Noname manuscript No.
(will be inserted by the editor)

Learning Influence from Heterogeneous Social Networks

**Lu Liu · Jie Tang · Jiawei Han ·
Shiqiang Yang**

Received: date / Accepted: date

Abstract Influence is a complex and subtle force that governs social dynamics and user behaviors. Understanding how users influence each other can benefit various applications, e.g., viral marketing, recommendation, information retrieval and etc. While prior work has mainly focused on qualitative aspect, in this paper, we present our research in quantitatively learning influence between users in heterogeneous networks. We propose a generative graphical model which leverages both heterogeneous link information and textual content associated with each user in the network to mine topic-level influence strength. Based on the learned direct influence, we further study the influence propagation and aggregation mechanisms: conservative and non-conservative propagations to derive the indirect influence. We apply the discovered influence to user behavior prediction in four different genres of social networks: Twitter, Digg, Renren, and Citation. Qualitatively, our approach can discover some interesting influence patterns from these heterogeneous networks. Quantitatively, the learned influence strength greatly improves the accuracy of user behavior prediction.

Keywords Social influence analysis · Social network analysis · Influence propagation · Topic modeling

1 Introduction

It is well known that influence is a complex and subtle force to govern user behaviors and relationship formation in social networks. With the power of influence, a company can market a new product by first convincing a small number of influential

L. Liu
Capital Medical University
E-mail: lu-liu@mails.tsinghua.edu.cn

J. Tang · S. Yang
Tsinghua University

J. Han
University of Illinois at Urbana-Champaign

users to adopt the product and then triggering further adoptions through the effect of "word of mouth" (also referred to as influence maximization [10,36,25,30]). In academic networks, thanks to the influence between research collaborators, novel ideas or innovations quickly spread and lead the blooming of new academic directions. On social websites, e.g., Facebook and Twitter, users are very likely to follow influential friends in their social circle to retweet a microblog or to "like" a picture.

An interesting question is: how friends in a social network influence each other and how the influence is spreading in the social network? Answering the question is non-trivial. Indeed, it is challenging on the following aspects.

First, what are the fundamental (micro-level) mechanisms of social influence in social networks? In particular, when social networks are heterogeneous (consisting of heterogeneous objects such as users, groups, and blogs), how the influence is affected by different types of objects on different topics (e.g., entertainment, marketing, and research)? Recently, web users enjoy sharing or spreading interesting User Generated Content (UGC), e.g., users re-tweet microblogs on Twitter and dig stories on Digg, etc. Social networks closely inosculate with UGC in result of many heterogenous networks. Thus besides the network structure, the content spreading on the top of networks becomes a key factor for social influence mining in heterogeneous networks. For example, students' research interests are greatly influenced by their advisors. While, their hobbies may be mainly influenced by their family members or close friends in their daily life. Thus influence strength varies with topics. The problem of jointly learning topic distribution associated with each user and topic-level influence between users has not been addressed before.

Second, can your friends' friends have some kind of influence on your behaviors? Interestingly, the answer is "Yes". For example, Fowler, Christakis [13] and Whitfield [46] 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. Then a straightforward question is: how the influence propagates in social networks? Existing works such as [13] and [46] merely qualitatively test indirect influence on two small data sets. A systematic investigation of this problem is still needed.

Social influence analysis has attracted considerable research interests and is becoming a popular research topic. However, most existing works have focused on validating the existence of influence [1,7], or studying the maximization of influence spread in the whole network [25,6], or modeling only direct influence in homogeneous networks [9,44,47]. The micro-level mechanisms of social influence w.r.t. topics and its propagation over social networks have been largely ignored.

Contributions In this article, we aim to systematically and quantitatively study how friends influence each other and how influence spreads in heterogeneous social networks. Our objective is to effectively and efficiently discover the underlying influence patterns in heterogeneous networks. Building on our previous research work [32], we aim to provide a more comprehensive analysis on this problem, which can be explained by using the example in Fig. 1. The input (left figure) is a heterogeneous network consisting of web documents, users, and links between them. To leverage both content information of web documents and social network structure, we propose a probabilistic generative model to jointly learn topics and to associate a topic distribution with each user which indicates his/her interests. Based on the modeling results, we can estimate

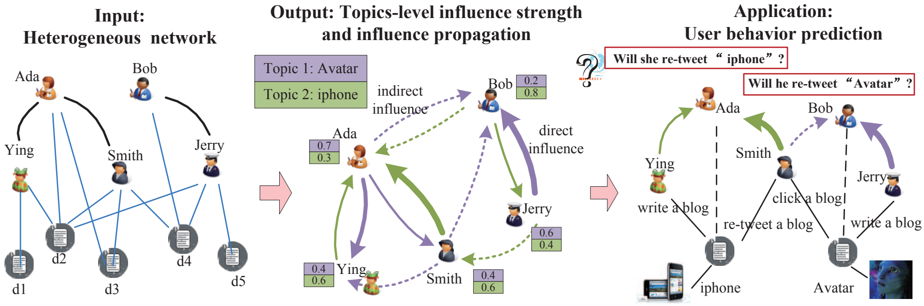


Fig. 1 Problem illustration of mining topic-level influence in heterogeneous networks and predicting user behaviors.

the influence strength between friends. We further investigate two kinds of diffusion models for conservative and non-conservative influence propagations in social networks, which uncover the indirect influence between non-connected users. The middle figure of Fig. 1 illustrates the output of topic discovery and influence propagation. The solid arrow indicates direct influence and the dashed arrow indicates indirect influence. Our last task is to validate how the discovered influence can really help. We apply the discovered influence to user behavior prediction. Extensive experiments in real social networks are conducted to evaluate the effect of influence in terms of user behavior prediction performance.

To summarize, this work contributes on the following aspects:

- We formulate the problem of topic-level influence mining and propose a generative model which utilizes both content and link information to mine direct influence strength in heterogeneous networks.
- We study two kinds of diffusion models for conservative and non-conservative influence propagations to learn indirect influence in social networks.
- We apply the discovered influence strength to user behavior prediction and validate how it can help some social applications.
- We conduct extensive experiments in four different types of data sets: Twitter[1], Digg[2], Renren[3] and Cora[4], and test the model performance in both qualitative and quantitative ways.

The rest of the paper is organized as follows: Sect. 2 formally formulates the problem; Sect. 3 illustrates some interesting observations on relationships between social influence and other factors in social networks; Sect. 4 proposes a probabilistic generative model to discover topics and direct influence strength; Sect. 5 defines influence propagation process and Sect. 6 studies conservative and non-conservative diffusion models to derive indirect influence in social networks; Sect. 7 introduces the application of user behavior prediction based on the discovered influence; Sect. 8 presents experimental

[1] http://www.twitter.com, a microblogging system.

[2] http://www.digg.com, a social news sharing and voting website.

[3] http://www.renren.com, one of the largest Facebook-like social networks in China.

[4] http://www.cs.umass.edu/mccallum/code-data.html, a bibliographic citation network

results that validate the effectiveness of our methodology; Sect. 9 discusses related work and Sect. 10 concludes.

2 Problem Formulation

In this section, we introduce several related concepts and then formulate the problem of mining topic-level influence in heterogeneous networks.

Definition 1 [Heterogeneous Social Network] Define a network as $G = (V, D, E)$, where V is a set of user nodes, D is a set of document nodes, and E denotes a set of edges that includes social relationships connecting users and links connecting users and documents. For each edge $e_{uv} = (u, v) \in E$, if there exists an edge between u and v, $e_{uv} = 1$; otherwise $e_{uv} = 0$. The edges can be directed or undirected.

Many online social networks are heterogeneous consisting of different types of object nodes. For example, Twitter is comprised of users and microblogs. Digg consists of users and website URL addresses. Citation network consists of authors and publication papers. Here, we use "document" to represent different types of associated content (e.g., microblog, website, and paper) to each user. Thus links in heterogeneous networks would contain friendships between users and authoring relationships between users and documents (links between documents are not considered in this paper). The links can be directed or undirected. For example, in Twitter and citation networks, the links between users are directed from normal users to their followers. In Digg social network, the links between users are undirected. Furthermore, we assume that influence can propagate 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 G, we denote $\delta_v(u) \in \{R^+ \cup 0\}$ as the influential strength of user u on user v. Furthermore, if $e_{uv} = 1$, we call $\delta_v(u)$ the direct influence of user u on v; if $e_{uv} = 0$, we call $\delta_v(u)$ the indirect influence of user u on v.

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

Definition 3 [Global Influence] Given a heterogeneous network, $\Lambda(v) \in \{R^+ \cup 0\}$ is defined as the global influence of v, which represents the global influential strength of user v in the network.

The global influence strength has a close relationship with the (local) direct/indirect influence. For example, if a user has a strong influence on her/his friends, it is very likely that she/he has a strong global influence.

Our formulation of topic-level influence mining is quite different from existing works on social influence analysis. Works [1] and [39] study how to qualitatively measure the existence of influence. Crandall etc. [7] study the correlation between social similarity and influence. However, they focus on qualitative identification of influence existence, but do not provide a quantitative measure of the influential strength. Tang et al. [44] try to learn the influence probabilities according to the network structure and the similarity between nodes. Works [15,47] further investigate 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 influence propagation. The challenge of our work is how to jointly learn the topics and the topic-level (direct and indirect) influence from heterogeneous networks. The learned social influence has a number of immediate applications such as influence maximization [10,25,36], social action prediction [20,42].

2.1 Intuitions and Our Approach

To summarize, we have two important intuitions for learning influence from heterogenous social networks: (1) influence between users varies over different topics; and (2) user behaviors are not only influenced by their friends but also their n-degree friends (e.g., friends' friends). Indeed, in real networks users may be interested in different topics, e.g., in the research network an author may be interested in topics "database" and "data mining". The influential strength from one's coauthors on her/him w.r.t. the two topics might be very different. Actually, this has been qualitatively verified in sociology [16,29] and quantitatively studied in [44]. More precisely, we can give the following descriptions for the intuitions:

1. Each node v is associated with a vector $\psi_v \in R^T$ of T-dimensional topic distribution ($\sum_z \psi_v(z) = 1$), where $\psi_v(z)$ indicates the interest probability of the node (user) on topic z.
2. Influence can propagate over social networks, thus the influence $\delta_v(u)$ 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 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 his friends have digged this story or simply because he is interested in this topic.

From the technique perspective, our objective is to design a method to learn user interests (the associated topic distribution) and to estimate user influence 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 mixtures of topics and direct influence between users simultaneously. Second, based on direct influence, we study two types of influence propagation mechanism, which are conservative and non-conservative influence propagations, to derive indirect influence between users.

Our definition of influence is different from other social factors (e.g., similarity and tie strength) on the following aspects:

- According to the above intuitions, our definition of influence is based on the dynamic process of user behaviors, which is related to both content and network structure in heterogeneous networks. But similarity is more likely to be defined based on content, while tie strength measures the kinship between two persons which is likely to be related to common neighborhood.

- We investigate influence propagation in this paper, which is an important property of social influence and can be used in applications such as influence maximization. Similarity or tie strength does not have the propagation property. Thus they are different from social influence.
- The obtained influence strength from our model is directed, which means the social influence from user A to user B is different from that from B to A. While, both similarity and tie strength are symmetric measurements.

In Section 8.3, we will compare the performance of user behavior prediction based on influence with the results based on other social factors.

3 Observations

In order to be fully aware of the effect of social influence, we first conduct a series of analysis before proposing our approach. We focus on four aspects: (1) *influence vs. activity*: how one's activity impacts his/her influence strength? (2) *influence vs. degree centrality*: how one's influence on his friends correlates with his/her degree centrality? (3) *influence vs. similarity*: how influence between friends correlates with their similarity? and (4) *influence vs. n-degree*: will a user influence his n-degree friends and how?

Influence vs. Activity In online heterogeneous networks, some users are much more active than some others. Taking Renren for example, some users share many web documents while some others are very silent. Then a question arise: "Is the influence of a user related to his/her activity?". To answer this question, we show some observations and analysis in Renren data set here.

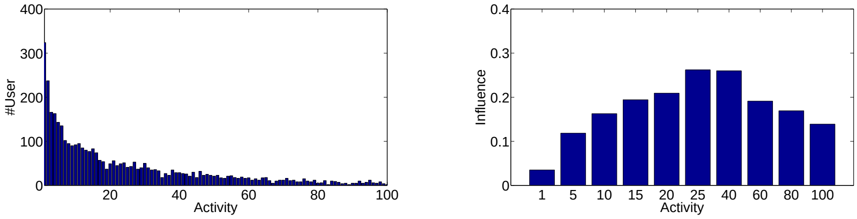
Suppose $I(v,d)$ denotes the connection between a user v and a document d, i.e., if user v shares or re-tweets document d, $I(v,d)=1$; otherwise $I(v,d)=0$. Then the influence strength of a user v is simply approximated as Eq. (1).

$$p_1(v) = \max_{d:I(v,d)=1} \frac{\sum_{u \in Nb(v)} I(u,d)}{|Nb(v)|} \tag{1}$$

Thus if v shares a document and his/her friends also shares this document, we think that the friends are influenced by v. Thus we use the ratio of v's friends who have the same actions to estimate the influence strength of v. The maximal influence strength w.r.t. document in Eq. (1) is used to approximate a user's influence strength in order to overcome the noise problem. On the other hand, the number of documents shared by a user is utilized to indicate the activity of this user, i.e., $Act(v) = \sum_d I(v,d)$.

We analyze 5000 users from Renren. Fig. 2(a) shows that the number of users with the same activity value decreases with the increase of user activity factor $Act(v)$ and most of users share about 0 to 30 web documents. We calculate the average influence strength of users who share 1, 5, 10, 15, 20, 25, 40, 60, 80, 100 web documents respectively as shown in Fig. 2(b), which demonstrates that with the increase of user activity, user influence first increases a lot then decreases a little. The result is interesting and also intuitive. An active user seems to be more likely to influence his/her friends to act in the same way.

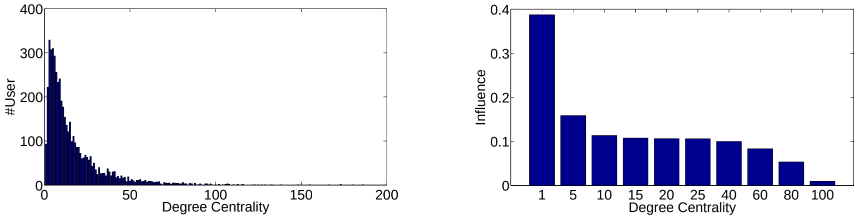
Influence vs. Degree Centrality Besides the connections between users and documents, there are also links between users in heterogeneous networks. Some users are



(a) User number v.s. activity

(b) User influence v.s. activity

Fig. 2 User number and user influence strength changing with user activity



(a) User number v.s. degree

(b) User influence v.s. degree

Fig. 3 User number and user influence strength changing with degree centrality

popular and have many connections with the others. Degree centrality is a measure in graph theory to determine the relative importance based on the number of links. Here we try to examine whether users with a higher degree centrality have a stronger influence in the social network. Again, we study this problem on the Renren data. The influence strength of a user is estimated by Eq. (1). Fig. 3(a) shows that the user number first increases and then decreases with the increase of user degree and most of users have about 10 friends in this social networks. Fig. 3(b) demonstrates that the user influence strength is weakening with degree, which is consistent to some sociological research results, i.e., when a user has more and more friends, his/her attention paid to each friend will be reduced and his/her friend connection will not be so close as before, which results in his/her influence strength weakening.

Influence vs. Similarity Does influence correlate with similarity in heterogeneous networks? When the similarity between two nodes increases, how does their potential influence strength change? In Renren data set, we analyze the relationship between influence and similarity among node pairs.

First, each user is represented as a keyword vector based on the document content he/she has shared. Then their similarity is estimated by the Cosine-distance of these two keyword vectors. The influence from user v to user u is estimated as the ratio of actions that u has followed v, i.e.,

$$Inf(v \rightarrow u) = \frac{\sum_{d:I(v,d)=1} I(u,d)}{\sum_{d:I(v,d)=1} I(v,d)} \tag{2}$$

Then the correlation coefficient between influence and similarity calculated by the above roughly estimation methods in Renren data set is about 0.24. This result demonstrates that user influence is positive correlated with similarity, but they are still dif-

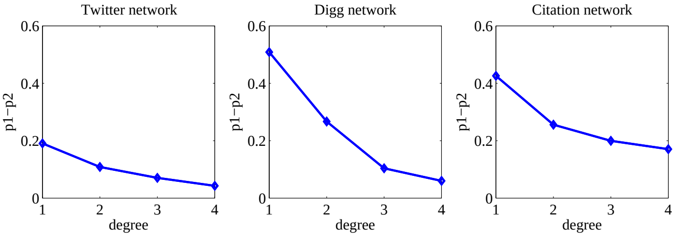


Fig. 4 $n(1 \leq n \leq 4)$-degree influence in three networks: Twitter, Digg, and Cora.

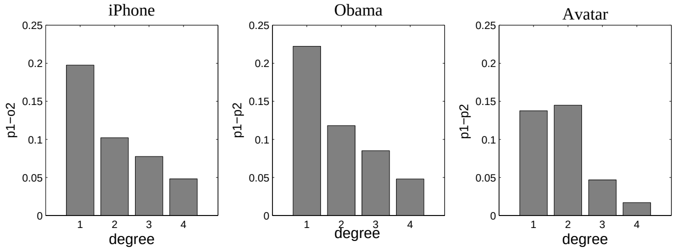


Fig. 5 Topic-level influence on three topics: “iPhone”, “Obama”, and “Avatar”.

ferent factors with different effects in social networks as their correlation coefficient is not so big.

Influence vs. n-degree Influence propagates over social networks as discussed in Sect. 2. In order to verify the existence of indirect influence and to study the influence propagation mechanism, we conduct an analysis on the influence strength changing with propagation length.

We estimate the influence strength after i-step propagation as Eq. (3)

$$p_1(v) = \max_{d:I(v,d)=1} \frac{\sum_{u \in Nb_i(v)} I(u,d)}{|Nb_i(v)|} \tag{3}$$

To test the effect of influence propagation, Eq. (3) extends Eq. (1) by enlarging v's neighborhood to $Nb_i(v)$. $Nb_i(v)$ includes v's possible accessed friends after i-step propagation. For example, when influence propagates one step, $Nb_1(v)$ includes v's friends' friends (named as two-degree friends in [13, 46]) which can be accessed through one of v's friends who has shared d, i.e., if $w \in Nb(v)$ and $I(w,d) = 1$ and $u \in Nb(w)$, then $u \in Nb_1(v)$.

In order to form a close community, the 5000 users in Renren data set are selected from a user's two-degree friend neighborhood. Thus we calculate each user's influence strength on one-degree friends as well as that on two-degree friends, which are 0.2 and 0.05 respectively. Thus the two-degree influence strength decreases a lot, which is only 25% of one-degree influence strength in Renren data set. Besides, to further study influence strength changing with propagation step, we calculate three and four-degree influence strength in other three heterogeneous networks – Twitter, Digg, and Cora. Fig. 4 demonstrates that indirect influence also exists in other social networks. For example, on Renren, when a user shares an interesting poster, his/her friends' friends (2-degree friends) averagely have a 20+% higher probability to follow him/her. However, the influence strength decreases with the increase of propagation length on

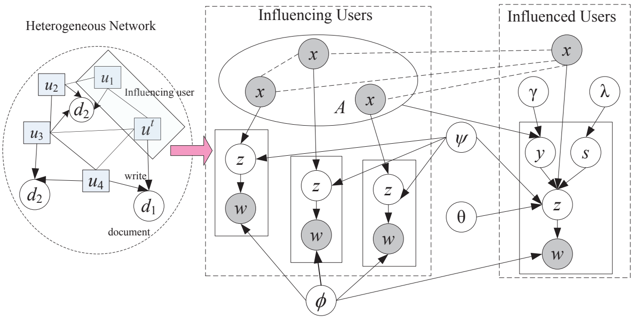


Fig. 6 Probabilistic generative model to estimate direct influence strength

average. Furthermore, the influence patterns on these networks are quite different. For example, influence on Twitter is small and gradually decreases with the increase of degree. While on Digg, influence tapers off quickly with the increase of propagation length. Furthermore, we analyze the topic-level influence on Twitter as shown in Fig. 5. We study the n-degree influence on three topics: “Obama”, “iphone” and “Avatar”. These influence changing patterns are different. An interesting phenomenon is that on some topics the two-degree influence is even stronger than the one-degree influence (e.g., on “Avatar”). his is because “Avatar” is a very popular topic, on which the users may be mainly influenced by the global trend (via conformity [24]).

In summary, according to the approximate analysis above, we have the following observations:

- Active users are likely to be more influential. But when the user activity increases to a certain level, it may be no longer the major factor to impact the user influence.
- When a user has more friends, his/her attention paid to each friend would be reduced and his/her friend connection would not be so close as before, which may result in his/her influence strength weakening.
- User influence is positive correlated with similarity, but they are still different factors with different effects in social networks.
- Indirect influence exists in social networks, which decreases with the increase of propagation length generally. And influence strength changing patterns vary with topics.

4 Mining Influence in Heterogeneous Networks

Influence is interacted with many potential factors, e.g., similarity and correlation [1, 7]. Here we have two general assumptions in order to model the influence strength quantitatively.

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

This assumption actually corresponds to the influence and selection theory [1]. We have observed that user influence is positive correlated with similarity in Sect 3. In real networks, the similarity can be calculated based on the content information associated with each user. Thus, influence can be represented as to which extent the

Table 1 Variable descriptions

Notation	Description
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_x	the user list who may influence x
y	the influencing user from A_x
s	the label denoting either influencing or not
W	the number of words in the data set
T	the number of topics to be extracted
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

textual content 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.

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

The co-occurrence frequency is often used to indicate the correlation strength between two nodes, which is denoted by the weights of edges in networks. Thus the influence strength between two nodes would be enlarged by their frequent co-occurrence. For example, if author a cites a number of papers of author b, then a should be strongly influenced by b. For another example on Twitter, if user a replies or re-tweets many microblogs posted by user b, then 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 direct influence strength between users quantitatively.

4.1 Probabilistic Generative Model

In Sect. 3 we have observed that influence strength varies with topics. Thus in this section we design a model to mine topics and influence strength simultaneously. The model combines the content information and network structure in heterogeneous networks as shown in Fig. 6. Based on the intuitions in Sect 2, we assume that the behavior of each influenced user can be generated in two ways, either depending on his/her own interests or influenced by one of his/her friends. E.g., when a user shares a blog on Renren, he/she may like its content or follow the action of one of his/her friends who also share it. Thus the proposed model consists of the following two parts, and the whole generative process are illustrated in Alg. 1 (Table 1 lists the descriptions of variables).

- **User interest modeling** As shown in the middle part of Fig. 6, each user x is represented as a multinomial distribution over topics ψ, which indicates user interests. We assume that topics of documents are generated based on user interests. Then each word w in documents is generated from one topic z selected from the

```
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 $A_x$;
                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

distribution. The details of the generative process are illustrated in the first iteration of Alg. 1.

- **Influence strength mining** The right part of Fig. 6 illustrates influence strength modeling. The parameter s, which is generated from a Bernoulli distribution with parameter λ, is used to control the influence situation. We assume that when $s = 1$, the behavior is generated based on his/her own interests, while when $s = 0$, the behavior of the user is influenced by one of his/her friends. Then another parameter γ is used to indicate the influence strength from candidate user set A_x to user x, based on which one influencing user y is selected from A_x. At last, a topic is generated from the mixture of topics of a user – the user himself/herself x or one of his/her friends y, based on which the word w is generated. This part corresponds to the second iteration of Alg. 1.

In the above generative process, A_x is the candidate influencing user set w.r.t. x, thus A_x changes with x. Besides, A_x is determined by real applications, which considers both directed and undirected links between users. For example, in Twitter network A_x denotes the users whom a blog is re-tweeted from while in citation networks it denotes the authors of cited papers. In these networks, the links between users are directed. In some other networks, such as Renren and Digg, A_x denotes the friends of user x who also share or dig the same story, and the links are undirected. Thus the proposed model is able to handle both types of cases.

4.2 Model Learning via Gibbs Sampling

We employ Gibbs sampling to estimate the model. 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, $N_{x,z,s}(x, z, 1)$ represents the number of samples of topics z which are supposed to be generated from user x when $s = 1$.

$$p(s_i = 0|\mathbf{s}_{-i}, x_i, z_i, .) \propto$$
$$\frac{N_{x',z'}(y_i,z_i)+N_{y,z,s}(y_i,z_i,0)+\alpha_\theta}{N_{x'}(y_i)+N_{y,s}(y_i,0)+T\cdot\alpha_\theta} \cdot \frac{N_{x,s}(x_i,0)+\alpha_{\lambda_{s_0}}}{N_x(x_i)+\alpha_{\lambda_{s_0}}+\alpha_{\lambda_{s_1}}} \tag{4}$$

$$p(s_i = 1|\mathbf{s}_{-i}, x_i, z_i, .) \propto$$
$$\frac{N_{x,z,s}(x_i,z_i,1)+\alpha_\psi}{N_{x,s}(x_i,1)+T\cdot\alpha_\psi} \cdot \frac{N_{x,s}(x_i,1)+\alpha_{\lambda_{s_1}}}{N_x(x_i)+\alpha_{\lambda_{s_0}}+\alpha_{\lambda_{s_1}}} \tag{5}$$

$$p(y_i|\mathbf{y}_{-i}, s_i = 0, d_i, x_i, z_i, A_x, .) \propto$$
$$\frac{N_{x,y,s}(x_i,y_i,0)+\alpha_\gamma}{N_{x,s}(x_i,0)+|A_x|\cdot\alpha_\gamma} \cdot \frac{N_{x',z'}(y_i,z_i)+N_{y,z,s}(y_i,z_i,0)+\alpha_\theta}{N_{x'}(y_i)+N_{y,s}(y_i,0)+T\cdot\alpha_\theta} \tag{6}$$

$$p(z_i|\mathbf{z}_{-i}, s_i = 0, w_i, .) \propto$$
$$\frac{N_{x',z'}(y_i,z_i)+N_{y,z,s}(y_i,z_i,0)+\alpha_\theta}{N_{x'}(y_i)+N_{y,s}(y_i,0)+T\cdot\alpha_\theta} \cdot \frac{N_{w,z}(w_i,z_i)+N_{w',z'}(w'_i,z'_i)+\alpha_\phi}{N_z(z_i)+N_{z'}(z_i)+W\cdot\alpha_\phi} \tag{7}$$

$$p(z_i|\mathbf{z}_{-i}, s_i = 1, w_i, .) \propto$$
$$\frac{N_{x,z,s}(x_i,z_i,1)+\alpha_\psi}{N_{x,s}(x_i,1)+T\cdot\alpha_\psi} \cdot \frac{N_{w,z}(w_i,z_i)+N_{w',z'}(w'_i,z'_i)+\alpha_\phi}{N_z(z_i)+N_{z'}(z_i)+W\cdot\alpha_\phi} \tag{8}$$

Through the Gibbs sampling process, we obtain the sampled coin s_i, influencing user y_i, and topic z_i for each word. Then the influence strength can be estimated by Eq.(9), which are averaged over the sampling chain after convergence. K denotes the length of the sampling chain.

$$\delta_x(y) = \gamma_x(y) = \frac{1}{K}\sum_{i=1}^{K}\frac{N_{x,y,s}(x,y,0)^i + \alpha_\gamma}{N_{x,s}(x,0)^i + |A| \cdot \alpha_\gamma} \tag{9}$$

The equation is consistent to our assumptions in a statistical way. Take citation networks for example. If author x cites more papers of author y and “copies” more content from y, $N_{x,y,s}(x, y, 0)$ will be larger, and thus the influence from y to x will be stronger. Besides, it is easy to get that $\sum_{y=1}^{|A_x|}\delta_x(y) = 1$, i.e., the sum of influence on user x from all the users obtained in the model equals to 1. And the model does not consider the influence between the nodes which are not connected, i.e., $\delta_x(y) = 0$ when x and y are not connected.

Furthermore, we can estimate the topic-level influence strength. Suppose $\delta_{x,z}(y)$ represents the influence strength from user y to user x on the topic z, which satisfy that $\delta_x(y) = \sum_{z=1}^{T}\delta_{x,z}(y)$. Thus the topic-level influence can be estimated by Eq. (10).

$$\delta_{x,z}(y) = \frac{1}{K}\sum_{i=1}^{K}\frac{N_{x,y,z,s}(x,y,z,0)^i + \frac{1}{T}\cdot\alpha_\gamma}{N_{x,s}(x,0)^i + |A| \cdot \alpha_\gamma} \tag{10}$$

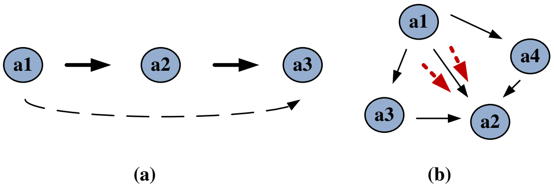


Fig. 7 Influence propagation

5 Influence Propagation and Aggregation

The above probabilistic model only discovers direct influence, but does not consider indirect influence. In reality, like information or virus, influence also propagates over networks, which produces different types of indirect influence. Take Fig. 7(a) for example. If $a1$ influences $a2$ and $a2$ influences $a3$, then $a1$ will influence $a3$ potentially, i.e., two-degree of influence. Fig. 7(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 $a2$ should be enhanced. The observations in Sect 3 have demonstrated the existence of indirect influence. Based on these observations, we study atomic and iterative influence propagation process over social networks in this section, via which indirect influence can be obtained from direct influence and global influence strength can be estimated.

5.1 Atomic Influence Propagation

As shown in Fig. 7, we observe there are two basic processes for influence propagation.

- **Concatenation** The indirect influence from $a1$ to $a3$ in Fig. 7(a) can be modeled as a concatenate result of the direct influence from $a1$ to $a2$ and the direct influence from $a2$ to $a3$.
- **Aggregation** The enhancement of the influence from $a1$ to $a2$ in Fig. 7(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:

$$\delta_v(u) = \Diamond(\forall w \in Nb(v) : \delta_v(w) \circ \delta_w(u)) \tag{11}$$

where $Nb(v)$ is the set of neighbors of node v. $\circ$ is the concatenation function and $\Diamond$ is the aggregation function.

In real processes, multiplication operation or minimum value is often used as concatenation function while addition operation or maximum value is used as the aggregation function. In particular, if we employ multiplication and addition operations to replace the concatenation and aggregation function in Eq. (11) respectively, then the atomic influence propagation can be instantiated as:

$$\delta_v(u) = \sum_{w \in Nb(v)} \delta_v(w) \cdot \delta_w(u) \tag{12}$$

Suppose Δ_v represents the vector of the influence strength from all the nodes in the network on node v, i.e., $\Delta_v = (\delta_v(u_1), \delta_v(u_2), ..., \delta_v(u_n))$. And we use superscript to denote the propagation step, i.e., $\mathit{\Delta}^0$ denotes the initial influence strength and $\mathit{\Delta}^1$ denotes the influence strength after the atomic propagation. Then the atomic influence propagation can be represented as the matrix multiplication.

$$\mathit{\Delta}_v^1 = \mathit{\Delta}_v^0 \cdot M \tag{13}$$

where M is the transition matrix and $M = (\mathit{\Delta}_{v_1}; \mathit{\Delta}_{v_2}; ...; \mathit{\Delta}_{v_n})$, i.e., each element in the transition matrix $M(v, u) = \delta_v(u)$.

5.2 Iterative Influence Propagation

In reality, the indirect influence along longer paths, e.g., three-degree or four-degree influence, also have effect on the nodes in a network. In another word, influence can propagate iteratively to collect the contribute of influence on longer paths. Thus the atomic influence propagation should be performed iteratively to propagate direct influence over the entire network. Thus the influence strength on k-length paths can be calculated by k steps of atomic propagations.

If the atomic propagation is defined as Eq. (13), the influence strength vector after k-step atomic propagation can be calculated by the matrix powering operation.

$$\mathit{\Delta}_v^k = \mathit{\Delta}_v^{k-1} \cdot M = \mathit{\Delta}_v^0 \cdot M^k \tag{14}$$

where $M^k = M^{k-1} \cdot M$. $\mathit{\Delta}^k$ denotes the influence strength vector on k-length paths.

Formally, we define the *iterative influence propagation* as following:

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

Suppose the final influence strength between two nodes after k-step iterative propagation is denoted as $\mathit{\Delta}^{f_k}$. Based on the above definition, it should collect all the contributes of the influence strength on paths with the length ranging from 0 to k, i.e.,

$$\mathit{\Delta}^{f_k} = \Diamond(\forall i \in \{0, 1, 2, ..., k\} : \mathit{\Delta}^i) \tag{15}$$

If addition operation is used as the aggregation function, $\mathit{\Delta}^{f_k}$ can be inferred from the sequences of propagation via a weighted linear combination [18]:

$$\mathit{\Delta}^{f_k} = \sum_{i=0}^{k} \beta_i \mathit{\Delta}^i \tag{16}$$

β_i denotes the weight for the influence strength on i-length paths, i.e., $\mathit{\Delta}^i$.

Intuitively, the effect of the influence on shorter paths should be larger than the one on longer paths as the iterative propagation process brings in more outside information. In Sect 3 we have also found that indirect strength decreases with the increase of propagation length generally. Therefore, β_i should decrease with the increase of iteration step i. Different strategies can be employed to assign the weights. In the next section, we will study two kinds of strategies, which are conservative propagation and non-conservative propagation respectively.

5.3 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 $\Lambda(u)$ should be related to its influence on all the other nodes in the network. 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 of its influence on the other nodes, specifically,

$$\Lambda(u) = \sum_{v} \delta_v(u) \tag{17}$$

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

6 Conservative and Non-conservative Propagation

In this section, we describe two types of diffusion process - conservative and non-conservative diffusion process, based on which we propose two kinds of methods to propagate influence over the network and to obtain indirect influence strength.

First, we formally define a propagation process over a network.

Definition 4 [Propagation Process] A propagation process over a network G is defined as a function $\{F_t(w) : (R^+ \cup \{0\})^{|V|} \to (R^+ \cup \{0\})^{|V|}\}$, where V is the set of nodes in G. w is a V-dimensional vector, which represents a weight distribution over the nodes in the network. t denotes propagation step.

Therefore, in a propagation process, each node in a network is first initialized with some mass, which is denoted as the weight of the node. Then via each step of propagation, some nodes transfer a part of the weights to their neighbors. Thus through a t-step propagation process, a $|V|$-dimensional non-negative vector is mapped to another $|V|$-dimensional non-negative vector. In particular, when $t = 1$, the propagation is atomic propagation.

6.1 Conservative Propagation

Definition 5 [Conservative Propagation] For a propagation process F, if $\forall w \in (R^+ \cup \{0\})^{|V|}$,$||w||_1 = ||F(w)||_1$,i.e., it preserves the sum of the entries, we call the propagation process conservative propagation.

Therefore, conservative propagation simply redistributes the weights among the nodes in the network and keeps the sum of weights constant. There are many conservative propagation examples in the real world. Take the circulation of money for example. At each step of propagation, some nodes transfer a fraction of their money to their neighbors. But the total money in the network does not change. Traffic transportation and energy cycle are also conservative propagations as the total traffic or energy does not change with the propagation process.

Mathematically, random walk is a canonical example of conservative propagation. In a random walk, a particle starts to locate on a node. Then at each step, the particle selects one of the out-neighbors at random and moves to that node. A weight vector is used to represent the probability with which the particle can be found on each node. Thus the sum of the weights equals to one. And after iterative propagations, the probabilities of finding the particle on the nodes change, but the sum remains to be one all the time.

PageRank is a classical random walk model, which is represented as:

$$pr(w) = (1-\beta) \cdot w_0 + \beta \cdot pr(w) \cdot M \tag{18}$$

M is a transition matrix, in which the element $M(a, b)$ denotes the transfer probability from node a to b. β is a damping factor which is used to ensure the stationary probability distribution of the propagation. $1-\beta$ is the restart probability, which gives the probability distribution when the random walk transition restarts. w_0 is the initial weight distribution, which is usually set to be uniform vector. Personalized PageRank [22] extends the model by setting w_0 to be a non-uniform starting vector.

6.1.1 Conservative Influence Propagation

We model the *conservative influence propagation* as a personalized PageRank in a network as Eq. (19).

$$\Delta^{f_t} = (1-\beta) \cdot \Delta^0 + \beta \cdot \Delta^{f_{t-1}} \cdot M \tag{19}$$

The propagation probability matrix M can be set in various ways. If we use direct influence strength to define the propagation probability, i.e., $M(v, u) = \delta_v^0(u)$, then $\sum_u M(v, u) = 1$. It is easy to prove that the sum of influence strength from all the nodes on one node v remains to be one after influence propagation, i.e., $||\Delta_v^{f_t}||_1 = 1$. Thus Eq. (19) defines a conservative influence propagation.

This conservative influence propagation provides a strategy for the combination process in the iterative propagation. From Eq. (19), it is easy to get that

$$\Delta^{f_t} = (1-\beta) \cdot \Delta^0 \cdot \sum_{i=0}^{t-1} (\beta^i \cdot M^i) + \Delta^0 \cdot \beta^t \cdot M^t \tag{20}$$

As the influence vector on t-length path is $\Delta^t = \Delta^0 \cdot M^t$,

$$\Delta^{f_t} = (1-\beta) \cdot \sum_{i=0}^{t-1} (\beta^i \cdot \Delta^i) + \beta^t \cdot \Delta^t \tag{21}$$

Thus the conservative influence propagation defined in Eq. (19) assigns different weights to the influences on different-length paths.

β is a damping factor, i.e., $0 \le \beta \le 1$. Thus when t increases, β^t decreases, which makes the effect of influence on longer paths smaller.

6.2 Non-conservative Propagation

Definition 6 [Non-conservative Propagation] For a propagation process F, if $\exists w \in (R^{+} \cup \{0\})^{|V|}, ||w||_1 \neq ||F(w)||_1$, we call the propagation process non-conservative propagation.

Compared with conservative propagation, non-conservative propagation does not keep the sum of weights constant. There are also many non-conservative propagation examples in the real world. Take the spread of a virus for example. Suppose a virus is propagating over the social network. When one infected node infects its neighbors, it is still infected. Thus the total number of infected nodes is increased with time. Therefore, the spread of virus is a kind of non-conservative process. Besides, information diffusion and oral advertising are also non-conservative propagations as the number of nodes which accept the information or advertisement increases with propagation step.

Alpha-Centrality, which was introduced by Bonacich [3,4], can be used to model non-conservative propagation. The Alpha-Centrality vector $c(w)$ is defined as the solution of the following equation:

$$c^{t}(w) = w_0 + \beta \cdot c^{t-1}(w) \cdot M \tag{22}$$

β is a damping factor. The starting vector w_0 is usually set to be in-degree centrality. And M uses the adjacency matrix.

When $\beta < \frac{1}{|\lambda_1|}$ (where λ_1 is the largest eigenvalue of M), we can get that $c(w) = w_0 \cdot (I - \beta M)^{-1}$, where I is the identity matrix of size n. Using the identity

$$\sum_{t=1}^{\infty} (\beta^{t} \cdot M^{t}) = (I - \beta \cdot M)^{-1} - I \tag{23}$$

we can get

$$c(w) = w_0 \cdot (I - \beta \cdot M)^{-1} = w_0 \cdot \sum_{t=0}^{\infty} (\beta^{t} \cdot M^{t}) \tag{24}$$

Besides Alpha-Centrality, Katz score [23], SenderRank [27] and eigenvector centrality [2] are other examples of non-conservative mathematical metrics.

6.2.1 Non-conservative Influence Propagation

We model the *non-conservative influence propagation* process in the form of Alpha-Centrality as Eq. (25).

$$\Delta^{f_t} = \Delta^{0} + \beta \cdot \Delta^{f_{t-1}} \cdot M \tag{25}$$

For Alpha-Centrality, M is usually set to be adjacency matrix. Here we also use direct influence strength to define the transition matrix M, i.e., $M(v,u) = \delta_v(u)$. It is easy to prove that the sum of influence strength from all the nodes on node v increases with non-conservative propagation step, i.e., $||\Delta_v^{f_t}||_1 > 1$. Thus Eq. (25) defines a non-conservative propagation for local influence.

This non-conservative influence propagation provides another strategy for the combination process in the iterative propagation. From Eq. (25), we can get

$$\Delta^{f_t} = \Delta^0 \cdot \sum_{i=0}^{t} (\beta^i \cdot M^i) = \sum_{i=0}^{t} \beta^i \cdot \Delta^i \tag{26}$$

Thus it assigns different weights to the influence strength on different-length paths. However, the weight assignment strategy is different from conservative propagation referring to Eq. (21).

6.3 Comparison and Explanation

Both conservative and non-conservative influence propagations collect all the contributes of direct and indirect influence on the propagating paths. And both of them define a weight assignment strategy to distinguish the effect of influence on different-length paths. The major difference between these two types of models is that conservative propagation keeps the sum of influence in the whole network constant while non-conservative propagation does not.

Intuitively, indirect influence strength on shorter-paths should be more reliable since there have been fewer propagation steps. The more iteration steps, the more outside information will be brought. Thus, both conservative and non-conservative propagations utilize a damping factor β to penalize larger t-step propagations. As $0 \leq \beta \leq 1$, when t increases, β^t decreases greatly, which makes the effect of influence on $(t+1)$-length paths very small. In another word, we do not need to iterate influence propagation for many times to obtain the final indirect influence, i.e., t can be set as a small number. Besides, when $\beta = 0$, both conservative and non-conservative propagations only utilize direct influence and ignore the effect of indirect influence.

7 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 Renren, Twitter and Digg. The user behavior is defined as one time connection between a user and a document. 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 [33] 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:

$$p(d|u) = \frac{1}{\sum_v \delta_u(v)} \sum_{v \in Nb(u)} \delta_u(v) p(d|v) \tag{27}$$

where $Nb(u)$ denotes the friends of u.

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.(28) for comparison, where the similarity between users $s(v,u)$ is calculated as the Euclidean distance of user distributions over topics.

$$p(d|u) = \frac{1}{\sum_v s(v,u)} \sum_{v \in Nb(u)} s(v,u) p(d|v) \tag{28}$$

We will test the user behavior prediction performance based on the above three methods in the following experiments and demonstrate the effect of influence strength obtained from both conservative and non-conservative influence propagations for social network applications.

8 Experiments

In this section, we present various experiments to evaluate the efficiency and effectiveness of the proposed approach. The data sets and codes are publicly available[5].

8.1 Experimental Setup

Data Sets We prepare four different types of heterogeneous networks for our experiments, including Renren, Twitter, Digg and citation networks. Renren is a very popular FaceBook-style social website in China, on which users (especially the undergraduate and graduate students) connect with their classmates or friends and share interesting web content. 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 collect user relationship and document content from these websites.

- **Renren social network** The data contains 5000 users and the web content shared by these users in one month which includes about 10000 documents and 30000 words.
- **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 contains about 1 million stories related to 10000 users and 30000 keywords, in which we aim to mine 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 four data sets. The algorithms are implemented in C++ and run on an Intel Core 2 T7200 and a processor with 2GB DDR2 RAM. The parameters of the model 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 in citation networks, we manually label the citation data and then test the influence prediction performance in it. We compare the results

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

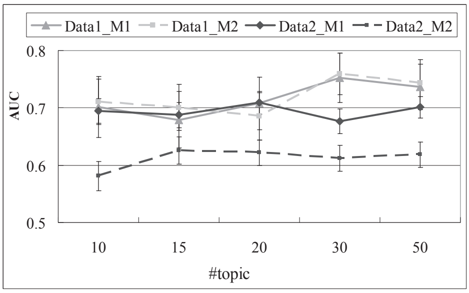


Fig. 8 Influence prediction performance comparison

of our approach with previous work [9] to demonstrate our model's better performance in terms of influence prediction.

User behavior prediction We use 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 Sect. 7. 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 how effectively our method can identify topic-level influence. In particular, we study the global influence of authors in citation networks to demonstrate semantic meaning of topic-level influence. And we compare the results with that of previous work [44] which can also be used to mine topic-level influence to demonstrate the better performance of our approach.

8.2 Influence Prediction

In work [9], researchers evaluated the document influence prediction performance in a manually labeled data set. We use the same data from the authors and also test the influence prediction performance of our model in 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 label document influence strength in a larger data set with about 1000 documents. We classify the influence strength into three levels: 1, 2, 3. Similar to [9], 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. 8 shows the comparative results in these two data sets, where Data1 is the small data set obtained from authors of [9] while Data2 is our larger labeled data set. $M1$ and $M2$ are used to denote our model and the model in [9] respectively. And we use the real and dash lines to distinguish the results of these two models in the figure. We calculate all the AUC values with the number of topics changing from 10 to 50. Thus this figure demonstrates that in the small data set our model can achieve as good prediction performance as the work in [9] while in the larger data set, our prediction performance is better than theirs.

Furthermore, we compare the influence prediction performance before and after influence propagation in our labeled data set. The results prove that the influence pre-

Table 2 Conservative and non-conservative influence propagation effect on user behavior prediction

method p	baseline	DI	influence propagation						
			steps	$\beta = 0.3$		$\beta = 0.5$		$\beta = 0.8$	
				CIP	NCIP	CIP	NCIP	CIP	NCIP
Avg	0.101	0.160	$t = 1$	0.168	0.168	0.168	0.168	0.172	0.168
			$t = 5$	0.168	0.168	0.170	0.170	0.180	0.175
			$t = 10$	0.168	0.168	0.170	0.170	0.180	0.178
Var	0.011	0.048	$t = 1$	0.044	0.045	0.041	0.044	0.039	0.042
			$t = 5$	0.044	0.045	0.042	0.043	0.041	0.041
			$t = 10$	0.044	0.045	0.043	0.042	0.041	0.041

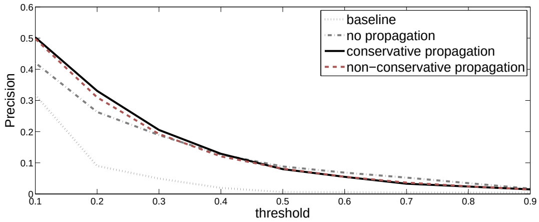


Fig. 9 User behavior prediction precision on Renren network

diction performance is enhanced after influence propagation (AUC values are enhanced from 0.69 to o.76). Moreover, the influence prediction performance is robust to the parameters t and β. In particular, when t changes, the performance changes little, which is consistent to the observation in Fig. 4. It means that influence does propagate over the network, but the effect of propagation is reduced with propagation step.

8.3 User Behavior Prediction

We employ our model to discover the concrete influence strength between the 5000 users in Renren social networks. Then we apply the learned influence strength to user behavior prediction as described in Sect. 7. In particular, the parameters which are the damping factor β and iteration step t for both conservative and non-conservative influence propagations are varied to test the effect of influence propagation process. About 36000 tuples in Renren data set are used as testing samples. Each tuple represents that a user shares a web document, whose probability is estimated as Eq. (27).

The average and variance values of the predicted probabilities for all the samples are calculated and shown in Table 2, where DI denotes direct influence, CIP and NCIP denote conservative and non-conservative influence propagations respectively. The results demonstrate that using influence, especially the propagated influence, can greatly improve the predicted probabilities. But the parameters t and β as well as the propagation mechanism do not affect the 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. 9 shows four curves of prediction precision changing with the threshold in Renren data set, which indicate the performance of baseline, using direct influence without influence propagation, conservative and non-conservative influence propagations with parameter $\beta = 0.8, t = 5$ respectively. The results demonstrate that influence-based behavior prediction approach outperforms the baseline. Thus it proves that the influence obtained from our

Table 3 Behavior prediction probability

Digg Social Network				
method p	baseline	similarity	DI	NCIF
average	0.112	0.121	0.366	0.405
variance	0.006	0.008	0.075	0.048

Twitter Social Network				
method p	baseline	similarity	DI	NCIF
average	0.215	0.222	0.319	0.310
variance	0.078	0.089	0.129	0.136

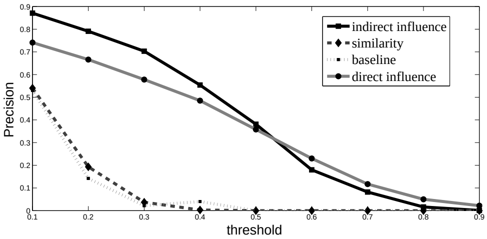


Fig. 10 User behavior prediction precision on Digg network

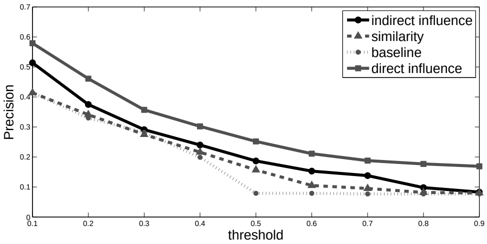


Fig. 11 User behavior prediction precision on Twitter network

model benefits the user behavior prediction greatly. Moreover both conservative and non-conservative influence propagations improve the prediction precision and almost achieve the same performance.

Besides, we apply our model to the application of user behavior predication in Twitter and Digg social networks. In this experiment, we employ non-conservative influence propagation with $t = 5, \beta = 0.8$ to obtain indirect influence. We randomly select about 3000 tuples from Digg and Twitter data sets as testing samples and estimate their probabilities. Table 3 shows the average and variance values of the predicted probabilities for all the samples. The prediction precision curves for these two data sets are shown in Fig. 10 and 11 respectively. The results demonstrate that influence-based behavior prediction approach outperforms the baseline and the similarity-based method. In particular, it shows that influence propagation process enhances the user behavior prediction performance in Digg social network but it takes little effect in Twitter social network. Furthermore, comparing these two figures, we can get that the effect of influence in Digg social network is larger than that in Twitter social network. The conclusion is consistent to the observation in Fig. 4.

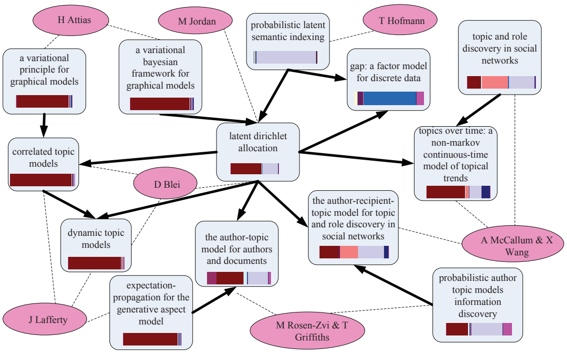

H Attias
M Jordan
probabilistic latent semantic indexing
T Hofmann
topic and role discovery in social networks
a variational principle for graphical models
a variational bayesian framework for graphical models
gap: a factor model for discrete data
latent dirichlet allocation
correlated topic models
D Blei
topics over time: a non-markov continuous-time model of topical trends
the author-recipient-topic model for topic and role discovery in social networks
dynamic topic models
the author-topic model for authors and documents
A McCallum & X Wang
probabilistic author topic models information discovery
expectation-propagation for the generative aspect model
J Lafferty
M Rosen-Zvi & T Griffiths

Fig. 12 Document influence case study

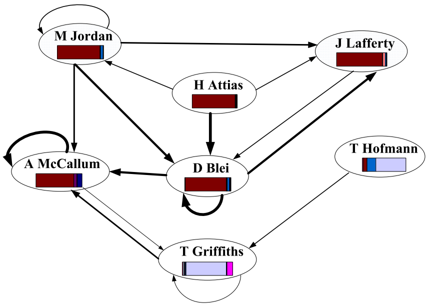


Fig. 13 Author influence case study

8.4 Topic-level Influence Case Study

Topic-level influence graph

We apply our model to the citation network which we crawled from the Internet and set the number of topics to be 10 empirically. Fig. 12 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 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. 12 also shows the connections between authors and documents by dash lines. The influences between these authors are visualized in Fig. 13. 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. 12 shows,

Table 4 Author ranking on "statistical topic models"

Direct Influence	Indirect Influence		Pagerank
	$t = 1$	$t = 5$	
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

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 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" (t 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 [35, 19] over 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 in more general areas, we select 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. 14 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

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

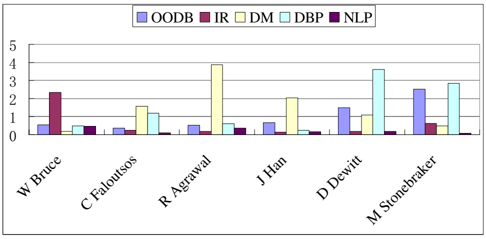


Fig. 14 Estimated global influence distribution on topics

the two topics related to database, D DeWitt is also very influential on topic “DM”. The reason should be that the area “DM” develops from database. Furthermore, Table 5 shows the maximal, minimal and average values of the estimated global influence in the whole network w.r.t. each topic, which demonstrates that these authors almost have the largest values in their domains. Thus it proves the validity of the way of global influence estimation.

Topic-level influence comparison

Work [44] 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 [44] ($M3$). 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 [44] 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. 15 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.

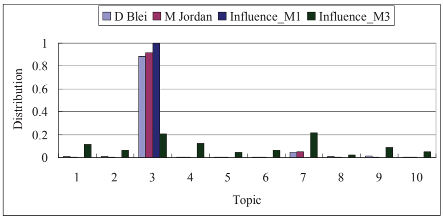


Fig. 15 Topic distributions of authors and influence

9 Related Work

9.1 Heterogeneous Network Analysis

With the information explosion in the real world, how to fuse and utilize heterogeneous source becomes an important research problem in many areas. E.g., Ye et al. [49] fused heterogeneous data sources to study the alzheimer's disease. In particular, various user generated content is spreading over social networks, which makes the integration of textual content and social network structure and forms types of heterogeneous networks. Many researchers study data mining problems in heterogenous networks. For example, Sun et al. [40, 41] investigated how to cluster different types of nodes jointly in heterogeneous networks based on the analysis of heterogeneous link characteristics. Furthermore, combining textual information with link structure becomes a feasible means to improve the performance of social network analysis. For example, Yang [48] proposed a discriminative approach to combine link and content to detect communities in networks. Chang [5] presented a probabilistic topic model to infer descriptions of entities in text corpora and their relationships. Zheleva [50] analyzed the co-evolution of social and affiliation networks. Tang [45] addressed the relational learning problem of social media based on their extracted latent social dimensions. Nallapati [34] proposed two topic models to jointly model text and citation relationships. However, the problem how to fully utilize heterogeneous information to mine social influence has not been well addressed yet.

9.2 Social Influence Analysis

Researchers have recognized that influence is a potential factor which affects user behavior and social network dynamics. Considerable work has been conducted to validate the existence of influence and study its effect from the global view of the whole network. For example, King et al. [26] analyzed influence factor among paper citation networks. 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 [39] 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. [7] further investigated the correlation between social similarity and influence. Cui et al.

[8] proposed a Hybrid Factor Non-Negative Matrix Factorization (HF-NMF) approach for item-level social influence modeling.

Besides the global effect of influence, many efforts have been made to estimate the concrete influence strength between individual nodes. Dietz et al. [9] proposed a citation influence topic model to discover the influential strength between papers. Tang et al. [44] introduced the problem of topic-based social influence analysis. And they proposed a Topical Affinity Propagation (TAP) approach to describe the problem via using a graphical probabilistic model. However, these works neither consider heterogeneous information nor learn topics and influence strength jointly. Tan et al. [42] studied how to track and predict users' action according to a learning model. However, they did not consider the topic-level influence and the indirect influence. Gerrish et al. designed a topic model to discover scholarly impact [14]. However, this paper aims at discovering the influence of documents instead of social influence. Furthermore, this topic model is based on the content changes and does not use the network structure.

This article is an extension of our previous work [32] and has new contributions on the following aspects:

- For the problem definition, this paper studies the influence propagation modeling in social networks. We newly define two types of diffusion, i.e., conservative and non-conservative propagation models for influence propagation in social networks.
- On technical part, we propose personalized PageRank and Alpha-Centrality models for conservative and non-conservative influence propagations. Both of them can be used to derive indirect influence strength. And we compare them and discuss the different parameters of the models.
- In experimental sections, we add a new data set, Renren, which is a very popular Facebook-styple website in China, and we analyze the influence effect in this dataset.

Besides, we conduct series of analysis on the relationship between influence and four kinds of social factors in Sect. 3. These observations are intuitive supportive for our social influence modeling. All these parts are our new contributions of the current article.

9.3 Propagation Process Modeling in Social Networks

Social network dynamics analysis is an important problem that attracts many researchers' interests [12]. And many propagation models in social networks have been proposed. For example, Random walk models [35] assume a particle's moving process in a network so as to estimate its emerging probability on each node. Various centrality metrics make implicit assumptions of propagation to study the properties of networks, such as degree, closeness, betweenness and etc. More recently, researchers study other types of propagation process, including information diffusion [28, 17], viral marketing [21], money exchange, e-mail forwarding [31], etc.

Like epidemics or information, influence also propagates over social networks and affects social network dynamics. For example, Scripps et al. [38] investigated how different pre-processing decisions and network forces such as selection and influence affect the modeling of dynamic networks. Rodriguez [37] developed a method to trace paths of diffusion and influence through networks so as to infer the networks over which contagions propagate. Furthermore, some researchers investigated the problem how to maximize influence on a person network for real applications, e.g., viral marketing

[10, 36, 25, 15, 6]. However, these works do not explore the effect of different types of propagation process on social influence mining.

10 Conclusions and Future Work

In this paper, we study a novel problem of mining topic-level influence in 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 different types of influence propagation methods to mine indirect and global influence. In the probabilistic model, we combine the textual content and heterogeneous link information into a unified generative process. Influence propagation methods further propagate influence along the links in the entire network. We have done extensive experiments in 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 in 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 interesting topic is to study the influence learning problem across heterogeneous networks [43]. Users' behaviors are distributed in different networks. It would be intriguing to merge the information from different networks and leverage the correlation between them to better the influence learning performance.

Acknowledgements The work was supported in part by the National Natural Science Foundation of China under grants 61103065, 61073073, 61035004 and 60933013, and by the U.S. National Science Foundation under grant IIS-09-05215 and the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA).

References

1. A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD'08*, pages 7–15, 2008.
2. P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972.
3. P. Bonacich. Power and centrality: A family of measures. *The American Journal of Sociology*, 92(5):1170–1182, 1987.
4. P. Bonacich and P. Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191–201, 2001.
5. 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.
6. W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD '10*, 2010.
7. 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.
8. P. Cui, F. Wang, S. Liu, M. Ou, and S. Yang. Who should share what? item-level social influence prediction for users and posts ranking. In *SIGIR'11*, 2011.
9. L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *ICML'07*, pages 233–240, 2007.
10. P. Domingos and M. Richardson. Mining the network value of customers. In *KDD' 01*, pages 57–66, 2001.

11. A. Doucet, N. de Freitas, K. Murphy, and S. Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In *UAI'00*, pages 176–183, 2000.
12. Y. Dourisboure, F. Geraci, and M. Pellegrini. Extraction and classification of dense communities in the web. In *WWW '07*, pages 461–470, 2007.
13. 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.
14. S. Gerrish and D. M. Blei. A language-based approach to measuring scholarly impact. In *ICML'10*, pages 375–382, 2010.
15. A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *WSDM'10*, pages 207–217, 2010.
16. M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.
17. D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04*, pages 491–501, 2004.
18. R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW '04*, pages 403–412, 2004.
19. T. H. Haveliwala. Topic-sensitive pagerank. In *WWW'02*, pages 517–526, 2002.
20. J. E. Hopcroft, T. Lou, and J. Tang. Who will follow you back? reciprocal relationship prediction. In *CIKM'11*, 2011.
21. J. L. Iribarren and E. Moro. Impact of human activity patterns on the dynamics of information diffusion. *Phys. Rev. Lett.*, 103(3):038702, Jul 2009.
22. G. Jeh and J. Widom. Scaling personalized web search. In *WWW'02*, pages 271–279, 2002.
23. L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953.
24. H. Kelman. Compliance, identification, and internalization: Three processes of attitude change. *Journal of Conflict Resolution*, (1):51–60, 1958.
25. D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03*, pages 137–146, 2003.
26. J. King. A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, 13(5), 1987.
27. C. Kiss and M. Bichler. Identification of influencers- measuring influence in customer networks. *Decision Support Systems*, 46(1):233–253, 2008.
28. G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. In *KDD '08*, pages 435–443, 2008.
29. 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.
30. T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW '10*, pages 601–610, 2010.
31. D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 2008.
32. L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining Topic-Level Influence in Heterogeneous Networks. In *CIKM'10*, pages 199–208, October 2010.
33. S. Macskassy and F. Provost. A simple relational classifier. In *Workshop on Multi-Relational Data Mining in conjunction with KDD'03*, 2003.
34. 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.
35. 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.
36. M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD '02*, pages 61–70, 2002.
37. M. G. Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *KDD '10*, 2010.
38. 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.
39. 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.

40. 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*, March 2009.
41. 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.
42. C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In *KDD'10*, pages 1049–1058, 2010.
43. J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *WSDM'12*, 2012.
44. J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD '09*, pages 807–816, 2009.
45. L. Tang and H. Liu. Relational learning via latent social dimensions. In *KDD '09*, pages 817–826, 2009.
46. J. Whitfield. The secret of happiness: grinning on the internet. In *Nature*, 2008.
47. R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *WWW '10*, pages 981–990, 2010.
48. 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.
49. 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.
50. E. Zheleva, H. Sharara, and L. Getoor. Co-evolution of social and affiliation networks. In *KDD '09*, pages 1007–1016, 2009.

A Gibbs Sampling Derivation

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 [11].

$$\begin{aligned}
& p(\mathbf{w}, \mathbf{w}', \mathbf{z}, \mathbf{z}', \mathbf{s}, \mathbf{y} | \boldsymbol{\alpha}_\phi, \boldsymbol{\alpha}_\theta, \boldsymbol{\alpha}_\psi, \boldsymbol{\alpha}_\lambda, \boldsymbol{\alpha}_\gamma) \\
\propto & \int p(\mathbf{s}|\boldsymbol{\lambda}, \mathbf{x}) p(\boldsymbol{\lambda}|\boldsymbol{\alpha}_\lambda) d\boldsymbol{\lambda} \int p(\mathbf{z}, \mathbf{z}' | \mathbf{y}, \mathbf{s}, \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}_\theta) p(\boldsymbol{\psi}|\boldsymbol{\alpha}_\psi) d\psi\theta \\
& \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\gamma}, A) p(\boldsymbol{\gamma}|\boldsymbol{\alpha}_\gamma) d\boldsymbol{\gamma} \int p(\mathbf{w}, \mathbf{w}' | \mathbf{z}, \mathbf{z}', \boldsymbol{\phi}) p(\boldsymbol{\phi}|\boldsymbol{\alpha}_\phi) d\boldsymbol{\phi}
\end{aligned} \tag{29}$$

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 $\mathbf{s}_{-i}$) and canceling factors that do not depend on $\mathbf{s}_{-i}$.

$$\begin{aligned}
& p(s_i = 0 | \mathbf{s}_{-i}, x_i, z_i, .) \\
= & \frac{p(\mathbf{w}, \mathbf{w}', \mathbf{z}, \mathbf{z}', s_i, \mathbf{y} | \boldsymbol{\alpha}_\phi, \boldsymbol{\alpha}_\theta, \boldsymbol{\alpha}_\psi, \boldsymbol{\alpha}_\lambda, \boldsymbol{\alpha}_\gamma)}{p(\mathbf{w}, \mathbf{w}', \mathbf{z}, \mathbf{z}', \mathbf{s}_{-i}, \mathbf{y} | \boldsymbol{\alpha}_\phi, \boldsymbol{\alpha}_\theta, \boldsymbol{\alpha}_\psi, \boldsymbol{\alpha}_\lambda, \boldsymbol{\alpha}_\gamma)} \\
= & \frac{\int p(s_i|\boldsymbol{\lambda}, \mathbf{x}) p(\boldsymbol{\lambda}|\boldsymbol{\alpha}_\lambda) d\boldsymbol{\lambda}}{\int p(\mathbf{s}_{-i}|\boldsymbol{\lambda}, \mathbf{x}) p(\boldsymbol{\lambda}|\boldsymbol{\alpha}_\lambda) d\boldsymbol{\lambda}} \cdot \frac{\int p(\mathbf{z}, \mathbf{z}' | \mathbf{y}, s_i, \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}_\theta) p(\boldsymbol{\psi}|\boldsymbol{\alpha}_\psi) d\psi\theta}{\int p(\mathbf{z}, \mathbf{z}' | \mathbf{y}, \mathbf{s}_{-i}, \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}_\theta) p(\boldsymbol{\psi}|\boldsymbol{\alpha}_\psi) d\psi\theta}
\end{aligned} \tag{30}$$

We derive the first fraction of Eq. (30) (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 $p(s_i|\boldsymbol{\lambda}, \mathbf{x}) = \prod_i \alpha_{\lambda_{s_0}}^{N_{x,s}(x_i,0)} \cdot \alpha_{\lambda_{s_1}}^{N_{x,s}(x_i,1)}$, 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. (30) has a closed form solution. Thus we can get that when $s_i = 0$, the first fraction can be derived as below

$$\frac{\int p(s|\boldsymbol{\lambda}, \mathbf{x})p(\boldsymbol{\lambda}|\boldsymbol{\alpha}_\lambda)d\boldsymbol{\lambda}}{\int p(\mathbf{s}_{-i}|\boldsymbol{\lambda}, \mathbf{x})p(\boldsymbol{\lambda}|\boldsymbol{\alpha}_\lambda)d\boldsymbol{\lambda}} = \frac{N_{x,s}(x_i, 0) + \alpha_{\lambda_{s_0}}}{N_x(x_i) + \alpha_{\lambda_{s_0}} + \alpha_{\lambda_{s_1}}} \tag{31}$$

Deriving $\frac{\int p(\mathbf{z},\mathbf{z}'|\mathbf{y},s_i,\mathbf{x},\boldsymbol{\theta},\boldsymbol{\psi})p(\boldsymbol{\theta}|\boldsymbol{\alpha}_\theta)p(\boldsymbol{\psi}|\boldsymbol{\alpha}_\psi)d\psi\theta}{\int p(\mathbf{z},\mathbf{z}'|\mathbf{y},\mathbf{s}_{-i},\mathbf{x},\boldsymbol{\theta},\boldsymbol{\psi})p(\boldsymbol{\theta}|\boldsymbol{\alpha}_\theta)p(\boldsymbol{\psi}|\boldsymbol{\alpha}_\psi)d\psi\theta}$ analogously, we can get:

$$p(s_i = 0|\mathbf{s}_{-i}, x_i, z_i, .) = \frac{N_{x,s}(x_i, 0) + \alpha_{\lambda_{s_0}}}{N_x(x_i) + \alpha_{\lambda_{s_0}} + \alpha_{\lambda_{s_1}}} \cdot \frac{N_{x',z'}(y_i, z_i) + N_{y,z,s}(y_i, z_i, 0) + \alpha_\theta}{N_{x'}(y_i) + N_{y,s}(y_i, 0) + T \cdot \alpha_\theta} \tag{32}$$