Mining Topic-level Influence in Heterogeneous Networks

Lu Liu[†⋆♯], Jie Tang[⋆], Jiawei Han[◦], Meng Jiang[†⋆], and Shiqiang Yang[†⋆]
[†]Tsinghua National Laboratory for Information Science and Technology
[⋆]Department of Computer Science and Technology, Tsinghua University, China
[♯]Capital Medical University, China
[◦]University of Illinois at Urbana-Champaign, USA
{lu-liu,jm06}@mails.tsinghua.edu.cn, {jietang,yangshq}@tsinghua.edu.cn, hanj@cs.uiuc.edu

ABSTRACT

Influence is a complex and subtle force that governs the dynamics of social networks as well as the behaviors of involved users. Understanding influence can benefit various applications such as viral marketing, recommendation, and information retrieval. However, most existing works on social influence analysis have focused on verifying the existence of social influence. Few works systematically investigate how to mine the strength of direct and indirect influence between nodes in heterogeneous networks.

To address the problem, we propose a generative graphical model which utilizes the heterogeneous link information and the textual content associated with each node in the network to mine topic-level direct influence. Based on the learned direct influence, a topic-level influence propagation and aggregation algorithm is proposed to derive the indirect influence between nodes. We further study how the discovered topic-level influence can help the prediction of user behaviors. We validate the approach on three different genres of data sets: Twitter, Digg, and citation networks. Qualitatively, our approach can discover interesting influence patterns in heterogeneous networks. Quantitatively, the learned topic-level influence can greatly improve the accuracy of user behavior prediction.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]; H.4.m [**Information Systems**]: Miscellaneous; G.3 [**Probability and Statistics**]: Models

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Social influence, Influence propagation, Behavior prediction

1. INTRODUCTION

It is well recognized that influence is a complex and subtle force that governs the dynamics of social networks. With the power of influence, a company can market a new product by first convincing a small number of influential users to adopt the product and then triggering cascade further adoptions through the effect of "word of mouth" in the social network (also referred to as influence maximization [6, 22, 14, 18]). In academic networks, with the influence between research collaborators, novel ideas or innovations may quickly spread and lead the blooming of new academic directions. In microblogging networks (e.g., Twitter), an influential user can induce her/his friends to post/re-tweet a blog on a specific topic, e.g., "Obama".

Recently, social influence analysis has attracted considerable research interests. However, most existing works have focused on validating the existence of influence [1, 4], or studying the maximization of influence spread in the whole network [14, 3], or modeling only direct influence in homogeneous networks [5, 28, 31]. The micro-level mechanisms of social influence in heterogeneous networks, e.g., the influence strength of a user on his/her friends at a specific topic, have been largely ignored. Moreover, besides the direct influence, another interesting question is "Does the influence exist between users who are not connected?" In another word, "Does a user have a certain indirect influence on his/her friends' friends in the social network?" Christakis and Fowler [8, 30] have studied a special case of this problem, i.e., influence of happiness, and showed that within a social network, happiness spreads among people up to three degrees of separation, which means when you feel happy, your friend's friend's friend has a higher likelihood to feel happy too. However, they only qualitatively test this finding on two small data sets. Therefore, to understand the underlying social dynamics, a systematic study on the problem of mining direct and indirect influence in heterogeneous networks is clearly needed.

Motivating Example To clearly motivate this work, we conduct an influence analysis on three different types of social networks: Twitter[1], Digg[2], and Cora[3]. On Twitter, the user action is defined as whether a user posts/re-tweets a blog on a specific topic. On Digg, the action is defined as whether a user submits/votes a story on a topic. On Cora, we define the action as whether a user publishes a paper on a topic. For example, on Twitter, if a user posts a tweet on "Obama" and his friend (or an n-degree friend) re-tweets it or also posts a tweet on this topic, we say that the friend is influenced by the user. In the analysis, the influence strength is estimated by the averagely increased probability $(p_1 - p_2)$ for all users, where p_1 is the probability of a user's n-degree friends performing the action when the user has already performed the action and p_2 is the average probability of users to perform the action over the whole

[1]http://www.twitter.com, a microblogging system.
[2]http://www.digg.com, a social news sharing and voting website.
[3]http://www.cs.umass.edu/mccallum/code-data.html, a bibliographic citation network

CIKM'10, October 25–29, 2010, Toronto, Ontario, Canada.

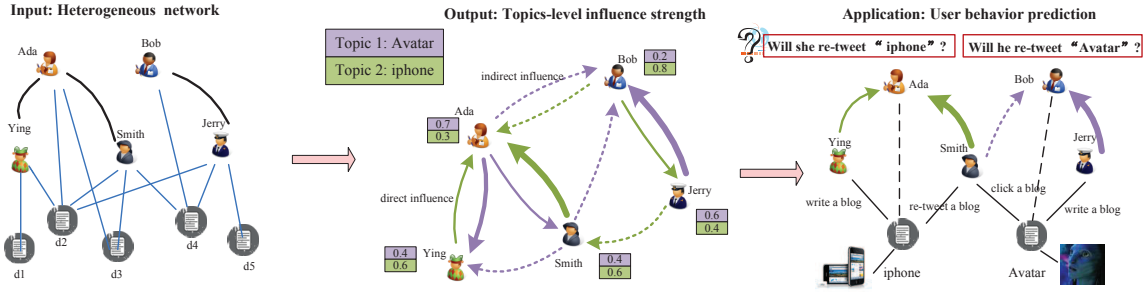

Figure 1: Problem illustration of mining topic-level influence in heterogeneous networks and predicting user behaviors.

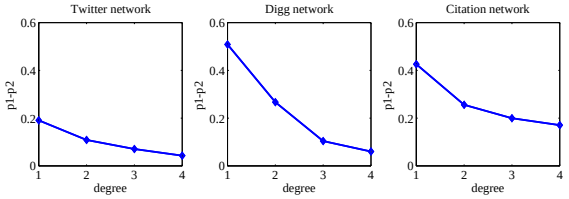

Figure 2: $n(1 \le n \le 4)$-degree influence in the three networks: Twitter, Digg, and Cora. The x-axis stands for the degree of friends and the y-axis stands for the probability $(p_1 - p_2)$, where p_1 is the probability of one user's n-degree friends performing the action when he has performed the action and p_2 is the average probability of users performing the action over the whole network.

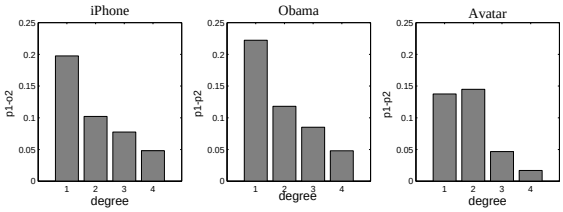

Figure 3: Topic-level influence on Twitter for three topics: "iPhone", "Obama", and "Avatar". x-axis stands for $n = 1, \cdots, 4$ degrees, and y-axis stands for the probability $(p_1 - p_2)$, with p_1 and p_2 having the same meanings as those in Fig. 2.

network, which is used to consider the factor of topic popularity. The analysis includes two aspects: (1) n $(1 \le n \le 4)$ degree influence; and (2) topic-level influence.

Fig. 2 shows the n-degree influence patterns for the three networks. We see that not only the actions of 1-degree friends would be influenced, but also 2-degree (even 4-degree) friends could be also influenced. For example, on Digg, when a user votes a story on a topic, his friends' friends (2-degree friends) averagely have a 20+% higher probability to vote (or submit) a story of this topic. However, the influence strength decreases with the increase of degree on average. For example, the 2-degree influence strength is almost half of 1-degree influence strength on Digg. It can be also seen that the three networks have very different influence patterns. In Fig. 3, we further analyze the topic-level influence on Twitter. We study the n-degree influence on three topics: "Obama", "iphone" and "Avatar". An interesting phenomenon is that on some topics the 2-degree influence is even stronger than the 1-degree influence (e.g., on "Avatar"). This is because "Avatar" is a very popular topic, on which the social users may be mainly influenced by the global trend (or more accurately, local community trend) in the social network, instead of one or two friends.

Problems and Contributions Thus our objective is to effectively and efficiently discover the underlying influence patterns in heterogeneous networks. The problem can be clearly explained by Fig. 1. The input is a heterogeneous network consisting of documents, users, and links between them. Topic-level influence mining can be decomposed into two subtasks: topic distribution modeling and direct (and indirect) influence strength estimation. The former is to associate a topic distribution with each node in the social network and the latter is to estimate the influence strength (including indirect) between users. The middle figure illustrates the output of topic-level influence mining. The solid arrow indicates the direct influence and the dashed arrow indicates the indirect influence. Our last task is to validate how the discovered topic-level influence can really help. A straightforward application is to utilize the discovered influence to help predict user behaviors, that is, to predict who will perform an action in the future.

To summarize, this work contributes on the follow aspects:

- We formally formulate the problem of topic-level influence mining and propose a generative model which utilizes both content and link information in heterogeneous networks.
- We propose a topic-level influence propagation process to mine indirect influence and to model the influence flow over networks.
- We apply the discovered topic-level influence to user behavior prediction and validate how it can help other social applications.
- We conduct experiments on three different types of data sets: Twitter, Digg, and Cora. Experimental results show that the learned influence model can greatly improve the accuracy of user behavior prediction. We also perform qualitative analysis to show interesting topic-level influence patterns discovered by the proposed approach.

The rest of the paper is organized as follows: Section 2 formally formulates the problem; Section 3 and Section 4 explain the proposed approach. Section 5 introduces the application of user behavior prediction based on the discovered influence. Section 6 presents experimental results that validate the effectiveness of our methodology. Finally, Section 7 discusses related work and Section 8 concludes.

2. PROBLEM DEFINITION

In this section, we introduce several related concepts and then

formally formulate the problem of mining topic-level influence in heterogeneous networks.

Definition 1. ***[Heterogeneous Network]*** *Define a network $G = (V, E; \Omega)$. V is a set of nodes, which are classified into T types $\Omega = \{X_t\}_{t=1}^{T}$, where X_t is a set of nodes with the $t-$th type. The edge set $E \subseteq V \times V$ denotes the connections between nodes. For $\forall e_{uv} = (u, v) \in E$, if there exists an edge between u and v, $e_{uv} = 1$; otherwise $e_{uv} = 0$. The edges can be directed or undirected.*

Most online social networks are heterogeneous (e.g., Twitter, Digg and citation networks), consisting of more than one type of nodes, e.g., user nodes and document nodes (stories, tweets, papers, and other objects). Thus links in heterogeneous networks are comprised of friendships between users, authoring relationships between users and documents and links between documents. The links can be directed or undirected. For example, on Twitter and citation networks, the links between nodes are directed, from cited papers to citing papers or from normal users to their followers. On Digg social network, the links between users are undirected. Furthermore, we assume that the influence of a user on other users can be propagated along social links, thus we have the following definition.

Definition 2. ***[Direct and Indirect Influence]*** *Given two user nodes u, v in a heterogeneous network, we denote $\Phi_v(u) \in \mathbb{R}$ as the influential strength of user u on user v. Furthermore, if $e_{uv} = 1$, we call $\Phi_v(u)$ the direct influence of user u on v; if $e_{uv} = 0$, we call $\Phi_v(u)$ the indirect influence of user u on v.*

Direct influence indicates the influence between two nodes which are connected while indirect influence indicates the influence of two nodes which are not connected. Please note that the influence is asymmetric, i.e., $\Phi_v(u) \neq \Phi_u(v)$. Based on the influence between node pairs, we can further define the concept of global influence.

Definition 3. ***[Global Influence]*** *Given a heterogeneous network, $\Lambda(v)$ is defined as the global influence of v, which represents the global influential strength of v over the whole network.*

The global influence strength has a close relationship with the direct/indirect influence. For example, if a user has a strong influence on other users, it is probably that he is very influential globally. In this work, we only consider influence between nodes with the same type, e.g., the influence between users. The influence among different types of nodes, e.g., the influence from authors to documents or from documents to authors, is not included in our goals due to the difficulty for meaning explanation and quantitative measure.

Our formulation of topic-level influence mining is quite different from existing works on social influence analysis. For social influence analysis, works [1] and [25] studied how to qualitatively measure the existence of influence. Crandall etc. [4] studied the correlation between social similarity and influence. However, they focus on qualitative identification of the existence of influence, but do not provide a quantitative measure of the influential strength. Works [9, 28, 31] investigated how to learn the influence probabilities from the history of user actions. However, these methods either do not consider the influence at the topic-level or ignore the indirect influence. Another challenge which has not been studied extensively before is how to learn the topic distributions and the topic-level influence jointly.

2.1 Intuitions and Our Approach

In most real networks, users may be interested in different topics, e.g., an author in citation networks may be interested in topics

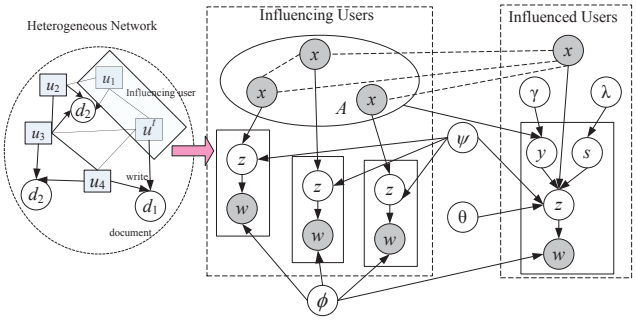

Figure 4: Probabilistic generative model

"database" and "data mining". The influence strength between two users may also vary with different topics, which has been qualitatively verified in sociology [10, 17]. Furthermore, the influence can be direct or indirect, which is referred to as n-degree of influence [8, 30]. Thus, to summarize, we have the following intuitions for topic-level influence mining:

1. Each node v is associated with a vector $\psi_v \in \mathbb{R}^T$ of T-dimensional topic distribution ($\sum_z \psi_v(z) = 1$). Each element $\psi_v(z)$ is the interest probability of the node (user) on topic z.
2. The influence $\Phi_v(u)$ of user u on v can be direct ($e_{uv} = 1$) or indirect ($e_{uv} = 0$).
3. The behavior of a user is either influenced by his/her n-degree friends who have the same behavior or generated depending on his/her interests.

The last intuition can be better explained by an example on Digg. A user may dig a story because he is following the action of his friends who have digged this story (or a story on the same topic) or simply because he is interested in this topic.

Therefore, from the technique aspect, our objective is to design a method to learn the user interests (the associated topic distribution) and to estimate the (in)direct influence between users simultaneously. In this paper, we propose a topic-level influence modeling framework. First, by combining both textual information and link information in heterogeneous networks, we present a probabilistic generative model to learn user interests which are represented as topic distributions and direct influence between users at the topic-level simultaneously. Second, based on the topic-level direct influence, we propose an influence propagation process to derive indirect influence between users.

3. MINING INFLUENCE ON HETEROGENEOUS NETWORKS

Influence is interacted with many potential factors, e.g., similarity, correlation and etc. [1, 4]. Commonsense knowledge is needed to quantitatively model the influence strength. Here we have two general assumptions.

Assumption 1. *Users with similar interests have a stronger influence on each other.*

This assumption actually corresponds to the influence and selection theory [1]. In real networks, the similarity can be calculated based on the textual content associated with each user. Thus, influence can be represented as to which extent the textual content

Table 1: Variable descriptions

x, x'	the influenced/influencing user
w, w'	words in the associated document
z, z'	topic assignment to each word
d, d'	document associated with influenced/influencing user
A	the user list who may influence x associated with d
y	the influencing user from A
s	the label denoting either influencing or not
V	the number of words in the data set
T	the number of topics to be extracted
X	the number of users in the data set
θ	the topic mixture of influencing users
ψ	innovative topic mixture of users
ϕ	word distribution for each topic
γ	the influence mixture of users
λ	the parameter to draw the label s
α	the Dirichlet prior for hidden variables

is "copied" from the influencing nodes. For example, in the citation network, if the content of document d_1 is very similar to that of document d_2, we may deem that d_1 "copies" a lot of ideas from d_2, thus d_1 is influenced by d_2 a lot. Another assumption is about correlation.

ASSUMPTION 2. *Users whose actions frequently correlate have a stronger influence on each other.*

In heterogeneous networks, the link weight is usually used to indicate the correlation strength between nodes, which can be calculated by the co-occurrence frequency of nodes. For example, if author a and author b jointly write a number of papers, then the two authors have a strong influence on each other. Another example on Twitter, if user a replies or retweets a number of microblogs posted by user b, then a is highly correlated with b and it is very likely that b has a strong influence on a.

Based on these considerations, we propose a probabilistic generative model to jointly learn user interests and the direct influence between users.

3.1 Probabilistic Generative Model

Fig. 4 shows the graphical structure of the model, which contains two types of links: links between users and links between users and documents. The details of the generative process are illustrated in Alg. 1. And Table 1 lists the descriptions of variables. The proposed model consists of two parts. First, we model the interests of each user in the corpus (as the left part of Fig. 4), which corresponds to the first iteration of Alg. 1. Specifically, we assume that the topics of documents are generated from users. We represent each user as a multinomial distribution over topics ψ, thus each word in documents is generated from one topic selected from the distribution.

Then, we assume that the behavior of each influenced user can be generated in two ways, either depending on his/her interests or influenced by one of his/her friends. E.g., when an author writes a paper, he/she may create the idea innovationally w.r.t. his/her research interests or "copy" it from one of the cited authors. In the model, we use a parameter s to control the influence situation. s is generated from a Bernoulli distribution whose parameter is λ. When $s = 1$, the behavior is generated based on his/her own interests. When $s = 0$, it means the behavior of the user is influenced by one of his/her friends. Thus we need another parameter γ to select one influencing user y from the candidate user list A. The last step of the generation process is to select a topic from the topic distribution of one user, the user himself/herself or one of his/her friends, based on which the word is generated.

In the above generative process, A is determined by real applications, which considers both directed and undirected links between users. For example, on Twitter network A denotes the users whom a blog is re-tweeted from while on citation networks it denotes the authors of cited papers. On these networks, the links between users are directed. In some other networks, for example Digg, A denotes the friends of user x who also dig the same story, and the links are undirected. Thus the proposed model is able to handle both types of cases.

3.2 Model Learning via Gibbs Sampling

The model can be estimated by Gibbs sampling. Gibbs sampling is an algorithm to approximate the joint distribution of multiple variables by drawing a sequence of samples, which iteratively updates each latent variable under the condition of fixing remaining variables. We list the update equations for each variable as below and the details of derivation can refer to the appendix. In all the update equations, $N(*)$ is the function which stores the number of samples during Gibbs sampling. For example, $N_{d,y,s}(d, y, 0)$ represents the number of topics/words in d which are supposed to be generated from user y.

$$p(s_i = 0|\vec{s}_{-i}, x_i, z_i, .) \propto \frac{N_{x',z'}(y_i,z_i)+N_{y,z,s}(y_i,z_i,0)+\alpha_\theta}{N_{x'}(y_i)+N_{y,s}(y_i,0)+T\cdot\alpha_\theta} \cdot \frac{N_{x,s}(x_i,0)+\alpha_{\lambda_{s_0}}}{N_x(x_i)+\alpha_{\lambda_{s_0}}+\alpha_{\lambda_{s_1}}} \tag{1}$$

$$p(s_i = 1|\vec{s}_{-i}, x_i, z_i, .) \propto \frac{N_{x,z,s}(x_i,z_i,1)+\alpha_\psi}{N_{x,s}(x_i,1)+T\cdot\alpha_\psi} \cdot \frac{N_{x,s}(x_i,1)+\alpha_{\lambda_{s_1}}}{N_x(x_i)+\alpha_{\lambda_{s_0}}+\alpha_{\lambda_{s_1}}} \tag{2}$$

$$p(y_i|\vec{y}_{-i}, s_i = 0, d_i, x_i, z_i, A, .) \propto \frac{N_{x,z,y,s}(x_i,z_i,y_i,0)+\alpha_\gamma}{N_{x,z,s}(x_i,z_i,0)+|A|\cdot\alpha_\gamma} \cdot \frac{N_{x',z'}(y_i,z_i)+N_{y,z,s}(y_i,z_i,0)+\alpha_\theta}{N_{x'}(y_i)+N_{y,s}(y_i,0)+T\cdot\alpha_\theta} \tag{3}$$

$$p(z_i|\vec{z}_{-i}, s_i = 0, w_i, .) \propto \frac{N_{x',z'}(y_i,z_i)+N_{y,z,s}(y_i,z_i,0)+\alpha_\theta}{N_{x'}(y_i)+N_{y,s}(y_i,0)+T\cdot\alpha_\theta} \cdot \frac{N_{w,z}(w_i,z_i)+N_{w',z'}(w'_i,z'_i)+\alpha_\phi}{N_z(z_i)+N_{z'}(z_i)+V\cdot\alpha_\phi} \tag{4}$$

$$p(z_i|\vec{z}_{-i}, s_i = 1, w_i, .) \propto \frac{N_{x,z,s}(x_i,z_i,1)+\alpha_\psi}{N_{x,s}(x_i,1)+T\cdot\alpha_\psi} \cdot \frac{N_{w,z}(w_i,z_i)+N_{w',z'}(w'_i,z'_i)+\alpha_\phi}{N_z(z_i)+N_{z'}(z_i)+V\cdot\alpha_\phi} \tag{5}$$

After the Gibbs sampling process, we will obtain the sampled coin s_i, influencing user y_i, and topic z_i for each word, and the influence strength can be then estimated by Eq.(6), which are averaged over the sampling chain after convergence. K denotes the length of the sampling chain.

$$\Phi_x(y|z) = \gamma_x(y|z) = \frac{1}{K}\sum_{i=1}^{K} \frac{N_{x,z,y,s}(x,z,y,0)^i + \alpha_\gamma}{N_{x,z,s}(x,z,0)^i + |A|\cdot\alpha_\gamma} \tag{6}$$

The equations reflect our assumptions in a statistical way. Take citation networks as an example. It indicates that if one author x cites more papers of author y on topic z, then y has a stronger influence on x w.r.t. topic z.

4. TOPIC-LEVEL INFLUENCE PROPAGATION & AGGREGATION

The above probabilistic model only discovers the direct influence, but does not consider indirect influence. In reality, there exist different types of indirect influence. Take Fig. 5(a) as an example. If $a1$ influences $a2$ and $a2$ influences $a3$, then $a1$ will influence $a3$ potentially, i.e., two-degree of influence. Fig. 5(b) demonstrates the influence enhancement: if $a1$ influences $a3$ and $a4$ while $a3$ and $a4$ also have an influence on $a2$, then the influence from $a1$ to $a2$

foreach *influencing user* x' **do**
  foreach *associated document* d' **do**
    foreach *word* $i \in d'$ **do**
      Draw a topic $z'_{d',i} \sim multi(\psi_x)$ from the topic mixture of user $x'_{d',i}$;
      Draw a word $w'_{d',i} \sim multi(\phi_{z_{d,i}})$ from $z'_{d',i}$-specific word distribution;
    end
  end
end
foreach *influenced user* x **do**
  foreach *associated documents* d **do**
    foreach *word* $i \in d$ **do**
      Toss a coin $s_{d,i} \sim bernoulli(\lambda_{x_{d,i}})$, where $\lambda_{x_{d,i}} = p(s = 0|x_{d,i}) \sim beta(\alpha_{\lambda_{s_0}}, \alpha_{\lambda_{s_1}})$ which indicates the proportion between the innovation and influenced probability of $x_{d,i}$;
      if $s_{d,i} = 0$ **then**
        Draw a influencing user $y_{d,i} \sim multi(\gamma_x)$ from the user list A;
        Draw a topic $z_{d,i} \sim multi(\theta_y)$ from the topic mixture of $y_{d,i}$;
      end
      if $s_{d,i} = 1$ **then**
        Draw a topic $z_{d,i} \sim multi(\psi_x)$ from the topic mixture of $x_{d,i}$;
      end
      Draw a word $w_{d,i} \sim multi(\phi_{z_{d,i}})$ from $z_{d,i}$-specific word distribution;
    end
  end
end

Algorithm 1: Probabilistic generative process

should be enhanced. In this section, we propose a topic-level influence propagation algorithm to derive the indirect influence based on the learned direct influence by the above probabilistic model.

4.1 Atomic Influence Propagation

We first introduce the notion of atomic propagation. The indirect influence from $a1$ to $a3$ in Fig. 5(a) can be modeled as a concatenate result of the direct influence from $a1$ to $a2$ and the influence from $a2$ to $a3$. The enhancement of the influence from $a1$ to $a2$ in Fig. 5(b) can be defined as an aggregate result of the direct influence among the neighborhood of $a1$ and $a2$. Therefore, the *atomic influence propagation* is defined as follows:

$$\Phi_v(u) = \diamond(\forall w \in Nb(v) : \Phi_v(w) \circ \Phi_w(u)) \tag{7}$$

where $Nb(v)$ is the set of neighbors of node v; $\circ$ is the concatenation function, e.g., multiplication and minimum value; $\diamond$ is the aggregation function, e.g., addition and maximum value. In particular, if we use multiplication as the concatenation function and addition as the aggregation function, then the atomic influence propagation can be instantiated as:

$$\Phi_v(u) = \sum_{w \in Nb(v)} \Phi_v(w) \cdot \Phi_w(u) \tag{8}$$

Matrix operation can be used to represent this atomic propagation process. Suppose M is the transition matrix, Φ^0 represents the initial values of influence strength. Then we have $\Phi^{new} = \Phi^0 \cdot M$. It can be easily seen that the direct propagation matrix is Φ^0 itself, thus we have $\Phi^{new} = \Phi^0 \cdot \Phi^0$, which is the matrix of all length-2 paths in our initial influence network.

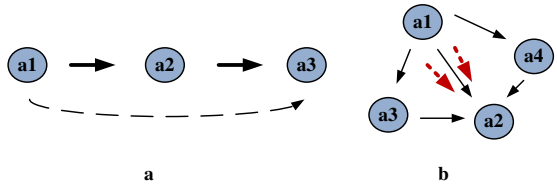

Figure 5: Influence propagation

4.2 Iterative Influence Propagation

The atomic influence propagation can be performed iteratively to propagate the direct influence on the entire network. Formally, we can define the influence propagation process as following:

- Enumerate all paths between each two nodes.
- Calculate the influence propagation strength on each path by applying a concatenation function.
- Combine the influence strength on all the paths by an aggregation function.

In this way, the influence strength on k-length paths can be calculated by k steps of atomic propagations.

We use Φ^k to denote the influence strength after k steps of atomic propagations and Φ^f as the final influence strength that we aim to obtain. Suppose each element $M(u, v)$ in the transition matrix M equals to $\Phi_v(u)$. Thus the iterative propagation can be represented by the matrix powering operation $M^k = M^{k-1} \cdot M$. And we can get $\Phi^k_v(u) = M^k(u, v)$. We assume that the matrix M^k for smaller values of k is more reliable, since there have been fewer propagation steps; while larger values of k may bring in more outside information. Thus the final influence can be inferred from the sequences of propagation via a weighted linear combination [12]:

$$\Phi^f \propto \sum_{k=1}^{K} \delta^{k-1} \cdot \Phi^k \tag{9}$$

where K is the number of iterations and δ can be viewed as the damping factor to penalize larger k step propagation. As $0 < \delta < 1, 0 \leq \Phi^k \leq 1$, when k increases, δ^{k-1} decreases greatly, which makes the effort of influence on k-length paths very small. In another word, we do not need to iterate the influence propagation for many times to obtain the final indirect influence, i.e., K can be set as a small number.

The computation complexity of the propagation process in Eq.(9) is a bit expensive $O(|V|^3 K)$. Fortunately, when K is small, the transition matrix M would be very sparse (many influence scores in the transition matrix are zero), thus the computation complexity actually is reduced to $O(E)$ for each iteration. The algorithm of iterative influence propagation is summarized in Alg. 2.

4.3 Topic-level Influence Propagation

We further extend the influence propagation to the topic-level. The input of the topic-level propagation is the topic distribution of each user learned by the probabilistic model (Section 3) and the topic-level direct influence (Eq. (6)). We perform topic-level influence propagation in the following steps:

- Step 0: Input a query topic z.
- Step 1: According to the topic, the related users are selected based on their topic distributions. The assigned topic is calculated as

$$z = \arg\max_{z'} \psi_x(z') \tag{10}$$

Input: 1) Network G, the Initial local influence Φ^0,
2) the number of iteration K
Output: Final local influence Φ^f

Initialize: $\Phi^f = \Phi^0$;
for k=*1* **to** K **do**
  foreach $\Phi_v^k(u)! = 0$ **do**
    foreach $w : \Phi_u^k(w)! = 0$ **do**
      $\Phi_v^k(w) = \Phi_v^{k-1}(u) \cdot \Phi_u^{k-1}(w)$;
      $\Phi_v^f(w) = \Phi_v^f(w) + \delta^{k-1} \cdot \Phi_v^k(w)$;
    end
  end
end
Return Φ^f;

Algorithm 2: Iterative influence propagation

- Step 2: The influence strength related to the topic is used as the weights of edges.
- Step 3: Employ Alg. 2 to propagate the influence over the network.

If the influence is propagated on the whole network, we have found that some popular nodes may dominate on all topics. Thus we choose users based on the topic z to reduce noise.

4.4 Global Influence Estimation

Global influence is to measure one's influential ability over the whole network. For example, some authors are very influential on the topic of "data mining". In this section, we propose one way to estimate one node's global influence over the whole network.

Intuitively, the global influence of one node on the network $\Lambda(u)$ should be related to its influence on all the other nodes. If one node strongly influences many other nodes, its global influence might be also strong. Therefore the global influence of a node is defined as an aggregation function of its influence on the other nodes, specifically,

$$\Lambda(u) = \sum_v \Phi_v(u) \tag{11}$$

The influence scores $\Phi_v(u)$ include both direct and indirect influences.

5. USER BEHAVIOR PREDICTION

The learned influence strength can be used to help with many applications. Here we illustrate one application on user behavior prediction, i.e., how the learned influence can help improve the performance of user behavior prediction.

We evaluate our approach for user behavior prediction on Twitter and Digg. On Twitter, the behavior is defined as whether a user retweets a friend's microblog and on Digg, the behavior is defined as whether a user digs a story. We here take Digg as the example for explanation. Intuitively, if more friends of a user dig a story, there is a larger probability that the user will also dig it. Thus a vote-based relational neighbor classifier [19] can be used as a baseline. Then, we use the influence strength obtained from our approach to distinguish different friends' weights and estimate the probability of users' digging stories as follows:

$$p(d|u) = \frac{1}{\sum_{v,z} \Phi_u(v|z)} \sum_{v \in Nb(u)} \sum_z \Phi_u(v|z) p(d|v) \tag{12}$$

where $Nb(u)$ denotes the friends of u. We can also apply obtained indirect influence weights for prediction. In this situation, $Nb(u)$ includes 2 or 3 degree of friends.

Besides, the similarity between users can also be used to distinguish different friends' weights in the above intuitive method for prediction. Thus the prediction probability is estimated as Eq.(13) for comparison, where the similarity between users $s(v,u)$ is calculated as the Euclidean distance of user distributions over topics.

$$p(d|u) = \frac{1}{\sum_v s(v,u)} \sum_{v \in Nb(u)} s(v,u) p(d|v) \tag{13}$$

We will test the user behavior prediction performance based on the above three methods in the following experiments and demonstrate the efforts of obtained influence strength on social network applications.

6. EXPERIMENTS

In this section, we present various of experiments to evaluate the efficiency and effectiveness of the proposed approach. The data sets and codes are publicly available [4].

6.1 Experimental Setup

Data Sets We prepare three different types of heterogeneous networks for our experiments, including Twitter, Digg and citation networks. Twitter is a microblog website, on which users can publish blogs and re-tweet friends' blogs. Digg is a different type of social website, on which users can submit, dig and comment on stories. Users also have links to their friends, which indicate their relationship. We collected user and document information from these websites.

- **Twitter social network** The dataset includes about millions of microblogs related to about 40000 users and 50000 keywords (removing the stop words and the infrequent words).
- **Digg social network** The data contain about 1 million stories related to 10000 users and 30000 keywords, on which we aim to mine the user influence as well.
- **Citation network** We crawled the citation data of about 1000 documents from the Internet on several specific topics, e.g., "topic models", "sentiment analysis", "association rule mining", "privacy security" and etc. Besides, the public citation data set Cora is also used in our experiments.

We apply our model to the above three data sets. The algorithms were implemented in C++ and run on an Intel Core 2 T7200 and a processor with 2GB DDR2 RAM. The parameters will be discussed in the following subsections.

Evaluation Aspects We evaluate our method on the following three aspects:

Influence strength prediction As it is more intuitive and easier for people to distinguish the influence strength on citation networks, we manually labeled the citation data and test the influence prediction performance on it. We compare the results of our approach with previous work [5] to prove our model's better performance in terms of influence prediction.

User behavior prediction We apply the derived influence strength to help predict user behaviors and compare the prediction performance with that of baseline as well as the method based on user similarity as described in Section 5. The results demonstrate how the quantitative measurement of the influence can benefit social network applications.

Topic-level influence case study We show several case studies to demonstrate concrete influence weights between users and show

[4]http://arnetminer.org/heterinf

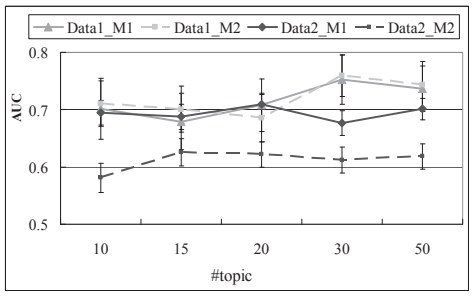

Figure 6: Influence prediction performance comparison

Table 2: Influence prediction performance

direct	indirect	$\delta = 0.5$	$\delta = 0.8$	$\delta = 0.9$	$\delta = 1$
0.6956	$K = 1$	0.7537	0.7555	0.7548	0.7546
	$K = 5$	0.7538	0.7551	0.7547	0.7545
	$K = 10$	0.7538	0.7551	0.7547	0.7545

how effectively our method can identify the topic-level influence. In particular, we study the global influence of authors on citation networks to demonstrate semantic meaning of topic-level influence. And we compare the results with that of previous work [28] which also mined topic-level influence to demonstrate the better performance of our approach.

6.2 Influence Prediction

In work [5], researchers evaluated the document influence prediction performance on manually labeled data set. We got the same data from the authors and also test the influence prediction performance of our model on it. However, the data set, which only contains 22 citing documents and 132 documents in all, is so small that the results could be ad-hoc sometimes. Therefore, besides using this data, we also manually labeled document influence strength on the larger data set with about 1000 documents. We classified the influence strength into three levels: 1, 2, 3. Similar to [5], we use the quality measure, averaged AUC (Area Under the ROC Curve) values for the decision boundaries "1 vs. 2, 3" and "1, 2 vs. 3" for each citing document, to evaluate the prediction performance.

Fig. 6 shows the comparative results on these two data sets, where Data1 is the small data set obtained from authors of [5] while Data2 is our larger labeled data set. $M1$ and $M2$ are used to denote our model and the model in [5] respectively. And we use the real and dash lines to distinguish the results of these two models in the figure. We calculated all the AUC values with the number of topics changing from 10 to 50. Thus this figure demonstrates that on the small data set our model can achieve as good prediction performance as the work in [5] while on the larger data set, our prediction performance is better than theirs.

Furthermore, we compare the influence prediction performance before and after influence propagation on our labeled data set. Table 2 shows the AUC values when damping factor δ and iteration number K changes, which proves that the influence prediction performance is enhanced based on indirect influence obtained by influence propagation. Moreover, the influence prediction performance is robust to the parameters K and δ. In particular, when K changes, the performance change little, which is consistent to the observation in Fig. 2. It means that influence do propagate over the network, but the effort of propagation is reduced a lot when the degree increases.

6.3 User Behavior Prediction

We apply our model on Twitter and Digg social networks and discover the concrete influence strength between users. Then we use the learned influence for user behavior prediction as described in Section 5 to demonstrate the efforts of obtained influence on social network applications. In this experiment, we empirically set the number of topics to be 30, $K = 5$ and $\delta = 0.8$.

Table 3: Behavior prediction probability

Digg Social Network					
method / p	baseline	similarity	direct influence	indirect influence	
				$r = 2$	$r = 3$
average	0.112	0.121	0.366	0.405	0.405
variance	0.006	0.008	0.075	0.048	0.046
Twitter Social Network					
method / p	baseline	similarity	direct influence	indirect influence	
				$r = 2$	$r = 3$
average	0.215	0.222	0.319	0.310	0.308
variance	0.078	0.089	0.129	0.136	0.134

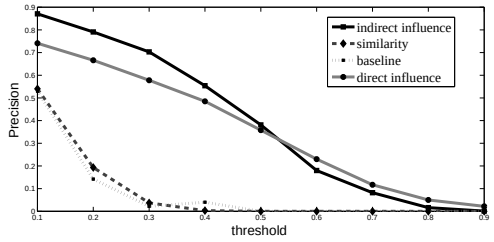

Figure 7: User behavior prediction precision on Digg network

We randomly select about 3000 tuples from Digg and Twitter data sets as testing samples. Each tuple represents that a user digs a story or re-tweets a microblog. We estimate each sample's probability by using the three prediction methods in Section 5. Table 3 shows the average and variance values of the predicted probabilities on all the samples, where r denotes the degree of indirect influence. The results demonstrate that using influence can improve the predicted probabilities a lot.

Then given a threshold, we calculate the prediction precision, which means how many testing samples' probabilities are larger than the threshold. Fig. 7 and 8 show four curves of prediction precision changing with the threshold on Digg and Twitter data sets respectively. The results demonstrate that influence-based behavior prediction approach outperforms the baseline and the similarity-based method. Thus it proves that the influence obtained from our model benefits the user behavior prediction greatly. In particular, it shows that on Digg social network the indirect influence enhances the user behavior prediction performance but on Twitter social network the indirect influence get lower performance than direct influence. Furthermore, comparing these two figures, we can get that the effort of influence on Digg social network is larger than that on Twitter social network. The conclusion is consistent to the observation in Fig. 2.

6.4 Topic-level Influence Case Study

Topic-level influence graph

We apply our model on the citation network which we crawled from the Internet and set the number of topics to be 10 empirically. Fig. 9 demonstrates the influence relationship between the papers on the topic "statistical topic models". The color bars show the topic distributions of these documents. In order to show the major influencing nodes clearly, we rank the influencing nodes according to each influenced node based on the influence strength and only display the top 2 most influencing ones in this figure. Thus we can

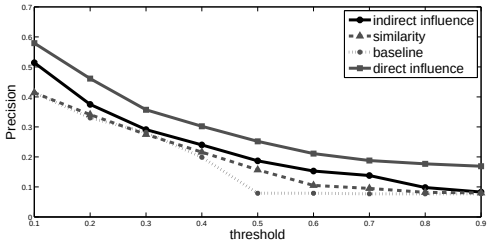

Figure 8: User behavior prediction precision on Twitter network

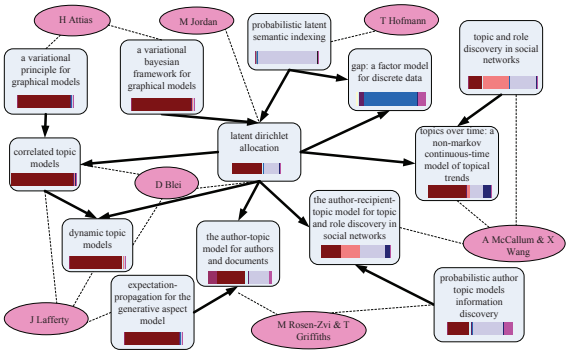

Figure 9: Document influence case study

get that the top 2 most influencing documents on document "LDA" are "PLSA" and "variational inference". Furthermore, the results demonstrate that there are many documents which are most influenced by "LDA", e.g., "the author-topic model", "correlated topic model", "dynamic topic model" and etc. Besides the influence from "LDA", strong influences also exist among these documents, e.g., "author-topic model" influences "author-recipient-model" strongly while "correlated topic model" influences "dynamic topic model" a lot.

Fig. 9 also shows the connections between authors and documents by dash lines. The influence between these authors is visualized in Fig. 10. We only draw the lines when the pointing nodes are the top 5 most influencing authors on the pointed nodes. The thickness of the lines indicates the influence strength. From the results, we can get some meaningful conclusions. For example, Jordan is one of the most influential researchers to Blei. Although "PLSA" strongly influences "LDA" as Fig. 9 shows, Hofmann does not have a great influence on Blei. The reason is that the area of Hofmann varies from the area of Blei (this can be observed from

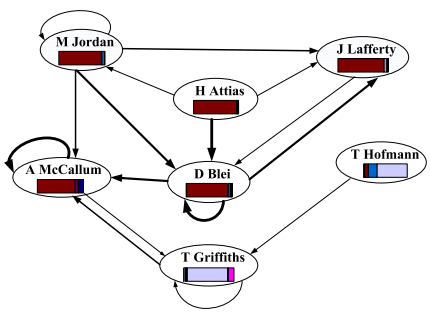

Figure 10: Author influence case study

Table 4: Author ranking on "statistical topic models"

Direct Influence	Indirect Influence		Pagerank
	$K = 1$	$K = 5$	
TM Cover	D Blei	D Blei	M Jordan
A McCallum	A McCallum	A McCallum	D Blei
D Blei	TM Cover	M Jordan	J Lafferty
M Jordan	M Jordan	TM Cover	A McCallum
P Kantor	P Kantor	P Kantor	Z Ghahramani

Table 5: Influence aggregation values on topics

Topic	OODB	IR	DM	DBP
Maximal value	2.525	2.333	3.877	3.607
Minimal value	0.0005	0.001	0.0006	0.0009
Average value	0.078	0.091	0.095	0.087
D DeWitt	1.487	0.181	1.087	**3.607**
M Stonebraker	**2.525**	0.632	0.481	**2.851**
C Faloutsos	0.357	0.242	**1.571**	**1.187**
W Bruce	0.538	**2.333**	0.172	0.483
R Agrawal	0.518	0.189	**3.877**	0.600
J Han	0.666	0.138	**2.029**	0.240

the topic distributions represented by colored bars) and furthermore Blei only cited few documents of Hofmann, i.e., correlation value is small. Other interesting results are also obtained, e.g., the influence of Blei on Lafferty is larger than the influence of Lafferty on Blei. Besides, the self-loop lines which indicate the self-influence show Jordan and Blei influence themselves greatly.

Topic-level global influence illustration

Table 4 shows an example of author ranking by estimated global influence on "statistical topic models" (K denotes the number of propagation steps). The results are very meaningful. If one node has a high reputation over the whole network, it can be treated as a key node which is very influential over the whole network. In another word, authority of one node can also be used to represent its global influence from some point of view. Therefore, we can employ Pagerank [21, 13] on topic-level networks to estimate the nodes' global influence on one topic. The author ranking based on the authority from Pagerank is also illustrated. We calculate the correlation coefficients between the global influence values estimated in the two ways, which ranges from 0.8 to 0.9 when the number of topics and iteration change. It proves that estimating global influence based on our framework can get highly-correlated results with Pagerank authority. Thus, to some extent, it demonstrates that the influence discovered by our model is consistent to the global characteristics of the whole network structure.

In order to show the influence results on more general areas, we selected five categories of documents in Cora data and set the number of topics to be 5. Five meaningful topics according to the five categories: data mining ("DM"), information retrieval ("IR"), natural language processing ("NLP"), object oriented database ("OODB") and database performance ("DBP") are obtained. Fig. 11 shows several famous authors' estimated global influence distributions on the five topics. The results are very telling. For example, W Bruce is most influential on topic "IR", while R Agrawal and J Han are most influential on topic "DM". It is interesting to find that C Faloustsos is influential on both topic "DM" and topic "DBP", which is consistent to the real situation. Besides on the two topics related to database, D DeWitt is also very influential on topic "DM". The reason should be that the area "DM" is developed from database. Furthermore, Table 5 shows the maximal, minimal and average values of the estimated global influence on the whole network w.r.t. each topic, which demonstrates that these authors almost have the

Table 6: Influencing author ranking w.r.t. several authors

D Blei		A McCallum		T Griffiths	
M1	M3	M1	M3	M1	M3
H Attias	D Blei	A McCallum	A McCallum	T Hofmann	T Griffiths
D Blei	M Stephens	D Blei	D Kauchak	M Steyvers	R Kass
M Jordan	J Pritchard	Andrew Ng	E Stephen	T Griffiths	N Chater
K Nigam	P Donnelly	T Griffiths	R Madsen	T Minka	D Lawson
T Jaakkola	C Meghini	M Jordan	C Elkan	A McCallum	H Neville

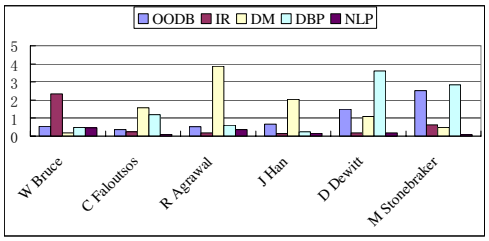

Figure 11: Estimated global influence distribution on topics

largest values in their domains. Thus it proves the validity of the way of global influence estimation.

Topic-level influence comparison

Work [28] also proposed a method to discover topic-level influence. We compare the author influence results obtained by our model ($M1$) with the results by the model in [28] ($M3$). As sometimes it is hard to label the author influence strength, we only show the top 5 most influencing authors on some well-known researchers: Blei, McCallum and Griffiths obtained by these two models in Table 6. The results demonstrate that our model can get meaningful results but $M3$ can not. For example, our model discovers that Jordan, Blei and Hofmann are one of the most influential researchers for Blei, McCallum and Griffiths respectively. But $M3$ does not get these results. As $M3$ only uses the link information of author citation, it will lose the information of relationships between authors and documents. And the assumption used in [28] which states that the node will be more influential if it has a great self-influence makes each person most influential on himself.

Similar to our model, $M3$ can also get the influence distributions on topics by inputting the nodes' topic mixtures. But the difference is that the topic information is used as an input prior instead of an integrated parameter in the method $M3$ while our method can obtain topics simultaneously. Fig. 12 shows an example of the influence from Jordan to Blei and compares the topic distributions of influence obtained by our model and $M3$ respectively. First, Jordan and Blei's distributions on topics are illustrated, which indicate that both of them mainly work on Topic 3. Then, we can see that the influence obtained by our model has the largest strength on Topic 3 but the influence distribution from $M3$ is flat, from which it is not obvious to tell the influence semantic meaning. Thus it is proved that our model can obtain more meaningful topic distributions of influence.

7. RELATED WORK

Heterogeneous Network Analysis Heterogeneous network and source analysis has attracted many researchers' interests recently [26, 27, 33]. For example, Sun et al.[26, 27] studied the clustering problem on heterogeneous networks. Ye et al.[33] fused heterogeneous data sources to study the alzheimer's disease. Many works have tried to combine the sufficient information on heterogeneous networks, e.g., the text and links, to detect communities, to analyze

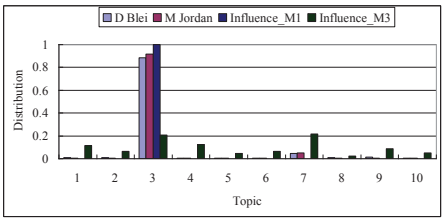

Figure 12: Topic distributions of authors and influence

the evolution of networks and to model relational learning [32, 2, 34, 29]. Besides, some researchers studied the problem of information diffusion over networks [16, 11].

Link Influence Analysis Link influence analysis has attracted tremendous interests from both academic and industry communities. Many efforts have been made for estimating link influence between individual pages. For example, Dietz et al. [5] proposed a citation influence topic model to model the influential strength between papers. Nallapati et al. [20] proposed two topic models to jointly model text and citation relationships. King et al. [15] analyzed the influence factor among paper citation networks. The goal of this kind of work is to estimate the influence of a citation in the whole citation collection, thus the objective differs largely from ours.

Social Influence Analysis Considerable work has been conducted to validate the existence of influence and study its effort from the global view of the whole network, e.g., influence maximization on a person network [6, 22, 14, 9, 3]; influence diffusion over networks [23]; influence and correlation on social activities [1]; correlation between influence and similarity [4]. Several efforts have been made to identify the existence of social influence in online social networks. For example, Anagnostopoulos et al. [1] gave a theoretical justification to identify influence as a source of social correlation when the time series of user actions are available. They proposed a shuffle test to prove the existence of social influence. Singla and Richardson [25] studied the correlation between personal behaviors and their interests. They found that in online systems people who chat with each other (using instant messaging) are more likely to share interests (their Web searches are the same or topically similar), and the more time they spend talking, the stronger this relationship is. Crandall et al. [4] further investigated the correlation between social similarity and influence. Tang et al. [28] introduced the problem of topic-based social influence analysis. They proposed a Topical Affinity Propagation (TAP) approach to describe the problem using a graphical probabilistic model.

More recently, some attempts have been made to analyze the dynamics in the social networks. For example, Scripps et al. [24] investigated how different pre-processing decisions and different network forces such as selection and influence affect the modeling of dynamic networks. Other similar work can be referred to [7]. However, most of the aforementioned methods only focus on homogeneous networks with the same type of nodes.

8. CONCLUSIONS AND FUTURE WORK

In this paper, we study a novel problem of mining topic-level influence on heterogeneous networks. Our approach to solve this problem primarily consists of two steps, i.e., a probabilistic model to mine direct influence between nodes and a topic-level influence propagation method to mine indirect and global influence. In the probabilistic model, we combine the text content and heteroge-

neous link information into a unified generative process. The topic-level influence propagation method further propagates the influence along the links in the entire network. We have done extensive experiments on different types of heterogeneous networks, show some interesting cases and demonstrate that using influence can benefit the prediction performance greatly.

The general problem of influence analysis on informative networks represents a new and interesting research direction in social network mining. There are many potential future directions of this work. One interesting issue is to employ more robust models to predict user behavior based on the obtained influence strength and study a semi-supervised learning framework to incorporate user feedbacks into our approach. Another potential issue is to scale up the approach to large data set.

9. ACKNOWLEDGEMENTS

The work was supported in part by 973 Program of China under grant 2011CB302206, and the State Key Program of National Natural Science of China under grants 60933013, 60833009, and 60703059, and National High-tech R&D Program 2009AA01Z138, and the U.S. National Science Foundation under grant IIS-09-05215 and by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA).

10. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD'08*, pages 7–15, 2008.

[2] J. Chang, J. Boyd-Graber, and D. M. Blei. Connections between the lines: augmenting social networks with text. In *KDD '09*, pages 169–178, 2009.

[3] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD '10*, 2010.

[4] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD'08*, pages 160–168, 2008.

[5] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *ICML'07*, pages 233–240, 2007.

[6] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD' 01*, pages 57–66, 2001.

[7] Y. Dourisboure, F. Geraci, and M. Pellegrini. Extraction and classification of dense communities in the web. In *WWW '07*, pages 461–470, 2007.

[8] J. H. Fowler and N. A. Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. In *British Medical Journal*, 2008.

[9] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *WSDM'10*.

[10] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[11] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04*, pages 491–501, 2004.

[12] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW '04*, pages 403–412, 2004.

[13] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW'02*, pages 517–526, 2002.

[14] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03*, pages 137–146, 2003.

[15] J. King. A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, 13(5), 1987.

[16] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. In *KDD '08*, pages 435–443, 2008.

[17] D. Krackhardt. *The Strength of Strong ties: the importance of philos in networks and organization in Book of Nitin Nohria and Robert G. Eccles (Ed.), Networks and Organizations*. Cambridge, Harvard Business School Press, Hershey, USA, 1992.

[18] T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW '10*, pages 601–610, 2010.

[19] S. Macskassy and F. Provost. A simple relational classifier. In *Workshop on Multi-Relational Data Mining in conjunction with KDD'03*, 2003.

[20] R. M. Nallapati, A. Ahmed, E. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *KDD '08*, pages 542–550, 2008.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.

[22] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD '02*, pages 61–70, 2002.

[23] M. G. Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *KDD '10*, 2010.

[24] J. Scripps, P.-N. Tan, and A.-H. Esfahanian. Measuring the effects of preprocessing decisions and network forces in dynamic network analysis. In *KDD '09*, pages 747–756, 2009.

[25] P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *WWW '08*, pages 655–664, 2008.

[26] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. RankClus: integrating clustering with ranking for heterogeneous information network analysis. In *EDBT'09: Proceedings of Int. Conf. on Extending Data Base Technology*, Saint-Petersburg, Russia, March 2009.

[27] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *KDD '09*, pages 797–806, 2009.

[28] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD '09*, pages 807–816, 2009.

[29] L. Tang and H. Liu. Relational learning via latent social dimensions. In *KDD '09*, pages 817–826, 2009.

[30] J. Whitfield. The secret of happiness: grinning on the internet. In *Nature*, 2008.

[31] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *WWW '10*, pages 981–990, 2010.

[32] T. Yang, R. Jin, Y. Chi, and S. Zhu. Combining link and content for community detection: a discriminative approach. In *KDD '09*, pages 927–936, 2009.

[33] J. Ye, K. Chen, T. Wu, J. Li, Z. Zhao, R. Patel, M. Bae, R. Janardan, H. Liu, G. Alexander, and E. Reiman. Heterogeneous data fusion for alzheimer's disease study. In *KDD '08*, pages 1025–1033, 2008.

[34] E. Zheleva, H. Sharara, and L. Getoor. Co-evolution of social and affiliation networks. In *KDD '09*, pages 1007–1016, 2009.

APPENDIX

Based on the generation process, we can get the posterior probability of the whole data set by integrating out the multinomial distributions $\lambda, \gamma, \psi, \theta, \phi$ because the model uses only conjugate priors.

$$\begin{aligned} & p(\vec{w}, \vec{w}', \vec{z}, \vec{z}', \vec{s}, \vec{y}|\vec{\alpha}_\phi, \vec{\alpha}_\theta, \vec{\alpha}_\psi, \vec{\alpha}_\lambda, \vec{\alpha}_\gamma) \\ \propto & \int p(\vec{s}|\vec{\lambda}, \vec{x}) p(\vec{\lambda}|\vec{\alpha}_\lambda) d\vec{\lambda} \int p(\vec{z}, \vec{z}'|\vec{y}, \vec{s}, \vec{x}, \vec{\theta}, \vec{\psi}) p(\vec{\theta}|\vec{\alpha}_\theta) p(\vec{\psi}|\vec{\alpha}_\psi) d\psi\theta \\ & \int p(\vec{y}|\vec{x}, \vec{\gamma}, A) p(\vec{\gamma}|\vec{\alpha}_\gamma) d\vec{\gamma} \int p(\vec{w}, \vec{w}'|\vec{z}, \vec{z}', \vec{\phi}) p(\vec{\phi}|\vec{\alpha}_\phi) d\vec{\phi} \end{aligned} \quad (14)$$

In the following, we exemplify the derivation of the update equation for s_i and the other variables are derived analogously. The conditional of s_i is obtained by dividing the joint distribution of all variables by the joint with all variables but s_i (denoted by $\vec{s}_{-i}$) and canceling factors that do not depend on $\vec{s}_{-i}$.

$$\begin{aligned} & p(s_i = 0|\vec{s}_{-i}, x_i, z_i, .) \\ = & \frac{p(\vec{w}, \vec{w}', \vec{z}, \vec{z}', s_i, \vec{y}|\vec{\alpha}_\phi, \vec{\alpha}_\theta, \vec{\alpha}_\psi, \vec{\alpha}_\lambda, \vec{\alpha}_\gamma)}{p(\vec{w}, \vec{w}', \vec{z}, \vec{z}', \vec{s}_{-i}, \vec{y}|\vec{\alpha}_\phi, \vec{\alpha}_\theta, \vec{\alpha}_\psi, \vec{\alpha}_\lambda, \vec{\alpha}_\gamma)} \\ = & \frac{\int p(s_i|\vec{\lambda}, \vec{x}) p(\vec{\lambda}|\vec{\alpha}_\lambda) d\vec{\lambda}}{\int p(\vec{s}_{-i}|\vec{\lambda}, \vec{x}) p(\vec{\lambda}|\vec{\alpha}_\lambda) d\vec{\lambda}} \cdot \\ & \frac{\int p(\vec{z}, \vec{z}'|\vec{y}, s_i, \vec{x}, \vec{\theta}, \vec{\psi}) p(\vec{\theta}|\vec{\alpha}_\theta) p(\vec{\psi}|\vec{\alpha}_\psi) d\psi\theta}{\int p(\vec{z}, \vec{z}'|\vec{y}, \vec{s}_{-i}, \vec{x}, \vec{\theta}, \vec{\psi}) p(\vec{\theta}|\vec{\alpha}_\theta) p(\vec{\psi}|\vec{\alpha}_\psi) d\psi\theta} \end{aligned} \quad (15)$$

We derive the first fraction of Eq. (15) (the second fraction is derived analogously). As we assume that s_i is generated from a Bernoulli distribution λ whose Dirichlet parameters are $\alpha_{\lambda_{s_0}}, \alpha_{\lambda_{s_1}}$, then we can get $p(s_i|\vec{\lambda}, \vec{x}) = \prod_i \alpha_{\lambda_{s_0}}^{N_{x,s}(x_i,0)} \cdot \alpha_{\lambda_{s_1}}^{N_{x,s}(x_i,1)}$, where $N(*)$ is the function which stores the number of samples during Gibbs sampling. For example, $N_{x,s}(x_i, 0)$ represents the number of samples when user x_i is influenced to generate a topic. Because we only use conjugate priors in the model, the multinomial-Dirichlet integral in Eq. (15) has a closed form solution. Thus we can get that when $s_i = 0$, the first fraction can be derived as below

$$\frac{\int p(s|\vec{\lambda}, \vec{x}) p(\vec{\lambda}|\vec{\alpha}_\lambda) d\vec{\lambda}}{\int p(\vec{s}_{-i}|\vec{\lambda}, \vec{x}) p(\vec{\lambda}|\vec{\alpha}_\lambda) d\vec{\lambda}} = \frac{N_{x,s}(x_i, 0) + \alpha_{\lambda_{s_0}}}{N_x(x_i) + \alpha_{\lambda_{s_0}} + \alpha_{\lambda_{s_1}}} \quad (16)$$

The other equations can be derived analogously.