Fuzzy Community Detection with Multi-View Correlated Topics

Lizhong Yao*
yaolizhong225@163.com
CPEE, Chongqing Normal University
SCE, Chongqing University of Science and Technology
China P.R.

Tiantian He*
he_tiantian@ihpc.a-star.edu.sg
IHPC, A*STAR
SCSE, Nanyang Technological University
Singapore

ABSTRACT

In this paper, we present a novel fuzzy framework, dubbed as Fuzzy Multi-View Featured Network Clustering (FMVFNC), for effectively uncovering overlapping communities in social network data. Unlike most previous efforts which utilize only edge structure and single view of vertex features to perform the community discovery task, the proposed FMVFNC is able to take advantage of both edge structure and correlated vertex features which may be collected from multiple views. As the uncovered social communities are described by both network structure and semantically correlated features from diverse modalities, their practical significance can be well revealed. We innovatively design a unified fuzzy objective for FMVFNC to perform the task. We then derive an iterative algorithm for the proposed framework to optimize the formulated objective function. FMVFNC has been tested with a number of well-established datasets and has been compared with a number of state-of-the-art baselines for community detection. The notable results obtained may validate the effectiveness of FMVFNC.

CCS CONCEPTS

• **Computing methodologies → Non-negative matrix factorization**; **Cluster analysis**; *Vagueness and fuzzy logic.*

KEYWORDS

Fuzzy clustering, Community detection, Matrix factorization, Graph clustering

ACM Reference Format:
Lizhong Yao and Tiantian He. 2021. Fuzzy Community Detection with Multi-View Correlated Topics. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT '21 Companion), December 14–17, 2021, ESSENDON, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3498851.3498971

1 INTRODUCTION

As a prime sub-category of network data, social networks can be used to effectively model the complex relations from every corner of social lives. For example, representing social network users and their social ties as nodes (vertices) and edges, a social network can be used to describe the online interactions among a group of people. Different from those artificially constructed networks, meaningful latent structures are always hidden in the social network. And discovering these latent structures may bring a better understanding to human social lives from a relational setting. Aiming at revealing the group cohesiveness among people in the social network, analyzing social communities (social network/graph clustering) has recently drawn an increasing interest and has become one of most important analytical tasks in social networks.

Community detection in social networks has been one of the most active areas in data mining and machine learning. Quite a few approaches have been proposed to tackle this problem. According to which kind of data are used to discover social communities, these methods can generally be categorized as structure-based ones, and feature-structure-based ones.

Structure-based approaches to community detection can perform the task by utilizing various topological properties of the network. As the cornerstone of the network, edge structure has been one of the major choices for detecting communities in social network data. Heuristic measures, such as modularity [3], and clique percolation [17] have been successfully used to perform the community discovery task. Besides, more model-based methods make use of different machine learning techniques to fulfill the task of community detection. For example, Bayesian inference [18, 19], Matrix factorization [22], spectral analysis [20], and information propagation [4] have been adopted to build different models to learn community structure from social networks.

In contrast, feature-structure-based methods additionally take into the consideration content features that are associated with vertices in the network, when unfolding communities in the social network. As two sources of data, i.e., edge structure and vertex features are concerned, feature-structure-based approaches always perform the task of community discovery, following a co-learning strategy. Many learning paradigms that are recently proposed, e.g., joint matrix factorization [8], co-spectral analysis[12], topic modeling [2], and latent block modeling [6] can be applied to learn community structure that is shared by edge structure and vertex features.

Although these proposed approaches are effective in uncovering communities in social network data, most of them still face the following problems. First, most feature-structure-based approaches to community detection assume the vertex features are collected from a single source, which might contradict the well-known fact that the vertex can be characterized by features collected from diverse sources. A typical example (see Fig. 1) can be found in the online social networking sites, where the users always have different categories (views) of features, e.g., social tags, comments,

[*] Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WI-IAT '21 Companion, December 14–17, 2021, ESSENDON, VIC, Australia
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-9187-0/21/12…$15.00
https://doi.org/10.1145/3498851.3498971

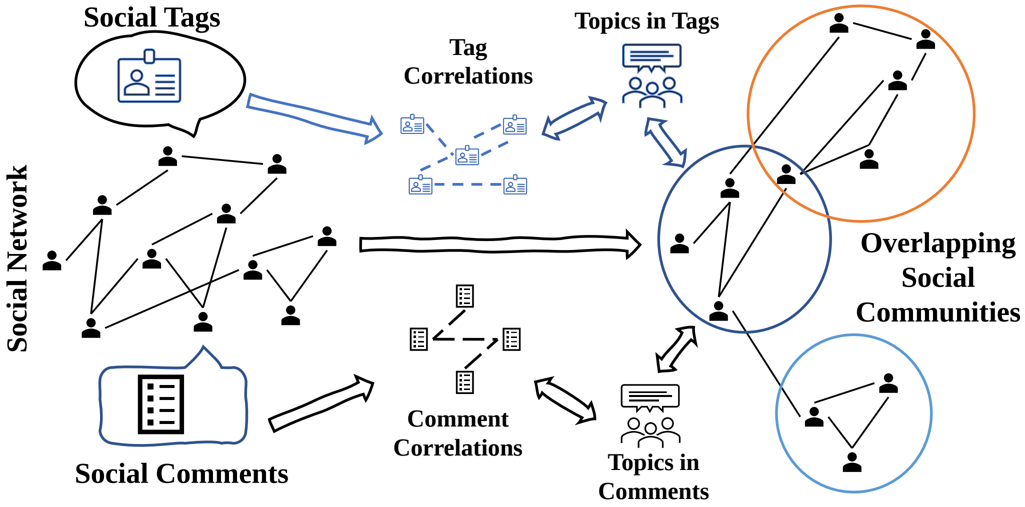

Figure 1: Graphical illustration on how FMVFNC detects social communities and a social community with multi-view correlated topics that can be detected by the proposed framework.

profiles, and hobbies. As these features describe the social network users from different perspectives, possibly they influence differently the community affiliation concerned by each user. However, such effect from multiple views of features is under-explored by most existing methods as they concatenate multi-view features as single-view ones, following the aforementioned assumption. Second, most existing approaches are incapable of uncovering overlapping groups from social networks, which deters them from obtaining better analytical performances. Overlapping communities widely exist in social network data as human-beings may play various roles online and consequently may belong to different social groups.

We propose a novel framework in this paper, labeled here as Fuzzy Multi-View Featured Network Clustering (FMVFNC) to overcome the aforementioned problems. Different from previous approaches, FMVFNC may effectively capture the influence brought by both edge structure and multi-view vertex features, when discovering communities in social network. Moreover, FMVFNC is capable of discovering overlapping communities by adopting fuzziness when inferring the community affiliation for each vertex in the network. In Fig. 1, how FMVFNC formulates the problem of community detection in social networks is graphically illustrated. Given a social network where vertices possess multi-view features, FMVFNC attempts to detect a predetermined number of social communities, which are possibly assigned with a particular number of social topics. For each view, FMVFNC pre-computes the contextual correlations in terms of vertex features, and further assumes social topics in each view are learned from the feature-feature correlations. Appropriately integrating above sub learning processes as a unified objective, the proposed FMVFNC can learn the community affiliation for each vertex which is jointly determined by edge structure, and multi-view correlated social topics constituted by multi-view vertex features. As a result, meaningful community structure can be identified by the proposed framework and its analytical performance can be further improved. The main contributions of the paper are summarized as follows.

- We propose FMVFNC, a novel framework for discovering overlapping communities in social network data. Different from most existing methods, the proposed framework is able to take advantage of both edge structure and correlated social topics learned from multi-view vertex features to perform the task of community discovery. As the view-wise effect brought by multi-view vertex features is considered, FMVFNC is able to discover more meaningful community structures.
- We formulate a generic objective function for the detection problem and let FMVFNC use it to qualify the process of community discovery. An iterative algorithm is then derived to optimize the formulated objective function.
- The proposed framework has been tested with several well-established datasets and compared with a number of prevalent baselines. The obtained results demonstrate the effectiveness of FMVFNC.

The rest of the paper is organized as follows. In Section 2, existing approaches to social community detection are briefly investigated. In Section 3, we elaborately introduce the proposed framework, i.e., Fuzzy Multi-View Featured Network Clustering (FMVFNC). In Section 4, the iterative algorithm for optimizing the proposed framework is derived and how FMVFNC extracts overlapping communities is introduced. How FMVFNC is tested and compared with prevalent baselines is introduced in Section 5. At last, we summarize the paper and conclude the future works which may further improve FMVFNC.

2 RELATED WORKS

Community detection has been an active research area, where many approaches have been proposed. Some of them can perform the task of community detection using edge structure in the network. For example, Clauset-Newman-Moore algorithm [3] and Fast unfolding [1] are classical approaches that are based on modularity optimization. Normalized cut [20] is an effective method for community detection, which is based on spectral analysis. Symmetric non-negative matrix factorization [22], and Hidden community detection [5] are popular approaches that are based on matrix factorizations. Stochastic block models [18, 19] are effective Bayesian methods for unfolding communities in social network data.

In contrast, many other approaches to community detection perform the task by concerning edge structure and single view of vertex features. For example, Community detection with node attributes [24], Semantic community identification [23], Mining interesting sub-graphs [8], and Vicinal vertex allocated matrix factorization [7] are effective matrix-factorization-based methods which can discover social communities using both edge structure and vertex features. Relational topic model [2], Manifold regularized block model [6], Content propagation Markov clustering [15], and Generative community detection [26] are popular probabilistic models for community detection. Discovery fuzzy structural patterns [9] and Fuzzy clustering algorithm for complex networks [11] are effective fuzzy clustering methods that can make use of both edge structure and vertex features to uncover communities in network data.

It is found that most existing methods for social community detection do not tackle the problem by concerning multi-view vertex features and fuzzy clustering techniques. This motivates us to propose the novel framework, Multi-View Featured Network Clustering (FMVFNC).

3 MULTI-VIEW FUZZY CLUSTERING IN SOCIAL NETWORK

In this section, the proposed Multi-View Featured Network Clustering (FMVFNC) is elaborated. Mathematical notations and preliminaries are firstly introduced. How the proposed FMVFNC formulates the task of community detection as an optimization problem, is illustrated next.

3.1 Mathematical preliminaries and notations

In this paper, we assume a social network contains N vertices, |E| edges, D views of vertex features, and K ground-truth social communities. In each view, there are M^i features, and $\sum_{i=1}^{D} M^i = M$. To represent the social network, we use two binary matrices, $\mathbf{Y} \in \{0,1\}^{N\times N}$ and $\mathbf{F}^i \in \{0,1\}^{M^i\times n}$ to stand for edge structure, and multi-view vertex features, respectively. To represent the aforementioned correlations among features in each view, we use an $M^i \times M^i$ matrix $\mathbf{X}^i$. $N \times K$, $M^i \times T$, and $T \times K$ matrices $\mathbf{V}$, $\mathbf{U}^i$, and $\mathbf{W}^i$ are used to represent the core variables for tackling the community detection problem, i.e., fuzzy vertex-community affiliation, social topics in each view, and corresponding topic-community affiliation. $\mathbf{X}_{ij}$ means the (i, j)-th element in matrix $\mathbf{X}$. The Frobenius norm is denoted as $\|\cdot\|_F$.

3.2 Edge modeling

As Fig. 1 depicts, FMVFNC simultaneously tackles two sub-tasks, i.e., edge structure and multi-view correlated topic modeling, so as to infer the community structure for each vertex in the social network. How FMVFNC solves the community detection problem given the aforementioned two sub-tasks will be mathematically illustrated here with more details.

Like most previous approaches to community detection, FMVFNC also models the edge structure for the community discovery task. To achieve this, FMVFNC follows the empirical method for edge structure modeling, i.e., each edge in Y is approximated by the community membership $\mathbf{V}$. Thus, we have the following loss function:

$$O_1 = \left\|\mathbf{Y} - \mathbf{V}\mathbf{V}^T\right\|_F^2. \tag{1}$$

By minimizing O_1 in Eq. (1), FMVFNC is able to infer the optimal community affiliation concerning the edge structure.

3.3 Multi-view correlated topic modeling

3.3.1 Identifying feature-feature correlations. As aforementioned, FMVFNC pre-computes $\mathbf{X}^i$ to capture the correlations among the features in each view, which is then used to model the correlated social topics. In this paper, FMVFNC adopts a simple but effective strategy, the normalized shifted point-wise mutual information (SPMI) method [14] to directly acquire the multi-view feature-feature correlations. Given two features, say f_j and f_k in view i, their correlations can be obtained by computing the normalized entropy that these two features are simultaneously observed:

$$\begin{aligned}
\mathbf{X}^i_{jk} &= \begin{cases} \frac{PMI(f_j,f_k)}{H(f_j,f_k)} & \text{if } PMI(f_j,f_k) \geq log\delta,\\ 0 & \text{otherwise,}\end{cases}\\
PMI(f_j,f_k) &= log\frac{obs(f_j,f_k)\cdot|obs|}{obs(f_j,+)\cdot obs(f_k,+)},\\
H(f_j,f_k) &= -log\frac{obs(f_j,f_k)}{|obs|},
\end{aligned} \tag{2}$$

where $obs(f_j, f_k)$, $obs(f_j, +)$, and $|obs|$ respectively represent the number of times that a vertex possesses both f_j and f_k, the number of times that a vertex possesses f_j, and the total number of times that a vertex possesses any pair of features. Previous works have proved that $\mathbf{X}^i \geq \log\delta$ and the value learned through word2vec with the negative sample value of δ are equivalent [14]. Thus, $\mathbf{X}^i$ may describe how a pair of features are correlated from the contextual perspective. Given $\mathbf{X}^i$, FMVFNC is possible to infer multi-view social topics that are constituted by those correlated features. To allow FMVFNC to use more feature-feature correlations, in this paper, we simply set δ to 1.

3.3.2 Correlated topic modeling. Given $\mathbf{F}^i$, FMVFNC assumes that there are T social topics among these features ($\mathbf{U}^i$) and these topics can be assigned to K communities ($\mathbf{W}^i$). Besides, FMVFNC further assumes that the T social topics in each view can also be learned from the corresponding feature-feature correlations ($\mathbf{X}^i$). Thus, we have $\mathbf{X}^i_{jk} \approx [\mathbf{U}^i\mathbf{U}^{iT}]_{jk}$ and $\mathbf{F}^i_{jk} \approx [\mathbf{U}^i\mathbf{W}^i\mathbf{V}^T]_{jk}$. And the topic modeling problem in each view of vertex features can be formulated as the following loss function:

$$O_2 = \sum_i [\left\|\mathbf{F}^i - \mathbf{U}^i\mathbf{W}^i\mathbf{V}^T\right\|_F^2 + \left\|\mathbf{X}^i - \mathbf{U}^i\mathbf{U}^{iT}\right\|_F^2]. \tag{3}$$

As Eq. (3) shows, O_2 allows the vertex-community affiliation to be learned from multi-view social topics, which are jointly inferred from multi-view vertex features and their corresponding correlations.

3.4 Unified objective function

Combining the two aforementioned sub-tasks and plugging in the fuzzy constraints and balancing parameters, we have the unified objective function for FMVFNC to detect communities in social networks:

$$\begin{aligned}&\text{minimize}\\ &O=\sum_i[\left\|\mathbf{F}^i-\mathbf{U}^i\mathbf{W}^i\mathbf{V}^T\right\|_F^2+\left\|\mathbf{X}^i-\mathbf{U}^i\mathbf{U}^{iT}\right\|_F^2]+\alpha\left\|\mathbf{Y}-\mathbf{V}\mathbf{V}^T\right\|_F^2,\\ &\text{subject to } \mathbf{V}\geq 0, \mathbf{U}^i\geq 0, \mathbf{W}^i\geq 0, \sum_k \mathbf{V}_{ik}=1, \text{for all } i,\end{aligned} \tag{4}$$

where α is a positive parameter used to adjust the relative significance between edge and correlated topic modeling. Minimizing the objective function above allows FMVFNC to infer the optimal fuzzy community membership for all the vertices from edge structure and multi-view correlated social topics constituted by contextually relevant features.

Based on the proposed unified objective function, FMVFNC can be evidently distinguished from existing works in the following two aspects. First, the generic objective function enables FMVFNC to model the problem of community detection using both edge structure and multi-view correlated social topics learned from multi-view vertex features. Such a learning paradigm is seldom considered by previous approaches to community detection. Second, the consideration of fuzzy community membership allows FMVFNC to discover overlapping communities in the social network, which further improve the analytical capability of FMVFNC and consequently the practical significance of the framework.

4 MODEL OPTIMIZATION AND LEARNING COMMUNITIES

4.1 Algorithm for optimization

The proposed objective (Eq. (4)) is a non-convex function in general. But it is convex to $\mathbf{V}$, $\mathbf{U}^i$ or $\mathbf{W}^i$ when fixing the others. Given this property, Eq. (4) can be optimized in an alternating manner. We derive the iterative algorithm for minimizing Eq. (4) as follows.

4.1.1 Updating $\mathbf{V}$. Denoting η_{jk} and ϕ_j as the Lagrange multipliers for $\mathbf{V}_{jk}\geq 0$ and $\sum_k \mathbf{V}_{jk}=1$, the Lagrange function regarding $\mathbf{V}$ is defined as follows:

$$L(\mathbf{V},\eta)=O-tr(\eta^T\mathbf{V})+\sum_j\phi_j[\sum_k\mathbf{V}_{ik}-1]. \tag{5}$$

According to the KKT conditions, we may obtain the following equation system:

$$\begin{aligned}&\frac{\partial L}{\partial \mathbf{V}_{jk}}=4\alpha[\mathbf{V}\mathbf{V}^T\mathbf{V}]_{jk}-4\alpha[\mathbf{Y}\mathbf{V}]_{jk}\\ &+2\sum_i[\mathbf{V}\mathbf{W}^{iT}\mathbf{U}^{iT}\mathbf{U}^i\mathbf{W}^i-\mathbf{F}^{iT}\mathbf{U}^i\mathbf{W}^i]_{jk}-\eta_{jk}+\phi_j=0,\\ &\eta_{jk}\cdot\mathbf{V}_{jk}=0, \eta_{jk}\geq 0, \sum_k\mathbf{V}_{jk}-1=0.\end{aligned} \tag{6}$$

Algorithm 1: Model optimization-FMVFNC

```
Input  : Social network data: $\mathbf{Y}$, $\{\mathbf{F}^i\}_{i=1}^D$
Output : Fuzzy community membership: $\mathbf{V}$
Compute $\{\mathbf{X}^i\}_{i=1}^D$ by Eq. (2);
for $i \leftarrow 1:D$ do
    Initialize $\mathbf{U}^i$;
    Initialize $\mathbf{W}^i$;
end
Initialize $\mathbf{V}$;
$t \leftarrow 0$;
while $t < T_{max}$ do
    $t \leftarrow t+1$;
    for $i \leftarrow 1:D$ do
        Update $\mathbf{W}^i$ by Eq. (9);
        Update $\mathbf{U}^i$ by Eq. (8);
    end
    Update $\mathbf{V}$ by Eq. (7);
    Compute objective value $O^{(t)}$ by Eq. (4);
    if $O^{(t-1)} - O^{(t)} \leq \epsilon$ then
        break;
    end
end
```

The rule for updating $\mathbf{V}$ can be derived via solving the above equation system:

$$\begin{aligned}&\mathbf{V}_{jk}\leftarrow\mathbf{V}_{jk}\cdot\frac{4\alpha[\mathbf{Y}\mathbf{V}]_{jk}+2\sum_i[\mathbf{F}^{iT}\mathbf{U}^i\mathbf{W}^i]_{jk}+\phi_j}{4\alpha[\mathbf{V}\mathbf{V}^T\mathbf{V}]_{jk}+2\sum_i[\mathbf{V}\mathbf{W}^{iT}\mathbf{U}^{iT}\mathbf{U}^i\mathbf{W}^i]_{jk}}\\ &\phi_j=\frac{1-\sum_k[\mathbf{V}_{jk}\cdot\frac{4\alpha[\mathbf{Y}\mathbf{V}]_{jk}+2\sum_i[\mathbf{F}^{iT}\mathbf{U}^i\mathbf{W}^i]_{jk}}{4\alpha[\mathbf{V}\mathbf{V}^T\mathbf{V}]_{jk}+2\sum_i[\mathbf{V}\mathbf{W}^{iT}\mathbf{U}^{iT}\mathbf{U}^i\mathbf{W}^i]_{jk}}]}{\sum_k[\frac{\mathbf{V}_{jk}}{4\alpha[\mathbf{V}\mathbf{V}^T\mathbf{V}]_{jk}+2\sum_i[\mathbf{V}\mathbf{W}^{iT}\mathbf{U}^{iT}\mathbf{U}^i\mathbf{W}^i]_{jk}}]}\end{aligned} \tag{7}$$

4.1.2 Updating $\mathbf{U}^i$ *and* $\mathbf{W}^i$. Similarly, we can obtain the rules for updating $\mathbf{U}^i$ and $\mathbf{W}^i$. These iterative rules are directly listed here due to the space limitation. The updating rule for $\mathbf{U}^i_{jk}$ is:

$$\mathbf{U}^i_{jk}\leftarrow\mathbf{U}^i_{jk}\cdot\frac{2\mathbf{X}^i\mathbf{U}^i+\mathbf{F}^i\mathbf{V}\mathbf{W}^{iT}}{2[\mathbf{U}^i\mathbf{U}^{iT}\mathbf{U}^i]_{jk}+[\mathbf{U}^i\mathbf{W}^i\mathbf{V}^T\mathbf{V}\mathbf{W}^{iT}]_{jk}} \tag{8}$$

The rule for updating $\mathbf{W}^i_{tk}$ is:

$$\mathbf{W}^i_{tk}\leftarrow\mathbf{W}^i_{tk}\cdot\frac{[\mathbf{U}^{iT}\mathbf{F}^i\mathbf{V}]_{tk}}{[\mathbf{U}^{iT}\mathbf{U}^i\mathbf{W}^i\mathbf{V}^T\mathbf{V}]_{tk}} \tag{9}$$

The objective function in Eq. (4) can achieve convergence by iteratively updating $\mathbf{V}$, $\mathbf{U}^i$ and $\mathbf{W}^i$ according to Eqs. (7), (8), and (9) in a finite epochs. We summarize the illustrated optimization process in Algorithm 1.

4.2 Learning communities

Having obtained the optimal fuzzy community membership $\mathbf{V}$, FMVFNC is then able to determine the community affiliation for each vertex in the social network. As each vertex may belong to more than one community, we adopt the following strategy [9] to finally extract the community affiliation for all the vertices in the social network:

$$\mathbf{V}_{jk} \geq \frac{\beta\sqrt{K-1}}{K}, \mathbf{V}_{jk} < \max_{p} \mathbf{V}_{jp}, \tag{10}$$

where β is a positive number which determines the extent of community overlapping in the social network. It should be noted that, the above strategy is valid only for those $\mathbf{V}_{jk}$s smaller than the highest. This ensures that each vertex belongs to one community at least, and it can belong to more than one if its fuzzy community membership satisfies the above condition. A smaller value of β allows FMVFNC to assign each vertex in the network to more communities, and vice versa. In this paper, we typically set β to 1.

5 EXPERIMENTS AND ANALYSIS

In this section, we test the proposed framework with several well-established social networks and compare it with a number of popular approaches to social community detection.

5.1 Compared baselines

Eight prevalent approaches, including NCut [20], AP [4], CoDA [25], *k*-means [16], CESNA [24], CP-SI [15], CP-PI [15], and MISAGA [8] are selected as baselines in our experiment.

NCut, AP, and CoDA are three effective structure-based approaches to community detection. NCut can detect communities via learning spectral representations from the Laplacian graph regarding edge structure. AP can detect communities in social network by maximizing the intra-cluster structural propagation. CoDA is a state-of-the-art method for community detection. It can perform the task via probabilistically factorizing the matrix in terms of edge structure.

k-means is a classical method for unsupervised learning. It can unfold social communities utilizing the single view of vertex features.

As for feature-structure-based community detection, we select four state-of-the-art baselines, including CESNA, CP-SI, CP-PI, and MISAGA. CESNA is able to discover social communities by jointly factorizing matrices of edge structure and vertex features. CP-SI and CP-PI can perform the task of community detection via maximizing the intra-community feature diffusion. MISAGA is an effective matrix-factorization-based method for community detection. It can discover social communities by jointly factorizing similarity matrices regarding edge structure and vertex features.

Table 1: Characteristics of testing datasets.

Dataset	N	$\|E\|$	M	D	K
Cal	769	16656	53	1	10
Ego	4039	88234	1283	1	191
Twitter	3687	49881	20905	2	242
Gplus	8725	972899	5913	5	130

5.2 Datasets description

To test the effectiveness of different approaches, four well-established social networks, including $Caltech$ [21], $Ego{-}facebook$ [13], $Twitter$ [24], and *Googleplus* [13] are selected as testing datasets.

$Caltech$ (Cal) is collected from social network users who are students in California Institute of Technology. This dataset contains 769 vertices, and 16656 edges. Additionally, 53 vertex features are collected from user profiles to semantically describe the vertices.

$Ego - facebook$ (Ego) is constructed according to 4039 facebook users and 88234 online social ties among them. 1283 features representing the user profiles are also collected to describe the vertices.

$Twitter$ is constructed according to 3687 users from twitter.com. Besides 49881 edges, there are 20905 features which are collected from social tags and locations and can describe these 3687 users in two views.

Googleplus (Gplus) is the largest testing dataset used in our experiment. It is constructed according to 8725 googleplus users and their online social ties (972899 edges). Besides, 5913 features, which are collected from jobs, locations, institutions, universities, and identity information, are used to describe these googleplus users in five views. The statistics of these testing datasets are summarized in Table 1.

5.3 Experimental set-up

To test the performance of different methods, we let them discover social communities from all the testing datasets, using the settings recommended in the previous works. Specifically, the number of communities, i.e., K, which is required as an input by all the approaches, except of AP, and CESNA, is set to equal the number of ground-truth communities in the testing dataset. All approaches are run on a workstation with a 6-Core 3.4GHz CPU and 64GB RAM. And each approach is run 10 times in each testing dataset to obtain its steady performance. For performance evaluation, two widely used metrics, Normalized Mutual Information (*NMI*) [7] and Accuracy (*Acc*) [10] are used in our experiment. Their detailed definitions can be referred in previous works.

5.4 Clustering performance

Community detection in various social networks may bring better understandings to the online behaviors of human beings. In our experiment, we used the aforementioned four social networks to test the effectiveness of different approaches. As the ground-truth communities are available for all the testing datasets, the effectiveness of all methods can be easily evaluated against them. The experimental results regarding *NMI* and *Acc* are summarized in Table 2 and 3. When *NMI* is considered, the proposed fuzzy clustering method can perform better than any other baselines in all datasets. Specifically, in dataset Cal, FMVFNC is better than Ncut by % 61.73. FMVFNC outperforms CESNA in datasets Ego by % 21.76. In Twitter, FMVFNC is better than MISAGA by % 8.03. In Gplus, the performance improvement against the second best, i.e., AP, is % 4.39.

When the detected communities are evaluated by *Acc*, the proposed framework still performs robustly. Specifically, FMVFNC outperforms AP in datasets Cal and Gplus by % 30.96, and % 19.92, respectively. In dataset Ego, the proposed approach is better than

Table 2: Community detection performance evaluated by *NMI* (%). The best performance on each dataset is highlighted in bold.

Datasets Approaches	Cal	Ego	Twitter	Gplus
NCut	41.113	53.646	48.421	12.915
AP	38.133	57.093	54.264	40.769
CoDA	33.517	55.505	60.878	18.900
k-means	21.064	40.461	14.929	39.735
CESNA	39.259	57.513	46.588	21.817
CP-SI	21.505	48.510	49.752	21.929
CP-PI	21.616	48.981	48.538	24.663
MISAGA	29.774	56.452	65.329	21.553
Our Approach	**66.492**	**70.029**	**70.576**	**42.557**

Table 3: Community detection performance evaluated by *Acc* (%). The best performance on each dataset is highlighted in bold.

Datasets Approaches	Cal	Ego	Twitter	Gplus
NCut	37.451	44.689	42.121	26.705
AP	45.774	41.619	56.453	57.996
CoDA	37.824	52.091	66.537	34.735
k-means	14.954	29.116	28.343	16.252
CESNA	38.429	46.124	51.340	24.038
CP-SI	15.735	37.905	51.587	25.616
CP-PI	14.434	36.915	51.179	23.381
MISAGA	25.618	45.159	68.619	53.009
Our Approach	**59.948**	**59.619**	**70.762**	**69.547**

CoDA by % 14.45. In Twitter dataset, FMVFNC outperforms MISAGA by % 3.12. Given the experimental results w.r.t. *NMI* and *Acc*, it is observed that the proposed approach is very effective in detecting communities in social network data.

5.5 Sensitivity test

As Eq. (4) shows, α is used to control the relative significance between edge and multi-view correlated topic modeling. Different settings of α may influence the performance (*NMI* and *Acc*) of the proposed model. We set $\alpha = [0.1, 0.5, 1, 5, 10, 20, 50, 100]$ to test the model sensitivity and the corresponding results obtained from dataset Ego are plotted in Fig. 2. As depicted, both *NMI* and *Acc* become steady when α is set to equal or be larger than 1. Given the obtained results, we set $\alpha = 1$ when using FMVFNC to discover social communities in all the testing datasets.

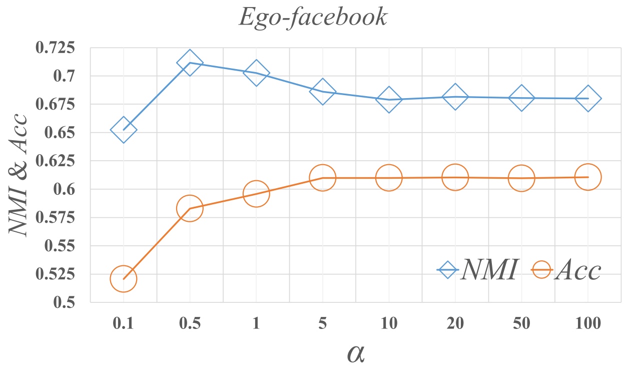

Figure 2: Sensitivity test in *Ego-facebook*.

5.6 Analysis on multi-view correlated topics

To further investigate whether the proposed framework can identify correlated topics from multi-view vertex features, we perform a detailed analysis on the discovered social communities and their topics. Here, we list the vertex features of correlated topics by which a discovered social community in Twitter can be semantically characterized (Table 4). As the table shows, in each view, there are two correlated topics inferred by the proposed framework. In both views, there are some feature IDs shared by both two identified social topics. Though many feature IDs are different, all the correlated topics are constituted by those features whose IDs are in the similar interval. This indicates the contents represented by these IDs can be found coinstantaneously, consequently meaning they are contextual correlated. Regularized by these relevant topics inferred from multi-view vertex features, the proposed FMVFNC is able to learn meaningful communities from social network data.

6 CONCLUSION

In this paper, a novel framework for discovering communities in social network data, Fuzzy Multi-View Featured Network Clustering (FMVFNC) is proposed. Different from existing approaches, FMVFNC is capable of inferring the community affiliation for each vertex in the network from edge structure and correlated topics learned from multi-view vertex features. As fuzzy membership is

Table 4: Multi-view correlated topics from Twitter dataset.

	Correlated Topics	
View 1	**Topic 284**	**Topic 423**
Feature ID	1939, 1942, 1945, 1965, 1980, 1981, 1985, 2049, 2107, 2111	1208, 1939, 1944, 1945, 1951, 1959, 1963,1989, 1995, 2034
View 2	**Topic 244**	**Topic 272**
Feature ID	9556, 9557, 9558, 9605, 9611, 9612, 9616, 9617, 9618, 9619	9547, 9566, 9615, 9619, 9642, 9643, 9644, 9646, 9648, 9650

considered when FMVFNC is learning community affiliation for each vertex, overlapping communities can be identified. The proposed framework has been tested with a number of popular social networks and compared with several prevalent approaches to community detection. The obtained satisfying results may validate the effectiveness of the proposed method. In future, we will further improve FMVFNC by considering cross-view and heterogeneously correlated topics.

ACKNOWLEDGMENTS

This paper is supported by the National Natural Science Foundation of China under Grant 61802317 and 51805059. The authors would like to thank the anonymous reviewers for their constructive suggestions and comments.

REFERENCES

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[2] Jonathan Chang and David Blei. 2009. Relational topic models for document networks. In *Artificial Intelligence and Statistics*. 81–88.

[3] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical review E* 70, 6 (2004), 066111.

[4] Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science* 315, 5814 (2007), 972–976.

[5] Kun He, Yingru Li, Sucheta Soundarajan, and John E Hopcroft. 2018. Hidden community detection in social networks. *Information Sciences* 425 (2018), 92–106.

[6] Tiantian He, Lu Bai, and Yew-Soon Ong. 2019. Manifold Regularized Stochastic Block Model. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 800–807.

[7] Tiantian He, Lu Bai, and Yew-Soon Ong. 2021. Vicinal Vertex Allocation for Matrix Factorization in Networks. *IEEE Transactions on Cybernetics* (2021).

[8] Tiantian He and Keith CC Chan. 2017. MISAGA: An algorithm for mining interesting subgraphs in attributed graphs. *IEEE transactions on cybernetics* 48, 5 (2017), 1369–1382.

[9] Tiantian He and Keith CC Chan. 2018. Discovering fuzzy structural patterns for graph analytics. *IEEE Transactions on Fuzzy Systems* 26, 5 (2018), 2785–2796.

[10] Tiantian He, Yang Liu, Tobey H Ko, Keith CC Chan, and Yew Soon Ong. 2019. Contextual Correlation Preserving Multiview Featured Graph Clustering. *IEEE transactions on cybernetics* (2019).

[11] Lun Hu and Keith CC Chan. 2015. Fuzzy clustering in a complex network based on content relevance and link structures. *IEEE Transactions on Fuzzy Systems* 24, 2 (2015), 456–470.

[12] Abhishek Kumar, Piyush Rai, and Hal Daume. 2011. Co-regularized multi-view spectral clustering. In *Advances in neural information processing systems*. 1413–1421.

[13] Jure Leskovec and Julian J Mcauley. 2012. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*. 539–547.

[14] Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems* 27 (2014), 2177–2185.

[15] Liyuan Liu, Linli Xu, Zhen Wangy, and Enhong Chen. 2015. Community detection based on structure and content: A content propagation perspective. In *2015 IEEE international conference on data mining*. IEEE, 271–280.

[16] David JC MacKay and David JC Mac Kay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

[17] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *nature* 435, 7043 (2005), 814–818.

[18] Chengbin Peng, Zhihua Zhang, Ka-Chun Wong, Xiangliang Zhang, and David Keyes. 2015. A scalable community detection algorithm for large graphs using stochastic block models. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[19] Maoying Qiao, Jun Yu, Wei Bian, Qiang Li, and Dacheng Tao. 2018. Adapting stochastic block models to power-law degree distributions. *IEEE transactions on cybernetics* 49, 2 (2018), 626–637.

[20] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22, 8 (2000), 888–905.

[21] Amanda L Traud, Peter J Mucha, and Mason A Porter. 2012. Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications* 391, 16 (2012), 4165–4180.

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

[23] Xiao Wang, Di Jin, Xiaochun Cao, Liang Yang, and Weixiong Zhang. 2016. Semantic community identification in large attribute networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[24] Jaewon Yang, Julian McAuley, and Jure Leskovec. 2013. Community detection in networks with node attributes. In *Data Mining (ICDM), 2013 IEEE 13th international conference on*. IEEE, 1151–1156.

[25] Jaewon Yang, Julian McAuley, and Jure Leskovec. 2014. Detecting cohesive and 2-mode communities indirected and undirected networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*. 323–332.

[26] Hadi Zare, Mahdi Hajiabadi, and Mahdi Jalili. 2019. Detection of Community Structures in Networks with Nodal Features based on Generative Probabilistic Approach. *IEEE Transactions on Knowledge and Data Engineering* (2019).