centralidad entrante pues generan información de interés para muchas personas.

Por otro lado el nodo con mayor centralidad de cercanía saliente es @1121958 (Milton) seguido por @Tiranos_Temblad y @chanchosUY. Es extraño el caso del primer usuario por no tratarse de un usuario conocido (al menos no para mi). Al parecer es alguien con la habilidad de seguir a los usuarios correctos para poder recibir noticias con pocos saltos. Los otros dos usuarios son seguidos por mucha gente de Uruguay y ellos mismos siguen a muchos de sus seguidores lo cual los ubica en una buena posición. Otra diferencia importante es que @1121958 sigue a menos de 700 personas (muchas de ellas son medios de comunicación) mientras que los otros siguen a varios miles.

### 4.4.2. Centralidad de grado

Se define como:

$$C_d(i, g) = \frac{d_i(g)}{n - 1}$$

donde $i$ es un nodo, $g$ es un grafo, $d_i(g)$ es el grado del nodo $i$ y $n$ es la cantidad de nodos del grafo $g$ [21]. Le da más peso a los nodos con mayor grado. Podemos definirlo como el número de caminos de largo uno que se originan desde un nodo sobre todos los posibles. En el caso de Twitter, al copiar la información por todas las aristas de salida, la probabilidad de recibir en el próximo período un mensaje que se está transmitiendo en forma aleatoria por la red es una función del grado de entrada del nodo (de salida en nuestro caso). Lo mismo sucede con el grado de salida (de entrada en nuestro caso) si pensamos en la probabilidad de que el resto de la red reciba un mensaje emitido por el nodo.

Otra forma de interpretarlo es como una medida de los efectos inmediatos desde un instante $t$ al $t + 1$. Por ejemplo si una cantidad de nodos en la red están infectados con una enfermedad y las aristas implican contagio, entonces la probabilidad de infección inmediata es una función del grado del nodo.

También podemos pensar que, si tenemos un proceso de Markov haciendo una recorrida aleatoria infinita sobre el grafo, la proporción de veces que cada nodo es visitado es función del grado del nodo.

En nuestro grafo separamos las medidas entre la centralidad de grado saliente y entrante. El nodo con mayor centralidad de grado entrante es @Tiranos_Temblad seguida por @ObservadorUY y @LuisSuarez9. Esto coincide exactamente con los tres nodos con mayor grado de entrada.

El nodo con mayor centralidad de grado saliente también es @Tiranos_Temblad seguido por @chanchosUY y @musicaENVIVO_Uy. Aquí no coincide con los nodos con mayor grado saliente porque al eliminar los nodos de grado de salida cero y las aristas hacia ellos muchos usuarios con alto grado de salida (que no son uruguayos) pierden grado al perder las aristas eliminadas a usuarios que no fueron procesados.

## 4.5. PageRank

El algoritmo de PageRank [32] está muy relacionado con otra medida de centralidad conocida como centralidad de valor propio [12]. Esta está dada por el vector propio principal de la matriz de adyacencia que define al grafo. La centralidad de valor propio es similar a la centralidad de grado con la diferencia que la primera mide relaciones directas e indirectas de largo plazo y la segunda mide relaciones inmediatas solamente.

PageRank también se puede ver como el valor propio de una matriz. Sea $A$ a la matriz donde valor en la posición $A_{uv}$ es $1/N_u$ donde $N_u$ es el grado de salida de $u$ si hay una arista entre $u$ y $v$ y 0 en otro caso. PageRank es entonces el vector propio principal de la matriz $A + E \times 1$ donde 1 es un vector de unos y $E$ es un vector que asigna un valor a cada página. Podemos ver a E como el vector que determina la probabilidad de saltar a cada página cuando se ejecuta una teletransportación.

En general estos modelos no son del todo apropiados para un modelo como el de Twitter dado que se basan en que la información se mueve aleatoriamente sin restricciones y puede recorrer las mismas aristas varias veces. Esto en general no es razonable con Twitter dado que los usuarios normalmente no comparten la misma información múltiples veces.

El nodo de nuestro grafo con mayor PageRank es también @BBCBreaking (seguido por @BBC-World y @BBCNews) y el nodo con mayor centralidad de valor propio es @ObservadorUY (seguido por @dcastro65 y @elpaisuy).

# 5. Conclusiones y trabajo futuro

Pudimos observar que los datos que se obtuvieron usando SNA a partir del grafo dan valores razonables lo cual valida la metodología usada y la efectividad del uso de SNA. También pudimos ver que los datos obtenidos coinciden en gran medida con otros estudios similares en la bibliografía lo cual también confirma que el grafo obtenido se comporta en forma similar a otras redes estudiadas.

Como trabajo futuro podemos comentar que sería interesante volver a evaluar alguna de estas estadísticas calculadas con el grafo completo de distancia cuatro para observar si hay diferencias significativas o no. Por otro lado se podría realizar un análisis de detección de comunidades para ver cuales son los subgrupos de nodos del grafo con más conectividad entre sí. Esto se intentó hacer para este trabajo pero no fue posible finalizarlo a tiempo dada la demora en la ejecución de estos algoritmos en un grafo de este tamaño en el hardaware que se tenía disponible.

# Referencias


[1] J. Abello, A. L. Buchsbaum, and J. Westbrook. *A functional approach to external graph algorithms.* In Proceedings of the 6th Annual European Symposium on Algorithms, pages 332–343, 1998.

[2] J. Abello, P. M. Pardalos, and M. G. C. Resende. *On maximum cliques problems in very large graphs.* In J. Abello and J. Vitter, editors, External memory algorithms, volume 50 of DIMACS Series on Discrete Mathematics and Theoretical Computer Science, pages 119–130. American Mathematical Society, 1999.

[3] J. Abello, M. G. C. Resende, and S. Sudarsky. *Massive quasi-clique detection.* In S. Rajsbaum, editor, LATIN 2002: Theoretical Informatics, volume 2286 of Lecture Notes in Computer Science, pages 598–612. Springer Verlag, 2002.

[4] R. Albert, H. Jeong, and A.-L. Barabasi. *Diameter of the world-wide web.* Nature, 401:130–131, September 1999.

[5] R. Albert and A.-L. Barabasi. *Statistical mechanics of complex networks.* Reviews of Modern Physics, 74(1):47–97, 2002.

[6] R. M. Anderson, R. M. May. *Infectious diseases of humans: Dynamics and control.* 2002.

[7] L. Backstrom, D. P. Huttenlocher, J. M. Kleinberg, and X. Lan. *Group formation in large social networks: membership, growth, and evolution.* In KDD, pages 44–54, 2006.

[8] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications.* 2nd edition, 1975.

[9] A.-L. Barabasi and R. Albert. *Emergence of scaling in random networks.* Science, 286:509–512, 1999.

[10] N. Biggs, E. Lloyd, R. Wilson, *Graph Theory.* Oxford University Press, 1986.

[11] B. Bollobas and O. Riordan. *The diameter of a scale-free random graph.* Combinatorica, 24(1):5–34, 2004.

[12] P. Bonacich. *Factoring and weighting approaches to status scores and clique identification.* Journal of Mathematical Sociology 2, 113–120, 1972

[13] S. P. Borgatti. *Centrality and AIDS.* Connections 18 (1), 112–114. 1995

[14] S. P. Borgatti, *Centrality and network flow.* Social networks 27.1 (2005): 55-71.

[15] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. *Graph structure in the web: experiments and models.* In WWW '00: Proceedings of the 9th international conference on World Wide Web, 2000.

[16] D. Chakrabarti, Y. Zhan, and C. Faloutsos. *R-mat: A recursive model for graph mining.* In SDM '04: SIAM Conference on Data Mining, 2004.

[17] A. D. Chepelianskii, *Towards physical laws for software architecture.* CoRR, abs/1003.5455, 2010.

[18] F. Chung and L. Lu. *The average distances in random graphs with given expected degrees.* Proceedings of the National Academy of Sciences, 99(25):15879–15882, 2002.

[19] É. Durkheim, *The Rules of Sociological Method* (1895) 8th edition, trans. Sarah A. Solovay and John M. Mueller, ed. George E. G. Catlin (1938, 1964 edition)

[20] M. Faloutsos, P. Faloutsos, and C. Faloutsos. *On power-law relationships of the internet topology.* In SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication, pages 251–262, 1999.

[21] L.C. Freeman. *Centrality in networks: I. Conceptual clarification.* Social Networks 1, 215–239. 1979

[22] S. J. Hardiman, L. Katzir. *Estimating clustering coefficients and size of social networks via random walk.* Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013.

[23] B. A. Huberman and L. A. Adamic. *Growth dynamics of the world-wide web.* Nature, 399:131, 1999

[24] J. M. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. *The web as a graph: Measurements, models and methods.* In COCOON '99: Proceedings of the International Conference on Combinatorics and Computing, 1999.

[25] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. *Trawling the web for emerging cybercommunities.* Computer Networks, 31(11-16):1481–1493, 1999.

[26] H. Kwak, C. Lee, H. Park and S. Moon *What is Twitter, a Social Network or a News Media?* Proc. WWW, 591-600. 2010

[27] J. Leskovec , C. Faloutsos, *Dynamics of large networks*, Carnegie Mellon University, Pittsburgh, PA, 2008

[28] M. Ley. *The DBLP computer science bibliography: Evolution, research issues, perspectives.* In Proc. Int. Symp. on String Processing and Information Retrieval, pages 1–10, 2002.

[29] S. Milgram. *The small-world problem.* Psychology Today, 2:60–67, 1967.

[30] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. *Growth of the flickr social network.* In Proceedings of the 1st ACM SIGCOMM Workshop on Social Networks (WOSN'08), August 2008.

[31] M. E. J. Newman, *The structure and function of complex networks.* Department of Physics, University of Michigan.

[32] L. Page, S. Brin, R. Motwani, T. Winograd. *The PageRank citation ranking: bringing order to the Web.* 1999.

[33] S. Redner. *How popular is your paper? an empirical study of the citation distribution.* European Physical Journal B, 4:131–134, 1998.

[34] G. Ritzer, *Modern Sociological Theory* (7th ed.). New York: McGraw–Hill 2007. ISBN 0073404101

[35] J. Scott, *Social Network Analysis*, SAGE publications, 2013. ISBN 978-1-4462-0903-5

[36] D. J. Watts and S. H. Strogatz. *Collective dynamics of 'small-world' networks.* Nature, 393:440–442, 1998.

[37] A. O. Zhirov, O. V. Zhirov, D. L. Shepelyansky, *Two-dimensional ranking of Wikipedia articles.* Eur. Phys. J. B v.77, 2010 p.523.

[38] http://www.statisticbrain.com/twitter-statistics

[39] https://www.r-project.org

[40] http://igraph.org/r

[41] https://gephi.org