Simplificación de interacciones y la detección de comunidades en una red social

Erick López-Ornelas[1], Rocío Abascal Mena[1]

[1]Universidad Autónoma Metropolitana – Cuajimalpa, México
Departamento de Tecnologías de la Información, México
{elopez, mabascal}@correo.cua.uam.mx
http://www.cua.uam.mx/

Resumen. Este artículo plantea la utilización del método de *"Mapas Jerárquicos"* para la simplificación de nodos a partir de un conjunto de interacciones en una red social. El corpus utilizado en este artículo está basado en la interacción realizada en Twitter por un conjunto de actores en la destitución del Secretario Mexicano de Hacienda y Crédito Público, el Sr. Luis Videgaray. Al aplicar los mapas jerárquicos es posible simplificar la red social y de esta forma identificar comunidades y actores importantes en el suceso. Un análisis basado en la visualización de las comunidades se realizó para corroborar la pertinencia de la detección.

Palabras clave. Twitter, Simplificación de redes sociales, Detección de comunidades, Visualización de información

Simplifying Interactions and Detection of Communities in a Social Network

Abstract. In this paper we present the use of the method of *"Hierarchical Maps"* for node simplification using a group of interactions in a social network. The corpus used in this paper is based on Twitter interaction by a set of actors in the destitution of the Mexican Secretariat of Finance and Public Credit, Luis Videgaray. By applying the *"Hierarchical Maps"*, it is possible to simplify the social network and thus identify communities and the major actors in the event. An analysis based on community visualization was conducted to confirm the relevance of the detection.

Keywords. Twitter, Social Network Simplification, Community Detection, Information Visualization

1. Introducción

El uso de plataformas de redes sociales en contextos políticos alrededor del mundo, no solo ha cobrado protagonismo para difundir movimientos sociales, sino que se ha convertido en una importante modalidad de socialización y difusión de información. En el caso de México, la integración de las Tecnologías de la Información y la Comunicación (TICs) y el acceso a las redes sociales digitales ha permitido que los jóvenes y la ciudadanía en general, tengan un medio alternativo de comunicación y poder, de esta manera, interactuar, organizarse y visibilizar sus causas.

Una de las principales redes sociales que se utiliza es Twitter, cuya característica principal es que los mensajes que se envían tienen una longitud máxima de 140 caracteres, llamados tuits, que pueden ser almacenados y categorizados en temas a partir del uso de etiquetas precedidas por el símbolo # (*Hashtag*), mismo que permiten seguir, buscar y encontrar conversaciones relacionadas con un tema en común. La generación de comunidades internas en la red social es un fenómeno natural que se gesta cuando existen muchos actores que generan información relacionada con algún tema en específico.

La identificación de comunidades dentro de una red social es un tema importante ya que, además de simplificar la gran cantidad de información contenida en la red, permite detectar grupos o sectores importantes que interactúan entre sí.

En este artículo hacemos un estudio sobre la destitución reciente del Secretario de Hacienda y Crédito Público, el Sr. Luis Videgaray. Las interacciones en Twitter se gestaron cuando el Presidente de la República, el Lic. Enrique Peña Nieto, anunció el día 7 de septiembre del 2016, cambios importantes en su gabinete, cuando faltan dos años para que termine su periodo en la Presidencia de la República.

Entre los múltiples elementos de análisis que se generaron en las redes sociales bajo el *Hashtag* #Videgaray, decidimos realizar un análisis de las principales comunidades que se gestaron dentro de la red social y de este modo mostrar de manera concisa y simplificada la gran cantidad de información generada en la red. A continuación, mostramos en la Fig. 1 la red generada por las intervenciones alrededor de evento mencionado y usando el *Hashtag* #Videgaray.

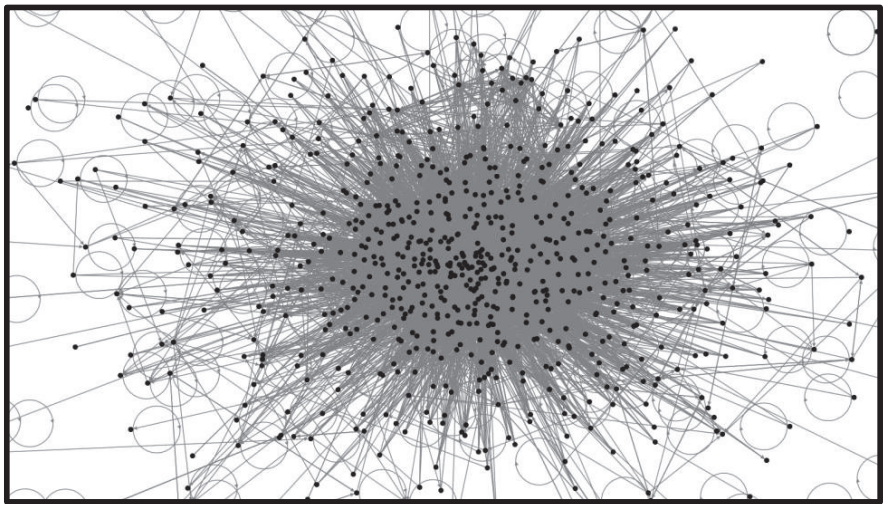

Fig. 1. Red de interacciones #Videgaray el 7 de septiembre del 2016

La información concentrada en la Fig. 1 corresponde a la extracción realizada el 7 de septiembre del 2016. En la red están representados 850 nodos (tuiteros) y 12666 arcos (interacciones) que fueron realizadas entre ellos. Esta es una red de interacción compleja, por lo que es importante simplificarla para poder realizar un mejor análisis de las interacciones que se gestaron en la red social.

El artículo está organizado de la siguiente manera. En la sección 2 se muestra el objetivo general de la investigación y se muestra el método utilizado para abordar la problemática. En la sección 3 se presenta el estado del arte con algunos de los esfuerzos realizados enfocados en la extracción y análisis de comunidades en una red social. En la sección 4 planteamos y explicamos la técnica utilizada para realizar la detección de comunidades utilizada en este artículo. En la sección 5 se muestra el resultado experimental aplicado y se muestra la simplificación de una red social. Finalmente en la sección 6 se describen y exponen algunas de las conclusiones y perspectivas de este trabajo.

2. Objetivo y método

El objetivo de este artículo, es el de poder analizar y simplificar las interacciones que existen en una red social, a partir de un *trending topic*. Para tal efecto la detección de comunidades es un elemento importante para posteriormente realizar una simplificación de las interacciones.

El método utilizado, tiene una relación directa con la extracción de la información proveniente de Twitter. Se realizó un conjunto de extracciones de tuits generados el día del evento anteriormente mencionado y utilizando el *Hashtag* #Videgaray. Se

utilizó la herramienta NodeXL [14] para realizar una primera extracción de los tuits. La visualización mostrada en la Fig. 1 fue realizada utilizando esta herramienta.

Posteriormente, se utilizó el algoritmo basado en los *Mapas Jerárquicos*, el cual se explicará posteriormente, para poder realizar el conjunto de agrupaciones y simplificaciones sobre el grafo inicial correspondiente a la información extraída. Finalmente, una identificación de usuarios relevantes en cada comunidad es propuesta, mostrando la simplificación de las interacciones. El método propuesto es esquematizado en la Fig. 2.

El corpus utilizado está compuesto por 12666 mensajes de Twitter en español escritos por periodistas, políticos y público en general realizados en México el día 7 de septiembre del 2016 entre las 18:00 y 18:05 horas.

Para cada mensaje, la información contenida en el corpus es: su identificador en la red social Twitter, nombre de usuario que lo ha escrito, nombre del usuario a quien hace mención, fecha y hora, el tuit con la descripción textual del mensaje, los hashtags a los que hace referencia en el tuit, el número de seguidores (*Followers*) con que cuenta cada usuario, el número de tuits realizados por cada usuario, la descripción en su perfil de usuario, la ubicación en donde realizó el tuit y la fecha en la que se dio de alta en Twitter.

De toda esta información contenida en el corpus, lo que nos interesa de manera prioritaria son las menciones realizadas, traducidas posteriormente como las interacciones realizadas por los usuarios, además del número de seguidores y frecuencia de publicaciones de sus tuits.

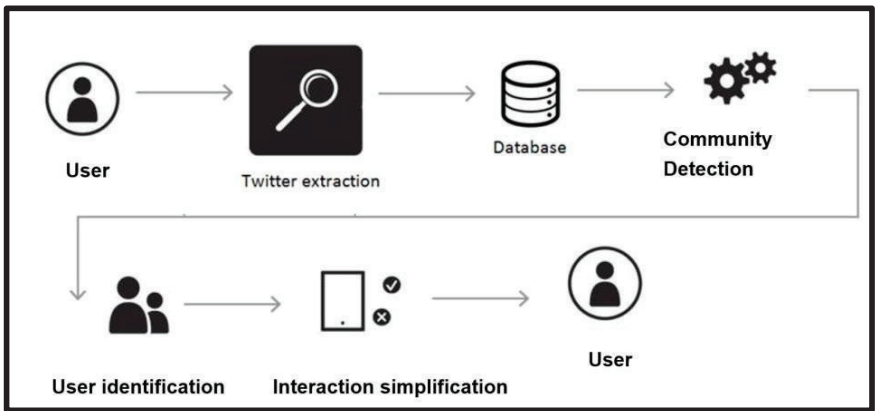

Fig. 2. Método de simplificación de interacciones propuesto

3. Estado del arte: la detección de comunidades

La estructuración de comunidades, es una propiedad de las redes sociales actuales [5]. Una comunidad puede ser definida como un conjunto de nodos que están más densamente conectados entre ellos que con el resto de la red o que existe una mayor comunicación entre ellos. La importancia de este planteamiento radica en que se espera que los nodos que están contenidos dentro de una misma comunidad compartan atributos, características comunes o relaciones funcionales [4]. Sin embargo, no existe una definición exacta de lo que es, o cómo debería ser particionada la red en una comunidad.

Una partición es la división de una red en comunidades o *clusters,* de modo que todo nodo pertenece a algún *cluster*. Además, estas comunidades pueden estar jerárquicamente estructuradas, es decir, dos o más comunidades al fusionarse pueden formar una comunidad de un nivel superior.

Este tipo de estructuras pueden ser representadas mediante un árbol o dendrograma [4]. Por otro lado, en el caso de que un nodo sea asignado a más de una comunidad hablamos de particiones empalmadas u *overlapping*. Es obvio que conforme crece el número de nodos, dificulta de manera extrema la selección de la mejor partición del grafo.

Cómo encontrar la partición óptima es, probablemente, el problema abierto más importante de la investigación en estructura de comunidades. Una gran variedad de métodos y algoritmos, cada uno de ellos con su propia definición intrínseca de comunidad, han sido desarrollados para intentar extraer la partición óptima de una red. Algunos de ellos tratan de optimizar un índice global de calidad de la partición, como puede ser su *Modularity* [7] o *Surprise* [1]. Otros, sin embargo, utilizan la matriz de adyacencia para extraer información del grafo, aplicando, por ejemplo, métodos espectrales [13]. Además, estimaciones de máxima verosimilitud [8], o elementos extraídos de la Teoría de la Información [12], son solo unos pocos ejemplos de métodos que han sido aplicados con relativo éxito a la búsqueda de comunidades.

En la literatura existen trabajos interesantes en torno a la detección de comunidades, por ejemplo [2] propuso un algoritmo basado en dos principios: (i) naturaleza intrínseca de las comunidades, y (ii) detección longitudinal, de igual manera en [10], desarrollaron un algoritmo de detección de comunidades solapadas basado en la idea de "amistad" entre los miembros de una comunidad, donde algunos de estos miembros se comportan como líderes de grupo. En [15] se propone un algoritmo capaz de indicar la influencia de los vértices del grafo contando el número de triángulos que cada vértice comparte con sus vértices adyacentes.

4. Mapas jerárquicos

La definición más general de la comunidad es la de que en la red exista un grupo de nodos que están densamente interconectados. Mientras tanto, desde el punto de vista de la propagación de la información, una comunidad es un grupo de nodos en los que es más probable que se conserve la información en lugar de extenderse. Teniendo

en cuenta que el modelo de propagación fundamental de la información es el "*random walk*" [12], entonces la estructura de una comunidad puede ser identificada mediante una búsqueda local en la estructura. Algunos estudios recientes [3, 6] han demostrado que la modularidad [9], la cual es una función de calidad, es utilizada para encontrar comunidades donde existen grupos de nodos densamente conectados.

En este artículo, utilizamos los "*mapas jerárquicos*" para describir la dinámica de los enlaces y nodos dirigidos, así como de las redes ponderadas para identificar las interacciones locales dentro de la red. Estas interacciones locales permiten calcular el flujo de la información que se puede trasmitir por el nodo, en otras palabras, el grado de interconexión que existe entre dos nodos [12, 11]. En consecuencia, es importante entender el flujo completo de la información en la red. Un grupo de nodos donde la información fluye de manera rápida y sencilla puede ser agregado y definido como un módulo bien conectado. Los enlaces entre los módulos y las veces que se comunicaron, permite identificar el grado de conexión entre los módulos, lo que permite realizar una simplificación del grupo generando un módulo y conservando al nodo principal de este módulo o comunidad (Fig. 3).

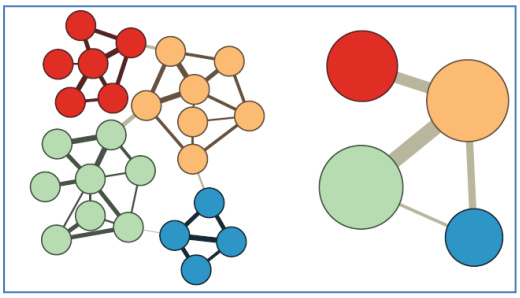

Fig. 3. Esquematización de las agregaciones de nodos en una red

Esta agrupación radica en la dualidad entre la búsqueda de la estructura comunitaria en las redes y el problema de codificación: encontrar un código eficiente, buscamos entonces, una partición M de nodos n dentro de los módulos m para reducir al mínimo la distancia del llamado "*random walk*". Utilizando el módulo de partición M, la distancia promedio para que la información pase de un nodo a otro está dado por la ecuación 1 [11], la cual está formada por dos términos, el primero es la entropía del movimiento entre los módulos, y el segundo es la entropía del movimiento en los módulos.

$$L(\mathbf{M}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^{i}_{\circlearrowright} H(\mathcal{P}^{i}). \tag{1}$$

Los resultados de este método son mostrados utilizando la red generada en el movimiento #Videgaray.

5. Simplificación de la red social

La detección de comunidades, resulta de gran interés para poder analizar las interacciones que se generan en la red. Esta permitirá predecir la intensidad de interacción, en qué etapa se encuentra y cuáles son los actores con mayor peso.

En esta sección aplicamos el algoritmo de *mapas jerárquicos* descrito en la sección anterior para simplificar la red social inicial, lo cual provoca una disminución considerable en el número de nodos visibles.

De los 850 nodos (tuiteros), fueron agrupados y categorizados 30 grandes comunidades. Estas fueron seleccionadas debido a que el porcentaje de la interacción era mayor a 1.0, lo que refleja un conjunto de interacciones importante en la red social. Todas las interacciones menores a 1.0 fueron descartadas.

De estas 30 comunidades identificadas, el actor más representativo fue seleccionado para abanderar la comunidad basado en número de seguidores y número de tuits realizados. De este modo, las 30 comunidades se muestran en la Tabla 1, donde tenemos el módulo de agrupación, el usuario más influyente, la cantidad de información que comparten entre ellos, los que corresponde al porcentaje de las interacciones en la red y el número de nodos que se encuentran agrupados o que son aglutinados dentro de esta gran comunidad.

Tabla 1. Principales comunidades identificadas con el *Hashtag* #Videgaray

Módulo	Actor representativo	Porcentaje de la interacción	Nodos en la comunidad
1	@epn	13.0	75
2	@cnnmex	16.1	56
3	@cnnee	10.8	66
4	@aristeguionline	7.1	80
5	@adnpolitico	6.0	44
6	@jrisco	1.2	3
7	@pictoline	1.5	2
8	@lvidegaray	5.0	21
9	@jernarovillamil	1.3	4
10	@denisedresserg	4.5	34
11	@lopezdoriga	3.1	12
12	@chumeltorres	1.1	2
13	@nytimes	2.0	85
14	@ap	1.3	2
15	@leonkrauze	2.0	4
16	@washingtonpost	1.2	70
17	@el_pais	1.0	67
18	@josecardenas1	2.3	5
19	@mzavalagc	2.3	14
20	@revistaproceso	2.0	12

21	@jshm00	1.0	1
22	@lydiacachosi	1.3	9
23	@leozuckerman	1.2	8
24	@el_universal_mx	1.2	24
25	@lasillarota	2.0	13
26	@jorgeramosnews	2.6	46
27	@reforma	1.2	79
28	@sncp_mx	3.1	3
29	@josemeadek	3.1	4
30	@sopitas	2.1	5

Al aplicar el algoritmo, existe una simplificación visual muy importante, reduciendo la carga cognitiva para la identificación de las interacciones de cada comunidad.

En la Fig. 4 mostramos la red original con el usuario más influyente en estas interacciones (@cnnmex). En este escenario resulta imposible identificar a los usuarios con los que interactúa y mucho menos identificar la importancia en las relaciones. En la Fig. 5 se muestra una simplificación utilizado los *mapas jerárquicos* donde únicamente se muestran las comunidades con las que interactúa. Esta visualización resulta más clara y concisa y permite realizar un mejor análisis de las interacciones existentes en la red.

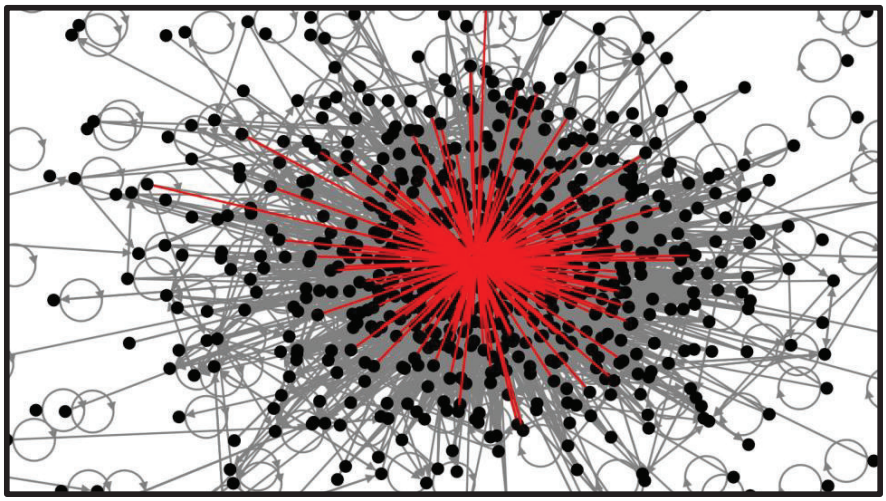

Fig. 4. Enlaces del usuario @cnnmex en la red original

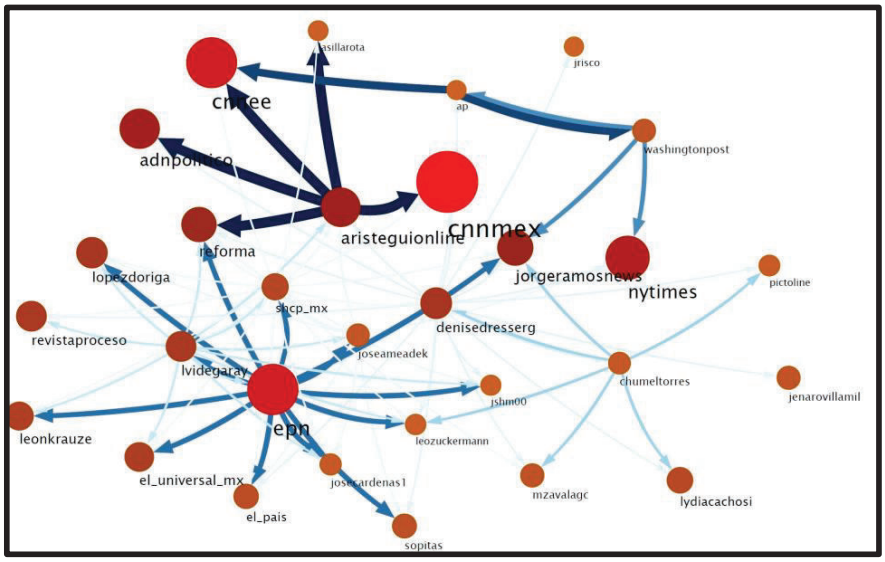

Fig. 5. Simplificación de la red #Videgaray en 30 Módulos

Al aplicar esta simplificación, inmediatamente podemos apreciar que, en esta interacción, existió una presencia muy importante de los medios de comunicación, incluso podemos determinar quiénes fueron los más activos y con mayor peso (definida por el porcentaje de interacción encontrado) en la red. De este modo es posible identificar los usuarios influyentes en esta interacción.

Otra característica importante a resaltar, es que podemos analizar sub redes de información, esto es, podemos analizar qué es lo que está pasando dentro de la comunidad, cuáles son las interacciones importantes, quiénes son los actores importantes en la sub red y qué fuerzas interactúan dentro de la comunidad. Este análisis es de gran ayuda para determinar la importancia de cada una de las comunidades y sobre todo el poder de influencia no solo de un usuario, sino de la comunidad.

6. Conclusiones

El desarrollo de métodos que puedan detectar estructuras de comunidades en redes sociales, son elementos muy importantes para poder realizar un análisis adecuado, además de que puede desvelar relaciones subyacentes entre los elementos de una red que difícilmente pueden ser vistos en la red original.

En este artículo, hemos realizado una simplificación de una red social generada a partir de un conjunto de interacciones sobre un evento importante (*trending topic*), #Videgaray en nuestro caso, el cual tomó gran importancia a nivel nacional a inicios del mes de septiembre del 2016. La detección de comunidades dentro de la red social permitió identificar un conjunto de características importantes:

- El número de comunidades que se generan internamente.
- La importancia de cada una de las comunidades determinada por el número de interacciones existentes.
- Identificación de actores importantes o influyentes dentro de la comunidad.

En general, el poder analizar las Redes Sociales, resulta un elemento muy poderoso para identificar comportamientos y actores, hacer diagnósticos, descubrir relaciones y también para detectar comunidades. Además, este análisis es susceptible de ser aplicado en áreas diversas como la biología, la política, la computación, la sociología, etc.

La identificación de comunidades presenta un desafío interesante en la precisión con la que se categorizan los múltiples actores o nodos de una red. Sin embargo, los algoritmos actualmente implementados muestran que se está avanzando por el buen camino.

El poder extraer *Hashtags* (#) a partir de una red social como Twitter, permite explorar una gran cantidad de información y determinar tendencias e interacciones existentes.

Como trabajo futuro es interesante el poder comparar la técnica utilizada en este artículo (*mapas jerárquicos*) con algunas otras para poder determinar la confiablidad y la precisión de la misma. Además de poder experimentar esta técnica con otros ejemplos de redes en áreas diversas.

Referencias

1. Aldecoa, R., Marín, I.: Deciphering network community structure by surprise. PLoS ONE, 6(9):8, (2011).
2. Cazabet, R., Amblard, F., Hanachi, C.: Detection of Overlapping Communities in Dynamical Social Networks, Proceedings of the 2010 IEEE International Conference on Social Computing, pp 309-314, (2010).
3. Delvenne, J. C., Yaliraki, S. N., Barahona, M.: Stability of graph communities across time scales, Proc. Natl. Acad. Sci. USA 107, 12755, (2010).
4. Fortunato, S.: Community detection in graphs. Physics Reports, 486(3-5):75–174, (2010).
5. Girvan, M., Newman, E. J.: Community structure in social and biological Networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12):7821–7826, (2002).
6. Kim, Y., Son, S. W., Jeong, H.: Finding communities in directed networks, Phys. Rev. E 81, 016103, (2010).
7. Newman, E. J., Girvan, M.: Finding and evaluating community structure in Networks. Physical Review E - Statistical, Nonlinear and Soft Matter Physics, 69(2 Pt 2):16, (2004).
8. Newman, E. J., Leicht, E. A.: Mixture models and exploratory data analysis in networks, Proc Natl Acad Sci USA, 104(23):9564–9569, (2007).
9. Newman, E. J.: Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA 103, 8577, (2006).
10. Palazuelos, C., Zorrilla, M.: FRINGE: A new Approach to the Detection of Overlapping Communities in Graphs. ICCSA, Lecture Notes pp 638-653, (2011).

11. Rosvall, M., Axelsson, D., Bergstrom, C. T.: The map equation. Eur Phys J Spec Top. 178:13 23, (2009).
12. Rosvall, M., Bergstrom, C.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences of the United States of America, 105(4):1118–1123, (2008).
13. Shen, H.-W., Cheng, X. Q.: Spectral methods for the detection of network community structure: a comparative analysis. Journal of Statistical Mechanics: Theory and Experiment, (2010)(10):13.
14. Smith, M. A., Shneiderman, B., Milic-Frayling, N., Mendes Rodrigues, E., Barash, V., Dunne, C., & Gleave, E.: Analyzing (social media) networks with NodeXL. In Proceedings of the fourth international conference on Communities and technologies (pp. 255-264). ACM. (2009)
15. Stanoev, A., Smikov, D., Kocarev, L.: Identifying Communities by influence dynamics in social networks. Physical Review. (2011).