ANÁLISIS COMPARATIVO ENTRE ALGORITMOS BETWEENNESS Y CLOSENESS CENTRALITY PARA IDENTIFICAR NODOS CENTRALES EN REDES SOCIALES.

Miguel Ángel Cohuo Avila[1]
José Manuel Lira Turriza[2]
José Luis Lira Turriza[3]
Yaqueline Pech Huh[4]

RESUMEN.

En la actualidad se demanda un desarrollo de software ágil, pero ello en ocasiones pone en riesgo la profundidad del estudio de dominio del negocio, por lo que el ingeniero de software se enfrenta al reto de dar respuestas a sus clientes en menor tiempo y con la calidad que por naturaleza se requiere. La tarea de interpretación de requerimientos es compleja y conlleva en muchas de las ocasiones a interpretaciones erróneas de lo que el cliente demanda como necesidad.
La presente investigación aborda un caso de estudio del sector restaurantero, que parte de la organización del instrumento de entrevista en las áreas fundamentales del modelo de negocio, pasando por el análisis de resultados y modelado de los procesos del negocio, priorizándolos para atenderlos por incrementos en su desarrollo. La aplicación de la Notación del Modelado de Procesos de Negocio (BPMN por sus siglas en inglés) en el análisis de sistemas, apoyó para lograr una correcta interpretación de los requerimientos, la optimización del proceso de desarrollo y una comunicación efectiva con el cliente. Por lo anterior, y la sencillez de la nomenclatura, se recomienda su uso en el desarrollo de sistemas.

Fecha de recepción: 01 de marzo, 2018.

Fecha de aceptación: 08 de mayo, 2018.

[1] Integrante de Cuerpo Académico ITESCAM-CA-04 Instituto Tecnológico Superior de Calkiní en el estado de Campeche, México (ITESCAM).macohuo@itescam.edu.mx
[2] Integrante de Cuerpo Académico ITESCAM-CA-04 Instituto Tecnológico Superior de Calkiní en el estado de Campeche, México (ITESCAM).jmlira@itescam.edu.mx
[3] Líder de Cuerpo Académico ITESCAM-CA-04 Instituto Tecnológico Superior de Calkiní en el estado de Campeche, México (ITESCAM).jlira@itescam.edu.mx
[4] Integrante de Cuerpo Académico ITESCAM-CA-04 Instituto Tecnológico Superior de Calkiní en el estado de Campeche, México (ITESCAM).ypech@itescam.edu.mx

INTRODUCCIÓN.

Análisis en redes sociales en línea (SNA)

La población de usuarios de las redes sociales desde equipos de cómputo o teléfonos celulares se ha incrementado en la república mexicana siendo esta de unos 44.1 millones de usuarios(Juárez and Menéndez 2013) de los cuales un 82% utilizan las plataformas más comunes como lo son Facebook y Twitter. Esto conlleva a la generación de grandes volúmenes de información debido a la facilidad de propagación que dichas plataformas ofrecen, pero dificultando la obtención de patrones a simple vista, surgiendo la necesidad de modelar las relaciones de los miembros de una comunidad o red social en línea. El incremento y la popularidad de las redes sociales on-line a gran escala han retomado importancia en los estudios del análisis de una red social en línea, utilizando los conocimientos basados en la teoría de grafos y algoritmos matemáticos, aunado a la inmensa información que se genera permitiendo establecer grandes oportunidades en la búsqueda del conocimiento. El análisis de la centralidad es uno de los métodos principales de investigación en SNA, en donde la centralidad de un sujeto refleja su posible desempeño en la red.(Liu, Chen et al. 2013)

De acuerdo al artículo realizado en(Aggarwal 2011), El SNA se pueden establecer desde dos puntos de vista:

A. Basado en el análisis de contenido (Adding Content-based Analysis), las plataformas de redes sociales contribuyen con grandes volúmenes de contenido que permitirán mejorar el SNA.
B. Basado en el vínculo (Linkage-based) y el análisis estructural (Structural Analysis), siendo este un análisis de comportamiento estructural de la red para poder determinar la importancia de los nodos, comunidades, enlaces, y la evolución de las regiones de la red. Este análisis proporciona una correcta visión general del comportamiento de la evolución global de la red subyacente; enfoque sobre el cual se basará el presente artículo.

JUSTIFICACIÓN.

La importancia de utilizar herramientas para el análisis de redes sociales en línea se debe a la necesidad de organizar grandes volúmenes de información generada por los usuarios en redes sociales (Facebook), como son los datos de sus contactos o amigos, las publicaciones realizadas y el conjunto de páginas de su interés; esta diversidad de información se enfrenta a los retos para determinar esquemas de visualización e interpretación. El poder conocer la estructura de la red y su interacción con sus nodos (teoría de grafos), así como determinar conocimiento a partir de la información procesada como lo establece (Kuz, Falco et al. 2016) es importante en el análisis de las redes sociales, este artículo se centra en la propiedad de centralidad de cada nodo y su grado de enlace recibido indegreed (Enterría 2012).

El artículo aborda el concepto de centralidad desde el punto de vista de nodos influyentes, conocido como prestigio del nodo y la forma en que interactúan en la red comparando dos métodos para determinar cuál arroja mejores resultados al identificar el nodo influyente abordando el análisis del uso de caminos más cortos de todos los vértices a todos los otros nodos que pasan a través de ese nodo(BC) y aquellos nodos que a pesar de tener pocas conexiones, sus arcos permiten llegar a todos los nodos de la red más rápidamente que desde cualquier otro punto (CC), además de poder experimentar con la identificación de los nodos que nos permitan una comunicación y difusión de información entre ellos de manera más eficiente, directa y a la mayoría de los integrantes de la red.

En este artículo se analizan los elementos principales del SNA de los Alumnos del Instituto Tecnológico Superior de Calkiní en el estado de Campeche (ITESCAM) utilizando herramientas de extracción de datos de la plataforma Facebook de los usuarios seleccionados con la herramienta Netvizz (Rieder 2013), para aplicar la teoría de grafos y visualizar los patrones de distribución, utilizando Layout forceAtlas2 (Jacomy, Heymann et al.), para obtener métricas básicas de la red que permitan identificar los nodos con mayor influencia para cada red de amigos de los sujetos a estudiar, con el propósito de identificar la efectividad de los métodos Betweenness Centrality (frecuencia de un nodo) y Closeness Centrality (distancia de un nodo) y se analizan en base a las métricas principales de la red de personas a estudiar, como se presenta en la sección de resultados de éste artículo.

METODOLOGÍA.

En la siguiente sección se abordan metodologías importantes para el análisis de redes sociales, así como diferentes herramientas y algoritmos de ayuda tanto para la visualización como el análisis.

La forma en que interactúan actualmente las personas en plataformas web como lo es Facebook dio mayor auge e importancia al SNA el cual se define como una estructura social que está conformada por un grupo de individuos a los cuales se les llama nodo, y están conectados mediante algún tipo de amistad, negocio o parentesco. (Olivares)

Con un grafo podemos representar una red social que consiste en un conjunto de nodos v_1 de personas y un conjunto de aristas ε que representan las relaciones de amistad entre ellos donde cada vértice debe ser $v \neq 0$, para trabajar exclusivamente con grafos dirigidos. (McGlohon, Akoglu et al. 2011), expresado en la ecuación 1

$$\varepsilon : G = (v, \varepsilon) \qquad (1)$$

Dentro del conjunto de software existente para la visualización, SNA y aplicación de los algoritmos tanto de código abierto o comercial como es: Tableau (Public), Weave, Many Eyes, NodeXL, Data-Driven Documents, Axiis, Google Fusion Tables y Gephi. (Ferrer-Sapena and Sánchez-Pérez 2013)

Seleccionamos Gephi por su entorno intuitivo y su sencilla plataforma de visualización de nodos que permite hacer grafos dinámicos y jerárquicos de la forma más simple. Dos ejemplos de uso de Gephi es aplicado al SEO donde nos demuestra como visualizar los datos de Open Site Explorer, otro ejemplo nos explica cómo conseguir análisis de redes y visualización con Gephi (Cherven 2013) y (Bastian, Heymann et al. 2009)

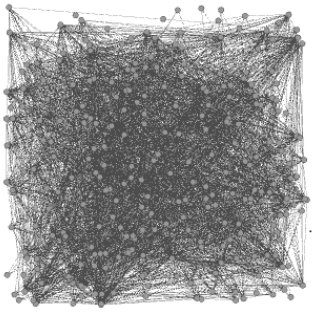

Figura 1 Representación del grafo con 1254 nodos y 63,157 aristas
Source: Ejecutado en Gephi.

Las características del equipo de cómputo donde se ejecutaron las pruebas son: una arquitectura de 64 bits, con un procesador Intel Core i7 de 8 núcleos cada uno a 1.87Ghz y una capacidad de memoria instalada de 8 Gb, además de una Tarjeta Gráfica Nvidia Quadro FX 1800M. Así mismo se configuró la máquina virtual de java versión 1.6 para poder obtener más memoria virtual en la ejecución del programa modificando el archivo de configuración gephi.conf. con los parámetros –J-Xms1024m y –J-Xms4096m, lo que representa el 50% de uso de la memoria física. El grafo resultante es similar al manejado en el estudio. (Soleimani-Pouri, Rezvanian et al. 2014)

Layout de visualización ForceAtlas2

Para un mejor entendimiento visual se utilizan algoritmos de visualización de grafos (Jacomy, Heymann et al.), en nuestro estudio seleccionamos ForceAtlas2 que nos permite graficar y obtener una vista del grafo más simple utilizando la fuerza de atracción clásica, donde nos indica que la fuerza de atracción f_α entre dos nodos conectados v_1 y v_2 no tiene nada destacable, sin embargo depende linealmente de la distancia entre los nodos ubicando en espacios bidimensionales o tridimensionales los nodos con similitudes expresado en la ecuación 2

$$f_\alpha(v_1, v_2) = d(v_1, v_2) \qquad (2)$$

El algoritmo ForceAtlas2 fue diseñado utilizando la repulsión por grado (Repulsion by degree) para interpretar gráficos web y redes sociales para obtener una visualización más simple siendo una de las características de las redes la presencia de muchos nodos con un solo vecino, debido a la ley de potencia de distribución del grado (Power-law distribution of degrees) que caracteriza muchos datos del mundo real. Utilizando la repulsión del grado del nodo de tal manera que los nodos pobremente conectados se repelen con los que tienen mayor fuerza de atracción. La fuerza de repulsión f_k es proporcional al producto de los grados más uno de dos nodos. El coeficiente k_r para un par de nodos se expresa en la ecuación 3

$$f_k(v_1, v_2) = k_r = \frac{(\deg(v_1)+1)(\deg(v_2)+1)}{d(v_1, v_2)} \qquad (3)$$

Un punto importante en el SNA es poder identificar el nodo central dentro de la red, considerando el propósito y el contexto.

Considerando una medida local tenemos Degree, y considerando el resto de la red podemos mencionar cercanía (closeness), intermediación (betweenness) y vector propio (eigenvector) basado en el poder de la centralidad de Bonacich (Borgatti 2005)

Lo que nos genera el siguiente cuestionamiento ¿cómo se distribuye la centralidad de manera uniforme en una red, para poder determinar la influencia en base a las relaciones interpersonales de una red social en este caso Facebook?

Nos basaremos en la fórmula de la centralidad propuesta por Freeman la cual se expresa en la ecuación 4. (Freeman 1978)

$$C_D = \frac{\sum_{i=1}^{g}\{(C_D(n^*) - C_D(i)\}}{\left[(N-1)(N-2)\right]} \qquad (4)$$

Para nuestro estudio seleccionamos betweenness: la frecuencia con la que un nodo se encuentra en una posición intermedia a lo largo de las trayectorias geodésicas que unen pares de otros nodos y Closeness Centrality: Que se definen solamente para redes en las que todos los nodos están mutuamente relacionados entre sí por caminos de distancia geodésica (O'malley and Marsden 2008)

Closeness Centrality (CC).

Se basa en la idea de que los nodos con una corta distancia a otros nodos pueden propagar información muy productiva a través de la red, con el fin de calcular el CC σC (V) de un nodo V, las distancias entre el nodo v_1 y v_2 todos los demás nodos de la red se resumen. Al utilizar el valor recíproco logramos que el valor de CC aumente cuando se reduce la distancia a otro nodo, es decir, cuando se mejora la integración en la red. (Landherr, Friedl et al. 2010)

Es decir, establece la distancia media de un nodo con el resto de los nodos de la red. Expresado en las ecuaciones 5 y 6

$$C_c(i) = \left| \sum_{j=1}^{N} d(i, j) \right|^{-1} \qquad (5)$$

Normalizado

$$C'_c(i) = \frac{c_c(i)}{(N-1)} \qquad (6)$$

Betweenness Centrality (BC)

Es el número que representa cómo es de frecuente que un nodo esté entre los caminos geodésicos (número de relaciones en el camino más corto posible de un nodo hacia otro nodo) de otros nodos expresado en la ecuación 7 y 8

$$C_B(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}} \qquad (7)$$

Donde g_{jk} denota el número de conexiones geodésicas jk y $g_{jk}(i)$ denota el número geodésico que el nodo se encuentra en el camino de i, Usualmente se normaliza para aplicar.

$$C'_B(i) = \frac{C_B(i)}{\left[\frac{(n-1)(n-2)}{2} \right]} \qquad (8)$$

DISCUSIÓN DE RESULTADOS.

Experimentación

Antes de realizar el análisis comparativo de los métodos de centralidad, se debe preparar los escenarios como es: la obtención de los datos, selección del software de análisis, graficación de los datos y visualización.

Se seleccionó una muestra aleatoria de 86 estudiantes de la comunidad estudiantil del ITESCAM de los diferentes carreras, semestres y grupos, luego se procedió a la obtención de sus datos usando la herramienta Netvizz apps versión 1.0.1 donde se seleccionó la opción para descargar la información del perfil del usuario de Facebook referente a su red personal, el cual extrae a sus amigos y la conexión de amistad entre ellos.

Obteniendo el archivo GDF del usuario tenemos dos maneras de representar la red por medio de matrices o grafos (O'malley and Marsden 2008) de las que se seleccionó la interpretación gráfica por medio de la teoría de grafos con el apoyo de técnicas de visualización que faciliten la identificación de nodos centrales con mayor claridad como lo realizado en (Crnovrsanin, Muelder et al. 2014) para proceder a ejecutar con la herramienta Gephi la visualización del grafo como se ilustró en la figura 1.

Posteriormente se procedió a la obtención de los datos generales de la red, como es diámetro de la red, longitud de camino medio, grados de entrada y salida de todos los nodos. Se ejecutó el algoritmo layout forceatlas2 para su registro.

Un ejemplo de la utilización del algoritmo forceAtlas2 con un rendimiento de velocidad de 0.1, escalado de 2.0, gravedad 1.0 y entre 600 y 1000 iteraciones se muestra en la figura 2

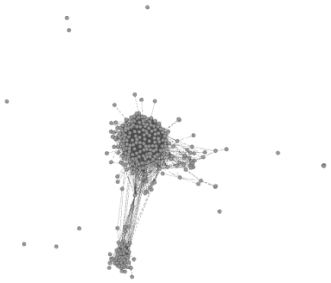

Figura 2 Ejecución de ForceAtlas2 con un grado medio de 50.36 y una longitud promedio de camino de 2.19 Source: Ejecutado en Gephi.

Al obtener el grafo resultante, se realizó la ejecución de los algoritmos de clasificación para cada una de las muestras obteniendo el valor calculado para BC y CC para registrar los nodos encontrados, de manera numérica y gráfica.

Como último paso del desarrollo se realizó el análisis comparativo de cada uno de los métodos de clasificación.

Resultados y discusión del SNA de los sujetos de estudio

Al obtener los datos para cada uno de las redes de amigos de los estudiantes a analizar se tiene entre un rango de [46, 4142], esto nos representa una muestra significativa en la estructura de la red en términos de 10^3 nodos, con una variación en el trabajo para identificar nodos influyentes en la redes sociales en línea en cual trabajaron una cantidad de nodos de 10^4 (Ilyas and Radha 2011)

Para el desarrollo de nuestro estudio se realizó un muestreo aleatorio generando números aleatorios y seleccionando 86 estudiantes, de las diferentes carreras semestres y grupos. Al realizar el análisis de los datos obtenidos se pudo observar que en la variable del diámetro de la red con mayor frecuencia con un diámetro de 7 representa el 25.58%, de diámetro 8 el 20.93% y con un diámetro de 6 el 17.4 %, esto significa que es la distancia del grafo más larga entre dos nodos de la red, como lo describe (McGlohon, Akoglu et al. 2011) y se muestra en la figura 3.

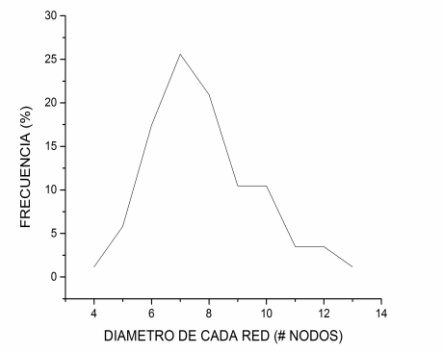

Figura 3 Diámetros de las redes analizadas.

ISSN: 2395-843X. Año VI. No. 07. Abril, 2018
FIME - UANL

La manera de ejecutar los métodos con la herramienta Gephi al obtener el grafo resultante con forceatlas2, se procede a obtener los datos generales de la red, como ejemplo se ejecutó un sujeto de estudio que su red personal contaba con 1254 nodos y 63, 157 aristas, el grado medio 10.83, diámetro de la red 7, la longitud de camino promedio de 2.43. El siguiente paso fue ejecutar los algoritmos de clasificación en primer lugar se ejecutó el algoritmo Closeness con factores de 4.3, 4.0 y 3.9 es decir se identificaron 3 nodos influyentes como se observa en la figura 4

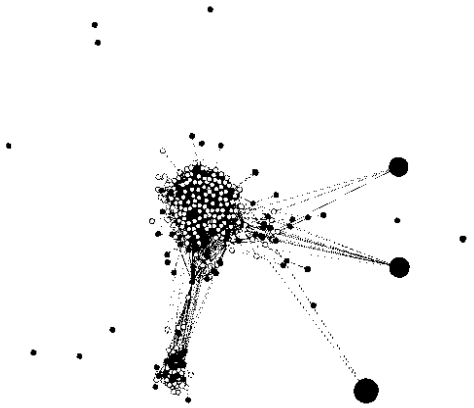

Figura 4. Identificación de 3 nodos Closeness Centrality.

Posteriormente se realizó la ejecución de Betweenness se pudo obtener un factor de 1,608.35 identificando un solo nodo como se muestra en la figura 5

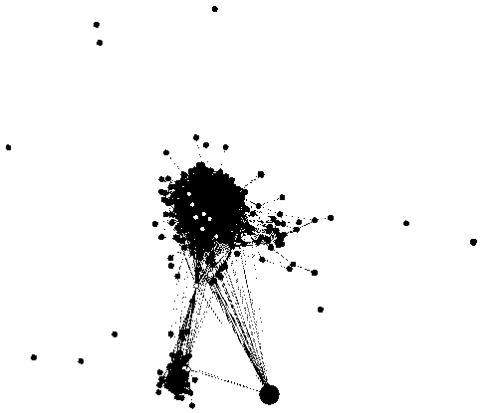

Figura 5. Ejecución de Betweenness Centrality al identificar un nodo.

Este procedimiento se ejecutó para cada uno de los casos donde se observó que el 98% de los casos el método BC obtuvo menos nodos que CC con su 2 % como se observa en la tabla 1.

Table 1. Frecuencia de casos con el número de nodos identificados por método.

NODOS	BC	CC
1	67	18
2	11	23
3	7	14
4	1	8
5	0	6
6	0	8
7	0	2
8	0	3
10	0	3

Source: Gephi

CONCLUSIONES Y/O RECOMENDACIONES.

La efectividad para encontrar un nodo con el método BC fue superior al CC, además el nodo encontrado con BC contaba con un mayor grado de relación que los obtenidos con el método CC como pudimos observar en las tablas 2 y 3. Gephi nos proporcionó una visualización clara del nodo influyente en conjunto con los algoritmos de la Centralidad

En las figuras 4 y 5, se observan los resultados de la ejecución de los dos métodos empleados para encontrar los nodos de mayor influencia en las redes de amistades de los estudiantes del ITESCAM. Con el método BC se encontró 67 veces lo que representa un 77.9 % de éxito en la identificación de un único nodo. Así mismo cuando se evaluó el método CC se identificaron 18 casos para un único nodo lo que representa un 20.9 % de éxito en la identificación del nodo con mayor influencia, lo que representa mayor efectividad del método BC. Podemos observar que se pueden realizar comparativas de los algoritmos de centralidad a través de su complejidad espacial y temporal como se describe en (Divya and Reghuraj 2014)

En las tablas 2 y 3 se muestran los resultados obtenidos al aplicar cada uno de los métodos que se están comparando en las mismas redes. En la tabla 2 podemos observar una muestra de los valores obtenidos en las redes con me método BC y sus valores de grados de entrada y de salida. En la tabla 3 se encuentran los resultados con las mismas redes, pero obteniendo los valores de los nodos centrales y sus grados de entrada y salida a través del método CC. Haciendo la comparación entre ambas tablas podemos observar que la cantidad de relacionados por el método BC es mucho mayor en comparación con los nodos obtenidos por el método CC lo que nos indica que el primero es más efectivo en referencia a que permite una mayor identificación con un solo nodo.

Table 2. Resultados del Análisis a través del método Betweenness Centrality

SUJETO	BC	GRADO DE ENTRADA	GRADO DE SALIDA	GRADO
1	329,492.60	118	598	716
2	165,093.65	535	129	664
3	95,900.79	168	74	242
4	27,309.71	605	149	754
5	8,951.68	243	130	373
6	24,384.25	258	459	717
7	41,257.07	176	12	188
8	10,540.92	147	37	184
9	6,929.34	112	163	275
10	4,648.86	174	120	294
11	1,549.24	186	83	269

ISSN: 2395-843X. Año VI. No. 07. Abril, 2018
FIME - UANL

12	3,919.41	139	54	193
13	1,839.39	38	26	64
14	988.99	12	30	42
15	497.65	23	39	62

Source: Gephi.

Table 3. Resultados del Análisis a través del método Closeness Centrality

SUJETO	CC	GRADO DE ENTRADA	GRADO DE SALIDA	GRADO
1	7.036	10	4	14
2	6.039	0	2	2
3	6.509	0	1	1
4	5.649	44	3	47
5	5.481	0	1	1
6	5.103	28	1	29
7	6.651	0	3	3
8	4.936	7	14	21
9	4.346	25	1	26
10	5.164	11	1	12
11	4.108	0	1	1
12	4.127	1	4	5
13	3.696	0	1	1
14	3.578	7	17	24
15	3.672	5	1	6
16	3.639	3	2	5

Source:Gephi

Se recomienda para futuros trabajos el determinar y estudiar nuevos mecanismos de extracción de datos en las redes sociales con resultados óptimos y aceptables al incrementar la cantidad de información a manejar y manipular, generar conocimiento con base a la interpretación de las características de los nodos y su interacción con otros nodos es otro campo de oportunidad en la aplicación e interpretación de nuevos modelos y técnicas, analizar las herramientas tecnológicas existentes para la visualización e interpretación e identificación de métricas del análisis de redes sociales, es decir que herramientas adicionales a Gephi existen para analizar su impacto y rendimiento en el procesamiento de cantidades enormes de información, la utilización y facilidad de cálculo de las métricas, es importante determinar los diferentes layouts de visualización de información de la red para una mejor comprensión. Existe una tendencia de investigación al poder ya identificar el nodo de centralidad desde el punto de vista del nodo más influyente su tipología de tipo: diseminadores, relacionales y líderes (García, Daly et al. 2016)

BIBLIOGRAFÍA.

Aggarwal, C. C. (2011). "Social network data analytics, Chapter An introduction to social network data analytics." IBM TJ Watson Research Center Hawthorne, NY 10532 **13**.

Bastian, M., et al. (2009). "Gephi: an open source software for exploring and manipulating networks." Icwsm **8**: 361-362.

Borgatti, S. P. (2005). "Centrality and network flow." Social networks **27**(1): 55-71.

Crnovrsanin, T., et al. (2014). "Visualization techniques for categorical analysis of social networks with multiple edge sets." Social networks **37**: 56-64.

Cherven, K. (2013). Network graph analysis and visualization with Gephi, Packt Publishing Ltd.

Divya, S. and P. Reghuraj (2014). "Eigenvector based approach for sentence ranking in news summarization." IJCLNLP, April.

Enterría, A. G. (2012). "El análisis de las redes sociales (ARS) con metodología pra el estudio del ciberespacio islámico español." Revista española de ciencia política(30): 121-131.

Ferrer-Sapena, A. and E. Sánchez-Pérez (2013). "Open data, big data:¿ hacia dónde nos dirigimos?" Anuario ThinkEPI 2013 **7**: 150-156.

Freeman, L. C. (1978). "Centrality in social networks conceptual clarification." Social networks **1**(3): 215-239.

García, M. d. F., et al. (2016). "Identificando a los nuevos influyentes en tiempos de Internet: medios sociales y análisis de redes sociales." Revista EspaÑola de Investigaciones Sociológicas (REIS) **153**(1): 23-40.

Ilyas, M. U. and H. Radha (2011). Identifying influential nodes in online social networks using principal component centrality. Communications (ICC), 2011 IEEE International Conference on, IEEE.

Jacomy, M., et al. Forceatlas2, A Graph Layout Algorithm for Handy Network Visualization, 2012.

Juárez, R. and P. Menéndez (2013). "Hábitos de los usuarios de internet en México 2013." México: Asociación Mexicana de Internet.

Kuz, A., et al. (2016). "Análisis de redes sociales: un caso práctico." Computación y Sistemas **20**(1): 89-106.

Landherr, A., et al. (2010). "A critical review of centrality measures in social networks." Business & Information Systems Engineering **2**(6): 371-385.

Liu, Y., et al. (2013). An Email Forensics Analysis Method Based on Social Network Analysis. Cloud Computing and Big Data (CloudCom-Asia), 2013 International Conference on, IEEE.

McGlohon, M., et al. (2011). Statistical properties of social networks. Social network data analytics, Springer**:** 17-42.

O'malley, A. J. and P. V. Marsden (2008). "The analysis of social networks." Health services and outcomes research methodology **8**(4): 222-269.

Olivares, C. P. M. Análisis de Redes Sociales a Gran Escala, TFC Centro De Investigación Y De Estudios Avanzados del Instituto Politécnico Nacional Departamento De Computación.

Rieder, B. (2013). Studying Facebook via data extraction: the Netvizz application. Proceedings of the 5th annual ACM web science conference, ACM.

Soleimani-Pouri, M., et al. (2014). An ant based particle swarm optimization algorithm for maximum clique problem in social networks. State of the art applications of social network analysis, Springer**:** 295-304.