Learning Nonparametric Relational Models by Conjugately Incorporating Node Information in a Network

Xuhui Fan, Richard Yi Da Xu, Longbing Cao, *Senior Member, IEEE*, and Yin Song

***Abstract*—Relational model learning is useful for numerous practical applications. Many algorithms have been proposed in recent years to tackle this important yet challenging problem. Existing algorithms utilize only binary directional link data to recover hidden network structures. However, there exists far richer and more meaningful information in other parts of a network which one can (and should) exploit. The attributes associated with each node, for instance, contain crucial information to help practitioners understand the underlying relationships in a network. For this reason, in this paper, we propose two models and their solutions, namely the node-information involved mixed-membership model and the node-information involved latent-feature model, in an effort to systematically incorporate additional node information. To effectively achieve this aim, node information is used to generate individual sticks of a stick-breaking process. In this way, not only can we avoid the need to prespecify the number of communities beforehand, the algorithm also encourages that nodes exhibiting similar information have a higher chance of assigning the same community membership. Substantial efforts have been made toward achieving the appropriateness and efficiency of these models, including the use of conjugate priors. We evaluate our framework and its inference algorithms using real-world data sets, which show the generality and effectiveness of our models in capturing implicit network structures.**

***Index Terms*—Bayesian nonparametrics, convergence rate, node information, relational model.**

I. Introduction

COMMUNITY detection and network partitioning is an emergent topic in various areas including social-media recommendation [1], customer partitioning [2], [3], social network analysis [4], [5], and partitioning protein interaction network tasks [6]–[10]. Many models have been proposed in recent years to address this problem by using link data (e.g., a person's view toward others). Some examples include the stochastic blockmodel [11] and in the case of infinite communities, the infinite relational model (IRM) [12], both aiming at partitioning a network of nodes into different groups based on their pairwise, directional binary observations. In most existing approaches, the "internodes" link data are lone contributors toward the understanding of the insights of social structures. That is to say, in classical relational models, we are given the nodes' interaction data (usually an $n \times n$ binary matrix, n is the number of nodes in a network, 0 for no interaction and 1 for interaction, and the task is then to infer the nodes's group belongings).

On the other hand, the "intranodes" information is a vital source of additional information to complement the link information. Let us take the Lazega data set [13] (detailed in Section VI), which is a social network within a lawyer firm, as an example. The node (i.e., attorney) here contains information such as ages, offices (Boston, Hartford, or Providence), and law schools (Harvard, Yale, Ucon, or other). Naturally, the attorneys with similar information (e.g., the same office) tend to have relationships and/or belong to same community. An appropriate modeling which also incorporates this information would provide us with a much more complete understanding of the network.

While some recent efforts have been directed to incorporate the node information, they all face several shortcomings mainly in terms of appropriateness and efficiency. For example, in terms of appropriateness, in latent feature relational model (LFRM) [14], although the direct and linear combination of node information and the latent feature have experimentally demonstrated its effectiveness in link prediction, it is hard to interpret the recovered features and their related social structure (also stated in [15]). In terms of efficiency, taking nonparametric metadata dependent relational (NMDR) model [15] as example, the logistic-normal transform was employed to integrate the node information into each node's mixed-membership distribution. However, this integration complicates the original structure and results in nonconjugacy during the inference. Especially, as we should notice, all of the previous work incorporating these node-information is utilized via an unconjugate way.

Two major branches of relational models have been developed in the last few years, namely the mixed-membership stochastic blockmodel (MMSB) [16] and the LFRM [14], where community memberships are modeled as mixed memberships and latent features, respectively. In order to

Manuscript received February 3, 2015; revised October 1, 2015; accepted January 10, 2016. Date of publication February 11, 2016; date of current version February 14, 2017. This paper was recommended by Associate Editor P. De Wilde.

X. Fan and R. Y. D. Xu are with the Faculty of Engineering and Information Technology, University of Technology Sydney, Chippendale, NSW 2008, Australia (e-mail: xhfan.ml@gmail.com; yida.xu@uts.edu.au).

L. Cao is with the Advanced Analytics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Chippendale, NSW 2008, Australia (e-mail: longbing.cao@uts.edu.au).

Y. Song is with Brandscreen, Sydney, NSW 2061, Australia (e-mail: yinsong1986@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2016.2521376

demonstrate the generality of our method, we have individually adapted our method to both of these frameworks and have produced two distinct models, which is the central theme of this paper: the node-information involved mixed-membership model (niMM) and the node-information involved latent-feature model (niLF). In both cases, methods similar to the stick-breaking process [17], [18] are proposed to model the unknown number of communities. In particular, niMM successfully obtains the conjugate property (in Bayesian probability theory, conjugate property refers to the case that the posterior distributions are in the same family as the prior probability distributions) during the Markov chain Monte Carlo (MCMC) inference procedure. This property enables us to efficiently and effectively inference the model, since the required posterior likelihoods can be presented as an analytical form. As discussed later, through these efforts, the existing models (MMSB and LFRM) can be seen as special cases of our proposed models. In this way, our models capture much richer information embedded in a network; hence, they result in better performance in modeling the communities' memberships as illustrated in the experiments.

In summary, our contributions can be stated as follows.

1) We have naturally extended the existing benchmark models (i.e., MMSB and LFRM) to incorporate the nodes' information. The experimental results seem quite promising while the nodes' information is closely related to the link data.
2) Our extension to MMSB has retrieved the conjugate property during the MCMC inference, which mixes much faster in the Markov Chain than the previous approaches. Also, we find that in the experiments, our method converges much earlier than the previous one.
3) Our model is under the Bayesian nonparametrics setting (achieved through the methods similar to the stick-breaking constructions), which can deal with the problem of an unknown number of communities.

The rest of this paper is organized as follows. We start with an introduction to the notations and preliminary knowledge. Then, we describe both our niMM and niLF models in details, as well as the detail inference procedure and a "collapsed" inference discussion of niMM. We also include the model's computational complexity analysis in the same section. In Section VI, we compare our methods with the previous work to validate the models performances. The conclusions and future work are given in the last section.

II. Notations and Preliminary Knowledge

A. Notations

We first give all the notations in this paper in Table I.

TABLE I
NOTATIONS FOR OUR NIMM AND NILF MODELS

n	number of the nodes in a network
K	number of discovered communities
F	number of attributes in node information
ϕ	an $n \times F$ binary matrix, $\phi_{if} = 1$ denotes the i^{th} data occupies the f^{th} attribute
η	an $F \times K$ positive matrix, η_{fk} indicates the importance of f^{th} attribute to k^{th} roles.
ψ_i	stick-breaking weights to constitute π_i
π_i	stick-breaking representation for node i, i.e., the communities' memberships for i
π_{ik}	the significance of community k for node i
s_{ij}, r_{ij}	membership indicators of e_{ij} in MMSB
z_i	latent feature vector of node i in LFRM
e_{ij}	directional, binary link data
B	asymmetric, role-compatibility matrix, B_{kl} indicates compatibility of communities k, l
m_{kl}	part of $m_{k,l}$ where the corresponding $e_{ij} = 1$, i.e. $m_{kl} = \sum_{s_{ij}=k, r_{ij}=l} e_{ij}$
N_{ik}	number of times that a node i has participated in community k (either sending or receiving), i.e. $N_{ik} = \#\{j : s_{ij} = k\} + \#\{j : r_{ji} = k\}$

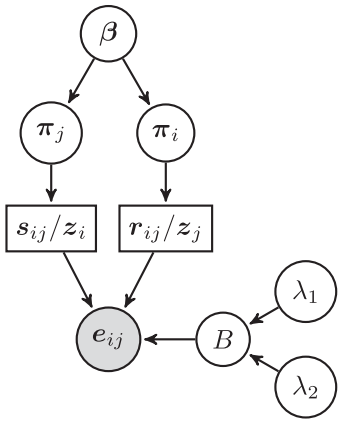


Fig. 1. Graphical model for the MMSB and the LFRM. Here, s_{ij} and r_{ij} in the rectangular nodes represent the latent variable in MMSB, and z_i and z_j are in the LFRM context.

B. Mixed-Membership Stochastic Blockmodel

A graphical model or probabilistic graphical model is a probabilistic model for which a graph expresses the conditional dependence structure between random variables. Among its various presentation, the MMSB [16] aims at modeling each node's individual mixed-membership distribution. In MMSB, each link data e_{ij} corresponds to two membership indicators: 1) s_{ij} from the sender i and 2) r_{ij} to the receiver j. (w.l.o.g., we assume $s_{ij} = k$ and $r_{ij} = l$). The link data's value is determined by the compatibility of two corresponding communities k and l. Fig. 1 shows the graphical model, which expresses the conditional dependence among these variables and the detailed generative process can be described as follows.

1) $\forall\{k, l\} \in \mathcal{N} > 0$, draw the communities' compatibility values $B_{k,l} \sim \text{Beta}(\lambda_1, \lambda_2)$.
2) $\forall i \in \{1, \ldots, n\}$, draw node i's mixed-membership distribution $\pi_i \sim \text{Dirichlet}(\beta)$.
3) $\forall\{i, j\} \in \{1, \ldots, n\}^2$, for link data e_{ij}.
   a) Sender's membership indicator $s_{ij} \sim \text{Multi}(\pi_i)$.
   b) Receiver's membership indicator $r_{ij} \sim \text{Multi}(\pi_j)$.
   c) The link data $e_{ij} \sim \text{Bernoulli}(B_{s_{ij}, r_{ij}})$.

It should be noted that each $\boldsymbol{\pi}_i$ is responsible for generating both the sender's labels $\{s_{ij}\}_{j=1}^n$ from node i and the receiver's labels $\{r_{ji}\}_{j=1}^n$ to node i.

C. Latent Feature Relational Learning

The LFRM [14] provides an alternative modeling method to infer the nodes' latent features. Compared to MMSB, the main

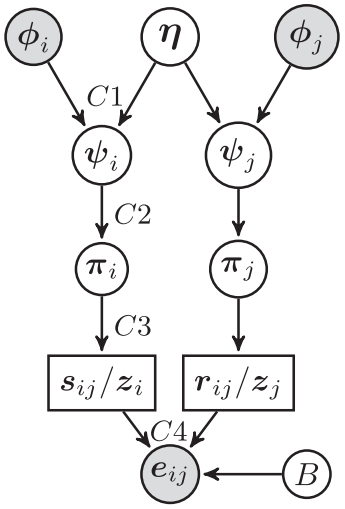


Fig. 2. Generative model for the niMM and niLF models.

difference of LFRM is that it assumes each node has links with others under one single binary vector, and this vector indicates the communities to which it belongs.

As shown in graphical model shown in Fig. 1, the detailed generative process can be described as follows.

1) $\forall\{k, l\} \in \mathcal{N} > 0$, draw the communities' compatibility values $B_{k,l} \sim \text{Normal}(0, 1)$.
2) $\forall i \in \{1, \ldots, n\}$, draw node i's stick-breaking representation $\pi_i \sim \text{Dirichlet}(\beta)$.
3) $\forall i \in \{1, \ldots, n\}$, draw node i's binary latent feature vector $z_i \sim \text{Bernoulli}(\pi_i)$.
4) $\forall\{i, j\} \in \{1, \ldots, n\}^2$, for link data e_{ij}.
   a) The link data $e_{ij} \sim \text{Bernoulli}(1/1 + \exp(-z_i B z_j))$.

D. Bayesian Nonparametrics

Bayesian nonparametrics is a popular tool to fit a single model that can adapt its complexity to the data, that is to say, the model complexity would grow as more data are observed. Upon this powerful technique, the hierarchical Dirichlet process (HDP) [19] is further developed to model the groups' correlation in the Bayesian nonparametrics contexture. For instance, the Bayesian nonparametrics methods can be used to infer the number of topics in one document in latent Dirichlet allocation of topic models. The HDP is used to model the topics' behavior in different documents and the topics are shared among these documents.

In our "dynamic" setting, we use the Bayesian nonparametric method allow the communities' numbers to vary across time and further use HDP to model the mixed-membership distribution $\{\pi_i\}_{i=1}^n$, where $\forall i \in \{1, \ldots, n\}, \pi_i \sim \text{DP}(\alpha, \boldsymbol{\beta})$ and $\boldsymbol{\beta}$ is generated from a stick-breaking construction $\boldsymbol{\beta} = \sum_{k=1}^{\infty} \boldsymbol{\beta}_k \delta_k$, $\boldsymbol{\beta}_k = \beta'_k \prod_{l=1}^{k-1}(1 - \beta'_l)$, $\beta'_l \sim \text{beta}(1, \gamma)$) [20].

III. Generative Model

Fig. 2 depicts the graphical models of all the variables used in this paper. As previously discussed, node information is incorporated into both branches of the relational models: iMMM and LFRM. Therefore, we illustrate both in the same figure, as most nodes are common to both graphical models.

A. Node-Information Involved Mixed-Membership Model

The generative process for the niMM model is defined as follows (w.l.o.g. $\forall i, j = 1, \ldots, n, k \in N^+$).

C1: $\psi_{ik} \sim \text{Beta}(1, \prod_f \eta_{fk}^{\phi_{if}})$.
C2: $\pi_{ik} = \psi_{ik} \prod_{l=1}^{k-1}(1 - \psi_{il})$.
C3: $s_{ij} \sim \text{Multi}(\pi_i)$, $r_{ij} \sim \text{Multi}(\pi_j)$.
C4: $e_{ij} \sim \text{Bernoulli}(B_{s_{ij} r_{ij}})$.

Here, $C1$ and $C2$ constitute the stick-breaking representation for our mixed-membership distribution π_i, which is similar to that of the Dirichlet process. While the Dirichlet process employs one single γ parameter to finish its stick-breaking construction, our representation uses different values for each component. The values are computed through exponential form $\eta_{fk}^{\phi_{if}}$ to further facilitate the conjugate design. $C3$ and $C4$ correspond to the membership indicator and link data generation, which follows the procedure as in Section II-B.

Correspondingly, the graphical model is shown in Fig. 2. Observational variables are colored in gray. $\{\phi_i\}_{i=1}^n$ is the nodes' attributes information (each node's feature value is transformed into one binary vector). For example, $\phi_i = [1, 0, 0, 1, 1]$ presents node i occupies first, fourth, and fifth features. $\{e_{ij}\}_{i,j}$ stands for the observational link data, where $e_{ij} = 1$ represents node i has an interaction with node j and $e_{ij} = 0$ represents node i does not interact with node j. ψ_i refers node i's weight values in constructing the stick-breaking weights π_i. s_{ij} and r_{ij} in the rectangular nodes represent the latent label and $B_{k,l}$ refers to the compatibility value between community k and l.

On $C1$, we replace the fixed γ parameter in the stick-breaking process (see in Section II-D) with $\prod_f \eta_{fk}^{\phi_{if}}$, where the positive, importance indicator η_{fk} is given a vague gamma prior $\eta_{fk} \sim \text{Gamma}(\alpha_\eta, \beta_\eta)$. Our method can successfully integrate the node information into the node's mixed-membership distribution and enjoy the conjugate property during the inference procedure. On the other hand, the previous approach [15], [21] uses the logistic normal distribution [with the mean value being the linear sum (i.e., $\sum_f \phi_{if} \eta_{fk}$)] to construct a stick-breaking weight ψ_{ik}, which makes the inference inefficient (i.e., slow mixing rate during the MCMC sampling).

We again use the attribute age (which will be "binarized" before use) in the Lazega data set to further explain the importance indicator η_{fk} used in $C1$. w.l.o.g., we let f_0th column of ϕ matrix denotes the age attribute, $\phi_{if_0} = 1$ implies that node i has age > 40 (in our experimental setting), and 0 otherwise. From $C1$, one can easily see that when $\eta_{f_0 k} \ll 1$, age would largely increase its impact on the kth community. Likewise, $\eta_{f_0 k} \gg 1$ reduces the influence of the age attribute on the kth community. $\eta_{f_0 k} = 1$ means that age does not have an impact on the kth community at all. Also, $\phi_{if_0} = 0$ makes age of the node i neutral toward all other communities.

Both the importance indicator η_{fk} and stick-breaking weight ψ_{ik} can enjoy the conjugate property. More specifically, the distributions of η_{fk} and ψ_{ik} are

$$p(\eta_{fk}|\alpha_\eta, \beta_\eta) \propto \eta_{fk}^{\alpha_\eta - 1} e^{-\beta_\eta \eta_{fk}}$$
$$p(\psi_{ik}|\eta_{\cdot k}, \phi_i) \propto \left[\prod_f \eta_{fk}^{\phi_{if}}\right] \cdot (1-\psi_{ik})^{\prod_f \eta_{fk}^{\phi_{if}} - 1}. \quad (1)$$

Thus, the posterior distribution of η_{fk} becomes

$$p(\eta_{fk}|\alpha_\eta, \beta_\eta, \psi_{\cdot k}, \phi) \propto p(\eta_{fk}|\alpha_\eta, \beta_\eta) \prod_i p(\psi_{ik}|\phi_i, \eta_{\cdot k})$$
$$\propto \eta_{fk}^{\alpha_\eta + \sum_i \phi_{if} - 1} e^{-\left(\beta_\eta - \sum_i \phi_{if} \ln(1-\psi_{ik}) \prod_{F \neq f} \eta_{Fk}^{\phi_{iF}}\right)\eta_{fk}}$$
$$\Longrightarrow \eta_{fk} \sim \text{Gamma}\left(\alpha_\eta + \sum_i \phi_{if}, \beta_\eta - \sum_i \phi_{if} \ln(1-\psi_{ik}) \prod_{q \neq f} \eta_{qk}^{\phi_{iq}}\right). \quad (2)$$

The joint probability of $\{s_{ij}, r_{ji}\}_{j=1}^n$ becomes

$$p\left(\{s_{ij}\}_{j=1}^n, \{r_{ji}\}_{j=1}^n | \psi_{i\cdot}\right) \propto \prod_{k=1}^K \left[\psi_{ik}^{N_{ik}} (1-\psi_{ik})^{\sum_{l=k+1}^K N_{il}}\right] \quad (3)$$

here $N_{ik} = \#\{j : s_{ij} = k\} + \#\{j : r_{ji} = k\}$.

The posterior distribution of ψ_{ik} becomes

$$p\left(\psi_{ik}|\phi, \eta_{\cdot k}, \{s_{ij}, r_{ji}\}_{j=1}^n\right)$$
$$\propto \psi_{ik}^{N_{ik}} (1-\psi_{ik})^{\sum_{l=k+1}^K N_{il} + \prod_f \eta_{fk}^{\phi_{if}} - 1}$$
$$\Longrightarrow \psi_{ik} \sim \text{Beta}\left(N_{ik} + 1, \sum_{l=k+1}^K N_{il} + \prod_f \eta_{fk}^{\phi_{if}}\right). \quad (4)$$

The posterior distribution of ψ_{ik} in (4) is consistent with the result in [20] and [22], where their result is conditioned on a single concentration parameter α instead of $\prod_f \eta_{fk}^{\phi_{if}}$.

Another interesting comparison is the placing of prior information for communities within different models. In iMMM, although the author claimed to use different α_i to model individual π_i, the stick-breaking weights $\{\psi_{ik}\}_{k=1}^\infty$ within one π_i are generated identically, i.e., from beta$(1, \alpha_i)$. This is obviously insufficient as each community may expect an individual prior in real application. Accordingly, NMDR has incorporated node information using a logistic normal function, as stated above. In a way, this approach has further generalized the model, such that each ψ_{ik} differs in their distributions.

Despite the model relaxation, empirical results show that NMDR has a slow convergence. It is therefore imperative for us to search for a more efficient way to incorporate the node information. Compared to iMMM, our niMM model replaces the simple set $\{\alpha_i\}$ with $\prod_f \eta_{fk}^{\phi_{if}}$ for the generation of ψ_{ik}. Its conjugate property makes our model appealing in terms of mixing efficiency, which is confirmed in the results shown in Section VI. What is more, our model can be seen as a natural extension of the popular iMMM model. By letting $\eta_{fk} = \alpha^{1/F}$ and $\phi_{if} = 1$, we obtain the classical iMMM. This makes sense, as without the presence of metadata, each feature is assumed to be counted equally, which implies that the model becomes the classical iMMM.

B. Node-Information Involved Latent Feature Model

The generative process for the niLF model is defined as follows.

C1: $\psi_{ik} \sim \text{Beta}(\prod_f \eta_{fk}^{\phi_{if}}, 1)$.
C2: $\pi_{ik} = \prod_{l=1}^k \psi_{il}$.
C3: $z_{ik} \sim \text{Bernoulli}(\pi_{ik})$.
C4: $e_{ij} \sim \text{Bernoulli}(1/(1+\exp(-z_i B z_j^T)))$.

$C1$ and $C2$ here also constitute our specialized stick-breaking representation π_i. However, we should note that these two are different from those of the niMM model while here they are based on the traditional stick-breaking process for the Indian buffet process [17], [18]. The π_is are used to generate the latent feature matrix z in $C3$. $C4$ corresponds to the link data generation, which is the same as the LFRM model. Similar to the niMM model, this paper can be seen as an extension of the traditional LFRM [14].

Correspondingly, the graphical model of niLF is similar to the one in niMM. One major difference is the setting of z_i and z_j, which is to replace s_{ij} and r_{ij} in the rectangular nodes. That is to say, the niLF model uses the binary vector of z_i and z_j (also B) to determine the probability of generating e_{ij}.

However, the structure of the stick-breaking representation in our niLF model differs from that of the LFRM model. In our niLF model, each ith node's latent feature is motivated by their own stick-breaking representation π_i, i.e., there are $\boldsymbol{n}$ stick-breaking representations in total. In this way, the individual node information of node i is contained in each corresponding representation π_i, which will consequently be reflected in the latent feature. On the contrary, the LFRM model uses one specialized beta process π as the underlying representation for all the n nodes' latent feature z. This process can be easily marginalized out π_i, beneficiated from the Beta-Bernoulli conjugacy [23].

We use the new transform, i.e., $\prod_f \eta_{fk}^{\phi_{if}}$, as the mass parameter [23] in the construction of the stick-breaking representation, as stated in $C1$. The importance indicator η here plays an opposite role when compared to the niMM model, i.e., a larger value of η_{fk} would make the presence of attribute f promote the kth community.

An interesting notation is that the stick-breaking representations in both our niMM and niLF models are no longer the Dirichlet process and Beta process individually, as the single-valued α parameter is replaced by a set of individually different valued $\{\prod_f \eta_{fk}^{\phi_{if}}\}$.

IV. Inference

A. Node-Information Involved Mixed-Membership Model

In niMM's sampling Algorithm 1, the variables of interest in our slice sampling are: node information weight $\{\eta_{fk}\}_{f,k}$, stick-breaking weight $\{\psi_{ik}\}_{i,k}$, latent feature indicator $\{s_{ij}, r_{ij}\}_{i,j}$, compatibility value B_{kl}, and the hyperparameters. Also, we

Algorithm 1 Inference Work for niMM

Input: $\{e_{ij}\}_{n\times n}$: observed relational binary matrix;
$\phi_\cdot$: feature matrix indicates the feature value for each node
T: the given iteration number.
Output: samples of $\{\eta_{fk}\}_{f,k}, \{\psi_{ik}\}_{i,k}, \{s_{ij}, r_{ij}\}_{i,j}$ and hyper-parameters $\alpha_\eta, \beta_\eta, \alpha_B, \beta_B$
Initialize variables $\{\eta_{fk}\}_{f,k}, \{\psi_{ik}\}_{i,k}, \{s_{ij}, r_{ij}\}_{i,j}$ and hyper-parameters $\alpha_\eta, \beta_\eta, \alpha_B, \beta_B$
for $t \in \{1, \cdots, T\}$ **do**
Updating $\{\eta_{fk}\}_{f,k}$ according to Eq. (5)
Updating $\{\psi_{ik}\}_{i,k}$ according to Eq. (6)
Updating $\{s_{ij}, r_{ij}\}_{i,j}$ according to Eq. (8)
Updating $\alpha_\eta, \beta_\eta, \alpha_B, \beta_B$ according to Eq. (9)(10)(11)(12)
end for

Algorithm 2 Inference Work for niLF

Input: $\{e_{ij}\}_{n\times n}$: observed relational binary matrix;
$\phi_\cdot$: feature matrix indicates the feature value for each node
T: the given iteration number.
Output: Samples of $\{\eta_{fk}\}_{f,k}, \{\psi_{ik}\}_{i,k}, \{u_i\}_i, \{z_{ik}\}_{i,k}, \{B_{kl}\}_{k,l}$ and hyper-parameters $\mu_f, \lambda_f, \lambda_v, \lambda_B$
Initialize variables $\{\eta_{fk}\}_{f,k}, \{\psi_{ik}\}_{i,k}, \{u_i\}_i, \{z_{ik}\}_{i,k}, \{B_{kl}\}_{k,l}$ and hyper-parameters $\mu_f, \lambda_f, \lambda_v, \lambda_B$
for $t \in \{1, \cdots, T\}$ **do**
Updating $\{\eta_{fk}\}_{f,k}$ according to Eq. (13)
Updating $\{\psi_{ik}\}_{i,k}$ according to Eq. (14)
Updating $\{u_i\}_i$ according to Eq. (15)
Updating $\{z_{ik}\}_{i,k}$ according to Eq. (17)
Updating $\{B_{kl}\}_{k,l}$ according to Eq. (18)
Updating $\mu_f, \lambda_f, \lambda_v, \lambda_B$ according to Eq. (19)(20)(21)(22)
end for

discuss here the Beta distribution as the generation distribution and the other ones can be trivially derived.

1) Sampling η_{fk}: $\forall f, k$, η_{fk}'s posterior distribution relies on node information $\{\phi_{if}\}_{i=1}^n$, stick-breaking weights $\{\psi_{ik}\}_{i=1}^n$, the other attribute importance indicator $\{\eta_{Fk}\}_{F\neq f}$, and its hyper-parameters α_η and β_η

$$\eta_{fk} \sim \text{Gamma}\left(\alpha_\eta + \sum_i \phi_{if},\ \beta_\eta - \sum_i \phi_{if} \ln(1-\psi_{ik}) \prod_{F\neq f} \eta_{Fk}^{\phi_{iF}}\right). \tag{5}$$

This part requires $\mathcal{O}(FKn)$ operations in each iteration.

2) Sampling ψ_{ik}: $\forall i, k$, ψ_{ik}'s posterior distribution relies on $\{N_{ik}\}_{k=1}^n$, $\{\eta_{fk}\}_{f,k}$, $\{\phi_{if}\}_{q=1}^q$

$$\psi_{ik} \sim \text{Beta}\left(N_{ik}+1,\ \sum_{l=k+1}^{K} N_{il} + \prod_f \eta_{fk}^{\phi_{if}}\right). \tag{6}$$

This part contains $\mathcal{O}(FKn)$ operations in the sampling of ψ_{ik}.

3) Sampling s_{ij}, and r_{ij}: $e_{ij's}$ posterior distribution is a Bernoulli distribution due to the Beta-Bernoulli conjugate, in front of Eq. (7) $\eta_{ij's}$ posterior distribution is a Gamma distribution, in front of Eq. (13)

$$\Pr\left(e_{ij}|Z_{\setminus e_{ij}}, \alpha_B, \beta_B\right) = \frac{m_{kl}^{1,-e_{ij}} + \alpha_B}{m_{kl}^{-e_{ij}} + \alpha_B + \beta_B} \tag{7}$$

here we assume $s_{ij} = k, r_{ij} = l$ and $m_{kl}^{1,-e_{ij}} = \sum_{i'j'\neq ij, s_{i'j'}=k, r_{i'j'}=l} e_{i'j'}$, $m_{kl}^{-e_{ij}} = \sum_{i'j'\neq ij, s_{i'j'}=k, r_{i'j'}=l} 1$.

Thus, we get

$$\Pr\left(s_{ij}=k, r_{ij}=l\right) \propto \pi_{ik}\pi_{jl} \cdot \frac{m_{kl}^{1,-e_{ij}} + \alpha_B}{m_{kl}^{-e_{ij}} + \alpha_B + \beta_B}. \tag{8}$$

When we sample $K+1$ to s_{ij} or r_{ij}, we need to resample the corresponding $\{\eta_{fK+1}\}_{f=1}^F$, ψ_{iK+1}(or ψ_{jK+1}) to the new $(K+1)$th component.

Since we have used the blocked sampling version of sampling (s_{ij}, r_{ij}) together, the computational cost $[\mathcal{O}(N^2K^2)]$ is a bit higher than the alternating sampling scheme [separately sample s_{ij} and r_{ij}, which is $\mathcal{O}(N^2K^2)$]. However, we should note that this blocked sampling version provides a significant running time speed up and better convergence behavior. This part occupies the majority of the computational cost.

4) Sampling Hyper-Parameters α_η, β_η, α_B, and β_B: The hyper-parameters we are sampling are α_η, β_η, α_B, and β_B.

For α_η, we set a vague prior Gamma$(\alpha_{\alpha_\eta}, \beta_{\alpha_\eta})$

$$p\left(\alpha_\eta|\{\eta_{fk}\}_{f,k}, \beta_\eta, \alpha_{\alpha_\eta}, \beta_{\alpha_\eta}\right) \propto \prod_{f,k}\left[\frac{{\beta_\eta}^{\alpha_\eta}}{\text{Gamma}(\alpha_\eta)}\eta_{fk}^{\alpha_\eta-1}\right] \cdot {\alpha_\eta}^{\alpha_{\alpha_\eta}-1} e^{-\beta_{\alpha_\eta}\alpha_\eta}. \tag{9}$$

As (9) is log-concave in α_η, we use adaptive rejection sampling (ARS) to finish its update.

For β_η, we set a vague prior Gamma$(\alpha_{\beta_\eta}, \beta_{\beta_\eta})$

$$\begin{aligned} p\left(\beta_\eta|\{\eta_{fk}\}_{f,k}, \alpha_\eta, \alpha_{\beta_\eta}, \beta_{\beta_\eta}\right) &\propto \prod_{f,k}\left[{\beta_\eta}^{\alpha_\eta} e^{-\beta_\eta\eta_{fk}}\right] \cdot {\beta_\eta}^{\alpha_{\alpha_\eta}-1} e^{-\beta_{\alpha_\eta}\beta_\eta} \\ &\propto {\beta_\eta}^{KF\cdot\alpha_\eta+\alpha_{\alpha_\eta}-1} \cdot e^{-\left(\sum_{f,k}\eta_{fk}+\beta_{\alpha_\eta}\right)\beta_\eta} \\ &\Longrightarrow \beta_\eta \sim \text{Gamma}\left(KF\cdot\alpha_\eta + \alpha_{\alpha_\eta}, \sum_{f,k}\eta_{fk} + \beta_{\alpha_\eta}\right) \end{aligned} \tag{10}$$

where α_B and β_B is similar as above, we set a vague prior Gamma$(\alpha_{\alpha_B}, \beta_{\alpha_B})$

$$p\left(\alpha_B|\{B_{kl}\}_{k,l}, \beta_B, \alpha_{\alpha_B}, \beta_{\alpha_B}\right) \propto \prod_{k,l}\left[\frac{{\beta_B}^{\alpha_B}}{\text{Gamma}(\alpha_B)}B_{kl}^{\alpha_B-1}\right] \cdot {\alpha_B}^{\alpha_{\alpha_B}-1} e^{-\beta_{\alpha_B}\alpha_B}. \tag{11}$$

As (11) is log-concave in α_B, we use ARS to finish its update.

For β_B, we set a vague prior Gamma$(\alpha_{\beta_B}, \beta_{\beta_B})$

$$\begin{aligned} p\left(\beta_B|\{B_{kl}\}_{k,l}, \alpha_B, \alpha_{\beta_B}, \beta_{\beta_B}\right) &\propto \prod_{k,l}\left[{\beta_B}^{\alpha_B} e^{-\beta_B B_{kl}}\right] \cdot {\beta_B}^{\alpha_{\alpha_B}-1} e^{-\beta_{\alpha_B}\beta_B} \\ &\propto {\beta_B}^{K^2\cdot\alpha_B+\alpha_{\alpha_B}-1} \cdot e^{-(\sum_{kl}B_{kl}+\beta_{\alpha_B})\beta_B} \\ &\Longrightarrow \beta_B \sim \text{Gamma}\left(K^2\cdot\alpha_B + \alpha_{\alpha_B}, \sum_{kl}B_{kl} + \beta_{\alpha_B}\right). \end{aligned} \tag{12}$$

B. Node-Information Involved Latent Feature Model

In niLF's sampling Algorithm 2, the variables of interest in our slice sampling are node information weight $\{\eta_{fk}\}_{f,k}$, stick-breaking weight $\{\psi_{ik}\}_{i,k}$, latent feature indicator $\{s_{ij}, r_{ij}\}_{i,j}$, compatibility vale B_{kl}, and the hyperparameters.

1) Sampling η_{fk}: $\eta_{ij's}$ posterior distribution is a Gamma distribution

$$\eta_{fk} \sim \text{Gamma}\left(\alpha_\eta + \sum_i \phi_{if}, \beta_\eta - \sum_i \phi_{if} \ln \psi_{ik} \prod_{F \neq f} \eta_{Fk}^{\phi_{iF}}\right). \tag{13}$$

This part requires $\mathcal{O}(FKn)$ operations in each iteration.

2) Sampling ψ_{ik}: We use Metropolis–Hastings sampling to obtain the ψ_{ik}'s value, so the acceptance ratio becomes that of $(\pi_{ik} = \prod_{l=1}^{k} \psi_{il})$

$$A\left(\psi_{ik}^*, \psi_{ik}^{(\tau)}\right) = \frac{\pi_{ik}^{*,z_{ik}}\left(1-\pi_{ik}^*\right)^{1-z_{ik}}}{\pi_{ik}^{(\tau),z_{ik}}\left(1-\pi_{ik}^{(\tau)}\right)^{1-z_{ik}}}. \tag{14}$$

This part requires $\mathcal{O}(Kn)$ operations in each iteration.

3) Sampling u_i: We introduce an auxiliary slice variable u_i for each node i

$$u_i | z_i, \pi \sim \text{Uniform}[0, \pi_i^*] \tag{15}$$

where $\pi_i^* = \min_{k:z_{ik}=1}\{\pi_{ik}\}$.

This part requires $\mathcal{O}(n)$ operations in each iteration.

4) Sampling $\mathbf{z}_{ik}$: We let $z_i^1 = z_{i,z_{ik}=1}$ and $z_i^0 = z_{i,z_{ik}=0}$, the likelihood term becomes

$$\Pr\left(\boldsymbol{e}_{ij} | Z_{\setminus i}, z_i^1, B\right) = \sigma\left(z_i^1 B z_j\right)^{\boldsymbol{e}_{ij}}\left(1-\sigma\left(z_i^1 B z_j\right)\right)^{1-\boldsymbol{e}_{ij}}. \tag{16}$$

Thus, we get

$$\Pr\left(z_{ik} | \pi_i, \left\{\boldsymbol{e}_{ij}\right\}_{j=1}^n, Z_{\setminus i}, B\right) \propto \begin{cases} \pi_{ik} \prod_j \left[\Pr\left(\boldsymbol{e}_{ij} | Z_{\setminus i}, z_i^1, B\right) \Pr\left(\boldsymbol{e}_{ji} | Z_{\setminus i}, z_i^1, B\right)\right], & z_{ik} = 1 \\ (1-\pi_{ik}) \prod_j \left[\Pr\left(\boldsymbol{e}_{ij} | Z_{\setminus i}, z_i^0, B\right) \Pr\left(\boldsymbol{e}_{ji} | Z_{\setminus i}, z_i^0, B\right)\right], & z_{ik} = 0. \end{cases} \tag{17}$$

This part requires $\mathcal{O}(K^2 n^2)$ operations in each iteration.

5) Sampling B_{kl}: Due to the nonconjugacy of $\sigma(\cdot)$ function, we use the Metropolis–Hastings method to do the sampling. Setting the proposal distribution, the same as the prior distribution Normal$(0, \sigma_B)$, we have the acceptance ratio as

$$A\left(B_{kl}^*, B_{kl}^\tau\right) = \min\left\{1, \frac{f\left(B_{kl}^*\right)}{f\left(B_{kl}^\tau\right)}\right\}. \tag{18}$$

This part requires $\mathcal{O}(K^2)$ operations in each iteration.

6) Sampling Hyper-Parameters λ_f, $\boldsymbol{\mu}_f$, λ_v, and λ_B: For $\boldsymbol{\mu}_f$, we set the prior as Gaussian prior Normal$(0, \lambda_\mu)$, which leads to

$$p\left(\boldsymbol{\mu}_f | \lambda_\mu, \boldsymbol{\eta}, \lambda_f\right) \propto \text{Normal}\left(\boldsymbol{\mu}_f; 0, \lambda_\mu\right) \prod_k \text{Normal}\left(\boldsymbol{\eta}_{fk}; \boldsymbol{\mu}_f, \lambda_f\right) \propto \text{Normal}\left(\boldsymbol{\mu}_f; \frac{\sum_k \boldsymbol{\eta}_{fk}}{\lambda_f^2 + K}, 1 + \frac{K}{\lambda_f^2}\right). \tag{19}$$

For the rest of the hyperparameters, we set the vague gamma prior $\mathcal{G}(a., b.)$ on them and the corresponding update can be done accordingly.

For λ_f, we give the prior on λ_f^{-2}

$$p\left(\lambda_f | a_f, b_f, \eta, \boldsymbol{\mu}_f\right) \propto \mathcal{G}\left(\lambda_f^{-2}; a_f, b_f\right) \prod_f \prod_k \text{Normal}\left(\boldsymbol{\eta}_{fk}; \boldsymbol{\mu}_f, \lambda_f\right) \propto \mathcal{G}\left(\lambda_f^{-2}; a_f + \frac{1}{2}KF, b_f + \frac{1}{2}\sum_k \sum_f \left(\boldsymbol{\eta}_{fk} - \boldsymbol{\mu}_f\right)^2\right). \tag{20}$$

For λ_v, we give the prior on λ_v^{-2}

$$p\left(\lambda_v | a_v, b_v, \eta, \phi\right) \propto \mathcal{G}\left(\lambda_v^{-2}; a_v, b_v\right) \prod_i \text{Normal}\left(v_{i:}; \boldsymbol{\eta}\boldsymbol{\phi}_{i:}^T, \lambda_v\right) \propto \mathcal{G}\left(\lambda_v^{-2}; a_v + \frac{1}{2}KN, b_v + \frac{1}{2}\sum_k \sum_i \left(v_{ik} - \boldsymbol{\eta}_k \phi_i\right)^2\right). \tag{21}$$

For λ_B, we give the prior on λ_B^{-2}

$$p\left(\lambda_B | a_B, b_B, B\right) \propto \mathcal{G}\left(\lambda_B^{-2}; a_B, b_B\right) \prod_k \prod_l \text{Normal}\left(B_{kl}; 0, \lambda_B\right) \propto \text{Gamma}\left(\lambda_B^{-2}; a_B + \frac{1}{2}K^2, b_B + \frac{1}{2}\sum_k \sum_i B_{kl}^2\right). \tag{22}$$

C. π_i-Collapsed Sampling for the niMM Model

When the community number is know in advance, inferencing the niMM model by collapsing the mixed-membership distributions $\{\pi_i\}_i^n$ is a promising solution. W.l.o.g., the membership indicators' joint probability for node i is

$$\Pr\left(\left\{s_{ij}\right\}_{j=1}^n, \left\{r_{ji}\right\}_{j=1}^n | \phi, \eta\right) \propto \frac{\prod_k \text{Gamma}\left(\text{Gamma}\left(N_{ik} + \prod_f \eta_{fk}^{\phi_{if}}\right)\right)}{\text{Gamma}\left(2n + \sum_k \prod_f \eta_{fk}^{\phi_{if}}\right)}. \tag{23}$$

$\forall k \in \{1, \ldots, K\}$, the conditional probability of the membership indicator s_{ij} (the same to r_{ij}) is

$$\Pr\left(s_{ij} = k | \left\{s_{ij_0}\right\}_{j_0 \neq j}, \left\{r_{ji}\right\}_{j_0=1}^n, \phi, \eta\right) \propto N_{ik}^{\setminus s_{ij}} + \prod_f \eta_{fk}^{\phi_{if}}. \tag{24}$$

Compared to its counterpart in MMSB

$$\Pr\left(s_{ij} = k | \left\{s_{ij_0}\right\}_{j_0 \neq j}, \left\{r_{ji}\right\}_{j_0=1}^n, \alpha, K\right) \propto N_{ik}^{\setminus s_{ij}} + \frac{\alpha}{K}. \tag{25}$$

Our collapsed niMM (cniMM) model replaces the term of (α/K) in (25) with $\{\prod_f \eta_{fk}^{\phi_{if}}\}_{k=1}^K$. In fact, while the MMSB generates the mixed-membership distribution π_i through the Dirichlet distribution with parameters $((\alpha/K), \ldots, (\alpha/K))$, our cniMM's corresponding one is the Dirichlet distribution with unequal parameter $(\prod_f \eta_{f1}^{\phi_{if}}, \ldots, \prod_f \eta_{fk}^{\phi_{if}})$.

Due to the unknown information on the undiscovered communities, we limit our cniMM model into this finite communities' number case. The extension on the infinite communities' case remains an interesting future task.

TABLE II
COMPUTATIONAL COMPLEXITY FOR DIFFERENT MODELS

Models	Computational complexity
IRM	$\mathcal{O}(K^2n^2)$ [24]
LFRM	$\mathcal{O}(K^2n^2)$ [24]
MMSB	$\mathcal{O}(Kn^2)$ [15]
NMDR	$\mathcal{O}(Kn^2 + Kn + FKn) = \mathcal{O}(Kn^2)$
niMM	$\mathcal{O}(K^2n^2 + FKn + FKn) = \mathcal{O}(K^2n^2)$
niLF	$\mathcal{O}(K^2n^2 + Kn + FKn) = \mathcal{O}(K^2n^2)$

D. Computational Complexity

We estimate the computational complexities for each model and present the results in Table II. Our niMM and niLF are $\mathcal{O}(K^2n^2 + 2FKn)$ and $\mathcal{O}(K^2n^2 + Kn + FKn)$, respectively, with $\mathcal{O}(Kn)$ for the sampling of $\{\pi_i\}_{i=1}^n$ and $\mathcal{O}(FKn)$ for the incorporation of node information.

V. RELATED WORK

A. Relational Models

The stochastic blockmodel [11] assumes that each node has a latent variable that directly represents its community membership. Each of the fixed number of communities is associated with a weight, and the whole weight vector can be seen as a draw from a K-dimensional Dirichlet distribution. Naturally, the community memberships are realized from the multinomial distribution parameterized by this weight vector. The binary link data between two nodes are determined by the communities to which they belong. This model has been extended to an infinite K community, i.e., IRM [12], where the Dirichlet distribution has been replaced by a Dirichlet process.

Various recent work has been proposed to capture the complex link data among nodes based on the stochastic blockmodel, which can be categorized into two notable branches. The first branch features the LFRM [14]: instead of associating a node with only a single feature, it allows a binary features vector to be associated with each node. All the presented features will be used to generate the link data. The second branch follows the MMSB, in which each node has its own community distribution. In the link data between two nodes, each node chooses one membership indicator from its community distribution and these two indicators consequently fix the parameters to generate the link data.

The LFRM-like work was originated from [25] and [26], while it assumes a latent real-valued feature vector for each node. The LFRM in [14] uses a binary vector to represent the latent features of each node, and the number of features of all the nodes can potentially be infinite by using an Indian buffet process prior [27]. The work in [24] further uncovers the substructure within each feature and uses the "coactive" features from two nodes during the generation of their link data. On the MMSB-type work, a few variants have been subsequently proposed, including infinite mixed membership models (iMMM) [28] which extends the MMSB into the infinite community case with a Dirichlet process prior and [29] which uses the nested Chinese restaurant process [30] to build the hierarchical structure of communities. In this paper, we use iMMM to replace MMSB as our discussion is within the HDP [19] prior. This paper can be trivially applied to the MMSB case.

The idea of modeling an individual mixed-membership distribution for each node bears some resemblance to the author-topic model [31], [32]. However, their model cannot distinguish the differences between components, due to the absence of design in involving node information, as well as a different mechanism in terms of link data generation.

B. Stick-Breaking Process Review

The stick-breaking method [17], [20] has provided us an explicit construction of a draw G from a Dirichlet process

$$G = \sum_k^{\infty} \pi_k \delta_{\theta_k}, \pi_k = \psi_k \prod_{l=1}^{k-1}(1-\psi_l)$$
$$\psi_k \overset{iid}{\sim} \text{Beta}(1,\gamma), \theta_k \overset{iid}{\sim} G_0. \tag{26}$$

The concentration parameter γ controls the diversity of θ in G, whereas G_0 is regarded as the base measure generating $\{\theta_k\}_{k=1}^{\infty}$. A larger γ encourages the weights distribution to be more "flat," whereas a smaller γ stimulates the weights to be "sharper," i.e., only a few weights have appreciable values and the others are relatively small. As an indication of the importance of this concentration parameter γ, a vague gamma prior distribution is usually placed on it.

Based on this ingenious construction, more flexible constructions have been proposed, the recent examples being the logistic stick-breaking process [33], the probit stick-breaking process [34], the kernel stick-breaking process [35], and the discrete infinite logistic normal process [36]. While being elastic in describing the Bayesian nonparametric prior in different situations, one common problem is that they cannot form a prior-posterior conjugate design, which caused difficulties for both the MCMC sampling inference (using Metropolis–Hastings sampling instead would greatly slow down the mixing rate) and variational inference (having to find an approximate distribution to replace this distribution).

Another interesting topic is the stick-breaking construction of the Indian buffet process [27] and its underlying Beta process [23], [37]. As we have already seen, the underlying representation under the Indian buffet process is one Beta process, with the concentration parameter specialized to 1. Teh *et al.* [18] gave us a stick-breaking construction for this specialized Beta process as

$$G = \sum_k^{\infty} \pi_k \delta_{\theta_k}, \pi_k = \prod_{l=1}^{k} \psi_l, \psi_k \overset{iid}{\sim} \text{Beta}(\gamma, 1), \theta_k \overset{iid}{\sim} G_0. \tag{27}$$

Regarding to this special construction for a simplified beta process, a construction of a general Beta process was proposed by Paisley *et al.* [37], which was later followed by an improved version [38].

VI. EXPERIMENTS

We analyze the performance of our models (niMM and niLF) on two real-world data sets: 1) the Lazega data set [13]

TABLE III
PERFORMANCE ON REAL-WORLD DATA SETS (MEAN ∓ STANDARD DEVIATION)

Datasets	Models	Training error	Testing error	Testing log likelihood	AUC
Lazega	IRM	0.0987 ∓ 0.0003	0.1046 ∓ 0.0012	-201.7912 ∓ 3.3500	0.7056 ∓ 0.0167
	LFRM	0.0566 ∓ 0.0024	0.1051 ∓ 0.0064	-222.5924 ∓ 16.1985	0.8170 ∓ 0.0197
	iMMM	0.0487 ∓ 0.0068	0.1096 ∓ 0.0026	-202.7148 ∓ 5.3076	0.8074 ∓ 0.0141
	NMDR	0.0640 ∓ 0.0055	0.1133 ∓ 0.0018	-207.7188 ∓ 3.4754	0.8285 ∓ 0.0114
	niMM	$\mathbf{0.0334 \mp 0.0056}$	0.1067 ∓ 0.0021	$\mathbf{-196.0503 \mp 4.3962}$	$\mathbf{0.8369 \mp 0.0122}$
	niLF	0.0389 ∓ 0.0126	$\mathbf{0.1012 \mp 0.0034}$	-213.5246 ∓ 12.3249	0.8123 ∓ 0.0135
	cniMM	0.0466 ∓ 0.0092	0.1119 ∓ 0.0020	-205.0673 ∓ 4.5321	0.8314 ∓ 0.0119
Reality	IRM	0.0627 ∓ 0.0002	0.0665 ∓ 0.0004	-133.8037 ∓ 1.1269	0.8261 ∓ 0.0047
	LFRM	0.0397 ∓ 0.0017	0.0629 ∓ 0.0037	-143.6067 ∓ 10.0592	0.8529 ∓ 0.0179
	iMMM	0.0297 ∓ 0.0055	0.0625 ∓ 0.0015	$\mathbf{-126.7876 \mp 3.4774}$	0.8617 ∓ 0.0124
	NMDR	0.0386 ∓ 0.0040	0.0668 ∓ 0.0013	-139.5227 ∓ 2.9371	0.8569 ∓ 0.0138
	niMM	$\mathbf{0.0269 \mp 0.0047}$	$\mathbf{0.0621 \mp 0.0015}$	-127.7377 ∓ 3.1313	0.8507 ∓ 0.0134
	niLF	0.0379 ∓ 0.0046	0.0732 ∓ 0.0049	-131.0326 ∓ 9.4521	$\mathbf{0.8645 \mp 0.0139}$
	cniMM	0.0553 ∓ 0.0023	0.0641 ∓ 0.0011	-126.9091 ∓ 2.6459	0.8597 ∓ 0.0099

and 2) the MIT reality mining data set [39]. The comparison models we are using include IRM [12], LFRM [14], iMMM [28] (an infinite community case of MMSB [16]), and NMDR [15].

We have independently implemented the above baseline models to the best of our understanding. There has been a slight variation to NMDR, in which we have employed Gibbs sampling to sample the unknown cluster number, instead of the retrospective MCMC [40] used in the original paper. This setting is to ensure a fair comparison as all of our sampling schemes are under the Gibbs sampling pipeline.

To validate our models' link prediction performance, we use a tenfold cross-validation strategy. For each node's link data, we randomly select one out of ten from them as the test data. Then, we remove these test data and keep the remaining ones as the training data. The corresponding evaluation criteria [15] are the training error ($0-1$ loss) on the training data, the testing error ($0-1$ loss), the testing log likelihood, and the area under the ROC curve (AUC) score on the test data. Specifically, AUC values indicate the expectation that a uniformly drawn random positive is ranked before a uniformly drawn random negative. That is to say, the AUC equals to the probability that our model will rank a randomly chosen interaction 1 higher than a randomly chosen interaction 0. Apart from this, we also conduct a study on learning the node information's importance indicator in the Lazega data set.

At the beginning of the learning process, we set the vague Gamma prior Gamma$(1, 1)$ for the hyperparameters $\alpha_\eta, \beta_\eta, \alpha_B$, and β_B. For B's setting, we set Beta$(1, 1)$ as the prior distribution. For the attributes values that are not in binary form, we have to do the binary transform. The initial states are of random guesses on the hidden labels (membership indicators in MMSB and latent feature in LFRM). For all the experiments, we run chains of 10 000 MCMC samples 30 times, assuming the first 5000 samples are used for burn-in. The average statistics of the remaining 5000 samples are reported.

A. Lazega Data Set [13], [41]

The Lazega data set is on the social network links within a U.S. firm from 1988 to 1991. The data set contains a cowork network for 71 attorneys, in which each directional link data are labeled as 1 (exist) or 0 (absent). Apart from this 71×71 binary asymmetric matrices, the data set also provides information on each node (i.e., attorneys), including the status (partner or associate), gender, office (Boston, Hartford, or Providence), years (with the firm), age, practice (litigation or corporate), and law school (Harvard, Yale, Ucon, or other). After binarizing these attributes, we obtain a 71×11 binary information matrix.

TABLE IV
IMPORTANCE INDICATOR η IN THE LAZEGA DATA SET

Community		1	2	3	4
Office	boston	**0.3103**	1.3139	**0.0877**	2.7415
	hartford	**0.4061**	0.6547	**0.2601**	0.9010
Age	young	1.1884	1.0649	0.8954	1.2016
	middle	0.8562	0.7420	0.7078	0.9639
Years	long	**0.3684**	0.5422	**0.2089**	1.8316
	middle	0.7429	0.7164	0.6534	1.3045
School	Yale	0.9733	0.6881	0.9465	0.7372
	Ucon	1.4117	1.0636	1.1856	0.8408
Status	partner	0.9822	0.8203	0.9583	0.6359
Practice	litigation	**0.2971**	0.9731	**0.3405**	0.9884
Gender	man	**0.3972**	1.1592	0.7156	0.8653

We start the link prediction task here and the result is shown in Table III. Notably, the performance of our implementation of the NMDR model is inferior compared to its original performance in [15]. The reason for this may due to a suboptimal metadata binarization process. However, we have shown that with the same attributes, our niMM model performs better than the NMDR model, as well as the other relational models without the involvement of node information. On the cniMM, its performance is also quite competitive.

1) Node-Information Importance Learning: Another interesting topic here is the learning of the importance indicator η for the node-information. We fix the communities and use the cniMM model to observe the node-information's effect on each individual community. The number of communities is set as 4 and the detail result is shown in Table IV. Also, we should note smaller value indicates larger influence.

Each feature value here represents a column of binary values. Thus, the feature matrix is $n \times 11$, where n denotes the number of attorneys and 11 feature values are boston(office), harford(office), young(age), middle(age), long(year), middle(year), Yale(school), UCon(school), partner(status), litigation(practice), and man(gender). If a middle-aged women attorney comes from Yale, office in boston, long year

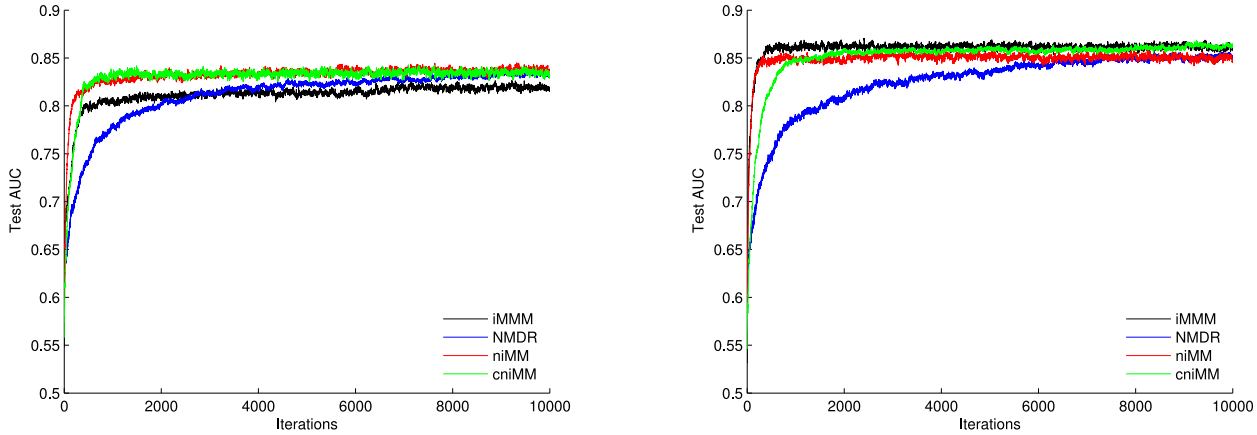


Fig. 3. Trace plot of the AUC value versus iteration time in different MMSB type models (left: Lazega data set and right: MIT reality data set).

TABLE V
MIXING RATE (MEAN ∓ STANDARD DEVIATION) FOR DIFFERENT MODELS, WITH THE BOLD TYPE DENOTING THE BEST WITHIN EACH ROW

Datasets	Criteria	iMMM	LFRM	NMDR	niMM	niLF
Lazega	$\hat{\tau}$	166.2 ∓ 90.37	310.6 ∓ 141.95	179.8 ∓ 156.96	$\mathbf{39.1 \mp 40.58}$	149.2 ∓ 126.12
	ESS	77.6 ∓ 38.71	40.7 ∓ 26.26	134.3 ∓ 133.12	$\mathbf{341.8 \mp 132.00}$	61.2 ∓ 59.93
Reality	$\hat{\tau}$	184.9 ∓ 78.88	113.4 ∓ 77.35	142.8 ∓ 129.99	$\mathbf{27.8 \mp 22.49}$	134.2 ∓ 163.23
	ESS	62.5 ∓ 22.70	125.5 ∓ 71.93	185.0 ∓ 206.12	$\mathbf{449.7 \mp 181.37}$	71.24 ∓ 48.74

service in the company, is not a partner in status and is a practice in litigation, then her feature vector $\phi_i = [1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0]$.

As we can see, the attributes office, long years with firm, and litigation in practice are the smallest among all these attributes. This implies that they are more important than others. Most of this is consistent with our common sense. For instance, people in the same office would usually have more communications in everyday life; the employees would be more familiar with each other if they together have a long time with the firm. An explanation for the surprised result of the importance of the litigation in practice is that it needs corporation.

B. MIT Reality Mining [39]

Based on the MIT reality mining data set, we obtain a proximity matrix describing each node's proximity toward the others, i.e., e_{ij} represents the proximity from i to j based on participant i's opinion. With the same setting of the previous model [28], we manually set the proximity value to be larger than 10 min per day as 1, and 0 otherwise. We hence obtain a 73×73 asymmetric matrix.

Alongside this directional link data, we also have survey data on the participants' information (i.e., node information), including the transport choice to work, social activity, the communication method, and satisfaction of university life. As we can see in Table III, we find our niMM and niLF models' performances are competitive in relation to the ones in iMMM; however, we do not achieve a significant improvement compared to the baseline models. When we trace back to the node information, we find it does not have a direct correlations with the link data. This may be the main reason for our models' less significant result.

C. Convergence Behavior

1) Trace Plot for AUC: Since the AUC value assesses the model's predictability over unseen links, it is natural to use its trace plot to diagnose the inference's convergence behavior, which could also help us choose an appropriate burn-in length. An earlier reach to the stable status of MCMC is desirable as it indicates fast convergence. Fig. 3 shows the detailed results. As we can see, except for NMDR, all the other models reach the stable status quite fast. On the Lazega data set, our niMM and cniMM outperform all the others. On the MIT reality data set, our niMM and cniMM's performances are still quite competitive.

2) Mixing Rate for Stable MCMC: In addition to the MCMC trace plot, another interesting observation is the mixing rate of the stable MCMC chains. We use the number of active communities K as a function of the updated variable to monitor the mixing rate of the MCMC samples, whereas the efficiency of the algorithms can be measured by estimating the integrated autocorrelation time τ and effective sample size (ESS) for K. τ is a good performance indicator as it measures the statistical error of Monte Carlo approximation on a target function f. The smaller the τ, the more efficient the algorithm. Also, the ESS of the stable MCMC chains informs the quality of the Markov chains, i.e., a larger ESS value indicates more independent useful samples, which is our desired property.

On estimating the integrated autocorrelation time, different approaches are proposed in [42]. Here, we use an estimator $\widehat{\tau}$ [22] and the ESS value is calculated based on $\widehat{\tau}$ as

$$\widehat{\tau} = \frac{1}{2} + \sum_{l=1}^{C-1} \widehat{\rho}_l; \ \text{ESS} = \frac{2M}{1 + \widehat{\tau}}. \tag{28}$$

Here, $\widehat{\rho}_l$ is the estimated autocorrelation at lag l and C is a cut-off point which is defined as $C \coloneqq \min\{l : |\widehat{\rho}_l| < 2/\sqrt{M}\}$, and M is equal to half of the original sample size, as the first half is treated as a burn in phase. The detailed results are shown in Table V. As we can see, our model niMM performs the best among all the models.

VII. Conclusion

Increasing applications with natural and social networking behaviors request the effective modeling of hidden relations and structures. This is beyond the currently available models, which only involve limited link information in binary settings. In this paper, we have proposed a unified approach to incorporate the rich node information into the relational models. The proposed niMM model and niLF model have been demonstrated to be effective in learning the structure and have shown advanced performance on learning implicit relations and structures.

We are extending this paper to investigate the following.

1) How to integrate the multirelational networks and unify them into the niMM framework to deeply understand network structures.
2) As there are more advanced constructions for the beta process [37], [38], what are more flexible ways to incorporate the node information into LFRM.
3) When the node information goes beyond the binary scope and becomes the continuous form, how can we utilize such information.
4) Since the multiagent-based method can be used to model the node autonomy in the networks [43], how can we incorporate this node autonomy into the community detection problem.

Appendix
Acronyms and Their Explanations

IRM	Infinite Relational Model
MMSB	Mixed-Membership Stochastic Blockmodel
LFRM	Latent Feature Relational Model
NMDR	Nonparametric Metadata Dependent Relational Model
HDP	Hierarchical Dirichlet Process
LDA	Latent Dirichlet Allocation
niMM	node-information involved Mixed-Membership model
niLF	node-information involved Latent-Feature model
iMMM	Infinite Mixed-Membership Model
MCMC	Markov Chain Monte Carlo
AUC	Area Under the ROC curve
ESS	Effective Sample Size

References

[1] L. Tang and H. Liu, "Community detection and mining in social media," *Synth. Lect. Data Mining Knowl. Disc.*, vol. 2, no. 1, pp. 1–137, 2010.

[2] B. Li, Q. Yang, and X. Xue, "Transfer learning for collaborative filtering via a rating-matrix generative model," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, Montreal, QC, Canada, 2009, pp. 617–624.

[3] B. Li, X. Zhu, R. Li, and C. Zhang, "Rating knowledge sharing in cross-domain collaborative filtering," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1068–1082, May 2015.

[4] X. Fan, L. Cao, and R. Y. D. Xu, "Dynamic infinite mixed-membership stochastic blockmodel," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 2072–2085, Sep. 2015.

[5] X. Fan, B. Li, Y. Wang, Y. Wang, and F. Chen, "The Ostomachion process," in *Proc. AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, 2016.

[6] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Academy Sci.*, vol. 99, no. 12, pp. 7821–7826, 2002.

[7] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, 2010.

[8] P. De Meo, E. Ferrara, D. Rosaci, and G. M. L. Sarne, "Trust and compactness in social network groups," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 205–216, Feb. 2015.

[9] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Distributed fault detection and isolation resilient to network model uncertainties," *IEEE Trans. Cybern.*, vol. 44, no. 11, pp. 2024–2037, Nov. 2014.

[10] W. Wang and Y. Jiang, "Community-aware task allocation for social networked multiagent systems," *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1529–1543, Sep. 2014.

[11] K. Nowicki and T. A. B. Snijders, "Estimation and prediction for stochastic blockstructures," *J. Amer. Stat. Assoc.*, vol. 96, no. 455, pp. 1077–1087, 2001.

[12] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and U. Naonori, "Learning systems of concepts with an infinite relational model," in *Proc. 21st Nat. Conf. Artif. Intell. (AAAI)*, Boston, MA, USA, Jul. 2006, pp. 381–388.

[13] E. Lazega, *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford, U.K.: Oxford Univ. Press, 2001.

[14] K. Miller, M. I. Jordan, and T. L. Griffiths, "Nonparametric latent feature models for link prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2009, pp. 1276–1284.

[15] D. I. Kim, M. Hughes, and E. Sudderth, "The nonparametric metadata dependent relational model," in *Proc. 29th Annu. Int. Conf. Mach. Learn.*, Edinburgh, U.K., 2012 pp. 1559–1566.

[16] 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.

[17] J. Sethuraman, "A constructive definition of Dirichlet priors," *Stat. Sinica*, no. 4, pp. 639–650, 1994.

[18] Y. W. Teh, D. Görür, and Z. Ghahramani, "Stick-breaking construction for the Indian buffet process," in *Proc. Int. Conf. Artif. Intell. Stat.*, vol. 11. San Juan, PR, USA, 2007, pp. 556–563.

[19] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.

[20] H. Ishwaran and L. F. James, "Gibbs sampling methods for stick-breaking priors," *J. Amer. Stat. Assoc.*, vol. 96, no. 453, pp. 161–173, 2001.

[21] D. I. Kim and E. B. Sudderth, "The doubly correlated nonparametric topic model," in *Proc. Adv. Neural Inf. Process. Syst.*, Granada, Spain, 2011, pp. 1980–1988.

[22] M. Kalli, J. E. Griffin, and S. G. Walker, "Slice sampling mixture models," *Stat. Comput.*, vol. 21, no. 1, pp. 93–105, 2011.

[23] R. Thibaux and M. I. Jordan, "Hierarchical beta processes and the Indian buffet process," in *Proc. Int. Conf. Artif. Intell. Stat.*, San Juan, Puerto Rico, USA, 2007, pp. 564–571.

[24] K. Palla, D. A. Knowles, and Z. Ghahramani, "An infinite latent attribute model for network data," in *Proc. 29th Int. Conf. Mach. Learn. (ICML)*, Edinburgh, U.K., Jul. 2012, pp. 1607–1614.

[25] 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.

[26] P. D. Hoff, "Bilinear mixed-effects models for dyadic data," *J. Amer. Stat. Assoc.*, vol. 100, no. 469, pp. 286–295, 2005.

[27] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2006, pp. 475–482.

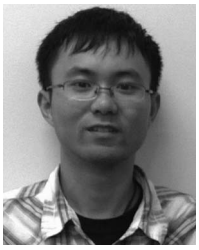
[28] P.-S. Koutsourelakis and T. Eliassi-Rad, "Finding mixed-memberships in social networks," in *Proc. AAAI Spring Symp. Soc. Inf. Process.*, Palo Alto, CA, USA, 2008, pp. 48–53.

[29] Q. Ho, A. P. Parikh, and E. P. Xing, "A multiscale community blockmodel for network exploration," *J. Amer. Stat. Assoc.*, vol. 107, no. 499, pp. 916–934, 2012.

[30] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *J. ACM*, vol. 57, no. 2, pp. 1–30, Feb. 2010.

[31] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proc. 20th Conf. Uncertainty Artif. Intell.*, Banff, AB, Canada, 2004, pp. 487–494.

[32] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, New York, NY, USA, 2004, pp. 306–315.
[33] L. Ren, L. Du, L. Carin, and D. B. Dunson, "Logistic stick-breaking process," *J. Mach. Learn. Res.*, vol. 12, pp. 203–239, Jan. 2011.
[34] A. Rodríguez and D. B. Dunson, "Nonparametric Bayesian models through probit stick-breaking processes," *Bayesian Anal.*, vol. 6, no. 1, pp. 145–177, 2011.
[35] D. B. Dunson and J.-H. Park, "Kernel stick-breaking processes," *Biometrika*, vol. 95, no. 2, pp. 307–323, 2008.
[36] J. Paisley, C. Wang, and D. M. Blei, "The discrete infinite logistic normal distribution," *Bayesian Anal.*, vol. 7, no. 4, pp. 997–1034, 2012.
[37] J. Paisley, A. Zaas, C. W. Woods, G. S. Ginsburg, and L. Carin, "A stick-breaking construction of the beta process," in *Proc. Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 847–854.
[38] J. Paisley, D. M. Blei, and M. I. Jordan, "Stick-breaking beta processes and the poisson process," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2012, pp. 850–858.
[39] N. Eagle and A. Sandy, "Reality mining: Sensing complex social systems," *Pers. Ubiquit. Comput.*, vol. 10, no. 4, pp. 255–268, 2006.
[40] O. Papaspiliopoulos and G. O. Roberts, "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models," *Biometrika*, vol. 95, no. 1, pp. 169–186, 2008.
[41] T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock, "New specifications for exponential random graph models," *Sociol. Methodol.*, vol. 36, no. 1, pp. 99–153, 2006.
[42] C. J. Geyer, "Practical Markov chain Monte Carlo," *Stat. Sci.*, vol. 7, no. 4, pp. 473–483, 1992.
[43] Y. Jiang and J. Jiang, "Understanding social networks from a multiagent perspective," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 10, pp. 2743–2759, Oct. 2014.



Xuhui Fan received the bachelor's degree in mathematical statistics from the University of Science and Technology of China, Hefei, China, in 2010, and the Ph.D. degree in computer science from the University of Technology Sydney, Chippendale, NSW, Australia, in 2015.

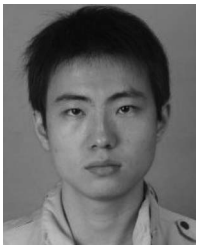
His current research interests include stochastic random partition and Bayesian nonparametrics.

Richard Yi Da Xu received the B.Eng. degree in computer engineering from the University of New South Wales, Sydney, NSW, Australia, in 2001, and the Ph.D. degree in computer sciences from the University of Technology Sydney (UTS), Chippendale, NSW, Australia, in 2006.

He is currently a Senior Lecturer with the School of Computing and Communications, UTS. His current research interests include machine learning, computer vision, and statistical data mining.

Longbing Cao (SM'06) received the Ph.D. degree in pattern recognition and intelligent systems from Chinese Academy of Science, Beijing, China, and the Ph.D. degree in computing sciences from the University of Technology, NSW, Sydney, Australia.

He is currently a Professor, the Founding Director of the Advanced Analytics Institute, and the Data Mining Research Leader of the Australian Capital Markets Cooperative Research Center, University of Technology Sydney, Chippendale, NSW, Australia. His current research interests include big data analytics, data mining, machine learning, behavior informatics, complex intelligent systems, agent mining, and their applications.



Yin Song received the bachelor of science degree in science and technology of electronic information from Beijing Normal University, Beijing, China, the master's degree of engineering in integrated circuit engineering from Tsinghua University, Beijing, and the Ph.D. degree in computer science from the University of Technology Sydney, Ultimo, NSW, Australia, in 2014.

He is a Data Scientist with Brandscreen Ltd., Sydney, NSW, Australia, a world-leading online advertising agency. He is currently focusing on data science applications in real world. He is leading research and development of analytics system and predictive modeling on computational advertising. His current research interests include machine learning, pattern recognition, and data mining.