Nonnegative Matrix Tri-Factorization with Graph Regularization for Community Detection in Social Networks

Yulong Pei[1] **Nilanjan Chakraborty**[2], **Katia Sycara**[1]
[1]School of Computer Science, Carnegie Mellon University
{yulongp,katia}@cs.cmu.edu
[2]Department of Mechanical Engineering, Stony Brook University
nilanjan.chakraborty@stonybrook.edu

Abstract

Community detection on social media is a classic and challenging task. In this paper, we study the problem of detecting communities by combining social relations and user generated content in social networks. We propose a nonnegative matrix tri-factorization (NMTF) based clustering framework with three types of graph regularization. The NMTF based clustering framework can combine the relations and content seamlessly and the graph regularization can capture user similarity, message similarity and user interaction explicitly. In order to design regularization components, we further exploit user similarity and message similarity in social networks. A unified optimization problem is proposed by integrating the NMTF framework and the graph regularization. Then we derive an iterative learning algorithm for this optimization problem. Extensive experiments are conducted on three real-world data sets and the experimental results demonstrate the effectiveness of the proposed method.

1 Introduction

Community detection in social networks is a classical and challenging problem. A community can be defined as a group of users that (1) interact with each other more frequently than with those outside the group and (2) are more similar to each other than to those outside the group. The research on community detection is beneficial for a variety of real-world applications such as online marketing and recommendation systems.

Many existing works on community detection focus only on social relations [Girvan and Newman, 2002; Newman, 2006; Wang *et al.*, 2011] or content [Lee *et al.*, 2013]. However, neither social relations nor content alone can indicate the community membership accurately. On one hand, in the real-world social media such as Twitter, compared with the large amount of users, the social relations for each user are extremely sparse and two users may belong to the same community even if there are no relations between them. On the other hand, the content on social media is diverse and noisy which will influence the content analysis and may lead to failure in detecting communities. Therefore, combining the relations and content may be a better strategy for community detection. Several combination strategies for community detection have been proposed [Ruan *et al.*, 2013; Sachan *et al.*, 2012; Yang *et al.*, 2009]. However, there are several shortcomings in these methods. In heuristic linear combination method [Ruan *et al.*, 2013], the strategy lacks theoretical basis and the parameter for combining relations and content is difficult to determine. In topic model based method [Sachan *et al.*, 2012], the modeling process depends on content information and may be misled by the noisy and irrelevant information. In the discriminative model [Yang *et al.*, 2009], relations and content are modeled by two individual models and the user similarity is not modeled explicitly.

In summary, there are three challenges in community detection task: (1) *Combination Strategy*. As explained above, it is insufficient to determine the community membership using only social relations or only content. For example, there is no relation between user u_5 and u_6 in Figure 1, but they should be grouped into the same community because they published similar content related to *Technology*. Thus, the model should combine social relations and content in detecting communities. (2) *Model Flexibility*. This challenge requires the model can capture different social information without changing the model form so the model can be generalized in modeling different social networks. A negative example is the topic model. If incorporating new social information, the generative process will be different. (3) *User Similarity*. Community detection essentially is a user clustering problem in which the user similarity plays an important role. Thus, the model should consider the user similarity explicitly. Beside, the user similarity calculation should use both user relations and content. For instance, user u_2 and u_3 in Figure 1 are similar if only considering the message similarity, but in fact u_2 and u_3 belong to different communities.

In this paper, we organize users and messages in a user-word-message tripartite graph shown in Figure 1. Our aim is to cluster the users into different communities using not only the user-word-message relations and the user pairwise relation shown in Figure 1 but user similarity, message similarity and user interaction. In order to cluster the users, we employ a constrained nonnegative matrix tri-factorization (NMTF) framework [Ding *et al.*, 2006] to cluster users and messages simultaneously by combining the relations and content, and

propose three types of graph regularization [Smola and Kondor, 2003] to model user similarity, message similarity and user interaction explicitly. This proposed method can deal with the three challenges introduced above.

1. We utilize two NMTF components to model the user-word relation and the message-word relation respectively and one NMTF component to model the user pairwise relation. NMTF method performs well in co-clustering tasks with multiple relations [Gu and Zhou, 2009] so the combination of these NMTF components can fuse social relations and content.
2. By integrating graph regularization, the NMTF framework is flexible in incorporating rich social information such as retweet and citation in social networks. In particular, we introduce three types of graph regularization based on user similarity, message similarity and interaction respectively in this paper. Other social information can also be integrated using similar graph regularization without changing the form of this framework.
3. We model the user similarity and message similarity explicitly in the graph regularization. To exploit the similarities, we construct a two-layer graph based on user pairwise relations, message pairwise relations and user-message relations, and then propose a random walk method on this graph to calculate the user similarity and message similarity. This method employs both user relations and content to calculate the user similarity and message similarity in the networks.

Furthermore, in order to validate the effectiveness of the proposed method, experiments are conducted on three real-world data sets.

2 Related Work

The methods for community detection can be categorized into three types: relation-only methods, content-only methods and methods combining relations and content. For more details, please refer to the survey papers [Tang and Liu, 2010; Fortunato, 2010]. Nonnegative matrix factorization (NMF) [Lee and Seung, 2001] has been shown to be useful in many research areas. By introducing orthogonality constraints, NMF can perform well in clustering [Gu and Zhou, 2009]. [Wang *et al.*, 2011] applied NMF to model the networks and cluster users into communities, but they did not take into consideration the content generated by users. Since traditional 2-factor factorization $X = FG^T$ can only capture two types of relations, Ding et al. [Ding *et al.*, 2006] extended NMF to 3-factor factorization $X = FSG^T$, i.e., NMTF, and this 3-factor factorization can capture more types of relations. NMTF has been employed in sentiment classification in [Li *et al.*, 2009] and [Zhu *et al.*, 2014]. In order to incorporate prior knowledge in sentiment, [Li *et al.*, 2009] introduced the sentiment lexicon based regularization. To capture the user interaction in Twitter, [Zhu *et al.*, 2014] also applied retweet based regularization in the NMTF framework. However, different from sentiment analysis, in this paper, we focus on community detection which is based on user similarity, so we consider not only the interaction based regularization, but also user similarity based regularization and message similarity based regularization. Besides, we do not include constraint on words in our study.

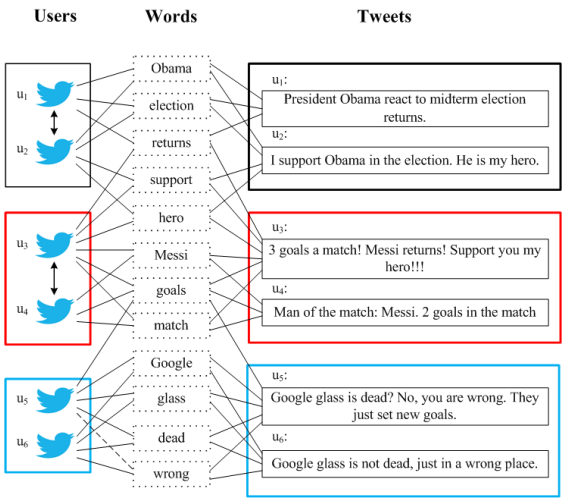

Figure 1: The user-word-message tripartite graph. There are 6 users and 6 messages published by these users. The user-word link denotes a user uses a word and the message-word link denotes a message contains a word. The arrows between u_1 and u_2, and between u_3 and u_4 indicate the social relations between these users. The users and messages in the black rectangle, in the red rectangle and in the blue rectangle belong to the *politics* community, the *soccer* community and the *technology* community, respectively.

3 Problem Statement

First we introduce the notations used in this paper, which are listed in Table 1. We use m to denote the number of users, n to denotes the number of messages, w to denote the number of words in all the messages and k to denote the number of communities. M_{u-u} is a binary matrix to denote the user pairwise relation, i.e., if there is relation between user i and user j, $M_{u-u}(i,j) = 1$, otherwise $M_{u-u}(i,j) = 0$. M_{u-f} and M_{t-f} are both binary matrices. $M_{u-f}(i,j) = 1$ denotes the ith user used the jth word and $M_{t-f}(i,j) = 1$ denotes the ith message contains the jth word. S_{u-u} and S_{t-t} are user similarity matrix and message similarity matrix respectively and the elements in both matrices are nonnegative real numbers. U and V are the binary cluster matrices for users and messages, i.e., $U(i,j) = 1$ denotes that the ith user belongs to jth cluster and $V(i,j) = 1$ denotes the ith message belongs to jth cluster. W is the word (soft-)cluster matrix in which the elements are the real values since we do not constrain that one word must belong to only one cluster. R is the binary interaction matrix in which $R(i,j) = 1$ denotes user i has interacted with user j and $R(i,j) = 0$ otherwise. With the notations introduced above, the community detection problem in this paper is formally defined as follows:

Problem 1 *The community detection problem is defined as a co-cluster problem using constrained nonnegative matrix tri-*

Table 1: Notations used in this paper and the corresponding explanations and dimensions.

Notations	Explanations	Dimension
m	number of users	-
n	number of messages	-
w	number of words	-
k	number of communities	-
M_{u-u}	user relation matrix	$m \times m$
M_{u-f}	user-words matrix	$m \times w$
M_{t-f}	message-words matrix	$n \times w$
S_{u-u}	user similarity matrix	$m \times m$
S_{t-t}	message similarity matrix	$n \times n$
U	user cluster matrix	$m \times k$
V	message cluster matrix	$n \times k$
W	word (soft-)cluster matrix	$w \times k$
R	interaction matrix	$m \times m$
$H_1/H_2/H_3$	associated matrix	$k \times k$
L^u	Laplacian matrix for user	$m \times m$
L^t	Laplacian matrix for message	$n \times n$
L^r	Laplacian matrix for interaction	$m \times m$

factorization (NMTF):

$$\begin{aligned} &\min_{U,V,W,H_1,H_2,H_3} \|M_{u-u} - UH_1U^T\|_F^2 \quad (1)\\ &+ \|M_{t-f} - VH_2W^T\|_F^2 + \|M_{u-f} - UH_3W^T\|_F^2 \\ &s.t.\ UU^T = I, VV^T = I \end{aligned}$$

where α, β, and γ are the parameters[1] *to control the proportion of the three types of graph regularization in the optimization problem, $tr(\cdot)$ is the trace function and $\|\cdot\|_F$ denotes the Frobenuis norm of a matrix.*

Note that in Eq (1), the user cluster matrix U is most important because its element $U(i,j)$ denotes whether the ith user belongs to the jth community. The component $\|M_{u-u} - UH_1U^T\|_F^2$ in line 1 captures the user relation to measure the distance of the ideal user clusters from the real user clusters. The components in line 2 captures the tripartite graph shown in Figure 1. Since the $X - Y - Z$ relation in a tripartite graph can be represented by $X - Y$ and $Y - Z$ relations in two bipartite graphs [Cheng *et al.*, 2007; Zhu *et al.*, 2014], we divide the user-word-message relation in the tripartite graph into user-word relation and message-word relation corresponding to the term $\|M_{u-f} - UH_3W^T\|_F^2$ and $\|M_{t-f} - VH_2W^T\|_F^2$ respectively. The constraints $UU^T = I, VV^T = I$ are used to ensure that matrices U and V can represent the clusters of users and messages, and a user or message can only belong to one cluster. Different from NMTF for sentiment classification [Li *et al.*, 2009; Zhu *et al.*, 2014], since words used in messages can belong to multiple communities in community detection rather than two polarities in sentiment analysis, we do not include constraint on word cluster matrix W in our study.

Based on the problem definition above, the community detection essentially is a user and message co-clustering problem in which user and message similarity may play an important role. However, in Eq (1), both user and message similarity are not captured explicitly. Besides, user interactions are also good indicators for clustering users. Also, the experimental results which is shown in Table 3 demonstrate only the NMTF cannot perform well in community detection. Therefore, we propose three types of graph regularization to capture user similarity, message similarity and user interaction, respectively. By integrating these graph regularization into Eq (1), the new optimization problem becomes:

$$\begin{aligned} &\min_{U,V,W,H_1,H_2,H_3} \|M_{u-u} - UH_1U^T\|_F^2 \quad (2)\\ &+ \|M_{t-f} - VH_2W^T\|_F^2 + \|M_{u-f} - UH_3W^T\|_F^2 \\ &+ \alpha \cdot tr(U^T L^u U) + \beta \cdot tr(V^T L^t V) + \gamma \cdot tr(U^T L^r U) \\ &s.t.\ UU^T = I, VV^T = I \end{aligned}$$

The details of the components in this optimization problem will be presented in Section 4.

4 Proposed Method

In this section, the proposed NMTF based clustering framework with different types of regularization for community detection is presented. The proposed method consists of the NMTF based clustering component, user similarity based regularization, message similarity based regularization and interaction based regularization. The details about these components are presented as follows.

NMTF based clustering component. The NMTF based clustering component is applied to co-cluster users and messages by combining social relations and content. This component can be represented as

$$\begin{aligned} &\min_{U,V,W,H_1,H_2,H_3} \|M_{u-u} - UH_1U^T\|_F^2 \quad (3)\\ &+ \|M_{t-f} - VH_2W^T\|_F^2 + \|M_{u-f} - UH_3W^T\|_F^2 \\ &s.t.\ UU^T = I, VV^T = I \end{aligned}$$

The NMTF based clustering component consists of two parts: (1) user relation part (Line 1 in Eq (3)); and (2) user-word-message relation part (Line 2 in Eq (3)) shown in Figure 1 . By integrating these two parts, the social relations and content can be combined. The constraints $UU^T = I, VV^T = I$ are used to ensure that matrices U and V can represent the clusters of users and messages, and a user/message can only belong to one cluster.

User similarity based regularization. Intuitively two users that are very similar are more likely to belong to the same community. Formally, the user similarity based regularization is represented as:

$$S_{u-u}(i,j)\|l_i^u - l_j^u\|_F^2 \quad (4)$$

where $S_{u-u}(i,j)$ denotes the similarity between user u_i and u_j and l_i^u and l_j^u denote the community which user u_i and u_j belong to, respectively. It is easy to transform this formula into matrix format as follows:

$$\frac{1}{2}\sum_i\sum_j S_{u-u}(i,j)\|l_i^u - l_j^u\|_F^2 = tr(U^T L^u U) \quad (5)$$

where $L^u = D^u - S_{u-u}$ is the Laplacian matrix of the user similarity based graph. D^{uu} is the degree matrix of S_{u-u} and it is a diagonal matrix.

[1] In the experiments, the three parameters are set to be 0.5 and the parameter tuning is omitted due to the page limitation.

Message similarity based regularization. Similarly, messages with similar content should be categorized into the same cluster. So the message similarity based regularization is defined as:

$$S_{t-t}(i,j)\|l_i^t - l_j^t\|_F^2 \tag{6}$$

where $S_{t-t}(i,j)$ denotes the similarity between message t_i and t_j. l_i^t and l_j^t denote the cluster which message t_i and t_j belong to, respectively. Similarly, the matrix format is:

$$\frac{1}{2}\sum_i\sum_j S_{t-t}(i,j)\|l_i^t - l_j^t\|_F^2 = tr(V^T L^t V) \tag{7}$$

where $L^t = D^t - S_{t-t}$ is the Laplacian matrix of the message similarity based graph and D^{uu} is the degree matrix of S_{t-t}.

Interaction based regularization. Based on the definition of a community introduced in Introduction, interaction is an effective indicator to determine the community for a user. It is also straightforward that if two users have interaction, they are more likely to belong to the same community. Therefore, the interaction based regularization is represented as:

$$R(ij)\|l_i^u - l_j^u\|_F^2 \tag{8}$$

where $R(ij)$ denotes the user interaction and l_i^u and l_j^u denote the community user u_i and u_j belong to, respectively. Then, the matrix form is calculated as:

$$\frac{1}{2}\sum_i\sum_j R_{ij}\|l_i^u - l_j^u\|_F^2 = tr(U^T L^r U) \tag{9}$$

where $L^r = D^r - R$ is the Laplacian matrix of the interaction based graph and D^{uu} is the degree matrix of R.

Now combining all the components, we have the objective function shown in Eq (2) in Section 3.

4.1 Learning Algorithm

The optimal solution to the optimization problem in Eq (2) can be achieved using an iterative update algorithm [Ding *et al.*, 2006] and the updating rules are shown as follows.

$$U \leftarrow U\circ \sqrt{\frac{M_{u-u}UH_1^T + M_{u-f}WH_3^T + \alpha S_{u-u}U + \gamma RU}{UH_1U^TUH_1^T + UH_3W^TWH_3^T + \alpha D^uU + \gamma D^rU + U\Psi_U}} \tag{10}$$

$$V \leftarrow V \circ \sqrt{\frac{M_{t-f}WH_2^T + \beta S_{t-t}V}{VH_2W^TWH_2^T + \beta D^tV + V\Psi_V}} \tag{11}$$

$$W \leftarrow W \circ \sqrt{\frac{M_{t-f}^TVH_2 + M_{u-f}^TUH_3}{WH_2^TV^TVH_2 + WH_3^TU^TUH_3}} \tag{12}$$

where $\Psi_U = U^TM_{u-u}UH_1^T + U^TM_{u-f}WH_3^T - H_1U^TUH_1^T - H_3W^TWH_3^T - \alpha U^TL^uU - \gamma U^TL^rU$ and $\Psi_V = V^TM_{t-f}WH_2^T - H_2W^TWH_2^T - \beta V^TL^tV$.

$$H_1 \leftarrow H_1 \circ \sqrt{\frac{U^TM_{u-u}U}{U^TUH_1U^TU}} \tag{13}$$

$$H_2 \leftarrow H_2 \circ \sqrt{\frac{V^TM_{t-f}W}{V^TVH_2W^TW}} \tag{14}$$

$$H_3 \leftarrow H_3 \circ \sqrt{\frac{U^TM_{u-f}W}{U^TUH_3W^TW}} \tag{15}$$

where $\circ$ denotes element-wise product, $\frac{[\cdot]}{[\cdot]}$ denotes element-wise division and $\sqrt{\cdot}$ denotes element-wise square root.

With these updating rules, the optimization algorithm for the optimization is presented in Algorithm 1. In this algorithm, we update one matrix and fix the other matrices in each step (Line 3-8) and the iterative process is stopped if these cluster matrices converge or the number of iteration exceeds a given threshold.

Algorithm 1 Optimization Algorithm

Input:
user relation matrix M_{u-u}, user-word matrix M_{u-f}, message-word matrix M_{t-f}
user similarity matrix S_{u-u}, message similarity matrix S_{t-t}, interaction matrix R
parameters: α, β, and γ

Output:
user cluster matrix U and message cluster matrix V

1: initialize $U, V, W, H_1, H_2, H_3 \geq 0$
2: **while** not converge **do**
3: update U according to Eq. (10)
4: update V according to Eq. (11)
5: update W according to Eq. (12)
6: update H_1 according to Eq. (13)
7: update H_2 according to Eq. (14)
8: update H_3 according to Eq. (15)
9: **end while**

4.2 Complexity Analysis

In this method, the major operations are the matrix multiplication. For convenience, we assume that the time complexity of multiplication for two matrices, e.g., a $m \times k$ matrix and a $k \times n$ matrix, is $O(mkn)$. Therefore, the time complexity for Algorithm 1 is $O(rk(mn + mw + nw + m^3 + n^2))$ where r is the iteration times. m, n, k, and w denote the number of users, messages, features and communities respectively which are shown in Table 1.

5 Similarity Measure

As introduced in Section 1, how to model similarity between users and messages is an important issue in community detection. However, conventional relation based user similarity and word based message similarity cannot perform well in social media. For example, two users should be similar if they published similar tweets even there is no relation between them. And two messages towards the same topic should be similar if they were published by users who are friends even there are not many overlapping words in these messages. Therefore, in this section, we propose a novel measure to calculate user similarity and message similarity by fusing user relations, user-message relations and message relations. The user similarity S_{u-u} and message similarity S_{t-t} are applied in the user similarity based regularization and message similarity based regularization introduced in Eq. (5) and Eq. (7), respectively.

First, we build a two-layer graph based on user relation, user-message relation and message relation. For the user layer, the link between two users denotes the social relation

between these two users, e.g., friendship in Twitter. For the message layer, the link denotes the cosine similarity between two messages exceeds a given value. In detail, motivated by [Mihalcea and Tarau, 2004] which constructs a graph based on the content similarity, if the similarity between two messages exceeds a given threshold, there will be a link between these two messages and the basic message similarity is calculated using standard cosine similarity:

$$sim_{cosine}(\mathbf{w}_i, \mathbf{w}_j) = \frac{\mathbf{w}_i \cdot \mathbf{w}_j}{|\mathbf{w}_i| \times |\mathbf{w}_j|} \tag{16}$$

where $\mathbf{w}_i$ and $\mathbf{w}_j$ to denote the feature vectors for message t_i and t_j. Element w_{ij} in $\mathbf{w}_i$ is set to be 1 message t_i contains the jth word and 0 otherwise. The link between user layer and message level indicates a user publishes a message.

Then we propose a novel random walk method to calculate the user similarity and message similarity based on the two-layer graph. Traditional PageRank [Page *et al.*, 1999] is:

$$p^{(t+1)} = (1-\alpha)p^{(t)}M + \alpha q \tag{17}$$

where M denotes the transition matrix, $p^{(t)}$ denotes the vector of PageRank value at (t)th iteration and α is the damping factor. q is set to be the vector with the same value $1/N$ where N is the number of nodes in the graph. PageRank can also be used to calculate the similarity between nodes. In this scenario, q denotes the information of the query node and we can calculate the similarity between the query node and all the other nodes using PageRank. By setting every node as the query node in each time, the similarity between any two nodes can be obtained. We generalize this method to the two-layer graph. Given three types of relations, i.e., E_{uu}, E_{ut} and E_{tt} indicate user relation, user-message relation and message relation respectively, the transition matrix $Tran$ in the two-layer graph is defined as:

$$Tran = \begin{pmatrix} E_{uu} & E_{ut} \\ E_{ut}^T & E_{tt} \end{pmatrix}$$

where all elements in these three matrices are binary, i.e., the element is 1 if there is a link between two nodes and 0 otherwise. Then given user a as the query, the vector for the query node q_a is defined as:

$$q_a = \begin{pmatrix} E_{uu}^T(a) \\ E_{ut}^T(a) \end{pmatrix}$$

where $E_{uu}(a)$ and $E_{ut}(a)$ denote the ath row of matrix E_{uu} and E_{ut}, i.e., the relation between user a and other users and the relation between user a and all the messages. Therefore, given a user a, the random walk based similarity vector $p_a^{(t)}$ between a and other users can be calculated as:

$$p_a^{(t+1)} = (1-\alpha)p_a^{(t)}Tran + \alpha q_a \tag{18}$$

where α is the e damping factor. Similarly, the new message similarity between message e and other messages can be calculated as:

$$p_e^{(t+1)} = (1-\alpha)p_e^{(t)}Tran + \alpha q_e \tag{19}$$

where q_e is the query vector for message e and $p_e^{(t)}$ is denotes the similarity value between e and other messages at (t)th iteration.

6 Experiments

6.1 Data Collection

In the experiments, we use three data sets including two Twitter data sets[2] [Greene and Cunningham, 2013], i.e., Politics-UK and Politics-IE, and one bibliography data set, i.e., DBLP[3]. They are described as follows:

Table 2: A brief statistics of the data sets.

	Politics-UK	Politics-IE	DBLP
# of users	413	331	6604
# of messages	72693	49546	8293
# of words	7314	5805	2110
# of communities	5	7	4
# of social relations	37369	20253	17029
# of interactions	3271	3122	0

Politics-UK: This data set consists of 419 Members of Parliament from the United Kingdom and they belong to five different political groups.

Politics-IE: This data set has 348 Irish politicians and political organizations. These user are assigned to seven disjoint groups according to their affiliation.

DBLP: This data set contains 6604 authors 8293 papers from 16 top conferences which cover 4 research fields including *Machine Learning*, *Information Retrieval*, *Data Mining* and *Database*.

For each user in Politics-UK and Politics-IE data sets, we collected her social relations, i.e., the users she follows and the users that follow her, and her most recent 200 tweets. For each data set, users and tweets are preprocessed in following steps: (1) removing the users who have not published tweets; (2) remove the non-English tweets; (3) removing the stop words in the tweets; and (4) keep the words which occur more than 10 times in the data set as the features. For DBLP data set, we use the paper titles as the content and co-author relation as the social relation. Since there is no citation information in DBLP data set, we ignore the interaction regularization, i.e., set $\gamma = 0$ in Eq (1), for this data. Similarly, the content is preprocessed by removing stop words and words occurring less than 5 times. After preprocessing, a brief statistics of the data sets is shown in Table 2.

6.2 Baseline

In order to demonstrate the effectiveness of our method, three types of community detection methods, i.e., the relation-only methods, the content-only methods and the methods use the combination of relations and content, are compared in this study. These methods are introduced as follows:

Relation-only method. Two relation-only methods have been used in the comparison, i.e., Girvan-Newman algorithm [Girvan and Newman, 2002] and Louvain method [Blondel *et al.*, 2008]. In these relation-only methods, we use the social

[2]The data sets can be found at http://mlg.ucd.ie/aggregation/index.html. We only use the user lists provided in [Greene and Cunningham, 2013] and the information including social relations and tweets are collected via Twitter API.

[3]http://www.informatik.uni-trier.de/~ley/db/

Table 3: The *Purity* on Politics-UK, Politics-IE and DBLP data sets.

	Method	Politics-UK	Politics-IE	DBLP
Relation-only	Girvan-Newman	0.6683	0.6733	0.7548
	Louvain	0.6852	0.6770	0.7462
Content-only	Kmeans	0.5981	0.5776	0.6845
	LDA	0.6247	0.6408	0.6692
Relations + content	RTM	0.6949	0.6859	0.7706
	NMTF	0.6923	0.6805	0.7822
	NMTF + regularization	**0.7453**	**0.7322**	**0.8014**

relations, i.e., following relation in Twitter and co-author relation in DBLP, to construct the graph and then partition the graph to detect communities.

Content-only method. We use Kmeans and LDA based clustering method as the baselines in content-only methods. In these methods, all the messages published by a user are viewed as one document and then the similarity between two users are measured by the similarity between two documents belong to each user. The standard cosine similarity is applied for the similarity calculation. In Kmeans, the word list for a user is used as the feature vector and in LDA the topic distribution for a user is used as the feature vector.

Combination of relation and content. In the type of methods which use the combination of relations and content, we use the Relational Topic Model (RTM) [Chang and Blei, 2009] as the baseline which models the link between two as a binary random variable conditioned on the contents. Additionally, to validate the effectiveness of the regularization, we also compare the performance of the NMTF based clustering method without regularization in the comparison.

6.3 Evaluation Measures

In this study, *Purity* is applied to measure the quality of the communities detected by the approaches and the *Purity* is widely used in evaluating the performance of community detection [Lin *et al.*, 2012].

The *Purity* is defined as: each cluster is first assigned with the most frequent class in the cluster, and then the purity is measured by computing the number of instances assigned with the same labels in all clusters [Lin *et al.*, 2012]. Formally, let $\mathcal{C} = \{C_i, \ldots, C_k\}$ be the k communities detected by the algorithm and $\mathcal{G} = \{l_i, \ldots, l_t\}$ be the set of communities in the ground truth. The *Purity* is calculated as:

$$Purity = \frac{1}{n} \sum_{i=1}^{k} \max_j |C_i \cap l_j| \tag{20}$$

where n is the number of instances in the data set. The value of *purity* ranges from 0 to 1 and the higher purity value means better performance.

6.4 Experimental Results

The *Purity* scores for different methods in three data sets are shown in Table 3. Some conclusions can be drawn from the results reported in the table.

The proposed NMTF based clustering method with graph regularization performs best among all the methods in all three data sets. For example, the proposed method can get 7% improvement compared with the RTM method which is the second-best method, and the improvement is about 20% compared with the content-only method in **Politics-UK** data set. The methods which combine the relations and content perform better than the methods use relations or content only. Even removing the regularization, the NMTF based clustering method performs better than relation-only and content-only methods.

An interesting observation is that relation-only methods perform better than the content-only methods in all data sets. This result may due to the following reasons: (1) It is intuitive that the social relations reflect the user interests directly. For example, an Obama supporter will be more likely to follow the Democrats. Therefore, social relations can serve as a good indicator for communities. (2) The user generated content in social networks such as Twitter is diverse, and therefore detect communities from only content may be influenced by the diverse and noisy information. In DBLP data set, we use the paper titles as the messages and these short texts may not profile authors well.

7 Conclusions

In this paper, we propose a NMTF based clustering framework with three types of regularization for community detection in social networks. The NMTF based clustering framework can capture the relations in the user-word-message tripartite graph and the three types of regularization explicitly model the user similarity, message similarity and user interaction, respectively. This method integrates social relations and content seamlessly, is flexible in incorporating different social information and model user similarity based on both relations and content explicitly. Experiments on three real-world data sets have been conducted to validate the performance of the proposed method and experimental results illustrated the effectiveness of our method. For the future work, we plan to exploit more types of regularization such as topic and sentiment information from user and message level. We also plan to design faster learning algorithm for the matrix tri-factorization.

Acknowledgments

This work has been partially supported by ARO award W911NF-13-1-0416.

References

[Blondel *et al.*, 2008] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[Chang and Blei, 2009] Jonathan Chang and David M Blei. Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics*, pages 81–88, 2009.

[Cheng *et al.*, 2007] Haibin Cheng, Pang-Ning Tan, Jon Sticklen, and William F Punch. Recommendation via query centered random walk on k-partite graph. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 457–462. IEEE, 2007.

[Ding *et al.*, 2006] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135. ACM, 2006.

[Fortunato, 2010] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[Girvan and Newman, 2002] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[Greene and Cunningham, 2013] Derek Greene and Pádraig Cunningham. Producing a unified graph representation from multiple social network views. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 118–121. ACM, 2013.

[Gu and Zhou, 2009] Quanquan Gu and Jie Zhou. Co-clustering on manifolds. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 359–368. ACM, 2009.

[Lee and Seung, 2001] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[Lee *et al.*, 2013] Kyumin Lee, James Caverlee, Zhiyuan Cheng, and Daniel Z Sui. Campaign extraction from social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):9, 2013.

[Li *et al.*, 2009] Tao Li, Yi Zhang, and Vikas Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 244–252. Association for Computational Linguistics, 2009.

[Lin *et al.*, 2012] Wangqun Lin, Xiangnan Kong, Philip S Yu, Quanyuan Wu, Yan Jia, and Chuan Li. Community detection in incomplete information networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 341–350. ACM, 2012.

[Mihalcea and Tarau, 2004] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. Association for Computational Linguistics, 2004.

[Newman, 2006] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[Page *et al.*, 1999] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[Ruan *et al.*, 2013] Yiye Ruan, David Fuhry, and Srinivasan Parthasarathy. Efficient community detection in large networks using content and links. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1089–1098. International World Wide Web Conferences Steering Committee, 2013.

[Sachan *et al.*, 2012] Mrinmaya Sachan, Danish Contractor, Tanveer A Faruquie, and L Venkata Subramaniam. Using content and interactions for discovering communities in social networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 331–340. ACM, 2012.

[Smola and Kondor, 2003] Alexander J Smola and Risi Kondor. Kernels and regularization on graphs. In *Learning theory and kernel machines*, pages 144–158. Springer, 2003.

[Tang and Liu, 2010] Lei Tang and Huan Liu. Community detection and mining in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1):1–137, 2010.

[Wang *et al.*, 2011] Fei Wang, Tao Li, Xin Wang, Shenghuo Zhu, and Chris Ding. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery*, 22(3):493–521, 2011.

[Yang *et al.*, 2009] Tianbao Yang, Rong Jin, Yun Chi, and Shenghuo Zhu. 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*, pages 927–936. ACM, 2009.

[Zhu *et al.*, 2014] Linhong Zhu, Aram Galstyan, James Cheng, and Kristina Lerman. Tripartite graph clustering for dynamic sentiment analysis on social media. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 1531–1542. ACM, 2014.