Analítika

Redes sociales evolutivas: un ejercicio descriptivo y predictivo

Rolando Mantilla

Redes sociales evolutivas

Un ejercicio descriptivo y predictivo

Rolando Mantilla

Servicio de Rentas Internas, Quito, Ecuador
rpmantilla@sri.gob.ec

Resumen

Se exponen una serie de elementos de la teoría de grafos para el análisis de redes sociales tanto descriptivo como predictivo, para luego analizar un conjunto de mensajes obtenidos de la Red Social Twitter en varios periodos de tiempo consecutivos que conforman una red de menciones dinámicas. Se realiza un análisis descriptivo y evolutivo de las principales medidas de estructura del grafo y de varias medidas de centralidad; a continuación, de manera predictiva se aplica un modelo basado en una matriz estocástica promedio para estimar las menciones que podría tener un usuario en un periodo futuro.

Palabras clave: Redes sociales, teoría de grafos, matrices estocásticas.

Abstract

A set of elements of the graph theory and the social network analysis are discussed in the context of both descriptive and predictive statistics, to then analyze a set of messages obtained from the Twitter social network in several consecutive periods of time that make up a dynamic network of mentions. An evolutionary descriptive analysis is made of the main measures of structure of the graph and of several centrality measures. Then, a predictive network model is proposed based on an average stochastic matrix to estimate the mentions that a user could have in a future period.

Keywords: Social networks, graph theory, stochastic matrices.

Clasificador MSC: 91D30, 68R10, 15B51

1 Introducción

Los análisis de las relaciones o interacciones entre individuos en un contexto social permiten detectar e interpretar los patrones de los vínculos sociales entre sus actores. Es en este ámbito, la teoría de redes sociales provee fundamentos y herramientas basadas principalmente, en las matemáticas, los grafos y la computación, para analizar y tener una apreciación de las cualidades colectivas de la red, y medir la importancia que tienen los individuos que la conforman conforme su relacionamiento.

En tal sentido, la teoría de redes sociales ha tenido un gran impacto dentro de varias disciplinas, principalmente dentro de la Sociología y específicamente en la denominada Sociometría (Moreno, 1951). En la actualidad ha encontrado sus principales aplicaciones en problemas vinculados al Internet, a menudo en el ámbito del Big Data, que van desde grandes redes de comunicación, pasando por las conocidas redes sociales como Facebook, LinkedIn, Twitter, hasta en algoritmos de búsqueda como el Google Page Rank.

Uno de los retos a afrontar es el tamaño de las redes que plantean las aplicaciones mencionadas, para lo cual se pueden encontrar bases de datos de grafos y paquetes computacionales que tienen implementados algoritmos que permiten analizar estas redes o grafos; por ejemplo, el paquete *igraph* disponible para R o el programa *Pajek* especializado en redes extensa, además de la base de datos especializada en grafos *Neo4J*.

El análisis de redes sociales abarca temas como la identificación de individuos con opiniones influyentes en medios electrónicos. En la administración de impuestos, se puede utilizar para el análisis de transacciones financieras e identificar clústeres económicos, transacciones anómalas o inusuales o individuos de especial relevancia en tramas de evasión fiscal. Un ejemplo de especial relevancia constituye el caso de los "Panamá Papers" en donde las herramientas del análisis de redes sociales permitieron identificar a los individuos clave en las tramas de evasión basados en el uso de paraísos fiscales (Lion, 2016).

Por otra parte, resulta novedoso e interesante observar el comportamiento de estas redes en el tiempo y proponer ejercicios predictivos sobre la formación de relaciones entre actores, que pueden resultar útiles para la aplicación de acciones tempranas sobre personas con mejores probabilidades de convertirse en actores relevantes en las redes, ya sea en el ámbito de las ventas, el mercadeo o inclusive la prevención del fraude fiscal.

Actualmente, existe una aplicación extensa en empresas comerciales que tratan de explotar estas redes, desde el punto de vista del Marketing, pues proveen de un excelente medio para propagar recomendaciones a través de grupos y personas con intereses similares (Huberman *et al.*, 2008). En la administración pública también se puede hacer uso de grandes redes de transacciones y sus partícipes, para entender las dinámicas económicas.

En concordancia a esta temática, este trabajo busca realizar un aporte en la aplicación de las herramientas de las redes sociales que van desde su ámbito tradicional descriptivo y de carácter estático, a un ámbito dinámico y finalmente predictivo; esto, basado en una adaptación de carácter estocástico del algoritmo Page Rank de Google.

Para ello, se parte exponiendo una serie de elementos básicos del análisis de redes sociales y de la teoría de grafos, los mismos que permiten conocer varias características de los grafos, como las medidas de centralidad (grados, cercanía, intermediación), así como otras medidas estructurales de los grafos o redes, como el diámetro, la densidad, la transitividad y reciprocidad. Luego se mencionan varias formas de modelar el comportamiento futuro de una red social, profundizando en un método basado en matrices estocásticas y procesos de Markov.

Estos conceptos son aplicados al análisis de una red de menciones construida a partir de un conjunto de tuits obtenidos directamente de la web de la red social Twitter y clasificados en varios intervalos de tiempo para analizar su dinámica. Los tuits obtenidos corresponden a la temática Kin Jong-un a propósito de la Cumbre Intercoreana realizada el 27 de abril de 2018.

Se concluye con varios resultados de análisis de la red, principalmente con la identificación de usuarios clave en la red, tanto en los periodos de tiempo observados, así como en los futuros, buscando ser un aporte en cuanto a las posibilidades predictivas que presentan los problemas de menciones.

2 Las Redes Sociales

Una red social, en general, es un mecanismo en el que confluyen las interacciones que realizan usuarios o actores, y que ha cobrado capital importancia gracias al uso masivo de internet. En esencia, las redes permiten a sus usuarios estar al tanto de la vida de sus familiares, conocidos y amigos; no obstante, muchos personajes públicos lo usan para expresar ideas o información, pudiendo inclusive determinar tendencias, reunir expertos a partir de determinadas ideas, e incluso establecer relaciones comerciales con sus miembros.

La gran cantidad de información que genera Twitter, y su fácil accesibilidad ha permitido que el mundo académico vuelque sus esfuerzos al entendimiento de la formación, estructura y dinámica de las redes que se pueden conformar, siendo un terreno fértil también las aplicaciones relativas a la minería de texto, el análisis de sentimientos, etc. (Smith *et al.*, 2014).

Twitter posee posibilidades para acceder a los tuits a través de Twitter Search API (Application Programming Interface) que es un servicio gratuito que permite enviar búsquedas a Twitter y obtener los tuits que coincidan con los criterios de búsqueda, con determinadas limitaciones. También existen revendedores oficiales de estos datos que pueden ofrecer los datos de Twitter de forma más completa como Gnip. Además, existen otros proveedores que ofrecen servicios analíticos a partir de la información de Twitter como Topsy o Hootsuite.

La vía gratuita, a través del servicio API Twitter tiene diferentes métodos de acceso a los tuits (Thelwall, 2015); sin embargo, a pesar de que actualmente, por motivos técnicos, no se permite un acceso amplio a la totalidad de tuits de un usuario o temas, existe la

información suficiente para entender y estudiar redes reales. Adicionalmente, Twitter no restringe el uso de estos datos para investigación siempre que cumpla con lo estipulado en los acuerdos y políticas para desarrolladores[1] vigentes al momento de la elaboración de este trabajo. Evidentemente existe el riesgo de que Twitter pueda en un futuro cambiar sus políticas de acceso libre y gratuito a los tuits.

Para la presente aplicación se utiliza el paquete *twitte*R para R, el cual permite una accesibilidad a Twitter API para obtener información de los tuits y colocarla en variables u objetos de R. El proceso de obtención de tuits por este método requiere de la creación de un usuario desarrollador de Twitter y que las descargas de tuits se realicen a través de un proceso de autenticación.

3 Análisis de Redes Sociales

Para autores como Otte y Rousseau (2002) el Análisis de Redes Sociales (SNA, por sus siglas en inglés) no es una teoría formal de la sociología sino más bien una estrategia para investigar estructuras sociales. Algunos autores consideran que el objetivo principal del SNA es detectar e interpretar patrones de vínculos sociales entre sus actores (De Nooy *et al.*, 2005).

Se había mencionado que la característica principal de esta teoría es que permite analizar los comportamientos sociales de un actor en su contexto, lo que representa una diferencia total con los enfoques sociales y de análisis de datos *individualistas* ya que el objetivo principal constituye el análisis de las relaciones entre los actores. Otro aspecto importante del SNA es que estudia cómo las regularidades estructurales que se puedan detectar influyen sobre los individuos. De hecho, el análisis de las redes sociales puede tener dos formas generales: el análisis "ego" donde se estudian las relaciones de una persona y el análisis global de las redes. En particular las redes sociales presentes en internet representan una red de gran escala donde los conceptos de esta teoría pueden ser ensayados.

3.1 Elementos de la Teoría de Grafos

Un **grafo** o red, no es más que un conjunto de vértices unidos por arcos, los cuales también son llamados nodos y links respectivamente.

Matemáticamente un grafo es un par ordenado $G = (V, E)$ donde V es el conjunto de vértices y E es el conjunto de arcos que unen estos vértices, asumiendo que V es un conjunto finito. Los arcos pueden ser *únicos* entre un par de vértices y si hubieran más arcos entre el mismo par se denominan *multiarcos*. Pueden existir además arcos de un vértice a sí mismo, estos arcos se denominan *autoarcos*.

[1] Twitter (2018), Recuperado de: `https://developer.twitter.com/en/developer-terms/agreement-and-policy` (revisado el 11-06-2018).

De esta manera también se puede hablar de unas primeras clasificaciones de grafos: los *grafos simples* que no contienen ni multiarcos, ni autoarcos y los *grafos no simples* que sí tienen multiarcos y arcos. La figura 1 muestra un ejemplo:

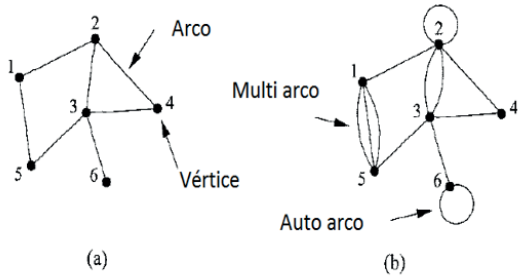

Figura 1: Ejemplos de un grafo simple (a) y uno no simple (b)
Fuente: Newman, 2010

3.1.1 Matrices de Adyacencia

En términos matemáticos hay diferentes maneras de representar una red, particularmente en forma de una matriz, considerando que tenemos n vértices denotados por números enteros únicos para cada vértice como en la Figura 1 y refiriéndonos a los arcos entre dos vértices $i, j \epsilon V$ por $(i, j) \epsilon E$ entonces se puede definir la matriz de adyacencia como sigue:

La *matriz de adyacencia* de un *grafo simple* es la matriz A con elementos $A_{i,j}$ tales que:

$$A_{i,j} = \begin{cases} 1 \text{ si hay un arco entre i y j} \\ 0 \text{ en caso contrario} \end{cases}$$

De manera que el ejemplo de la Figura 1(a) la matriz de adyacencia resultante es:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Es posible también representar gráficos con *multiarcos y autoarcos* usando una matriz de adyacencia colocando en $A_{i,j}$ de la matriz la multiplicidad de los arcos. Por otra parte, los autoarcos pueden representarse colocando para $A_{i,i} = 2$, esto, porque el arco tiene dos extremos. Así la matriz de adyacencia para la Figura 1(b) es:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{pmatrix}$$

3.1.2 Grafo Ponderado

En las redes sociales se pueden representar también la frecuencia del contacto o valorar un vínculo entre actores, lo que se puede representar en una matriz de adyacencia poniendo a los elementos de la matriz de adyacencia valores en el conjunto de los números reales. Si bien las valoraciones suelen ser positivas se pueden tener también valoraciones negativas, por ejemplo en una interpretación de un vínculo entre dos actores en función de la falta de cordialidad.

Un ejemplo de una matriz de adyacencia de una red ponderada es el siguiente:

$$A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 0{,}5 \\ 1 & 0{,}5 & 0 \end{pmatrix}$$

En A se ve que la relación entre los vértices 1 y 2 está ponderada el doble que la relación entre los vértices 1 y 3.

3.1.3 Grafos Dirigidos

Un *grafo dirigido* o *digrafo* es un grafo cuyos arcos tienen una dirección de un vértice a otro. Los arcos se denominan arcos dirigidos y se los representa con líneas con una flecha en el sentido requerido como se aprecia en la Figura 2.

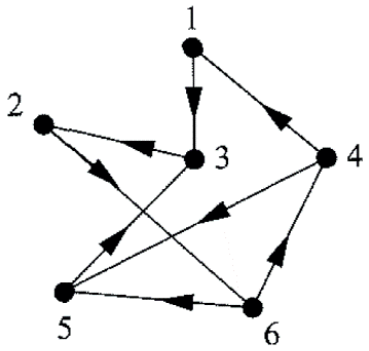

Figura 2: Ejemplo de grafo dirigido
Fuente: Newman, 2010

En lo relativo a la matriz de adyacencia para un grafo dirigido, por convención (Scott y Carrington, 2012) se colocarán en las filas al vértice o nodo que envía el link y en las columnas al receptor de manera que:

$$A_{i,j} = \begin{cases} 1 \text{ si hay un arco desde i hacia j} \\ 0 \text{ en caso contrario} \end{cases}$$

Resultando la matriz no simétrica:

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Al igual que en el caso de los grafos no dirigidos, un grafo dirigido puede tener multiarcos o autoarcos, que se representarán en la matriz como elementos de la matriz mayores que 1 o como valores mayores que cero en la diagonal.

3.1.4 Caminos en un grafo

Si es posible llegar de un vértice i a un vértice j recorriendo los arcos de un grafo cualquiera se dice que existe un *camino* de i a j, y si se atravesaron k arcos se dice que se trata de un camino de longitud k de i hacia j. Sea A la matriz de adyacencia de este grafo y A^k su producto k veces consecutivas se puede apreciar que el elemento $A^k_{i,j}$ representa el número de caminos de longitud k que van de i hacia j (Remus). Una prueba de este hecho puede ser observada en (Glickenstein, 2008).

Para el ejemplo de la Figura 2 en el caso $k = 2$ se tiene que:

$$A^2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Y, por ejemplo, se ve que del vértice 1 al 3 se tiene 1 camino de longitud y del vértice 4 al 2 se tienen dos caminos de longitud 2.

Un grafo se llama *conectado* si para todo par de vértices i, j hay un camino conectado de i hasta j o de j hasta i. Por otra parte, un grafo se llama *fuertemente conectado* si empezando en un nodo i se puede alcanzar cualquier nodo j caminando por sus arcos. Mayores referencias acerca de los tipos de grafos y sus propiedades pueden ser consultadas en Newman (2010).

3.1.5 Medidas y métricas

La mayoría de las medidas para entender la estructura de una red han sido desarrolladas en el contexto de las ciencias sociales, pero actualmente han sido adoptadas por otras disciplinas como la computación, la física, la biología, etc., constituyéndose en una caja de herramientas de la que buscan valerse profesionales de muchas ramas.

a) **Centralidad por grados**

La centralidad por grados es la medida más sencilla de las medidas de centralidad, pues corresponde a los grados de un vértice, esto es, el número de arcos conectados a éste. En específico para grafos dirigidos, los vértices presentan *grados de entrada y grados de salida* que a pesar de ser medidas muy sencillas pueden resultar muy útiles dependiendo del contexto en el que se planteen estas medidas de centralidad. Por ejemplo, en el contexto de citas bibliográficas, el número de citas o grado de entrada que tiene un artículo serían una buena medida de lo influyente que resulta un autor.

b) **Centralidad por valores propios**

Una extensión natural de la centralidad por grados es la *centralidad por valores propios*, pues al calcular los grados de un vértice valoramos una vecindad de éste; no obstante, no todos los vecinos son iguales pues las conexiones más relevantes serán aquellas que sean con vértices en sí mismo importantes. Este es el concepto que está detrás de la *centralidad*

por valores propios, pues en lugar de valorar a un vértice por sus vecinos directos, da a cada vértice una valoración proporcional a la suma de las valoraciones de sus vecinos. De la siguiente manera:

Sea x_i la centralidad de un vértice i, para arrancar se hace $x_i = 1$ para todos los $i = 1, \ldots, n$ siendo n el número de vértices. Luego se define,

$$x'_i = \sum_j A_{ij} X_j$$

Siendo los A_{ij} elementos de la matriz de adyacencia y en esencia define la suma de las centralidades de los vecinos. Esta idea puede ser escrita en notación vectorial como:

$$x' = Ax$$

Así una mejor aproximación puede ser obtenida al repetir este cálculo t veces de manera que:

$$x(t) = A^t x(0)$$

siendo $x(0)$ como una combinación lineal apropiada de los vectores V_i propios de la matriz A, así:

$$x(0) = \sum_t c_i V_i$$

Luego, notando κ_i a los valores propios de A y $\kappa_1 = max_i \kappa_i$ se tiene:

$$x(t) = A^t \sum_i c_i V_i = \sum_i c_i \kappa_i^t V_i = \kappa_1^t \sum_i c_i [\frac{\kappa_i}{\kappa_1}]^t V_i$$

Aquí, como cada elemento de $\frac{\kappa_i}{\kappa_1} < 1$ siempre que $i \neq 1$ se puede mostrar que $x(t) \to c_1 \kappa_1^t V_1$ que según se detalla en Newman (2010) será equivalente a decir que la centralidad x satisface:

$$Ax = \kappa_1 x$$

Cabe mencionar que la centralidad por valores propios es una de las medidas denominadas *de influencia* de los nodos; no obstante, se pueden mencionar también otras medidas que se interpretan como la influencia como la centralidad de Hubbell, Katz, Taylor (Introduction to social network methods). Es necesario tener en cuenta la condición necesaria y suficiente para la existencia y unicidad de x es la que A corresponda a un grafo fuertemente conectado.

c) **Centralidad por Intermediación**

El concepto de centralidad por intermediación, denominado en inglés *beetweness centrality*, mide qué tan frecuentemente un vértice aparece en los caminos entre otros vértices. La idea de esta medida inicia suponiendo que en la red hay alguna información o mensaje fluyendo a través de la red de una persona a otra y que en un inicio todo par de vértices conectados por un camino en la red intercambian mensajes con igual probabilidad por unidad de tiempo y que estos mensajes siempre tomarán el camino más corto[2] a través de la red o algún camino aleatorio si hubieran varios candidatos a caminos más cortos, de esta manera, al pasar una misma tasa de mensajes, en promedio el número que hubieran pasado por este vértice es proporcional al número de caminos más cortos que pasen por él; precisamente a este número se le denomina *centralidad por intermediación.*

Aquellos vértices con alta *intermediación* pueden ser considerados importantes en contexto de la red en virtud de que controlan la información que pasa hacia otros vértices y removerlos de la red evidentemente provocará desconexiones en la comunicación de una cantidad importante de elementos en la red. Evidentemente los supuestos discutidos no reflejan un problema del mundo real principalmente porque los vértices no intercambian información a la misma tasa, pero se considera que es una aproximación aceptable (Newman, 2010).

Matemáticamente, sea n_{st}^i tal que es igual a 1 si el vértice i esta en el camino más corto entre los vértices s y t; y, 0 en caso contrario, de manera que la centralidad por intermediación para el vértice i-ésimo viene dado por la expresión:

$$x_i = \sum_{st} n_{st}^i$$

Esta expresión es ajustada para el caso en que múltiples caminos más cortos de s a t pasen a través de i, introduciendo este número g_{st}, así:

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}}$$

En este último indicador se asume la convención de que $\frac{n_{st}^i}{g_{st}} = 0$ si n_{st}^i y g_{st} son cero. En particular el paquete *igraph* de R adiciona las condiciones $i \neq t$ y $s \neq t$ (inside-R).

c) **Cercanía**

Otra medida, más natural es la centralidad por cercanía de un vértice, que se basa en un promedio de las distancias de este vértice a otros vértices, donde una distancia $d_{i,j}$ es la longitud del camino más corto del vértice i al vértice j, luego este promedio excluyendo al caso $i \neq j$ será

[2]El camino o caminos más cortos entre dos vértices se denominan *caminos geodésicos* (Newman, 2010).

$$l_i = \frac{1}{n-1} \sum_{j \neq i} d_{ij}$$

Posteriormente, para que esta medida de centralidad sea consistente, es decir que a mayor valor refleje mayor cercanía, se calcula el indicador:

$$c_i = \frac{1}{l_i}$$

En el caso de que no existan caminos entre dos vértices *igraph* de R utiliza en lugar de la distancia, el número total de vértices (inside-R).

3.2 Modelado Estadístico de Redes en el Tiempo.

El análisis de redes en el tiempo tiene varios enfoques, algunos de los más relevantes se describen a continuación.

Uno de los enfoques busca relacionar las representaciones de las matrices de adyacencia de los grafos a los procesos estocásticos a través de la matriz estocástica. Este enfoque, descrito en Glickenstein (2008), se basa en el hecho de que dada la matriz de adyacencia A de un grafo $G(V, E)$, la entrada (i, j) de A^n con $n \geq 1$ representa el número de caminos diferentes entre los vértices v_i, v_j de longitud n en G, donde $v_i, v_j \, \epsilon V = \{v_1, v_2, \ldots, v_n\}$, el conjunto de vértices de G. Lo anterior, permite construir una matriz estocástica relacionando el problema analizado a partir de una interpretación que lleve a una compatibilidad con las cadenas de Markov homogéneas discretas. Este enfoque es el que se ensayará en este trabajo, por lo que se lo profundizará más adelante.

Para analizar la **dependencia de la red a través del tiempo**, se puede recurrir al ámbito estocástico a través del análisis de cadenas de Markov, asumiendo que las matrices de adyacencia constituyen un panel de datos $A(t_1), A(t_2), \ldots, A(t_n)$ de n observaciones consecutivas y que conforma un proceso de Markov discreto. Inclusive si se tiene que un proceso es tal que $\{A(t)|t_1 \leq t \leq t_2\}$ se podría hablar de un proceso de Markov continuo. Esto, con la suposición principal de que en las cadenas de Markov el estado futuro $(t + 1)$ solamente dependerá de su estado presente (t), lo cual implica suponer que independencia entre los links del futuro y anteriores al presente.

Esta aproximación es utilizada en el problema del Ranking de Páginas Web, aplicado con alguna variación por Google y presentado en (Remus) en su trabajo "describiendo el álgebra detrás de Google Search".

Por otro lado, el análisis de **dependencia analizada a través de los links** puede salvar el supuesto de independencia entre los links mencionado en una cadena de Markov; para ello se enfoca el problema en los pares de links o diadas $(X_{ij}(t), X_{ji}(t))$. Así, para capturar la dependencia deseada se formula un modelo donde un par aleatorio (i, j) es escogido para analizar la probabilidad de conformar un link de modo que X_{ij} cambie de 0 a 1 o de eliminar

el link cambiando X_{ij} de 1 a 0, y definiendo una función basada en estadísticos de la red X en cada tiempo sobre el número actual de links en la red $L(X)$, el número de diadas recíprocas $M(X)$, el número de tripletas transitivas $T(X)$ y la varianza de los grados de entrada $V_{in}(X)$, de la siguiente manera:

$$f(x;\beta) = \beta_1 L(x) + \beta_2 M(x) + \beta_3 T(x) + \beta_4 V_{in}(x)$$

Esta función para cada tiempo mide las tendencias de estos estadísticos para posteriormente incorporarlos en el cálculo de la probabilidad de que exista una red con el link entre i y j.

La función de probabilidad mencionada mide define a partir de una red x dos redes: $x^{(ij+)}$ que tiene el link ij y $x^{(ij-)}$ que no tiene el link (i,j).

$$p_{i,j} = \frac{exp(f(x^{(ij+)};\beta))}{exp(f(x^{(ij+)};\beta)) + exp(f(x^{(ij-)};\beta))}$$

Este enfoque incorpora en el modelo aspectos estructurales de la red.

Otra aproximación sobre la formación de las redes sociales, puede apreciarse en el campo de la investigación econométrica según explican Bramoullé y Fortin en (The Econometrics of Social Networks, 2009), donde se han realizado trabajos que permiten enfocar la creación de las relaciones o links entre actores de una red social a través de regresiones que tienen como variable respuesta los links formados entre actores y que tienen por variables explicativas a las características de relación misma. A este tipo de modelamiento se le denomina pairwise regression y está definido por la siguiente expresión:

$$Y_{ij} = X_{ij}\zeta + \varepsilon_{ij}$$

$$g_{ij} = 1 \text{ si } Y_{ij} \geq 0 \text{ y } 0 \text{ si } Y_{ij} < 0$$

donde Y_{ij} es la propensión para formar el link ij, X_{ij} es un vector de características del link g_{ij}, ε_{ij} un el error relativo al link y ζ es un vector de coeficientes que se a estimarse con alguna técnica de regresión. Los fundamentos teóricos de este tipo de modelo contemplan casos para cuando los links son efectuados por mutuo acuerdo o por decisión unilateral, siendo consistente la idea de la decisión unilateral con la idea de un grafo dirigido.

Luego, el **enfoque basado en los actores** está más relacionado con la idea de que los cambios en los links son iniciados por los actores y se formula un modelo como si los actores tuvieran el control de sus links. Así, la especificación del modelo emplea lo que se denomina un función ratio $\lambda_i = (x;\alpha)$ dependiente del actor i y del estado de la red actual x, que indica la frecuencia por unidad de tiempo con la que el actor tiene la oportunidad de cambiar un link y una función $f_i(x;\beta)$ similar a la del enfoque basado en links que se interpreta como un medida de qué tan atractiva resulta el estado de una red para el actor i. Así la probabilidad

de que un cambio de un estado x a un estado $x^{(ij\pm)}$, que es un estado idéntico a x pero contrario en la existencia link (i, j), es:

$$\pi_{i,j} = \frac{exp(f_i(x^{(ij\pm)}; \beta))}{\Sigma_{h=1}^{n} exp(f_i(x^{(ih\pm)}; \beta))}$$

El detalle de estas técnicas puede ser observado en Scott y Carrington (2012).

3.2.1 Procesos de Markov

Según se expone en Steward (2009) un *proceso estocástico* se define como una familia de variables aleatorias $\{X_t, t\epsilon T\}$. El índice t representa usualmente al tiempo, de manera que $X(t)$ denota el valor de una variable aleatoria en este tiempo. T es el conjunto de índices y es un subconjunto de los números Reales $(-\infty, +\infty)$. Si T es discreto, por ejemplo, $T = \{0, 1, 2, \ldots\}$ se dirá que el proceso estocástico es *discreto en el tiempo*, por otra parte si T es continuo, por ejemplo, $T = \{t|0 \leq t < +\infty\}$ el proceso estocástico será *continuo en el tiempo*. Los valores asumidos por las variables $X(t)$ serán denominadas *estados*. Si el espacio de estados es discreto, el proceso se suele denominar *cadena* y a los estados se los identifica con el conjunto de los números naturales.

Un proceso estocástico se dice *estacionario* si su distribución adjunta es invariante a cambios en el tiempo, esto es:

$$\begin{aligned} P(X(t_1) \leq x_1, X(t_2) \leq x_2, \ldots, X(t_n) \leq x_n) = P(X(t_1 + \alpha) & \\ \leq x_1, X(t_2 + \alpha) & \\ \leq x_2, \ldots, X(t_n + \alpha) \leq x_n) & \end{aligned}$$

para todo n, todo t_i y x_i con $i = 1, 2, \ldots, n$.

Una *cadena Markov* es una *cadena* a tiempo discreto $\{X_n, n = 0, 1, 2, \ldots\}$ es un proceso estocástico que satisface la siguiente relación, denominada la *propiedad de Markov*.

$$P(X_{n+1} = x_{n+1}|X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_{n+1} = x_{n+1}|X_n = x_n)$$

para todos los estados x_n.

Notar que en la cadena de Markov la probabilidad condicional conjunta de los estados indica que un estado x_{n+1} depende solamente del estado x_n, a esta propiedad se le denomina falta de memoria.

La probabilidad condicional $P(X_{n+1} = x_i|X_n = x_j)$, notada por facilidad $p_{ij} = P(X_{n+1} = j|X_n = i)$ se denomina la *probabilidad de transición* en un paso del estado $i = x_i$ al $j = x_j$.

La matriz $P(n)$, formada de colocar $p_{ij}(n)$ en la fila i y en la columna j se denomina la *matriz de probabilidades de transición* y es:

$$P(n) = \begin{pmatrix} p_{00} & \dots & P_{0n} \\ \vdots & \ddots & \vdots \\ p_{n0} & \dots & p_{nn} \end{pmatrix}$$

Notar que en esta matriz $0 \leq p_{ij} \leq 1$ y para todo i se tiene $\Sigma_{j=1}^{n} p_{ij}(n) = 1$. A una matriz con estas propiedades se la denomina *matriz de Markov o matriz estocástica*.

Las cadenas de Markov pueden ser expresadas a través de los denominados *diagramas de transición* que son grafos dirigidos que tienen como vértices a los estados del proceso y como links a las probabilidades de pasar entre éstos.

Un proceso estocástico se dice homogéneo si para todo par de estados i y j se tiene que:

$$p_{ij} = P(X_{n+1} = j | X_n = i) = P(X_{n+m+1} = j | X_{n+m} = i)$$

Para una cadena de Markov discreta en el tiempo, homogénea, se puede mostrar que para cualquier $n = 0, 1, 2, \dots,$

$$p_{i,j}^{(m)} = P(X_m = j | X_0 = i) = \Sigma_k p_{kj}^{m-l} p_{ik}^{l} \quad \text{para} \quad 0 < l < m$$

Esta última expresión es conocida como la ecuación de *Kolmogorov-Chapman* que puede ser escrita en notación matricial como:

$$P^{(m)} = P^{(l)} P^{(m-l)}$$

En particular para $l = 1$,

$$P^{(m)} = P P^{(m-1)}$$

De aquí la matriz de transición para m pasos es obtenida de la multiplicación de la matriz de transición de un paso $(m-1)$ veces consecutivas pues $P^{(m)} = P^m$. Se debe definir también $P^{(0)} = I$.

Sea $\pi_i^{(0)}$ la probabilidad de que una cadena de Markov empiece por el estado i y el vector $\pi^{(0)}$ de probabilidades inicial de que un proceso empiece por un estado i se tiene que:

$$\pi^{(1)} = \pi^{(0)} P$$

Y sucesivamente, en este escenario homogéneo en el tiempo se tiene que:

$$\pi^{(n)} = \pi^{(n-1)} P = \pi^{(0)} P^n$$

La distribución de probabilidad al límite π puede ser encontrada si existe el límite

$$\lim_{n \to \infty} P^{(n)} = \lim_{n \to \infty} P^n$$

Y calculando, $\pi = \pi(0) \lim_{n \to \infty} P^n$

Una distribución de probabilidad al límite π se denomina *distribución de estado estable*, si converge independientemente del vector de distribución inicial $\pi(0)$ a un vector cuyos componentes son estrictamente positivos ($\pi_i > 0$). Si la distribución de estado existe, es única. Notar que también el vector de estado estable satisface en particular $\pi = \pi P$.

4 Aplicación SNA a una red evolutiva de menciones en Twitter

Una red de menciones entre usuarios de Twitter puede ser vista desde un enfoque estocástico, suponiendo que una información general o particular transmitida por la red sobre un tema específico puede ser modelada mediante matrices estocásticas y así analizar su distribución en distintos periodos de tiempo a través de menciones.

El modelado en base a matrices estocásticas de una red de menciones tiene por detrás las ideas de las cadenas de Markov y de la propiedad de Markov que implica que la transmisión de un mensaje a otro individuo a través menciones en la red Twitter depende únicamente del individuo actual y de la probabilidad que éste tenga de transmitirlo a otro individuo, implicando que sin importar el origen ni el tipo de mensaje, el usuario actual lo pasará mediante una mención.

La forma utilizada para construir un grafo a partir de una lista de tuits es necesario realizar una extracción de los usuarios @mencionados en cada tuit de un @usuario y colocarlos en una lista, por ejemplo, en los tuits:

- *"@deportesTV: @barcelona venció al @realmadrid gracias a un gol de último minuto de @messi"*
- *"@messi: gracias a toda la hinchada del @barcelona por la victoria de hoy"*
- *"@deportesTV: RT*[3] *@ messi: gracias a toda la hinchada del @barcelona por la victoria de hoy"*

La manera de transformar estas menciones de Twitter en una lista es extrayendo los usuarios y los mencionados de manera que resulte en:

[3]RT: significa que un usuario re-tuiteó o re-publicó un mensaje de otro usuario.

Tabla 1: Ejemplo de tuits transformados en lista

Usuario	Mencionado
@deportesTV	*@barcelona*
@deportesTV	*@realmadrid*
@deportesTV	*@messi*
@messi	*@barcelona*
@deportesTV	*@messi*
@deportesTV	*@barcelona*

Esta misma lista luego debe ser agregada para representar a un grafo con links ponderados:

Tabla 2: Ejemplo de tuits como un grafo en formato lista

Usuario	Mencionado	Peso
@deportesTV	*@barcelona*	*2*
@deportesTV	*@realmadrid*	*1*
@deportesTV	*@messi*	*2*
@messi	*@barcelona*	*1*

Finalmente, esta lista representará al siguiente grafo:

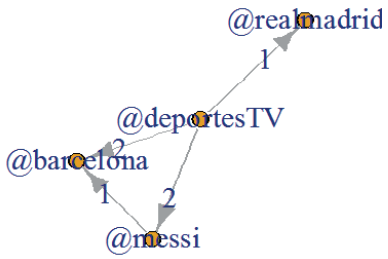

Figura 3: Ejemplo de grafo a partir de menciones

En notación de grafos, el ejemplo puede presentarse como el grafo $G(V, E)$ donde $V = \{@deportesTV, @barcelona, @realmadrid, @messi\}$ y E se conformaría son los pares descritos en la Tabla 1. De esta manera se tiene un grafo cabalmente establecido.

Para establecer la matriz estocástica se parte del número de menciones. La manera propuesta en este trabajo es tomando las menciones promedio realizadas entre cada par de individuos i, j en cada uno de los periodos de tiempo en los que se observe el gráfico. Así, se establecerá la relación entre la matriz de transición P asociada a un grafo G:

$$P = \begin{pmatrix} p_{11} & \dots & P_{0n} \\ \vdots & \ddots & \vdots \\ p_{n0} & \dots & p_{nn} \end{pmatrix}$$

y la matriz de transición M:

$$M = \begin{pmatrix} \overline{m}_{11}/\Sigma_j \overline{m}_{0j} & \dots & \overline{m}_{1n}/\Sigma_j \overline{m}_{1j} \\ \vdots & \ddots & \vdots \\ \overline{m}_{n1}/\Sigma_j \overline{m}_{nj} & \dots & \overline{m}_{nn}/\Sigma_j \overline{m}_{nj} \end{pmatrix}$$

donde $\overline{m}_{ij}$ es el promedio en las menciones entre los usuarios i y j, si un usuario k no ha realizado menciones se imputa[4] $\overline{m}_{kk} = 1$. Finalmente, para cualquier probabilidad de distribución de menciones iniciales π_0 se puede que ver la distribución de menciones a tiempos posteriores.

$$\pi^{(t+1)} = \pi^{(t)} P = \pi^{(0)} P^{t+1}$$

Este uso de las matrices nos permitirá, de manera predictiva, identificar la probabilidad de cada usuario de ser mencionado en periodos futuros.

Notar que en una red de menciones no necesariamente tenemos garantizada la existencia de un estado estable, no obstante, en caso de tenerlo, este representaría un indicador o ranking de probabilidades de mención en el tiempo, muy similar a la medida de influencia o índice de centralidad.

4.1 Ejemplo de análisis de una red de menciones

4.1.1 Descripción de la Red

Mediante una conexión al servicio API Twitter (Twitter Developers) que utiliza un mecanismo de autorización OAuth y usando la cuenta *@rolandomantilla* se realizó una descarga de 10.000 tuits el 27/04/2018, en idioma inglés con el método "search" del API de Twitter. Estos tuits son filtrados para contener a la palabra "Kim Jong-un" con la finalidad de tener un conjunto de tuits relativos la cumbre histórica del 26-04-2018 que podría llevar a un acuerdo de paz entre Korea de Norte y Korea de Sur. Mediante la nube palabras generada por el paquete *tm* de R se puede tener una idea de lo discutido según las palabras que se hablan con más frecuencia en los tuits (ver Figura 4). En este caso, podemos ver que se discute también sobre el líder de Corea del Sur, Moon Jae-in además de términos como la desnuclearización de la península motivados en la reunión en la frontera de ambos países.

[4]Esta condición supone un estado .ªbsolvente", en el cual ya no hay un cambio de estado, i.e., menciones a otros usuarios. Este problema es conocido como "dangling node" en el problema del ranking de páginas de Google, pero es resuelto introduciendo un salto aleatorio a otra página.

El conjunto de tuits registra horas de extracción de 12h17 y 15h33, este intervalo es dividido en 10 intervalos iguales para estudiar su evolución.

Figura 4: Nube de palabras

En la siguiente tabla se presenta una descripción de la composición de la red en cada intervalo. Las estadísticas presentadas son estimadas con el paquete igraph de R.

Tabla 3: Análisis descriptivo

Intervalo	Número de Usuarios	Número de links	Número de menciones	Menciones por Usuario
1	993	883	903	0,90936556
2	1.896	1.768	1.807	0,95305907
3	2.837	2.698	2.746	0,96792386
4	3.754	3.628	3.688	0,98241875
5	4.642	4.534	4.600	0,99095218
6	5.505	5.464	5.546	1,00744777
7	6.375	6.408	6.525	1,02352941
8	7.254	7.344	7.476	1,0306038
9	8.105	8.289	8.463	1,04417027
10	8.980	9.237	9.439	1,05111359

Como punto interesante se aprecia en la Tabla 3 que un inicio en el intervalo 1, las hay en promedio 0,91 menciones por usuario, indicando que puede haber auto menciones, en el tiempo 10 se observa que los usuarios mencionan cada vez más a otros usuarios.

Es interesante también observar que estas menciones se realizan en comunidades o clústeres (Csárdi y Nepuzs, 2006), este número de clústeres evoluciona en el tiempo entre 142 y 404 (ver Figura 5), pero al ser el gráfico cóncavo se puede entender que cada vez los clústeres son más grandes, en lugar de que los clústeres sean aislados.

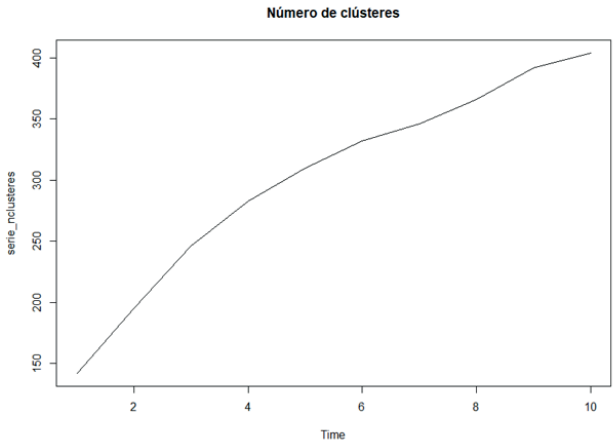

Figura 5: Clústers

Por otra parte, al analizar el diámetro[5] del grafo, se observa que no presenta una mayor variación al empezar en 4 y al terminar en 5. Además, se presenta la *densidad del grafo*[6] en la Figura 6, estos son los links establecidos sobre los links posibles. La tendencia decreciente puede leerse como una tendencia a no conectarse más en el tiempo, sino a la aparición de menciones aisladas.

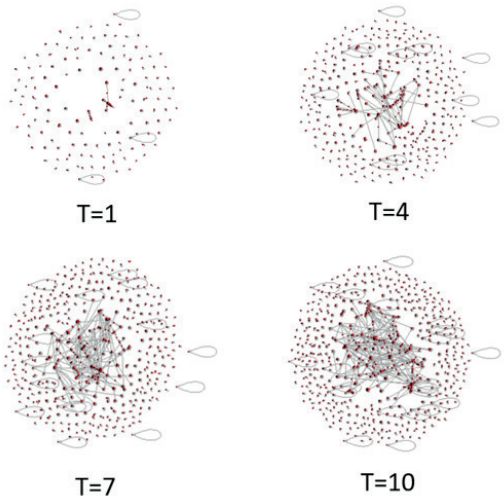

Figura 6: Evolución de la red

[5] `http://www.inside-r.org/packages/cran/igraph/docs/farthest.nodes`
[6] `http://www.inside-r.org/packages/cran/igraph/docs/graph.density`

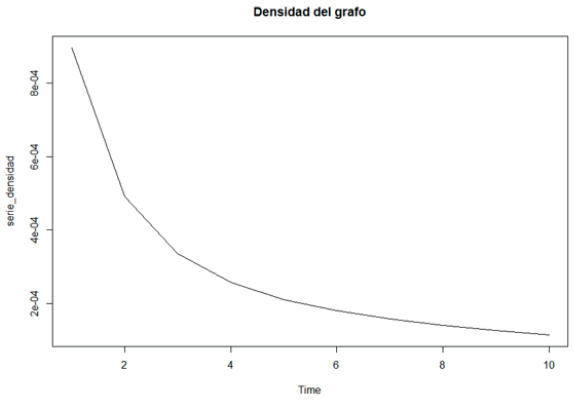

Figura 7: Densidad del grafo

Luego, se puede apreciar la transitividad[7] que mide la probabilidad de que los vértices adyacentes un vértice común estén a su vez conectados entre sí, observando la Figura 8, se ve un bajo nivel de que los usuarios comunes logren conectar a los otros usuarios.

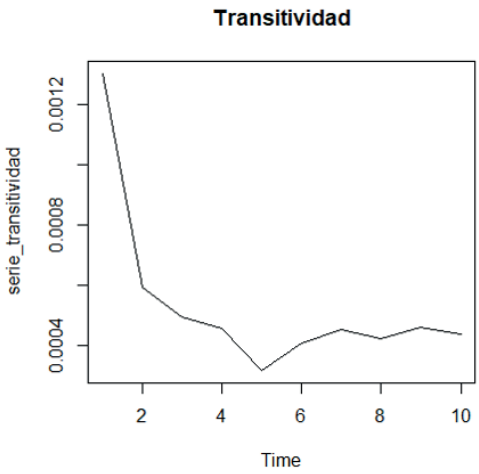

Figura 8: Transitividad

4.1.2 Medidas de centralidad

Las medidas de centralidad que se analizan son básicamente de carácter individual, pretendiendo identificar a los elementos más relevantes en la red de acuerdo con estas medidas para luego ver su evolución en el tiempo.

[7] `http://www.inside-r.org/packages/cran/igraph/docs/transitivity`

a) **Centralidad por grados**

Los grados de salida representan a la cantidad de usuarios a los que ha mencionado un usuario y los grados de entrada representa al número de usuarios que han mencionado a un usuario.

Usuario	Grados de Entrada		Usuario	Grados de Salida
@eugenegu	876		**@nannugent**	12
@malonebarry	771		**@JeffGrose59**	11
@Telegraph	588		**@OLibcrusher**	11
@BBCBreaking	493		**@laneylane25**	10
@Kasparov63	471		**@dleiben_debby**	10
@StephenMilIer	407		**@MichelleRubio68**	10
@CaliConsrvative	383		**@Moxi_Mimi**	9
@1776Stonewall	267		**@EasyMode243**	9
@AJEnglish	262		**@Jwally54**	8
@MothershipSG	250		**@iambo**	8

La evolución de los grados de entrada para @eugenegu se aprecia en la Figura 9, observando que apareció en el intervalo 6 y seguramente debe su alto grado de menciones a que genera muchos retuits.

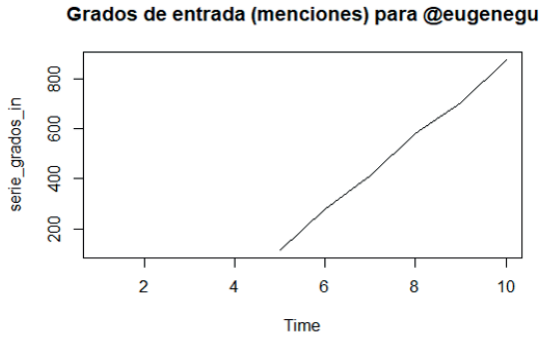

Figura 9: Evolución grados de entrada

En el caso del usuario @nannugent, se ve que aparece en el periodo 6 y realiza 12 menciones.

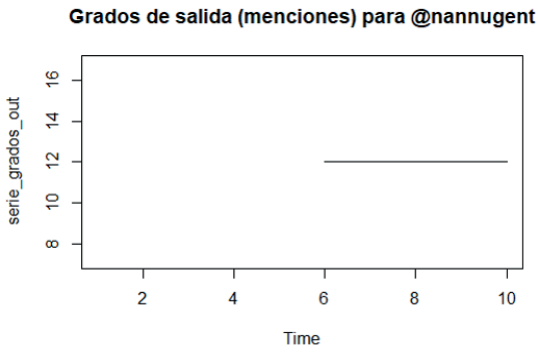

Figura 10: Evolución grados de salida

b) **Centralidad por valores propios**

La centralidad por valores propios es a menudo utilizada como una medida de influencia en la red. La manera en la que se explica su cálculo en el paquete *igraph* se asemeja al Page Rank de Google. En el tiempo 10 se tiene que el más influyente de la red aparece @malonebarry quien es el editor en de Al Jazeera on line. Observar que en onceavo lugar aparece @realDonaldTrump (Tabla 4).

Tabla 4: Centralidad por valores propios

Usuario	**Influencia**
@malonebarry	1,00E+00
@Telegraph	8,30E-01
@BBCBreaking	1,55E-01
@JeffreyGuterman	7,64E-02
@Kasparov63	5,94E-02
@orangeseahorse	8,15E-16
@eugenegu	4,24E-16
@bdnews24	1,54E-10
@JustSchmeltzer	6,01E-18
@WashTimes	2,75E-18

c) **Intermediación**

La intermediación identifica a los usuarios que si desaparecieran causarían mayor desconexión en la red. Nuevamente el mejor intermediario en la red es @eugenegu.

Tabla 5: Intermediación

Usuario	Intermediación
@eugenegu	870
@JustSchmeltzer	68
@trtworld	13
@EasyMode243	12
@VICE	10
@AmericaNewsroom	10
@atanasi_	8
@JNilssonWright	7
@ResistanceZone	5
@BretBaier	3

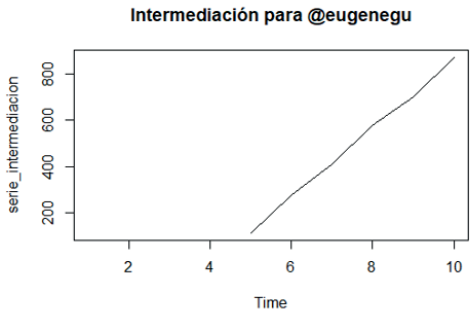

Figura 11: Evolución de la intermediación

Cercanía

La centralidad por cercanía puede ser por menciones que le hagan (entrada) o por menciones que realice (salida). En este caso analizaremos el caso de cercanía por entrada, que indica cuáles usuarios pueden ser mencionados de manera más directa. Así @JeffreyGuterman es el usuario más cercano al resto.

Usuario	**Cercanía**
@JeffreyGuterman	1,38E-02
@eugenegu	1,37E-02
@malonebarry	1,36E-02
@Telegraph	1,33E-02
@Kasparov63	1,31E-02
@BBCBreaking	1,31E-02
@CaliConsrvative	1,30E-02
@StephenMillIer	1,29E-02
@1776Stonewall	1,28E-02
@AJEnglish	1,28E-02

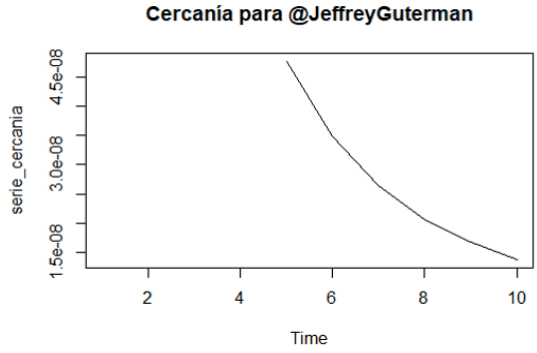

Figura 12: Evolución de la cercanía

4.1.3 Análisis Predictivo

En particular, se aplicará la metodología de la matriz estocástica (con pesos promedio) para predecir las menciones que los usuarios recibirán. Como primer paso, es necesario realizar una imputación a esta matriz, colocando valores de 1 en la diagonal de aquellos elementos en los que la su fila sume 0, lo cual es equivalente a colocar en aquellos elementos que solo han recibido una auto mención. Esto no salva los problemas de conectividad, pero permite calcular la matriz estocástica, a la vez que identifican en la red usuarios que reflejan la tendencia (promedio) a no transmitir mensajes mediante menciones en una red (análogos a los estados absorbentes).

En particular, la matriz M obtenida de la red analizada es una matriz no conectada fuertemente, por lo que no tendremos un vector de probabilidad de estado estacionario único para la red, pero como habíamos mencionado anteriormente, más bien lo que se buscará es obtener vectores de distribución para diferentes pasos o periodos, o los vectores de distribución a tiempos grandes para determinados estados iniciales. Se debe mencionar que para este ejercicio predictivo se supone que no se incorporarán nuevos usuarios a la red en los periodos posteriores ($t > 10$), ni otros eventos que puedan suponer cambios en las probabilidades de la matriz estocástica, esto implicaría que los usuarios seguirán su comportamiento en cuanto al número de menciones y a quien mencionan. Con estas consideraciones previas por ejemplo podemos hacer estimaciones del tipo:

Si una información de dominio público tiene igual probabilidad de entrar a la red con una mención a cualquiera de sus usuarios (esto se representa con un vector de distribución inicial $\pi^{(0)}$con todos sus valores $1/$*número de usuarios*), se tendrá que las probabilidades de recibir menciones en $t + 1$ y en $t + 99$ serán:

Tabla 6: Predicción de probabilidades

Usuario	t+1		Usuario	t+99
@eugenegu	9,54E-02		**@JeffreyGuterman**	9,67E-02
@malonebarry	8,47E-02		**@malonebarry**	8,48E-02
@Telegraph	6,26E-02		**@Telegraph**	6,26E-02
@BBCBreaking	5,26E-02		**@Kasparov63**	5,42E-02
@Kasparov63	5,23E-02		**@BBCBreaking**	5,26E-02
@CaliConsrvative	4,12E-02		**@CaliConsrvative**	4,12E-02
@StephenMillIer	3,93E-02		**@StephenMillIer**	3,93E-02
@1776Stonewall	2,79E-02		**@MothershipSG**	2,85E-02
@MothershipSG	2,76E-02		**@1776Stonewall**	2,79E-02

Se puede apreciar que en el primer periodo futuro en atraer menciones sería el usuario @eugenegu, mientras que al *límite* el más mencionado sería el usuario @JeffreyGuterman, esto sería que de cada 1000 menciones que se realicen en la red éste podría recibir 97.

Evidentemente, si se establecen probabilidades iniciales diferentes, considerando que una información puede ser introducida a la red mediante varios usuarios con una probabilidad mayor de mención, se podrían tener otros resultados.

Para valorar la certeza de la predicción se estimará el Error medio cuadrático (MSE), a través de los siguientes pasos:

1. Se estima un vector de distribución inicial $\pi(t)$ a partir de las menciones que hacen los usuarios de la red en el tiempo t (grados de salida a tiempo t).

2. Se realiza la estimación $\widehat{\pi}(t+1) = \pi(t).M$.

3. Se calcula el vector de errores $e = \widehat{\pi}(t+1) - \pi(t+1)$ donde $\pi(t+1)$ corresponde a la distribución de menciones recibidas por los usuarios en el tiempo.

4. Para apreciar el error E en el número de menciones estimadas, usamos la expresión $E = NT(t+1) \times e$, donde $NT(t+1)$ es un escalar que representa el número de menciones multiplicado por el vector de errores.

5. Se estima el MSE para el número de menciones:

$$MSE = \frac{1}{n}\sum_{i}(E)^2$$

Usando este método para estimar las menciones a $t = 10$ y compararlas con las observadas en el mismo periodo, se obtiene:

Tabla 7: Predicciones al periodo 10

Usuario	Predicción a t=10	Observado a t=10	Error	Error %
@malonebarry	810,4	771	39,4	5,10 %
@eugenegu	788,2	876	-87,8	-10,00 %
@Telegraph	627,6	606	21,6	3,60 %
@BBCBreaking	525,6	499	26,6	5,30 %
@Kasparov63	489,1	471	18,1	3,80 %
@StephenMillIer	447,4	435	12,4	2,90 %
@CaliConsrvative	403,7	383	20,7	5,40 %
@AJEnglish	312	319	-7	-2,20 %
@MothershipSG	252,2	250	2,2	0,90 %
@1776Stonewall	224,4	267	-42,6	-16,00 %

El MSE total asciende a 3,42. Resulta también interesante analizar los estadísticos descriptivos de E que se aprecian en la Tabla 8, observando que en el rango intercuartílico para los errores individuales es 0 en el total de los valores, incluyendo al promedio y la mediana, hablando de que la mayoría de los errores de la predicción son 0. No obstante, también se pueden ver errores de magnitud considerable pero que al ser poco frecuentes no afectan mayormente al estadístico MSE de errores global del método predictivo. En general se puede ver que el método presenta una buena capacidad predictiva

Tabla 8: Estadísticos del los errores

Estadístico	Valor E
Mín.	-125,0
1er. Cuartil	0,0
Mediana	0,0
Promedio	0,0
2do. Cuartil	0,0
Máx.	39,4

5 Conclusiones

Los análisis de redes sociales han cobrado mucha vigencia debido a la interacción que generan las nuevas formas de comunicación en la sociedad. Es así como Twitter, tiene mucha vigencia el ámbito político, empresarial, social, etc. Por ello, es necesario entender, en el ámbito colectivo, el impacto de los usuarios respecto a los temas que resulten de interés.

Para esto, la teoría de grafos resulta de gran utilidad ya que permite entender la composición de las redes e identificar a los usuarios más importantes, sobre todo a través de las medidas de centralidad. Adicionalmente, la combinación de estas técnicas con técnicas predictivas, como la de la matriz estocástica que ensayamos, puede llevar a tener una gran

aplicabilidad en diversos ámbitos, al tener un conocimiento de los usuarios con elevado potencial de menciones o con posibilidades de ser influyentes, en distintos ámbitos de análisis.

Como ejemplo de aplicación, se exploró la red de menciones formadas por la discusión generada en las redes sociales sobre el político Kim Jong-un con motivo de la cumbre Inter Coreana realizada el 23 de abril de 2018 en Corea del Sur y que da cuenta de mensajes que hablan consistentemente de paz, desnuclearización, entre otros; logrando identificar una serie de usuarios con características importantes como los más mencionados, los que más mencionan, los mejores intermediarios, los más cercanos (populares); e incluso los más influyentes (no únicos). Así, entidades políticas o comerciales podrían estar interesadas en que usuarios influentes apoyen mensajes determinados en la red social. Una idea similar ocupa el algoritmo Page Rank del buscador de Google para estimar la importancia de las páginas en sus búsquedas.

La aplicación predictiva aplicada vía matrices estocásticas, se enfrenta a supuestos como la propiedad de Markov, que involucra que las menciones inmediatas futuras dependen únicamente de quién realiza menciones en el tiempo actual sin importar de quien recibió la mención, lo cual puede implicar que el usuario que realiza la mención no discierne entre la veracidad de mensaje y lo transmite sin importar de quien lo reciba. Igualmente, se debe suponer que otros usuarios no se incorporarán a la conversación.

Adicionalmente, también se debe mencionar que los métodos de obtención de tuits provistos por Twitter pueden devolver información limitada porque dependen del número de usuarios descargando datos haciendo que la red no refleje la realidad y no necesariamente porque enfrenten alguna censura importante (Thelwall, 2015).

Este método resulta de aplicación adecuada, teniendo en cuenta que Twitter puede generar redes de menciones poco densas (con matrices de adyacencia mayormente pobladas de cero), dado que otros métodos destinados a predecir la formación de nodos pueden presentar desbalances predictivos hacia la no formación de links.

Referencias

Csárdi, G. y Nepuzs, T. (2006). The igraph software package for complex network research. *Interjournal, Complex Systems*, 1965.

De Nooy, W., Mrvar, A., y Vladimir, B. (2005). *Exploratory Network Analysis with Pajek*. New York: Cambridge University Press.

Glickenstein, D. (2008). Math 443/543 Graph Theory Notes 8: Graphs. Retrieved Agosto 2015, 30, from `http://math.arizona.edu/~glickenstein/math443f08/notes8.pdf`.

Huberman, B., Romero, D., y Fang, W. (2008). Social networks that matter: Twitter under the microscope.

inside-R (n.da). Closeness centrality of vertices. (inside-R). Retrieved Agosto 28, 2015, from `http://www.inside-r.org/packages/cran/igraph/docs/closeness`.

inside-R (n.db). Vertex and edge betweenness centrality (inside-R). Retrieved Agosto 2015, 28, from `http://www.inside-r.org/packages/cran/igraph/docs/betweenness`.

Introduction to social network methods (n.d). Retrieved Agosto 2015, 28, from `http://faculty.ucr.edu/~hanneman/nettext/C10_Centrality.html`.

Lion, H. (2016). Analyzing the Panama Papers with Neo4j: Data Models, Queries & More. Retrieved Junio 11, 2018, from `https://neo4j.com/blog/analyzing-panama-papers-neo4j/`.

Moreno, J. (1951). Sociometry, Experimental Method and the Science of Society: An Approach to a New Political Orientation. *Beacon House*.

Newman, M. (2010). *Networks. An Introduction.* New York: Oxford University Press.

Otte, E. y Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, (28):441–453.

Remus, R. (n.d). The Mathematics of Web Search. Retrieved julio 30, 2015, from `http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture2/lecture2.html`.

Scott, J. y Carrington, P. (2012). *The SAGE Handbook of Social Network Analysis*. Cornwall: SAGE Publications.

Smith, M., Lee, R., Himelboim, I., y Schneiderman, B. (2014). *Mapping Twitter Topic Netwoks: From Polarized Crows to Community Clusters*.

Steward, W. (2009). *Probability, Markov Chains, Queues and Simulation.* New Jersey: Princeton University Press.

The Econometrics of Social Networks (2009). Retrieved Junio 2015, 8, from `http://www.cirpee.org/fileadmin/documents/Cahiers_2009/CIRPEE09-13.pdf`.

Thelwall, M. (2015). *Evaluating the comprhensiveness of Twitter Search.* API. Wolverhampton.

Twitter Developers (n.d). The search API. Retrieved Julio 28, 2015, from `https://dev.twitter.com/rest/public/search`.