IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 23, 2019, accepted November 12, 2019, date of publication November 22, 2019,
date of current version December 11, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2955161*

# Detecting Overlapping Community Structure With Node Influence

QIANG ZHOU[1], SHIMIN CAI[1,2], AND YICHENG ZHANG[2,3]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Institute of Fundamental and Frontier Science, University of Electronic Science and Technology of China, Chengdu 611731, China
[3]Department of Physics, University of Fribourg, 1700 Fribourg, Switzerland

Corresponding author: Shimin Cai (shimin.cai81@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61673086, and in part by the Science Promotion Programme of UESTC, China, under Grant Y03111023901014006.

**ABSTRACT** Discovering the underlying overlapping community divisions can guide us in better exploring and predicting the structure and properties of the network. However, a large number of existing methods assume that nodes belong only to a single community. In this paper, we designed a posterior probabilistic prediction model under the Mixed-Membership Stochastic Blockmodel framework to accurately detect the overlapping community structure that exists in the network. In order to capture the degree of nodes that exhibit heterogeneous characteristics in the network, the model takes into account the influence of the nodes. In addition, we developed a non-conjugated stochastic variational inference to deduce the link probability prediction model with node influence. The key strategy is to use the mean-domain variational family with variable distribution to approximate the posterior community strengthen and node influence distribution in the prediction model. We compared the performance of this model with the previous algorithm models on computer-generated and real-world networks and found that it gives better results, especially when the heterogeneity of the network is very serious. In general, the combination of node influence and link probabilistic predictive model provides a new idea for us to use a statistical model to explore large-scale overlapping networks.

**INDEX TERMS** Community overlap, heterogeneous network, node influence, probability prediction model, structure detection.

## I. INTRODUCTION

The diverse entity relationships that exist in nature can be abstracted into a complex system with special associations, which can be described by a network or a graph. Common examples include society, the Internet, technology and biological network [1]–[5]. Because of its attractiveness in statistics and forecasting, it has become the focus of many current research fields. From the perspective of local attributes, similarity [6], [7], link distribution prediction [8], [9], clustering [10] and correlation [11] play an important role. However, the most interesting of these is the functional unit with a specific structure, community [2], [12]. Many networks in reality are revealed to have a community structure that nodes within a group have more dense connections than between them [13]–[15].

Communities have a fundamental interest in complex networks because of its potential functional implications. Therefore, the deep excavation and exploration of the properties determined by its topological structure and additional attributes is a challenging task, which can help us to understand its essence through the appearance of the network. It is a common method in community detection to solve some special problems by establishing an algorithm model for particular application environment, for example, in the course of the spread of epidemic over time [16]. Modularity maximization is one of the most widely used models in the current algorithmic prototypes. Modularity [4] is an object function that evaluates the fineness of network partitioning, with higher score means that the detected community partitions internally correspond to more edges than between them. Unfortunately, the exhaustive modularity maximization over all community partition on the network is a NP-Hard problem [17], and the algorithm itself has been proved to have a resolution defect [18].

The discovery of node-based community structures often manifests empirically two properties. The first is the similarity of nodes, that is, the nodes with similar characteristics

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.

should be classified into the same community according to the network topology or the attributes of the nodes themselves. The second is the influence of nodes, whose definition is based on the preferential attachment principle [5] of nodes in the network. The basic definition will give priority to join the community where the nodes with higher degree distribution, which is mathematically characterized by a power law distribution feature. Therefore, when designing a probabilistic prediction model of community structure, the key attributes of nodes should be fully considered so as to achieve a balance in accuracy and efficiency. In addition, for networks with different structural characteristics, the attributes of nodes should be determined according to specific situation(because the network characteristics are different, there is no uniform standard), especially the influence of the node.

With the evolution of the network, its properties are changing as well. In the research of one-dimension network model, it is revealed that there are two different types of nodes in the network, which is called bipartite network [19]. Although edges in bipartite networks exist only between different types of nodes, they can be converted to a one-dimension network through the common neighbor mapping relationship between nodes. Thus, we herein focus on the discovery of community structure in one-dimension model of network accurately. As the characteristic of the network is mainly determined by its topology and the node itself, the influence of the nodes in this paper refers to the fact that in the heterogeneous network condition, a few nodes have a larger node degree distribution, while most nodes have only a few connections. Such nodes have the advantage of preferential connection [5], [20] in the network, which makes the nodes that already have many connections become fatter, so it is inevitable that some nodes belong to multiple communities at the same time. As a result, the concept of overlapping nodes among communities has been proposed. For example, a person is usually associated with many social organizations, such as family, friends, relatives and colleagues. He plays an active member simultaneously in the fields of physics, mathematics, biology, computer science, et al. [14], [21]. For PPI networks(protein-protein interaction), proteins may also belong to more than one functional unit and play a bridging role that allows for information transmission [22]. Therefore, the nodes influence with overlapping characteristics illustrates a very important practical value in the research of community detection.

The earliest overlapping community discovery algorithm based on node attributes is CPM algorithm proposed by Palla *et al.* [23], whose goal is to find the adjacent $k$-clique, which is also the overlapping community structure searched by CPM algorithm. The disadvantage is that only overlapping community structures based on $k$-clique can be found. On this basis, Zhang et al. proposed MOHCC algorithm [24], by looking for the $k$-clique and combining coupling strength to guide the merger of $k$-clique, and finally selected the best hierarchical division for the obtained tree graph by using the extended segmentation density index. Lancichinetti et al. also proposed an OSLOM algorithm with random perturbation to express the local optimization fitness function of the statistical importance of the community [25]. The algorithm first looks for important communities based on the adaptability function, then discovers the internal structure or possible integration among them, and finally detects the hierarchy of potential communities. Eustace et al. proposed a neighbor scale matrix model that filters the relationships between nodes in the network that are below the average number of neighbor nodes, the Person clusters are used to determine the number of communities in the network and combine non-negative matrix decomposition algorithms for overlapping community discovery [26]. The Poisson model based on statistical inference is an implicit overlapping community probability statistical model that contains the correct node degree [27]. However, its inherent maximum likelihood may make the community parameters dependent on its nodes very small, so it cannot work well in model prediction. By analyzing the topological structure attributes of binary networks, Cui et al. proposed an overlapping community discovery algorithm that assigns free nodes to bi-communities structure according to given rules in the bipartite network. The disadvantage is that the relevant bi-communities and free nodes need to be extracted initially [28]. According to the rank of the node popularities within communities, Jin et al. proposed an efficient Bayesian optimization objective function based on the stochastic generation model to discovery the community structure with overlapping phenomenon [29]. Its shortcoming is that the parameter selection in the modeling process will directly affect the final number of communities. As these algorithms only focus on the topological structure of the network and some additional attributes of nodes(such as similarity), but ignore the key attribute of node influence, they cannot effectively simulate overlapping network communities with high heterogeneity in the real world.

In this paper, in order to make better use of statistical methods to detect the key attributes existing in the community structure, we proposed a link probabilistic model to capture the influence of the nodes on the division of communities with overlapping phenomena. The model is based on the Mixed-Membership Stochastic Blockmodel(MMSB) [6], a community detection algorithm that allows nodes to belong to more than one community at the same time, which provides better fit to the real world network, but ignores the influence of nodes in the overlapping community. Considering the nodes influence can further strengthen the detection of communities with overlapping phenomena [30], therefore, the advantages of the algorithm proposed in this paper are as follows: (1) the number of potential communities in the network can be accurately detected; (2) it can effectively capture nodes with great influence in the network. In addition, we developed a novel method of non-conjugate stochastic variational inference to deduce the algorithm model, that is, the mean-domain variational family with variable distribution is used to approximate the posterior community strengthen and node influence distribution in the link probability prediction model, so as to predict the potential community structure in the network efficiently. We tested it on the computer-generated and real-world networks, which can achieve higher community detection accuracy than some other advanced algorithms.

The structure of this paper is organized as follows. In Section II, we gave the definition of node influence and some relevant detail description. Along the clue, the formal derivation of the link prediction algorithm based on non-conjugate stochastic variational inference appears in Section III. Performance test evaluations of algorithms for computer-generated and real-world networks are presented in Section IV. The summary is in Section V.

## II. DESCRIPTION OF THE NODE INFLUENCE

The mixed-membership stochastic blockmodel(MMSB) [6] indicates the effects of the interactions between nodes $p$ and $q$ in a network on the probability of the edges appearing between them. Assume that the community memberships of node $p$ is represented by $\pi_p$, a distribution over communities, then the probability that two nodes are connected is determined by the *homophily*(similar properties between nodes are more likely to be connected together [31], [32]) of their community members and the strength of the communities they share. Given a network with $C$ latent groups, the edge indicator vectors $L_{pq}$ are independent for a pair of nodes $(p, q)$. We draw $L_{pq}$ by selecting a community allocation $(z_{p\rightarrow q}, z_{p\leftarrow q})$ for a pair of nodes $(p, q)$. Thus, a community can be represented by constructing a binary matrix and the probability of an edge appearing in MMSB can be expressed as Eq.(1),

$$\rho(L_{pq}=1|z_{p\rightarrow q,i}, z_{p\leftarrow q,j}, B)=\sum\nolimits_{i=1}^{C}\sum\nolimits_{j=1}^{C} z_{p\rightarrow q,i}z_{p\leftarrow q,j}\beta_{ij}, \tag{1}$$

where **B** is a matrix can parameterize any kind of distribution, here we treat it as a blockmodel matrix to be evaluated. In order to capture the homogeneity of the nodes, the off-diagonal elements of the blockmodel are set close to zero in MMSB, which means that if there is a connection between two nodes, their potential community indicators may be the same.

In this paper, we combined the MMSB with the node influence to make it possible to take into account the characteristics of the priority connection while capturing the homogeneity of the node. The aim is to make its potential edge appeal independent of its community membership. Combined with Eq.(1), we use a logit model to represent this node influence in Eq.(2),

$$logit(\rho(L_{pq}=1|z_{p\rightarrow q}, z_{p\leftarrow q}, B, \boldsymbol{\kappa}))\equiv\kappa_p+\kappa_q+\sum\nolimits_{c=1}^{C}\delta_{pq}^{c}\beta_c, \tag{2}$$

where $\kappa$ is used to capture the node influence and $\delta_{pq}^{c} = z_{p\rightarrow q,c}, z_{p\leftarrow q,c}$. If all nodes are in the same community $c$, then $\delta_{pq}^{c} = 1$. We note that Eq.(2) is similar to the random effect matrix [33] in principle, that is, $\sum_{c=1}^{C}\delta_{pq}^{c}\beta_c$ represents the interaction of the potential communities and $\kappa_p$ implies the node influence. If necessary, Eq.(2) can also be extended to include the node covariates.

In addition, Eq.(1) itself is a logarithmic-linear model, which means that the expected prediction probability of edge connection has a multiplicative dependence on the observed node covariables. Based on this statistical rule, we can convert the community strength parameter $\beta_c$ and node influence parameter $\kappa$ related to the prediction model into a distribution corresponding to the real world, such as the Gaussian distribution, which also helps simplify the subsequent calculation. In summary, the stochastic variational inference algorithm based on node influence in this paper can be summed up as follows:

**Algorithm 1** Stochastic Variational Inference Based on Node Influence

```
**Input**: Node pairs $(p, q)$ in a given network $G$, the community potential strength $C$, the community membership $\pi_p$ and the node influence $\kappa_p$
**Output**: The potential community structure $C_p$
Step (1): Specify the potential community strength $C$ as $\beta_c \sim \mathcal{N}(\mu_0, \sigma_0^2)$.
Step (2): For any node $p$, assign the community membership $\pi_p \sim$ Dirichlet$(\alpha)$.
Step (3): For any node $p$, define the node influence as $\kappa_p \sim \mathcal{N}(0, \sigma_0^2)$.
Step (4): For each pair of nodes $(p, q)$, assign interaction indicator $z_{p\rightarrow q} \sim \pi_p$ and $z_{p\leftarrow q} \sim \pi_q$.
Step (5): Calculate the probability of an edge $(L_{pq} \mid z_{p\rightarrow q}, z_{p\leftarrow q}, \kappa, \beta) \sim logit^{-1}(z_{p\rightarrow q}, z_{p\leftarrow q}, \kappa, \beta)$.
Step (6): If there are nodes in the network that are not visited, repeat steps (2)-(5).
Step (7): According to the edge probability between nodes and node influence captured, the nodes are assigned to the appropriate community and the final potential community structure $C_p$ is obtained.
```

Under the computational framework given by Algorithm 1, we can further carefully analyze the posterior empirical distribution based on the potential variable $\rho(\pi_{1:N}, \kappa_{1:N}, \beta_{1:C}|L, \alpha, \mu_0, \sigma_0^2, \sigma_1^2)$(please refer to Section III for detailed mathematical derivation of the model), where the posterior over $\pi_{1:N}$ denotes the community memberships of nodes, $\kappa_{1:N}$ denotes the node influence. Because these potential structural variables can be estimated in many ways, there are a number of possibilities for describing the network. For simplicity, we can replace the potential community strength $\boldsymbol{\beta}$ of $C$ with a single community strength $\beta$ which gives an excellent result on small network. In general, the prediction of this potential network structure can help us to study the individual links, the similarity between the nodes and the node influence on the overall effect.

## III. STOCHASTIC VARIATIONAL INFERENCE IN NON-CONJUGATE MODEL

For the posterior distribution $\rho(\pi_{1:N}, \kappa_{1:N}, \beta_{1:C}|L, \alpha, \mu_0, \sigma_0^2, \sigma_1^2)$, the exact calculation is intractable, so we use stochastic variational inference [21] to approximate the solution. However, there are some problems here. Traditional variational inference [34] is a kind of coordinate ascent algorithm, which requires all nodes in the network to satisfy the conditional conjugate, which is different from the method in this paper. Take the strength of community memberships and node influence for example, they present a Gaussian prior, so they do not meet the need of conjugate conditions. Secondly, the coordinate ascent algorithm iterates all pairs of nodes when performing variational inference, which makes it an impossible task for large-scale network. Therefore, in the experiment, we adopted a sub-sampling strategy to reduce the number of iterations in the calculation, by establishing a stochastic gradient model to limit the target object to a lower boundary.

### A. OPTIMIZATION OBJECT OF VARIATIONAL INFERENCE

In a complex probabilistic model, variational inference is a very effective method to approximate the posterior inference [34]. It looks for a near-posterior fitting parameter distribution by defining a parameterized family based on the distribution of hidden variables. And the approximate degree is usually measured by Kullback-Leibler(KL) divergence [7]. Thus, the parametric posterior inference problem is transformed into a model optimization problem.

In this paper, we defined a family of distributions based on the hidden variable $h(\boldsymbol{\pi}, \boldsymbol{\kappa}, \boldsymbol{\beta}, z)$ and find the family members that are closest to the posterior. As the mean-domain variational [14] family independently considers each hidden variable with a different parameterized distribution, the obtained variation distribution object is shown in Eq.(3).

$$h\left(z_{p\to q} = i, z_{p\leftarrow q} = j\right) = \phi_{pq}^{ij}; \quad h(\beta_c) = N(\beta_c; \mu_c, \sigma_\beta^2);$$
$$h\left(\pi_n\right) = Dirichlet(\pi_n; \gamma_n); \quad h(\kappa_n) = N(\kappa_n; \lambda_n, \sigma_\kappa^2). \quad (3)$$

Here, the *per-interaction memberships* $\phi_{pq}$ represents the posterior over the joint distribution of edges assigned by the community to per node pair $(p, q)$, the community memberships $\gamma$, the community strength distributions $\mu$ and the node influence distributions $\lambda$. Finding the optimization problem of the family distribution of $h$ close to the true posterior can be transformed into minimizing KL divergence, i.e., optimizing an evidence lower bound $\mathcal{L}$, an observation-based log-likelihood. Thus, we use a little trick that constructs it through Jensen's inequality [14], the specific process is as follows:

$$\begin{aligned}\mathcal{L} = &\sum\nolimits_n \mathbb{E}_h[\log \rho(\pi_n|\alpha)] - \sum\nolimits_n \mathbb{E}_h[\log h(\pi_n|\gamma_n)] \\ &+ \sum\nolimits_n \mathbb{E}_h[\log \rho(\kappa_n|\sigma_1^2)] - \sum\nolimits_n \mathbb{E}_h[\log h(\kappa_n|\lambda_n, \sigma_\kappa^2)] \\ &+ \sum\nolimits_c \mathbb{E}_h[\log \rho(\beta_c|\mu_0, \sigma_0^2)] - \sum\nolimits_c \mathbb{E}_h[\log h(\beta_c|\mu_c, \sigma_\beta^2)] \\ &+ \sum\nolimits_{p,q} \mathbb{E}_h[\log \rho(z_{p\to q}|\pi_p)] + \sum\nolimits_{p,q} \mathbb{E}_h[\log \rho(z_{p\leftarrow q}|\pi_q)] \\ &- \sum\nolimits_{p,q} \mathbb{E}_h[\log h(z_{p\to q}, z_{p\leftarrow q}|\phi_{pq})] \\ &+ \sum\nolimits_{p,q} \mathbb{E}_h[\log \rho(L_{pq}|z_{p\to q}, z_{p\leftarrow q}, \boldsymbol{\kappa}, \boldsymbol{\beta})]. \end{aligned} \quad (4)$$

Noting that the first three lines in Eq.(4) are summation operations for all communities and nodes, which can be considered as global update operations for the network. The remaining three lines are the sum operation for all node pairs, which can be considered as local update operations of the network. Distinguishing between global update operations and local update operations is a necessary task because local update operations rely only on a few global terms that correspond to them, whereas global update operations require the participation of all local update operations.

Since the coordinate ascent algorithm takes into account each pair of nodes in each iteration, it is very strict to the cost of system resources. In [7], a conditional crossover strategy is adopted to solve this problem. However, the method in this paper is not conjugate, so we can only turn to find other approximate solutions.

### B. REDUCE THE BOUNDARY OF VARIATIONAL OBJECT

Optimizing the Eq.(4) directly increases the computational complexity, especially the global update operations. In order to simplify the calculation further, the last line in Eq.(4) can be rewritten as Eq.(5),

$$\begin{aligned}&\mathbb{E}_h[\log \rho(L_{pq}|z_{p\to q}, z_{p\leftarrow q}, \kappa, \beta)] \\ &\quad = L_{pq}\mathbb{E}_h[x_{pq}] - \mathbb{E}_h[\log(1 + \exp(x_{pq}))], \end{aligned} \quad (5)$$

where we let $x_{pq} \equiv \kappa_p + \kappa_q + \sum_{c=1}^{C} \beta_c \delta_{pq}^c$. In order to make it easier to expand the discussion of Eq.(5), we still use Jensen's inequality to reduce its boundary (see Eq.(6)), where let $s_{pq} \equiv \sum_{c=1}^{C} \phi_{pq}^{cc} \exp\left\{\mu_c + \sigma_\beta^2/2\right\} + \left(1 - \sum_{c=1}^{C} \phi_{pq}^{cc}\right)$. Considering that $h_{\kappa_n}$ has a Gaussian distribution, which allows us to further simplify Eq.(6). Then according to the logarithmic normal distribution, we find $\mathbb{E}_h[\exp(\kappa_n)] = \exp\left(\lambda_n + \sigma_\kappa^2/2\right)$. The same alternative is also applied to $\beta_c$.

$$\begin{aligned}&-\mathbb{E}_h[\log(1 + \exp(x_{pq}))] \\ &\quad \geq -\log\left[\mathbb{E}_h(1 + \exp(x_{pq}))\right] \\ &\quad = -\log\left[1 + \mathbb{E}_h\left[\exp(\kappa_p + \kappa_q + \sum\nolimits_{c=1}^{C} \beta_c \delta_{pq}^c)\right]\right] \\ &\quad = -\log[1 + \exp(\lambda_p + \sigma_\kappa^2/2)\exp(\lambda_q + \sigma_\kappa^2/2)s_{pq}], \end{aligned} \quad (6)$$

Finally, by replacing Eq.(6) with Eq.(4), we can obtain a low boundary $\mathcal{L}'$, which can be easily processed, and this also allows us to infer a coordinate ascent strategy by updating the global and local operation iteratively.

### C. GLOBAL UPDATE OPERATION

It can be clearly seen from Eq.(4) that the parameter related to the global update operation are$(\gamma, \lambda, \mu)$. When Eq.(6) is brought into Eq.(4), it becomes the update of the above global parameters with stochastic gradient of lower boundary on $\mathcal{L}'$.

Similar to the practice in [7], we use the natural gradient of $\mathcal{L}'$ after each iteration to update all the node influences and community memberships, but adopt different stochastic optimization methods to maintain the independent learning rate of each node, which limits the nodes that need to be updated for each iteration to a smaller range.

Considering that the variational optimization object is the summation of the terms, we can first sub-sampling the subset and then extend the gradient according to the appropriate proportions. Specifically, in each iteration, the $N$ nodes of the network are sampled randomly and evenly(we sampled a "mini-batch "$S$ of nodes during each iteration). The purpose of doing like this is not only to reduce the impact of noise [14], but also to make full use of the characteristics of network sparsity to improve efficiency. In many real networks, only a small subset of nodes are linked. Therefore, for each sampling point, it contains only all links associated with its observations and a small number of non-links that are evenly sampled.

Suppose $\partial\gamma_p^t$ is the natural gradient of $\mathcal{L}'$ with respect to $\partial\gamma_p$, $\partial\lambda_p^t$ is the gradient with respect to $\lambda_p$ and $\partial\mu_c^t$ is the gradient with respect to $\mu_c$, according to [6], [14], [21], we find Eq.(7),

$$\begin{aligned}\partial\gamma_{p,c}^t = \alpha_c - \gamma_{p,c}^{t-1} + \sum\nolimits_{(p,q)\in links(p)}\phi_{pq}^{cc}(t)\\ + \sum\nolimits_{(p,q)\in nonlinks(p)}\phi_{pq}^{cc}(t),\end{aligned} \tag{7}$$

where links($p$) and non-links($p$) correspond to the links and non-links sets of node $p$ in the training set. And we also can get an unbiased estimate of the sum over the non-links by sub-sampling the non-links of the node. Because this is a scale contraction of the original natural gradient, it not only maintains the characteristics of the original set, but also reduces the noise impact in the calculation.

Similarly, the nature gradient of $\mathcal{L}'$ with respect to the node influence $\lambda_p$ is given in Eq.(8),

$$\partial\lambda_p^t = -\frac{\lambda_p^{t-1}}{\sigma_1^2} + \sum\nolimits_{(p,q)\in links(p)\cup nonlinks(p)}\left(L_{pq} - r_{pq}s_{pq}\right), \tag{8}$$

where we designate $r_{pq}$ as

$$r_{pq} \equiv \frac{\exp\left\{\exp\left\{\lambda_p + \sigma_\kappa^2/2\right\}\exp\left\{\lambda_q + \sigma_\kappa^2/2\right\}\right\}}{1 + \exp\left\{\lambda_p + \sigma_\kappa^2/2\right\}\exp\left\{\lambda_q + \sigma_\kappa^2/2\right\}s_{pq}}. \tag{9}$$

Further, the community strength parameter $\mu_c$ over $\mathcal{L}'$ can be inferred with Eq.(10).

$$\begin{aligned}\partial\mu_c^t = \frac{\mu_0 - \mu_c^{t-1}}{\sigma_0^2} + \frac{N}{2|S|}\sum\nolimits_{(p,q)\in links(S)\cup nonlinks(S)}\phi_{pq}^{cc}\\ \times\left(L_{pq} - r_{pq}\exp\left\{\mu_c + \sigma_\beta^2/2\right\}\right).\end{aligned} \tag{10}$$

As with the gradient of community membership, the unbiased estimates of Eq.(8) and Eq.(10) can also be obtained from the sub-sampling of the non-links. In order to get an unbiased estimate of $\mu_c$, it is necessary to scale the links and non-links according to the inverse probability of sub-sampling the nodes in Eq.(10). Because each node pair is shared by two nodes, for a mini-batch with $S$ nodes, the sum over the pairs of nodes is marked as $\frac{N}{2|S|}$. In Eq.(8), $(L_{pq} - r_{pq}s_{pq})$ is the residual for node pair($p$, $q$), while $(L_{pq} - r_{pq}\exp\{\mu_c + \sigma_\beta^2/2\})$ is the residual for the pair($p$, $q$) under a situation that both nodes $p$ and $q$ are assigned to the latent community $c$ in Eq.(10). In addition, we also note that the updates for global operations of nodes $p$ and $q$ and even for parameter $\boldsymbol{\mu}$ rely only on the diagonal elements of the variational matrix indicator $\phi_{pq}$.

The approximate step of a global update operation accompanied by a noise gradient can be expressed as

$$\gamma_p \leftarrow \gamma_p + \rho_p(t)\partial\gamma_p^t;\ \lambda_p \leftarrow \lambda_p + \rho_p(t)\partial\lambda_p^t;\ \boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \rho'(t)\partial\boldsymbol{\mu}^t. \tag{11}$$

where $\rho_p$ represents the separate learning rate for any node $p$, and only $\gamma$ and $\lambda$ are updated for each node that exists in the mini-batch $S$ in each iteration. For the global learning rate $\rho'$ with respect to the community strength $\boldsymbol{\mu}$, it needs to be updated during each iteration.

### D. LOCAL UPDATE OPERATION

Associated with the local update operation is parameter $\phi_{pq}$ in Eq.(4), which is a interaction variational parameter of dimension $C \times C$ for each pair of node in a sub-sampling network. It means which pair of communities are activated to determine the posterior approximation of links and non-links. According to the coordinate ascent algorithm, we can deduce Eq.(12) and Eq.(13),

$$\begin{aligned}\phi_{pq}^{cc} \propto \exp\left\{\mathbb{E}_h[\log\pi_{p,c}] + \mathbb{E}_h[\log\pi_{q,c}]\right\}\\ + L_{pq}\mu_c - r_{pq}\left(\exp\left\{\mu_c + \sigma_\beta^2/2\right\} - 1\right),\end{aligned} \tag{12}$$

$$\phi_{pq}^{ij} \propto \exp\left\{\mathbb{E}_h[\log\pi_{p,i}] + \mathbb{E}_h[\log\pi_{q,j}]\right\}, (i \neq j). \tag{13}$$

where $r_{pq}$ is defined in Eq.(9).

Based on the above derivation process, a complete link probability prediction algorithm with non-conjugated variational inference based on node influence can be described in Algorithm 2.

The parameter derivation process is actually the detection step of the community discovery algorithm based on node influence used in this paper. It can not only capture the node influence in the network, but also capture the characteristics of network homogeneity(combined with MMSB algorithm model). The scalable sub-sampling of the network not only ensures the detection accuracy, but also greatly improves the execution efficiency of the algorithm. The computational complexity of traditional coordinate ascent algorithm is $O(n^2)$ when it faces a network with $n$ nodes, which makes it impossible to extend to large scale networks. Then the algorithm proposed in this paper makes the overall algorithm complex is much lower than $O(n^2)$(the mini-batch $S \ll n$) by optimizing a low-bound variational inference object and combining the mini-batch sampling method.

## IV. RESULTS AND ANALYSIS

In this section, we present examples of applications to test the performance of the proposed algorithm model, including computer-generated networks and real-world networks,

**Algorithm 2** Non-Conjugate Variational Inference Based on Node Influence

```
Input: The influence of node degree $\lambda$, the $\gamma$ obtained by posterior community membership of stochastic variational inference and the community strength $\mu$
Output: Final community structure $C_f$ obtained by the posterior link probability prediction model with node influence
Step (1): Initialize the community memberships $\boldsymbol{\gamma} = (\gamma_n)_{n=1}^{N}$. For the node influence, initialize the $\boldsymbol{\lambda}$ as the logarithm of the normalized node degree, and then initialize the community strength $\boldsymbol{\mu}$ to zero.
Step (2): Sub-sampling a mini-batch $S$ of nodes pairs.
Step (3): Local update operation. For each node pair $(p, q)$, compute the parameter $\phi_{pq}$ using Eq.(12) and Eq.(13).
Step (4): Global update operation. For each node $p \in S$, updating the community memberships $\gamma_p$ and node influence $\lambda_p$ using stochastic nature gradient and stochastic gradient in Eq.(7) and Eq.(8), respectively. The same stochastic gradient is used to update the community strength $\boldsymbol{\mu}$ in Eq.(10).
Step (5): For each node $p \in S$, let $\rho_p(t) = (\tau_0 + t_p)^{-\zeta}(t_p \leftarrow t_p + 1)$ and $\rho'(t) = (\tau_0 + t)^{-\zeta}(t \leftarrow t + 1)$, where $\tau_0 \geq 0$ is the step length and $\zeta \in (0.5, 1]$ is the learning rate. For $\rho$, in order to guarantee fast convergence to a local optimum, so $\sum_t \rho(t) = \infty$ and $\sum_t \rho(t)^2 \leq \infty$.
Step (6): If there are nodes in the network that are not visited, repeat steps (2)-(5).
Step (7): According to the edge probability between nodes and node influence captured, the nodes are assigned to the appropriate community and the final community structure $C_f$ is obtained.
```

and the related results and phenomena are analyzed and discussed.

### A. COMPUTER-GENERATED NETWORKS

The computer-generated benchmark model generally contains a predetermined community structure, and our goal is to restore the underlying structure of the network in a more precise way. In the tests, we used the LFR model [35], [36] as the basis for the experiment compared to the more widely used stochastic block model [3], [37], because the node degree distribution of this model is closer to the real world network. They organize the network based on the different probabilities that appear between the edges of the nodes. The tests include using the LFR model to generate a number of benchmark networks, using our algorithms to analyze them, and then finding the predefined planted community structure.

In the stochastic variational inference of MMSB with node influence parameter $\hat{\boldsymbol{\kappa}}$, we use a point estimate of the posterior community memberships $\hat{\boldsymbol{\pi}}$ and the posterior community strength $\hat{\boldsymbol{\beta}}$ to predict the distribution of the link, which are calculated as the average of the variational posterior parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$, respectively. Thus, the prediction of the edge distribution for a pair of nodes in a sub-sampling test set can be approximated as Eq.(14).

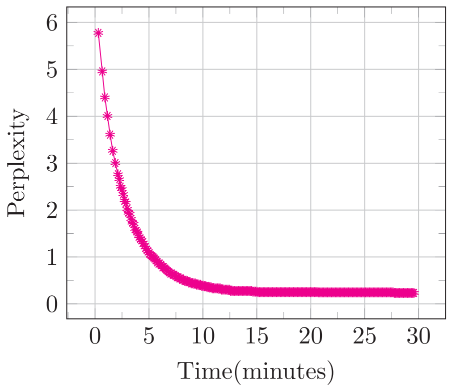


**FIGURE 1.** The perplexity computed by Eq.(15) on LFR network with 10, 000 nodes and 30 overlapping communities. It can be seen that it gradually converges with time. We used the mini-batch sampling method in the test and set 15% nodes who have memberships in three communities at the same time manually to increase the difficulty of overlapping community detection. The lower the score of the perplexity, the closer the detected model is to the ground-truth. In addition, the graph shows a power-law distribution characteristic, which indicates that the community divisions calculated by our algorithm based on node influence are consistent with the planted community structure with real-world network distribution in LFR.

$$\rho\left(L_{pq}|L_{observed}\right) \approx \sum_{z_{p\to q}} \sum_{z_{p\leftarrow q}} \rho\left(L_{pq}|z_{p\to q}, z_{p\leftarrow q}, \hat{\boldsymbol{\kappa}}, \hat{\boldsymbol{\beta}}\right)\rho \times \left(z_{p\to q}|\hat{\pi}_p\right) \rho\left(z_{p\to q}|\hat{\pi}_p\right). \quad (14)$$

It is clear that Eq.(14) is a legal approximation. Based on Eq.(14), the perplexity [7] is introduced according to the average predictive log likelihood of a test set of node pairs $Y$.

$$perplexity\,(Y) = \exp\left\{\frac{\sum_{p,q\in Y} \log \rho\left(L_{pq}|L_{observed}\right)}{Y}\right\}. \quad (15)$$

Perplexity is a measure of the adaptability of the model. The lower its value, the more consistent it is with the standard model. Fig.1 shows the change of the perplexity over time after applying the algorithm of this paper to LFR benchmark. This LFR test network contains 10, 000 nodes and 30 covered communities, and 15% of the nodes are shared by three communities at the same time. From Fig.1, we can see that the value of the perplexity decreases with time, which indicates that the number of communities detected with coverage nodes is basically the same as the number of communities in the standard model, which is approximate the posterior community model distribution. In addition, the graph illustrates a distribution trend of power-law [5], which can also prove that the results detected by our algorithm are consistent with the real world networks(LFR itself is a simulation of the real world network). Since the sub-sampling non-links are obtained by the way of characteristic scaling contraction,

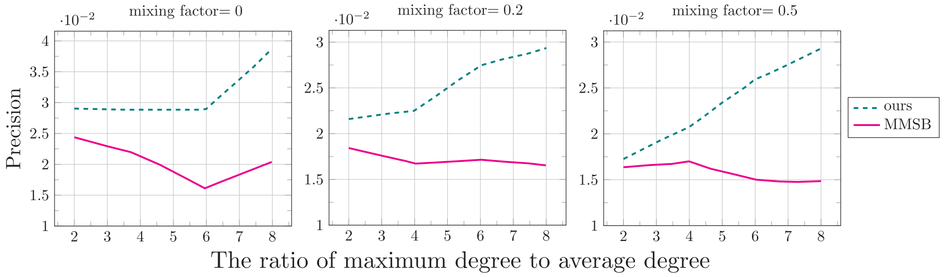


**FIGURE 2. The comparison results of MMSB [6] and the algorithm in this paper on LFR network with 10, 000 nodes and 20% community coverage. Each plot represents an experiment under five different networks with a right-skewness distribution. From the comparison of the precision, it can be seen that with the increase of the mixing factor(the increase of the factor indicates that the connections between dissimilar nodes are increasing, the community will become more and more obscure), the algorithm in this paper is better than the MMSB in correctly estimating the classification nodes. The abscissa describes the ratio of the maximum node degree to the average node degree. As it increases, the heterogeneity in the network becomes more and more serious, that is, a few nodes will have more connections. In this case, our algorithm not only ensures the accuracy of the community test results, but also captures this heterogeneity in node degrees when the community strength is correctly learned. Here, we set the average node degree to 15, the community size range of [50, 200] and the network power-law exponent is 2.**

it is an unbiased estimation method (see Section III-C), so the model in this paper can predict most of the non-links with high accuracy.

The precision-recall [38] is also a strategy for evaluating the accuracy of algorithms, which has been proved to be better than ROC ACC [39]. In a binary decision problem, a classifier labels samples as either positive or negative. According to different scoring standards, it mainly contains four categories: True positives(TP) indicate that the sample was correctly marked as positive. False positives(FP) correspond to negative sample was incorrectly marked as positive. True negatives(TN) refer to negative sample was correctly marked as negative. And the false negatives(FN) represent that the positive sample was incorrectly marked as negative. Thus, the precision and the recall can be computed by $\frac{TP}{TP+FP}$ and $\frac{TP}{TP+FN}$, respectively. Fig.2 shows how the accuracy of the two algorithms varies with the node degree. The horizontal axis indicates the ratio of the maximum degree to the average degree of the node, and as the ratio increases, the heterogeneity of the network becomes more and more serious, that is, the proportion of the node influence in this paper is increasing, and the community shows a serious overlapping phenomenon. Under this right-skewness degree distribution, our algorithm predicts better than MMSB. The main reason is that the right-skewness of the degree distribution may cause the community strength of MMSB to be overestimated or underestimate the link mode within the community. The algorithm in this paper, however, considering the weight of node influence within community, so it can capture the characteristics of the heterogeneity distribution of the node degrees, and thus learn the correct community strength.

We further tested the performance of the algorithm on a computer generated network with overlapping communities. In quantifying the similarities between communities, we employed a standardized means of normalized mutual information(NMI) [40], [41]. As a comparison of algorithm performance, several algorithm models with relatively novel concepts and high detection accuracy in community detection are selected in this paper, including Poisson model [27], information map model [42] and label belief propagation model [39]. Fig.3 is a comparison of algorithm performance on LFR network with overlapping communities. From a global point of view, the NMI of this paper is higher than the other three algorithms. However, as the mixing factor increases, the overall performance decreases. The performance of the Poisson model is very close the algorithm in this paper except for the other two models, which benefits from the nature of its principle derivation. From left to right in Fig.3, the panel can be seen as the process of increasing the heterogeneity of the network. In this process, since the degree of a few nodes grows very fast, a serious community overlapping phenomenon is formed, which further leads to a drastic reduction in the performance of most algorithms. When the mixing factor exceeds 0.5, the concept of the community is blurred.

### B. REAL-WORLD NETWORKS

Consistent with the aforementioned computer-generated network, we use similar evaluation criteria to test the performance of the algorithm. Table 1 compares the application of the MMSB and the algorithms of this paper to many real-world networks(the parameters are the same for all datasets). It can be clearly observed from the perplexity that the MMSB performs worse in predicting performance. Since the first four networks are small, they are fitted with a single community strength parameter. And for the remaining networks, the $C$ standard community strength parameter are used. The networks in Table 1 cover many areas that can be used for performance testing to make the results more representative. The score of the HEP-TH looks a bit different and much higher than other networks. The possible reason is that the degree distribution in the network has a very serious heterogeneity, and the number of connections between nodes

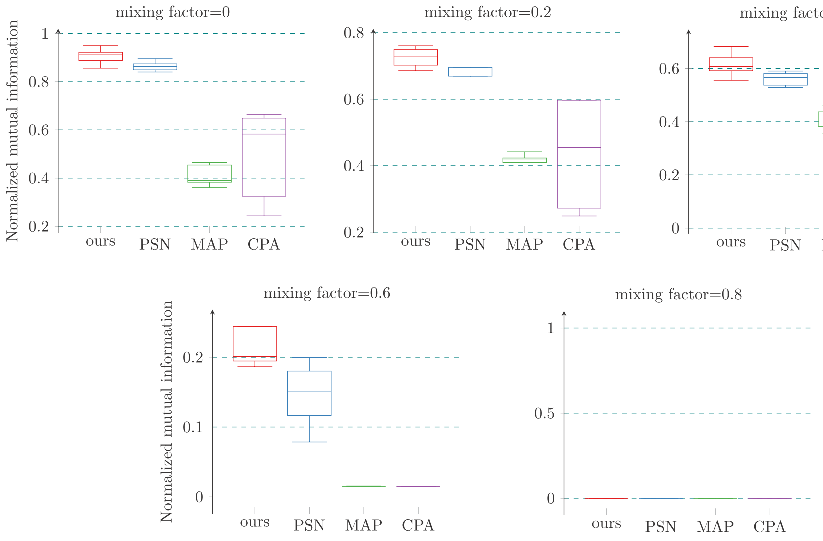


**FIGURE 3.** The comparative analysis of the NMI obtained by different algorithms, including the Poisson model(PSN) [27], information map algorithm(MAP) [42] and label belief propagation model(CPA) [39]. Each graph represents a repeat test under fifteen different LFR networks with 10, 000 nodes and 10% of the nodes are assigned to three overlapping communities. Obviously, as the mixing factor increases, the overall performance of the algorithm is degraded, however, the algorithm in this paper maintains a relatively high community detection accuracy. From a single graph, the other two algorithms perform poorly except for the Poisson model. When the mixing factor is beyond 0.5, the performance of all algorithms is drastically reduced, and more serious is the result of the failure of the algorithm [37]. From left to right, it can be seen as a process of increasing noise in the network, which can affect the performance of community detection algorithms. In other words, under the network conditions with severe heterogeneity, whether the underlying community structure can be correctly captured can be used as an objective criterion for judging the merits of the algorithm.

**TABLE 1.** Test results of perplexity on real-world datasets.

| Network | Ref. | Nodes | Edges | $P_{ours}$ | $P_{MMSB}$ | Type |
|---|---|---|---|---|---|---|
| Football | [43] | 115 | 613 | $1.87 \pm 0.09$ | $2.23 \pm 0.04$ | Social |
| Email | [44] | 1133 | 5451 | $2.71 \pm 0.05$ | $3.63 \pm 0.26$ | Social |
| Political Blogs | [43] | 1490 | 19090 | $3.16 \pm 0.12$ | $3.54 \pm 0.04$ | Internet |
| PPI-CP | [22] | 4626 | 14801 | $3.53 \pm 0.06$ | $8.29 \pm 0.21$ | Biological |
| HEP-TH | [45] | 8638 | 24806 | $14.16 \pm 0.12$ | $22.17 \pm 0.41$ | Collaboration |
| HEP-PH | [45] | 11204 | 117619 | $3.72 \pm 0.14$ | $3.51 \pm 0.27$ | Collaboration |
| Collaboration | [46] | 27519 | 116181 | $10.16 \pm 0.47$ | $16.27 \pm 0.15$ | Collaboration |
| Brightkite | [45] | 58228 | 772933 | $10.52 \pm 0.17$ | $48.19 \pm 0.58$ | Social |

Nodes and Edges represent the total number of nodes and edges of the network, respectively. $P$ is the average perplexity over the sampled mini-batch $S$ of node pairs in the test set. From Table 1, it can be seen clearly that the posterior estimation of node influence based on stochastic variational inference in this paper outperforms the MMSB in predicting performance. The first four networks were fitted with a single community strength parameter, while the rest were fitted with the community strength parameters $C$.

is significantly different, resulting in extreme overlapping phenomenon. We also calculated precision and recall for some networks in Table 1, and the results are shown in Fig.4. During the experiment, we generated the top $n$ node pairs for each node and sorted them according to the probability that a link between them. Then, we picked out the number of the top $m$ suggested node pairs of each node from the $n$ node pairs to calculate precision and recall, where $m$ takes a range from 10 to 100. As can be seen from Fig.4, with the increase in the number of nodes(the panel is arranged from left to right according to the increase in the number of nodes), the gap between precision and recall corresponding to these two algorithms is also increasing. In particular, the MMSB does not consider the node influence in the calculation, so it can only rely on the node memberships and the community strengths to predict the link, and its performance is inevitably lower than the algorithm of this paper. In addition, as $C$ is fixed, with the increase in the number of nodes $N$, the communities are likely to contain more nodes, which further increases the link prediction difficulty of MMSB, which can also be found in Table 1.

Finally, let us introduce a more interesting example. The arXiv network consists of the scientific literature and references between them [47]. Until August 2011, the total number of articles uploaded on this website reached 694, 000. After filtering, we retained 570, 000 physics-related articles

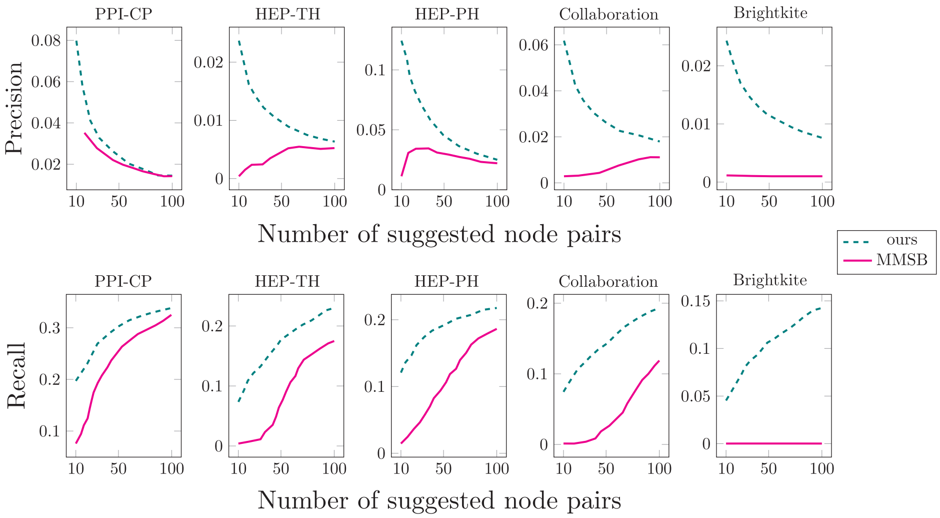


**FIGURE 4.** The results of applying the algorithm in this paper to real-world networks. The panel corresponds to several datasets in Table 1 from left to right. Overall, the model in this paper outperforms the MMSB algorithm in predicting precision and recall. Although both models use variational inference to fit, because the model in this paper considers node influence in the community division, it avoids overestimation or underestimation when predicting links. By sampling the number of the top *m* suggested node pairs of each node, We can make quantitative judgments about the correctness of the node classification. In this case, the precision evaluates the *m* suggestion that appears in the test set, while recall captures the node pairs in the top *m* suggestion that appear in the test set.

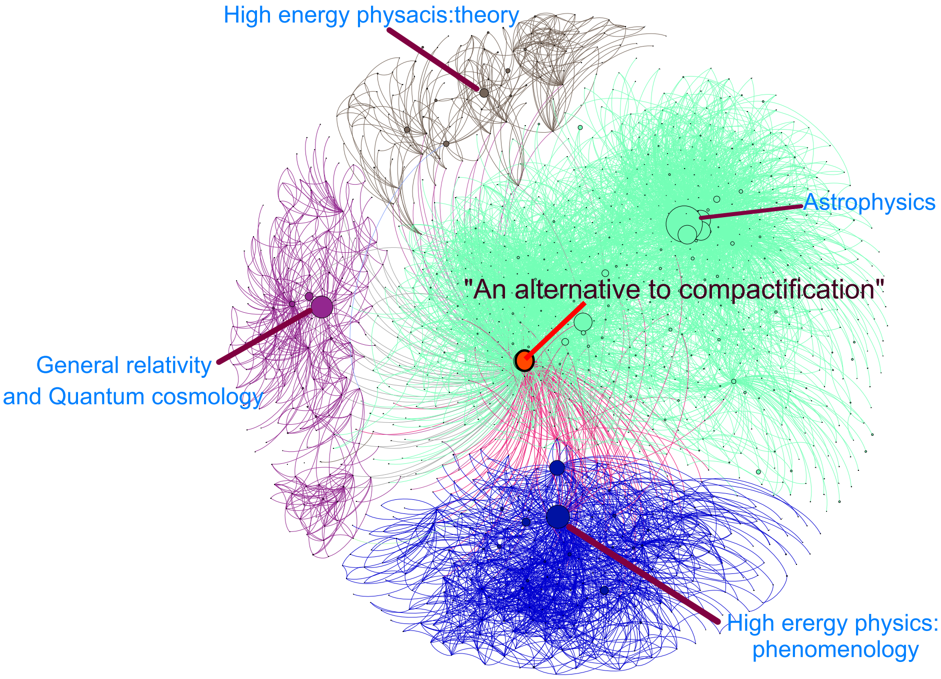


**FIGURE 5.** The community structure detected on the arXiv network [47] with 570, 000 articles using the method of this paper. The nodes in the network represent the articles, and the edges imply the citation relationships between them, which are colored according to the posterior estimates of the detected communities. This figure demonstrates the four top-level community structures and is marked with their specific domain. We use the posterior node influence to mark the size of the nodes and color them according to its community assignment. The node showing "An alternative to compactification" [48] is a node with highly overlapping features and can be viewed as a bridging node between many communities.

as the main research object. Fig.5 is a community partitioning result detected by the posterior link estimation method in this paper. Since the number of detected communities is close to 200, and many communities contain few nodes, which does not affect the accuracy and overlapping of community partitioning, only the top four communities are shown here.

According to our model, each node $i$ has a community membership $\pi_i$, and each link associated with the node pair $(i, j)$ is assigned to one of the $C$ communities. In addition, the links are colored according to the approximate posterior probability $\rho(z_{i\rightarrow j}, z_{i\leftarrow j}|\boldsymbol{L})$, which also implies the relationship between references and references between articles. The highly cited article in the middle of the figure is "An alternative to compactification" [48], as the author said it was originally a purely theoretical article published in 1999. However, with the development of society, it has been labeled with a technical label, so that it is referenced by more other research fields, which also makes it have the characteristic of heterogeneous node degree distribution to some extent.

For convenience, we examined the additional label attributes(the category to which the article belongs) of the nodes to name the communities that were detected, which are "High energy physics: Phenomenology", "High energy physics: Theory", "Quantum cosmology" and "General relativity". The link color corresponds to the relative community associated with the links, and the size of the node is used to distinguish heterogeneous nodes in the network(the degree of such nodes is very large and has varying degrees of overlapping). It should be pointed out here that the citation itself does not reflect the role of the article in the citation. Other operations on the network, such as calculating the overlapping of nodes, the centrality of the edges or nodes that act as mediators are all based on the results of community detection by the algorithm of this paper.

In the analysis of the network, articles belonging to multiple sub-domains can be found by our method. These nodes build their own membership between different communities and get a high a posterior bridge score, which is a strategy for judging the strength of bridging nodes [12]. Take the article "Cosmological constant-the weight of the vacuum"as an example [49], it contains 1117 references, mainly introduces cosmology related knowledge. By calculating, we found that the community members associated with it were concentrated in two communities, which meant a very low posterior bridge score. Nevertheless, the two communities it is associated with are labeled "General relativity and Quantum cosmology "and "Astrophysics"at the same time, which shows that it really provide a bridge between the two larger communities, and this is a very meaningful activity. By exploring potential community structures, articles with interdisciplinary influences can be separated from their respective specific areas, making the hierarchy more clear.

## V. CONCLUSION

Revealing underlying overlapping communities in large-scale networks can help us better explore, interpret and predict the structure and properties of the network. However, a large number of existing methods assume that nodes belong to only one single community, without considering the "multiple identities"that a node might have. In this paper, we interpret this "multiple identity"as the influence of a node, which is similar to the characteristics of preferential connection [5], [20], so that a few nodes have a large degree distribution, and most nodes have only a few links. In order to simulate the behavior of this kind of node effectively, we derived a probabilistic prediction model to capture the nodes influence that exist in the community detection process.

Our approach is based on a Bayesian network model called Mixed-Membership Stochastic Blockmodel(MMSB) [6], which allows nodes to belong to multiple different communities at the same time. Compared to models that only consider a single community, MMSB can better simulate most networks in the real world, but does not explain the node influence. In view of this, the influence model parameters of nodes are added to improve the algorithm, and summarized the stochastic variational inference algorithm based on the node influence. The advantage of doing this is not only can make full use of the probability prediction model of MMSB, but also deeply excavate the community overlapping phenomenon in the detection process, uncovering the nodes with the bridge function among the communities, and maximizing the information income. In addition, we developed a non-conjugated stochastic variational inference to derive a link probability prediction model with node influence. The key strategy is to use the mean-domain variational family with variable distribution to approximate the posterior community strengthen and node influence distribution in the link probability prediction model. By sub-sampling the test network, it is convenient to make a posterior inference to the parameters of community strength and node influence, and then re-estimate the potential community structure.

We applied the posterior link prediction model with node influence to the computer-generated and real-world networks, and made performance evaluation and analysis. Experimental results show that the algorithm in this paper is better than the MMSB in overall performance, especially when the heterogeneity of network is very serious. Through the performance comparison test of several algorithms on the LFR network, when the network contains a clear community structure, the algorithm of this paper can achieve higher community identification accuracy than other algorithms. However, when the network structure mixed factor reaches a certain critical value, all the detection performance of the algorithm will be affected, and even lead to failure. For real-world networks that conform to power-law distribution, as the algorithm in this paper considers the influence of nodes, the accuracy of community detection is better than that of MMSB(the node degree that exhibits a right-skewness distribution may lead MMSB to underestimate or overestimate the link behavior within the community). In addition, the method of this paper is applied to the arXiv network [47] with $570,000$ articles. It can be clearly seen from the experimental results that it can not only detect the hidden community structure with overlapping features correctly, but also captures nodes with greater influence in the network. The test also further demonstrates the importance of node influence in predicting the accuracy of the degree of node that exhibits a power-law distribution. In general, this paper combines the node influence with probability model of link prediction, providing a new idea for using statistical models to explore real-world networks.

## REFERENCES

[1] X. Liu, H.-M. Cheng, and Z.-Y. Zhang, "Evaluation of community detection methods," *IEEE Trans. Knowl. Data Eng.*, to be published.

[2] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Phys. Rep.*, vol. 659, pp. 1–44, Nov. 2016.

[3] V. Lyzinski, M. Tang, A. Athreya, Y. Park, and C. E. Priebe, "Community detection and classification in hierarchical stochastic blockmodels," *IEEE Trans. Netw. Sci. Eng.*, vol. 4, no. 1, pp. 13–26, Jan./Mar. 2017.

[4] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 23, pp. 8577–8582, 2006.

[5] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[6] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *J. Mach. Learn. Res.*, vol. 9, pp. 1981–2014, Sep. 2008.

[7] P. K. Gopalan and D. M. Blei, "Efficient discovery of overlapping communities in massive networks," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 36, pp. 14534–14539, 2013.

[8] A. Ozcan and S. G. Oguducu, "Link prediction in evolving heterogeneous networks using the NARX neural networks," *Knowl. Inf. Syst.*, vol. 55, no. 2, pp. 333–360, 2018.

[9] A. Ozcan and S. G. Oguducu, "Multivariate time series link prediction for evolving heterogeneous network," *Int. J. Inf. Technol. Decis. Making*, vol. 18, no. 1, pp. 241–286, 2019.

[10] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[11] M. E. J. Newman, "Communities, modules and large-scale structure in networks," *Nature Phys.*, vol. 8, no. 1, pp. 25–31, 2012.

[12] H. G. Russell and B. A. Graybeal, "Ultra-high performance concrete: A state-of-the-art report for the bridge community," U.S. Dept. Transp., Washington, DC, USA, Tech. Rep. FHWA-HRT-13-060, 2013.

[13] Y. Lei and P. S. Yu, "Cloud service community detection for real-world service networks based on parallel graph computing," *IEEE Access*, vol. 7, pp. 131355–131362, 2019.

[14] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, pp. 1303–1347, May 2013.

[15] M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett.*, vol. 89, no. 20, Oct. 2002, Art. no. 208701.

[16] G. Ren and X. Wang, "Epidemic spreading in time-varying community networks," *Chaos Interdiscipl. J. Nonlinear Sci.*, vol. 24, no. 2, 2014, Art. no. 023116.

[17] X. Que, F. Checconi, F. Petrini, and J. A. Gunnels, "Scalable community detection with the Louvain algorithm," in *Proc. IEEE Int. Parallel Distrib. Process. Symp.*, May 2015, pp. 28–37.

[18] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 1, pp. 36–41, 2007.

[19] Y. Cui and X. Wang, "Detecting one-mode communities in bipartite networks by bipartite clustering triangular," *Phys. A, Stat. Mech. Appl.*, vol. 457, pp. 307–315, Sep. 2016.

[20] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 2, 2001, Art. no. 025102.

[21] C. Wang and D. M. Blei, "Variational inference in nonconjugate models," *J. Mach. Learn. Res.*, vol. 14, pp. 1005–1031, Jan. 2013.

[22] V. Colizza, A. Flammini, A. Maritan, and A. Vespignani, "Characterization and modeling of protein-protein interaction networks," *Phys. A, Stat. Mech. Appl.*, vol. 352, no. 1, pp. 1–27, 2005.

[23] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[24] Z. Zhang and Z. Wang, "Mining overlapping and hierarchical communities in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 421, pp. 25–33, Mar. 2015.

[25] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," *PLoS ONE*, vol. 6, no. 4, 2011, Art. no. e18961.

[26] J. Eustace, X. Wang, and Y. Cui, "Overlapping community detection using neighborhood ratio matrix," *Phys. A, Stat. Mech. Appl.*, vol. 421, pp. 510–521, Mar. 2015.

[27] B. Ball, B. Karrer, and M. E. Newman, "Efficient and principled method for detecting communities in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 84, no. 3, 2011, Art. no. 036103.

[28] Y. Cui and X. Wang, "Uncovering overlapping community structures by the key bi-community and intimate degree in bipartite networks," *Phys. A, Stat. Mech. Appl.*, vol. 407, pp. 7–14, Aug. 2014.

[29] D. Jin, H. Wang, J. Dang, D. He, and W. Zhang, "Detect overlapping communities via ranking node popularities," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 172–178.

[30] J. Li, X. Wang, and J. Eustace, "Detecting overlapping communities by seed community in weighted complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 392, no. 23, pp. 6125–6134, Dec. 2013.

[31] C. Zhe, A. Sun, and X. Xiao, "Community detection on large complex attribute network," in *Proc. 25th SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 2041–2049.

[32] J. Xi, W. Zhan, and Z. Wang, "Hierarchical community detection algorithm based on node similarity," *Int. J. Database Theory Appl.*, vol. 9, no. 6, pp. 209–218, 2016.

[33] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network analysis," *J. Amer. Stat. Assoc.*, vol. 97, no. 460, pp. 1090–1098, 2002.

[34] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, Nov. 1999.

[35] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 78, no. 4, 2008, Art. no. 046110.

[36] A. Arenas and A. Díaz-Guilera, "Synchronization and modularity in complex networks," *Eur. Phys. J. Special Topics*, vol. 143, no. 1, pp. 19–25, 2007.

[37] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 83, Jan. 2011, Art. no. 016107.

[38] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233–240.

[39] T. Martin, B. Ball, and M. E. J. Newman, "Structural inference for uncertain networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 93, no. 1, 2016, Art. no. 012306.

[40] M. Cordeiro, R. P. Sarmento, and J. Gama, "Dynamic community detection in evolving networks using locality modularity optimization," *Social Netw. Anal. Mining*, vol. 6, no. 1, p. 15, 2016.

[41] X. Zhang and M. E. J. Newman, "Multiway spectral community detection in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 92, no. 5, 2015, Art. no. 052808.

[42] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 2, pp. 1118–1123, 2008.

[43] M. E. J. Newman, "Spectral methods for community detection and graph partitioning," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 88, no. 4, 2013, Art. no. 042822.

[44] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, "Modularity from fluctuations in random graphs and complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 2, 2004, Art. no. 025101.

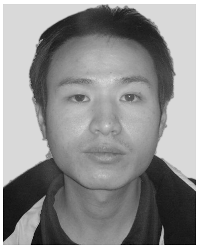
[45] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Math.*, vol. 6, no. 1, pp. 29–123, 2009.

[46] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 2, pp. 404–409, 2001.

[47] P. Ginsparg, "ArXiv at 20," *Nature*, vol. 476, no. 7359, pp. 145–147, 2011.

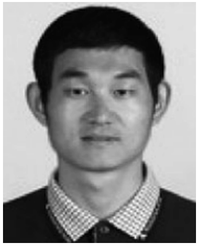
[48] L. Randall and R. Sundrum, "An alternative to compactification," *Phys. Rev. Lett.*, vol. 83, no. 23, p. 4690, 1999.

[49] T. Padmanabhan, "Cosmological constant—The weight of the vacuum," *Phys. Rep.*, vol. 380, nos. 5–6, pp. 235–320, 2003.



**QIANG ZHOU** received the B.S. and M.S. degrees in the major of computer science and technology from Guizhou University, Guiyang, China, in 2012. He is currently pursuing the Ph.D. degree in computer software and theory with the University of Electronic Science and Technology, Chengdu, China.

Since 2014, he has been with the Institute of Fundamental and Frontier Science, University of Electronic Science and Technology. His main interests include social networking, big data analytics, complex network community detection, and recommendation algorithm research. His main awards and honors include National Scholarships, National Motivational Scholarships, and provincial outstanding graduates.



**SHIMIN CAI** received the B.S. degree in electrical engineering from the Hefei University of Technology, in 2004, and the Ph.D. degree in circuit and systems from the University of Science and Technology of China, in 2009.

He currently serves as an Associate Professor with the University of Electronic Science and Technology of China. He is interested in complex network theory and its application for mining and modeling of real large-scale networked systems, time series analysis, and personalized recommendation systems. At present, he has published nearly 100 high-level academic articles, including nearly 70 SCI articles, nearly 400 SCI quotations, and completed more than ten national projects supported by the National Natural Science Foundation of China and the Military Commission for Science and Technology.

**YICHENG ZHANG** received the B.S. degree from the Physics Department, University of Science and Technology of China, Hefei, China, in 1980, the M.S. degree in physics from the Graz University of Technology, Graz, Austria, in 1981, and the Ph.D. degree in physics from the University of Rome La Sapienza, Roma, Italy, in 1984.

He is currently a Full Professor with the University of Fribourg, Switzerland, a Distinguished Professor with the University of Electronic Science and Technology of China. So far, he has published more than 130 SCI articles, in which 35 published in Physics Reports, PNAS and Physical Review Letters, and got more than 24297 times cites according to Google Scholar. His main research interests include complex networks, cloud computing, and big data processing.