Clustering in Networks with Multi-Modality Attributes

Tiantian He
Department of Computing
The Hong Kong Polytechnic University
Hong Kong SAR
tiantian.he@outlook.com

Keith C.C. Chan
Department of Computing
The Hong Kong Polytechnic University
Hong Kong SAR
cskcchan@comp.polyu.edu.hk

Libin Yang
School of Automation
Northwestern Polytechnical University
Xi'an, China
libiny@nwpu.edu.cn

***Abstract*—Network clustering is one of the most significant tasks of network analytics. To discover network clusters, there have been many approaches proposed, utilizing network topology, or node attributes. However, there are no effective approaches that are able to discover clusters in the network with multiple modalities of attributes. In this paper, we propose a novel clustering model, called CNMMA, to discover network clusters using edge structure, and multi-modality attributes associated with vertices. Assuming edge structure, and node attributes are generated by corresponding low dimensional latent spaces (matrices), CNMMA can learn an optimal latent matrix representing the cluster membership for each vertex in the network. Besides, CNMMA makes use of an effective method to regulate the latent spaces w.r.t. edge structure and node attributes so that those vertices sharing similar edges and modality-wise attributes are more possible to be assigned with the same cluster labels. CNMMA has been tested with several real-world networks, which contain multiple modalities of node attributes, and has been compared with state-of-the-art approaches to network clustering. The experimental results show that CNMMA outperforms most approaches in most datasets. The clusters discovered by CNMMA are better matched with the ground truth.**

Keywords—network clustering, multi-modality attributed network, multi-modality attributed graph, community detection, graph clustering, complex network, social network

I. Introduction

Many real-world data can be represented as networks (graphs) which contain vertices (nodes) and edges (links) representing the data entities, and the interrelationship between them, respectively. Different from network data that are randomly generated, real-world networks always possess some latent features beneath the chaotic structure. Amongst them, network clusters, or network communities are of significance. How to discover them in the network has drawn much attention in the recent [1] as network cluster discovery is related to many real-world applications, e.g., social community detection [2], document classification [3], and biological graph analytics [4].

To effectively discover clusters in the network, several approaches have been proposed, utilizing different information carried by the network data. Unsurprisingly, many of these algorithms may perform the task by taking into the consideration the properties of network topology, which include edges or edge weights between pairwise vertices. For examples, CNM [5] and Lovain [6] are two representative algorithms for community detection in networks, which are able to perform the task via modularity optimization. In [7], the authors propose an approach to discovering clusters in social and biological networks through a clique percolation method (CPM). In [8], a clustering algorithm called affinity propagation (AP) is proposed to detect clusters based on the similarities between candidate cluster centers and other vertices. In [9], a framework for community detection is proposed. The proposed approach can perform the task of network clustering taking into the consideration edge weights representing the topological correlation between vertices.

Besides, spectral clustering [10] is also proved to be effective in analyzing the network data. By analyzing the spectral properties of the similarity matrix constructed based on different measures, e.g., network transitivity [11] and normalized cut [12], the cluster structure can be revealed. In [13], a model-based algorithm called CoDa is proposed to detect clusters in the network. Modeling the discovering of clusters as identifying the cluster affiliations of each vertex, the best affiliation can be identified by optimizing the posterior probabilities representing the possibility that vertices belong to a cluster in a generative model.

Besides edge structure, other information associated with vertices can also be collected, and they can be seen as attributes characterizing the vertices. Therefore, there are some attempts to discover network clusters considering using node attributes. For example, *k*-means clustering [14] is an effective method which discover clusters in the network via grouping vertices sharing higher similarities of attributes. In [15], an algorithm (MAC) is proposed to discover clusters in network data in which vertices are with Boolean attribute values.

Though effective to some extent, the aforementioned approaches may not discover more meaningful clusters in the network as they either consider network topology, or node attributes only.

To discover clusters in network data taking into the consideration both topological and attribute properties, there have been several approaches proposed. In [16], EDCAR is proposed to mine clusters by grouping vertices with higher local density of links and similarity of vectorized attributes. In [17], a new criterion for discovering communities in networks with node attributes is proposed. The proposed metric can be used to generate edge weights taking into the consideration attributes associated with vertices. In [18] and [29], two algorithms (CESNA and aMMSB) are proposed to use statistical models to learn the posterior probability that pairwise vertices are connected given edge structure and attributes in a cluster. In [19], an attribute random-walk kernel is proposed to combine the link structure and content similarity in network clustering. The cluster membership can be learned via spectral clustering. In [20], a diffusion model for network analysis is proposed. The cluster membership learned by the model may consider both structural diversity and the node contents between vertices and each of the clusters. In [21] and [22], two effective approaches (MISAGA and FSPGA) to analyzing attributed graphs are proposed. These two methods are able to discover network clusters via optimizing objective functions which evaluate both local network density and attribute similarity between pairwise vertices.

Given the prevalent works for cluster discovery in the network, we have the following findings that motivate us to redefine the problem and propose an effective approach to it. First, many prevalent approaches to clustering in network data do not fully utilize the information carried by the network, just like the ones that are either topology or attribute based. Second, almost all the previous approaches to clustering in network data utilizing attribute information, do not consider the attributes may come from multiple modalities. Examples of such attributes from multiple modalities are easily found in real-world networks: an online social network user can be characterized by his profiles, hobbies, topics concerned, etc. Such attributes from different sources may take effect on the relationship between vertices, and cluster membership of vertices. While, on the other hand, multi-modality learning [27], has shown to be very effective when compared with traditional single-view learning approaches. Motivated by the multi-modality learning framework, we propose a novel model for clustering the network with multi-modality attributes (CNMMA) to address the mentioned challenges. Different from previous works, CNMMA is able to learn an optimal cluster assignment for each vertex associated with multi-modality attributes in the network. Moreover, CNMMA also utilizes an effective method to regularize the similarity of latent spaces learned from network topology and different attribute modalities, so that those vertices sharing similar link structure and attributes are more possible to be assigned with the same cluster membership. CNMMA has been tested with several sets of real-world network data and compared with both classical and state-of-the-art approaches. The experimental results show that CNMMA outperforms those compared baselines in most testing datasets, which means that clustering in the network with multiple modalities of attributes may lead to a better latent structure discovery.

II. Mathematical preliminaries and notations

In this section, the mathematical preliminaries and notations used in this paper are introduced. As attributes characterizing the vertices may come from more than one view, the definition of the network in this manuscript is different from the previous. Given a network containing n vertices and |E| edges, it can be represented as N = {V, E, Λ}, where V, E, and Λ represent the vertex, edge, and attribute set, respectively. For the vertex set, it is defined as V = $\{v_i \mid 1 \leq i \leq n\}$. The edge set is defined as E = $\{e_{ij}=1 \mid v_i$ and v_j are connected$\}$. As for the attribute set, it is defined as $\Lambda = \{\Lambda^i_j \mid 1 \leq i \leq d,\ 1 \leq j \leq m^i\}$, where d and m^i represent the total number of attribute modalities, and the total number of attributes in modality i, respectively. It should be noted that d is always larger than 1 and $m^1+\ldots+m^d=m$, where m is the total number of attributes. CNMMA uses **G** to represent the link structure (node adjacency) between pairwise vertices in N. Apparently, **G** is an n-by-n symmetric matrix each element of which, say $\mathbf{G}_{ij}$, is equal to 1 if vertex i and vertex j are connected (e_{ij}=1), and vice versa. CNMMA uses an m^i-by-n binary matrix, $\mathbf{A}^i$ to represent whether a vertex is labeled with an attribute from Λ^i, i.e., $\mathbf{A}^i_{jk}$, is equal to 1 if vertex k is labeled with attribute j from Λ^i, and vice versa. For notations, we use a subscript, e.g., $\mathbf{G}_i$, to represent the ith column of a given matrix, say **G**. We use $\mathbf{G}_{ij}$, to represent the entry of **G**, in ith row, jth column. We use a superscript, e.g., $\mathbf{U}^i$, to represent a matrix related to ith view. $tr(\cdot)$ represents the matrix trace. $\|\cdot\|_F$ represents the matrix Frobenius norm. All these mentioned preliminaries and notations are used by CNMMA to model the problem of discovering clusters in the network with multiple modalities of attributes.

III. The clustering algorithm

In this section, how CNMMA defines the problem of clustering in the network with multi-modality attributes and how it solves the problem are introduced in details.

A. The objective function

The main task performed by CNMMA is to discover a set of clusters hidden in the network. These clusters are subnetworks possessing the feature that all the vertices within the same subnetwork share similar topological structure and node attributes from multiple modalities. To fulfill the task, CNMMA assumes that the link structure **G**, the multi-modality attributes associated with each vertex, are generated by corresponding low-dimensional latent spaces, represented as corresponding low-dimensional matrices. The optimal cluster assignment for each vertex can be obtained when the difference between the original data and the generated is minimized.

To model the link structure of N, CNMMA utilizes an n-by-k matrix, **C** to represent weight of affiliation that each vertex belongs to each of the k latent components. Based on the assumption mentioned above, the difference between **G** and the link structure generated by the latent space, can be represent as

$$\left\|\mathbf{G}-\mathbf{C}\mathbf{C}^T\right\|_F^2 \tag{1}$$

It can be seen from (1), CNMMA assumes that $\mathbf{G}_{ij}$ is generated by $[\mathbf{C}\mathbf{C}^T]_{ij}$, which means the product of ith and jth row in **C**. When (1) is optimized, the link density within each cluster can be maximized.

To model the attributes that are associated with each vertex, CNMMA constructs a m^i-by-k latent space, $\mathbf{S}^i$, to represent as the weight factors that each attribute from Λ^i contributes to each of k latent components. It is easy to verify that, a higher value of $\mathbf{S}^i_{jk}$, means attribute j in modality i contributes more to component k. In addition, CNMMA constructs an n-by-k matrix, $\mathbf{D}^i$ to represent the strength that each vertex belongs to each of the k latent components, when Λ^i is considered. CNMMA assumes that whether or not each attribute in Λ^i, is associated with vertex, is generated by $\mathbf{S}^i$ and $\mathbf{D}^i$, so that the following function can be used to evaluate the difference between the original attributes associated and the generated in all d attribute modalities

$$\sum_{i=1}^{d}\left\|\mathbf{A}^i-\mathbf{S}^i\mathbf{D}^{iT}\right\|_F^2 \tag{2}$$

When (2) is optimized, $\mathbf{A}^i$ is best approximated by $\mathbf{S}^i$ and $\mathbf{D}^i$, and $\mathbf{D}^i$ may reveal the strength that each vertex belongs to each of k latent components, in terms of attributes from Λ^i.

As CNMMA attempts to cluster network data with multi-modality attributes, it needs to identify a common latent space representing the cluster assignment for each vertex. CNMMA is able to learn such latent space using the following objective function

$$\|\mathbf{C}-\mathbf{B}\|_F^2+\sum_{i=1}^{d}\left\|\mathbf{B}-\mathbf{D}^i\right\|_F^2 \tag{3}$$

It is seen that CNMMA follows a hierarchical consensus rule when learning the cluster membership. It firstly learns a latent space **B** as the common cluster membership taking into the consideration the attributes from all modalities, and then regulates the structure of **B** and **C** to be similar. Thus, the matrix **C** can be seen as the common latent space representing the cluster affiliation between each vertex and each of the *k* clusters.

As mentioned above, both link structure and vertex attributes may affect the interrelationship between pairwise vertices and the cluster membership of each vertex in the network. Such interrelationship can be seen as affinity between pairwise vertices, which stands for how similar/related two vertices are. For CNMMA, we propose to use the following method to better capture the pairwise affinity in terms of both topology and attribute from multiple modalities. Let **X** be a matrix representing the degree of affinity between pairwise vertices in *N*. For each entry in **X**, CNMMA assumes that it can be generated by **C** and **B**. Hence, we may obtain the following function evaluating the difference between **X** and the one generated by **C** and **B**

$$\left\|\mathbf{X}-\mathbf{B}\mathbf{C}^T\right\|_F^2 \tag{4}$$

Given (4), we find that how affinitive that two vertices are is determined by both **C** and **B**, which are the latent spaces learned from (1) and (3). This means both structural and attributes of pairwise vertices take effect on their interrelationship. If (4) can be optimized, latent spaces **C** and **B** is regularized to assign those closely affinitive vertices into the same latent components, which means those vertices with similar latent structural and attribute attributes are more possible to be assigned into the same clusters. As for **X**, it can be obtained by computing pairwise similarity between vertices, using either link or attribute information in *N*. In this paper, **X** is obtained by computing the cosine similarity of local neighbors between pairwise vertices.

Given (1)-(4), we proposed CNMMA to make use of the following objective function to evaluate the overall clustering quality when it is discovering the clusters in the network with multi-modality attributes

$$\begin{aligned} O=&\left\|\mathbf{G}-\mathbf{C}\mathbf{C}^T\right\|_F^2+\alpha\sum_{i=1}^{d}\left[\left\|\mathbf{A}^i-\mathbf{S}^i\mathbf{D}^{iT}\right\|_F^2+\left\|\mathbf{B}-\mathbf{D}^i\right\|_F^2\right]\\ &+\|\mathbf{C}-\mathbf{B}\|_F^2+\left\|\mathbf{X}-\mathbf{B}\mathbf{C}^T\right\|_F^2 \\ &\mathbf{C}\geq 0, \mathbf{S}^i\geq 0, \mathbf{D}^i\geq 0, \mathbf{B}\geq 0\end{aligned} \tag{5}$$

where α is a positive real which is used to control the effect which the node attributes take on the formulation of cluster membership. Given (5), we may find out CNMMA has the following peculiarities when clustering in the network with multi-modality node attributes. First, it considers the effect from multiple modalities of attributes when performing the task, which is not considered by almost all the other related works. Second, CNMMA considers modeling the affinity between pairwise vertices using both topological and attribute properties of the network, so that vertices sharing similar latent structural and attribute attributes are more possible to be assigned into the same clusters. For such a modelling method, this is also the first attempt, compared with previous works. When (5) is optimized, CNMMA is able to find the optimal cluster membership for each vertex in the network. Such membership is obtained by considering edge structure, modality-wise node attributes and pairwise affinity. According to (5), we may also see that the problem tackled by CNMMA is to discover clusters in the network with multi-modality node attributes, not to discover clusters in multiple networks, e.g., clustering in heterogeneous networks [28].

B. Model optimization

How to search for an optimal cluster assignment for each vertex in *N* is essential to CNMMA. As (5) can be seen as a constrained optimization problem, we may derive a series updating rules to infer the optimal variables of CNMMA, based on KKT conditions.

Let β_{jk} be the Lagrange multiplier for the constraint $\mathbf{C}_{jk}\geq 0$, and the Lagrange function *L* for **C** is

$$L(\mathbf{C},\boldsymbol{\beta})=O-tr(\boldsymbol{\beta}^{\mathbf{T}}\mathbf{C}) \tag{6}$$

where $\boldsymbol{\beta}=[\beta_{jk}]$ is the matrix of Lagrange multipliers for the non-negativity of **C**. Based on the KKT condition for constrained optimization, we have the following element-wise equation system

$$\begin{aligned}&\frac{\partial L}{\partial \mathbf{C}_{jk}}=\left[-4\mathbf{G}\mathbf{C}+4\mathbf{C}\mathbf{C}^T\mathbf{C}\right]_{jk}\\ &+\left[-2\mathbf{X}\mathbf{B}+2\mathbf{C}\mathbf{B}^T\mathbf{B}+2\mathbf{C}-2\mathbf{B}-\boldsymbol{\beta}\right]_{jk}\\ &\boldsymbol{\beta}_{jk}\mathbf{C}_{jk}=0\\ &\boldsymbol{\beta}\geq 0\end{aligned} \tag{7}$$

Based on (7), we may obtain the following iterative rule for updating the variables in **C**, by making some mathematical transformations

$$\begin{aligned}&\mathbf{C}_{jk}\leftarrow\mathbf{C}_{jk}\frac{\sqrt{\sqrt{\Phi_{jk}}-\left[\mathbf{C}\mathbf{B}^T\mathbf{B}+\mathbf{C}\right]_{jk}}}{\sqrt{4\left[\mathbf{C}\mathbf{C}^T\mathbf{C}\right]_{ij}}}\\ &\Phi_{jk}=\left[\mathbf{C}\mathbf{B}^T\mathbf{B}+\mathbf{C}\right]_{jk}^2+8\left[\mathbf{C}\mathbf{C}^T\mathbf{C}\right]_{jk}\left[2\mathbf{G}\mathbf{C}+\mathbf{X}\mathbf{B}+\mathbf{B}\right]_{jk}\end{aligned} \tag{8}$$

Similarly, we are able to derive the iterative rules for updating the variables in **B**, $\mathbf{S}^i$ and $\mathbf{D}^i$. Here, we directly present the derived updating rules due to the space limitation. For updating **B**, we have

$$\mathbf{B}_{jk}\leftarrow\mathbf{B}_{jk}\frac{\left[\mathbf{X}\mathbf{C}+\alpha\sum_i\mathbf{D}^i+\mathbf{C}\right]_{jk}}{\left[\mathbf{B}\mathbf{C}^T\mathbf{C}+(\alpha+1)\mathbf{B}\right]_{jk}} \tag{9}$$

For updating $\mathbf{D}^i$, we have

$$\mathbf{D}_{jk}^i\leftarrow\mathbf{D}_{jk}^i\frac{\left[\mathbf{A}^{iT}\mathbf{S}^i+\mathbf{B}\right]_{jk}}{\left[\mathbf{D}^i\mathbf{S}^{iT}\mathbf{S}^i+\mathbf{D}^i\right]_{jk}} \tag{10}$$

For updating $\mathbf{S}^i$, we have

$$\mathbf{S}_{jk}^i\leftarrow\mathbf{S}_{jk}^i\frac{\left[\mathbf{A}^i\mathbf{D}^i\right]_{jk}}{\left[\mathbf{S}^i\mathbf{D}^{iT}\mathbf{D}^i\right]_{jk}} \tag{11}$$

In each iteration of optimization, CNMMA iteratively updates the latent variables in **C**, **B**, $\mathbf{S}^i$ and $\mathbf{D}^i$, while keeping the others fixed. Thus, (5) will converge to local optima and CNMMA is able to identify an optimal cluster assignment for each vertex in *N*. One may theoretically verify the convergence of the updating algorithms following the analysis shown in [23].

C. Model complexity

The model complexity determines the efficiency of the clustering process. Based on the derived updating rules, we may obtain the complexity of CNMMA as the following. Based on (8), updating the variables in **C** in each iteration follows the order of $O(2(n^2+n)(k^2+k))$. Based on (9), updating the variables in **B** in each iteration follows the order of $O(n^2(k^2+k)+nk^2+(d+2)nk)$. Based on (10), updating variables in each $\mathbf{D}^i$ in each iteration follows the order of $O(nm(k^2+k)+n(k^2+2k))$. Based on (11), updating variables in each $\mathbf{S}^i$ follows the order of $O(nm(k^2+k)+mk^2)$. It is seen that CNMMA is a model having the computational complexity that is approximate to $O(n^2)$, as *k* is much smaller than *n*, and *m* is near to *n*. Such a computational complexity ensures CNMMA costs an acceptable amount of time for model optimization.

IV. Experimental analysis

In this section, we present the details of how we test the effectiveness of CNMMA using the real-world datasets.

A. Experimental set-up and evaluation metrics

For performance comparison, we have selected 6 approaches to network clustering as baselines, including CNM [5], Spectral clustering (SC) [10], *k*-means clustering [14], CESNA [18], MISAGA [21], and FSPGA[22]. These selected baselines are either classical approaches, or state-of-the-art ones to network clusterin. Specifically, these 6 approaches can be categorized into 3 classes.

CNM and SC are able to discover network clusters utilizing topological information. CNM is a classical and effective method based on modularity optimization. SC performs the task by grouping vertices with similar local structures into the same cluster.

k-means is able to discover clusters in the network data taking into the consideration node attributes.

CESNA, MISAGA, and FSPGA are three methods utilizing both network topology and node attributes. CESNA is a very effective probabilistic model for network clustering. Those vertices sharing similar latent structures of both local connectivity and node attributes are more possible to be assigned with the same cluster label. MISAGA, which is based on matrix factorization, can perform the task of network clustering using link structure and node similarity. FSPGA is a very effective fuzzy clustering method, which is able to discover overlapping clusters in attributed networks.

We either used the source code, or executables of the compared baselines released by the authors. Together with CNMMA, all the approaches were executed for discovering network clusters under the same environment which included a workstation with 4-core 3.4GHz CPU and 16GB RAM. As for the parameter settings of different approaches, we used default settings of CNM and CESNA as these two approaches are able to automatically perform the clustering task. For other ones that require the number of clusters, *k* to be determined, we set *k* to be the number of ground truth classes in each testing dataset. For CNMMA, we set *α*, and *max_iteration* to be 1 and 300. Under these settings, we compared the clustering performance of different approaches, using real-world network data.

TABLE I. Characteristics of Testing Datasets

	PUK	Olym	Twitter	Gplus
n	419	464	2511	8725
|E|	23003	9749	37154	972899
m	22747	21552	9067	5913
d	2	2	2	5
Ground truth *k*	5	28	132	130

For performance testing, we used 4 real-world networks whose ground truth classes have been verified in the previous works. These networks are with different sizes and their modalities of attributes have been verified in the previous works. It should be noted that, to allow those baselines, including *k*-means, CESNA, MISAGA, and FSPGA, to utilize the same attributes as CNMMA uses, we have concatenated the multi-modality attributes in each dataset as a single attribute matrix. The details of these 4 testing datasets are listed below.

PoliticsUK (PUK) [24]-It is a politician network containing 419 vertices and 23003 edges, which represent the English politicians, and the social interactions between them. The key points of their comments, and social tags have been verified as the two modalities of attributes. There are 5 ground truth clusters verified in this dataset, representing the political parties to which the politicians belong.

Olympics (Olym) [24]-It is an athletic social network which contains athletes joining the Olympic Games. There are 464 vertices and 9749 edges representing the athletes, and the social relationship between them. In addition, two modalities of attributes, which are the key points of the comments and social tags, are used to characterize these 464 athletes. In Olym dataset, there are 28 social communities that are verified as ground truth clusters.

Twitter [25]-The Twitter dataset is constructed based on a number of social circles extracted from twitter.com. For this dataset, there are 2511 vertices, and 37154 edges, representing the twitter users and the friendship between them, respectively. Two modalities of attributes are collected, which represent the social tags of the twitter users, and common topics concerned by these users. There are 132 social circles verified as ground truth clusters in this dataset.

Googleplus (Gplus) [18]-Gplus is a set of online social network data constructed based on the sub-networks from plus.google.com. There are 8725 vertices, and 972899 links representing the social network users and their social relationship, respectively. In addition, there are 5 modalities of attributes that are collected and used to describe the job titles the users used to have, universities the users graduated from, the organizations the users are working at, the places the users used to visit, and the last names of the users, respectively. The number of ground truth classes in this dataset is 130. The characteristics of these 4 datasets have been summarized in Table I.

TABLE II. CLUSTERING PERFORMANCE EVALUATED BY *NMI* (%)

	CNMMA	CNM	SC	*k*-means	CESNA	MISAGA	FSPGA	Improvement (%)#
PUK	64.468	**80.809**	47.380	27.527	60.850	54.643	51.314	--
Olym	**83.709**	40.343	46.905	48.257	41.941	82.437	80.227	1.543
Twitter	**69.526**	37.791	47.940	28.591	57.673	65.854	63.584	5.576
Gplus	**44.142**	11.847	12.915	9.735	22.425	21.553	27.557	60.184

TABLE III. CLUSTERING PERFORMANCE EVALUATED BY *ACC* (%)

	CNMMA	CNM	SC	*k*-means	CESNA	MISAGA	FSPGA	Improvement (%)
PUK	**95.465**	**95.465**	80.907	75.895	83.204	90.215	90.931	--
Olym	**85.560**	34.989	42.672	46.983	43.644	82.112	77.371	4.199
Twitter	50.777	28.297	30.506	23.935	47.254	50.258	**50.855**	--
Gplus	**68.356**	21.410	26.705	16.252	7.893	53.009	60.224	13.503

Improvement (%): The percentage of improvement when CNMMA is compared with the second-best approach in each dataset.

For performance evaluation, we used two evaluation measures, including the Normalized Mutual Information (*NMI*), and Accuracy (*Acc*) [26]. Both of them are widely used for evaluating the validity of clusters in network data.

The *NMI* measures the overall accuracy of the matches between clusters that are detected and those that are considered "ground truth". It is defined as

$$NMI = \frac{\sum_{C,C^*} \Pr(C_i, C_j^*) \log \frac{\Pr(C_i, C_j^*)}{\Pr(C_i)\Pr(C_j^*)}}{\max(H(C_i), H(C_j^*))} \qquad (12)$$

where $\Pr(C_i, C_j^*)$ denotes the probability that vertices are in both the detected cluster *i* and the ground truth cluster *j*, and $\Pr(C_i)$ denotes the probability that a vertex is found to exist in detected cluster *i*. The *NMI* considers both the size of each discovered cluster and the ground-truth clusters by computing a fraction ratio between the clusters that are identified and those that are available in the ground-truth database. If the *NMI* measure is high, it means that the clusters detected match well with the ground-truth ones.

Contrary to the *NMI*, the *Acc* [26] measure evaluates individually detected sub-graphs. It is defined as

$$Acc = \sum_i \frac{|C_i|}{n_V} f\left(C_i, C^*\right) \qquad (13)$$

where $|C_i|$ means the size of a detected cluster, and *f*(.) stands for a particular mapping function between a detected cluster *i* and the ground truth. For our purpose, we define *f*(.) to be the maximum overlap between detected cluster *i* and a ground-truth cluster. As a result, *Acc* evaluates the weighted average of maximum overlapping between each discovered cluster and a ground-truth cluster. A higher value of *Acc* therefore means that each detected cluster has a better match with the ground truth. The higher the *Acc* is, the more effective the algorithm can be considered to be. Given its definition, *Acc* emphasizes more on the discovered cluster when evaluating.

B. *Clustering performance in real-world data*

For performance evaluation of the proposed model, we used the 4 testing datasets, including PUK, Olym, Twitter, and Gplus. All these datasets have known ground-truth communities that have been verified in the previous works. The experimental results of *NMI* and *Acc* obtained in the testing datasets are summarized in Table II and III.

As Table II shows, CNMMA obtains a robust performance in all the 4 datasets. CNMMA outperforms any other baseline in all the datasets except of PUK, when the discovered clusters are evaluated by *NMI*. CNMMA outperforms MISAGA in Olym, and Twitter by 1.543%, 5.576%, respectively. CNMMA is better than FSPGA in Gplus by 60.184%.

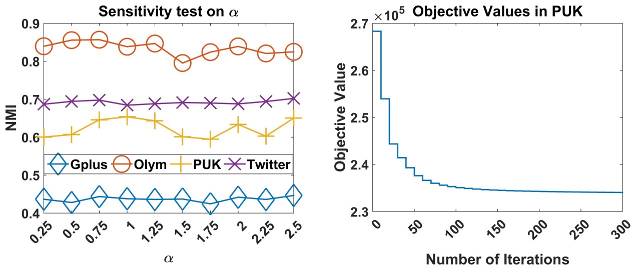
When *Acc* measure is considered, CNMMA is able to outperform all the other compared baselines in all the testing datasets, except of Twitter. As Table III shows, CNMMA outperforms MISAGA in Olym by 4.199%. In Gplus dataset, CNMMA is better than FSPGA by 13.503%.

Given the experimental results shown in Table II and III, it is said that CNMMA is a very effective model for clustering in the network with multi-modality attributes. The clusters discovered by CNMMA are better matched with the ground truth. By taking into the consideration edge structure and multi-modality attributes and modeling the pairwise affinity between vertices, CNMMA is able to learn the optimal cluster membership for each vertex in the network.

C. *Parameter sensitivity of CNMMA*

As mentioned in Section III, α is used to control the effect of node attributes on the clustering process. The clustering performance of CNMMA might be different when one varies the setting of α. To investigate how the settings of α may affect the clustering performance of CNMMA, we used CNMMA to perform the network clustering tasks in all the testing datasets, varying α from 0.25 to 2.5, with a 0.25 increment step. The clusters discovered by CNMMA using these settings are evaluated by *NMI*. The corresponding results are shown in Fig. 1 (a).

When *NMI* is considered, CNMMA performs relatively steadily in most datasets when α changes. The clustering performance evaluated by *NMI* has relatively evident variations in datasets PUK and Olym. In other words, some attributes in some modalities can be seen as noise which may degrade the clustering performance. By varying the value of α, one may use CNMMA to obtain a better cluster assignment for each vertex in the network, while minimize the side-effect brought by the noisy attributes. Based on the



(a) (b)

Fig. 1. Sensitivity test on α and model convergence in PUK dataset

results shown in Fig. 1 (a), we recommend α to be set ranging from 0.5 to 1.5 to ensure CNMMA to perform robustly.

D. Convergence of the model

Whether the proposed model is able to convergent in a finite number of iterations of optimization may determine the quality of the discovered clusters. To verify the convergence of CNMMA, we used it to perform the task of cluster discovery in all the real-world datasets by setting the model parameters as mentioned in Section IV.A, and we recorded the objective values obtained by CNMMA after each iteration of optimization. Here we take the variations of objective values obtained from PUK dataset as an example (See Fig. 1 (b)). As the figure shows, CNMMA may converge to the local optima around 200 iterations in PUK. Noted that CNMMA may achieve convergence after about 200 iterations of optimization in all the 4 testing datasets, the fast convergence of CNMMA ensures it can find the optimal cluster assignment for each vertex in the network in a shorter time. And this is also one of the reasons that CNMMA is very effective in clustering in the network with multiple modalities of node attributes.

V. Conclusion

In this paper, a novel model for clustering the network data, CNMMA is proposed. Different from the previous approaches, CNMMA is able to discover clusters in the network with multiple modalities of node attributes. By modeling the affinity between pairwise vertices within the clustering process, CNMMA is more possible to assign those vertices sharing higher topological and attribute similarities with the same cluster labels. CNMMA has been tested with 4 sets of real-world data and compared with both classical and state-of-the-art approaches to network clustering. The experimental results show that the clusters discovered by CNMMA are better matched with the ground truth. This means using a multi-modality setting is a promising way for clustering the network data. In future, we will further improve the efficiency of CNMMA and develop a novel version of the model that is able to effectively discover overlapping clusters in the network with multi-modality attributes.

References

[1] S. Fortunato, “Community detection in graphs,” Phys. Rep., vol. 486, no.3-5, pp. 75-174, 2010.

[2] T. He, and K.C.C. Chan, “Evolutionary community detection in social networks,” CEC, pp. 1496-1503, 2014.

[3] J. Chang, and D.M. Blei, “Relational topic models for document networks,” AISTATS, pp. 81-88, 2009.

[4] T. He, and K.C.C. Chan, “Evolutionary graph clustering for protein complex identification,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 15, no. 3, pp. 892-904, 2018.

[5] A. Clauset, M.E.J. Newman, and C. Moore, “Finding community structure in very large networks,” Phys. Rev. E, vol. 70, no. 6, pp. 066111, 2004.

[6] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” J. Stat. Mech., vol. 2008, no. 10, pp. P10008, 2008.

[7] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, “Uncovering overlapping community structure of complex networks in nature and society,” Nature, vol. 435, pp. 814-818, 2005.

[8] B.J. Frey, and D. Dueck, “Clustering by passing messages between data points,” Science, vol. 16, pp. 972-976, 2007.

[9] N. Veldt, D.F. Gleich, and A. Wirth, “A Correlation Clustering Framework for Community Detection,” WWW, pp. 439-448, 2018.

[10] U. Luxburg, “A tutorial on spectral clustering,” Stat. Comput., vol. 17, no. 4, pp. 395-426, 2007.

[11] B. Yang, J. Liu, and J. Feng, “On the spectral characterization and scalable mining of network communities,” IEEE Trans. Knowl. Data Eng., vol. 24, no. 2, pp. 326-337, 2012.

[12] J. Shi, and J. Malik, “Normalized cuts and image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 888-905, 2000.

[13] J. Yang, J. McAuley, and J. Leskovec, “Detecting Cohesive and 2-mode Communities in Directed and Undirected Networks,” WSDM, 2014, pp. 323-332.

[14] D.J.C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge: Cambridge Univ. Press, 2003.

[15] M. Frank, A.P. Streich, D. Basin, and J.M. Buhmann, “Multi-assignment clustering for boolean data,” J. Mach. Learn. Res., vol. 13, pp. 459-489, 2012.

[16] S. Gunnermann, B. Boden, I. Farber, and T. Seidl, “Efficient mining of combined subspace and subgraph clusters in graphs with attribute vectors,” PAKDD, pp.261-275, 2013.

[17] Y. Zhang, E. Levina, and J. Zhu, “Community detection with node attributes,” Electron. J. Stat., vol. 10, pp. 3153-3178, 2016.

[18] J. Yang, J. McAuley, and J. Leskovec, “Community detection in networks with node attributes,” ICDM, pp. 1151-1156, 2013.

[19] T. Guo, J. Wu, X. Zhu, and C. Zhang, “Combining structured node content and topology information for networked graph clustering,” ACM Trans. Knowl. Discov. Data, vol. 11, no. 3, article 29, 2017.

[20] Q. Bao, W.K. Cheung, Y. Zhang, and J. Liu, “A component-based diffusion model with structural diversity for social networks,” IEEE Trans. Cybern., vol. 47, no. 4, pp. 1078-1089, 2017.

[21] T. He, and K.C.C. Chan, “MISAGA: an algorithm for mining interesting sub-graphs in attributed graphs,” IEEE Trans. Cybern., vol. 48, no. 5, pp. 1369-1382, 2018.

[22] T. He, and K.C.C. Chan, “Discovering fuzzy structural patterns for graph analytics,” IEEE Trans. Fuzzy Syst., in press.

[23] F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding, “Community discovery using nonnegative matrix factorization,” Data Min. Knowl. Disc., vol. 22, pp. 493-521, 2011.

[24] D. Greene, and P. Cunningham, “Producing a unified graph representation from multiple social network views,” WebSci, pp. 1-4, 2013.

[25] J. McAuley, and J. Leskovec, “Discovering social circles in ego networks,” ACM Trans. Knowl. Discov. Data, vol. 8, no. 1, article 4, 2014.

[26] G. Qi, C.C. Aggarwal, and T. Huang, “Community detection with edge content in social media networks,” ICDE, pp. 534-545, 2012.

[27] C. Xu, D. Tao, and C. Xu, “A survey on multi-modality learning,” arXiv preprint, 2013.

[28] G. Pio, F. Serafino, D. Malerba, and M. Ceci, “Multi-type clustering and classification from heterogeneous networks,” Info. Sci., vol. 425, pp. 107-126, 2018.

[29] M. E. Newman and A. Clauset, “Structure and inference in annotated networks,” Nature communications, vol. 7, article: 11863, 2016.