Identificación de Nodos Influyentes en una Red Social de Microblogging

Diego Alonso, Luis Berdun, Ariel Monteserin y Marcelo Campo
ISISTAN Research Institute (CONICET-UNICEN) Campus Universitario, Paraje Arroyo Seco, B7001BBO, Tandil, Bs. As
{diego.alonso; luis.berdun; ariel.monteserin; marcelo.campo}
@isistan.unicen.edu.ar

Resumen. En la actualidad, el éxito de las redes sociales y microblogs ha dirigido el interés de los especialistas de diversas áreas (marketing, sociología, ciencias de la computación, entre otras) a la identificación de usuarios influyentes. En este contexto, uno de los problemas que se intenta resolver es el de maximizar la propagación de la influencia en una red social mediante la estimulación de un conjunto reducido de nodos. En este trabajo proponemos un enfoque para la identificación de usuarios influyentes que permite analizar la evolución de la influenciabilidad de los usuarios en tiempo real en una plataforma de microblogging. La viabilidad de este enfoque fue evaluada en un caso de estudio comparativo que mostró resultados alentadores.

Palabras Claves: Maximización de Influencia; Marketing Viral; Evolución de Influenciabilidad.

1 Introducción

El fenómeno de la influencia ejercida por los usuarios y su propagación en una red social ha atraído recientemente el interés de los especialistas de diversas áreas como las ciencias de la computación y el marketing [1,2,3]. Esto se debe al éxito de las redes sociales y microblogs como Facebook[1] y Twitter[2]. En particular, el servicio de Microblogging permite a sus usuarios publicar y enviar mensajes breves a través de sitios web, mensajería instantánea o smartphones. En este contexto, uno de los problemas claves del análisis de la influencia en redes sociales es la identificación de los usuarios influyentes, es decir, aquellos usuarios cuya opinión impacta, con mayor fuerza, en los gustos o acciones de otros usuarios. Para determinar cuáles son estos usuarios influyentes es necesario analizar la propagación de la influencia en las redes sociales. Este análisis es de gran utilidad ya que permite comprender, entre otras cosas, cómo se difunde la información y cómo se realiza la adopción de innovaciones en una red social [4].

[1] Facebook: https://www.facebook.com/
[2] Twitter: https://twitter.com/

En primer lugar, es importante definir qué es considerado influencia en las redes sociales. Para aclarar este punto, tomamos como definición de influencia una perspectiva basada en el seguimiento de las relaciones de los usuarios en dichas redes. Es decir, si observamos a un usuario v realizando una acción a en un tiempo t, y un usuario u (que tiene una relación con v) realiza la misma acción en un plazo corto de tiempo, digamos $t + \Delta$, entonces podemos pensar que la acción a se propagó de v a u. En caso de observar que esto sucede frecuentemente para diferentes acciones, entonces podemos concluir que el usuario v está ejerciendo influencia sobre el usuario u, convirtiéndose el usuario v en un usuario influyente [4].

Una de las diversas aplicaciones que hacen uso del fenómeno de la influencia en las redes sociales es el marketing viral. Este tipo de marketing tiene como objetivo producir incrementos exponenciales en el número de personas conocedoras de una marca con el menor esfuerzo posible [5]. Motivados por esto, en [6] plantean un problema algorítmico para las redes sociales: "si podemos intentar convencer a un subconjunto de individuos de adquirir un nuevo producto o innovación, y el objetivo es desencadenar una cascada de adopciones futuras, ¿a qué conjunto de individuos debemos apuntar?". De manera formal, Kempe define lo antes mencionado como el problema de maximización de influencia [6]. Dicho problema busca determinar k nodos (llamados semillas) en la red, de manera tal que al activarlos se maximice la propagación de la influencia esperada.

Para resolver el problema de maximización de influencia se utilizan modelos de propagación de influencia. Uno de estos modelos es el denominado Credit Distribution (CD) que calcula directamente la propagación de influencia mediante la utilización de datos históricos .

Por otra parte, el tiempo es un factor fundamental en estos análisis de influencia. Está probado que un tiempo sublogarítmico es suficiente para propagar una novedad a todos los nodos de la red [7]. También está argumentado que la naturaleza instantánea de estas redes influye en la velocidad en la que estos eventos se desarrollan [8]. Sin embargo, en la actualidad, se proponen enfoques de análisis estático de la influencia de los usuarios como los que resuelven el problema de ranking o clasificación. Es por ello que, dado el gran dinamismo de las redes sociales actuales, surge la necesidad de analizar la evolución de los usuarios influyentes en un determinado período de tiempo y sobre una temática específica.

En el contexto anterior, el presente trabajo presenta un enfoque basado en el Credit Distribution para identificar los usuarios influyentes de una red social, y analizar la evolución de la influencia de dichos usuarios de manera dinámica. Para analizar el desempeño de nuestro enfoque se realizó un monitoreo sobre la red social Twitter, bajo una temática referida a la Revolución de Mayo. El resultado de este experimento proveyó evidencia alentadora sobre la viabilidad del enfoque propuesto para la resolución del problema de maximización de influencia. El enfoque logró mejorar notoriamente la propagación de la influencia respecto de la propagación lograda por los rankings basados en seguidores.

El resto del artículo se organiza de la siguiente manera. La sección 2 describe distintos modelos de propagación de influencia existentes y explica en detalle el modelo utilizado. La sección 3 detalla el enfoque propuesto. La sección 4 muestra el caso de

estudio utilizado para evaluar el enfoque. Finalmente, la sección 5 presenta las conclusiones y trabajos futuros.

2 Propagación de Influencia en Redes Sociales

El problema de maximización de influencia es definido de la siguiente forma: dada una red social, la cual es representada por un grafo dirigido denominado Grafo Social, seleccionar un conjunto de semillas (usuarios influyentes) que maximicen la propagación de la influencia en dicha red. En este contexto, para determinar cómo se propaga la influencia, qué factores tener en cuenta y con qué probabilidad se propaga de un nodo a otro existen diversos modelos que fueron evolucionando en el tiempo. En un principio analizamos modelos como el Linear Threshold (LT) [6], el Independent Cascade (IC) [9] y el Trivalency (TR) [10]. Éstos tienen en común que en un tiempo dado (pasos discretos) cada nodo está activo o inactivo (se considera que un nodo está activo cuando el mismo ha sido influenciado o convencido de lo que se desea propagar) y su tendencia a activarse aumenta de manera monótona a medida que es mayor el número de vecinos activos. Además, estos modelos asignan probabilidades según criterios. Por ejemplo, en el LT a cada nodo se le adjunta un valor al azar entre 0 y 1 que indica el porcentaje de vecinos que deben ser convencidos para que este se active. Por su parte, en el IC cada nodo tiene una sola chance de activar a un nodo vecino con una probabilidad determinada. Por ejemplo, a cada arista que se establece con el nodo v se le asigna una probabilidad p=1/d, siendo d el grado (cantidad de aristas que lo relacionan) del nodo v. Por último, el TR asigna probabilidades a las aristas uniformemente al azar sobre un conjunto de valores que indican niveles de influencia. Por ejemplo, el conjunto {0.1; 0.01; 0.001} podría indicar alta, media y baja influencia.

Para nuestro enfoque decidimos utilizar el modelo Credit Distribution (CD). Este modelo, a diferencia de los anteriores, mediante la utilización de datos históricos calcula directamente la propagación de la influencia y obtiene mejores resultados en cuanto a precisión de semillas predichas, calidad de la selección de las mismas, tiempo de ejecución y escalabilidad que los otros modelos mencionados (IC, LT, TR) [11]. Además una de las características más importantes a destacar es que el CD tiene en cuenta los tiempos en que las acciones se propagan y los promedios de los mismos influyen al momento de determinar la influencia de los distintos usuarios. Este factor es fundamental para cumplir el objetivo del trabajo.

El modelo CD estima directamente la propagación de la influencia en base a información histórica. De esta forma, el problema de maximización de influencia para ser resuelto bajo el modelo CD es reformulado de la siguiente forma: dado un Grafo Social dirigido $G=(V,E)$, un Log de Acciones L, y un entero $k \leq |V|$, encontrar un conjunto $S \subseteq V$, $|S| = k$, tal que $\sigma_{CD}(S)$ sea maximizada. $\sigma_{CD}(S)$ es computado utilizando la ecuación $\sigma_{CD}(S) = \sum_{u \in V} K_{S,u}$, donde $K_{S,u}$ representa al total de créditos asignados a S por influenciar al usuario u en todas las acciones. Como resultado, $\sigma_{CD}(S)$ representa la influencia total propagada por los nodos incluidos en el conjunto S.

Para resolver este problema, utilizamos un algoritmo que inicialmente escanea el Log de Acciones L para aprender las probabilidades de influencia en la red social, computando los valores de influenciabilidad de los usuarios [11]. Finalmente, el con-

junto *S* (*Seed Set*) es seleccionado bajo el modelo CD utilizando un algoritmo greedy con la optimización CELF [12]. Para detalles adicionales sobre la implementación del algoritmo ver [11].

3 Enfoque

Para atacar la problemática de la identificación y evolución de nodos influyentes en una red social de Microblogging, nuestro trabajo propone un enfoque que determina minuto a minuto los usuarios más influyentes que maximizan la propagación de la influencia desde la óptica de una determinada temática. De esta forma, nuestro enfoque permite analizar la evolución de la influencia de los usuarios a medida que transcurre el tiempo. En la Figura 1 se ilustra el detalle del enfoque propuesto.

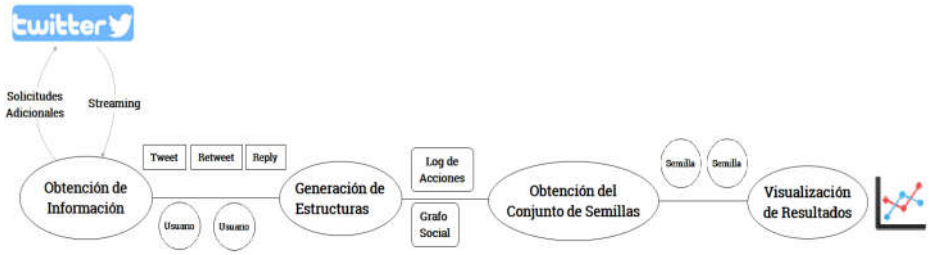

Fig. 1. Enfoque propuesto para identificación de nodos influyentes que maximicen influencia.

Como podemos observar, el enfoque parte de obtener información de una red social en tiempo real. En particular, la red social de Microblogging utilizada es Twitter. Por ende, lo que respecta a la obtención de información corresponde a escuchar *tweets, retweets* y *replies* en tiempo real y procesar dichas interacciones. Los *tweets* son las publicaciones o mensajes de los usuarios, los *retweets* son re-publicaciones que realizan los usuarios de los mensajes de otros, y las *replies* son las respuestas de usuarios a las publicaciones de otros o a sus propios mensajes. De esta forma un *tweet* constituye una acción, mientras que un *retweet* o una *reply* será considerado una acción propagada.

Sin embargo, como se va a trabajar sobre un tema y no sobre la totalidad de las publicaciones, no todas las interacciones escuchadas son consideradas válidas sino que deben respetar un conjunto de filtros especificados por un experto para definir el tema sobre el que se trabajarà. Para ello, se establece un conjunto de palabras claves o *keywords* que define la categoría de interés, de manera que sólo serán válidas las interacciones que contengan dichas palabras. Por ejemplo, si se desea analizar la influencia en un dominio relacionado a la Selección Argentina de fútbol, el conjunto de palabras podría ser {fútbol; selección argentina; messi}. Por otro lado, se realiza un filtrado relacionado directamente con el factor temporal del enfoque. Con este se busca sólo analizar la información que se considera reciente y dentro de una ventana temporal de validez. Por ejemplo, considerar como válida una interacción cuyo origen no supere los 60 minutos respecto del tiempo actual. Es decir, en caso de identificar una interac-

ción (*retweet* ó *reply*) que refiera a un *tweet* realizado en un plazo mayor a los 60 minutos previos, dicha interacción será descartada.

Una vez obtenida la información de la red social se procede a establecer las relaciones entre los usuarios involucrados en las interacciones en un grafo dirigido (Grafo Social) y adjuntar dichas interacciones en un registro de propagaciones (Log de Acciones). Esta etapa corresponde a la Generación de Estructuras en la Figura 1. El Grafo Social es una representación basada en nodos (usuarios) y aristas (relación). Donde las aristas se establecen acorde a la dirección de origen de la acción propagada. Es decir, "interacción" → "acción original". El grafo se construye de manera incremental con el arribo de las interacciones. Tras el arribo de una nueva interacción, primero se verifica si el usuario (que realiza la acción) se encuentra ingresado en el grafo. En caso de no estar ingresado, es creado y agregado al mismo como un nodo aislado (sin relaciones). En caso de ser un *retweet* o un *reply* (interacciones que hacen referencia a un *tweet*), se procede a verificar también la existencia del usuario de la acción original. Análogamente al caso del usuario que realiza la interacción, el usuario que llevó a cabo la acción original es agregado en caso de no existir en el grafo. Adicionalmente, en este caso se establece la arista dirigida entre el usuario que propagó la acción y el que la realizó originalmente. Sin embargo, en la definición de influencia descrita en la sección 1 se enuncia que debe existir una relación entre los usuarios para considerar que se propaga influencia. Es decir, no alcanza con que sólo se propague una acción de uno al otro para relacionarlos en el grafo, sino que hay que verificar que exista una conexión entre ambos. En la plataforma Twitter, las relaciones se basan en seguidores y seguidos, que es similar a una suscripción a las publicaciones del usuario que se sigue. De esta forma, previo a establecer las aristas entre los nodos, se verifica si el usuario de la interacción está socialmente vinculado con el usuario de la acción original. En caso de ser así, se establece efectivamente la arista como se mencionó con anterioridad. Sin embargo, debido a que en la plataforma Twitter se permite realizar un *retweet* de otro *retweet*, puede que no exista un vínculo directo entre el usuario que interactúa y el original, sino que esto se realiza a través de uno o más usuarios intermedios. Es por ello que, se propone analizar el camino establecido entre el usuario de la interacción y el original, agregando los usuarios intermedios que sean necesarios. A fines prácticos se analiza si el usuario que interactúa está relacionado socialmente con otro usuario que haya realizado el mismo *retweet*. Este proceso se repite recursivamente hasta encontrar un vínculo directo entre un usuario que interactúa y el que realizó la acción originalmente.

Mientras se lleva a cabo la incorporación del usuario al Grafo Social, se registra también la interacción en cuestión en el Log de Acciones o Registro de Propagaciones. Cada entrada en dicho Log está compuesta por los campos: usuario, acción realizada y tiempo (en que fue realizada). Por ejemplo, se identifica que el usuario D realiza la acción 20 (un *tweet*) a las 18.45hs y al mismo tiempo un usuario F realiza la acción 21 (un *tweet*). Luego, el usuario I realiza la acción 20 (un *retweet* del *tweet* 20) a las 19.05hs. El Log de Acciones quedaría representado por: <D, 20, 18.45>, <F, 21, 18.45>, <I, 20, 19.05>.

Con el Grafo y el Registro de Propagaciones se procede a la etapa de Obtención de semillas (ver Figura 1). Es en dicha etapa donde se ejecuta el algoritmo de maximiza-

ción de influencia bajo el modelo CD con los datos contenidos en el Grafo Social y el Log de Acciones, y se obtiene el conjunto de semillas S de la red que maximiza $\sigma_{CD}(S)$. El conjunto de semillas se obtiene cada cierta ventana temporal. Es por ello, que con el avance del tiempo, el enfoque permite visualizar la evolución de la posición de los usuarios en el conjunto de semillas. Dicha posición representa la influenciabilidad del usuario. El orden en el conjunto de semillas se establece según la influencia total alcanzada por el nodo (semilla).

De esta forma el enfoque permite, minuto a minuto, identificar nodos influyentes en una determinada temática con el objetivo de maximizar la propagación de la influencia y además visualizar la evolución de la influenciabilidad de los usuarios con el transcurso del tiempo.

4 Resultados Experimentales

Para analizar la viabilidad del enfoque propuesto sobre el problema de la maximización de la influencia en redes sociales decidimos contrastar los resultados obtenidos por el presente enfoque con la perspectiva de los rankings. Es decir, mientras seleccionamos el conjunto de semillas que maximiza la influencia según el enfoque propuesto, simultáneamente, se selecciona un conjunto de semillas acorde al ranking de los usuarios. Para realizar esto último, se define un ranking R con los usuarios observados con más seguidores, dado que es éste uno de los criterios más utilizados para su conformación. Finalmente, para el ranking R calculamos $\sigma_{CD}(R)$, es decir, calculamos la influencia total propagada por los usuarios de R bajo el modelo CD. De esta forma constatar que el enfoque propuesto mejora el aporte por los rankings de seguidores.

Para poder realizar un estudio comparativo entre ambos enfoques previamente definimos una categoría de interés para monitorear interacciones en tiempo real. Con el objetivo de mostrar un caso reciente determinamos una temática relacionada con el 25 de mayo y lo que a esta fecha refiere. Para ello, seleccionamos las *keywords* pertenecientes al siguiente conjunto: {revolución de mayo; mayo de 1810; cabildo; virreinato del río de la plata; primera junta; fernando VII; hidalgo de cisneros; french y beruti; plaza de la victoria; plaza de mayo; cornelio saavedra; miguel de azcuénaga; mariano moreno; manuel belgrano; juan josé castelli; juan josé paso}.

El día 25 de mayo se monitoreó en dos brechas temporales consecutivas de aproximadamente 12 horas y se compararon los valores de influencia obtenidos por cada conjunto de usuarios influyentes según cada enfoque.

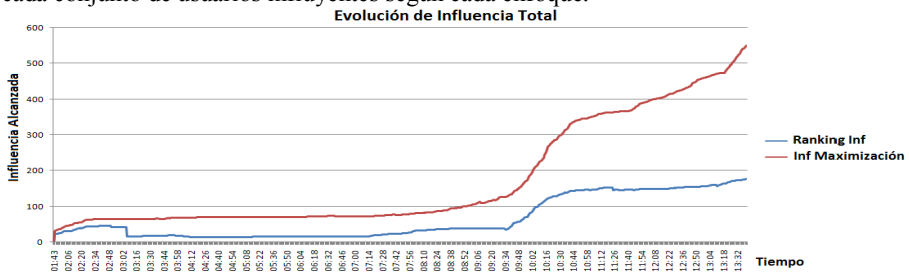

Fig. 2. Evolución de influencia total alcanzada en el primer monitoreo.

En la Figura 2 se pueden observar los resultados obtenidos tras la primera escucha de interacciones. Durante las primeras horas, el flujo de interacciones escuchadas fue relativamente bajo, y la diferencia entre la influencia alcanzada por el ranking de seguidores y el conjunto de semillas del enfoque propuesto es más estrecha (ver Figura 2). Sin embargo, en las horas siguientes, cuando el flujo de interacciones escuchadas aumentó notablemente, la diferencia entre la influencia total de un conjunto y otro es notoriamente mayor.

En la Figura 3, se muestran los resultados obtenidos tras la segunda escucha de interacciones. Cabe aclarar que, durante este monitoreo, el flujo de interacciones escuchadas fue parejo. En la Figura 3, se puede apreciar que la diferencia entre la influencia alcanzada por un conjunto de usuarios y otro es significativamente grande pero similar durante el transcurso del tiempo. Aunque es interesante destacar la disminución notoria que se identifica en el lapso de las 19.35 y las 19.49. Este descenso está relacionado con que se añade al conjunto de influyentes un usuario cuya cantidad de seguidores es mayor a la de alguno de los que se encuentran en el ranking. El problema es que la influencia que este usuario aporta no supera la del usuario que cede su lugar en el ranking. Esto ilustra acentuadamente que un usuario con muchos suscriptores no indica una mejora en la propagación de la influencia en un tema.

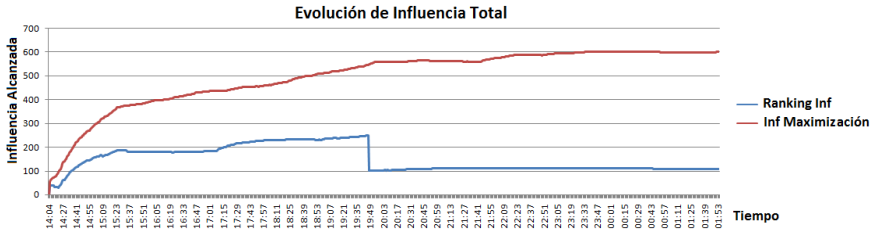

Fig. 3. Evolución de influencia total alcanzada en el segundo monitoreo.

Los resultados obtenidos en ambos monitoreos evidencian la viabilidad de la utilización del presente enfoque para la resolución del problema de la maximización de influencia en tiempo real. Basados en ello, destacamos que el enfoque es además viable para analizar la evolución de los usuarios influyentes en el transcurso del tiempo, ya sea en brechas temporales cortas o extensas. Esto se debe a que nuestro enfoque permite visualizar los crecimientos y decrecimientos de la influenciabilidad de los usuarios de forma dinámica, ilustrando las interacciones que realiza y la influencia alcanzada sobre la red minuto a minuto.

5 Conclusiones

En este trabajo presentamos un enfoque orientado a la resolución del problema de la maximización de la influencia en las redes sociales con el objetivo de poder analizar la evolución de los usuarios influyentes de forma dinámica.

Los estudios y experimentos realizados demuestran la viabilidad del enfoque para la determinación en tiempo real de los usuarios que maximizan la propagación de la

influencia. De este modo, este trabajo tiene una aplicación directa, por ejemplo, en marketing viral, siendo útil al momento de seleccionar los usuarios que se deben convencer para maximizar la propagación de una novedad o producto a través de una red social. Por otra parte, al permitir visualizar la evolución de los usuarios influyentes, el presente trabajo también resulta de utilidad si la selección de usuarios a convencer se desea realizar luego de una brecha temporal y no momentáneamente.

Como trabajos futuros, se planea continuar con distintas líneas de investigación que se focalicen en la obtención de información complementaria de interés sobre los usuarios influyentes. Entre otras cosas, se propone el agregado de clasificadores de texto para analizar en mayor profundidad el contenido textual de las interacciones de los usuarios influyentes, con el fin de determinar si existen palabras que ayudan a influenciar a otros usuarios o si hay cierto contenido en particular en el que el usuario es más referenciado. Por otro lado, se propone añadir diversas fuentes de información (no limitarse a la red social Twitter) con el objetivo de realizar un análisis más robusto de los usuarios. Para esta propuesta se debe previamente definir qué acciones en la red a añadir serán de interés para analizar su propagación. Por ejemplo, si la red fuese Facebook, las acciones podrían ser compartir o comentar una publicación.

Referencias

1. J. Weng, E.P. Lim, J. Jiang, and Q. He, *"Twitterrank: finding topic sensitive influential twitterers"* in Proc. of the Third Int. Conf. on Web Search and Web Data Mining, 2010.
2. E. Bakshy, J. M. Hofman, W. A. Mason and D. J. Watts, *"Everyone's an influencer: quantifying influence on twitter"* in Proc. of the Fourth Int. Conf. on Web Search and Web Data Mining, 2011.
3. D. M. Romero, B. Meeder and J. M. Kleinberg, *"Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter"* in Proc. of the 20th Int. Conf. on World Wide Web, 2011.
4. Francesco Bonchi, *"Influence Propagation in Social Networks: A Data Mining Perspective"* , Barcelona, Spain, Dec 2011.
5. I.B. López, A. D. Shewak, N. M. Sánchez y C. T. Coma, *"Proyecto Kimbi: Marketing Viral"*, Facultad Cs Económicas y Empresariales, Universidad Pompeu Fabra, España, 2008.
6. David Kempe, Jon Kleinberg, and Éva Tardos, *"Maximizing the Spread of Influence through a Social Network"*, Dept. of Computer Science Cornell University, Ithaca NY.
7. Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich, *"Why Rumors Spread Fast in Social Networks"*, Max-Planck-Institut für Informatik and Universität des Saarlandes, Saarbrücken, Germany, 2012.
8. P. Beaumont, *"The truth about twitter, facebook and the uprisings in the arab world"*, 2011.
9. J. Goldenberg, B. Libai and E. Muller, *"Using Complex Systems Analysis to Advance Marketing Theory Development"*, Academy of Marketing Science Review, 2001.
10. W. Chen, C. Wang, and Y. Wang, *"Scalable Influence Maximization for Prevalent Viral Marketing in Large Scale Social Networks"*, in KDD'10, 2010.
11. Amit Goyal, Francesco Bonchi, and Laks V. S. Lakshmanan, "*A Data Based Approach to Social Influence Maximization"*, University of British Columbia, Canada, 2011.
12. J. Leskovec, "*Cost-effective outbreak detection in networks"*, in KDD'07, 2007.