Manifold Regularized Stochastic Block Model

Tiantian He
SCSE
Nanyang Technological University
Singapore
tiantian.he@ntu.edu.sg

Lu Bai
SCSE
Nanyang Technological University
Singapore
bailu@ntu.edu.sg

Yew-Soon Ong
SCSE
Nanyang Technological University
Singapore
asysong@ntu.edu.sg

***Abstract*—Stochastic block models (SBMs) play essential roles in network analysis, especially in those related to unsupervised learning (clustering). Many SBM-based approaches have been proposed to uncover network clusters, by means of maximizing the block-wise posterior probability that generates edges bridging vertices. However, none of them is capable of inferring the cluster preference for each vertex through simultaneously modeling block-wise edge structure, vertex features, and similarities between pairwise vertices. To fill this void, we propose a novel SBM dubbed manifold regularized stochastic model (MrSBM) to perform the task of unsupervised learning in network data in this paper. Besides modeling edges that are within or connecting blocks, MrSBM also considers modeling vertex features utilizing the probabilities of vertex-cluster preference and feature-cluster contribution. In addition, MrSBM attempts to generate manifold similarity of pairwise vertices utilizing the inferred vertex-cluster preference. As a result, the inference of cluster preference may well capture the comparability in the manifold. We design a novel process for network data generation, based on which, we specify the model structure and formulate the network clustering problem using a novel likelihood function. To guarantee MrSBM learns the optimal cluster preference for each vertex, we derive an effective Expectation-Maximization based algorithm for model fitting. MrSBM has been tested on five sets of real-world network data and has been compared with both classical and state-of-the-art approaches to network clustering. The competitive experimental results validate the effectiveness of MrSBM.**

***Index Terms*—Stochastic block model (SBM), generative model, network clustering, community detection, complex network, social network, biological network, document network, manifold regulation, vertex features**

I. Introduction

Real-world data containing relations are ubiquitous. To model them, the network which has a set of vertices and edges is always used to represent the data samples, and sample-sample relations. As vertices and edges can represent real data collected from every regard of real life, e.g., social players on social networking sites and their friendship, and cited documents in the corpus, network is ideal for preserving illustrative information on the data samples and complex relationships between them. Due to their importance and universality, how to effectively analyze the networked data has drawn great interest in the recent. Cluster analysis is one of the most important analytical tasks in network data, and it is directly related to many real applications, such as social group detection [12], [14], social recommendation [21], [28], biological module discovery [11], [13], and topic modeling in cited documents [5].

Cluster analysis in network data has ever been a challenging problem in machine learning community. And there have been a number of approaches proposed, aiming at effectively unfolding the hidden clusters. Whether these methods are heuristic, or model-based, all of them attempt to discover network clusters by means of maximizing the group-wise vertex cohesiveness, regarding edge density, vertex features, or both aforementioned. For example, modularity [6], which measures the difference in terms of density of vertices in the same group and the density if vertices are randomly grouped, is a popular benchmark. And it has been adopted by many heuristic approaches, such as Clauset-Newman-Moore algorithm (CNM) [6], and Fast unfolding algorithm [4], to uncover clusters in network data. In addition, machine learning techniques, including matrix factorization [12], [28], [29], spectral clustering [22], and Bayesian generative models [5], [27] have also been used to build effective model-based methods for network cluster analysis. Amongst these approaches, stochastic block models [1], [2], [15], [20] are one of the most desirable frameworks for network cluster analysis. SBMs firstly assume that there are K blocks hidden in the network and generating an edge connecting two blocks follows some probabilistic distribution. Then, each edge in the network is assumed to be generated according to the block affiliations of two endpoints and the corresponding probability of block-block connection. The cluster preference for each vertex and probability of edge generation inside/outside blocks can be learned by optimizing the posterior probability given the network structure and other hyper-parameters. SBMs have been important tools for effectively analyzing complex network data since it was proposed.

Though effective in capturing the vertex-wise cluster assignment and the characteristics of inter-block connectivity, most SBMs overlook the fact that higher-order similarities embedded in the network and vertex features may also influence the vertex-cluster preference. In this paper, we instead propose a novel SBM-based model, dubbed manifold regularized stochastic block model (MrSBM), to learn cluster assignment for each vertex concerning edge structure, vertex features, and manifold similarities. The contribution of this paper can be summarized as follows:

- We propose MrSBM, which is a novel stochastic block model for uncovering clusters in network data. Besides

TABLE I
COMPARISONS BETWEEN MRSBM AND OTHER APPROACHES. EDGE, VERTEX FEATURE AND SIMILARITY REPRESENTS WHETHER AN APPROACH CONSIDERS MODELING EDGE, VERTEX FEATURES AND VERTEX-VERTEX SIMILARITY WHEN PERFORMING CLUSTERING TASKS IN NETWORK DATA.

Approach	Model-based	Edge	Vertex Feature	Similarity
CNM	×	√	×	×
CoDA	√	√	×	×
NCut	√	√	×	√
CESNA	√	√	√	×
MISAGA	√	√	×	√
MrSBM	√	√	√	√

assuming that each edge is generated according to cluster preference and latent blocks, MrSBM also allows vertex-cluster preference to involve into the generation of vertex features and pairwise vertices similarities. Compared with prevalent SBMs and other approaches to network clustering, MrSBM utilizes multiple sources of information to infer the cluster membership.

- We design a novel generative process for generating network data, based on which, we formulate the clustering problem as optimizing a unified likelihood function. We also derive an Expectation-Maximization algorithm for inferring the optimal variables of the model.
- We extensively compared MrSBM with both classical and state-of-the-art approaches on five widely used network datasets. Experimental results show that MrSBM is very competitive, which validate the effectiveness of the model.

The rest of this paper is organized as follows. In Section II, the previous works related to network clustering and SBMs are investigated. In Section III, we elaborate the proposed MrSBM, derive the EM algorithm for learning the latent variables, and analyze the computational complexity of the proposed model. The extensive experiments that are used to verify the effectiveness of MrSBM are presented in Section IV. In the last section, we conclude the paper and discuss future works.

II. RELATED WORKS

Aiming at discovering latent groups where vertices are cohesive, network cluster analysis has been a fundamental problem in machine learning and data mining. Many approaches have been proposed to solve this problem. Some methods are heuristic-based. For example, Clauset-Newman-Moore algorithm (CNM) [6], and Fast unfolding algorithm [4] are two effective heuristic approaches which can uncover network clusters via modularity optimization. While, more approaches to network cluster analysis are model-based. For example, Communities from Edge Structure and Node Attributes (CESNA) [28], Communities through Directed Affiliations (CoDA) [29], Mining Interest Sub-Graphs (MISAGA) [12], Fuzzy Structural Patterns (FSPGA) [10], Semantic Community Identification (SCI) [25], and Contextual Correlation Preserving Multi-view Featured Graph Clustering (CCPMVFGC) [14] are effective matrix-factorization-based approaches to network clustering. Relational topic model (RTM) [5] and iTopic model [23] are two effective Bayesian models based on topic modeling [3]. Normalized cut [22] and Combining Structured Node Content and Topology (CSNCT) [9] are two effective models for network data, which are based on spectral clustering. As a powerful technique to network cluster analysis, several SBM based models are also proposed. For example, mixed membership stochastic block model (MMSB) [2], degree-corrected block model [16], and power-law degree SBM [20], are three effective Bayesian stochastic block models. Though these proposed methods are effective to some extent, most of them, especially those SBM based ones, do not consider the effect brought by vertex features and higher-order vertex similarities when modeling vertex-cluster preference. Their capability of discovering meaningful clusters in the network is thereby constrained. To address this issue, in this paper, we propose MrSBM, a novel stochastic block model that may infer cluster preference for each vertex concerning edge structure, vertex features, and manifold similarities. In Table I, we compare MrSBM with five representative approaches to network clustering, in terms of the essential features which a method for network clustering should possess. As the table shows, MrSBM is the only one that concerns the modeling of edge structure, vertex features, and vertex-vertex higher-order similarity when unfolding the hidden clusters in the network data.

III. STOCHASTIC BLOCK MODEL REGULARIZED BY MANIFOLD

In this section, we elaborate the details of the proposed manifold regularized stochastic block model (MrSBM). First, we introduce the mathematical notations used in this paper. We then introduce the structure of MrSBM according to a newly designed generative process. At last, we develop an Expectation-Maximization-based strategy to infer the optimal latent variables of MrSBM and analyze the computational complexity of the model.

A. Notations

Given a network composed of N vertices, $|\mathrm{E}|$ edges, and M vertex features, we use two binary matrices $\mathbf{Y} \in \{0,1\}^{N\times N}$, and $\mathbf{F} \in \{0,1\}^{N\times M}$, to represent whether two vertices are connected and whether a vertex has a corresponding feature, respectively. We denote $\mathbf{X} \in R_{+}^{N\times N}$ as the manifold matrix representing the similarity between pairwise vertices in the network. Similarity in manifold has been widely used in different machine learning tasks, e.g., feature extraction [31], dimensionality reduction [32], and discriminant analysis [30], it is able to guide a model to infer latent variables which more local information between pairwise data samples is embedded. In this paper, we let MrSBM concern more about local structural similarity as the topological information is the cornerstone of a network. Inspired by the diffusion theory [7],

TABLE II
NOTATIONS USED BY MRSBM

Notation	Meaning
N, M, K	Number of vertices, vertex features and blocks
$\mathbf{Y}$	Vertex adjacency matrix
$\mathbf{F}$	Vertex-feature matrix
$\mathbf{X}$	Manifold matrix representing vertex-vertex similarity
$\mathbf{V}_i$	K dimensional multinomial variables representing cluster preference for ith vertex
$\mathbf{U}_j$	K dimensional Gamma variables representing feature-cluster preference for jth feature
$\mathbf{B}$	$K \times K$ Bernoulli parameters representing probability of block-block edge generation
α, λ	Hyper-parameters of the model

we propose the following method to evaluate the vertex-vertex similarity in the network:

$$\mathbf{X}_{ij} = \begin{cases} \frac{[\mathbf{Y}^T\mathbf{Y}+\mathbf{Y}]_{ij}}{2d_i} + \frac{[\mathbf{Y}^T\mathbf{Y}+\mathbf{Y}]_{ij}}{2d_j} & \text{if } \mathbf{Y}_{ij} = 1, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where d_i denotes the degree of vertex i in the network. When two vertices, say v_i and v_j are connected, $\mathbf{X}_{ij}$ is computed by averaging the proportions of local structures shared by two end points.

To build the proposed model, we use K dimensional multinomial variables $\mathbf{V}_i$ to represent the cluster preference of vertex i. We use K dimensional Gamma variables $\mathbf{U}_j$ to represent the feature-cluster contribution of feature j. We use a Bernoulli parameter $\mathbf{B}_{kk'}$ to represent the generating probability of an edge bridging block k and k'. The (i, j)-th element of a matrix $\mathbf{Y}$ is denoted as $\mathbf{Y}_{ij}$. Table II summarizes the notations used in this paper.

B. Model structure

As mentioned, the proposed MrSBM takes into account the modeling of edge structure, vertex features, and vertex-vertex similarity. We, therefore, design a novel generative process for the model, which is depicted in Fig. 1. Given a network data contains N vertices, $|E|$ edges, and M vertex features, MrSBM assumes there are K latent blocks (clusters) in the network. And the network data are generated as follows:

- For each vertex, draw multinomial cluster preference $\mathbf{V}_{ik}$, $i = 1, ...N, k = 1, ..., K$;
- For each feature, draw Gamma variables as cluster contribution $\mathbf{U}_{jk}$, $j = 1, ...M, k = 1, ..., K$;
- For each pair of latent blocks, draw Bernoulli variables as probability of edge generation $\mathbf{B}_{kk'}$, $k = 1, ...K, k' = 1, ..., K$;
- For each pair of vertices in the network
  Draw Poisson likelihood as the probability of edge generation
  $p(\mathbf{Y}_{ij}|\mathbf{V}_i, \mathbf{V}_j, \mathbf{B}) =$
  $\alpha Poisson(\mathbf{Y}_{ij}|\sum_{k,k'} \mathbf{V}_{ik}\mathbf{B}_{kk'}\mathbf{V}_{jk'})$;
  Draw Gaussian likelihood as the probability of vertex-vertex similarity
  $p(\mathbf{X}_{ij}|\mathbf{V}_i, \mathbf{V}_j) = Gaussian(\mathbf{X}_{ij}|\sum_k \mathbf{V}_{ik}\mathbf{V}_{jk}, \lambda)$;

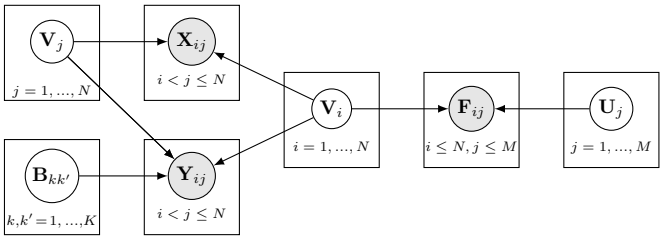

Fig. 1. Graphical representation of manifold regularized stochastic block model (MrSBM)

- For each vertex, draw Poisson likelihood as the probability of feature generation
  $p(\mathbf{F}_{ij}|\mathbf{V}_i, \mathbf{U}_j) = Poisson(\mathbf{F}_{ij}|\sum_k \mathbf{V}_{ik}\mathbf{U}_{jk})$,

where α and λ are two Gamma parameters which are used to control the relative weight of edge modeling in the model, and the precision of modeling manifold similarity, respectively. Given such a generative process, it is seen that MrSBM is fundamentally different from previous SBMs and many other approaches to network clustering, as it involves two additional likelihood functions, i.e., Poisson mass function for feature generation, and Gaussian density function for vertex-vertex similarity. Therefore, the estimation of vertex-cluster preference is simultaneously influenced by edge and feature generation, and similarity of pairwise vertices. It is expected that more meaningful cluster preferences will be inferred by MrSBM.

C. The joint likelihood

Based on the generative process introduced in Section III-B, the joint likelihood of the proposed model is:

$$\begin{aligned} p(\mathbf{Y}, \mathbf{F}, \mathbf{X}, \mathbf{V}, \mathbf{U}, \mathbf{B}|\alpha, \lambda) = & \alpha \prod_{i<j\leq N} p(\mathbf{Y}_{ij}|\mathbf{B}, \mathbf{V}_i, \mathbf{V}_j) \\ \cdot \prod_{i<j\leq N} p(\mathbf{X}_{ij}|\mathbf{V}_i, \mathbf{V}_j, \lambda) & \cdot \prod_{i\leq N, j\leq M} p(\mathbf{F}_{ij}|\mathbf{V}_i, \mathbf{U}_j) \end{aligned} \quad (2)$$

Taking the logarithm of the joint likelihood, we have:

$$\begin{aligned} & L(\mathbf{Y}, \mathbf{F}, \mathbf{X}, \mathbf{V}, \mathbf{U}, \mathbf{B}|\alpha, \lambda) = \\ & \alpha \sum_{i<j\leq N} [\mathbf{Y}_{ij} \log(\sum_{k,k'} \mathbf{V}_{ik}\mathbf{B}_{kk'}\mathbf{V}_{jk'}) - \sum_{k,k'} \mathbf{V}_{ik}\mathbf{B}_{kk'}\mathbf{V}_{jk'}] \\ & + \sum_{i\leq N, j\leq M} [\mathbf{F}_{ij} \log(\sum_k \mathbf{V}_{ik}\mathbf{U}_{jk}) - \sum_k \mathbf{V}_{ik}\mathbf{U}_{jk}] \\ & - \frac{\lambda}{2} \sum_{i<j\leq N} (\mathbf{X}_{ij} - \sum_k \mathbf{V}_{ik}\mathbf{V}_{jk})^2 + const \end{aligned} \quad (3)$$

where $const$ contains the terms that are irrelevant to all the latent variables. Base on Eq. (3), we may observe that the joint likelihood of the proposed model may increase when each edge is generated between an appropriate pair of blocks, each feature is generated by proper feature-cluster contributions, and vertex-vertex similarity is generated by appropriate cluster preferences of corresponding vertices. Therefore, when this joint likelihood function is maximized, optimal cluster preference for each vertex in the network can be inferred.

D. Learning MrSBM

As Eq. (3) shows, it is intractable to directly optimize the model through point estimation, as the summation operators are embedded into the logarithm operator. As a result, the latent variables, including $\mathbf{V}$, $\mathbf{B}$, and $\mathbf{U}$ can only be optimized via approximation methods. In this paper, we derive an alternative manner for updating the latent variables of the model, based on Expectation-Maximization (EM) framework [8]. In the E-step of each iteration, MrSBM constructs an auxiliary function which manifests the lower bound of the log-likelihood function regarding the latent variables. Then, this lower bound is maximized by MrSBM in the M-step. By iteratively updating the latent variables using EM algorithm, the proposed model can converge to the local optima.

To derive the lower bound of Eq. (3) in E-step, we use the following well known property of concavity of logarithmic functions:

$$\begin{aligned} &\log(\sum_k x_k) \geq \sum_k a_k \log(\frac{x_k}{a_k}) \\ &\text{if } \sum_k a_k = 1, x_k > 0 \end{aligned} \tag{4}$$

Given Eq. (4), we may derive the following auxiliary function as the lower bound of the joint likelihood of MrSBM (Eq. (3)):

$$\begin{aligned} &L(\mathbf{Y},\mathbf{F},\mathbf{X},\mathbf{V},\mathbf{U},\mathbf{B}|\alpha,\lambda) \geq Q(\theta,\phi,\gamma) = \\ &\alpha \sum_{i<j\leq N} [\mathbf{Y}_{ij} \sum_{kk'} \theta_{ij,kk'} \log(\frac{\mathbf{V}_{ik}\mathbf{B}_{kk'}\mathbf{V}_{jk'}}{\theta_{ij,kk'}}) \\ &-\sum_{k,k'} \mathbf{V}_{ik}\mathbf{B}_{kk'}\mathbf{V}_{jk'}] \\ &+ \sum_{i\leq N, j\leq M} [\mathbf{F}_{ij} \sum_k \phi_{ij,k} \log(\frac{\mathbf{V}_{ik}\mathbf{U}_{jk}}{\phi_{ij,k}}) - \sum_k \mathbf{V}_{ik}\mathbf{U}_{jk}] \\ &- \frac{\lambda}{2} \sum_{i<j\leq N} (\mathbf{X}_{ij}^2 - 2\mathbf{X}_{ij}\gamma_{ij} + (\sum_k \mathbf{V}_{ik}\mathbf{V}_{jk})^2) \\ &\theta_{ij,kk'} = \frac{\mathbf{V}_{ik}\mathbf{B}_{kk'}\mathbf{V}_{jk'}}{\sum_{k,k'} \mathbf{V}_{ik}\mathbf{B}_{kk'}\mathbf{V}_{jk'}}, \phi_{ij,k} = \frac{\mathbf{V}_{ik}\mathbf{U}_{jk}}{\sum_k \mathbf{V}_{ik}\mathbf{U}_{jk}} \\ &\gamma_{ij} = \exp\{\sum_k \omega_{ij,k} \log(\frac{\mathbf{V}_{ik}\mathbf{V}_{jk}}{\omega_{ij,k}})\}, \omega_{ij,k} = \frac{\mathbf{V}_{ik}\mathbf{V}_{jk}}{\sum_k \mathbf{V}_{ik}\mathbf{V}_{jk}} \end{aligned} \tag{5}$$

In E-step, we set the plug-in variables, including $\theta_{ij,kk'}$, $\phi_{ij,k}$, and γ_{ij} as Eq. (5) shows, so that these plug-in variables directly take effect on the corresponding latent variables of the model. Then, we are able to maximize Eq. (3) via pushing-up the lower bound $Q(\theta,\phi,\gamma)$.

In M-step, we maximize the lower bound $Q(\theta,\phi,\gamma)$, which is shown in Eq. (5), by simply performing point estimation. To optimize Q relevant to latent variable $\mathbf{V}_{ik}$, we construct the following Lagrangian function:

$$La(\mathbf{V}_{ik}) = Q(\theta,\phi,\gamma) - \nu[\sum_k \mathbf{V}_{ik} - 1] \tag{6}$$

Algorithm 1 Manifold regularized Stochastic Block Model (MrSBM)

Input: Network Data: $\mathbf{Y}$, $\mathbf{F}$
Output: Cluster preference for each vertex: $\{\mathbf{V}_i\}_{i=1}^N$; Feature-cluster contributions $\{\mathbf{U}_j\}_{j=1}^M$; Block-block edge generation $\{\mathbf{B}_{kk'}\}_{k,k'=1}^K$

```
Compute X;
Initialize U, V, B;
t ← 0;
while t < T_max do
    t ← t + 1;
    E-Step:
    Set lower bound of Eq. (3) by Eq. (5);
    M-Step:
    Maximize the lower bound via:
    Updating V_ik by Eq. (7);
    Updating B_kk' by Eq. (9);
    Updating U_jk by Eq. (8);
    Compute Log likelihood L^(t) by Eq. (3);
    if L^(t) − L^(t−1) ≤ ε then
        break;
    end if
end while
Identify cluster label for each vertex using V;
```

where ν denotes the Lagrange multiplier in terms of the unity constraint of $\mathbf{V}_i$. Taking partial derivative w.r.t. $\mathbf{V}_{ik}$ of Eq. (6) and let it equal to zero, we derive the updating rule for $\mathbf{V}_{ik}$:

$$\begin{aligned} &\mathbf{V}_{ik} \propto \frac{\Delta_{ik}}{\Lambda_{ik}} \\ &\Delta_{ik} = \alpha \sum_j \mathbf{Y}_{ij} \sum_{k'} \theta_{ij,kk'} + \sum_j \mathbf{F}_{ij}\phi_{ij,k} \\ &+ \lambda \sum_j \mathbf{X}_{ij}\omega_{ij,k} \exp\{\sum_k \omega_{ij,k} \log \frac{\mathbf{V}_{ik}\mathbf{V}_{jk}}{\omega_{ij,k}}\} \\ &\Lambda_{ik} = \alpha \sum_j \sum_{k'} \mathbf{V}_{jk'}\mathbf{B}_{kk'} + \sum_j \mathbf{U}_{jk} \\ &+ \lambda \sum_j (\sum_k \mathbf{V}_{ik}\mathbf{V}_{jk})\mathbf{V}_{jk} \end{aligned} \tag{7}$$

Similarly, we may obtain the updating rules for $\mathbf{U}_{jk}$ and $\mathbf{B}_{kk'}$. The rule for updating $\mathbf{U}_{jk}$ is:

$$\mathbf{U}_{jk} \propto \frac{\sum_i \mathbf{F}_{ij}\phi_{ij,k}}{\sum_i \mathbf{V}_{ik}} \tag{8}$$

The updating rule for $\mathbf{B}_{kk'}$ is:

$$\mathbf{B}_{kk'} \propto \frac{\sum_{i<j\leq N} \mathbf{Y}_{ij}\theta_{ij,kk'}}{\sum_{i<j\leq N} \mathbf{V}_{ik}\mathbf{V}_{jk'}} \tag{9}$$

By iteratively performing E-step and M-step, MrSBM will converge to local optima in a finite number of iterations. The process of variable learning of MrSBM has been summarized in Algorithm 1.

E. Analysis on model complexity

Based on the lower bound and updating rules shown in Eqs. (5), (7), (8) and (9), we can obtain the computational complexity of MrSBM as follows. Given Eq. (5), computing the auxiliary variables ($\theta_{ij,kk'}$, $\phi_{ij,k}$, and γ_{ij}) for the lower bound $Q(\theta, \phi, \gamma)$ follows the order of $O(K^2 + 4K)$. Given Eq. (7), updating the latent variable $\mathbf{V}_{ik}$ follows the order of $O(4NK + 2M)$. Given Eq. (8), updating the latent variable $\mathbf{U}_{jk}$ follows the order of $O(2N)$. Given Eq. (9), updating the latent variable $\mathbf{B}_{kk'}$ follows the order of $O(2N^2)$. Therefore, the overall complexity that MrSBM updates each latent variable is $O(2N^2)$.

IV. Experiments and Analysis

In this section, we conduct a series of experiments on real-world network datasets to validate the effectiveness of MrSBM against other classical or state-of-the-art methods.

A. Experimental Setup

1) Baselines for Comparison: We selected six approaches as baselines, including CNM [6], CoDA [29], NCut [22], k-means [18], CESNA [28], and MISAGA [12].

CNM, CoDA, and NCut are three representative approaches based on network topology. CNM is a typical method for community detection based on modularity optimization. CoDA performs network clustering via symmetric probabilistic matrix factorization. NCut is a spectral-based method for network clustering, which performs the task via assigning vertices sharing higher structural similarity into the same cluster.

k-means is a classical clustering method, which is able to discover clusters in the network using vertex features.

CESNA and MISAGA are two state-of-the-art approaches to network clustering by learning a shared latent space. CESNA is a probabilistic generative model, which learns a shared latent space as cluster preference for each vertex from edge structure and vertex features of a network. MISAGA is able to learn a shared latent space from the edge structure and pairwise similarity of vertex features.

In our experiments, we used the source codes of all the baselines provided by the authors for implementation. Algorithms including CNM and CESNA do not need any predefined parameter before they run. CoDA, NCut, k-means, and MISAGA need to pre-determine model parameters before they are executed, where we used the settings recommended in the corresponding papers. For the number of clusters, i.e., K, which have to be predetermined in CoDA, NCut, k-means, and MISAGA, we set it to be equal to the number of ground-truth clusters of the testing dataset. For the proposed MrSBM, we set $\alpha = 1$ and $\lambda = 1$. The setting of K in MrSBM is the same as that in the baselines. All of the experiments were performed on a workstation with 4-core 3.4GHz CPU and 16GB RAM and all approaches were executed 10 times to obtain a statistically steady performance.

TABLE III
Statistics of datasets used in the experiments. Soc, Bio, or Doc represents whether the dataset is a social, biological, or document network.

Dataset	Type	N	$\lvert E\rvert$	M	K
Cal	Soc	769	16656	53	10
Ego	Soc	4039	88234	1283	191
Gplus	Soc	8725	972899	5913	130
Wiki	Doc	2405	17981	4973	17
Biogrid	Bio	5640	59748	4286	200

2) Dataset Description: We used five real-world networks with verified ground-truth clusters as testing datasets, including three social networks, one biological network, and one document network. These real-world networks have different sizes and different numbers of vertex features. The detailed descriptions of these five datasets are as follows.

$Caltech$ (Cal) [24] is a college social network extracted from the social networking users in the California Institute of Technology. There are 769 vertices, 16656 edges and 53 vertex features representing the users, the social ties between them, and their profiles, respectively. In the Cal dataset, there are 10 large groups verified according to the college dorm system [24], which can be used as ground-truth clusters to evaluate the clustering performance of different approaches.

$Ego - facebook$ (Ego) [19] is a social network extracted from facebook.com. This dataset contains 4039 vertices, 88234 edges, and 1283 features which represent the Facebook users, the friendship, and the user profiles, respectively. There are 191 social circles that have been verified as ground-truth clusters.

$Googleplus$ (Gplus) [19] is also a social network constructed based on users from googleplus.com. It contains 8725 vertices, 972899 edges, and 5913 features, representing the users of googleplus, their social relationship, and their content characterizations, respectively. The vertex features are collected from five sources: jobs, locations, institutions, universities, and identity information. There are 130 social circles, which have been verified in previous studies and can be used as ground-truth clusters.

$Wiki$ [17] is an online document network collected from Wikipedia. There are 2405 vertices, 17981 edges, and 4973 vertex features, representing web-pages of wiki items, hyperlinks between web-pages, and keywords of wiki items, respectively. There are 17 document classes which have been verified as ground-truth clusters.

$Biogrid$ [26] is a biological network used to describe the interactions between proteins in $Saccharomyces\ cerevisiae$. There are 5640 vertices representing proteins, 59748 edges representing protein-protein interactions, and 4286 vertex features which are collected from three sources to characterize the proteins. In this dataset, there have been 200 protein complexes which have been discovered by laboratory experiments.

The statistics of these testing datasets are summarized in

TABLE IV
CLUSTERING PERFORMANCE EVALUATED BY NMI (%) AND ACC (%). THE BEST PERFORMANCE ON EACH DATASET IS HIGHLIGHTED IN BOLD.

Approaches \ Datasets	Cal		Ego		Gplus		Wiki		Biogrid	
	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc
CNM	42.298	30.949	48.266	37.979	11.847	21.410	30.572	41.788	76.274	2.128
CoDA	33.517	37.824	55.505	52.091	18.900	34.735	30.203	45.267	73.229	3.490
NCut	41.113	37.451	53.646	44.689	12.915	26.705	8.638	17.588	54.462	3.245
k-means	21.064	14.954	40.461	29.116	39.735	16.252	35.214	44.407	88.904	8.529
CESNA	39.259	38.429	57.513	46.124	23.158	24.038	9.374	24.738	80.088	2.570
MISAGA	29.774	25.618	56.452	45.159	21.553	53.009	35.109	45.114	89.485	9.167
MrSBM	**59.601**	**52.926**	**68.614**	**57.267**	**42.331**	**68.615**	**45.426**	**56.175**	**91.979**	**11.507**

Table III, where N is the number of vertices, $|E|$ is the number of edges, M is the number of vertex features, and K is the number of ground-truth clusters, respectively.

3) Evaluation Metrics: For performance evaluation, we selected two widely used metrics, i.e., the Normalized Mutual Information (NMI) [12] and the Accuracy (Acc) [10].

The NMI measures the overall accuracy of the matches between detected clusters and the ground-truth. It is defined as:

$$\begin{aligned} NMI &= \frac{\sum\limits_{C_i, C_j^*} Pr(C_i, C_j^*) \log \frac{Pr(C_i, C_j^*)}{Pr(C_i) Pr(C_j^*)}}{\max(H(C), H(C^*))}, \\ H(C) &= -\sum_i Pr(C_i) \log Pr(C_i), \\ H(C^*) &= -\sum_j Pr(C_j^*) \log Pr(C_j^*), \end{aligned} \tag{10}$$

where $Pr(C_i, C_j^*)$ denotes the probability that the vertices are shared in both the detected cluster i and the ground-truth cluster j, and $Pr(C_i)$ denotes the probability that a vertex belongs to cluster i. According to the definition in Eq. (10), a larger value of NMI indicates a better matching between the detected clusters and the ground-truth.

Unlike NMI, the Acc measures the accuracy of individually detected clusters. It is defined as follows:

$$Acc = \sum_i \frac{|C_i|}{|C|} f(C_i, C^*), \tag{11}$$

where $|C_i|$ and $|C|$ denote the size of detected cluster i and total number of data samples covered by the detected clusters, and $f(\cdot, \cdot)$ is defined as the maximum overlap between the detected cluster i and a cluster in the ground-truth database. Thus, Acc evaluates the best matching of each individual cluster. A larger value of Acc indicates a better matching between each detected cluster and the ground-truth. The larger the Acc values of all clusters detected by an algorithm, the better the performance of an algorithm. The properties of these two evaluation metrics enable them to evaluate the effectiveness of an approach in a complementary manner, so that all the clustering approaches can be evaluated comprehensively.

B. Clustering Performance Comparison

Social community detection, document segmentation, and biological module identification are typical applications of detecting clusters in network data. In our experiments, we used the aforementioned different types of networks to test the effectiveness of the aforementioned six approaches. As the ground-truth clusters of the five testing datasets have been verified in previous studies, we are able to validate the clusters discovered by different approaches against the ground-truth. The experimental results (in terms of NMI and Acc) of all algorithms are summarized in Table IV.

When NMI is considered, MrSBM outperforms all the other baselines in all the testing networks. In three datasets out of the five, the proposed approach significantly outperforms the second-best by at least 15%. Specifically, in the Cal dataset, MrSBM outperforms CNM by 40.907%; in the Ego dataset, the proposed approach outperforms CESNA by 19.302%; and in Wiki, the improvement of MrSBM against k-means is 28.999%.

Similarly, when using the Acc metric, MrSBM still outperforms all the other baselines in all the testing networks. In four datasets, the proposed method outperforms other baselines by more than 20%. Specifically, when detecting social communities in Cal and Gplus, MrSBM outperforms CESNA and MISAGA by 37.724% and 29.440%, respectively; when discovering document clusters in dataset Wiki, MrSBM outperforms the second-best, CoDA, by 24.097%; when identifying functional modules in Biogrid dataset, MrSBM is better than MISAGA by 25.526%.

From the experimental results in terms of NMI and Acc, we can observe that MrSBM is effective in network clustering. Besides modeling edge structure using vertex-preference and block-block connection generation, MrSBM also considers modeling vertex features utilizing vertex-preference and feature-cluster contributions. In addition, the inference of cluster preference for each vertex also benefits from manifold regularization regarding vertex-vertex similarities. As such

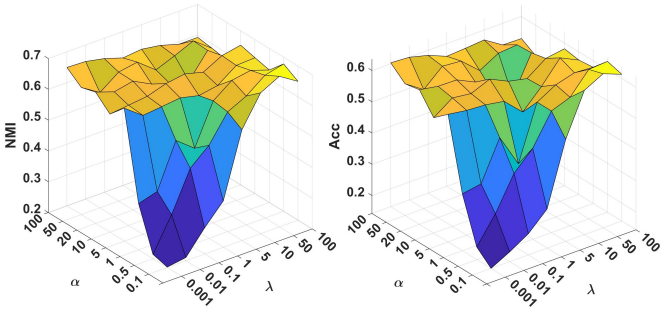

Fig. 2. Sensitivity analysis of MrSBM with respect to the parameter α and λ in Cal dataset

similarity is able to truly capture the local structural properties between pairwise vertices, MrSBM is propelled to learn similar cluster preferences for those vertices sharing higher comparability and is robust to noisy edges in the network data.

C. Parameter Sensitivity Analysis

We investigate the parameter sensitivity in this section to understand how the variation of α and λ, which lead to different weights of edge structure modeling and various precision of modeling manifold similarity respectively, will impact the clustering performance. Specifically, we set $\alpha = [0.1, 0.5, 1, 5, 10, 20, 50, 100]$ and $\lambda = [0.001, 0.01, 0.1, 1, 5, 10, 50, 100]$, and run MrSBM on all datasets. We then evaluate the clusters discovered by MrSBM using different settings of α and λ in terms of NMI and Acc. We present the results obtained from dataset Cal in Fig. 2 to show how the clustering performance of MrSBM is impacted by different settings of hyper-parameters. As depicted in Fig. 2, both NMI and Acc perform robustly when the value of α and λ is relatively large, e.g., $\alpha \geq 1$ and $\lambda \geq 1$. According to the results of sensitivity analysis shown in Fig. 2, the performance of the proposed MrSBM is relatively robust under a wide range of hyper-parameter combinations. For simplicity, we set $\alpha = 1$ and $\lambda = 1$ in all our experiments.

D. Model convergence test

In addition to derive the EM algorithm for updating the latent variables of MrSBM, we also investigated the convergence speed of MrSBM on real network datasets. Specifically, we recorded the value of log-likelihood function (Eq. (3)) for the first 300 iterations on all the five datasets. The variations of log-likelihood values collected from all the testing datasets have been shown in Fig. 3. As depicted, in any of the five testing datasets, the log-likelihood value may converge in less than 150 iterations, which showcases the capability of the derived EM algorithm to guarantee the model convergence and attain the optimal clustering results efficiently.

E. Scalability comparison between MrSBM and CESNA

Besides analyzing the model convergence of the proposed model, we also investigated the efficiency of MrSBM and compared it with CESNA, which is a well known efficient

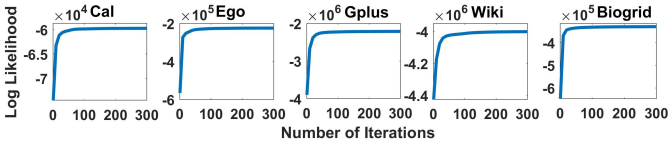

Fig. 3. Model convergence in testing datasets

model-based approach to network clustering. We recorded the optimization time used by MrSBM and CESNA in each testing dataset. For MrSBM, it costs 6.07 seconds in Cal, 808.25 seconds in Ego, 3349.53 seconds in Gplus, 261.68 seconds in Wiki, and 1728.64 seconds in Biogrid, respectively. For CESNA, it costs 2.72 seconds in Cal, 372 seconds in Ego, 2836 seconds in Gplus, 342 seconds in Wiki, and 4080 seconds in Biogrid, respectively. It is observed that the efficiency of MrSBM is competitive when compared with CESNA. Completing the model fitting in a relatively short time ensures MrSBM to uncover network clusters efficiently.

V. Conclusion

In this paper, we propose a novel model, dubbed manifold regularized stochastic block model (MrSBM), for cluster analysis in network data. Different from previous SBMs and other prevalent approaches to network clustering, MrSBM attempts to learn the cluster preference for each vertex concerning both edge structure and vertex features, and the learning of cluster preference is simultaneously regularized by vertex-vertex structural similarity in the manifold. The proposed MrSBM has been tested on five real-world networks and has been compared with six competitive baselines. The experimental results show that MrSBM outperforms all the baselines on all the datasets. In the future, we will attempt to develop Bayesian MrSBM via fully specifying the prior probabilities of latent variables. We will also further improve MrSBM's capability of dealing with large network data via developing scalable algorithms for inferring the optimal latent variables.

Acknowledgement

This research is supported by the Data Science and Artificial Intelligence Center at Nanyang Technological University, the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-RP-2018-004), and the National Natural Science Foundation of China under Grant 61802317.

References

[1] E. Abbe, "Community detection and stochastic block models: recent developments," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6446–6531, 2017.

[2] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *Journal of machine learning research*, vol. 9, no. Sep, pp. 1981–2014, 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[5] J. Chang and D. Blei, "Relational topic models for document networks," in *Artificial Intelligence and Statistics*, 2009, pp. 81–88.

[6] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.

[7] S. Coussi-Korbel and D. M. Fragaszy, "On the relation between social dynamics and social learning," *Animal behaviour*, vol. 50, no. 6, pp. 1441–1453, 1995.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[9] T. Guo, J. Wu, X. Zhu, and C. Zhang, "Combining structured node content and topology information for networked graph clustering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 11, no. 3, p. 29, 2017.

[10] T. He and K. C. Chan, "Discovering fuzzy structural patterns for graph analytics," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 5, pp. 2785–2796, 2018.

[11] T. He and K. C. Chan, "Evolutionary graph clustering for protein complex identification," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 3, pp. 892–904, 2018.

[12] T. He and K. C. Chan, "Misaga: An algorithm for mining interesting subgraphs in attributed graphs," *IEEE transactions on cybernetics*, vol. 48, no. 5, pp. 1369–1382, 2018.

[13] T. He and K. C. Chan, "Measuring boundedness for protein complex identification in ppi networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 16, no. 3, pp. 967–979, 2019.

[14] T. He, Y. Liu, T. H. Ko, K. C. Chan, and Y. S. Ong, "Contextual correlation preserving multiview featured graph clustering," *IEEE transactions on cybernetics*, 2019.

[15] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[16] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical review E*, vol. 83, no. 1, p. 016107, 2011.

[17] Q. Lu and L. Getoor, "Link-based classification," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003, pp. 496–503.

[18] D. J. MacKay and D. J. Mac Kay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[19] J. Mcauley and J. Leskovec, "Discovering social circles in ego networks," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 8, no. 1, p. 4, 2014.

[20] M. Qiao, J. Yu, W. Bian, Q. Li, and D. Tao, "Adapting stochastic block models to power-law degree distributions," *IEEE transactions on cybernetics*, no. 99, pp. 1–12, 2018.

[21] X. Shen, S. Pan, W. Liu, Y. S. Ong, and Q. S. Sun, "Discrete network embedding," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 3549–3555.

[22] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[23] Y. Sun, J. Han, J. Gao, and Y. Yu, "itopicmodel: Information network-integrated topic modeling," in *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, 2009, pp. 493–502.

[24] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, "Comparing community structure to characteristics in online collegiate social networks," *SIAM review*, vol. 53, no. 3, pp. 526–543, 2011.

[25] X. Wang, D. Jin, X. Cao, L. Yang, and W. Zhang, "Semantic community identification in large attribute networks." in *AAAI*, 2016, pp. 265–271.

[26] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic acids research*, vol. 30, no. 1, pp. 303–305, 2002.

[27] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng, "Gbagc: A general bayesian framework for attributed graph clustering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 9, no. 1, p. 5, 2014.

[28] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *Data Mining (ICDM), 2013 IEEE 13th international conference on*. IEEE, 2013, pp. 1151–1156.

[29] J. Yang, J. McAuley, and J. Leskovec, "Detecting cohesive and 2-mode communities indirected and undirected networks," in *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 2014, pp. 323–332.

[30] Q. Ye, L. Fu, Z. Zhang, H. Zhao, and M. Naiem, "Lp-and ls-norm distance based robust linear discriminant analysis," *Neural Networks*, vol. 105, pp. 393–404, 2018.

[31] Y. Zhang, Z. Zhang, J. Qin, L. Zhang, B. Li, and F. Li, "Semi-supervised local multi-manifold isomap by linear embedding for feature extraction," *Pattern Recognition*, vol. 76, pp. 662–678, 2018.

[32] Z. Zhang, T. W. Chow, and M. Zhao, "Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 5, pp. 1148–1161, 2012.