Extracting Top-k Most Influential Nodes by Activity Analysis

Myungcheol Doo
Applied Research Center, Arris
2400 Ogden Ave, Lisle, Illinois, USA
myungcheol.doo@arrisi.com

Ling Liu
School of Computer Science
Georgia Institute of Technology
266 Ferst Dr., Atlanta, Georgia, USA
ling.liu@cc.gatech.edu

Abstract

Can we statistically compute social influence and understand quantitatively to what extent people are likely to be influenced by the opinion or the decision of their friends, friends of friends, or acquaintances? An in-depth understanding of such social influence and the diffusion process of such social influence will help us better address the question of to what extent the 'word of mouth' effects will take hold on social networks. Most of the existing social influence models to define the influence diffusion are solely based on topological connectivity of social network nodes. In this paper, we presented an activity-base social influence model. Our experimental results show that activity-based social influence is more effective in understanding the viral marketing effects on social networks.

1 Introduction

A social network is typically modeled as a graph of people nodes connected through friend relationships or interactions among people. Facebook, LinkedIn, and Twitter are popular social networks that not only serve as a meeting point but also an important medium for the spread of information and influences among its members. An important property of social influence is the dynamics in terms of how influence evolves and which type of influence makes fast and persistent inroads into the population of the network or dies out quickly. Can we statistically compute social influence and understand quantitatively or qualitatively to what extent people are likely to be influenced by the opinion, the action, or the decision of their friends, friends of friends, or acquaintances? An in-depth understanding of such social influence and the diffusion process of such social influence will help us better address the question: to what extent the 'word of mouth' effects will take hold (i.e., one is being influenced by the social networks to which it belongs) on social networks.

1.1 Topological Diffusion Model

An intuitive approach to study the process of influence diffusion over social networks is the topological diffusion model. It examines how the spread of influence is carried out through the topological relationships among people nodes in a social network. For example, Kempe [11] characterizes the state of art influence research in social sciences into two basic diffusion models: Linear Threshold model (LT) and Independent Cascade model (IC). Both models classify people nodes into active and inactive. Given network $G = (V, E)$, LT model [3, 10, 11, 14, 16] requires each node u randomly to choose a weight $w(u, v)$ over an edge $E(u, v)$ from the interval [0,1]. $w(u, v)$ is the influence of u over v. Each node is either inactive or active. Then we set a system defined threshold θ which will be used to determine if a node switches from being inactive to active. An inactive node v can be switched to active when the total weight $\sum w(u, v)$ is greater than or equal to θ, where u is one of active neighbors of v. This process repeats on each newly activated node in V. In contrast, IC model [8, 9, 11] uses a probabilistic approach. A newly activated node u is given one chance to activate its inactive neighbor nodes with a probability $p_{u,v}$. This process repeats until no more activation is possible (reaching convergence).

In addition to two models, heat diffusion has been studied as another basic topological diffusion model. Mutual influence can readily be modeled as a heat diffusion process [12, 13, 19]. At initial time t_0, all nodes has zero heat. In a social network of n nodes, one node v_i is selected and given some amount of heat. At t_1, v_i diffuses its heat to all of its neighbors equally. At t_2, nodes with non-zero heat diffuse their heat to their neighbors. This routine repeats for a certain time period t. Then we know how many nodes are influenced from v_i by counting the number of nodes whose heat value is greater than or equal to a system defined value. By repeating this process for all n nodes, we find the influence of each node.

1.2 Problems with Topological Diffusion Models

Most of the existing social influence models, to the best of our knowledge, define the influence diffusion solely based on the topological connectivity of social network nodes. We argue that social influence among people is not only determined by the social connectivity but also the amount of activities carried out by a node and the volume of interactions between two social network nodes. It is evident in real world that those people that have more activities in social network typically have higher level of social influence on their neighbor nodes and the friends of their friends than those who are significantly less active. In addition, two people who interact more frequently in a social network will have higher influence on each other than two people who have not had many interactions. Bearing these observations in mind, we design and develop an activity-based social influence model and a suite of activity-based influence ranking algorithms.

2 Related Work

TwitterRank [18] measures the influence by taking into account both the topical similarity between users and the link structure. On the other hand, Anger [1] reveals that highly reciprocal social network structure cannot be observed with the top 10 Twitter users in Austria. This is due to the asymmetric phenomenon of social influence. The most popular super hubs are followed by many users, but the super hubs are seldomly following those users who are their followers.

Ma et al [13] and Bao et al [2] studied social influence in terms of the heat diffusion phenomena such that the influence spreading over the social network can be modeled as heat flowing along the links from a social network node to another at some diffusion rate. After several iterations of such topological heat diffusion, those users who spread heat most are selected as the most influential nodes in the give social network. Unfortunately, the heat diffusion process utilizes only the topological structure of the given social network. As a result, social network nodes with less neighbor but more activities will receive lower ranking values.

3 ACTIVITYINFLUENCE Model

3.1 Social Network Attributes

Intuitively, social network nodes that have much more activities typically have higher level of social influence on their neighbor nodes than those that are significantly less active even though they may have larger number of friends (neighbor nodes) and thus higher node degree. Thus, the number of activities that a node has performed should be an important indicator of social influence in addition to topological structure of nodes.

Furthermore, we argued that one pair of social network nodes that interact more frequently will have higher social influence on each other than another pair of nodes that have significantly fewer interactions recently. This leads us to introduce the number of interactive activities between two nodes as another important indicator of social influence. We call the activities performed at each node the non-interactive activities and the activities performed via interactions between a pair of nodes the interactive activities in this paper.

As a final remark, we would like to point out the importance of asymmetric influence between a pair of social network nodes. In the context of topology based social influence model, the simplest way of constructing a social network graph is to create a vertex for each user and connect two vertices if any two users are friends. All the edges have an equal weight and thus the edge weight is insignificant. This approach assumes that influence is symmetric between two nodes. However, the equality of influence between a pair of nodes does not always exist in real world. The social influence between a pair of social network nodes are not symmetric, no matter whether the two nodes are mutual friends or one is a follower of another. For example, a user s is a famous singer and f is one of her fans. Then f is keen on being informed about s's activities but not vice versa. Similarly, two friends s and f may not have equal influence on one another if f is less active and s is highly active in terms of conducting non-interactive and interactive activities.

Number of Friends Measuring the social influence based on the number of friends is the easiest and the most common way. We refer to this type of social influence as popularity based or friendship based social influence to differentiate it from the general concept of social influence, which should capture both topological structure and activity based semantics of social network nodes. Indeed, highly influential people often have more followers than less or non-influential people. Thus node degree (number of friends) is an important measure for computing social influence of nodes in a social network. However, the number of friends should not be used as the best and the only indicator of the level of influence. Some people report that they cannot reject friend requests from their clients [17] or they accept friend requests from even vaguely recognized people [4] to increase their popularity. Therefore, the influence diffusion model, which is based solely on the number of friends and the uniform distribution of influence among all friends of a node, will fail to capture how social influence is actually diffused in real life.

Activities In reality, friendship is not the only contextual feature that most social network services provide. Users can post photos, videos, and reviews on any subject. For friends' posting, users can vote for 'like' or write a comment. We categorize user activities into two groups: *interactive* activities (such as comments on friends postings) and *non-interactive* activities (such as reviews on a product

or tips for programming). Interactive activities are user activities that involves another social network node than self. Otherwise, activities are considered as non-interactive. For example, Alice posts a photo on her profile page. This activity is non-interactive because it involves only herself and no one else. If another social network node Bob leaves a comment on Alice's photo. Then this activity is an interactive activity between Bob and Alice because it involves two nodes in the social network.

Figure 1 shows an example of social network of 10 users, $v_1, v_2, \ldots, v_{10}$. Each user is represented as a vertex. Each node may have performed a number of non-interactive activities, such as posting of reviews, photos or videos. The underlined positive integer attached to each node represents the number of non-interactive activities performed by the node. If two users are friends, then there is an edge between two vertices. The positive integer attached to each edge connecting two nodes denotes the number of interactive activities performed between them. For example, v_1 in the middle of nodes has 70 postings, v_2 has 40 postings, and v_3 has 3 postings. There are 80 interactive activities between v_1 and v_2. However, there are only 4 activities between v_1 and v_3.

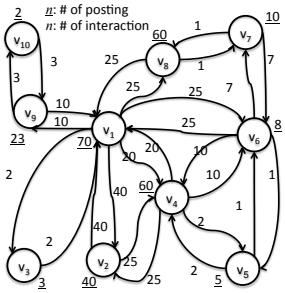


Figure 1. Example Social Network

3.2 Heat Diffusion Kernel

Heat diffusion is a physical phenomenon such that heat always flows from an object with high temperature to an object with low temperature. In physics, the heat capacity is a thermodynamic property; i.e., the heat capacity is a measure of how much an intensive thermodynamic variable (temperature) changes when a small amount of energy is added or subtracted from the sample. Thermal conductivity is a transport property; the thermal conductivity, α, is the linear transport coefficient that relates a temperature gradient to a heat flux. In most cases of practical interest, the α is a tensor and is diagonal. The ratio of the thermal conductivity and heat capacity per unit volume, C, is the thermal diffusivity, $D = \frac{\alpha}{C}$. We can get a rough idea of the time t it takes for heat to diffuse some distance L from dimensional analysis and t is increasing with increase of L and decrease of D. Thus, the time scales of heat diffusion in practical situations varies enormously. In scientific studies, the relevant time scales of heat diffusion span an amazing 27 orders of magnitude. For example, some physicists study heat diffusion in thin films on 10 picosecond time scales and planetary scientists are concerned with the diffusion of heat on the time scale of billions of years.

In a large social network graph, heat diffusion kernel can be used to model to social influence diffusion process in the social network graph. Let $G = (V, E)$ denote a social network graph where $V = \{v_1, v_2, ..., v_n\}$ is the set of vertices representing users and $E = \{(u, v)|u, v \in V\}$ is a set of edges representing friend relationship between users. Let α denote the thermal conductivity (the heat diffusion coefficient) on G. The heat on vertex v_i at time t is represented as a function $H_i(t)$ and heat flows from a high temperature node to a low temperature node following the edges between vertices. On a directed graph, for the duration Δt, v_i diffuses its heat, denoted by $DH_i(\Delta t)$, through its outgoing edges, and receives heat, denoted by $RH_i\ (\Delta t)$, through its incoming edges. The heat at vertex $v_i \in V$ between t and $t + \Delta t$ is defined by the sum of the differences between the heat that it receives from, and the heat it diffuses to, all its neighbors, and is formulated as follows:

$$H_i(t + \Delta t) - H_i(t) = RH_i(\Delta t) - DH_i(\Delta t) \quad (1)$$

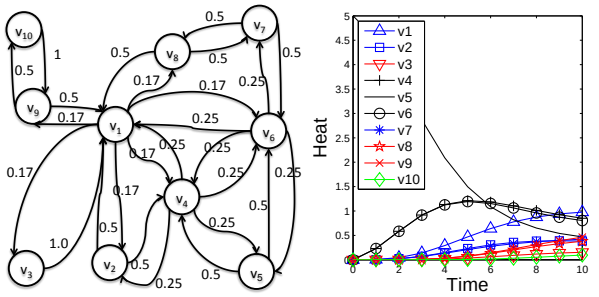
This formula shows that a number of factors impacts on $H_i(t + \Delta t) - H_i(t)$, including the heat conductivity α, the heat at vertex v_i, the duration of heat diffusion process Δt, and the number of friends of v_i, denoted by d_i.

Each heat distribution medium has different heat conductivity α, which is a real number between 0 and 1. If α is approaching zero, then the medium barely transfer any heat. If α is approaching the upper bound of 1, it transfers heat without loss and the speed of heat diffusion is faster than when α is close to the lower bound of 0. Each social network has different speed of information diffusion and so does each type of products or opinions. In this paper we set α to be 1 to allow us to focus more on the methods to define heat at a node and heat diffusion on weighted edges by taking the assumption that the more heat a social network node v_i has, the more heat v_i can diffuse.

The amount of heat transfer depends on the topology of graph, time of period, and the amount of heat at source node. If v_i is directly reachable from v_j through an edge in $E(v_j, v_i)$, then the more heat v_j has, the more heat v_i receives, and the longer the heat diffusion process takes, the more heat v_i receives from its neighbors and v_j diffuses to its neighbors. Lastly, the number of friends of v_i, denoted by d_i, is the out-degree of vertex v_i, another important factor for influence diffusion. In the topology based heat diffusion model where heat at a vertex is uniformly distributed to all its neighbors, each neighbor node of v_i receives $\frac{1}{d_j}$ of heat that v_i diffuses.

Figure 2(a) shows an example extracted from the sample social network given in Figure 1. Weight on edges are computed based on the uniform distribution, namely $\frac{1}{d_i}$ of the heat of vertex v_i will be diffused to each of its neigh-

bors. For example, v_5 in the lower right of vertices has two friends, v_4 and v_6. Hence, the weight of 0.5 is assigned to $E(v_5, v_4)$ and $E(v_5, v_6)$ respectively and one half of v_5's heat is transmitted to v_4 and the other half is transferred to v_6. v_{10} in the upper left of vertices has only one friend, v_9. Therefore, the weight on $E(v_{10}, v_9)$ is 1 and all of v_{10}'s heat is diffused to v_9. In summary, the amount of heat v_i dif-



(a) Constructed Graph (b) Heat

Figure 2. Heat Diffusion (Topology)

fused to its neighbors, $DH_i(\Delta t)$, is proportional to α, Δt, and $H_i(t)$. The amount of heat that v_i received from her neighbors, $RH_i, (\Delta t)$ is proportional to α, Δt, and $\frac{H_j(t)}{d_j}$. we can formulate $DH_i(\Delta t)$ and $RH_i(\Delta t)$ as follows:

$$DH_i(\Delta t) = \alpha \Delta t H_i(t) \tag{2}$$

$$RH_i(\Delta t) = \alpha \Delta t \sum_{j:(v_j,v_i)\in V} \frac{H_j(t)}{d_j}. \tag{3}$$

We can plug Eq. (2) and (3) into Eq. (1) then we have the following form:

$$H_i(t+\Delta t) - H_i(t) = \alpha \Delta t \Big(\sum_{j:(v_j,v_i)\in V} \frac{H_j(t)}{d_j} - H_i(t) \Big). \tag{4}$$

Consider the example in Figure 2(a), we have:

$$\begin{aligned}
H_1(t+\Delta t) - H_1(t) =& \alpha \Delta t \Big(- H_1(t) + 0.5H_2(t) + H_3(t) \\
&+ 0.25H_4(t) + 0.25H_6(t) + 0.5H_8(t) \\
&+ 0.5H_9(t) \Big) \\
H_2(t+\Delta t) - H_2(t) =& \alpha \Delta t \Big(0.17H_1(t) - H_2(t) + 0.25H_4(t) \Big) \\
H_3(t+\Delta t) - H_3(t) =& \alpha \Delta t \Big(0.17H_1(t) - H_3(t) \Big) \\
&\vdots \\
H_{10}(t+\Delta t) - H_{10}(t) =& \alpha \Delta t \Big(0.5H_9(t) - H_{10}(t) \Big).
\end{aligned}$$

The above heat difference can be represented in a matrix form as follows:

$$\begin{aligned}
&H(t+\Delta t) - H(t) = \\
&\alpha \Delta t \begin{pmatrix}
-1 & 0.5 & 1.0 & 0.25 & 0 & 0.25 & 0 & 0.5 & 0.5 & 0 \\
0.17 & -1 & 0 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.17 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.17 & 0.5 & 0 & -1 & 0.5 & 0.25 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.25 & -1 & 0.25 & 0 & 0 & 0 & 0 \\
0.17 & 0 & 0 & 0.25 & 0.5 & -1 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.25 & -1 & 0.5 & 0 & 0 \\
0.17 & 0 & 0 & 0 & 0 & 0 & 0.5 & -1 & 0 & 0 \\
0.17 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & -1
\end{pmatrix} H(t) \\
&= \alpha \Delta t K H(t)
\end{aligned} \tag{5}$$

where K is a $n \times n$ matrix whose element $K(i,j)$ is defined as in Eq. (6) and $H(t)$ is a column vector of size n defined in Eq. (7). In this example $n = 10$.

$$K(i,j) = \begin{cases} \frac{1}{d_j} & (v_j, v_i) \in E \\ -1 & i = j \text{ and } d_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$H(t) = \begin{Bmatrix} H_1(t) \\ H_2(t) \\ H_3(t) \\ \dots \\ H_{10}(t) \end{Bmatrix} \tag{7}$$

We can transform Eq. (5) as follows:

$$\begin{aligned}
& H(t+\Delta t) - H(t) = \alpha \Delta t K H(t) \\
\Leftrightarrow & \frac{H(t+\Delta t) - H(t)}{\Delta t} = \alpha K H(t) \\
\Leftrightarrow & \lim_{\Delta t \to 0} \frac{H(t+\Delta t) - H(t)}{\Delta t} = \alpha K H(t) \\
\Leftrightarrow & \frac{d}{dt} H(t) = \alpha K H(t)
\end{aligned} \tag{8}$$

.

Eq. (8) is a linear homogenous differential equation and can be solved as follows:

$$H(t) = e^{\alpha t K} H(0) \tag{9}$$

where K denotes the heat diffusion kernel of G and $H(0)$ denotes the initial heat distribution column vector at time zero on G. Eq. (9) defines the vertex's thermal capacity at time t by an exponential function $H(t)$ with independent variable t for the initial heat source $H(0)$. The matrix $e^{\alpha t K}$ is called the propagating heat diffusion kernel and can be represented as a Taylor series:

$$\begin{aligned}
H(t) =& e^{\alpha t K} H(0) \\
=& (I + \alpha t K + \frac{\alpha^2 t^2}{2!} K^2 + \frac{\alpha^3 t^3}{3!} K^3 + \dots) H(0) \\
=& I + \sum_{n=1}^{\infty} \frac{\alpha^n t^n}{n!} K^n H(0)
\end{aligned} \tag{10}$$

where $n!$ denotes the factorial of n and $0!$ is defined to be 1. $K^{(n)}$ denotes the nth derivative of K evaluated at the point t and the zeroth derivative of K is defined to be K itself.

Figure 2(b) shows the result of heat diffusion using Eq. (9). v_5 is selected as a heat source. x-axis is the time line and y-axis is the amount of heat at each node v_i. v_4 and v_6 are nodes that are 1-hop away from v_5. They evenly receive heat from v_5. Thus their heat graphs are the same. 2-hop away nodes, v_1, v_2, v_7, have less amount of heat than 1-hop away nodes but larger than 3 or more -hop away nodes. Although v_1 is 2-hop away node, it is receiving from three nodes. Thus its heat is higher than other two 2-hop away nodes.

3.3 Activity-based Heat Diffusion

3.3.1 Design Guidelines

Let IA_{ij} denote the number of interactive activities from node v_i to its neighbor node v_j and NIA_i denote the number of non-interactive activities at node v_i. There are several ways one can extend the heat diffusion kernel to incorporate both interactive and non-interactive activities. We choose to extend the basic heat diffusion kernel in two steps.

First, we argue that non-interactive activities at node v_i may play a role as heat source. The more non-interactive activities a node v_i has performed, the more heat is added to v_i and consequently v_i is losing its heat at a slower pace in the diffusion process. Concretely, let NA_i denote the number of non-interactive activities at node v_i. We define $MAX(NA)$ as the largest number of non-interactive activities in V. Thus, the diffused influence at vertex v_i, denoted by $DH_i(t)$, is augmented by the number of non-interactive activities at v_i normalized by $MAX(NA)$, denoted by $\frac{NA_i}{MAX(NA)}$. In order to express that v_i loses its heat slower when it has larger number of NA_i, we set $DH_i(t)$ is proportional to $1 - \frac{NA_i}{MAX(NA)}$.

Second, we argue that the amount of influence received by node v_i from one of its neighbors, say v_j, should be proportional to the number of interactive activities that v_j has performed with v_i normalized by the total number of interactive activities that v_j has performed with all of its neighbors, namely $\frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}}$. Thus $RH_i(t)$ is proportional to $\frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}}$.

Consider the example in Figure 1, v_2 has 80 interactions with v_1 and 50 with v_4. Instead of following topology based heat diffusion where v_1 and v_4 receive equal amount of heat from v_2, namely $\frac{H_2(t)}{2}$, we define that v_1 receives $\frac{80}{80+50}$ of $H_2(t)$ and v_4 receives $\frac{50}{80+50}$ of $H_2(t)$. Similarly, we can consider non-interactive activities. NA_1 is 70, NA_{10} is 2, and $MAX(NA)$ is 80. Then we can normalize NA_1 as $\frac{70}{80}$ and NA_{10} is normalized as $\frac{2}{80}$. Then the amount of heat v_1 loses is $(1-\frac{70}{80})H_1(t)$ and v_{10} loses $(1-\frac{2}{80})H_{10}(t)$ amount of heat. Thus v_{10} loses its heat faster than v_1.

3.3.2 Activity based Diffusion Kernel

In this section we formally define the activity based diffusion kernel by taking into account of both interactive and non-interactive activities.

Considering the above example we can formulate as follows:

$$RH_i(\Delta t) = \alpha \Delta t \sum_{j:(v_j,v_i)\in E} H_j(t)\left(\frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}}\right) \quad (11)$$

$$DH_i(\Delta t) = \alpha \Delta t H_i(t)\left(1 - \beta \frac{NA_i}{MAX(NA)}\right) \quad (12)$$

where β is a real number between 0 to 1 and served as the weight for non-interactive activities. If β is set to 0 then non-interactive activities are ignored. If β is set to 1, then $DH_i(\Delta t)$ may become zero when $MAX(NIA) = NA_i$.

By plugging Eq. (11) and (12) into the heat difference during Δt, defined in Eq. (1), we have the activity-based heat diffusion formula as follows:

$$\begin{aligned} & H_i(t+\Delta t) - H_i(t) \\ =& RH_i(\Delta t) - DH_i(\Delta t) \\ =& \alpha \Delta t \sum_{j:(v_j,v_i)\in E} H_j(t)\left(\frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}}\right) + \\ & \alpha \Delta t H_i(t)\left(1 - \beta \frac{NA_i}{MAX(NA)}\right) \end{aligned} \quad (13)$$

Eq. (13) is transformed into matrix form as follows:

$$\alpha \Delta t K H(t) \quad (14)$$

, where K is a $n \times n$ matrix as defined as follows:

$$K(i,j) = \begin{cases} \frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}} & (v_j,v_i) \in E \\ 1 - \beta \frac{NA_i}{MAX(NA)} & i = j \text{ and } d_i > 0 \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

We also define node activity-based diffusion kernel and node interaction-based diffusion kernel. If there are only node activity information and no node interaction, then $RH_i(\Delta t)$ is the same as one in the topological heat diffusion. In concrete, we use $DH_i(\Delta t)$ in Eq. (12) and $RH_i(\Delta t)$ in Eq. (3):

$$DH_i(\Delta t) = \alpha \Delta t H_i(t)\left(1 - \beta \frac{NA_i}{MAX(NA)}\right)$$

$$RH_i(\Delta t) = \alpha \Delta t \sum_{j:(v_j,v_i)\in V} \frac{H_j(t)}{d_j}.$$

. Then we modify K as follows:

$$K(i,j) = \begin{cases} \frac{1}{d_j} & (v_j,v_i) \in E \\ 1 - \beta \frac{NA_i}{MAX(NA)} & i = j \text{ and } d_i > 0 \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

On the other hand, if there are only node interaction information and no node activity information, then $DH_i(\Delta t)$ is the same as one in the topological heat diffusion. In concrete, we use $DH_i(\Delta t)$ in Eq. (2) and $RH_i(\Delta t)$ in Eq. (11):

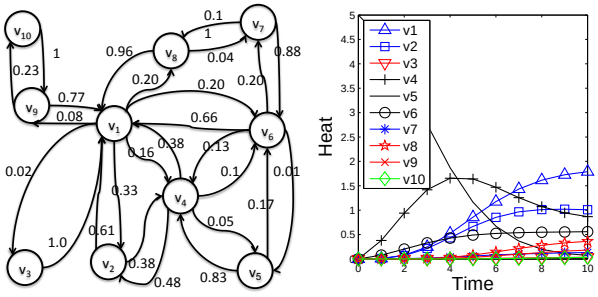
$$DH_i(\Delta t) = \alpha \Delta t H_i(t)$$

$$RH_i(\Delta t) = \alpha \Delta t \sum_{j:(v_j,v_i)\in E} H_j(t)\left(\frac{IA_{ji}}{\sum_{m:(v_j,v_m)\in E} IA_{jm}}\right)$$

and K is also defined as follows:

$$K(i,j) = \begin{cases} \frac{IA_{ji}}{\sum_{k:(v_j,v_k)\in E} IA_{jk}} & (v_j, v_i) \in E \\ 1 & i = j \text{ and } d_i > 0 \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

Figure 3(a) shows K as weights on edges computed using the above formula. Weights on Figure 3(a) is different from those in Figure 2(a). Figure 3(b) shows activity based heat diffusion result. From the heat source v_5, v_4 receives the largest amount of heat (83%) and its heat increases faster followed by v_6, which receives 17% of v_5's heat. Note that in Figure 2(a), v_4 and v_6 receive the same amount of heat from v_5, but now the ratio is changed. Also note that v_1 are v_2 have higher heat than v_6. This is because v_4 has higher heat than v_6 and v_4's heat is diffused more to v_1 and v_2 than v_6. Also v_1 is receiving heat from v_6. Therefore v_1 and v_2 have higher heat than v_6.



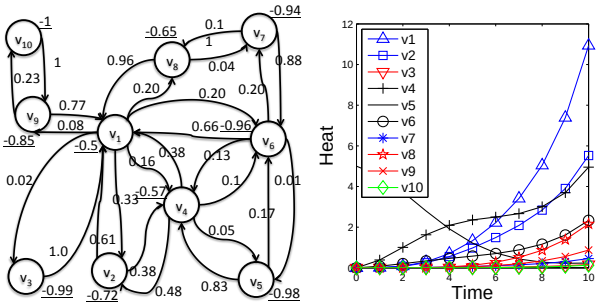
(a) Activity-DiffusionConstructed Graph (b) Heat Diffusion over time

Figure 3. Heat Diffusion (Interactive Activities Only)

Figure 4 shows the result of activity based heat diffusion process with information on both the number of interactive activities and the number of non-interactive activities. We set β to 0.5 in order to fully consider non-interactive activities. Each node receives the same heat as in Figure 3(a) but they lose heat differently. The rate of heat loss in Figure 4(b) is slower than the one in Figure 3(b). Note that some users such as v_1 and v_2 are actively posting once they are influenced by v_4.

4 INFLUENCERANK, INFLUENCECOVERAGE Algorithms

In this section we design the INFLUENCECOVERAGE, which can measure the coverage of nodes that are influenced by a given user using the heat diffusion process. We can define the INFLUENCECOVERAGE of a node v_i by the coverage of the nodes that are influenced by v_i. Given INFLUENCECOVERAGE, we assign INFLUENCERANK by descending order of INFLUENCECOVERAGE. This can be directly applied to viral marketing scenarios where a limited budget is allocated to some new products. Therefore marketing candidates should be carefully selected so that the marketing efficiency can be maximized.



(a) Constructed Graph (b) Heat

Figure 4. Heat Diffusion (Non-interactive and Interactive Activities)

Determining Influence Threshold: Given heat at v_i, v_i diffuses its heat to its neighbors. At some time t, we check the heat value at every vertex. Each vertex has an acceptance threshold θ. If the heat is greater than or equal to θ, we consider the user is influenced by the initial heat source, v_i.

Top K Influential Nodes by INFLUENCERANK: Given a social network G of size $|V|$, the top K influence rank based node selection algorithm performs the following four tasks in sequence: (a) activity based social influence computation using the activity based heat diffusion kernel, (b) INFLUENCECOVERAGE computation, (c) sorting social network nodes by INFLUENCECOVERAGE, and assigning INFLUENCERANK based on INFLUENCECOVERAGE, and (d) returning top-k nodes by their INFLUENCERANK.

It is well known that selecting top-k influential people from the network of $|V|$ nodes ($k < |V|$) to maximize the spread of their influence is NP-hard [11]. Thus we can present greedy algorithms for selecting top-k people in a social network graph.

There are several ways to compute the top k most influential nodes based on INFLUENCERANK. For example, if we simply rank all nodes in V based on their INFLUENCECOVERAGE, then the top k nodes are selected as they have the top k largest influence coverage. However, this approach may not give the highest overall node coverage by the top k nodes when there is a large overlap in terms of node coverage among the top k nodes. In this section we describe three influence ranking criteria: (a) independent INFLUENCECOVERAGE, (b) local optimal INFLUENCECOVERAGE and (c) global optimal INFLUENCECOVERAGE.

Top K by Independent INFLUENCECOVERAGE: The simplest approach to selecting top-k influential nodes in a social network is to compute their INFLUENCECOVERAGE and sort the nodes by their individual INFLUENCECOVERAGE. Algorithm 1 shows the pseudocode of the indepen-

dent INFLUENCECOVERAGE algorithm. For each v_i in G, we assign the same heat h_0 and compute INFLUENCECOVERAGE in terms of the number of nodes over which v_i has influenced.

ALGORITHM 1: Independent INFLUENCECOVERAGE

```
foreach v_i ∈ G do
    H(0) ← 0;                          /* initialize heat */
    H_i(0) ← h_0;                      /* assign heat h_0 to v_i */
    IC_i ← ∅;                          /* initialize IC_i */
    HeatDiffusion(t, H(0));
    foreach v_j ∈ G do
        if H_j(t) ≥ θ_j then
            Add v_j into set IC_i;
        end
    end
end
Sort {IC_i | v_i ∈ G} by set size;
return top-k v_i ordered by |IC_i|
```

ALGORITHM 2: Minimizing Local Overlap

```
Line 1 to 12 in Algorithm 1;
IC ← ∅;                                /* universal result set */
m ← 0;
V_local ← 0;
while m < k do
    foreach v_i ∈ (G − V_local) do
        Find IC_i that has max(IC_i − IC);
    end
    V_local ← V_local ∪ v_i;
    m ← m + 1;
    IC ← IC ∪ IC_i;
end
return marked IC_i
```

Algorithm 1 is easy to implement but it may select top k nodes that have very high level of overlapping in terms of node coverage. Given $k = 2$, we denote a set of influenced users by v_1 as IC_1. For example, given $IC_1 = \{v_{11}, v_{12}, v_{13}\}$, $IC_2 = \{v_{11}, v_{12}, v_{13}, v_{14}\}$, and $IC_3 = \{v_{20}, v_{21}\}$, Algorithm 1 returns v_2 and v_1 as top 2 influential users because $|IC_2| > |IC_1| > |IC_3|$. v_2 and v_1 influence four users, which is $IC_1 \cup IC_2 = \{v_{11}, v_{12}, v_{13}, v_{14}\}$. On the other hand, v_2 and v_3 influence 6 users, which is $IC_2 \cup IC_3 = \{v_{11}, v_{12}, v_{13}, v_{14}, v_{20}, v_{21}\}$. Simply returning first k largest IC_i does not guarantee the true top-k influential people. Therefore we present the second criteria in selecting top-k people, which is minimizing the local overlap.

Top K by Locally Optimal INFLUENCECOVERAGE**:** Instead of selecting top k nodes with the largest individual INFLUENCECOVERAGE, the locally optimal INFLUENCECOVERAGE algorithm adds node v_i to its top k list if it satisfies two conditions: (i) v_i has the high INFLUENCECOVERAGE and (ii) v_i has the minimal intersection with previously selected set of nodes as described in Algorithm 2. This algorithm extends the independent influence ranking Algorithm 1 by adding another round of computation to find the set of nodes in the social network, which minimizes local overlap, upon completion of the heat diffusion process. Each node found is added to the top-k list of influential nodes until every node in the social network is examined.

Top k By Globally Optimal INFLUENCECOVERAGE**:** Algorithm 1 and 2 computes INFLUENCECOVERAGE of v_i while assuming that v_i is the only heat source. But in reality, there are multiple heat sources at the same time. For example, a new iPhone is released, then several reviews are posted on the first day of release. Then people read multitple reviews and decide to buy based on these multiple reviews. Therefore, we need an algorithm that uses multiple heat sources to conduct the social influence computation and influence ranking.

Algorithm 3 gives a sketch of the top k selection by globally optimal influence ranking. It uses the Hill Climbing approach. In order to avoid local maximum, it performs Algorithm 1 and gets the sorted vertices by individual INFLUENCECOVERAGE. We then select v_i whose INFLUENCECOVERAGE, denoted by IC_i is the largest and initialize V_{global} by adding v_i into V_{global}. At each step in the outer loop, we give heat to vertices in V_{global}. By adding one more heat source at a time, we simulate the scenario that there are multiple heat sources. After the second inner loop, we find v_i, which has the minimal overlap with previous selected nodes in V_{global}.

ALGORITHM 3: Minimizing Global Overlap

```
Line 1 to 12 in Algorithm 1;
IC ← ∅; m ← 0;
Find v_i that has max(IC_i);
V_global ← v_i;
while m < k do
    H(0) ← 0;
    foreach v_i ∈ V_global do
        H_i(0) ← h_0;
    end
    foreach v_i ∈ (G − V_global) do
        H_i(0) ← h_0;
        HeatDiffusion(t, H(0));
    end
    Find v_i that has max(IC_i − IC);
    V_multi ← V_global ∪ v_i;
    m ← m + 1;
    return V_global
end
```

5 Experiments

We performed experimental evaluations in order to show the performance and effectiveness of the activity-based social influence against topology-based heat diffusion. First of all, we explain how we collect datasets, what properties these datasets have, what the effects of parameters, and the performance of our approaches. Our results show that activity-based approach has larger coverage so that with limited marketing budges marketing companies maximize the advertising.

5.1 Datasets

We used datasets from three sources, DBLP[5], Epinions[6], and and Facebook[7], to evaluate the effectiveness of our activity based social influence model and influence ranking algorithms. DBLP dataset provides bibliographic information on major computer science journals and proceedings. We parse DBLP data and extracted 5,000 authors and their co-authorship information. For example, if authors u and v wrote x number of papers together, then we create two edges $e_1(u, v)$ and $e_2(v, u)$ and set x as the number of interactive activities for both e_1 and e_2.

Epinions dataset, collected by Massa[15], contains consumer reviews and trust networks. Epinions is a platform for people to share their experiences and to maintain a trust network. Customers will be influenced by reviews when they consider buying products. These reviews are displayed after filtered using users' trust network. For example, if users u and v made some reviews and user w likes u's reviews and does not v's reviews, then w creates a trust list by adding u and does a block list by adding v. Now Epinions displays u's reviews first and hides v's reviews for w. Epinions' dataset has 49,289 users, 664,824 reviews, and 487,181 trust statements after 5-week crawl in 2003. Given a user u and her x number of reviews, we construct a vertex u and set x as the number of non-interactive activities for the vertex u. If a user v adds a user u into a trust network, then we create an edge $e(u, v)$.

Facebook is recognized by many as one of the largest social network services. Each user has her profile page where the owner can post photos, videos, and other owner specific information. For each posting, her friends can leave comments. When a user u posts photos, videos, or statuses, we consider it as non-interactive activities. When a user u leaves a comment on v's posting, then we consider it as an interactive activity. Given a user u and her x number of postings, we create a vertex u and set x as the number of non-interactive activities. Given a friendship between two users u and v, we create two edges $e_1(u, v)$ and $e_2(v, u)$. If node u has y number of comments on node v's posting, then we create an edge $e(u, v)$ and set y, the number of interactive activities, as the weight for this edge. We launched a Facebook app to analyze users' posting trends and statistics of friends. From 273 Facebook app users, we extract their friend relationship information which creates 76,954 nodes and 1,121,861 friendship edges. For example, one user may have 300 friends. Then we can create 301 users nodes. Repeating this procedure for all 273 users results in 76,954 nodes.

5.2 Effect of Various Parameters

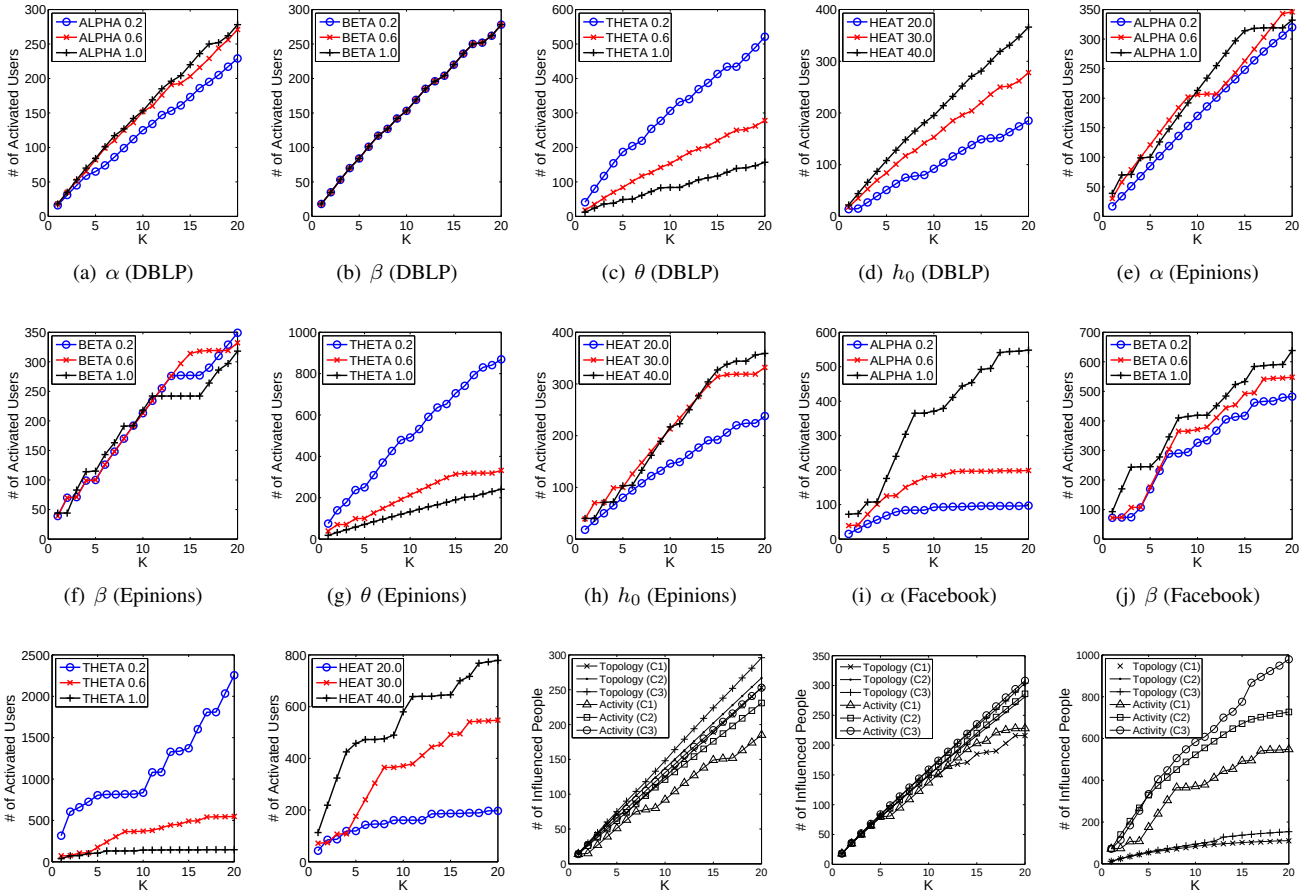
In our activity-based model, we use parameters such as a heat conductivity, α, an acceptance threshold, θ, an activity weight, β, and initial heat, h_0. Different settings of these parameters may significantly affect the heat diffusion process and consequently the influence computation result. Before assigning values for experiments, Figures 5(a) to (l) present how these values affect heat diffusion process. We initialize α as 1.0, β as 0.6, θ as 0.6, and initial heat h_0 as 30 and measure the number of influenced nodes by varying the value of each parameter.

The heat conductivity, α, controls the speed of heat spread. Some social networks spread news or rumors quickly while others do not. By setting the value of α, we can mimic the speed of spread. The value of α is a real number between 0 and 1. If it is set to 0, heat will not be diffused at all. On the other hand, if α is set to 1, all of the heat at the heat source will be diffused to its neighbors without any loss. Figures 5(a), 5(e), and 5(i) show how many users are influenced by top-k people while varying α from 0.2 to 1.0 for each dataset. As stated above, higher α value results in more influenced users because more heat will be diffused.

The activity weight, β, regulates how much weight will be given for non-interactive activities. If it is set to 0, the number of non-interactive activities does not play a role at all on the heat diffusion process. Then each user loses their heat as proportional to time. If β is high, then a user loses heat with regard to her activities in posting. Figure 5(j) shows that higher β value results in larger number of influenced users because heat sources will lose their heat much slower, which means the amount of heat to diffuse decreases slower. Figure 5(b) shows that the total number of influenced users does not change significantly while varying β. In the DBLP dataset, each node has no non-interactive activities. It has only interactive activities, which is co-authorship relation. Therefore, β does not affect the heat diffusion process. Each node loses its heat as the time goes by.

The acceptance threshold, θ, is used to determine if a user is influenced or not. If $H_i(t)$ is greater than or equal to θ_i, then we consider v_i is influenced by heat source users. θ_i value varies from 0 to 1. If it is low, more users will be influenced. Figures 5(c), 5(g), and 5(k) show that when we decrease θ, the number of influenced users increases. Especially, for Facebook dataset, the number of influence users are increased 6 times more when we decrease θ from 0.6 to 0.2.

The next parameter we measure is the initial amount of



(a) α (DBLP) (b) β (DBLP) (c) θ (DBLP) (d) h_0 (DBLP) (e) α (Epinions)

(f) β (Epinions) (g) θ (Epinions) (h) h_0 (Epinions) (i) α (Facebook) (j) β (Facebook)

(k) θ (Facebook) (l) h_0 (Facebook) (m) DBLP (n) Epinions (o) Facebook

Figure 5. Effect of Parameters: (a) to (l) and Accumulated Number of Influenced Users: (m) to (o)

heat, h_0. We varied h_0 from 20 to 40. The more heat that a user is assigned to, the more users are influenced by the top k nodes in Figures 5(d), 5(h), and 5(l).

In the following experiments, we initialize α as 1.0, β as 0.6, θ as 0.6, and h_0 as 30. Twenty most influential people are selected based on three criteria: independent INFLUENCECOVERAGE(C1), locally optimal INFLUENCECOVERAGE(C2), globally optimal INFLUENCECOVERAGE(C3). We compared activity-based heat diffusion with topology-based heat diffusion using three influence ranking criteria: independent INFLUENCECOVERAGE(C1), locally optimal INFLUENCECOVERAGE(C2), and globally INFLUENCECOVERAGE(C3). For each experiment, we show that global influence ranking is the best top k influence ranking algorithm for selecting the top k nodes that have the overall maximal influence in terms of non-overlapping node coverage.

5.3 DBLP Dataset

Experiments on DBLP dataset has display some interesting results. The number of influenced people nodes based on the influence ranks computed by topology-based heat diffusion is larger than thatthe one by activity-based heat diffusion.

DBLP dataset has only interactive activities information. Therefore heat loss rate, β, does not affect the result of heat diffusion. The only difference between topology-based and activity-based approach is the heat distribution. In the topology-based approach, user u_i diffuses its heat evenly to its neighbors. For example, if u_i has 10 co-authors and 90% of papers were written with u_j and 10% were done with 9 authors evenly, then $\frac{1}{10}H_i(t)$ is distributed to each of u_i's neighbors evenly. On the other hand, in the activity-based approach, u_i diffuses its heat to its neighbors based on the rate of interaction. For example, due to the 90% of interaction with u_j, u_i diffuses $\frac{9}{10}H_i(t)$ to u_j and $\frac{1}{10\times 9}H_i(t)$ is distributed to other 9 neighbors evenly. In this case, u_j Therefore, the topology-based heat diffusion has larger

number of influenced people as shown in Figure 5(m).

Experiments on DBLP dataset display some interesting results. The number of influenced nodes based on the influence ranks computed by topology-based heat diffusion is larger than that by activity-based heat diffusion.

5.4 Epinions Dataset

Experiments on Epinions dataset show that the activity-based approach has larger influence coverage than the topology-based approach. Epinions dataset has non-interactive activities only. Therefore, users in Epinions dataset diffuse heat evenly to its neighbors like in the topological approach. However, the number of influenced people computed by the activity based heat diffusion approach is different from that by the topology-based approach due to the number of non-interactive activities. Users lose its heat at a slower rate based on the number of non-interactive activities, which is regulated by heat loss factor β. Users with more non-interactive activities lose its heat much slower than the ones with smaller number of non-interactive activities. Thus the number of influenced people by using the activity-based approach is larger than that by the topology-based diffusion as shown in Figure 5(n).

5.5 Facebook Dataset

Experiments on Facebook dataset show some of the most interesting results. Facebook dataset has both interactive and non-interactive activities. Figure 5(o) shows the experimental results for Facebook datasets.

It is interesting to note that for each influence rank based top k selection criteria, the activity-based heat diffusion process influence more users for two reasons. First, in the topology-based approach, users lose heat based on the time period. On the other hand, in the activity-based approach, users loose heat based on the number of non-interactive activities. Thus, some users lose heat quickly while others lose slowly, which means these users have high amount of heat sources to diffuse. Second, users in Facebook dataset have much more non-interactive activities than the ones in Epinions dataset. Therefore, the gap in Facebook dataset between the topological and activity-based approach is much higher than that in Epinions dataset.

6 Conclusion

We have presented the activity-base social influence model based on activity enhanced heat diffusion kernel and a suite of activity influence rank based top k algorithms. Our activity-based heat diffusion model has made three unique contributions. First, we introduced a novel mechanism to extend the heat diffusion model by effectively incorporating both interactive and non-interactive activities. Second, we develop a suite of top k influence rank based node selection algorithms by minimizing the overlapping in the node coverage of top k most influential nodes, including independent influence rank, (b) locally optimal influence rank and (c) globally optimal influence rank using Hill Climbing algorithm. Finally we conduct an extensive series of experiments on three representative real-world social network datasets to show the effectiveness of our activity-based social influence model and influence rank algorithms. Compared to the existing topology-based influence diffusion model, the activity-based social influence model considers not only topology of a social network but also activity sensitive attributes such as interactive activities and non-interactive activities.

References

[1] I. Anger and C. Kittl. Measuring Influence on Twitter. In *i-KNOW*, 2011.

[2] H. Bao and E. Y. Chang. AdHeat: an influence-based diffusion model for propagating hints to match ads. In *WWW*, 2010.

[3] E. Berger. Dynamic Monopolies of Constant Size. *Journal of Combinatorial Theory Series*, 2001.

[4] D. Boyd and J. Heer. Profiles as Conversation: Networked Identity Performance on Friendster. In *HICSS*, 2006.

[5] DBLP. Computer science bibliography. In *http://dblp.uni-trier.de/db/*, 2006.

[6] Epinions. General consumer review site. In *http://www.epinions*, 1999.

[7] Facebook. Online social networking service. In *http://www.facebook.com*, 2004.

[8] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 2001.

[9] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development. *Academy of Marketing Science Review*, 2001.

[10] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 1978.

[11] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the Spread of Influence through a Social Network. In *SIGKDD*, 2003.

[12] R. I. Kondor and J. D. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *ICML*, 2002.

[13] H. Ma, H. Yang, M. R. Lyu, and I. King. Mining social networks using heat diffusion processes for marketing candidates selection. In *CIKM*, 2008.

[14] M. W. Macy. Chains of Cooperation: Threshold Effects in Collective Action. *American Sociological Review*, 1991.

[15] P. Massa and P. Avesani. Trust Metrics in Recommender Systems. In *Computing with Social Trust*, 2009.

[16] D. Peleg. Local majority voting, small coalitions, and controlling monopolies in graphs: A review. In *3rd Colloq. on Structural Information and Communication*, 2002.

[17] M. M. Skeels and J. Grudin. When social networks cross boundaries: a case study of workplace use of facebook and linkedin. In *ACM GROUP*, 2009.

[18] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: finding topic-sensitive influential twitterers. In *WSDM*, 2010.

[19] H. Yang, I. King, and M. R. Lyu. DiffusionRank: a possible penicillin for web spamming. In *SIGIR*, 2007.