

User community discovery from multi-relational networks

Zhongfeng Zhang^{a,1}, Qiudan Li^{a,*}, Daniel Zeng^{a,b,2}, Heng Gao^{a,1}

^a State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

^b Department of Management Information Systems, University of Arizona, Tucson, AZ, USA

ARTICLE INFO

Article history:

Received 18 July 2010

Received in revised form 2 April 2012

Accepted 18 September 2012

Available online 28 September 2012

Keywords:

Community discovery

Multi-relational network

Author topic model

Non-negative matrix factorization

ABSTRACT

Online social network services (SNS) have been experiencing rapid growth in recent years. SNS enable users to identify other users with common interests, exchange their opinions, and establish forums for communication, and so on. Discovering densely connected user communities from social networks has become one of the major challenges, to help understand the structural properties of SNS and improve user-oriented services such as identification of influential users and automated recommendations. Previous work on community discovery has treated user friendship networks and user-generated contents separately. We hypothesize that these two types of information can be fruitfully integrated and propose a unified framework for user community discovery in online social networks. This framework combines the author-topic (AT) model with user friendship network analysis. We empirically show that this approach is capable of discovering interesting user communities using two real-world datasets.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Web 2.0 technology has enabled massive online social networks and made sharing of user-generated contents easy and almost costless. Two-thirds of Americans now use Facebook, Twitter, Myspace, and other social media sites; and 43% are visiting these sites more than once a day.³ By May 2010, social networks have become more popular than search engines in U.K., accounting for 11.88% of all U.K. Internet visits.⁴ Usually, a social network involves multiple types of relations among different social actors. For instance, on Twitter, a user can specify whom to follow to construct an explicit friendship network. At the same time, this user's posted tweets provide important clues about her interests and such interests across the user community can be used to derive implicit "similarity" relationships among these users. The network embedding multiple types of relations, either explicit or implicit, is called a multi-relational network. Studying multiple relationships is gaining momentum in the literature recently. A case study of homophily on LiveJournal [1] has shown that users' friendships and interests are strongly interlinked. Hernandez [4] introduced some social network principles in online communities. Researchers have also investigated how to combine the friendship network and user-generated contents to cluster users [16,29].

User community changes the way people communicate and affects social interaction [37,38]. Discovering user communities may assist the setup of efficient recommender systems for targeted marketing, improving the quality of social information retrieval, among others [3,25–27,30,40]. For instance, Nie et al. [30] utilized the relevance of communities to improve web page ranking. Online user communities have also emerged as a thriving force in e-Commerce [2]. Spaulding [25] explained how firms could successfully interact with user communities using social contract and trust theory. Ganley et al. [4] examine a popular website Slashdot to test users' social network structure, which would potentially increase the opportunities for monetization. Chiu et al. [27] investigate people's knowledge sharing behavior in virtual communities to help identifying their motivations in communities. In [3], the authors investigated how consumers take advantage of virtual communities as social and information networks, and how this influences their decision making. The identified user communities can also help understand the structural properties of the social network and find the influential users about certain topics, which in turn will help users locate the latent friends they may be interested in.

Most prior work on user community detection has focused on analyzing either user friendship networks [6–8,17] or user-generated contents [12,14] but not both at the same time. The former techniques usually ignore the content generated by users. However, intuitively, two users who have posted similar contents might share common interests and join the same communities, even if no explicit friendship connection exists between them. On the other hand, the latter strategies do not take the friendship connections among users into consideration. Such explicit friendship networks can provide important clues to community discovery.

Being friends "makes a pair of users more likely to share common interests" [1]. In the study of recommender systems, it was shown

* Corresponding author. Tel.: +86 10 62558794.

E-mail addresses: zhongfeng.zhang@ia.ac.cn (Z. Zhang), qiudan.li@ia.ac.cn (Q. Li), zeng@email.arizona.edu (D. Zeng), heng.gao@ia.ac.cn (H. Gao).

¹ Tel.: +86 10 62636334.

² Tel.: +1 520 621 4614.

³ <http://www.socialnetworkingwatch.com/2010/06/social-media-up-230-since-2007.html>

⁴ <http://eu.techcrunch.com/2010/06/08/report-social-networks-overtake-search-engines-in-uk-should-google-be-worried/>

that more accurate recommendations could be made by taking into account both friendship networks and user-generated contents [31]. Similar strategies were also proven to be effective in document retrieval [32] and document classification [33]. This research stream suggests the practical value of a multi-relational approach. In the context of community discovery, work on multi-relational approaches is recently emerging (e.g., [16,28,29]).

This paper focuses on the problem of discovering user communities from multi-relational networks of SNS. We present a unified framework, which combines the author-topic (AT) model with social network analysis (SNA). The AT model, which deals with user-generated content information, is a domain sensitive model, while the SNA methods focus on user friendship networks. Users in the community identified with our approach have dense friendship connections as well as share common content interest. The efficacy of the proposed framework is evaluated using two real-world social network datasets, one from Delicious, a popular social bookmarking site, and the other from Twitter, the most popular microblogging site. Empirical analyses have shown that our algorithm could discover meaningful communities and the topics discussed by these communities in a unified way. Compared to the state-of-the-art, our new framework has resulted in comprehensive performance of closer friendship and higher content interest similarity in the extracted communities.

The rest of the paper is organized as follows. The literature review is presented in Section 2. Section 3 presents in detail the problem definition and our framework. Section 4 introduces the detailed algorithm. The empirical analysis is conducted in Section 5. Finally, Section 6 concludes the paper with a summary and discussion of the future work.

2. Literature review

Most user community detection methods fall into two categories, the network-based and the content-based. The network-based methods construct network structures among users and then split the network into different sub-networks. Technique-wise, these methods are based on graph partitioning in graph theory. The content-based methods discover users with common interests by analyzing the similarity between these users' posted contents. In this section, we review both types of methods.

2.1. Network-based methods

In many social networks, individuals form communities by specifying and establishing friendship connections with each other. The network-based methods aim to find communities such that the friendship connections are dense within communities and sparse between them. Traditional graph partition methods, such as degree-based methods and max-flow min-cut methods, are used to divide the network into groups of predefined size, such that the number of connections lying between the groups is minimal [5]. Spectral clustering techniques partition the network into clusters using the eigenvectors of its related matrices (e.g., Laplacian matrix) [5]. The GN [6] method selects links among users according to edge centrality. Palla et al. [8] proposed a clique percolation method (CPM) based on the concept that the internal links of a community are likely to form cliques due to their high density. Shen et al. [9] proposed to identify overlapping community structures from the maximal clique network of the original network, using modularity optimization methods. Evans et al. [10] introduced a link partition approach for overlapping community structure discovery. Lee and Seung [18] firstly investigated the algorithm of non-negative matrix factorization and it became widely used soon afterwards. Zhang et al. [11] proposed to discover fuzzy community structures in complex networks based on non-negative matrix factorization (NMF). The complex network theory was applied to analyze open-source software systems and structural properties of social interaction in collaborative tagging systems, respectively [41,42].

2.2. Content-based methods

The content-based methods link users and their posted contents via latent topics. Users interested in the same topic are grouped into a community. Steyvers et al. [12] proposed the author-topic (AT) model to explore the relationships among users, documents, topics, and words. It represents a topic as a multinomial distribution over words and models a user as probability distribution over different topics. McCallum et al. [13] presented the author-recipient-topic (ART) model to discover users with similar topic interests, which conditions the topic distribution on the sender-recipient relationships. Based on the ART model, Pathak et al. [15] introduced a community-author-recipient-topic (CART) for community extraction from the Eron email corpus, by leveraging both topic and document link information from the social network. Peng et al. [43] proposed a unified user profiling scheme which makes good use of all types of co-occurrence information in the tagging data. Relying on people's information in database, [39] developed an intelligent secretary agent system to help arrange efficient meetings among people who share similar interests. These models often ignore the explicit friendship connections among users, and may not properly predict users' community memberships.

In this work, we propose a new framework which utilizes both friendship networks and content analysis to discover user communities. At the core of this framework is the NMF-AT algorithm, which performs matrix factorization on the friendship network and author-topic analysis on the user-generated contents. One research closely related to ours combines topic modeling with network regularization [16]. In their work, pLSA was adopted for topic extraction and the graph harmonic function was used for community analysis. Finally, the topical community was extracted by performing topic mapping. However, the authorship of contents was not taken into consideration during the topic extraction procedure in [16]. Instead of performing topic mapping, we tend to extract community topics directly with author-topic analysis.

3. Problem definition

In this section, we first define the terminologies related to community discovery. We then present the framework of our approach for discovering user communities from multi-relational networks.

3.1. Terminology definition

Fig. 1(a) and (b) show examples of multi-relational networks from Twitter and Delicious, respectively. In Twitter, each user is called a twitterer, who can post tweets with a limit of 140 characters, or reply tweets posted by her friends. Each twitterer could follow any other twitterer she is interested in without securing permission. Conversely, she may also be followed by other twitterers. On Delicious, a user could bookmark a url with his own tags, and add interested users into his network to make friends. The friendship networks are marked with dashed lines in Fig. 1.

Given a multi-relational network as shown in Fig. 1, we represent it as a graph $G = (V, E)$, where V is the set of actors in the network, and E is a set of edges indicating the connections among actors in V . For instance, in Fig. 1(a), $V = \{U, T, R, W\} = \{<U1, \dots, U5>, <T1, \dots, T4>, <R1, R2>, <w1, w2, w3>\}$, where U_i ($i = 1, \dots, 5$) represents a twitterer, T_j ($j = 1, \dots, 4$) a tweet, w_k ($k = 1, 2, 4$) a word in the vocabulary, and $R1/R2$ reply to other tweets (which is also called a tweet). $E = \{<U1, U2>, \dots, <U1, T1>, \dots, <T1, w1>, \dots\}$ indicates the relationship among twitterers, tweets, and words. The edge $<U1, U2>$ indicates that twitterer $U1$ has followed $U2$. The edge $<U1, T1>$ implies the twitterer $U1$ has posted tweet $T1$. The edge $<T1, w1>$ indicates that tweet $T1$ is composed of word $w1$.

The friendship network associated with G is a subgraph $F = \langle U, Eu \rangle$, where $Eu \subseteq E$ is a set of edges among users. In the twitter case, $Eu = \{<U1, U2>, \dots, <U5, U4>\}$. The fact that a twitterer $U1$ follows $U2$ does not necessarily imply that $U2$ has followed $U1$. In such a

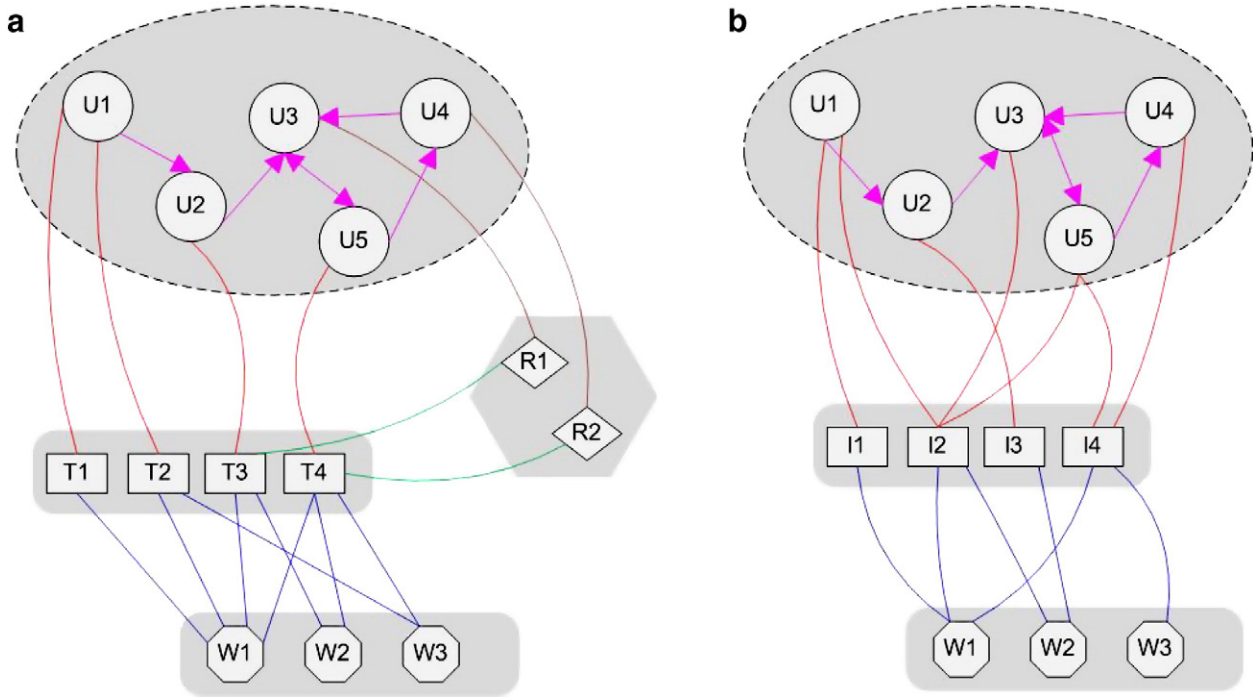


Fig. 1. Multi-relational network representation of interactions in (a) Twitter and (b) Delicious. (a) Interaction process on Twitter, where U_i ($i = 1, \dots, 5$) represents user i , T_j ($j = 1, \dots, 4$) tweets posted by users, w_k ($k = 1, 2, 3$) the vocabulary words, $R1$ and $R2$ the reply to tweets. (b) Interaction process on Delicious, where U_i ($i = 1, \dots, 5$) represents user i , I_j ($j = 1, \dots, 4$) the url of web pages, t_k ($k = 1, 2, 3$) the tags used to label urls.

scenario, the friendship network is a directed network, and is represented by directed graph F . For a directed network, the corresponding adjacent matrix M is asymmetric, with $M_{ij} = 1$ indicating user j has marked user i as friend. Previous work in [36] has shown that incorporating the direction information contained in edges will allow us to make more accurate determination of the community structures.

The notations used throughout the paper are summarized in Table 1.

With the graph-theoretic representation of the multi-relational network in a SNS, we have the following definitions:

Definition 1. User community

A user community is a group of users who form dense friendship connections during their participation in a specified topic. It is represented as $C = \langle \theta_C, \Omega_C \rangle$, where θ_C is the community distribution over users and Ω_C is the community topic distribution over words. A higher θ_{uc} implies user u gains higher authority in community C , while a higher Ω_{wc} indicates the topic of C can be better described using word w .

In our definition, a community is described using two factors, i.e. the user composition θ_C and the community topic Ω_C . Since θ_C and Ω_C are unique for each community, for simplicity, we will hereafter also use θ_C or Ω_C to represent a community.

Definition 2. Community structure

The community structure of a network G is the set of identified communities in G , represented as $S = \langle C1, C2, \dots, C\gamma \rangle = \langle \theta, \Omega \rangle$, where γ is the number of communities, $\theta = [\theta_{C1}, \dots, \theta_{C\gamma}]$, $\Omega = [\Omega_{C1}, \dots, \Omega_{C\gamma}]$.

Given the community structure in G , the community membership for user u is defined as:

Definition 3. Community membership

Community membership for user u is the probability of user u being interested in each community, represented using ϕ_u .

3.2. Overall solution framework

Fig. 2 shows the overall framework of the proposed approach. The input is a multi-relational network constructed from social network services. The output is detected community structures and community memberships. The approach itself consists of the friendship network analysis part and the content analysis part. For friendship network analysis, the social network analysis is performed, and the community distribution over users θ and community membership ϕ are obtained. For content analysis, a content-based community detection model is used to investigate the user-generated contents, with ϕ used as prior knowledge. The community topics Ω and users' authorities to these topics φ are derived in this step. Finally, user-topic

Table 1
Notations.

G	Graph representation of a multi-relational network
F	Friendship network embedded in G .
M	Adjacent matrix representation of the friendship network
θ	Community distribution over users
Ω	Community topic distribution over words
ϕ	Community membership for users
φ	Users' involvement in community topics
W	Vocabulary set of the corpus
w_d	Words in the d th document
w_{id}	The i th word in the d th document
D	Number of documents in the corpus
N_d	Number of words in the d th document
γ	Number of communities
U	Users in the corpus
U_d	Users associated with a document
α	Dirichlet prior for φ
β	Dirichlet prior for Ω
λ	Adjustable weight factor
n	Number of users

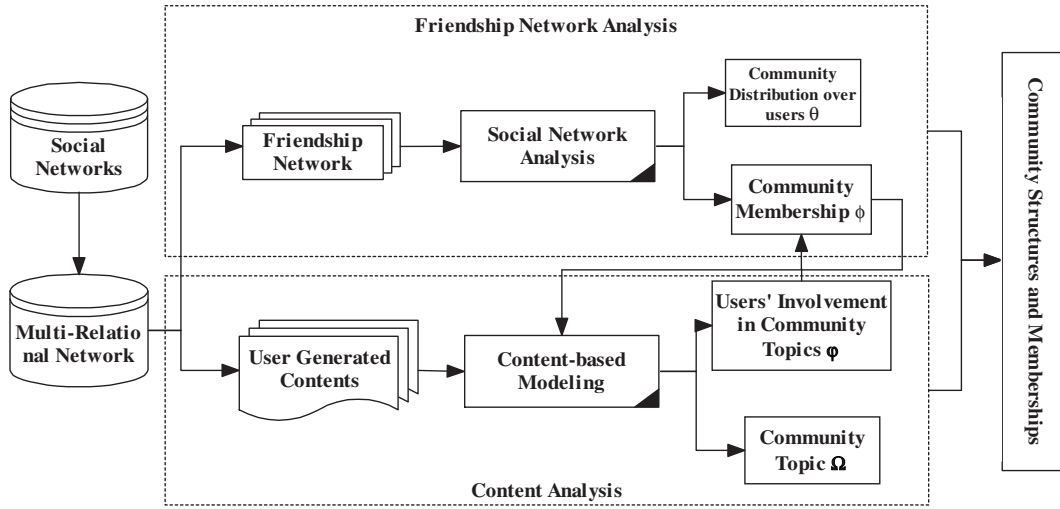


Fig. 2. Solution framework.

distribution φ is combined with ϕ to compute the final community memberships for individual users.

4. User community discovery algorithm

In this section, we present in detail the proposed NMF-AT method based on the framework discussed in Section 3. We have chosen NMF as the social network analysis method due to its two appealing characteristics: rapid convergence, which makes NMF suitable for analyzing large scale networks; and the sparsely distributed representation of the results, which can naturally be interpreted as users' community memberships. The AT model is employed to analyze user-generated contents. By explicitly modeling the relationship between users and topics, the AT model groups users with common interests into the same community and represent the community topic with a list of keywords.

4.1. NMF for friendship network analysis

Given the adjacency matrix representation M of a user friendship network, the basic idea of NMF is to construct approximate factorization of the following form

$$M \approx \theta \cdot \phi.$$

θ and ϕ are of dimension $n \times \gamma$ and $\gamma \times n$, respectively, where γ is the number of communities. NMF is unique in that it does not allow negative entities in the matrices. By normalizing the columns of θ and ϕ , $\theta_k = p(u|z=Ck)$ indicates the probabilistic distribution over users of community Ck , and ϕ indicates the community membership of user i . The factorization on M can be explained as follows: the adjacent vector for each user M_i is approximated by a linear combination of θ weighed by her community membership ϕ , i.e., $M_i \approx \theta \cdot \phi$. Fig. 3 shows a network-based illustration of NMF. The top layer represents community $\theta_1, \dots, \theta_r$ and nodes in the bottom layer are users u_1, \dots, u_n (column of M).

The computation of NMF can be stated as solving a minimization problem on the squared error:

$$\min_{\theta, \phi} E(\theta, \phi) = \min_{\theta, \phi} \|M - \theta\phi\|^2 = \min_{\theta, \phi} \sum_{ij} (M_{ij} - (\theta\phi)_{ij})^2. \quad (1)$$

The parameters θ and ϕ can be computed by iteratively updating based on the multiplicative rule [19]:

$$\phi = \frac{\phi\theta^T M}{\theta^T \theta\phi + \varepsilon} \quad (2)$$

$$\theta = \frac{\theta M \phi^T}{\theta\phi\phi^T + \varepsilon} \quad (3)$$

where $\varepsilon = 10^{-9}$, is a small positive parameter to avoid division by zero. The multiplicative method ensures that the lost function (1) converges to the local minimum efficiently.

4.2. AT model for content analysis

In the context of content analysis, the AT model assumes that the community topics can be modeled as multinomial distribution over words [23], represented as Ω , and that a user's interests in different topics is a probabilistic distribution over community topics, represented as φ . The topic structure of a document is a mixture of the topic distribution associated with users. With these assumptions, the process of generating a document is as follows: firstly, a user is chosen at random for each word token in the document; then, a community topic is picked from the user's topical interests; finally, a word is sampled from the multinomial distribution over words of the selected community topic. This sampling process is repeated N_d times to form document d . The probability of the corpus contents conditioned on φ and Ω can be calculated as:

$$P(w|\varphi, \Omega, U) = \prod_{d=1}^D P(w_d|\varphi, \Omega, U) = \prod_{d=1}^D \prod_{i=1}^{N_d} P(w_{di}|\varphi, \Omega, U_d) = \prod_{d=1}^D \prod_{i=1}^{N_d} \frac{1}{|U_d|} \sum_{a \in U_d} \sum_{k=1}^r \varphi_{w_{di} a} \Omega_{ka} \quad (4)$$

where w is word tokens in the content of the dataset; w_d is the d th document; U_d is the user set associated with document d . Parameters φ and Ω can be estimated by maximizing Eq. (4). Instead of estimating the model parameters directly, we evaluate the posterior distribution over z (the assignment of word tokens to community topic) and user u to infer φ and Ω using Gibbs sampling algorithm. The

community topic and user assignment for each word are sampled from:

$$p(z_i = k, x_i = u | w_i, z_{-i}, x_{-i}, \alpha, \beta, U) \propto \Omega_{wk} \varphi_{uk} \quad (5)$$

$$= \frac{C_{wk}^{W\gamma} + \beta}{\sum_w C_{wk}^{W\gamma} + |U|\beta} \frac{C_{ku}^{\gamma U} + \alpha}{\sum_{u' \in U} C_{ku'}^{\gamma U} + |U|\alpha}$$

where $z_i = k$ and $x_i = u$ indicate that the i th word is assigned to community topic k and user u respectively, while z_{-i} and x_{-i} are community topic and user assignment for all other word tokens. α and β are prior parameters for Dirichlet distribution.

From Eq. (5), φ and Ω can be estimated as follows:

$$\Omega_{wk} = \frac{C_{wk}^{W\gamma} + \beta}{\sum_w C_{wk}^{W\gamma} + |U|\beta}, \quad \varphi_{uk} = \frac{C_{ku}^{\gamma U} + \alpha}{\sum_{u' \in U} C_{ku'}^{\gamma U} + |U|\alpha}. \quad (6)$$

4.3. NMF-AT for user community discovery

We now present our NMF-AT approach which integrates both the user generated content and friendship information. The community membership information ϕ from F is used as prior knowledge for content analysis. Given ϕ as prior knowledge, the AT model can be trained by maximizing Eq. (7).

$$P(w|\varphi, \Omega, U, \phi) = \prod_{d=1}^D P(w_d|\varphi, \Omega, U, \phi) = \prod_{d=1}^D \prod_{i=1}^{N_d} P(w_{id}|\varphi, \Omega, U_d, \phi) \quad (7)$$

where ϕ is prior knowledge, and φ and ϕ are of the same dimension.

We can obtain the community topics Ω and users' interests in the community topics φ from Eq. (7). The final community membership for each user is calculated by integrating φ into ϕ , to take content information into consideration. Various integration strategies can be employed, including linear regularization, multiplication, among others. We have experimented with several of these strategies and selected a non-linear strategy because of its outstanding performance. Under this strategy, the combined community membership for a user u within a community k is linearly proportional to the membership derived from his friendship network, and exponentially proportional to his topical interests. The combined community membership matrix is calculated as:

$$\phi' = \frac{\phi \theta^T M}{\theta^T \theta \phi + \varepsilon} \circ \exp(\lambda \varphi) \quad (8)$$

where λ is the adjustable weighting factor, and $A \circ B$ is the Hadamard product between matrices A and B . We note that the detected community membership is identical with original NMF model when $\lambda = 0$.

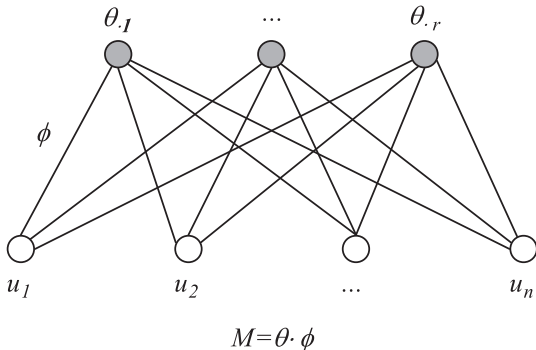


Fig. 3. A network illustration of NMF for community detection.

There are two major compositions in the combined community membership. The first one is derived from the friendship network, which plays a major role in determining the users' community activity. The second one measures the users' interest in a community from her content information. It is the exponential of the users' interest in the community topics, such that users who have posted a lot about a topic would be given more weight and regarded as experts in the community. The weight factor λ is used to adjust the importance of the content information in determining the authority of a user in a community.

An iterative procedure is adopted for parameter estimation of the NMF-AT model. To solve Eq. (7), the Gibbs sampling method is employed. Given ϕ as prior knowledge, the community topic and user assignment for word token w_i can be calculated as:

$$p(z_i = \Omega_{.k}, x_i = \varphi_{.u} | w_i, z_{-i}, x_{-i}, \alpha, \beta, U, \phi) \propto \Omega_{wk} \varphi_{uk} \phi_{ku} \quad (9)$$

$$= \frac{C_{wk}^{W\gamma} + \beta}{\sum_w C_{wk}^{W\gamma} + |U|\beta} \frac{C_{ku}^{\gamma U} + \alpha}{\sum_{u' \in U} C_{ku'}^{\gamma U} + |U|\alpha} \phi_{ku}.$$

The pseudo-code for parameter estimation is given in Algorithm 1.

Algorithm 1. Parameter estimation for the NMF-AT model

Input: The multi-relational network G ; Number of communities γ .

Output: Extracted community structure $S = \langle \theta, \Omega \rangle$; Community membership matrix ϕ .

Algorithmic Steps:

/* Initialization */

- (1) Initialize θ and ϕ with nonnegative values, and scale the columns of ϕ to unit norm.
- (2) Initialize user and community assignments for each word token w_i based on ϕ .
- (3) Iterate until convergence or after l iterations

(a) for $i = 1$ to $|W|$

- ① Randomly sample a word token w_i from the document corpus.
- ② Randomly sample a community k for w_i , conditioned on Ω .
- ③ Randomly sample a user u_i for w_i conditioned on ϕ , and assign user u_i to community k .

(b) Sample the matrix of word assignment to community topics $C^{W\gamma}$ and the matrix of user assignment to communities $C^{\gamma U}$.

(c) Calculate word distribution for community topics by $\Omega_{wk} =$

$$\frac{C_{wk}^{W\gamma} + \beta}{\sum_w C_{wk}^{W\gamma} + |U|\beta}, \quad \text{and user distribution for communities by}$$

$$\varphi_{uk} = \frac{C_{ku}^{\gamma U} + \alpha}{\sum_{u' \in U} C_{ku'}^{\gamma U} + |U|\alpha}$$

(d) update ϕ : $\phi = \frac{\phi \theta^T M}{\theta^T \theta \phi + \varepsilon} \circ \exp(\lambda \varphi)$

(e) update θ : $\theta = \frac{\theta M \phi^T}{\theta \phi \phi^T + \varepsilon}$

(f) Scale the columns of ϕ to unit form.

(g) Calculate the sum of squared error with Eq. (1).

(4) Scale the columns of θ and ϕ to unit form.

(5) Output the detected community structure $\langle \theta, \Omega \rangle$ and community membership ϕ .

5. Empirical evaluation

In this section, we empirically assess the efficacy of the proposed method using two real world datasets. To illustrate the benefit of combining the friendship network and content information, pure NMF and AT methods were used as the benchmark methods to compare against our NMF-AT algorithm. The MetaFac [29] model was also selected as benchmark, which performs tensor factorization on the multi-relational network and shares a similar high-level design as ours.

In the following subsections, we first introduce evaluation measures. Then, we report the experimental findings on the Delicious dataset and the Twitter dataset, respectively.

5.1. Performance evaluation

User and topic are two major components of any characterization of communities. To evaluate the quality of the given communities detected by any algorithms, we use the following user- and topic-related measures. The mean value $\bar{\mu}$, which is the average value of soft modularity Q_s and community's user-content similarity S_{ui} , measures the comprehensive performance of friendship density and content similarity in the extracted communities, of which higher Q_s indicating that users in the same community are densely connected with each other, and higher S_{ui} indicating that users in the same community share more similar content information with each other. We use community user divergence D_U and community topic divergence D_T to evaluate the diversity of detected communities from different perspectives. A higher divergence value generally implies that the communities are better distinguished from each other. We also use a composite divergence measure D , a harmonic mean of D_U and D_T .

5.1.1. Mean value measurements

For a good community partition, users belonging to the same community should be densely connected with each other and share common content interest. These could be measured by integrating two measurements, which include evaluating users' friendship connections with the well-known soft modularity method and calculating users' content similarity within the same community. We define the mean value $\bar{\mu}$ based on soft modularity Q_s and user-content similarity S_{ui} , to measure the quality of the extracted communities. The mean value $\bar{\mu}$ can be calculated as $\bar{\mu} = \frac{Q_s + S_{ui}}{2}$, where soft modularity Q_s and user-content similarity S_{ui} are explained as follows.

a. Soft modularity. Newman and Girvan developed a widely-used modularity function Q [6] to measure the goodness of a community structure. The soft modularity Q_s as defined in [20] extended the concept of modularity to evaluate overlapping community structures in undirected networks. Based on these concepts, we define soft modularity suitable for directed networks as follows:

Definition 4. Soft modularity Q_s .

Give the matrix representation of a network M , community distribution over users θ and the community membership matrix ϕ , the soft modularity value for overlapping community structures is calculated as: $Q_s = Tr[\theta^T \cdot M \cdot \phi] - E^T \cdot M \cdot \phi^T \cdot \phi \cdot M^T \cdot E$, where E is a vector whose elements are all ones.

Table 2
The basic statistics of the Delicious dataset.

Dataset	# of users	# of links among users	# of urls	# of tags
Delicious	749	9995	106,611	51,106

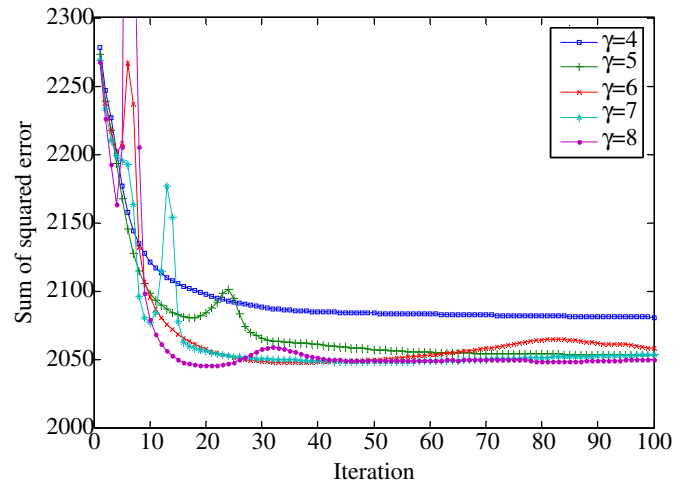


Fig. 4. Sum of squared errors as the number of iteration increases on the Delicious dataset.

b. Users' content similarity. We represent the content feature of each user as a tf-idf vector over the related words. Then we calculate user-content similarity as follows:

Definition 5. User-content similarity S_u .

Given the tf-idf vector representation of each user V_i ($i = 1, 2, \dots, m$) the users' content similarity value for overlapping community structures is calculated as:

$$S_u = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^{\gamma} R_{ki} R_{kj} \text{sim}(V_i, V_j)}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=1}^{\gamma} R_{ki} R_{kj}}$$

where $R_{ki} = 1$ if user i belongs to community k , otherwise, $R_{ki} = 0$; $\text{sim}(V_i, V_j)$ indicates the cosine similarity of contents between user i and user j .

5.1.2. Divergence measurements

For a good community partition, the extracted communities should be distinguished from each other. This could be measured by calculating the distance among communities. Jenson-Shannon (JS) divergence has been a popular method for measuring the distance between two probability distributions [24,35]. Based on JS divergence, we define D_U and D_T to measure the distance of detected communities from user distribution and community topic distribution, respectively.

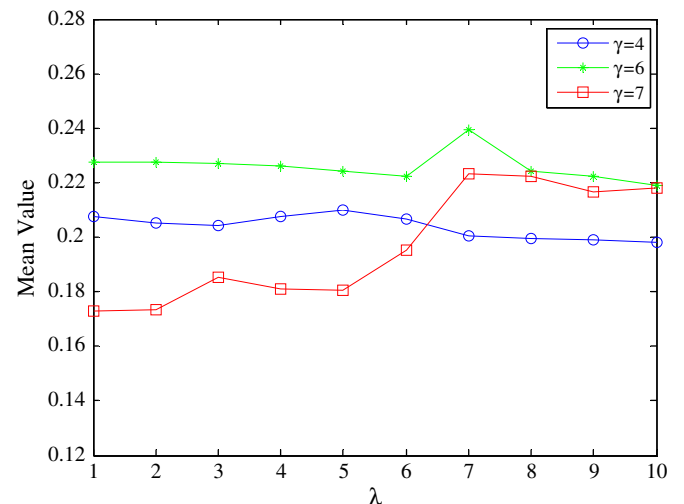


Fig. 5. Mean value $\bar{\mu}$ varying with λ on the Delicious dataset.

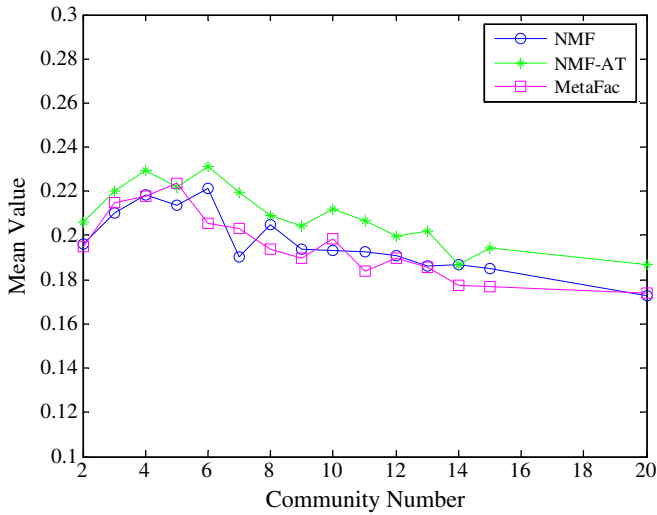


Fig. 6. Comparison of mean value $\bar{\mu}$ on the Delicious dataset t.

Definition 6. Community user divergence D_U

Given two communities C_1 and C_2 , the community user divergence between C_1 and C_2 is defined as: $D_{U(C_1,C_2)} = JS(\theta_{C_1}, \theta_{C_2})$, where $JS(\theta_{C_1}, \theta_{C_2})$ is the JS divergence [24] between the two probabilistic distributions $\theta_{C_1}, \theta_{C_2}$.

Definition 7. Community topic divergence D_T

Given two communities C_1 and C_2 , the community topic divergence between C_1 and C_2 is defined as: $D_{T(C_1,C_2)} = JS(\Omega_{C_1}, \Omega_{C_2})$.

To gain balance between the user and topic divergence in measuring an identified community structure, we calculate the overall community divergence as the harmonic mean of D_U and D_T :

Definition 8. Community divergence D

Given two communities C_1 and C_2 , the community divergence between C_1 and C_2 is calculated as: $D_{(C_1,C_2)} = \frac{D_U \cdot D_T}{D_U + D_T}$.

The divergence values are measured pairwise among communities in the detected community structure S . To avoid cluttering, we only report the average divergence value calculated as: $\bar{div} = \frac{1}{\gamma(\gamma-1)} \sum_{C_i \in S} \sum_{C_j \in S, C_j \neq C_i} div_{(C_i,C_j)}$, where div could be D_U, D_T or D .

5.2. Empirical evaluation using a Delicious dataset

In this subsection, we evaluate the performance of our method on a Delicious dataset [34]. Delicious is a popular social tagging system,

Table 3
Community topics on the Delicious dataset.

Community	Top ranked keywords	Top ranked users
C1	rate 0.04777, slash 0.03779, supernatur 0.03269, dean 0.01863, wincest 0.01425, jare 0.01274, jensen 0.01246, fiction 0.01244, panicatthedisco 0.00897, merlin 0.00717	fandomdirectory, wolfgrin, lilynjudus, zing_och, ispahan, dossier1013, amithereyet, ancientsavvy, bookbindbound, fiercynn
C2	dean 0.03684, slash 0.03346, supernatur 0.02253, rate 0.01909, jensen 0.01727, jare 0.0159, bigbang 0.01267, wincest 0.01139, sheppard 0.01022, mckai 0.00958	morgandawn, maygra, tzikeh, calime, ninasis, melina123, black_samvara, callaoressene, rosaw, InDenial88
C3	polit 0.01347, blog 0.00659, fandom 0.00649, book 0.00635, supernatur 0.00598, race 0.00591, elect 0.00527, obama 0.00517, fanfict 0.00479, stargat 0.00476	ibarw, mattbastard, ursamajor, vermilion, coffeeandink, ratcreature, xanphibian, calime, thebratq, Itrasbiel
C4	slash 0.05126, sheppard 0.04369, mckai 0.04006, fanfict 0.02596, ianto 0.0246, torchwood 0.02177, jack 0.01823, romanc 0.01753, humour 0.01614, rodnei 0.01498	stasha2g, thdancingferret, jfritsche, tabulaxrasa, fleurdeleo, amberlynn, madam_minnie, prurient_badger, kangeiko, turloughishere
C5	bandom 0.03522, bandslash 0.02281, fiction 0.0205, brendon 0.01768, spencer 0.0164, ryan 0.01563, frank 0.01299, gerard 0.01289, pete 0.01115, panic 0.01084	iris.eyes, sinsense, dirtykicks, prettykitty_aya, fandom.clare, lattara, stepps, mchandfan, ubastira, ScorpionofWater
C6	brendon 0.07715, spencer 0.06672, ryan 0.06185, bandom 0.05225, band 0.0376, bandslash 0.0308, pete 0.02889, patrick 0.02609, panicatthedisco 0.025, panic 0.02237	beachan18_recs, fickle_goddess, chexmix, blonde_cecile, bonibaru, lilitchiilde, Manderkitty, flamingsword, zarq, aishia

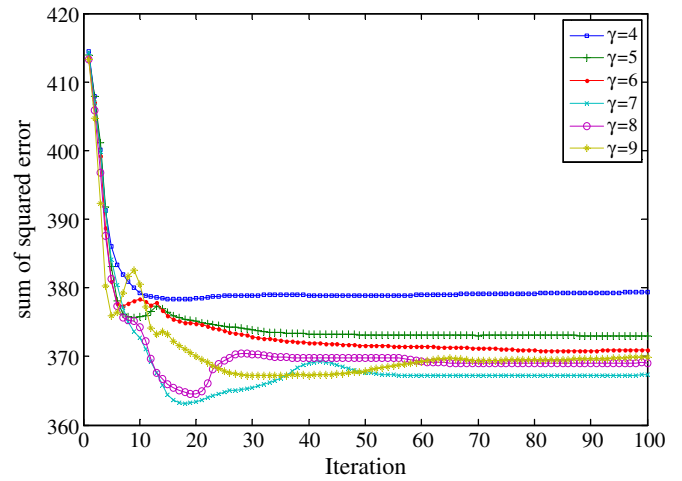


Fig. 7. Sum of squared errors as the number of iteration increases on the Twitter dataset.

allowing users to bookmark various web urls with individually selected tags. Additionally, it allows a user to create her social network by adding other users into her friendship network. The raw dataset consists of 5000 users and turns out to be extremely large and sparse. To reduce the size of the raw dataset, we dropped users who did not post any urls and urls that had received less than 5 tags during the period of June 2008 to June 2009. Finally, a smaller dataset consisting of 749 users was obtained. The key statistics are summarized in Table 2.

5.2.1. Convergence of NMF-AT on the Delicious dataset

The updating rules of NMF are guaranteed to converge to a locally optimal matrix factorization. We provide empirical evidence that this convergence property also holds for the NMF-AT model. Fig. 4 plots the sum of the squared errors varying as the number of iteration increases for different community settings. It can be seen that NMF-AT converges quite fast, and gets very close to the minimum error in 60 iterations. In the following experiments, we fix the iteration number at 100. Fig. 5 plots the mean value $\bar{\mu}$ varying with λ when the number of community set to be 4, 6 and 7. It can be noted that NMF-AT model achieves its best performances around the point $\lambda = 7$. In the following analysis, we set $\lambda = 7$.

5.2.2. Method comparison on the Delicious dataset

We have compared the performance of our approach with benchmark methods on the Delicious dataset. Fig. 6 shows comparison of the mean value $\bar{\mu}$ among NMF-AT, NMF and MetaFac models under different community numbers. Two relations are identified as the input of MetaFac model, including a 2-order tensor R1 describing the user friendship network, and another 3-order tensor R2 describing the relation among users, urls and tags. A triple (U_i, l_j, tk) in R2 indicates

that user U_i has bookmarked web page l_j with tag tk . It can be seen that the NMF-AT model constantly extracts higher quality communities than NMF, indicating the benefits of incorporating the content information for community discovery. MetaFac shows similar performance with NMF as γ increases. The pairwise t -test shows that NMF-AT model outperforms NMF and MetaFac with $p < 0.001$.

From Fig. 6, we observe that the peak value is achieved when $\gamma = 6$ by our model. The evaluation results on divergence measures when $\gamma = 6$ are listed in Table 4. Since NMF does not rely on the content information, it is not evaluated on \overline{D}_T and \overline{D} . We can see that best community divergence value $\overline{D} = 0.503$ is achieved by the NMF-AT model. NMF-AT and MetaFac show similar performance on community user divergence \overline{D}_U , indicating that our method can better detect users' community membership than NMF and AT models. The AT model gains highest community topic divergence \overline{D}_T and performs best in grasping the topics of the dataset. MetaFac and AT models seem to emphasize on the user friendship network and user-generated contents, respectively. On the other hand, our model gains better balance on the overall performance.

5.2.3. Topic analysis on the Delicious dataset

We further analyze the community topics on the Delicious dataset when $\gamma = 6$, to illustrate if significant communities have been explored. Table 3 lists the top ranked keywords and users for each community. We verify these communities by manually checking the topic keywords and web pages bookmarked with these tags by top ranked users.

Users in community C1 are probably fond of television drama series or fantasy shows. Web pages about *Merlin* and *Supernatural* (two popular magic fantasy shows) have been frequently tagged. C2 seems to focus on the TV series *Supernatural* solely. Keyword “dean” is the fiction character, while “Jare” and “Jenson” are two actors in the show. Many urls bookmarked in C2 link to fan fiction stories about the show, with a portion of stories involving a sexual relationship between Sam and Dean—known as “wincest” [22]. Users in C3 have tagged a great portion of web pages on political and fan fiction, while users in C4 mainly focus on science fiction, such as *Torchwood* and *Stargate: Atlantis*. C5 and C6 are both about bandslash. Users in C6 seem to focus on Panic! At The Disco (a rock band) only, while users in C5 also collect resources other than Panic! At The Disco.

With previous analyses, it can be concluded that NMF-AT model could enhance the modularity of discovered communities by incorporating the user-content information. The community topics are also explored in a unified way. As can be seen from the community topics, the overlapping community structure may sometimes cause topic ambiguities. For instance, both C1 and C2 contain users who are interested in

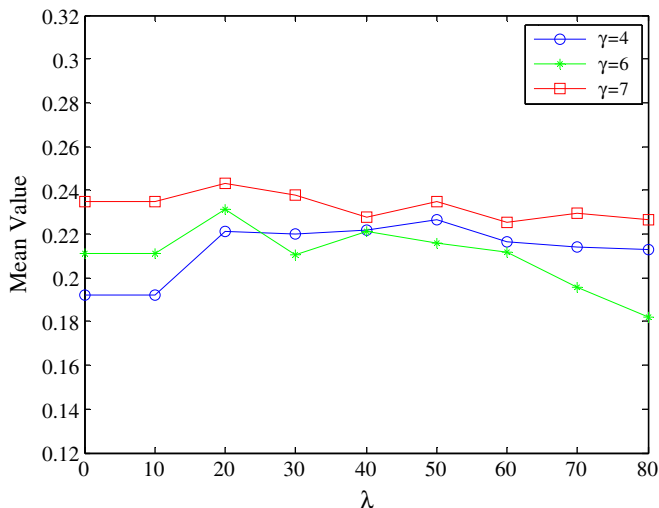


Fig. 8. Mean value $\bar{\mu}$ varying with λ on the Twitter dataset.

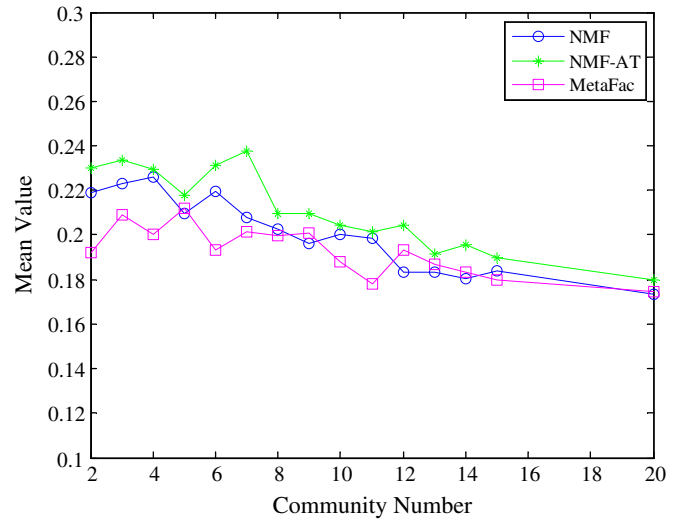


Fig. 9. Comparison of mean value $\bar{\mu}$ on the Twitter dataset.

Supernatural, and both C5 and C6 are about bandslash. Further studies have been planned to perform community disambiguation.

5.3. Empirical evaluation using a Twitter dataset

In this subsection, we evaluate our method on a Twitter dataset [21]. Twitter has been a representative microblogging service and has been experiencing rapid growth. Tweets during June 2009 and September 2009 are used for content analysis, with each tweet regarded as a document. By filtering out users who have less than five follower/following relationships, the final Twitter dataset covers 791 users, with 3437 follower/following links, and 59,104 tweets.

5.3.1. Convergence of NMF-AT on the Twitter dataset

We first study computationally the convergence of the proposed NMF-AT method. Fig. 7 plots the sum of squared errors under different number of iterations on the Twitter dataset ($\lambda = 20$). Again, it can be seen that NMF-AT converges quite fast and reaches the neighborhood of the minimum error in 60 iterations. The variation in minimal squared errors is also slow when $\gamma \geq 5$. We fixed the iteration number at 100 in the following settings. Fig. 8 plots the mean value $\bar{\mu}$ varying with λ when $\gamma = 4, 6$ and 7 , respectively. As can be seen, the best performance is achieved when $\lambda = 20$. In the following analysis, we set $\lambda = 20$ for the Twitter dataset.

5.3.2. Method comparison on the Twitter dataset

Fig. 9 shows the comparison of mean value $\bar{\mu}$ among NMF-AT, NMF and MetaFac models under different community numbers. We observe that NMF-AT improves the performance of community detection by considering the tweets information. The overall curve trends of the three models maintain the same, they go up to the peak and then drop as γ increases. The difference is that NMF-AT performs better

Table 4 Comparison of divergence measures on Delicious and Twitter (Higher values indicate better divergency performances).

	Delicious dataset			Twitter dataset				
	NMF	AT	MetaFac	NMF-AT	NMF	AT	MetaFac	NMF-AT
\overline{D}_U	0.512	0.465	0.637	0.633	0.54	0.277	0.259	0.67
\overline{D}_T	-	0.578	0.276	0.366	-	0.584	0.719	0.339
\overline{D}	-	0.494	0.439	0.503	-	0.329	0.327	0.494

Table 5
Community topics on the Twitter dataset.

Community	Top ranked keywords	Top ranked users
C1	twitter 0.0084 tinyurl 0.00799 free 0.00783 great 0.00738 make 0.00735 follow 0.00711 blog 0.00671 market 0.00657 check 0.00539 love 0.00517	BOBBYandNIKOent, dominoesstars, 1909ramon, SalesChoice, BraLadies, tEntrepreneur, Consoleskint, eggwhisk, Loony_Luna, taliamarie77
C2	forex 0.0144 new 0.01093 miss 0.01018 trade 0.00944 social 0.00935 glennbeck 0.00894 visit 0.00836 copyright 0.00836 onsugar 0.00828 beautyandthebudget 0.00828	2010inVancouver, 247success, 37nats, 7×20, CashWiz, 4enMoneyTrade, 123Print, BigDaddyNoyd, itsjenjen, companyofprayer
C3	free 0.01173 tinyurl 0.01122 make 0.01031 twitter 0.00974 monei 0.00824 tweet 0.00813 great 0.00783 look 0.00688 market 0.00685 busi 0.00569	AngelaAshby, 90dayexit, itouchinsanity, 3900income, 21 T, SkyeKing, 401kExpert, abspecken, 1sourceconsult, 3_step_success
C4	sale 0.0279 free 0.02621 peopl 0.02368 program 0.02199 profit 0.01945 halt 0.01691 foreclosure 0.01438 make 0.01269 life 0.01184 come 0.011	BarterCoach, changmin_boh, 5le, 3brothersbakery, 59minlearning, 4stepformula, twithealth, altpowerblog, Brettconyers, 5MinuteFitness
C5	morn 0.03109 rock 0.02286 new 0.02195 lori 0.01829 klzx 0.01829 cain 0.01738 zybsg 0.01646 classic 0.01555 feelright 0.01463 show 0.01463	959KLZX, SFBrawnyBear, 2015wpg, dominoesstars, 7_ly, datasnatchers, 5le, DDPokerPlayer, Abunzawg, cashflow4yu
C6	updat 0.03429 cell 0.03412 wallpap 0.03395 busi 0.01426 melkweg 0.0129 sbdc 0.01205 free 0.01052 nokia 0.00968 tinyurl 0.00968 ericsson 0.00951	2020plus1, IdahoSBDC, AAObserver, jennifer_dubow, 00sw, cell49, melkweg, TremFarmMarket, 140Conf, 07forcada
C7	tinyurl 0.01854 make 0.01691 twitter 0.01094 follow 0.01041 free 0.01009 movi 0.00738 wefollowfriday 0.00669 onlin 0.00674 followfriday 0.00666 monei 0.00655	2DBarcodenow, 2abetterU, 005dabrown, 2020Obits, 121Fitness, 2frog, rivalin, 07BondGirlXX, mailinks, 18twit

than NMF and Metafac throughout the interval range. Additionally, it can be noted that the best community partition occurs when $\gamma = 7$.

The divergence comparison of NMF-AT with benchmark methods on the Twitter dataset is shown in Table 4. It can be seen that NMF-AT achieves highest community user divergence value \bar{D}_U , indicating our method can group users better into different communities. The MetaFac model performs best in grasping community topics and gains highest community topic divergence \bar{D}_T . Again, the NMF-AT model shows its advantage in the combined community divergence measure \bar{D} , implying the value of combining the friendship network factorization and user content analysis.

5.3.3. Topic analysis on the Twitter dataset

We examined the community topics when $\gamma = 7$, to illustrate if significant communities have been identified. Table 5 lists the top ranked keywords and users for each community. We verify these community topics by manually checking the topic keywords and tweets posted by top ranked users.

Community C1 seems to be about network marketing on Twitter. One top-ranked community member, 1909ramon, has been broadcasting his website for reporting contactors. Another user, BraLadies, has been sharing his digs. The third user, tEntrepreneur, seems to be advocating online marketing, and so on. Community C2 may be about some websites for trades and social news, such as beautyandthebudget (an online shopping and fashion trend site), onsugar (a free blogging platform) and glennbeck (a syndicated talk show host). C3 is probably about doing business on Twitter. C4 may be about foreclosures and bartering. With the burst of the real estate bubble, foreclosure houses have become a great concern for many users. C5 mainly discusses rock music, with klzx being a radio station broadcasting classic rock format and Lori a rock music artist. C6 talks about cell phones. C7 is concerned with some twitter applications for following relationships, such as wefollowfriday.

6. Conclusions and future research

In this study, we have proposed a unified framework for user community discovery from multi-relational networks. The preliminary evaluation with two real-world datasets indicates that this model can detect significant communities, which have dense friendship connections and common interests among community members. We have illustrated that user-generated contents also play a significant role in discovering user community structures from online social networks. By utilizing both the friendship network and content information, we are able to improve the modularity properties of discovered community structures. Additionally, community topics can be explored in a unified procedure, which potentially saves a lot of manual efforts trying to interpret the discovered communities.

There has been much effort in analyzing the temporal evolution of communities in the literature. Utilizing our model to track the evolution of communities and their topics would be a promising future research direction. The further application of extracted communities in personalized services also represents a significant extension of our work, with applications in personalized resource recommendation and personalized information retrieval, among others.

Acknowledgment

This research is supported by NSFC grants (No. 61172106, No. 71025001, No. 91024030, No. 90924302, No. 60921061, and No. 70890084), the MOH Grant (No. 2012ZX10004801), a grant from U.S. DHS Center of Excellence in Border Security and Immigration (No. 2008-ST-061-BS0002), and BJNSF (No. 4112062).

References

- [1] H. Lauw, J.C. Shafer, R. Agrawal, A. Ntoulas, Homophily in the digital world: a live journal case study, *IEEE Internet Computing* 14 (2) (2010) 15–23.
- [2] W. Duan, Special issue on online communities and social network: an editorial introduction, *Decision Support Systems* 47 (3) (2009) 167–168.
- [3] K. Valck, G.H. Bruggen, B. Wierenga, Virtual communities: a marketing perspective, *Decision Support Systems* 47 (3) (2009) 185–203.
- [4] D. Ganley, C. Lamp, The ties that bind: social network principles in online communities, *Decision Support Systems* 47 (3) (2009) 266–274.
- [5] S. Fortunato, Community detection in graphs, *Physics Reports* 486 (2010) 75–174.
- [6] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, *Physical Review E* 69 (2004) 026113.
- [7] S. White, P. Smyth, A spectral clustering approach to finding communities in graph, in: *Proceedings of the Fifth SIAM International Conference on Data Mining*, 2005, pp. 274–285.
- [8] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, *Nature* 435 (2005) 814–818.
- [9] H. Shen, X. Cheng, J. Guo, Quantifying and identifying the overlapping community structure in networks, physics and society, *Journal of Statistical Mechanics* (2009) P07042.
- [10] T.S. Evans, R. Lambiotte, Line graphs, link partitions, and overlapping communities, *Physical Review E* 80 (1) (2009) 016105.
- [11] S. Zhang, R. Wang, X. Zhang, Uncovering fuzzy community structure in complex networks, *Physical Review E* 76 (4) (2007) 046103.
- [12] M. Steyvers, P. Smyth, M. Rosen-Zvi, T. Griffiths, Probabilistic author-topic models for information discovery, in: *The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 306–315.
- [13] A. McCallum, X. Wang, A. Corrada-Emmanuel, Topic and role discovery in social networks with experiments on Enron and academic email, *The Journal of Artificial Intelligence Research* 30 (2007) 249–272.
- [14] D. Zhou, I. Councill, H. Zha, C. Giles, Discovering temporal communities from social network documents, in: *Proceedings of the 7th IEEE International Conference on Data Mining*, 2007, pp. 745–750.
- [15] N. Pathak, C. Delong, A. Banerjee, K. Erickson, Social topic models for community extraction, in: *The 2nd SNA-KDD Workshop '08*, 2008.
- [16] Q. Mei, D. Cai, D. Zhang, C. Zhai, Topic modeling with network regularization, in: *The 17th International World Wide Web Conference*, 2008, pp. 101–110.
- [17] X. He, H. Zha, C.H.Q. Ding, H.D. Simon, Web document clustering using hyper-link structures, *Computational Statistics and Data Analysis* 41 (1) (2002) 19–45.
- [18] D. Lee, H. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature* 401 (1999) 788–791.

- [19] D. Lee, H. Seung, Algorithms for non-negative matrix factorization, in: *Advances in Neural Information Processing Systems* 13, 2001, pp. 556–562.
- [20] Y. Lin, Y. Chi, S. Zhu, H. Sundaram, B. Tseng, Analyzing communities and their evolutions in dynamic social networks, *ACM Transactions on Knowledge Discovery from Data*, Special Issue on Social Computing, Behavioral Modeling, and Prediction 3 (2) (2009) 1–31.
- [21] M. Choudhury, Y. Lin, H. Sundaram, K. Candan, L. Xie, A. Kelliher, How does the data sampling strategy impact the discovery of information diffusion in social media? in: *The 4th Int'l AAAI Conference on Weblogs and Social Media*, 2010.
- [22] http://en.wikipedia.org/wiki/Supernatural_TV_series.
- [23] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, M. Steyvers, Learning author-topic models from text corpora, *ACM Transactions on Information Systems* 28 (1) (2010).
- [24] M. Steyvers, T. Griffiths, Probabilistic topic models, in: *Handbook of Latent Semantic Analysis*, 2007.
- [25] T. Spaulding, How can virtual communities create value for business? *Electronic Commerce Research and Applications* 9 (1) (2010) 38–49.
- [26] R. Arakji, R. Benbunan-Fich, M. Koufaris, Exploring contributions of public resources in social bookmarking systems, *Decision Support Systems* 47 (3) (2009) 245–253.
- [27] C. Chiu, M. Hsu, E. Wang, Understanding knowledge sharing in virtual communities: an integration of social capital and social cognitive theories, *Decision Support Systems* 42 (3) (2006) 1872–1888.
- [28] Z. Zhang, Q. Li, D. Zeng, Mining user communities from online social network services, in: *Proceedings of the 20th Annual Workshop on Information Technologies and Systems*, 2010, pp. 91–96.
- [29] Y. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, A. Kelliher, MetaFac: community discovery via relational hypergraph factorization, in: *The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 527–536.
- [30] L. Nie, B. Davison, B. Wu, From whence does your authority come? Utilizing community relevance in ranking, in: *The 22nd AAAI Conference on Artificial Intelligence*, 2007, pp. 1421–1426.
- [31] H. Ma, H. Yang, M. Lyu, I. King, SoRec: social recommendation using probabilistic matrix factorization, in: *ACM 17th Conference on Information and Knowledge Management*, 2008, pp. 931–940.
- [32] D. Zhou, S. Zhu, K. Yu, X. Song, B.L. Tseng, H. Zha, C.L. Giles, Learning multiple graphs for document recommendations, in: *Proceedings of the 17th international Conference on World Wide Web*, 2008, pp. 141–150.
- [33] S. Zhu, K. Yu, Y. Chi, Y. Gong, Combining content and link for classification using matrix factorization, in: *The 30th Annual International ACM SIGIR Conference*, 2007, pp. 487–494.
- [34] P. Jing, D. Zeng, Topic-based web page recommendation using tags, in: *IEEE International Conference on Intelligence and Security Informatics*, 2009, pp. 269–271.
- [35] J. Lin, Divergence measures based on the Shannon Entropy, *IEEE Transactions on Information Theory* 37 (1) (1991) 145–151.
- [36] E.A. Leicht, M.E.J. Newman, Community structure in directed networks, *Physical Review Letters* 100 (11) (2008).
- [37] D. Zeng, F.Y. Wang, K.M. Carley, Social computing, *IEEE Intelligent Systems* 22 (5) (2007) 20–22.
- [38] F.Y. Wang, K.M. Carley, D. Zeng, W. Mao, Social computing: from social informatics to social intelligence, *IEEE Intelligent Systems* 22 (2) (2007) 79–83.
- [39] K.P. Sycara, D. Zeng, Towards an intelligent electronic secretary, *International Conference on Information and Knowledge Management, Intelligent Information Agents Workshop*, 1994.
- [40] F.Y. Wang, D. Zeng, J.A. Hendler, Q. Zhang, Z. Feng, Y. Gao, H. Wang, G. Lai, A study of the human flesh search engine: crowd-powered expansion of online knowledge, *Computer* 43 (8) (2010) 45–53.
- [41] X. Zheng, D. Zeng, H. Li, F. Wang, Analyzing open-source software systems as complex networks, *Physica A: Statistical Mechanics and its Applications* 387 (24) (2008) 6190–6200.
- [42] D. Zeng, H. Li, How useful are tags?—an empirical analysis of collaborative tagging for Web page recommendation, *Intelligence and Security Informatics, Lecture Notes in Computer Science* 5075 (2008) 320–330.
- [43] J. Peng, D.D. Zeng, H. Zhao, F. Wang, Collaborative filtering in social tagging systems based on joint item-tag recommendations, in: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010, pp. 809–818.

Zhongfeng Zhang is a Ph.D. candidate in the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. He received his B.S. degree in Automation from Nanjing University of Posts and Telecommunications, China, 2006. His research interests include information retrieval, web/text mining and community question answering.

Qiudan Li is an Associate Professor in the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. She received a Ph.D. in Computer Science from Da Lian University of Technology, China, 2004. Her research interests include web mining and mobile commerce applications. Her articles have been published in *Communications of the AIS*, *Decision Support Systems*, *IEEE Transactions on SMC*, *Expert Systems with Applications*, as well as in *WWW*, *CIKM*, *ECIR*, *IJCNN* international conference proceedings.

Daniel Zeng received the M.S. and Ph.D. degrees in industrial administration from Carnegie Mellon University, Pittsburgh, PA, and the B.S. degree in economics and operations research from the University of Science and Technology of China, Hefei, China. Currently, he is a Professor and Honeywell Fellow in the Department of Management Information Systems at the University of Arizona, Tucson, Arizona, U.S.A. He is also a Research Professor in the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interests include social computing, software agents, spatial-temporal data analysis, and intelligence and security informatics. He has co-edited fifteen books and published more than 170 peer-reviewed articles on information systems, computer science, and public health informatics journals, edited books, and conference proceedings. His research has been mainly funded by the U.S. NSF, U.S. DHS, MOST, and NNSFC. He serves on editorial boards of ten Information Technology-related journals. He is active in information systems, and public health and security informatics professional activities and is Vice President for Technical Activities for the IEEE Intelligent Transportation Systems Society and Chair of INFORMS College on Artificial Intelligence.

Heng Gao is a Master's Candidate in the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. He received his B.E. degree in Computer Science from China University of Mining and Technology, China, 2010. His research interests include information retrieval, web/text mining and community question answering.