Recommending Users and Communities in Social Media

LEI LI, School of Computer Science & Technology, Nanjing University of Posts and Telecommunications (NJUPT) & School of Computing and Information Sciences, Florida International University
WEI PENG, SAURABH KATARIA, and TONG SUN, Xerox Corporation
TAO LI, School of Computer Science & Technology, Nanjing University of Posts and Telecommunications (NJUPT) & School of Computing and Information Sciences, Florida International University

Social media has become increasingly prevalent in the last few years, not only enabling people to connect with each other by social links, but also providing platforms for people to share information and interact over diverse topics. Rich user-generated information, for example, users' relationships and daily posts, are often available in most social media service websites. Given such information, a challenging problem is to provide reasonable user and community recommendation for a target user, and consequently, help the target user engage in the daily discussions and activities with his/her friends or like-minded people. In this article, we propose a unified framework of recommending users and communities that utilizes the information in social media. Given a user's profile or a set of keywords as input, our framework is capable of recommending influential users and topic-cohesive interactive communities that are most relevant to the given user or keywords. With the proposed framework, users can find other individuals or communities sharing similar interests, and then have more interaction with these users or within the communities. We present a generative topic model to discover user-oriented and community-oriented topics simultaneously, which enables us to capture the exact topical interests of users, as well as the focuses of communities. Extensive experimental evaluation and case studies on a dataset collected from Twitter demonstrate the effectiveness of our proposed framework compared with other probabilistic-topic-model-based recommendation methods.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Information Filtering; H.3.5 [**Online Information Services**]: Web-Based Services

General Terms: Algorithms, Design, Experimentation

Additional Key Words and Phrases: User recommendation, community recommendation, social media, user-community-topic model

ACM Reference Format:
Lei Li, Wei Peng, Saurabh Kataria, Tong Sun, and Tao Li. 2015. Recommending users and communities on social media. ACM Trans. Knowl. Discov. Data 10, 2, Article 17 (October 2015), 27 pages.
DOI: http://dx.doi.org/10.1145/2757282

The work is supported in part by a Xerox University Affair Committee (UAC) Award and by US National Science Foundation under grants DBI-0850203, CCF-0939179, HRD-0833093, CNS-1126619, IIS-1213026, and CNS-1461926.

Authors' addresses: L. Li and T. Li are with School of Computer Science & Technology, Nanjing University of Posts and Telecommunications (NJUPT), Nanjing 210003, China; and School of Computing and Information Sciences, Florida International University, Miami, FL 33199 USA; emails: {lli003, taoli}@cs.fiu.edu; W. Peng, S. Kataria, and T. Sun are with Xerox Innovation Group, Xerox Corporation, 800 Phillips Rd, Webster, NY 14580; emails: {Wei.Peng, Saurabh.Kataria, Tong.Sun}@xerox.com.

DOI: http://dx.doi.org/10.1145/2757282

1. INTRODUCTION

Social media has gradually integrated into the daily activities of cyber citizens, allowing people at different geographical locations to connect socially over the Internet. Popular social media service websites, such as Twitter and Facebook, are attracting a myriad of online users to share information and interact with each other. Their popularity can be gauged from the fact that Twitter has nearly 125 million users while Facebook has nearly 500 million users [Sachan et al. 2012]. The rich data generated by online users and the affinity of people to these services have attracted substantial attention in both academia and industry. For example, companies often devise targeted marketing campaigns over a group of online users from social media websites to expand their business and exploit cross sell opportunities.

Due to the diversity of topics discussed in social media, one critical issue of utilizing social media data is how to precisely identify users' personal interest and the interest of communities, in which these users are connected to or frequently interact with. Naturally in social media, a community is often formed by a group of users with social connections as well as similar topic preferences. Taking online marketing campaign as an example, marketers not only target individuals with certain interest, but also hope the marketing messages could be cascaded to more audience sharing similar interests. Thus, it is very important to capture both user-oriented interest and community-oriented topics.

Automated discovery of topics and communities has received widespread attention in academia and has been addressed differently in previous works [Sachan et al. 2012; Yang et al. 2009; Zhang et al. 2007; Zhou et al. 2006; Peng and Li 2011]. One common approach is to use generative Bayesian models to capture the relationships among users, communities, and topics, with the assumption that these three types of entities are naturally correlated. However, prior approaches cannot make a distinction between user-oriented and community-oriented topics. Taking a query "campaign + economy" as an example, the task is to identify users and communities that are interested in US presidential campaign and also often discuss the topic of economy related to the campaign. "campaign" is discussed by a lot of people as it is relevant to the presidential selection, whereas "economy" often appears in users' general posts and may not be related to "campaign." In this case,

—if we only consider user-oriented topics, these two keywords are often treated independently since they might belong to different latent topics factors. Even if the recommended users post more relevant content, their topic interest toward these two keywords remain vague. In addition, the recommended users might not be involved in any community discussion on the query topic, and therefore, they have very limited interactions, which is not helpful to expand the social network;

—if we only consider community-oriented topics, the detected topic structure might be very loose. In general, a user's preference is often diverse, which may contain different topic interests. Detecting topics in an indiscriminate way will result in a lot of noise since all the user-generated content will contribute to the community topics. Therefore, we cannot distinguish whether "economy" is originated from community discussion or it comes from users' general posts.

The advantage of modeling user-oriented and community-oriented topics simultaneously is that it could identify high-quality community topics by sampling the topic for each word from either the community topic-word distribution or the user topic-word distribution. Thus, the noises induced by a wide variety of user interests that could contaminate the community topics can be naturally mitigated.

In our work, we identify the latent relationships among social objects, that is, users and communities, by distinguishing a user's interest from the focuses of communities. We propose a generative topic model to capture these two types of interests as topics in the same parameter space with an additional parameter that identifies the association of interests to either a given user or a given community. Our proposed model makes use of the communities derived from the social links of users to avoid the expensive computation of combining the community discovering process with the topic modeling process. By introducing a switch variable to control the generative process with respect to a given word, the model is capable of distinguishing a user's personal topical interests from his/her community-wise concerns. Extended from our previous work [Li et al. 2013], we further provide a novel recommendation framework, named `FRec`, based upon the derived relationships, which is able to recommend topic-related influential users and topic-cohesive interactive communities for a given user's profile or a set of keywords. To provide high-quality recommendation, the proposed framework utilizes both structural information of the social network and topic probabilities of the information network. In this way, the deficiencies of recommendation either by link prediction or by pure content analysis can be alleviated.

1.1. Our Contribution

The contribution of our work is threefold.

—A modeling approach to distinguishing community versus user interests (cf. Section 5): Unlike previous approaches focusing on analyzing topics in communities, our proposed model distinguishes community-oriented topics from users' personal topic interests. Community-topic distribution is separated from user-topic distribution by an additional Bernoulli variable, which controls the distribution a word is drawn from.

—A principled framework for recommendation (cf. Section 6): Based upon the inference results of the topic model, our proposed framework, `FRec`, is capable of recommending topic-related influential users and topic-cohesive interactive communities given a user profile or a set of keywords as input. The framework comprehensively considers the properties of social media data, for example, the amiable connections and the influential power of users.

—Extensive evaluation on social media data (cf. Section 7): We conduct extensive experimental evaluation and case studies on a dataset related to "presidential campaigns" obtained from Twitter. The results demonstrate that (1) our proposed topic model is superior to several baselines and existing topic models in capturing user and community topics in terms of generalization errors and computation complexity; and (2) our proposed recommendation framework provides more reasonable recommendations of users and communities compared with other baselines and existing works.

1.2. RoadMap

The rest of the paper is organized as follows. Section 2 presents a brief summary of prior work relevant to community-based topic models and recommender systems. Section 3 presents an overview of our framework. Section 4 briefly discusses the strategy used for community detection. Section 5 discusses the proposed topic modeling approach, and Section 6 describes the recommendation strategy. Empirical evaluation of our method is reported in Section 7. Finally Section 8 concludes the paper.

2. RELATED WORKS

Recommendation in social media, for example, user and community recommendation, has been well studied in previous research works. In this section, we highlight the ones that are most related to our work.

2.1. User Recommendation

User recommendation, often referred to as friendship recommendation or link prediction, focuses on recommending users to a target user based on diverse criteria. From a network perspective, user recommendation refers to finding missing edges in a user network. Typical approaches to solving this problem often utilize the network structure and node connections, for example, proximity measures that are derived from network topological features [Liben-Nowell and Kleinberg 2007], supervised learning methods [Al Hasan et al. 2006], relational learning methods [Popescul and Ungar 2003; Taskar et al. 2003], and so on. A detailed survey of link prediction can be found in Lü and Zhou [2011].

In social media, the content generated by users, for example, user relationships or posts, is a valuable information source to model users' preference. Recently, several methods have been proposed to resolve user recommendation in social media by employing latent Dirichlet allocation (LDA) alike topic models [Pennacchiotti and Gurumurthy 2011; Ramage et al. 2010]. These approaches, however, only consider interest similarity, and ignore whether the recommended user is interactive or not, which is essential for expanding the entire social network. In our work, we try to recommend users with influence abilities, given the fact that these users can help enrich the interactions among users. Note that our work is related to identifying influential users and their "network impacts" in social networks [Domingos and Richardson 2001; Kempe et al. 2003; Li et al. 2014]. However, different from general influential user identification, we identify the latent relationships among users and communities. In addition, our model can distinguish users' personal interests from the topics discussed within communities.

2.2. Community Recommendation

In social media environment, people often form communities intentionally or unconsciously. Some services, for example, Orkut,[1] provide functionalities to allow users to join different communities; other services, for example, Twitter, do not have explicit communities in their websites. Instead, implicit communities would be generated as people share information and discuss topics with their friends or users of similar interests.

Automated community discovery has been well studied by researchers. Most previous approaches in community discovery use only the social linking structure among users to identify communities, for example, min-cut-based partitioning, matrix-factorization, centrality-based, and Clique percolation methods [Fortunato 2010; Porter et al. 2009; Wang et al. 2011]. However, they fail to take into account the content generated by users in social network, which might result in the irrationality of the identified communities. For example, two users in a community are reasonably connected through several links, but they might have no common topic interest at all. Another work proposed in Mei et al. [2008] regularizes a statistical topic model with a harmonic regularizer based on a graph structure in the data.

To address this issue, probabilistic models are often employed to capture the topics being discussed by users and within communities [Rosen-Zvi et al. 2010; Zhang et al. 2007]. Several solutions are proposed to simultaneously perform topic modeling and community discovery. For instance, Zhou et al. [2006] proposes two generative Bayesian models for semantic community discovery in social networks, by combining probabilistic modeling with community detection. The sender–recipient relations within social networks are regarded as the basis for modeling. Liu et al. [2009] focuses on the topic

[1] http://www.orkut.com.

and author communities, and proposes a joint model that quantifies the effect of topic and community to the formation of a link. Yin et al. [2012] incorporates community discovery into topic analysis in text-associated graphs to guarantee the topical coherence in the communities.

In social media, other types of information, for example, the type of interactions, can also be taken into account. Yang et al. [2009] combines link and content analysis for community detection by proposing a conditional model for link analysis and a discriminative model for content analysis. Sachan et al. [2012] assumes a user's membership in a community is conditioned on its social relationship, the type of interaction and the information communicated with other members of that community. One limitation of these methods is that they assume all the content generated by a user will contribute to the community detection. In reality, however, an online user often posts his/her personal information, for example, moods and activities, which might not be related to any community. Comparatively, our model distinguishes community-oriented topics from users' personal topics within the content, which is more reasonable in modeling the topic interests of users.

Given the detected communities, a further step for online community management is community recommendation. Chen et al. [2008] proposes a collaborative filtering method for personalized community recommendation, by considering multiple types of cooccurrences in social data, for example, semantic and user information. Chen et al. [2009] uses association rule mining to discover associations between sets of communities that are shared across many users, and LDA [Blei et al. 2003] to model user-community cooccurrences via latent aspects. Both works performed experimental evaluation on Orkut dataset.

3. THE PROPOSED FRAMEWORK

The recommendation framework proposed in our work includes three interleaved modules, described as follows:

Community Detection (cf. Section 4): This module discovers the implicit communities among users in social media. Such communities are constructed based on the virtual friendships between users, and can represent topic-oriented user groups based on the fact that users who share similar topic interest form a friendship.

Topic Modeling (cf. Section 5): In this module, a probabilistic topic model is proposed to capture users' interests and communities' interests simultaneously. By modeling users' interest, we can understand users' general preferences; by modeling communities' interest, we can have an overview on what topics are discussed within a community. The proposed model is able to achieve a better generalization compared with other existing topic models.

Recommendation (cf. Section 6): In this module, we provide functionalities of recommending users and communities given a user profile or a set of keywords. The recommendation is achieved by analyzing the learned interest likelihood from the topic model. We also integrate the structural information of the social network, for example, the affinity among users and the diffusion capability of users, into the recommendation process.

4. COMMUNITY DETECTION

Most methods discussed in Section 2 try to identify communities and find topics discussed within the communities, simultaneously. Generative models are often employed, and the learning processes are conducted using Gibbs sampling. Due to the involvement of social connections among users in the models, and sampling from thousands of neighbors for each user, learning will become computationally expensive, especially when applied to social media data. Thus, they can only be employed in small sets of users with a limited number of connections. In our work, to alleviate this situation, we separate the process of community discovery from the generative topic modeling. We perform community discovery on the users' friendship network, and allow a user to belong to multiple communities by using soft clustering based methods. To this end, we employ the algorithm introduced in Yu et al. [2006] to obtain the community memberships of users.

Traditionally, data similarity relations can be conveniently encoded by a graph, in which vertices denote data objects and adjacency weights represent data similarities. However, it is not trivial to partition such a graph in a probabilistic way with the assumption that a node can belong to different clusters. To resolve this issue, we use a bipartite graph to encode the conditional probability of transitions between node, that is, the probability of a data object belonging to a cluster. Formally, let $K(U, C, E)$ be a bipartite graph, where $U = \{v_i\}_{i=1}^{n}$ contains all the data objects (in our case, all the users), $C = \{c_p\}_{p=1}^{m}$ contains all the clusters (in our case, all the communities), and E contains all the edges indicating that v_i belongs to c_p. Let $B = \{b_{ip}\}$ denote the $n \times m$ adjacency matrix with $b_{ip} \geq 0$ being the weight for edge $[v_i, c_p]$. The bipartite graph induces a similarity between data object v_i and v_j by Zhou et al. [2005]

$$\delta_{ij} = \sum_{p=1}^{m} \frac{b_{ip} b_{jp}}{\lambda_p} = (B\Lambda^{-1}B^T)_{ij}, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \tag{1}$$

where $\lambda_p = \sum_{i=1}^{n} b_{ip}$ denotes the degree of vertex $c_p \in C$. Let $\Delta = \{\delta_{ij}\}$ be the adjacency matrix. Without loss of generality, we normalize Δ to ensure $\sum_{ij} \delta_{ij} = 1$. Our goal is to construct such a bipartite graph to approximate the original data similarity graph. To obtain such approximation, we need to minimize $d(\Delta, B\Lambda^{-1}B^T)$ based on a given distance measurement $d(\cdot, \cdot)$ between two adjacency matrices. To simplify the problem, we decouple B and Λ via $H = B\Lambda^{-1}$, and consequently, have

$$\min_{H,\Lambda} d(\Delta, H\Lambda H^T), \quad \text{s.t.} \sum_{i=1}^{n} h_{ip} = 1, H \in \mathbb{R}_{+}^{n\times m}, \Lambda \in \mathbb{D}_{+}^{m\times m}, \tag{2}$$

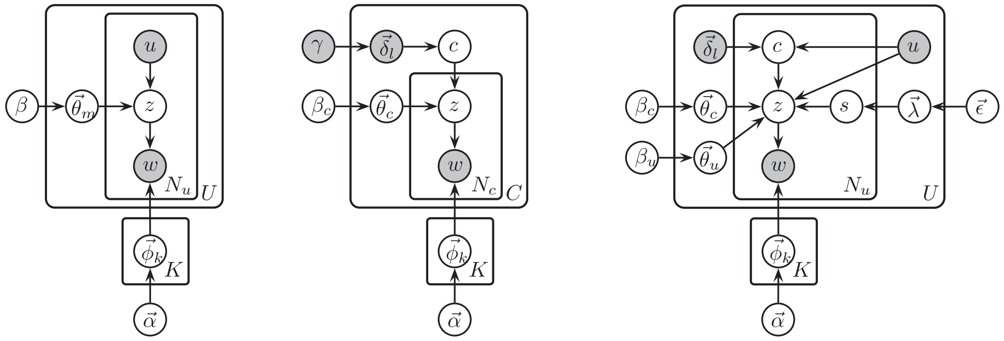
where $\mathbb{D}_{+}^{m\times m}$ denotes the set of $m \times m$ diagonal matrices with positive diagonal entries. This problem is a symmetric variant of nonnegative matrix factorization, and the solution can be found in Seung and Lee [2001] and Yu et al. [2006].

5. USER-COMMUNITY-TOPIC MODEL

In this section, we first discuss two basic topic models used for tracking topic interests of online users or online communities. Based on the discussion, we come up with our User-Community-Topic (UCT) model in order to resolve the issues in the two basic models. We then describe how to learn the hyperparameters using Gibbs sampling.

5.1. Discussion on Topic Models

Figure 1(a) shows the graphical model for what we refer to as the "user-topic model" (UT). UT aims to capture the correlation between users and topics. In UT, the generation of a document (containing all the posts of a user) is affected by the topic factor, that is, a document is considered as a mixture of topics. Each topic corresponds to a multinomial distribution over the vocabulary. The existence of observed word w in document N_u is considered to be drawn from the word distribution $\vec{\phi}_k$, which is specific to topic z. Similarly, the topic z is drawn from the document-specific topic distribution $\vec{\theta}_m$. Modeling latent factors as variables in the Bayesian network provides the



(a) User-Topic Model. (b) Community-Topic Model. (c) User-Community-Topic Model.

Fig. 1. Plate diagram for three topic models.

capacity of grouping the words used by a user into semantic topics. Based on the learned posterior probability, a document (a user's posts) can be denoted as a mixture of topic distribution. Each user's preference of using words and involvement in topics can also be discovered. Hence, UT model is equivalent to the standard LDA model.

However, in most cases, users might have diverse interests over topics. By using UT model, the obtained posterior probability of a user over a specific topic might be affected by the general topic interests of this user. In addition, users in social media often share common interests over topics, which cannot be captured in UT model.

Another model is called "community-topic model" (CT) [as shown in Figure 1(b)], in which the generation of a document is affected by both the topic factor and the community factor in a hierarchical manner. In CT model, we treat all the posts within a community as a document, and then, model the relations among words, topics, and communities. Similar to the plate notation of UT, a document here is considered as a mixture of topics, and each topic is represented by a multinomial distribution over the vocabulary. The difference is the community factor c, by which topics within a document would be affected. Here, the factor c denotes the community set that generates the document N_c.

One major problem of the CT model is that user posts in a community could include various topics, rendering the community document highly inconsistent. The posterior probabilities can be estimated through iteratively sampling instances from the repository and updating the corresponding probabilities. However, in this model, the content contained in an online community might be gigantic since a community could involve a myriad of users along with their posts. Sampling for all the words in a community document would result in uncontrolled generalization error for inference due to the noisy feature of social media data. In addition, there is no way to capture a specific user's interests using CT model, since no user factor is involved.

5.2. The Proposed Model

Our goal is to model the relations among users, topics and communities within the environment of social media. Taking Twitter as an example, we have the tweets posted by users and the follower–followee relations of users; however, we do not have the explicit community membership of users. We therefore assume there is a community factor c that captures the user-community memberships with respect to user u. Also, within each community, users might discuss different topics, and hence, we have a topic factor z that characterizes the topic-community relations. For topic mixture and term

Table I. Notations for Quantities in the Model

U	The user set in the community data.
V	The dictionary of texts in the community data.
L	The number of communities predefined.
N_u	The term set of texts posted by user u.
$\vec{\alpha}$	Dirichlet prior hyperparameter (known) on the term distribution.
$\vec{\beta}$	Dirichlet prior hyperparameter (known) on the mixture topic distribution.
$\vec{\gamma}$	Prior hyperparameter (known) on the mixture community distribution.
$\vec{\epsilon}$	Prior hyperparameter on the binary mixture.
$\vec{\phi_k}$	$p(t\|z = k)$, the mixture component of topic k.
$\vec{\theta_m}$	$p(z\|u = m)$, the topic mixture proportion for user m.
$\vec{\delta_l}$	$p(u\|c = l)$, the user proportion for community l. (observed)
$\vec{\lambda}$	Binary mixture for word generation.
c	The community mixture.
u	Mixture indicator that chooses a user from a community.
z	Mixture indicator that chooses the topic for the term from a user.
w	Term indicator for the word from a user.
s	Binary factor for word generation.

mixture, we give them Dirichlet priors; for community mixture, we use the distribution derived by analyzing the follower–followee relations. Table I lists the notations used in our model.

We denote our proposed topic model as "UCT" model. As described in Figure 1(c), UCT has a similar general structure to the LDA model, but with additional machinery to distinguish user-oriented topics and community-oriented topics. We fix the probability of $p(c|u)$, which is derived from the soft clustering result. In this way, we are not concerned with the relations between the community and the user, but focus more on the relations between the community and the topic (i.e., $p(z|c)$), and the relations between the user and the topic (i.e., $p(z|u)$). We add a latent Bernoulli variable s (a binary factor) to indicate whether a word is related to a user itself or to a community. In particular, s takes value 0 if the word w is generated via the user-topic route, value 1 if the word is generated from the community-topic route. The variable s in our model acts as a switch: if $s = 0$, words are sampled from a user-specific multinomial $\vec{\theta}_u$, whereas if $s = 1$, words are sampled from a community-specific multinomial $\vec{\theta}_c$ (with different symmetric Dirichlet priors parameterized by β_u and β_c). s is sampled from a document-specific Bernoulli distribution $\vec{\lambda}$, which in turn has a prior ϵ. The joint probability of the UCT model can be written as:

$$\begin{aligned} &p(w, z, u, c, s, \phi_k, \theta_u, \theta_c, \delta_l, \lambda|\vec{\alpha}, \vec{\beta}_c, \vec{\beta}_u, \vec{\epsilon}) \\ &\quad = p(w|z, \phi_k)p(z|u, c, s, \theta_u, \theta_c)p(c|u, \delta_l) \\ &\qquad \cdot p(s|\lambda)p(\lambda|\vec{\epsilon})p(\phi_k|\vec{\alpha})p(\theta_u|\vec{\beta}_u)p(\theta_c|\vec{\beta}_c), \end{aligned}$$

where $p(z|u, c, s, \theta_u, \theta_c) = p(z|u, s = 0, \theta_u)$ (where $s = 0$), and $p(z|u, c, s, \theta_u, \theta_c) = p(z|c, s = 1, \theta_c)$ (where $s = 0$). Here, $p(z|u, s = 0, \theta_u)$ is the probability of a user-specific topic, whereas $p(z|c, s = 1, \theta_c)$ is the probability of a community-specific topic. Given the graphical model described in Figure 1(c), the generative scheme is shown in Algorithm 1.

ALGORITHM 1: Generative scheme of UCT model.

```
for each topic $z \in (1, \ldots, K)$ do
  Sample $\phi_k \sim \mathrm{Dir}(\cdot|\vec{\alpha})$
end for
for each user $u \in (1, \ldots, U)$, do
  Sample $\lambda_u \sim \mathrm{Beta}(\cdot|\vec{\epsilon})$
  for each word $w \in (1, \ldots, N_u)$, do
    Sample $s \sim \mathrm{Bern}(\cdot|\vec{\lambda}_u)$
    Choose a community assignment $c_u \sim \mathrm{Mult}(\cdot|\vec{\delta}_l)$
    if(s==0): then
      Choose a topic assignment $z \sim \mathrm{Mult}(\cdot|\vec{\theta}_u)$
    else
      Choose a topic assignment $z \sim \mathrm{Mult}(\cdot|\vec{\theta}_c)$
    end if
    Choose a term $w \sim \mathrm{Mult}(\cdot|\vec{\phi}_k, z)$
  end for
end for
```

5.2.1. Gibbs Updates. Gibbs sampling is an procedure to approximate the joint distribution of multiple variables by iteratively drawing a sequence of samples. As a special case of the Metropolis–Hastings algorithm [Robert and Casella 2013], Gibbs sampling is a Markov Chain Monte Carlo algorithm, in which the parameters are often integrated out and the corresponding posterior probability can be estimated. To estimate the model, we use the collapsed Gibbs sampling [Griffiths and Steyvers 2004]. For our UCT model, we are interested in the latent user-topic portions $\vec{\theta}_u$, the latent community-topic portions $\vec{\theta}_c$, the topic-word distributions $\vec{\phi}_k$, and the topic index assignments for each word z_i. Also in the learning process, the value of s will be generated based on a Bernoulli distribution and be updated through the Gibbs sampling for each word. $\vec{\theta}_u$, $\vec{\theta}_c$, and $\vec{\phi}_k$ can be calculated using just the topic index assignments z_i, that is, $\mathbf{z}$ is a sufficient statistic for the three distributions. Therefore, we can integrate out the multinomial parameters and simply sample z_i and s_i.

Given the set of users U, the user-community membership distribution $\vec{\delta}_l$, the set of integrated posts D, and the number of desired topics T, Gibbs sampling starts with randomly assigning words to a community/user and topic, with different values of the switch variable s. A Markov chain can be constructed to converge to the target distribution through iteratively sampling. In each trial of the Monte Carlo simulation, a tuple of (*community / user, topic, switch*) can be assigned to an observed word w_i in each document. The collapsed Gibbs sampler needs to compute the probability of a topic z being assigned to a word w_i, given all other topic assignments to all other words, with respect to a specific value of s (0 or 1). Similarly, it needs to calculate the probability of s being assigned to a word w_i, given all other s assignments to all other words. Let $\mathbf{z}_{-i}$ denote all topic allocation except for z_i and $\mathbf{s}_{-i}$ represent all s assignments except for s_i. The probabilities that we need to update include: (1) $p(s_i = 0|\mathbf{s_{-i}}, w_i, z_i, u_i, c_i)$, (2) $p(s_i = 1|\mathbf{s_{-i}}, w_i, z_i, u_i, c_i)$, (3) $p(z_i|\mathbf{z_{-i}}, w_i, s_i = 0, u_i, c_i)$, and (4) $p(z_i|\mathbf{z_{-i}}, w_i, s_i = 1, u_i, c_i)$.

To illustrate how to update these probabilities, let us take $p(z_i|\mathbf{z_{-i}}, w_i, s_i = 0, u_i, c_i)$ as an example to derive its updating formula. $p(z_i|\mathbf{z_{-i}}, w_i, s_i = 0, u_i, c_i)$ denotes the probability that w_i is generated by user-related topic z_i when the switch variable s_i equals to 0, which is conditioned on all the assignments of words excluding the current observation of w_i, for example, $\mathbf{z_{-i}}$ represents all the assignments of topic not including

Table II. Gibbs Updates for UCT Model

$$
\begin{aligned}
p(s_i = 1|\mathbf{s_{-i}}, w, z, u, c) &\propto \frac{p(s_i = 1, \mathbf{s_{-i}}, w, z, u, c)}{p(\mathbf{s_{-i}}, w, z, u, c)} \propto p(s_i = 1|z_i, c_i) = p(z_i|s_i = 1, c_i) \cdot p(s_i = 1|u_i) \\
&\propto \frac{n_{z_i, c_i, s_i=1} + \beta_c(z_i)}{\sum_{z_i} n_{z_i, c_i, s_i=1} + \sum_{z_i} \beta_c(z_i) - 1} \cdot \frac{n_{s_i=1, u_i=u} + \epsilon_{s=1}}{\sum_{s_i} n_{s_i=1, u_i=u} + \epsilon_{s=0} + \epsilon_{s=1} - 1}. \\
p(s_i = 0|\mathbf{s_{-i}}, w, z, u, c) &\propto \frac{p(s_i = 0, \mathbf{s_{-i}}, w, z, u, c)}{p(\mathbf{s_{-i}}, w, z, u, c)} \propto p(s_i = 0|z_i, u_i) = p(z_i|s_i = 0, u_i) \cdot p(s_i = 0|u_i) \\
&\propto \frac{n_{z_i, u_i, s_i=0} + \beta_u(z_i)}{\sum_{z_i} n_{z_i, u_i, s_i=0} + \sum_{z_i} \beta_u(z_i) - 1} \cdot \frac{n_{s_i=0, u_i=u} + \epsilon_{s=0}}{\sum_{s_i} n_{s_i=0, u_i=u} + \epsilon_{s=0} + \epsilon_{s=1} - 1}. \\
p(z_i|\mathbf{z_{-i}}, w, s_i = 0, u, c) &\propto \frac{p(\mathbf{z_i}, w, s_i = 0, u, c)}{p(\mathbf{z_{-i}}, w, s_i = 0, u, c)} \propto p(z_i, s_i = 0, w_i, u_i, c_i) = p(w_i|z_i) \cdot p(z_i|s_i = 0, u_i) \\
&\propto \frac{n_{w_i, z_i} + \alpha(w_i)}{\sum_V n_{w_i, z_i} + \sum_V \alpha(w_i) - 1} \cdot \frac{n_{z_i, u_i, s_i=0} + \beta_u(z_i)}{\sum_{z_i} n_{z_i, u_i, s_i=0} + \sum_{z_i} \beta_u(z_i) - 1}. \\
p(z_i|\mathbf{z_{-i}}, w, s_i = 1, u, c) &\propto \frac{p(\mathbf{z_i}, w, s_i = 1, u, c)}{p(\mathbf{z_{-i}}, w, s_i = 1, u, c)} \propto p(z_i, s_i = 1, w_i, u_i, c_i) = p(w_i|z_i) \cdot p(z_i|s_i = 1, c_i) \\
&\propto \frac{n_{w_i, z_i} + \alpha(w_i)}{\sum_V n_{w_i, z_i} + \sum_V \alpha(w_i) - 1} \cdot \frac{n_{z_i, c_i, s_i=1} + \beta_c(z_i)}{\sum_{z_i} n_{z_i, c_i, s_i=1} + \sum_{z_i} \beta_c(z_i) - 1}.
\end{aligned}
$$

the current assignment of word w_i. This probability can be calculated by

$$
\begin{aligned}
p(z_i|\mathbf{z_{-i}}, w, s_i = 0, u, c) &\propto \frac{p(\mathbf{z_i}, w, s_i = 0, u, c)}{p(\mathbf{z_{-i}}, w, s_i = 0, u, c)} \\
&\propto p(z_i, s_i = 0, w_i, u_i, c_i) = p(w_i|z_i) \cdot p(z_i|s_i = 0, u_i) \\
&\propto \frac{n_{w_i, z_i} + \alpha(w_i)}{\sum_V n_{w_i, z_i} + \sum_V \alpha(w_i) - 1} \cdot \frac{n_{z_i, u_i, s_i=0} + \beta_u(z_i)}{\sum_{z_i} n_{z_i, u_i, s_i=0} + \sum_{z_i} \beta_u(z_i) - 1}, \qquad (3)
\end{aligned}
$$

where $p(w_i|z_i)$ and $p(z_i|s_i = 0, u_i)$ can be estimated via

$$
p(w_i|z_i) \propto \frac{n_{w_i, z_i} + \alpha(w_i)}{\sum_V n_{w_i, z_i} + \sum_V \alpha(w_i) - 1}, \quad p(z_i|s_i = 0, u_i) \propto \frac{n_{z_i, u_i, s_i=0} + \beta_u(z_i)}{\sum_{z_i} n_{z_i, u_i, s_i=0} + \sum_{z_i} \beta_u(z_i) - 1}. \qquad (4)
$$

In Eq. (4), n_{w_i, z_i} is the number of times that word w_i is assigned to topic z_i, not including the current instance (assignment). $n_{z_i, u_i, s_i=0}$ is the number of times that topic z_i is assigned to user u_i with the switch variable $s_i = 0$, not including the current instance. To calculate Eq. (4), we need to maintain a $V \times T$ matrix, each entry of which is the number of times that the corresponding word w_i is assigned to topic z_j; we also need to keep a $T \times U$ matrix, each entry of which records the number of times that the corresponding topic z_i is assigned to user u_j.

The derivations of the updates for the expected probabilities are described in Table II, and the notations are introduced in Table III.

We analyze the computational complexity of Gibbs sampling in the proposed UCT model. As discussed previously, in Gibbs sampling, we need to compute the posterior probability $p(z_i|\mathbf{z_{-i}}, w_i, s_i, u_i, c_i)$ for user-word pairs $(U \times V)$ and community-word pairs $(C \times V)$, where V is the total number of words. Each $p(z_i|\mathbf{z_{-i}}, w_i, s_i, u_i, c_i)$ consists of K topics, and requires a constant number of arithmetic operations, resulting in $O(V \cdot K \cdot U)$, assuming $U \gg C$, for a single sampling.

Table III. Descriptions for Notations in the Formula

Notations	Descriptions
$n_{z_i,c_i,s_i=1}$	# of times that z_i is assigned to c_i with $s_i = 1$ for a word w_i.
$n_{z_i,c_i,s_i=0}$	# of times that z_i is assigned to c_i with $s_i = 0$ for a word w_i.
$n_{z_i,s_i=1}$	# of times that $s_i = 1$ is assigned to z_i.
$n_{z_i,s_i=0}$	# of times that $s_i = 0$ is assigned to z_i.
n_{w_i,z_i}	# of times that w_i is assigned to z_i.
$n_{z_i,u_i,s_i=0}$	# of times that z_i is assigned to u_i with $s_i = 0$ for a word w_i.

6. RECOMMENDATION STRATEGIES

In our work, we are trying to solve the problem of recommending users and communities for a given user's profile or a set of keywords within social media environments. In general, user recommendation can be addressed by selecting the most similar users in terms of the topic interest; however, the interactiveness of users is often being ignored, which is essential for expanding a social network. The scenario of community recommendation is different, in which a community can either be a public page that are liked by a lot of users, or a list of people that are often interacting with each other to discuss similar topics. In our work, we focus on the latter, that is, to recommend a list of users with relevant topic interests and cohesive discussions. The target user can select some of the recommended users as friends, and then, start to involve the discussion among these users.

Our recommendation framework, FRec, provides various recommendation mechanisms based on our UCT model. In our problem setting, the target can be either a user profile or a list of keywords, and the recommended objects can be either a list of users or a list of communities. Therefore, we can have four different recommendation strategies based on the target and the recommended objects.

We also consider the user influence with respect to a topic. For each topic in the topic list, we can use the derived probabilities $p(u|z)$ as the initialization of the PageRank algorithm [Lawrence et al. 1998], and run PageRank on the friendship network to obtain the influence scores of users toward a specific topic z. Then, the topic-relevant user influence can be denoted as $R(u|z)$. We can setup a threshold for $p(u|z)$ to filter out low probabilities (in the experiments, we empirically set $p(u|z)$ as 0.01).

6.1. User-to-User Recommendation

Given a target user $\hat{u}$, we can rank other users based on $p(u_i|\hat{u})$, and then, select top ranked ones as $\hat{u}$'s recommendation. $p(u_i|\hat{u})$ can be calculated using Eq. (6).

$$\begin{aligned} p(u_i|\hat{u}) &= \frac{\sum_z \sum_c p(u_i \hat{u} c z)}{p(\hat{u})} \\ &= \sum_z \sum_c p(z|\hat{u}) p(z|u_i, s=0) p(u_i) p(s=0) p(z|c, s=1) p(c) p(s=1) p(c|u_i) p(c|\hat{u}) \\ &\propto p(u_i) \sum_z \sum_c p(z|\hat{u}) p(z|u_i, s=0) p(z|c, s=1) p(c|u_i) p(c|\hat{u}) p(c) \\ &\propto p(u_i) \sum_z \left(p(z|\hat{u}) p(z|u_i, s=0) \sum_c p(z|c, s=1) p(c|u_i) p(c|\hat{u}) p(c) \right). \end{aligned} \tag{6}$$

Here, $p(z|\hat{u})$ is the probability of topics given a test user $\hat{u}$, which can be obtained by extending Gibbs iterations over the test users after the hyperparameters are learned. Note that in Eq. (6), we consider both user-based topics ($p(z|u_i, s == 0)$) and community-based topics ($p(z|c, s == 1)$). The user-based topics often include a user's personal interest. To make the recommendation more community oriented, we can focus on community-based topics by removing the user-based component. The recommendation can be refined as

$$p(u_i|\hat{u}) \propto p(u_i)\sum_z \left(\frac{p(z|\hat{u})}{p(z)}\sum_c p(z|c, s == 1)p(c|u_i)p(c|\hat{u})p(c)\right). \tag{7}$$

By integrating the user influence into $p(u_i|\hat{u})$, we can have

$$p(u_i|\hat{u}) \propto p(u_i)\cdot\sum_z \left(\frac{p(z|\hat{u})R(u_i|z)}{p(z)}\sum_c p(z|c, s == 1)p(c|u_i)p(c|\hat{u})p(c)\right). \tag{8}$$

In this strategy, the user-to-user relations residing in the friendship network are not considered. In order to make the recommendation more reasonable, we incorporate the neighborhood similarity between u_i and the target user $\hat{u}$ into the recommendation. The neighborhood similarity can be calculated as

$$\mathrm{sim}(u_i, \hat{u}) = \frac{|\mathrm{neighborhood}(u_i) \cap \mathrm{neighborhood}(\hat{u})|}{|\mathrm{neighborhood}(u_i) \cup \mathrm{neighborhood}(\hat{u})|}, \tag{9}$$

where neighborhood$(\cdot)$ denotes all the neighbors of the user. By integrating $\mathrm{sim}(u_i, \hat{u})$ into Eq. (8), we have

$$\tilde{p}(u_i|\hat{u}) \propto p(u_i)\cdot sim(u_i, \hat{u})\cdot\sum_z \left(\frac{p(z|\hat{u})R(u_i|z)}{p(z)}\sum_c p(z|c, s == 1)p(c|u_i)p(c|\hat{u})p(c)\right). \tag{10}$$

6.2. User-to-Community Recommendation

Given a target user $\hat{u}$, we can also recommend communities to $\hat{u}$ based on the derived correlations among users, topics and communities. Given a community c, we can measure the relevance between $\hat{u}$ and c by

$$\begin{aligned} p(c|\hat{u}) &= \frac{\sum_z p(c, \hat{u}, z)}{p(\hat{u})} \\ &\propto \sum_z \frac{p(z|\hat{u}, s = 0)p(z|c, s = 1)p(c)}{p(z)} \\ &\propto p(c)\sum_z \frac{p(z|\hat{u}, s == 0)p(z|c, s == 1)}{p(z)}. \end{aligned} \tag{11}$$

A community with more influential users is likely to be more interactive, that is, it may involve more activities of sharing information and discussing topics. Therefore, we consider the user influence for community recommendation. By integrating the user influence into $p(c|\hat{u})$, we have

$$\tilde{p}(c|\hat{u}) \propto p(c)\sum_z \frac{p(z|\hat{u}, s = 0)p(z|c, s = 1)\cdot\left(\sum_{u_j\in c} R(u_j|z)\right)}{p(z)}. \tag{12}$$

6.3. Keyword-to-User Recommendation

Besides recommendation based on user profiles, we allow the user to input some keywords and then return this user a list of users who are relevant to the keywords set.

User recommendation based on keywords is a typical functionality in social media websites. Users often prefer to obtain recommendations based on a specific topic (e.g., a set of keywords), rather than passively receive recommendation without any explanation. For example, a user might feed "US economy" into the website, and then receive a list of users that often discuss this topic. Then, the target user can send friendship invitation to some of them for further discussion.

Formally, given a set of keywords $W = \{w_1, \ldots, w_k\}$, we can calculate the relevance between a user u and this keywords set W by

$$p(u|W) = \frac{\sum_z p(uWz)}{p(W)} \propto \sum_z \left(p(z|u)p(u) \prod_{w_i} p(w_i|z) \right) \propto p(u) \sum_z \left(p(z|u) \prod_{w_i} p(w_i|z) \right). \tag{13}$$

Here, $p(w_i|z)$ is the probability of a query term w_i given a topic z, and it can be calculated as

$$p(w_i|z) \propto \frac{n_{w_i} + \alpha}{\sum_{w_i'} (n_{w_i'} + \alpha)}.$$

By integrating the user influence into $p(u|W)$, we can have

$$\tilde{p}(u|W) \propto p(u) \sum_z \left(p(z|u) R(u|z) \prod_{w_i} p(w_i|z) \right). \tag{14}$$

6.4. Keyword-to-Community Recommendation

Similar to the aforementioned recommendation strategy, we allow a user to input some keywords and then recommend a list of communities to him/her. The recommendation can be obtained by evaluating the relevance between a community c and the given keywords set W as

$$p(c|W) = \frac{\sum_z p(cWz)}{p(W)} \propto \sum_z \left(p(z|c)p(c) \prod_{w_i} p(w_i|z) \right) \propto p(c) \sum_z \left(p(z|c) \prod_{w_i} p(w_i|z) \right). \tag{15}$$

By integrating the user influence into $p(c|W)$, we have

$$\tilde{p}(c|W) \propto p(c) \sum_z \left(p(z|c) \sum_{u_m \in c} R(u_m|z) \prod_{w_i} p(w_i|z) \right). \tag{16}$$

6.5. Discussion

In our proposed recommendation framework, we reason on combining both topic probabilities derived from content and user relations contained within the network structure. The way we integrate these two aspects for recommendation provides natural complement to the deficiencies of each aspect. In the following, we initially discuss the defects of utilizing a single aspect (either content analysis or structural analysis) for recommendation, and then explain how the local relations of users (i.e., user affinities) and the global influence of users (i.e., users' influence cascading) will affect the recommendation results.

6.5.1. Content Analysis Versus Structural Analysis. In the context of social media, like-minded people often discuss about similar topics, and form friendships with each other. It is straightforward to utilize the content posted by online users when performing recommendation. However, if we recommend users to the target user purely from the content perspective, the recommended users may have no indirect connections to the target user, and hence, they might not form friendships. On the other hand, structural information is often used in analyzing networks. A lot of works have been published along this direction, which is known as link prediction. Recommending users purely based on link information may achieve higher recall; however, in some cases, it is not reasonable since the recommended user might not share any topic interests with the target user, even though they connect with each other through a limited number of links.

In our work, we propose to integrate content analysis with structural analysis when performing the recommendation. To this end, we first derive the correlation among users, communities, and topics through generative topic model (i.e., $p(z|u, s = 0)$ and $p(z|c, s = 1)$), and then, calculate the user-user similarities and the community-user affinities via these probabilities. To include the structural information, we consider a user's topical influence over the network (i.e., global influence cascading) and a user's affinity with other users (i.e., local user similarities), that is, in Eqs. (10) and (12). Such a recommendation strategy is able to alleviate the shortcomings of content analysis or structural analysis, and hence, to achieve more reasonable recommendation results, as shown in the experiments described in Section 7.

6.5.2. Local Effect of User Affinities. The major goal of most social media services is to provide users a making-friend platform, on which users are freely sharing their thoughts or information with other users, and becoming companions to each other. In reality, it is more possible that a user would make friends with his/her friends-of-friends, that is, through the one-hub connections [Roth et al. 2010]. Therefore, in our work, when recommending users to a target user, we explicitly consider user affinities from a network perspective, by calculating users' pairwise similarities through their common friends, as described in Eq. (9). Such a strategy evaluates the local effect of how a user might link to his/her potential friends, and hence, provide more reasonable recommendation results. The empirical study in Section 7.3.1 demonstrates that the local similarity can slightly improve the performance of user recommendation.

6.5.3. Global Effect of User Influence. In social media services, users typically have numerous connections to other site members through a virtual friendship; however, studies show that only a fraction of those so-called friends may affect a user's site usage [Cha et al. 2010]. In addition, it is easy to observe that most users would prefer to follow a member with certain popularities on some topics that those users are interested in, that is, the topical influential users [Weng et al. 2010]. Based on these intuitions, we propose to incorporate a user's topical influence into our recommendation framework [as defined in Eq. (8, 12, 14, and 16)], via a topic-sensitive PageRank model [Haveliwala 2002]. As evaluated in Section 7.3.1, the quality of the recommendation results can be improved via recommending topically influential users to the target user.

7. EMPIRICAL EVALUATION

In this section, we evaluate the proposed topic model and the recommendation framework on a real-world Twitter dataset. For topic modeling, we use quantitative measures, for example, Perplexity, to evaluate the predictive power of the model. We also compare the identified user topics and community topics by sampling words of $s = 0$ and $s = 1$. For recommendation, we focus on three aspects: (1) Is the recommended user or community relevant to the given input? (2) Does the recommended user or community have more influential power? and (3) Does the recommended community have more cohesive topics with respect to the input?

7.1. Real-World Data

The dataset used in the experiment is a collection of tweets related to "presidential campaigns" between Barack Obama and Mitt Romney, ranging from March 1st, 2012 to May 31st, 2012. We crawled the tweets through Twitter Streaming API by feeding a list of keywords related to the campaign (e.g., campaign, Obama, Romney, and so on) into the API request. We then crawled the follower relationships of each user within the tweets dataset. Due to the property of microblogging services, the crawled tweets might contain a lot of noise, which would hinder the topic modeling. Therefore, we did a series of preprocessing to alleviate the negative impact of noise data, including: (1) removing short tweets (with the word count less than 10); (2) removing tweets with hashtags more than 3; (3) removing tweets whose author has no more than five tweets; and (4) removing usernames (starting with "@") and URLs. After preprocessing, the tweets data contain 133,465 users, 5,558,763 mutual-following relationships, and 5,079,994 tweets.

7.2. Comparison of Topic Models

For topic modeling, we concatenate the tweets of each user in the dataset as a document. We process the tweets data by removing stopwords, tokenizing, and stemming using MALLET [McCallum 2002]. To further reduce the noise of the tweets and expedite the learning process, we calculate the TF-IDF score of each word and then select the top ranked 10,000 words as features. After processing, the total number of word tokens in the tweets data is 6,643,278. We compare UCT model with four baselines: (1) Combinational Collaborative Filtering (CCF) [Chen et al. 2008]: this method combines the bag-of-users model and the bag-of-words model to capture the relations among topics, communities and users within the network, and formalizes the recommendation using the inferred probabilities. (2) Topic User Community Model (TUCM) [Sachan et al. 2012]: this approach assumes that a user's membership in a community is conditioned on its social relationship, the type of interaction and the shared information with other members. TUCM focuses on the information transition between the author and the recipient, which renders much more complex sampling in the training process; comparatively, our model simplifies the relationship between online users by only considering their community memberships. We use the simplest TUCM model in Sachan et al. [2012] for comparison in order to fit it into our large data set. (3) Topic-Link (Topic-Link LDA) [Liu et al. 2009]: a joint model that quantifies the effect of topic and community to the formation of a link. In this method, whether a link exists between two documents follows a binomial distribution parameterized by the similarity between topic mixtures and community mixtures as well as a random factor. (4) NetLDA (Topic Model with Network Structure) [Mei et al. 2008]: This method combines a statistical topic model with a harmonic regularizer based on a graph structure in the data. Through regularization, the method presents a general framework for community discovery. In the experiment, we replace the original PLSA topic model with the basic LDA model.

We also include the models shown in Figure 1(a) (UT) and Figure 1(b) (CT) in the comparison. Notice that the research efforts related to community discovery and recommendation are far from the ones considered in our experiments. For example, Zhou et al. [2006] proposed two generative Bayesian models for semantic community discovery in social networks. However, the models in Zhou et al. [2006] are extremely complex due to the sampling of friends if we apply them to our problem setting.

7.2.1. Perplexity Comparison. We compare the predictive performance of our proposed UCT model with other baselines by computing the perplexity of unseen words in test documents. The perplexity of a test set $\mathcal{D}^{\text{test}}$ under a model is defined as the reciprocal

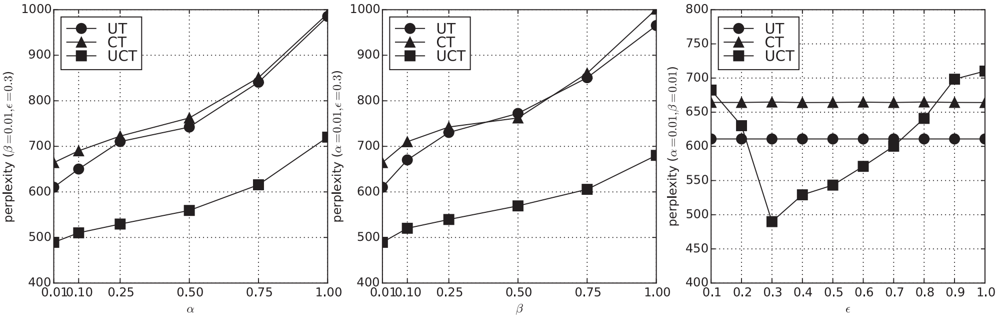


Fig. 2. Perplexity comparison of different topic models through grid search.

geometric mean of the likelihood of $\mathcal{D}^{\text{test}}$, given the training document set $\mathcal{D}^{\text{train}}$ [Heinrich 2005]:

$$P(\mathbf{w}_{\text{test}}|\mathcal{D}^{\text{train}}) = \exp\left(-\frac{\sum_{d=1}^{\mathcal{D}^{\text{test}}} \log p(\mathbf{w}_d|\mathcal{D}^{\text{train}})}{\sum_{d=1}^{\mathcal{D}^{\text{test}}} N_d}\right),$$

where $\mathbf{w}_{\text{test}}$ is a vector of words in the test dataset and $\mathbf{w}_d$ is a vector of words in document d of the test set. When a fraction of words of a test document d is observed, a Gibbs sampler runs on the observed words to update the document-specific parameters, and these updated parameters are used in the computation of perplexity.

For the models of UT, CT, and UCT, we empirically set the number of communities as 500. To decide the hyperparameters, such as α, β, and ϵ, a common approach is to use a validation set of data to explore various settings of these hyperparameters through grid search [Asuncion et al. 2009]. To this end, we use a grid of hyperparameters for each of α, β, and ϵ, and run the algorithm for each combination of hyperparameters. Specifically, we choose α and β from [0.01, 0.1, 0.25, 0.5, 0.75, 1], and ϵ from $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$. We perform grid search on a validation set that contains 10% of the entire dataset, and compare the resulting perplexity. For each model, we find the best hyperparameter settings (with the lowest perplexity). Through this comparison, we find that when $\alpha = 0.01$, $\beta = \beta_u = \beta_c = 0.01$, and $\epsilon = 0.3$, the corresponding perplexity of each model is lower than the one of other settings. Figure 2 shows the perplexity changes of different models with different parameter settings. We hence use these hyperparameter values in the following experiments. For other baselines, we use similar strategies to obtain the optimal values of hyperparameters. We run 200 iterations of Gibbs sampling for training and extend the chain with 100 iterations over the test set. We calculate the averaged perplexity for tenfold validation on the entire tweets dataset (excluding the validation set). The results are shown in Figure 3.

As is depicted in Figure 3, the predictive performance of two basic models (UT and CT) are not comparable with the other baselines. The reason is straightforward: in both models, only one aspect (either u or c) is considered, which violates the characteristics of the data, since in social media, people post information not only for their own purpose, but also expecting to interact with each other. NetLDA takes into account the network structure and models it using a regularizer. However, the basic LDA model utilized in NetLDA has fewer parameters to capture both user topics and community topics, and hence, the resulted performance of the perplexity is relatively higher than other baselines. CCF combines the word factor and the user factor to capture the correlation between users and communities, and TUCM takes into account the type of interactions between users. Topic-Link explicitly quantifies the formation of a link through the

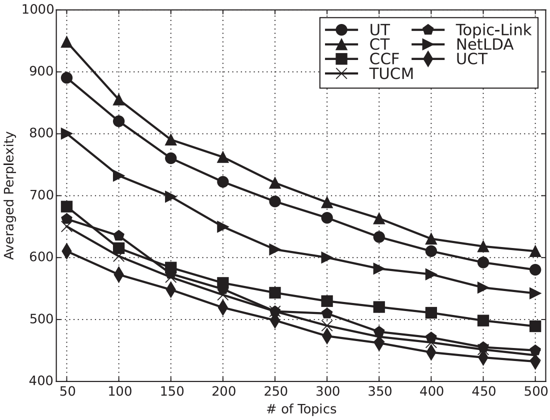


Fig. 3. Averaged perplexity of different topic models as # of topics increases.

Table IV. Sampled Topics for Evaluating *s*

Condition	Topics	
s=0	1	Sign white business bill jobs small today pass big insider fix thursday send week matt weekend signs group passover bio infographic tran call major favor tuesday breaking news...
	2	Company repeal healthfirm dollar save buy worker plan average ACA credit deserve senior free frank mil reform act receive paper afford hurt people card individual change...
	3	Wall drug street marijuana pot raid president urge journal fed state promise smoke lee assault federal complete national mexico penn loss examiner...
s=1	1	Obama president house act congress Barack employment ban government subsidy job expect increase legislation board startup yemen funding...
	2	Million Obama care debt insurance plan bankrupt affordable public health american cost relief explain policy reform coverage benefit...
	3	Obama white house economy bank debt loan increase petition administration JPMorgan company dollar investment wall street market...

similarities between topic mixture and community mixture. These three models achieve better predictive performance compared with UT and CT. Our model distinguishes community-oriented topics and user-oriented topics. Such a distinction indeed exists in most real-world scenarios, that is, a user has his/her personal topic interests, and is also often involved in the discussion within a specific community. In the recommendation experiments, we set the number of topics as 500 for all the models.

7.2.2. Effect of the Bernoulli Variable "s". In our proposed UCT model, we use a Bernoulli variable "*s*" to indicate whether a word is related to a user topic or to a community topic (cf. Section 5). We are concerned with whether such a distinction can be achieved in real-world cases. To evaluate the effect of s, we choose a sample of topics for both $s = 0$ and $s = 1$, and compare the corresponding representative words under different topics. Table IV shows some sampled topics learned from our model under both conditions and the representative words for each topic. As is observed, the topic words of $s = 1$ are more cohesive with respect to the underlying topic, for example, in Topic 2 when $s = 1$, it describes a topic of the relation between economics and debt; Comparatively, the

Table V. Sampled Community Topics for Different Topic Models

CT		CCF	
Topic 1	Topic 2	Topic 1	Topic 2
Obama 0.0085	White 0.0076	Obama 0.0095	Market 0.0101
Major 0.0076	House 0.0073	Care 0.0091	Reduce 0.0093
Change 0.0073	News 0.0069	Insurance 0.0087	Job 0.0085
Business 0.0045	Loss 0.0047	Bank 0.0082	Housing 0.0081
Worker 0.0043	Care 0.0041	Economy 0.0076	Cut 0.0073
Plan 0.0039	Debate 0.0038	People 0.0069	Euro 0.0068
Debt 0.0038	Obama 0.0034	State 0.0045	Labor 0.0054
Load 0.0034	Washington 0.0032	Romney 0.0038	Pay 0.0048
Romney 0.0032	Explain 0.0029	Fiscal 0.0035	Worsen 0.0045
National 0.0029	Call 0.0027	Promote 0.0031	Dollar 0.0039

TUCM		NetLDA	
Topic 1	Topic 2	Topic 1	Topic 2
Obama 0.0105	President 0.0126	People 0.0131	Health 0.0154
Barack 0.0101	State 0.0121	Plan 0.0102	Birth 0.0142
White 0.0089	Federal 0.0120	Insurance 0.0098	Reform 0.0123
House 0.0087	National 0.0118	Obama 0.0096	Money 0.0121
Bush 0.0063	Mexico 0.0093	Care 0.0083	Company 0.0109
Bill 0.0060	Policy 0.0082	Act 0.0064	Unemployment 0.0106
Poll 0.0049	Obama 0.0077	Banker 0.0062	Job 0.0104
Mitt 0.0034	Barack 0.0076	Administration 0.0056	Romney 0.078
Romney 0.0033	White 0.0065	Legislation 0.0053	Oppose 0.0076
Plan 0.0020	House 0.0062	Funding 0.0042	Family 0.0073

Topic-Link		UCT	
Topic 1	Topic 2	Topic 1	Topic 2
Unemployment 0.0185	Tax 0.0096	Care 0.0192	Economy 0.0201
Trend 0.0181	Cash 0.0090	Obama 0.0180	Financial 0.0183
Obama 0.0173	Rate 0.0087	Health 0.0167	Rise 0.0175
Job 0.0156	Loss 0.0079	Reform 0.0152	Market 0.0170
Worker 0.0142	Romney 0.0060	Public 0.0140	Increase 0.0166
Office 0.0109	Economy 0.0056	Insurance 0.0139	Investment 0.0159
Trillion 0.0098	Cut 0.0051	Benefit 0.0115	Loan 0.0132
Fall 0.0082	Policy 0.0045	Coverage 0.0108	Raise 0.0130
Romney 0.0067	Crisis 0.0036	Plan 0.0101	Market 0.0124
Education 0.0062	Job 0.0033	Quality 0.0089	Dollar 0.0103

topic words of $s = 0$ contain more noise, for example, in Topic 2 when $s = 0$, although it discusses some aspects of economics, it looks to be less compact than the one of $s = 1$.

To compare the discovered topics from UCT with the ones of other baselines (including CT, CCF, TUCM, NetLDA, and Topic-Link), we randomly sample several community topics and rank the terms based on their corresponding probabilities with respect to the community-topic mixture. The results are shown in Table V. We can observe that UCT is able to generate cohesive topics. For example, the first topic by UCT describes exactly "Obama Care," whereas the second topic by UCT is about the economic changes. The results by other baselines may represent somewhat cohesive topics; however, most of them are not quite "clean" due to the involvement of different topical terms. Hence, by explicitly distinguishing the user topics (which may contain noise) from the community topics, our proposed model performs better than the other baseline topic models.

7.2.3. Runtime Analysis. One major drawback with probabilistic models for social media analysis is their scalability. As discussed in Section 5.2.1, the time complexity of our proposed model is $O(K \cdot U \cdot V)$ for a single sampling process, where V is the dictionary size, K is the number of topics, and U is the number of users. Comparatively, the model CCF requires $O(K \cdot U \cdot C \cdot V)$ for a single iteration, where C represents the number of communities. This is due to the computation of the posterior probability $P(z|c, u, w)$ in the model. Another model, TUCM, requires $O(K \cdot U \cdot C \cdot V \cdot X)$ for a single iteration, where X denotes the number of interaction types for a single post of a user. NetLDA contains an extra process of evaluating the regularizer, which involves multiple iterations of Newton–Raphson updating in order to solve a linear system of $|U| \times (K+1)$ variables, which is quite complex. Topic-Link is performed based on each single post of users, and it requires $O(K \cdot U \cdot C \cdot M \cdot V)$, where M is the averaged number of posts per user. Apparently, without considering the involved constant, the time complexity of our model is lower than the ones of the baselines.

7.3. User-Profile-Based Recommendation

To evaluate the user-profile-based recommendation, we adopt the leave-one-out or leave-*n*-out strategy. Specifically, we randomly deleted one or more links between each user and his/her friends or links of this user to a community, and then, evaluate whether the deleted links could be recommended. *precision* and *recall* are used to measure the recommendation effectiveness, defined as follows:

$$\text{Precision} = \frac{|\{\text{recommended list}\} \bigcap \{\text{existing list}\}|}{|\{\text{recommended list}\}|}$$
$$\textit{Recall} = \frac{|\{\text{recommended list}\} \bigcap \{\text{existing list}\}|}{|\{\text{existing list}\}|}$$

where "existing list" represents the friends list or the community list that the user has connection with. *Precision* is calculated at a given cutoff rank, considering only the topmost results recommended by the approach, for example, top@30. As it is possible to achieve higher recall by recommending more results, we limit the size of our recommendation list to at most 30.

7.3.1. Recommending Users. We compare the user recommendation strategy introduced in Section 6.1 with several topic-model-based recommendation approaches: (1) UT: the recommendation can be achieved using the strategy similar to Eq. (6), by removing the components related to c; (2) CT: by considering the identified community membership, we can select a list of top ranked users, based on $p(u|c)$, from the community that the target user belongs to; (3) CCF: this method provides user recommendation by calculating the user similarity introduced in Chen et al. [2008]; (4) TUCM: the recommendation can be achieved using the strategy similar to Eq. (6); (5) Topic-Link: the recommendation can be obtained calculating the probability of a link using the link formation parameter derived from the training process, as shown in Liu et al. [2009].

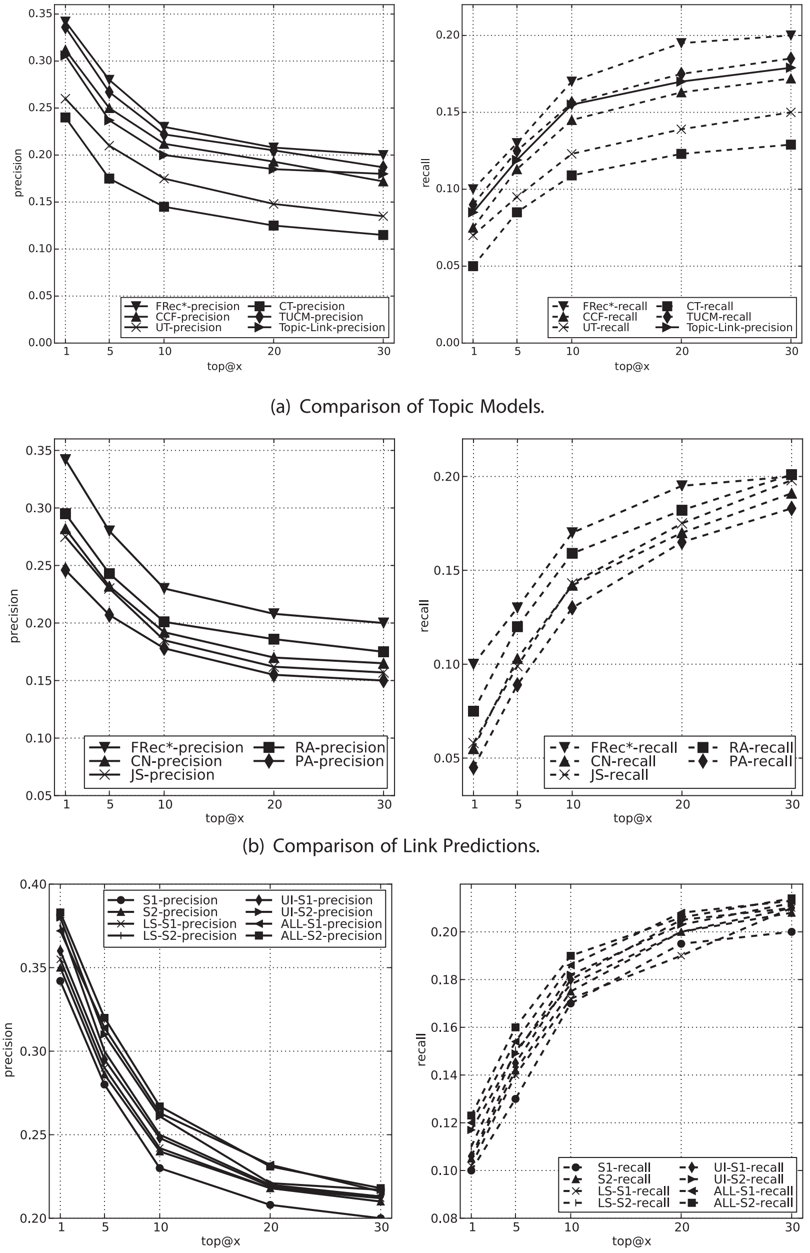
In addition to the aforementioned methods, we also compare several simple link prediction approaches [Yan and Gregory 2011] with FRec: (1) Common Neighbors (CN): the number of common neighbors that two vertices have is a basic idea that suggests a mutual relationship between them, that is, score$(u, v) = |\Gamma(u) \cap \Gamma(v)|$; (2) Jaccard Similarity (JS): score$(u, v) = |\Gamma(u) \cap \Gamma(v)|/|\Gamma(u) \cup \Gamma(v)|$; (3) Resource Allocation Index (RA) [Zhou et al. 2009]: it assumes that the common neighbors could transmit resources from one vertex to the other one, that is, score$(u, v) = \sum_{s \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(s)|}$; (4) Preferential Attachment (PA) [Newman 2001]: this shows that the probability of a connection between two arbitrary vertices is related to the number of neighbors of each vertex, that is, $\text{score}(u, v) = |\Gamma(u)| \cdot |\Gamma(v)|$.

Our goal is to select a list of users whose topic interests are close to the target user. By removing the user-oriented components from Eq. (6), we can make the recommendation results more community oriented, as defined in Eq. (7). Note that in Eq. (7), we consider the community information of both the target user $\hat{u}$ and the recommended user u_i. To this end, we randomly select 2,000 users from the user repository as the test set and randomly delete a set of links of each test user: (1) $S1$: removing 20% links; and (2) $S2$: removing 2% links. We conduct experiments based on these two setups. Figure 4 shows the averaged precision and recall for these users. For comparisons with topic models and link prediction methods, the experiments use setup $S1$; To evaluate the effect of different components in Eq. (10), we use setup $S1$ and $S2$.

From Figure 4(a) and 4(b), we observe that our proposed framework FRec achieves the best recommendation performance in terms of *precision* and *recall*. Simply using one information channel, for example, topics [UT in Figure 4(a)] or links [link prediction methods in Figure 4(b)] cannot guarantee high-quality recommendation results. If we only consider topics, then two users might share similar interests but they do not have connections in the social graph; If we only consider links, then two users might connect through several links but they do not share common topic interests. We also observe that the performance of some link prediction algorithms, for example, CN and RA, is comparable to, or even better than the one of topic models, for example, UT and CT. The reason here is straightforward: as we feed all the content generated by a user into the topic models, it is possible that the performance of these topic models could be affected due to the noise within the user content. Comparatively, link predictions are purely based on the friendships among users, which require the mutual appreciation and hence contain less noise compared with the user content. In addition, we observe that the performance of CCF, TUCM and Topic-Link is close to the one of FRec. However, some of them fail to explicitly distinguish the user topic from the community topic, and hence, the performance is relatively inferior to our proposed framework.

In Figure 4(c), we evaluate how user influence (UI) and users' local similarity (LS) affect the recommendation performance. We compare the basic model of FRec, the model with UI, the model with LS and the model with UI and LS for two different settings $S1$ and $S2$. Based on the comparison, we observer that: (1) User influence component and local similarity component slightly improved the performance of user recommendation. Intuitively, a user will prefer to make friends with influential people, since through these people he/she can reach more friends. Also, a user will be likely to interact with friends-of-friends. (2) The user recommendation has more accurate results if more social links of users are reserved. This is primarily because social links can help identify the underlying communities and then enrich the recommendation model through the user-community relations.

7.3.2. Recommending Communities. For community recommendation, we treat the communities identified from the module of community detection (Section 4) as the ground truth. We randomly sample 2,000 users from the user repository and recommend communities for these users. Note that in Eq. (11), we take into account the topic aspects, for example, user-oriented topics ($p(z|u, s = 0)$) and community-oriented topics ($p(z|c, s = 1)$), without considering the relations between users and communities. This setup indicates that the community recommendation can be obtained through the content analysis. The comparison includes: (1) FRec: the basic strategy described in Eq. (11); (2) FRec-s1: removing the factor of $p(z|u, s = 0)$ from Eq. (11), that is, only considering the community-oriented topics for recommendation; (3) FRec-IN: the strategy described in Eq. (12); and (4) FRec-IN-s1: removing the factor of $p(z|u, s = 0)$ from



(a) Comparison of Topic Models.

(b) Comparison of Link Predictions.

(c) Effect of Different Components.

Fig. 4. Comparison for profile-based user recommendation.

Eq. (12), that is, considering user influence and the community-oriented topic factor. We also compare FRec with several recommendation approaches, including CCF and TUCM as introduced previously. These two approaches use the inferred probabilities of $p(z|u)$ and $p(z|c)$ for community recommendation. We report the comparison in Figure 5.

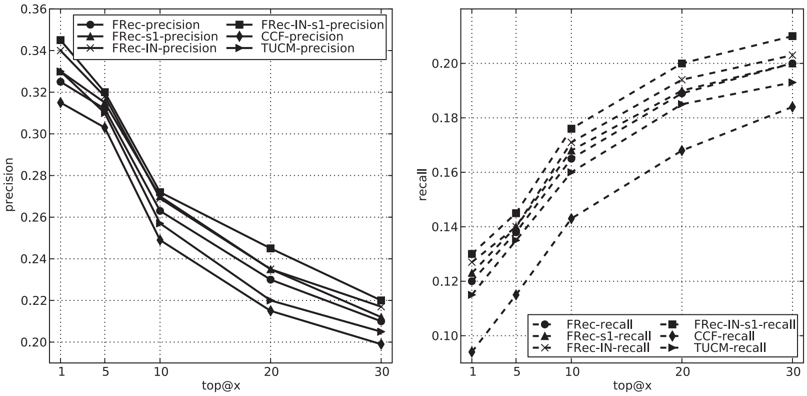


Fig. 5. Comparison for profile-based community recommendation.

As observed in Figure 5, the model of FRec-IN-s1 has the best performance against other baselines, which explains that users in social media would like to interact with influential users, and prefer to share information that is often discussed within a community, that is, by a group of people. The community-oriented topic factor $p(z|c, s = 1)$ has superior power over user-oriented topic factor $p(z|u, s = 0)$ in dominating the results of community recommendation.

7.4. Keyword-based Recommendation

Given a set of keywords, FRec is capable of recommending a list of influential users or a list of topic-cohesive communities. To validate keyword-based recommendation, we select ten sets of keywords related to "presidential campaign," and conduct case studies for each set of keywords. They all show similar trends on the results. We use "Obama, economy" as an example to compare the results with the ones provided by TUCM.

7.4.1. Recommending Users. For user recommendation, we focus on validating two aspects: (1) Is the recommended user relevant to the given keywords? and (2) Does the recommended user have more influential power to help enrich the interactions? To this end, we manually check the profiles of the top ranked recommended users, for example, how many followers that each user has, and select a list of posts for each user based on the inferred word probabilities ($p(w|z)$). Table VI shows the profiles of the top three recommended users.

From Table VI, we observe that: (1) The posts by the recommended users from FRec are more related to "Obama, economy," whereas the content of users by TUCM is more related to the marginal effect of the economy, for example, "missing from the job market," "earning less," and so on. (2) The number of followers of the users from FRec is higher than the ones from TUCM. In social media environment, we can simply treat the number of followers as a metric to evaluate the influence of a specific user. We checked the profiles of the recommended users through Twitter REST API. The screenname of the top first user by FRec is "JasonPlummer," who is currently the Vice President of Corporate Development at R.P. Lumber Company, a family-owned and operated business. Therefore, this person is more concerned with the economic policies discussed in presidential campaigns and is more worthy of following. Comparatively, we cannot obtain the identity of the recommended users by TUCM, even for the third user with 1,020 followers. Perhaps the third user is more influential in terms of the other topics, but not the one related to "Obama, economy."

Table VI. Top Ranked Users of "Obama, Economy"

User ID	User Posts	Followers
15924025 (FRec)	—Three year after Obama took office, #smallbiz is still struggling with the effect of the Obama economy. —Former Obama advisor : #smallbiz hit hardest by recession.	1,301
23712877 (FRec)	—This is the Obama economy: Millions of Americans is suffering in silence. —More bad news for Obama : Gallup has Romney winning on economy.	645
15700513 (FRec)	—Obama orders press blackout after US credit rating cut. —Obama stimulus dollars funded Soros Empire.	247
97041451 (TUCM)	—RT: 5 million Americans are missing from the job market,giving up finding work in the Obama economy. —RT: Obama's reelect is hiding the truth: billion of taxpayer $'s wasted on failed projects overseas.	30
110207321 (TUCM)	—New graduates earning less, owing more in Obama economy. —RT: Obama stimulus dollars funded Soros Empire.	502
111978318 (TUCM)	—#Obama says tax subsidies for oil companies should be use to invest in clean energy technology. —RT: Romney: "The real war on women has been the job losses as a result of the Obama economy."	1,020

Table VII. Top Ranked Users of "Obama, Economy" by # of Followers

User ID	User Posts	Followers
891991	—Perhaps this is Romney's clever new plan to revive the economy?	4,956
5484532	—RT: Gallup Today: 55% think economy will be better over the next 4 years if Romney wins, 46% if Obama.	3,577
15924025	—Three year after Obama took office, #smallbiz is still struggling with the effect of the Obama economy. —Former Obama advisor: #smallbiz hit hardest by recession.	1,301

Table VIII. Top Ranked Users of "Obama, Economy" without Considering User Influence

User ID	User Posts	Followers
150133790	—RT: Mitt Romney understands the economy, ready to promote long term fiscal growth. —RT: Romney has advantage over Obama because of sluggish economy.	539
397648178	—RT: This is the Obama economy: Millions of Americans is suffering in silence.	133
464896020	—RT: To given decent standard of living for the American people and restore the vibrancy of the US economy... —RT: We need to stop allowing secretive banking cartels to endlessly enslave us through monetary policy trickery.	31

We further looked into the recommendation results: ordering the top ten recommended users by the number of followers and selecting top three users to check if their posts are related to the given keywords. Table VII reports the results. We can see that the first two users have much more followers than the third user (the one ranked 1st in Table VI); however, their representative posts are not quite relevant to the keywords, or at least less relevant than the posts of the third user.

We also investigate the effect of user influence on the recommendation result. To do so, we remove the influence factor from Eq. (14), and check the profiles of the recommended users. Table VIII shows the results of top three users. We can observe that although the content posted by these users is somehow relevant to the given query, they have very limited followers. This observation, to a certain extent, demonstrates that intentionally considering user influence over topics can help recommend users with more influential power.

Table IX. The First Community of "Obama, Economy"

	Topics	
FRec	Topic1	economy highest require choice make money required fee Obama struggle attack deficit earn corporate pay appeal debt rich system economic tax plan return labor worsen...
	Topic2	great economy depression grow unemployment recession job Obama investment responsibility focus state trillion investor fall buddy typical money dollar...
	Topic3	wall street loss Obama market crisis romney economy reform side stock economic decline JPMorgan euro campaign increase raise program banker business...
TUCM	Topic1	tax rate lower cut spending economy reduce loan debt job level iowa education financial pres rise economic jobless housing Obama recovery recent program budget development deduction...
	Topic2	grow wealth country people economy Obama give trickle government choice economics problem future political reason financial bail hard excuse job reveal...

Table X. Comparison of Influence Power

	Averaged # of Followers		
	First community	Second community	Third community
TUCM	1,530	1,895	1,067
FRec-NoIN	1,923	1,130	1,752
FRec	2,563	2,124	1,923

7.4.2. Recommending Communities. For keyword-based community recommendation, we are concerned with three aspects: (1) Is the recommended community relevant to the given keywords? (2) Does the recommended community have more cohesive topics? and (3) Does the recommended community have more influential power? To answer these questions, we select the top ranked community from FRec and TUCM, and manually investigate the topics within the community. Table IX shows the top ranked topics (representative words) based on $p(z|c)$ (Note that for FRec, we consider $p(z|c, s = 1)$), in which the colored terms are relevant to the given keywords.

As depicted in Table IX, we list the top three topics for the FRec community and top two topics for the TUCM community. It is trivial to see that all the topics within the communities are relevant to the given keywords; they all describe some aspects related to "Obama, economy." For example, the second topic from the FRec community discusses how the depression on economy influences the investment and consequently, results in growing unemployment. We further observe that the topics of the FRec community are more cohesive compared with the ones of the TUCM community. Each of the three topics of FRec describes one aspect related to "Obama, economy"; Comparatively, the topical aspects of each topic of the TUCM community are more scattered, for example, the first topic of TUCM is related to "debt," "job," and "housing." Therefore, FRec is able to recommend relevant and topic-cohesive communities based on the given keywords.

We further investigate the influence power of communities by calculating the averaged number of followers within each community. To do so, we allow each user to belong to at most three communities, and then, select the top ranked three communities recommended to "Obama, economy" for analysis. To reduce noise, we exclude users with abnormally high amount of followers by setting a threshold of 10,000. We compare FRec with FRec-NoIN [removing influence factor from Eq. (15)] and TUCM. The results are shown in Table X. We observe that the averaged number of followers in the communities recommended by FRec is more than the two baselines. This, to some extent, shows that FRec can recommend communities with more influence power.

Table XI. The Evaluation Criteria of User Study

Tasks	Criteria	
Recommending Users	Relevant	Which model recommends more relevant users?
	Influential	Which model recommends more influential users?
Recommending Communities	Relevant	Which model recommends more relevant communities?
	Cohesive	Which model recommends communities with more cohesive topics?
	Influential	Which model recommends more influential communities?

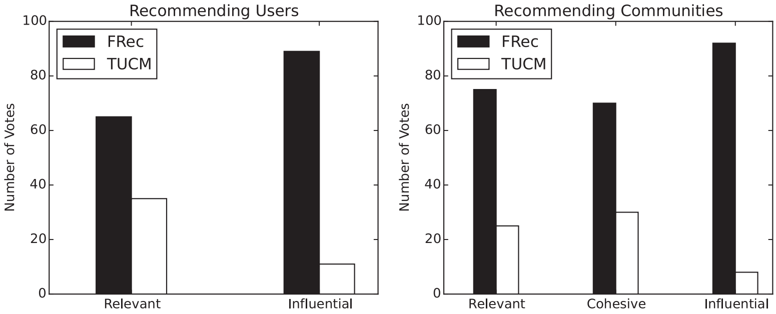


Fig. 6. User study on keyword-based recommendation.

7.4.3. A User Study on Keyword-based Recommendation. In the aforementioned, we have qualitatively examined and compared the performance of the models FRec and TUCM with respect to top recommended users and communities. Due to the lack of ground truth for keyword-based recommendations, it is difficult to provide a quantitative evaluation for the tasks of recommending users and communities based on the input keywords. To further demonstrate the efficacy of our proposed recommendation framework, we conduct a formal user study on the models of FRec and TUCM. Specifically, we hire five graduate students who have daily usage of Twitter to conduct this study. Each student is asked to feed 20 queries related to the presidential campaign into our system, and then manually examine the recommended users and communities based on specific criteria. The criteria are summarized in Table XI.

Each student is then asked to vote the models based on the criteria in Table XI for the queries he/she provides. Throughout the use study, we collect 100 queries in total, and hence for each criterion, we collect 100 votes. We plot the number of votes for both models, and the result is shown in Figure 6. From the comparison, we observe that our proposed model is superior to TUCM of all the criteria being considered in the study. In particular, FRec provides influential users and communities based on the input keywords, which is beneficial for expanding the social network. In addition, based on the feedback of subjects, our model is able to recommend communities with more cohesive topics.

8. CONCLUDING REMARKS

We have introduced a generative graphical model, UCT model, for capturing user-oriented topics and community-oriented topics simultaneously in social media data. UCT employs a Bernoulli variable to distinguish these two types of topics in the generative process of the model. Based on the model inference, we further proposed a novel recommendation framework, FRec. Given a user's profile or a set of keywords as input, FRec is able to recommend a list of topic-related influential users or a list of topic-cohesive interactive communities. Experiments on a Twitter dataset demonstrated the effectiveness of our proposed topic model and the recommendation framework. Our framework gives us the capability of recommending users and communities in most social media environments with implicit community memberships. The significant reduction in training time of the topic model and the elegant recommendation strategy enable the framework to be seamlessly integrated into real-life social networks.

There are several directions for future research. First, since the data in social media is very noisy, it is interesting if we incorporate a background distribution [Chemudugunta and Steyvers 2007] relevant to social media data into the topic modeling process. Such background distribution would be able to help filter the noise within the data, for example, some words specific to social media environment. Second, in our analysis, we only consider the topics discussed by users. An interesting extension would be to consider the sentiment of users toward topics, since in most cases, people tend to discuss with each other if they have similar opinions toward a specific topic.

REFERENCES

Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. 2006. Link prediction using supervised learning. In *SDM06: Workshop on Link Analysis, Counter-Terrorism and Security*. 1–10.

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 27–34.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P. Krishna Gummadi. 2010. Measuring user influence in Twitter: The million follower fallacy. *ICWSM* 10, 10–17.

C. Chemudugunta and P. S. M. Steyvers. 2007. Modeling general and specific aspects of documents with a probabilistic topic model. In *Advances in Neural Information Processing Systems*, Vol. 19. MIT Press, Cambridge, MA, 241–248.

W. Y. Chen, J. C. Chu, J. Luan, H. Bai, Y. Wang, and E. Y. Chang. 2009. Collaborative filtering for orkut communities: Discovery of user latent behavior. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, 681–690.

W. Y. Chen, D. Zhang, and E. Y. Chang. 2008. Combinational collaborative filtering for personalized community recommendation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Las Vegas, NV, USA, 115–123.

Pedro Domingos and Matt Richardson. 2001. Mining the network value of customers. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*. ACM, New York, NY, USA, 57–66. DOI:http://dx.doi.org/10.1145/502512.502525

S. Fortunato. 2010. Community detection in graphs. *Phys. Rep.* 486, 3, 75–174.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* 101, 5228–5235.

Taher H. Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web*. ACM, 517–526.

G. Heinrich. 2005. Parameter estimation for text analysis. *Retrieved from http://www.arbylon.net/publications/textest*.

David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*. ACM, New York, NY, USA, 137–146. DOI:http://dx.doi.org/10.1145/956750.956769

Page Lawrence, Brin Sergey, Rajeev Motwani, and Terry Winograd. 1998. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical report. Stanford University.

Jingxuan Li, Wei Peng, Tao Li, Tong Sun, Qianmu Li, and Jian Xu. 2014. Social network user influence sense-making and dynamics prediction. *Expert Sys. Appl.* 41, 11 (September 2014), 5115–5124.

Lei Li, Wei Peng, Saurabh Kataria, Tong Sun, and Tao Li. 2013. FRec: A novel framework of recommending users and communities in social media. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*. ACM, 1765–1770.

D. Liben-Nowell and J. Kleinberg. 2007. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* 58, 7, 1019–1031.

Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. 2009. Topic-link LDA: Joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 665–672.

L. Lü and T. Zhou. 2011. Link prediction in complex networks: A survey. *Physica A, Stat. Mech. Appl.* 390, 6, 1150–1170.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit.

Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web*. ACM, 101–110.

Mark E. J. Newman. 2001. Clustering and preferential attachment in growing networks. *Phys. Rev. E* 64, 2, 1–13.

Wei Peng and Tao Li. 2011. Temporal relation co-clustering on directional social network and author-topic evolution. *Know. Inf. Syst.* 26, 3, 467–486.

M. Pennacchiotti and S. Gurumurthy. 2011. Investigating topic models for social media user recommendation. In *Proceedings of the 20th International Conference Companion on World Wide Web*. ACM, 101–102.

Alexandrin Popescul and Lyle H. Ungar. 2003. Statistical relational learning for link prediction. In *Proceedings of the IJCAI Workshop on Learning Statistical Models From Relational Data*. 1–7

M. A. Porter, J. P. Onnela, and P. J. Mucha. 2009. Communities in networks. *Notices of the AMS* 56, 9, 1082–1097.

D. Ramage, S. Dumais, and D. Liebling. 2010. Characterizing microblogs with topic models. In *Proceedings of the 4th International Conference on Weblogs and Social Media*. 130–137.

Christian P. Robert and George Casella. 2013. *Monte Carlo Statistical Methods*. Springer Science & Business Media.

Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. 2010. Learning author-topic models from text corpora. *ACM Trans. Inf. Sys.* 28, 1, 1–38.

Maayan Roth, Assaf Ben-David, David Deutscher, Guy Flysher, Ilan Horn, Ari Leichtberg, Naty Leiser, Yossi Matias, and Ron Merom. 2010. Suggesting friends using the implicit social graph. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 233–242.

M. Sachan, D. Contractor, T. A. Faruquie, and L. V. Subramaniam. 2012. Using content and interactions for discovering communities in social networks. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, Lyon, France, 331–340.

D. Seung and L. Lee. 2001. Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* 13, 556–562.

Ben Taskar, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller. 2003. Link prediction in relational data. In *Advances in Neural Information Processing Systems*. 1–8.

Fei Wang, Tao Li, Xin Wang, Shenghuo Zhu, and Chris Ding. 2011. Community discovery using nonnegative matrix factorization. *Data Min. Knowl. Discovery* 22, 3 (May 2011), 493–521.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. ACM, 261–270.

Bowen Yan and Steve Gregory. 2011. Finding missing edges and communities in incomplete networks. *J. Phys. A* 44, 49, 1–15.

T. Yang, R. Jin, Y. Chi, and S. Zhu. 2009. Combining link and content for community detection: A discriminative approach. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 927–936.

Zhijun Yin, Liangliang Cao, Quanquan Gu, and Jiawei Han. 2012. Latent community topic analysis: Integration of community discovery with topic modeling. *ACM Trans. Intell. Syst. Technol. (TIST)* 3, 4, 1–21.

K. Yu, S. Yu, and V. Tresp. 2006. Soft clustering on graphs. *Adv. Neural Inf. Process. Syst.* 18, 1553–1560.

H. Zhang, B. Qiu, C. L. Giles, H. C. Foley, and J. Yen. 2007. An LDA-based community structure discovery approach for large-scale social networks. In *Proceedings of IEEE International Conference on Intelligence and Security Informatics*. IEEE, New Brunswick, USA, 200–207.

D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha. 2006. Probabilistic models for discovering e-communities. In *Proceedings of the 15th International Conference on World Wide Web*. ACM, Edinburgh, Scotland, 173–182.

D. Zhou, B. Schölkoph, and T. Hofmann. 2005. Semi-supervised learning on directed graphs. *Adv. Neural Inf. Process. Syst.* 17, 1633–1640.

T. Zhou, L. Lü, and Y. C. Zhang. 2009. Predicting missing links via local information. *Eur. Phys. J. B, Condens. Matter Complex Syst.* 71, 4, 623–630.

Received September 2013; revised December 2014; accepted April 2015