Dynamic Infinite Mixed-Membership Stochastic Blockmodel

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

***Abstract*— Directional and pairwise measurements are often used to model interactions in a social network setting. The *mixed-membership stochastic blockmodel* (MMSB) was a seminal work in this area, and its ability has been extended. However, models such as MMSB face particular challenges in modeling dynamic networks, for example, with the unknown number of communities. Accordingly, this paper proposes a *dynamic infinite mixed-membership stochastic blockmodel*, a generalized framework that extends the existing work to potentially infinite communities inside a network in dynamic settings (i.e., networks are observed over time). Additional model parameters are introduced to reflect the degree of persistence among one's memberships at consecutive time stamps. Under this framework, two specific models, namely *mixture time variant* and *mixture time invariant* models, are proposed to depict two different time correlation structures. Two effective posterior sampling strategies and their results are presented, respectively, using synthetic and real-world data.**

***Index Terms*— Bayesian nonparametric, dynamic, Gibbs sampling, Markov Chain Monte Carlo (MCMC) inference, mixed-membership stochastic blockmodel (MMSB), slice sampling.**

I. Introduction

NETWORKING applications with dynamic settings (i.e., networks observed over time) are widely seen in real-world environments, such as link prediction and community detection in social networks, social media interactions, capital market movements, and recommender systems. A deep understanding of such dynamic network mechanisms relies on latent relation analysis and latent variable modeling of dynamic network interactions and structures. This presents both challenges and opportunities to existing learning theories. The intricacy associated with the time-varying attributes makes learning and inference a difficult task, but at the same time, one can explore the evolutionary behavior of a network structure more realistically in this time-varying setting. The various dynamic characteristics of such a network can therefore be revealed in real applications.

A number of researchers have recently attempted to address this issue. Some notable earlier examples include *stochastic blockmodel* [1] and its infinite community case *infinite relational model* (IRM) [2] where the aim is to partition a network of nodes into different groups on the basis of their pairwise and directional binary interactions. It was extended in [3] to infer the evolving community's behavior over time. Their work assumes that a fixed number of K communities exist to which one node can potentially belong. However, in many applications, an accurate estimate of K beforehand may be impractical and its value may also vary during time stamps.

A *dynamic* IRM [4] is an alternative way to address the same problem, where K can be inferred from data itself. However, just as described in [2], its drawback is that this model assumes each node i must belong to only one single community. Therefore, an interaction between nodes i and j can only be determined by their community indicators. This approach can be inflexible in many scenarios, such as the monastery example depicted in [5], where one monk can belong to different communities. To this end, Airoldi *et al.* [5] introduce the concept of mixed-membership, where they assume each node i might belong to multiple communities. The membership indicators of one's interaction are no longer a fixed value of a special community. Instead, they are sampled from the nodes' mixed-membership distributions.

The aforementioned work addresses some aspects (such as infinite, dynamic, mixed-membership, and data-driven inference) of relational modeling. An emergent need is to effectively unify these models to provide a flexible and generalized framework which can encapsulate the advantages of most of this paper and address multiple aspects of complexities in one model. This is certainly not an easy thing to do because of the need to understand the relations among aspects and to build a seamless approach to aggregate the challenges. Accordingly, we propose a *dynamic infinite mixed-membership stochastic blockmodel* (DIM3).

DIM3 has the following features: 1) it allows a network to have an infinite number of latent communities; 2) it allows mixed-membership associated with each node; 3) the model adapts to dynamic settings and the number of communities varies with the time; and 4) it is apparent that in many social networking applications, a node's membership may become consistent (i.e., unchanged) over consecutive time stamps. For example, a person's opinion of a peer is more likely to be consistent in two consecutive time stamps.

Manuscript received July 31, 2013; revised August 28, 2014 and November 4, 2014; accepted November 5, 2014. This work was supported by the Australian Research Council Discovery Project under Grant DP130102691.

X. Fan and L. Cao are with the Faculty of Engineering and Information Technology, Advanced Analytics Institute, University of Technology, Sydney, NSW 2007, Australia (e-mail: xuhui.fan@student.uts.edu.au; long-bing.cao@uts.edu.au).

R. Y. D. Xu is with the Faculty of Engineering and Information Technology, University of Technology, Sydney, NSW 2007, Australia (e-mail: yida.xu@uts.edu.au).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author.

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/TNNLS.2014.2369374

To model this persistence, we devise two different implementations. The first is to have a single mixed-membership distribution for each node at different time intervals. The persistence factor is dependent on the statistics of each node's interactions with the rest of the nodes. The second implementation is to allow a set of mixed-membership distributions to associate with each node, and they are time-invariant. The number of elements in the set varies nonparametrically, as reported in [6]. The persistence factor is dependent on the value of the membership indicator at the previous time stamp.

Consequently, two effective sampling algorithms are designed for our proposed models, using either the Gibbs or slice sampling technique for efficient model inference. Their convergence behavior and mixing rate are analyzed and displayed in the first part of the experiment. In the experimental analysis, we show that we can assess nodes' positions in the network and their developing trends, predict unknown links according to the current structure, understand the network structure and identify change points. The techniques proposed can be used for forecasting the political tendencies of senators [7], predicting the function of a protein in biology [8], and tracking authors' community cooperation in academic circles [9].

The rest of the article is organized as follows. Section II introduces the preliminary knowledge for our work, including a brief introduction to mixed-membership stochastic blockmodel (MMSB) and Dirichlet processes. Section III details our main framework and explains how it can incorporate infinite communities in a dynamic setting. The related work is reviewed in Section IV. The inference schemes for the two proposed models are detailed in Section V. In Section VI, we show the experimental results of the proposed models by using both synthetic and real-world social network data. The conclusion is given in Section VII.

II. Preliminary Knowledge

A. Notations

For notational clarity, we first define the key terms and their meanings, as shown in Table I.

B. Introduction to MMSB and Bayesian Nonparametrics

1) Mixed-Membership Stochastic Blockmodel: MMSB [5] aims to model each node's individual mixed-membership distribution. In MMSB, each interaction e_{ij} corresponds to two membership indicators: s_{ij} from the sender i and r_{ij} to the receiver j (w.l.o.g. (Without Loss Of Generality), we assume $s_{ij} = k$, $r_{ij} = l$). The interaction's value is determined by the compatibility of two corresponding communities k and l. Fig. 1 shows the graphical model, and the detailed generative process can be described as:

1) $\forall\{k, l\} \in \mathcal{N} > 0$, draw the communities' compatibility values $W_{k,l} \sim Beta(\lambda_1, \lambda_2)$;
2) $\forall i \in \{1, \cdots, n\}$, draw node i's mixed-membership distribution $\pi_i \sim Dirichlet(\beta)$;
3) $\forall\{i, j\} \in \{1, \cdots, n\}^2$, for interaction 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 interaction $e_{ij} \sim$ Bernoulli $(W_{s_{ij},r_{ij}})$.

TABLE I

Notations for DIM3

n	number of nodes
K	number of discovered communities
T	number of whole time stamps
t	the specific time stamp
e^t_{ij}	directional, binary interactions at time t
$\boldsymbol{\beta}$	a stick-breaking representation to denote the "significance" of all existing communities at all times
γ, α	concentration parameters for HDP
κ	a sticky parameter representing the time-persistence effect
s^t_{ij}	sender's (from i to j) membership indicator at time t
r^t_{ij}	receiver's (from j to i) membership indicator at time t
Z	all the membership indicators, i.e. $Z = \{s^t_{ij}, r^t_{ij}\}_{i,j,t}$
$z^t_{i\cdot}$	node i's membership indicators at time t, i.e. $\{s^t_{ij}, r^t_{ji}\}^n_{j=1}$
$\boldsymbol{m}^t_{ik}$	in the Chinese Restaurant Franchise analogy, the number of tables having dish k at restaurant i and time t
π^t_i	mixed-membership distribution for node i at time t, it generates $s^t_{\boldsymbol{i}1}, \cdots, s^t_{\boldsymbol{i}n}, r^t_{1\boldsymbol{i}}, \cdots, r^t_{n\boldsymbol{i}}$
π^t_{ik}	the "significance" of community k for node i at time t
W	role-compatibility matrix
$W_{k,l}$	compatibilities between communities k and l
$n^t_{k,l}$	number of links from communities k to l at time t i.e. $n^t_{k,l} = \#\{ij : s^t_{ij} = k, r^t_{ij} = l.\}$
$n^{t,1}_{k,l}$	part of $m_{k,l}$ where the corresponding $e^t_{ij} = 1$ at time t, i.e. $n^{t,1}_{k,l} = \sum_{s^t_{ij}=k, r^t_{ij}=l} e^t_{ij}$
$n^{t,0}_{k,l}$	part of $m_{k,l}$ where the corresponding $e^t_{ij} = 0$ at time t, i.e. $n^{t,0}_{k,l} = n^t_{k,l} - n^{t,1}_{k,l}$
N^t_{ik}	number of times that a node i has participated in community k (either sending or receiving message) at time t, i.e. $N^t_{ik} = \#\{j : s^t_{\boldsymbol{i}j} = k\} + \#\{j : r^t_{j\boldsymbol{i}} = k\}$

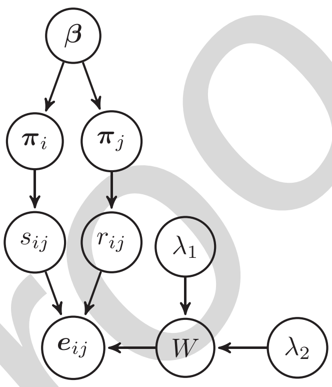


Fig. 1. MMSB model.

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

W is the communities' compatibility matrix as described previously. The prior $P(W)$ is elementwise beta distributed, which is a conjugate to the Bernoulli distribution $P(e_{ij}|.)$. Therefore, a marginal distribution of $P(e_{ij})$, that is, $\int_W p(e_{ij}|W)p(W)d(W)$ can be obtained on the basis of data analysis, and hence there is no need to explicitly sample the values of W.

2) Bayesian Nonparametrics: In the dynamic setting, the Bayesian nonparametric method is a perfect tool for allowing the communities' numbers to vary across time periods. In our case, we use variants of the hierarchical Dirichlet

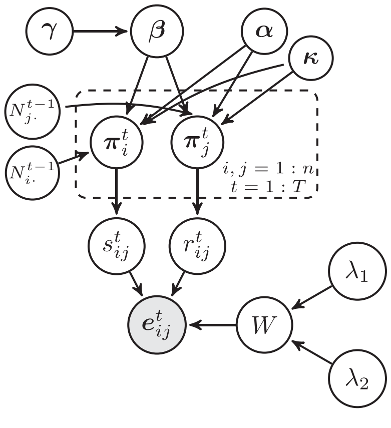


Fig. 2. MTV model.

process (HDP) [10] 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 $\boldsymbol{\beta} = \sum_{k=1}^{\infty} \boldsymbol{\beta}_k \delta_k, \boldsymbol{\beta}_k = \beta_k' \prod_{l=1}^{k-1}(1 - \beta_l'), \beta_l' \sim Beta(1, \gamma))$ [11].

III. Dynamic Infinite Mixed-Membership Stochastic Blockmodel

A. General Settings

In DIM3, we allow each node's membership indicators to change across time periods. Additionally, it is imperative that these indicators should contain the time-persistence property with past values, through which the reality of social behavior can be reflected. Here, we use the strategy of incorporating a sticky parameter κ into the mixed-membership distributions to overcome this issue [6], [12]. Different detailed designs are proposed for the mixture time variant (MTV) and mixture time invariant (MTI) models; however, the common idea is that the current mixed-membership distributions are influenced by the corresponding distributions at the previous time.

Once the current mixed-membership distributions have been selected, the interaction data is generated in the same way as MMSB. Thus, this paper is focused on the details of mixed-membership distribution constructions following the main route of the HDP [10]. Also, we should note that the intermediate variable $\boldsymbol{\beta}$ is identical for both models, representing the significance of all the communities across time periods, and its construction is the same as the stick-breaking construction as described in Section II-B2.

B. Mixture Time Variant (MTV) Model

Fig. 2 shows the graphical model of the MTV model. Here we only show all the variables involved for time t, and omit those for the other time points, where the structure is identical at any other time $\tau \neq t$.

Let us focus on the mixed-membership distribution's construction in the MTV model, which is

$$\boldsymbol{\pi}_i^t \sim DP\left(\alpha + \kappa, \frac{\alpha\boldsymbol{\beta} + \frac{\kappa}{2n} \cdot \sum_k N_{ik}^{t-1} \boldsymbol{\delta}_k}{\alpha + \kappa}\right) \tag{1}$$

$$s_{ij}^t \sim \boldsymbol{\pi}_i^t, r_{ij}^t \sim \boldsymbol{\pi}_j^t \quad \forall i, j \in \mathcal{N}, t \geq 1. \tag{2}$$

The mixed-membership distribution $\{\boldsymbol{\pi}_i^t\}_{1:n}^{1:T}$ is sampled from the Dirichlet process with a concentration parameter $(\alpha + \kappa)$ and a base measure $(\alpha\boldsymbol{\beta} + \frac{\kappa}{2n}\sum_k N_{ik}^{t-1}\boldsymbol{\delta}_k/\alpha + \kappa)$. There will be $N \times T$ of these distributions. They jointly describe each node's activities.

In the base measure, the introduced sticky parameter κ stands for each node's time influence on its mixed-membership distribution. In other words, we assume that each node's mixed-membership distribution at time t will be largely influenced by its activities at time $t-1$. This is reflected in the hidden label's multinomial distribution whereby the previous explicit activities will occupy a fixed proportion $\kappa/\alpha + \kappa$ of the current distribution. The larger the value of κ, the more weight the activities at $t - 1$ will have at time t.

As our method is largely based on the HDP framework, we use the popular Chinese Restaurant Franchise (CRF) [6], [10] analogy to explain our model. Using the CRF analogy, the mixed-membership distribution associated with a node i at time t can be seen as a restaurant $\boldsymbol{\pi}_i^t$, with its dishes representing the communities. If a customer s_{ij}^t(or r_{ji}^t) eats the dish k at the ith restaurant at time t, then $s_{ij}^t(r_{ji}^t) = k$. For all $t > 1$, the restaurant $\boldsymbol{\pi}_i^t$ will have its own specials on the dishes served, representing the sticky configuration in the graphical model. In contrast to the sticky HDP–hidden Markov model (HMM) [6] approach, which places emphasis on one dish only, we allow multiple specials in our work, where the weight of each special dish is adjusted according to the number of dishes served at this restaurant at time $t-1$, that is, $(\kappa/2n)\sum_k N_{ik}^{t-1}\boldsymbol{\delta}_k$. Therefore, we can ensure that the special dishes are served persistently across time in the same restaurant.

C. Mixture Time Invariant (MTI) Model

We show the MTI model in Fig. 3. Here we only show the interaction e_{ij}^1 and omit the other interactions, whose structure is directly derived.

The $\boldsymbol{\beta}$ in the MTI model is identical to that in the MTV model, and we sample the mixed-membership distribution and membership indicators as follows:

$$\boldsymbol{\pi}_i^{(k)} \sim \text{DP}\left(\alpha + \kappa, \frac{\alpha\boldsymbol{\beta} + \kappa\boldsymbol{\delta}_k}{\alpha + \kappa}\right) \quad \forall i, k \in \mathcal{N} \tag{3}$$

$$s_{ij}^t \sim \boldsymbol{\pi}_i^{\left(s_{ij}^{t-1}\right)}, r_{ij}^t \sim \boldsymbol{\pi}_j^{\left(r_{ij}^{t-1}\right)} \quad \forall i, j \in \mathcal{N}, t \geq 1. \tag{4}$$

We assign uninformative priors on sampling the initial membership indicators $\{s_{ij}^0, r_{ij}^0\}_{i,j}$, that is, $\{s_{ij}^0, r_{ij}^0\}_{i,j}$ are sampled from a multinomial distribution, with each category having an

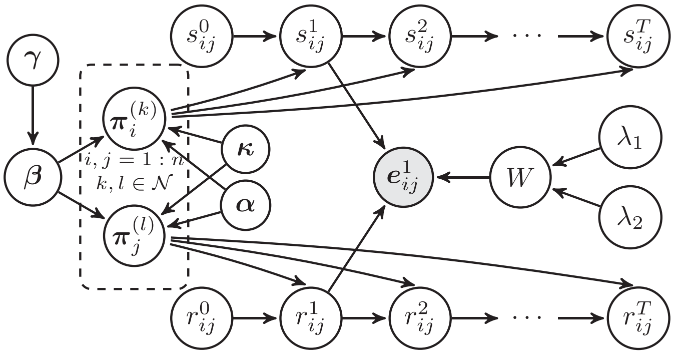


Fig. 3. MTI model.

equalized success probability. The dimension of this multinomial distribution is automatically adjusted according to the current number of communities in the model.

On each node's membership distribution, our MTI model is essentially a Sticky HDP–HMM [6], [12], [13]. In this model, each node has a variable number of mixed-membership distributions associated with it, which may be infinite. At time $t \geq 2$, its membership indicator s_{ij}^t (or r_{ij}^t) is generated from $\boldsymbol{\pi}_i^{(s_{ij}^{t-1})}$ (or $\boldsymbol{\pi}_j^{(r_{ij}^{t-1})}$). To encourage persistence, each π_{ik} is generated from the corresponding $\boldsymbol{\beta}$, where κ is added to $\boldsymbol{\beta}$'s kth component [6], [12], [13].

Returning to the CRF [10] analogy, we have $N \times \infty$ matrix, where its (i, k)th element refers to $\pi_i^{(k)}$, which can be seen as the weights of eating each of the available dishes. A customer s_{ij}^t (or r_{ji}^t) can therefore only travel between restaurants located at the ith row of the matrix. When $\pi_i^{(k)}$'s kth component is more likely to be larger, it means that the dish k is a special dish for restaurant k. Therefore, a customer at restaurant k at time $t - 1$ is more likely to eat the same dish (i.e., kth dish), and hence to stay at restaurant k at time t.

D. Discussion and Comparison

Here, we discuss the difference between the two models in the design of the time-persistence property. The MTV model allows the mixed-membership distribution itself to change over time stamps. However, there is only a single (but different) distribution for each node at each individual time stamp. The membership indicator of a node at time t is dependent on the statistics of all membership indicators of the same node at $t - 1$ and $t + 1$. With a larger value of the sticky parameter κ, the current mixed-membership distribution tends to be more similar to that of the previous time stamp.

In contrast, the MTI model requires the mixed-membership distributions to stay invariant over time. However, there may be an infinite number of possible distributions associated with each node, due to a HDP prior, often only a few distributions will be discovered. In this case, the membership indicator at the current time is dependent and more likely to have the same value as it has in the previous time stamp.

IV. Related Work

We here provide a detailed review of some of the current state-of-the-art in relational learning and at the same time, distinguish our paper from existing ones. In general, we categorize the relational learning models into two major frameworks: the *latent feature model* (LFM) and *latent class model* (LCM). Both frameworks assume that a node's interaction is a Bernoulli draw, which is parameterized by an entry from the role-compatibility matrix. Their main difference is hence in the way the entry is indexed. For LCM, it is assumed that the indices for each pair of nodes are derived from the two associated hidden class labels; in case of LFM, it is assumed that the indices are, however, determined from a set of latent features associated with the pair of nodes.

A representative work for LFM is the *latent feature relational model* (LFRM) [14], which uses a latent feature matrix and a corresponding link generative function to define the model. To account for the variable number of features associated with each node, it uses the Indian Buffet Process [15], [16] as a prior. The *max-margin latent feature relational model* (Med-LFRM) [17] uses the *maximum entropy discrimination* (MED) [18] technique to minimize the hinge loss which measures the quality of link prediction. The *infinite latent attribute* (ILA) model [19] uses a Dirichlet process to construct a substructure within each feature, and all the features are used through the LFRM model.

On the LCM front, the classical approach is the MMSB which enables each node to be associated with multiple membership indicators, and an interaction is formed using one of these indicators. Several variants are subsequently proposed from MMSB, with examples including [20] which extends the MMSB into the infinite communities case [21], which uses the nested Chinese Restaurant Process [22] to build a communities' hierarchical structure, and [23] which incorporates the node's attribute information into its membership indicator construction in MMSB.

Like any data modeling problem, interaction data may also change over time; therefore, dynamic extensions are found in both the LCM and LFM frameworks. Examples such as [24] and [25] describe the time dependency by using Gaussian linear motion models. The *dynamic relational infinite feature model* (DRIFT) [26], which employs an independent Markov dynamic transition matrix to correlate consecutive time interaction data, is a natural extension of the LFRM. *Latent feature propagation* (LFP) [9] directly integrates observed interactions, rather than the latent feature matrix, in the current time to model the distribution of latent features at the next time stamp. On the dynamic setting of MMSB, Xing *et al.* [8] and Fu *et al.* [27] place a parameter (the mean)-dependent Gaussian distribution to consider the time correlation, whereas Ho *et al.* [7] consider hierarchical communities modeling that evolves. However, as both of these two models require predefinition of the number of communities, additional techniques, such as cross-validation, are necessary when choosing the number of communities. Furthermore, their implicit description of the time dependency may not be sufficiently intuitive.

V. Inference

Two sampling schemes are implemented to complete the inference on the MTV model: standard Gibbs sampling and slice-efficient sampling, which both target the same posterior distribution.

A. Gibbs Sampling for the MTV Model

The Gibbs sampling scheme is largely based on [10]. The variables of interest are: $\boldsymbol{\beta}$, Z and auxiliary variables $\hat{\boldsymbol{m}}$, where $\hat{\boldsymbol{m}}$ refers to the number of tables having dish k as in [6] and [10] without counting the tables that are generated from the sticky portion, that is, κN_{ik}^{t-1}. Note that we do not sample $\{\boldsymbol{\pi}_i^t\}_{1:n}^{1:T}$, as it gets integrated out.

1) Sampling $\boldsymbol{\beta}$: $\boldsymbol{\beta}$ is the prior for all $\{\boldsymbol{\pi}_i^t\}s$, which can be viewed as the ratios between the community components for all communities. Its posterior distribution is obtained through the auxiliary variable $\hat{\boldsymbol{m}}$

$$(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\beta}_\mu) \sim \mathrm{Dir}(\hat{\boldsymbol{m}}_{\cdot 1}, \cdots, \hat{\boldsymbol{m}}_{\cdot K}, \gamma) \tag{5}$$

where its detail can be found in [10].

2) Sampling $\{s_{ij}^t\}_{n\times n}^{1:T}$, $\{r_{ij}^t\}_{n\times n}^{1:T}$: Each observation e_{ij}^t is sampled from a fixed Bernoulli distribution, where the Bernoulli's parameter is contained within the role-compatibility matrix W whose rows and columns are indexed by a pair of corresponding membership indicators $\{s_{ij}^t, r_{ij}^t\}$. W.l.o.g., $\forall k, l \in \{1, \cdots, K+1\}$, the joint posterior probability of $(s_{ij}^t = k, r_{ij}^t = l)$ is

$$\begin{aligned}
&\Pr\left(s_{ij}^t = k, r_{ij}^t = l | Z \backslash \{s_{ij}^t, r_{ij}^t\}, e, \boldsymbol{\beta}, \alpha, \lambda_1, \lambda_2, \kappa\right)\\
&\quad \propto \Pr\left(s_{ij}^t = k | \{s_{ij_0}^t\}_{j_0 \neq j}, \{r_{j_0 i}^t\}_{j_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t-1}\right)\\
&\quad\quad \cdot \prod_{l=1}^{2n} \Pr\left(z_{il}^{t+1} | z_{i\cdot}^t / s_{ij}^t, s_{ij}^t = k, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t+1}\right)\\
&\quad\quad \cdot \Pr\left(r_{ij}^t = l | \{r_{i_0 j}^t\}_{i_0 \neq i}, \{s_{j i_0}\}_{i_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t-1}\right)\\
&\quad\quad \cdot \prod_{l=1}^{2n} \Pr\left(z_{jl}^{t+1} | z_{j\cdot}^t / r_{ij}^t, r_{ij}^t = l, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t+1}\right)\\
&\quad\quad \cdot \Pr\left(e_{ij}^t | E \backslash \{e_{ij}^t\}, s_{ij}^t = k, r_{ij}^t = l, \boldsymbol{Z} \backslash \ \{s_{ij}^t, r_{ij}^t\}, \lambda_1, \lambda_2\right).
\end{aligned} \tag{6}$$

The first two terms of (6)

$$\begin{aligned}
&\Pr\left(s_{ij}^t = k | \{s_{ij_0}^t\}_{j_0 \neq j}, \{r_{j_0 i}^t\}_{j_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t-1}\right)\\
&\quad \cdot \prod_{l=1}^{2n} \Pr\left(z_{il}^{t+1} | z_{i\cdot}^t / s_{ij}^t, s_{ij}^t = k, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t+1}\right)\\
&\propto \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k + N_{ik}^{t+1} + \kappa N_{ik}^{t,-s_{ij}^t} + \kappa\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + N_{ik}^{t+1} + \kappa N_{ik}^{t,-s_{ij}^t}\right)} \cdot \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t,-s_{ij}^t}\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t,-s_{ij}^t} + \kappa\right)}\\
&\quad \cdot \begin{cases} \alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t-1} + N_{ik}^{t,-s_{ij}^t}, & k \in \{1, \ldots, K\};\\ \alpha\boldsymbol{\beta}_\mu, & k = K+1 \end{cases}
\end{aligned} \tag{7}$$

where $N_{ik}^0 = 0$, $N_{ik}^{T+1} = 0$, $\forall i \in \{1, \ldots, n\}, k \in \{1, \ldots, K\}$.

The following two terms of (6) are:

$$\begin{aligned}
&\Pr\left(r_{ij}^t = l | \{r_{i_0 j}^t\}_{i_0 \neq i}, \{s_{j i_0}^t\}_{i_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t-1}\right)\\
&\quad \cdot \prod_{l=1}^{2n} \Pr\left(z_{jl}^{t+1} | z_{j\cdot}^t / r_{ij}^t, r_{ij}^t = l, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t+1}\right)\\
&\propto \frac{\Gamma\left(\alpha\boldsymbol{\beta}_l + N_{jl}^{t+1} + \kappa N_{jl}^{t,-r_{ij}^t} + \kappa\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_l + N_{jl}^{t+1} + \kappa N_{jl}^{t,-r_{ij}^t}\right)} \cdot \frac{\Gamma\left(\alpha\boldsymbol{\beta}_l + \kappa N_{jl}^{t,-r_{ij}^t}\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_l + \kappa N_{jl}^{t,-r_{ij}^t} + \kappa\right)}\\
&\quad \cdot \begin{cases} \alpha\boldsymbol{\beta}_l + \kappa N_{jl}^{t-1} + N_{jl}^{t,-r_{ij}^t}, & l \in \{1, \ldots, K\}\\ \alpha\boldsymbol{\beta}_\mu, & l = K+1. \end{cases}
\end{aligned} \tag{8}$$

The last term, that is, the likelihood term, is calculated as

$$\begin{aligned}
&\Pr\left(e_{ij}^t | E \backslash \{e_{ij}^t\}, s_{ij}^t = k, r_{ij}^t = l, \boldsymbol{Z} \backslash \{s_{ij}^t, r_{ij}^t\}, \lambda_1, \lambda_2\right)\\
&\quad = \begin{cases} \dfrac{n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^t} + \lambda_1}{n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} + \lambda_1 + \lambda_2}, & \boldsymbol{e}_{ij}^t = 1\\[2ex] \dfrac{n_{k,l}^{t,0,-\boldsymbol{e}_{ij}^t} + \lambda_2}{n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} + \lambda_1 + \lambda_2}, & \boldsymbol{e}_{ij}^t = 0 \end{cases}
\end{aligned} \tag{9}$$

where $n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} = n_{k,l}^t - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l) = \sum_{i'j'} \mathbf{1}(s_{i'j'}^t = k, r_{i'j'}^t = l) - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l)$, $n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^t} = n_{k,l}^{1,t} - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l)\boldsymbol{e}_{ij}^t = \sum_{i'j':s_{i'j'}^t=k, r_{i'j'}^t=l} \boldsymbol{e}_{i'j'}^t - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l)\boldsymbol{e}_{ij}^t$, and $n_{k,l}^{t,0,-\boldsymbol{e}_{ij}^t} = n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} - n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^t}$.

The detailed derivation of (7)–(9) is given in Assuming the current sample of $\{s_{ij}^t, r_{ij}^t\}$ has values ranging between $1 \ldots K$, we let the undiscovered (i.e., new) community be indexed by $K+1$. Then, to sample a pair (s_{ij}^t, r_{ij}^t) in question, we need to calculate all $(K+1)^2$ combinations of values for the pair.

3) Sampling $\hat{\boldsymbol{m}}$: Using the restaurant-table-dish analogy, we denote $\boldsymbol{m}_{ik}^t$ as the number of tables having dish k, $\forall i, k, t$. This is related to the variable $\hat{\boldsymbol{m}}$ used in sampling $\boldsymbol{\beta}$; it also includes the counts of the unsticky portion, that is, $\alpha\boldsymbol{\beta}_k$.

The sampling of $\boldsymbol{m}_{ik}^t$ incorporates a similar strategy as in [6] and [10], which is independently distributed from

$$\Pr\left(\boldsymbol{m}_{ik}^t = m | \alpha, \boldsymbol{\beta}_k, N_{ik}^{t-1}, \kappa\right) \propto S\left(N_{ik}^t, m\right)\left(\alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t-1}\right)^m \tag{10}$$

where $S(\cdot, \cdot)$ is the Stirling number of the first kind.

For each node, the ratio of generating new tables is the result of two factors: 1) a Dirichlet prior with parameter $\{\alpha, \boldsymbol{\beta}\}$ and 2) the sticky configuration from membership indicators at $t-1$, that is, κN_{ik}^{t-1}.

To sample $\boldsymbol{\beta}$, we need to only include tables generated from the unsticky portion, that is, $\hat{\boldsymbol{m}}$, where each $\hat{\boldsymbol{m}}_{ik}^t$ can be obtained from a single binomial raw

$$\hat{\boldsymbol{m}}_{ik}^{t} \sim \text{Binomial}\left(\boldsymbol{m}_{ik}^{t}, \frac{\alpha \boldsymbol{\beta}_k}{\frac{\kappa}{2n} N_{ik}^{t-1} + \alpha \boldsymbol{\beta}_k}\right). \tag{11}$$

$$\hat{\boldsymbol{m}}_k = \sum_{i,t} \hat{\boldsymbol{m}}_{ik}^{t}. \tag{12}$$

B. Adapted Slice-Efficient Sampling for the MTV Model

We also incorporate the slice-efficient sampling [28], [29] to our model. The original sampling scheme was designed to sample the Dirichlet process mixture model. To adapt it to our framework, which is based on a HDP prior and also has pairwise membership indicators, we use the auxiliary variables $U = \{u_{ij,s}^{t}, u_{ij,r}^{t}\}$ for each of the latent membership pairs $\{s_{ij}^{t}, r_{ij}^{t}\}$. With Us, we are able to limit the number of components in which $\boldsymbol{\pi}_i$ needs to be considered, which is otherwise infinite.

Under the slice-efficient sampling framework, the variables of interest are now extended to: $\boldsymbol{\pi}_i^t, \{u_{ij,r}^t, u_{ij,s}^t\}, \{s_{ij}^t, r_{ij}^t\}, \boldsymbol{\beta}, \boldsymbol{m}$:

1) Sampling $\boldsymbol{\pi}^t$: For each node $i = 1, \ldots, N$; $t = 1, \ldots, T$: we generate $\boldsymbol{\pi}_i^{'t}$ using the stick-breaking process [11], where each kth component is generated using $\boldsymbol{\pi}_{ik}^{'t} \sim \text{beta}(\boldsymbol{\pi}_{ik}^{'t}; a_{ik}^t, b_{ik}^t)$ where

$$\begin{aligned} a_{ik}^t &= \alpha \boldsymbol{\beta}_k + N_{ik}^t + \kappa N_{ik}^{t-1} \\ b_{ik}^t &= \alpha\left(1 - \sum_{l=1}^{k} \boldsymbol{\beta}_l\right) + N_{i,k_0>k}^t + \kappa N_{i,k^0>k}^{t-1} \end{aligned} \tag{13}$$

where $\boldsymbol{\pi}_k^t = \boldsymbol{\pi}_k^{'t} \prod_{i=1}^{k-1}(1 - \boldsymbol{\pi}_i^{'t})$.

2) Sampling $u_{ij,s}^t, u_{ij,r}^t, s_{ij}^t, r_{ij}^t$: We use $u_{ij,s}^t \sim U(0, \boldsymbol{\pi}^t_{is_{ij}^t})$, $u_{ij,r}^t \sim U(0, \boldsymbol{\pi}^t_{jr_{ij}^t})$. The hidden label subsequently obtained is then independently sampled from the finite candidates

$$\begin{aligned} &P\left(s_{ij}^t = k, r_{ij}^t = l | Z, e_{ij}^t, \boldsymbol{\beta}, \alpha, \kappa, N, \boldsymbol{\pi}, u_{ij,s}^t, u_{ij,r}^t)\right) \\ &\quad \propto \mathbf{1}\left(\boldsymbol{\pi}_{ik}^t > u_{ij,s}^t\right) \cdot \mathbf{1}\left(\boldsymbol{\pi}_{jl}^t > u_{ij,r}^t\right) \\ &\qquad \cdot \prod_{l=1}^{2n} \Pr\left(z_{il}^{t+1} | z_{i\cdot}^t / s_{ij}^t, s_{ij}^t = k, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t+1}\right) \\ &\qquad \cdot \prod_{l=1}^{2n} \Pr\left(z_{jl}^{t+1} | z_{j\cdot}^t / r_{ij}^t, r_{ij}^t = l, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t+1}\right) \\ &\qquad \cdot \Pr\left(e_{ij}^t | E \backslash \{e_{ij}^t\}, s_{ij}^t = k, r_{ij}^t = l, \boldsymbol{Z} \backslash \{s_{ij}^t, r_{ij}^t\}, \lambda_1, \lambda_2\right). \end{aligned} \tag{14}$$

We refer the reader to (7)–(9) for the detailed calculation of each term in (14).

3) Sampling $\boldsymbol{\beta}$: An obvious choice for the proposal distribution of $\boldsymbol{\beta}$ used in M-H is its prior $p(\boldsymbol{\beta}|\gamma) = \text{stick} - \text{breaking}(\gamma)$. However, this proposal may be noninformative, which results in a low acceptance rate. We sample $\boldsymbol{\beta}^*$ conditioned on an auxiliary variable $\hat{\boldsymbol{m}}$: $(\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*, \boldsymbol{\beta}_{K+1}^*) \sim Dir(\hat{\boldsymbol{m}}_1, \ldots, \hat{\boldsymbol{m}}_K, \gamma)$, to increase the M-H's acceptance rate, where $\hat{\boldsymbol{m}}$ are sampled in accordance with the method proposed in Section V-A3 [(10)–(12)]. However, instead of sampling $\boldsymbol{\beta}$ directly from $\boldsymbol{m}$ as described in Section V-A3, we only use it for our proposal distribution, as we explicitly sample $\{\pi_i\}_{i=1}^n$. The acceptance ratio is hence (τ indexes the iteration time)

$$A(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(\tau)}) = \min(1, a) \tag{15}$$

$$a = \frac{\prod_{t,i}\left[\prod_{d=1}^{K+1} \Gamma\left(\alpha \boldsymbol{\beta}_d^{(\tau)}\right) \cdot \left[\pi_{id}^t\right]^{\alpha \boldsymbol{\beta}_d^*}\right]}{\prod_{t,i}\left[\prod_{d=1}^{K+1} \Gamma\left(\alpha \boldsymbol{\beta}_d^*\right) \cdot \left[\pi_{id}^t\right]^{\alpha \boldsymbol{\beta}_d^{(\tau)}}\right]} \cdot \frac{\prod_{d=1}^{K} \left[\boldsymbol{\beta}_d^{(\tau)}\right]^{\hat{\boldsymbol{m}}_d - \gamma}}{\prod_{d=1}^{K} \left[\boldsymbol{\beta}_d^*\right]^{\hat{\boldsymbol{m}}_d - \gamma}}. \tag{16}$$

C. Hyperparameter Sampling

The hyperparameters involved in the MTV model are γ, α, and κ. However, it is impossible to compute their posterior individually. Therefore, we place three prior distributions on some combination of the variables. A vague gamma prior $\mathcal{G}(1, 1)$ is placed on both $\gamma, (\alpha + \kappa)$. A beta prior is placed on the ratio $\kappa/\alpha + \kappa$.

To sample γ value, since $\log(\gamma)$'s posterior distribution is log-concave, we use the adaptive rejection sampling (ARS) method [30].

To sample $(\alpha + \kappa)$, we use the auxiliary variable sampling [10], and this needs the auxiliary variable $\boldsymbol{m}$ in (10), as proposed in [10].

To sample $\kappa/(\alpha + \kappa)$, we place a vague beta prior $\mathcal{B}(1, 1)$ on it, with a likelihood of $\{\boldsymbol{m}_{ik}^t - \hat{\boldsymbol{m}}_{ik}^t, \forall i, k, t > 1\}$ in (11). The posterior is in an analytical form that can be sampled, owing to its conjugate property.

D. Gibbs Sampling for the MTI Model

The variables of interest are: $\boldsymbol{\beta}$, Z and auxiliary variables $\hat{\boldsymbol{m}}$, where $\hat{\boldsymbol{m}}$ refers to the number of tables having dish k as used in [6] and [10] without counting the tables generated from the sticky portion, that is, κN_{ik}^{t-1}. As the hyperparameters in the MTI model are quite similar to those in [12], we do not present the hyperparameters here. Interested readers can refer to [6], [12], and [13] for the detailed implementation.

1) Sampling $\boldsymbol{\beta}$: $\boldsymbol{\beta}$'s sampling is the same as (1).

2) Sampling s_{ij}^t, r_{ij}^t: The posterior probability of s_{ij}^t, r_{ij}^t is

$$\begin{aligned} &\Pr\left(s_{ij} = k, r_{ij} = l | \alpha, \boldsymbol{\beta}, \kappa, \{N_{\cdot\cdot}^{(i)}\}, \{N_{\cdot\cdot}^{(j)}\}, \boldsymbol{e}, \lambda_1, \lambda_2, Z\right) \\ &\quad \propto \Pr\left(s_{ij}^t = k | \alpha, \boldsymbol{\beta}, \kappa, N_{s_{ij}^{t-1}\cdot}^{(i)}, s_{ij}^{t-1}\right) \\ &\Pr\left(r_{ij}^t = l | \alpha, \boldsymbol{\beta}, \kappa, N_{r_{ij}^{t-1}\cdot}^{(j)}, r_{ij}^{t-1}\right) \\ &\qquad \cdot \Pr\left(\boldsymbol{e}_{ij}^t | \boldsymbol{e} / \{\boldsymbol{e}_{ij}^t\}, s_{ij}^t = k, r_{ij}^t = l, Z / \{s_{ij}^t, r_{ij}^t\}, \lambda_1, \lambda_2\right). \end{aligned} \tag{17}$$

The first term of (17) is

$$\begin{aligned} &\Pr\left(s_{ij}^t = k | \alpha, \boldsymbol{\beta}, \kappa, N_{s_{ij}^{t-1}\cdot}^{(i)}, s_{ij}^{t-1}\right) \\ &\propto \left(\alpha \boldsymbol{\beta}_k + N_{s_{ij}^{t-1}k}^{(i)} + \kappa \delta\left(s_{ij}^{t-1}, k\right)\right) \\ &\quad \cdot \left(\frac{\alpha \boldsymbol{\beta}_{s_{ij}^{t+1}} + N_{ks_{ij}^{t+1}}^{(i)} + k\delta\left(k, s_{ij}^{t+1}\right) + \delta\left(k, s_{ij}^{t-1}\right)\delta\left(k, s_{ij}^{t+1}\right)}{\alpha + N_{k\cdot}^{(i)} + \kappa + \delta\left(s_{ij}^{t-1}, k\right)}\right). \end{aligned} \tag{18}$$

$$\begin{vmatrix} 0.95 & 0.05 & 0 \\ 0.05 & 0.95 & 0.05 \\ 0.05 & 0 & 0.95 \end{vmatrix} \qquad \begin{vmatrix} 0.95 & 0.2 & 0 \\ 0.05 & 0.95 & 0.05 \\ 0.2 & 0 & 0.95 \end{vmatrix} \qquad \begin{vmatrix} 0.05 & 0.95 & 0 \\ 0.05 & 0.05 & 0.95 \\ 0.95 & 0 & 0.05 \end{vmatrix} \qquad \begin{vmatrix} 0.05 & 0.95 & 0 \\ 0.2 & 0.05 & 0.95 \\ 0.95 & 0 & 0.2 \end{vmatrix}$$

Fig. 4. Four cases of the compatibility matrix. Left (Case 1): large diagonal values and small nondiagonal values. Left-middle (Case 2): large diagonal values and mediate nondiagonal values. Right-middle (Case 3): large nondiagonal values and small diagonal values. Right (Case 4): small diagonal values and mediate nondiagonal values.

The second term of (17) is

$$\begin{aligned} &\Pr\left(r_{ij}^t = l|\alpha, \boldsymbol{\beta}, \kappa, N_{r_{ij}^{t-1}\cdot}^{(j)}, r_{ij}^{t-1}\right) \\ &\propto \left(\alpha\boldsymbol{\beta}_l + N_{r_{ij}^{t-1}l}^{(j)} + \kappa\delta\left(r_{ij}^{t-1}, l\right)\right) \\ &\quad\cdot\left(\frac{\alpha\boldsymbol{\beta}_{r_{ij}^{t+1}} + N_{lr_{ij}^{t+1}}^{(i)} + l\delta\left(l, r_{ij}^{t+1}\right) + \delta\left(l, r_{ij}^{t-1}\right)\delta\left(l, r_{ij}^{t+1}\right)}{\alpha + N_{l\cdot}^{(i)} + \kappa + \delta\left(r_{ij}^{t-1}, l\right)}\right). \end{aligned} \tag{19}$$

The likelihood of $\Pr(\boldsymbol{e}_{ij}^t|\boldsymbol{e}/\{\boldsymbol{e}_{ij}^t\}, s_{ij}^t = k, r_{ij}^t = l, Z/\{s_{ij}^t, r_{ij}^t\}, \lambda_1, \lambda_2)$ is the same as (9).

3) Sampling $\hat{\boldsymbol{m}}$: $\hat{\boldsymbol{m}}$ is similar to that in the MTV model; however, it differs in the incorporation of κ

$$\Pr\left(\boldsymbol{m}_{qk}^{(i)} = m|\alpha, \boldsymbol{\beta}_k, \kappa, N_{qk}^{(i)}\right) \propto S\left(N_{qk}^{(i)}, m\right)\left(\alpha\boldsymbol{\beta}_k + \kappa\right) \tag{20}$$

$$\hat{\boldsymbol{m}}_{qk}^{(i)} \sim \text{Binomial}\left(\boldsymbol{m}_{qk}^{(i)}, \frac{\alpha\boldsymbol{\beta}_k}{\kappa + \alpha\boldsymbol{\beta}_k}\right) \tag{21}$$

$$\hat{\boldsymbol{m}}_{\cdot k} = \sum_{q,i}\hat{\boldsymbol{m}}_{qk}^{(i)}. \tag{22}$$

E. Inference Discussions

Both the Gibbs sampling and slice-efficient sampling are two feasible ways to accomplish our task. They have different advantages and disadvantages.

As mentioned previously, Gibbs sampling in our MTV model integrates out the mixed-membership distribution $\{\boldsymbol{\pi}_i^t\}$. It is the marginal approach [31]. The property of community exchangeability makes it simple to implement. However, theoretically, the obtained samples mix slowly as the sampling of each label is dependent on other labels.

Slice-efficient sampling is a conditional approach [28] whereas the membership indicators are independently sampled from $\{\boldsymbol{\pi}_i^t\}$. In each iteration, given $\{\boldsymbol{\pi}_i^t\}$ and the role-compatibility matrix W, we can parallelize the process of sampling membership indicators, which may help to improve the computation, especially when the number of nodes (N) becomes larger, and the number of communities (k) becomes smaller.

VI. Experiments

The performance of our DIM3 model is validated by experiments on both synthetic and real-world datasets. On the synthetic datasets, we implement the finite-communities cases of our models as baseline algorithms, namely as the f-MTV and f-MTI model. On the real-world datasets, we individually implement three benchmark models: MMSB, IRM, and LFRM to the best of our understanding. Also, we compare DRIFT with our models on real-world datasets, and the source code is provided by [26].

A. Synthetic Datasets

For the synthetic data generation, the variables are generated by following [7]. We use $N = 20, T = 3$, and hence E is a $20 \times 20 \times 3$ asymmetric and binary matrix. The parameters are set up in a way so that 20 nodes are equally partitioned into four groups. The ground-truth of the mixed-membership distributions for each of the groups are $[0.8, 0.2, 0; 0, 0.8, 0.2; 0.1, 0.05, 0.85; 0.4, 0.4, 0.2]$.

We consider four different cases to fully assess DIM3 against the ground-truth; all lie in the three-role-compatibility matrix.

The detailed results of the role-compatibility matrix on these four cases are shown in Fig. 4.

1) Markov Chain Monte Carlo Analysis: The convergence behavior is tested in terms of two quantities: the cluster number K, that is, the number of different values Z can take, and the deviance D of the estimated density [28], [31], which is defined as

$$D = -2\sum_{i,j,t}\log\left(\sum_{k,l}\frac{N_{ik}^t \cdot N_{jl}^t}{4n^2T}p(e_{ij}^t|Z, \lambda_1, \lambda_2)\right). \tag{23}$$

In our Markov Chain Monte Carlo (MCMC) stationary analysis, we run five independent Markov chains and discard the first half of the Markov chains as a burn-in. With the random partition of three initial classes as the starting point, 20 000 iterations are conducted in our samplings.

The simulated chains satisfy the standard convergence criteria, when the test was implemented using the CODA package [32]. In Gelman and Rubin's diagnostics [33], the value of the proportional scale reduction factor is 1.09 (with upper C.I. 1.27) for k, 1.03 (with upper C.I. 1.09) for D in the Gibbs sampling, and 1.02 (with upper C.I. 1.06) for k, 1.02 (with upper C.I. 1.02) for D in slice sampling. Geweke's convergence diagnostics [34] are also employed, with the proportion of the first 10% and last 50% of the chain for comparison. The corresponding z-scores are calculated in the interval $[-2.09, 0.85]$ for five chains. In addition, the stationarity and half-width tests of the Heidelberg and Welch Diagnostic [35] are both passed in all cases, with the p-value higher than 0.05. On the basis of all these statistics, the Markov chain's stationarity can be safely ensured in our case.

The efficiency of the algorithms can be measured by estimating the integrated autocorrelation time τ for K and D. 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 is.

Referenece [28] used an estimator $\widehat{\tau}$ as

$$\widehat{\tau} = \frac{1}{2} + \sum_{l=1}^{C-1}\widehat{\rho}_l \tag{24}$$

TABLE II

INTEGRATED AUTOCORRELATION TIMES ESTIMATOR $\widehat{\tau}$ FOR K AND D

		K					D				
Sampling	γ \ α	0.1	0.3	0.5	1	2	0.1	0.3	0.5	1	2
MTV-g	0.1	177.2	93.65	26.91	50.21	11.24	358.8	148.3	23.94	84.75	4.31
	0.3	260.5	54.00	9.18	5.31	6.56	389.5	315.0	3.11	26.32	4.78
	0.5	1.83	8.33	7.54	3.95	5.24	2.88	79.34	90.93	3.17	3.82
	1.0	5.57	6.45	3.44	3.64	4.56	3.19	2.78	1.76	8.14	5.74
	2.0	4.30	2.87	3.35	2.98	3.28	95.48	1.91	3.29	8.74	6.55
MTV-s	0.1	248.6	90.63	161.3	9.58	17.69	8.67	59.90	57.57	1.87	3.70
	0.3	120.6	66.23	44.35	11.40	7.28	29.05	20.64	30.01	45.57	3.40
	0.5	18.99	27.27	6.08	8.76	10.40	39.66	3.87	5.30	3.17	5.83
	1.0	5.79	9.19	11.85	8.46	7.25	40.51	4.85	3.12	6.88	10.51
	2.0	3.17	8.41	5.35	5.48	5.05	25.54	34.82	4.61	35.61	12.68

where $\widehat{\rho}_l$ is the estimated autocorrelation at lag l and C is a cutoff point, which is defined as $C := \min\{l : |\widehat{\rho}_l| < 2/\sqrt{M}\}$, and M is the number of iterations.

We test the sampling efficiency of the MTV-g and MTV-s models on Case 1 with the same setting as [31]. Of the whole 20 000 iterations, the first half of the samples is discarded as a burn-in and the remainder are thinned 1/20. We manually try different values of the hyperparameters γ and α and show the integrated autocorrelation time estimator in Table II. Although some outliers exist, we can see that there is a general trend that, with a fixed α value, the autocorrelation function decreases when the γ value increases. This same phenomenon happens on α while γ is fixed. This result confirms our empirical knowledge. The larger value of γ, α will help to discover more clusters, followed by a smaller autocorrelation function.

On the other hand, we confirm that MTV-g and MTV-s models do not show much difference in the mixing rate of the Markov Chain, as shown in Table II. As mentioned in the previous section, slice sampling provides a mixed-membership distribution-independent sampling scheme, which enjoys the time efficiency of parallel computing in one iteration. For large-scale datasets, it is a feasible solution. In Gibbs sampling, parallel computing is impossible as the sampling variables are in a dependent sequence.

Fig. 5 shows the trace plot of the training log-likelihood against the iterations on Case 1. As we can see, the sampler in the MTI model converges to the high training log-likelihood region faster than the MTV model. Also, the MTI model reaches a higher training log-likelihood than the MTV model.

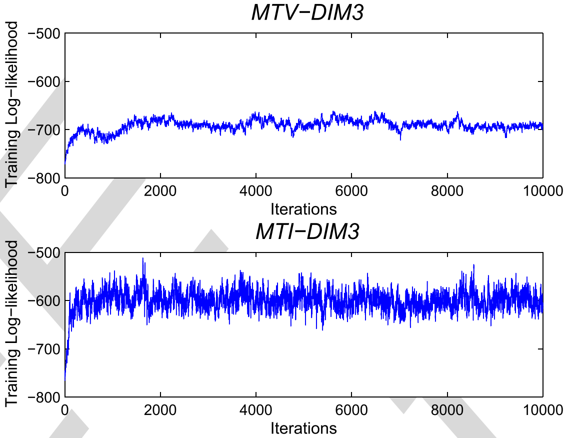


Fig. 5. Top: training log-likelihood trace plot on the MTV-g model. Bottom: training log-likelihood trace plot on the MTI-g model.

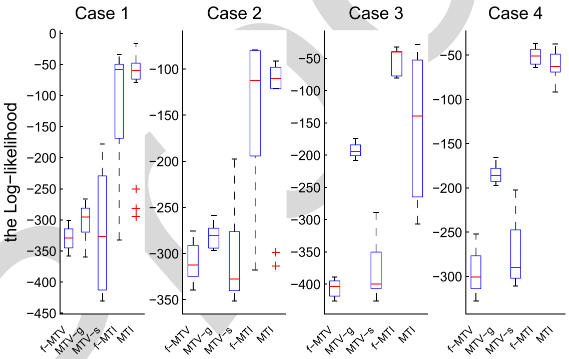


Fig. 6. Log-likelihood performance on all the four cases.

2) Further Performance: We will compare the models in terms of the log-likelihood (Fig. 6); the average l_2 distance between the mixed-membership distributions and its ground-truth; and the l_2 distance between the posterior role-compatibility matrix and its ground-truth (Table III).

From the log-likelihood comparison shown in Fig. 6, we can see that the MTI model performs better than the MTV model in general. On the average l_2 distance to the ground-truth performance, the MTI model also performs better. The superiority of the MTI model's performance over that of the MTV model is within our expectation, as the MTI model describes the membership indicator's time consistency more accurately (i.e., integrating the sticky parameter κ on the specific membership indicator, rather than the mixed-membership distribution). Also, the hidden Markov property enables the MTI model to categorize membership indicators into the same mixed-membership distributions on the basis of its previous value. This seems to be a more effective method than the time-based grouping in the MTV model. However, in situations where there are dramatic changes amongst the membership distributions over time, the MTI model will not respond well. The MTV model is much more effective and robust under these settings as the distribution consistency is a more robust modeling strategy. In addition, the assumption

TABLE III

AVERAGE l_2 DISTANCE TO THE GROUND-TRUTH

Cases	Role-compatibility matrix						Mixed-memberships					
	f-MTV	MTV-g	MTV-s	f-MTI	MTI	MMSB	f-MTV	MTV-g	MTV-s	f-MTI	MTI	MMSB
1	0.239	0.243	0.259	0.114	**0.086**	0.271	0.366	0.384	0.403	0.199	**0.191**	0.411
2	0.206	0.225	0.240	**0.195**	0.204	0.285	0.355	0.355	0.319	**0.207**	0.227	0.398
3	0.134	0.201	0.246	0.117	**0.087**	0.280	0.278	0.289	0.589	0.208	**0.187**	0.329
4	**0.195**	0.214	0.267	0.220	0.219	0.246	0.258	0.285	0.277	0.192	**0.182**	0.310
large size	0.220	0.239	0.215	0.142	**0.059**	0.237	0.243	0.316	0.246	0.147	**0.120**	0.296

TABLE IV

RUNNING TIME (SECONDS PER ITERATION)

N.	f-MTV	MTV-g	MTV-s	f-MTI	MTI	p-MTV-s[1]
20	0.20	0.28	0.23	0.15	0.31	0.29
50	1.03	1.52	1.29	0.95	1.91	1.79
100	3.69	5.76	4.81	3.74	7.49	5.06
200	15.61	24.17	19.87	15.82	30.19	21.64
500	106.96	154.45	119.82	105.61	202.09	132.43
1000	493.44	888.86	642.28	597.29	1102.90	393.24

[1] p-MTV-s denotes the parallel implementation of the MTV-s inference.

that there exist different membership distributions at each time instance makes it possible to parallelize the MTV model to some extent, making it suitable for dealing with large-scale problems.

Here, we compare the computational complexity (running time) of the models in one iteration, with K discovered communities, and show the results in Table IV. We discuss the MTV-g and MTV-s models as an instance. In the MTV-g model, the number of variables to be sampled is $(2K + 2n^2T)$, whereas a total of $(2K + 4n^2T + nT)$ variables are sampled in the MTV-s model. However, the posterior calculation of Z in the MTV-s model can be directly obtained from the mixed-membership distribution, while we need to calculate the ratio for each of Z in the MTV-g model. Also, the U value at each time can be sampled in one operation as its independency in the MTV-s model. The result shows that the MTV-s model runs faster than the MTV-g model, which is in accordance with our assumption.

We also tried a parallel implementation of the slice variables $\{u_{ij,s}^t, u_{ij,r}^t\}_{i,j,t}$'s in the MTV-s model. During each iteration, these slice variables are partitioned into four parts (as our machine has four cores) and are sampled independently, while other variables are still sampled in a sequence. Its corresponding running time is shown in the last column of Table IV. It shows that the parallel design costs even more time when the dataset size is small ($N \leq 500$). This may be due to the time spent on transferring the variables. However, it needs less time when the dataset size becomes larger ($N > 500$). This verifies that our parallel slice sampling method is a promising approach in achieving scalability.

3) Larger Data Size Results: We also conduct the experiments with a larger synthetic dataset ($N = 100$, $T = 20$). With the same construction as previous ones, we increase the role number to 5 and set the role-compatibility matrix as shown in Fig. 7.

$$\left|\begin{array}{ccccc} 0.95 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.1 & 0.9 & 0.1 & 0.1 & 0.1 \\ 0.05 & 0.1 & 0.9 & 0.05 & 0 \\ 0.05 & 0 & 0.05 & 0.9 & 0.15 \\ 0 & 0.05 & 0.1 & 0.05 & 0.9 \end{array}\right|$$

Fig. 7. Larger dataset's role-compatibility matrix.

We set five groups in this network, with the group sizes as $[35, 20, 20, 20, 5]$ and the mixed-membership distributions for each of the groups as $[0.8, 0.1, 0, 0.05, 0.05; 0.02, 0.85, 0.05, 0.03, 0.05; 0.1, 0, 0.9, 0, 0;\ 0.05, 0.1, 0, 0.85, 0; 0, 0.2, 0, 0.4, 0.4]$. The detailed results are also given in Table III. As we can see, our MTI model still achieves the best performance of all the models.

B. Real-World Datasets Performance

We select ten real-world datasets for benchmark testing. Their detailed information, including the number of nodes, the number of edges, edge types, and time intervals, is given in Table VII. Following a general test on the training log-likelihood of the training data and area under the ROC (Receiver Operating Characteristic) curve (AUC) of the test data, we elaborate the results on three selected datasets in the following.

We use a fivefold cross validation method to certify our model's performance on the real-world datasets. The hyperparameters γ, κ, α are sampled according to the sampling strategy mentioned in Section V. Each experiment is run ten times and we report their mean and standard deviation in Tables V and VI.

In these two tables, the bold type denotes the best value in each row. As we can see, our MTI model performs best in eight of the ten datasets on the training log-likelihood and six of the ten datasets on the AUC value. In the remaining datasets, although our MTI model's performance is still quite competitive, the DRIFT model has the best values, possibly because, in these datasets, all associated communities from both nodes are considered in generating the link between these two nodes [14]. The MTV models still do not perform well enough, for the reason previously given. The IRM's results are the worst, which reflects that the simple structure (i.e., each node occupies only one class) may not be enough to capture the full structure in relational learning.

C. Kapferer Tailor Shop

The Kapferer Tailor Shop data [1] records interactions in a tailor shop at two time points. In this time period, the employees in the shop negotiate for higher wages. The dataset

TABLE V

TRAINING LOG-LIKELIHOOD PERFORMANCE (95% CONFIDENCE INTERVAL = MEAN ∓1.96× STANDARD DEVIATION)

Dataset	MTV-g	MTV-s	MTI	MMSB	IRM	LFRM	DRIFT
Kapferer	−673.7 ∓ 15.9	−698.9 ∓ 15.2	**−501.5 ∓ 0.0**	−618.4 ∓ 59.8	−658.6 ∓ 70.3	−865.1 ∓ 70.1	−783.2 ∓ 92.3
Sampson	−347.6 ∓ 23.4	−350.4 ∓ 22.2	**−242.0 ∓ 0.0**	−353.0 ∓ 16.3	−366.8 ∓ 0.6	−332.2 ∓ 16.9	−275.2 ∓ 52.0
Student-net	−1054.4 ∓ 48.5	−1059.3 ∓ 46.2	**−594.3 ∓ 0.0**	−881.4 ∓ 29.9	−1201.2 ∓ 1.6	−1069.6 ∓ 42.2	−905.8 ∓ 46.3
Enron	−2274.2 ∓ 25.6	−2154.4 ∓ 43.3	**−1335.7 ∓ 17.1**	−1512.5 ∓ 6.5	−2264.8 ∓ 26.2	−1742.9 ∓ 36.0	−1492.3 ∓ 13.2
Senator	−897.3 ∓ 16.2	−887.4 ∓ 43.2	**−657.4 ∓ 12.3**	−713.2 ∓ 64.2	−843.6 ∓ 23.5	−673.2 ∓ 43.6	−678.6 ∓ 48.5
DBLP-link	−1923.9 ∓ 19.4	−2124.6 ∓ 26.4	**−1049.6 ∓ 7.5**	−2082.0 ∓ 12.0	−2953.1 ∓ 4.9	−1746.5 ∓ 15.4	−1426.1 ∓ 46.2
Hypertext	−5276.7 ∓ 9.6	−5281.4 ∓ 10.3	**−2923.2 ∓ 0.0**	−4083.5 ∓ 77.8	−5432.7 ∓ 19.6	−3747.5 ∓ 94.3	−3942.3 ∓ 48.5
Newcomb	−1075.0 ∓ 47.6	−1098.1 ∓ 48.0	−876.7 ∓ 0.0	−1835.2 ∓ 14.2	−1965.9 ∓ 1.8	−1203.0 ∓ 14.7	**−789.3 ∓ 63.2**
Freeman	−658.5 ∓ 19.6	−664.1 ∓ 19.2	**−405.2 ∓ 0.0**	−673.5 ∓ 73.9	−728.9 ∓ 66.9	−917.2 ∓ 35.7	−794.2 ∓ 66.2
Coleman	−1500.8 ∓ 63.7	−1532.8 ∓ 64.2	−1003.9 ∓ 0.0	−1302.8 ∓ 130.2	−689.5 ∓ 3.2	−606.7 ∓ 65.1	**−546.1 ∓ 26.9**

TABLE VI

AUC PERFORMANCE (95% CONFIDENCE INTERVAL = MEAN ∓1.96× STANDARD DEVIATION)

Dataset	MTV-g	MTV-s	MTI	MMSB	IRM	LFRM	DRIFT
Kapferer	0.816 ∓ 0.074	0.816 ∓ 0.011	**0.928 ∓ 0.000**	0.893 ∓ 0.001	0.751 ∓ 0.016	0.891 ∓ 0.034	0.905 ∓ 0.013
Sampson	0.804 ∓ 0.000	0.821 ∓ 0.098	**0.927 ∓ 0.000**	0.836 ∓ 0.002	0.738 ∓ 0.005	0.841 ∓ 0.012	0.855 ∓ 0.029
Student-net	0.867 ∓ 0.030	0.877 ∓ 0.095	0.934 ∓ 0.000	0.938 ∓ 0.001	0.809 ∓ 0.004	0.862 ∓ 0.076	**0.949 ∓ 0.015**
Enron	0.834 ∓ 0.097	0.853 ∓ 0.143	0.920 ∓ 0.001	0.907 ∓ 0.013	0.820 ∓ 0.082	0.894 ∓ 0.073	**0.956 ∓ 0.079**
Senator	0.849 ∓ 0.129	0.839 ∓ 0.046	**0.931 ∓ 0.001**	0.880 ∓ 0.022	0.829 ∓ 0.064	0.892 ∓ 0.056	0.925 ∓ 0.076
DBLP-link	0.831 ∓ 0.046	0.816 ∓ 0.017	**0.926 ∓ 0.000**	0.918 ∓ 0.000	0.817 ∓ 0.010	0.891 ∓ 0.062	0.891 ∓ 0.034
Hypertext	0.861 ∓ 0.029	0.843 ∓ 0.027	**0.901 ∓ 0.023**	0.844 ∓ 0.008	0.788 ∓ 0.015	0.853 ∓ 0.042	0.871 ∓ 0.010
Newcomb	0.814 ∓ 0.049	0.795 ∓ 0.090	0.931 ∓ 0.000	0.836 ∓ 0.001	0.765 ∓ 0.013	0.879 ∓ 0.041	**0.960 ∓ 0.027**
Freeman	0.875 ∓ 0.133	0.862 ∓ 0.041	**0.915 ∓ 0.000**	0.867 ∓ 0.001	0.790 ∓ 0.008	0.883 ∓ 0.026	0.897 ∓ 0.022
Coleman	0.891 ∓ 0.067	0.872 ∓ 0.052	0.928 ∓ 0.000	0.928 ∓ 0.001	0.888 ∓ 0.004	0.929 ∓ 0.018	**0.945 ∓ 0.052**

TABLE VII

DATASET INFORMATION

Dataset	Nodes	Edge	Time	Link Type
Kapferer [36]	39	256	2	friends
Sampson [37], [38]	18	168	3	like
Student-net	50	351	3	friends
Enron [39]	151	1980	12	email
Senator [7]	100	5786	8	vote
DBLPlink [40], [41]	100	5706	10	citation
Hypertext [42]	113	7264	10	contact
Newcomb [43]	17	1020	15	contact
Freeman [44]	32	357	2	friends
Coleman [45]	73	506	2	co-work

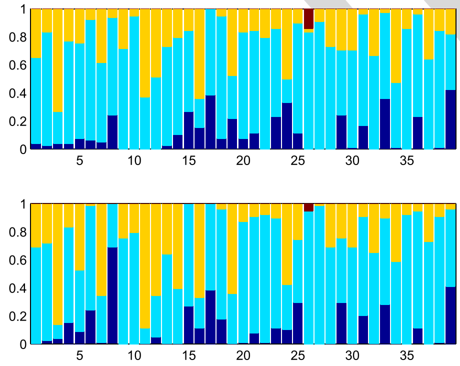


Fig. 8. MTI model's performance on Kapferer Tailor Shop dataset. The x-axis stands for the nodes, while the y-axis represents the mixed-membership distribution. Different colors represent various communities we discovered. Top bar chart: all the employees' mixed-membership distributions in Time 1. Bottom bar chart: all the employees' mixed-membership distributions in Time 2.

is of particular interest because two strikes occur after each time point, with the first failing and the second successful.

We mainly use the work–assistance interaction matrix in the dataset. The employees have eight occupations: head tailor (19), cutter (16), line 1 tailor (1-3, 5-7, 9, 11-14, 21, 24), button machiner (25-26), line 3 tailor (8, 15, 20, 22-23, 27-28), ironer (29, 33, 39), cotton boy (30-32, 34-38), and line 2 tailor (4, 10, 17-18).

In Fig. 8, we can see that the yellow communities at Time 2 are larger than those at Time 1, which means that people tend to have another community at Time 2, rather than being mostly dominated by one large group at Time 1. This larger yellow community may be the result of the first failed strike, after which employees start to shift to the minor (yellow) community for a successful strike.

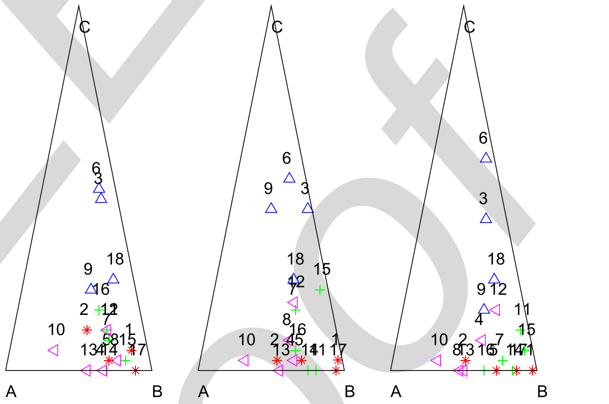


Fig. 9. Nodes' mixed-membership distribution of the MTI model on Sampson Monastery dataset. Left to right: time 1–3. Blue: *loyal opposition*. Red: *outcasts*. Green: *young Turks*. Magenta: *interstitial group*.

$$\begin{vmatrix} 0.09 & 0 & 0.0 \\ 0.05 & 0.99 & 0.02 \\ 0.01 & 0 & 0.96 \end{vmatrix} \quad \begin{vmatrix} 0.01 & 0 & 0.03 \\ 0.02 & 0.78 & 0 \\ 0.02 & 0 & 0.67 \end{vmatrix}$$

Fig. 10. Role-compatibility matrix. Left: MTV-g. Right: MTI.

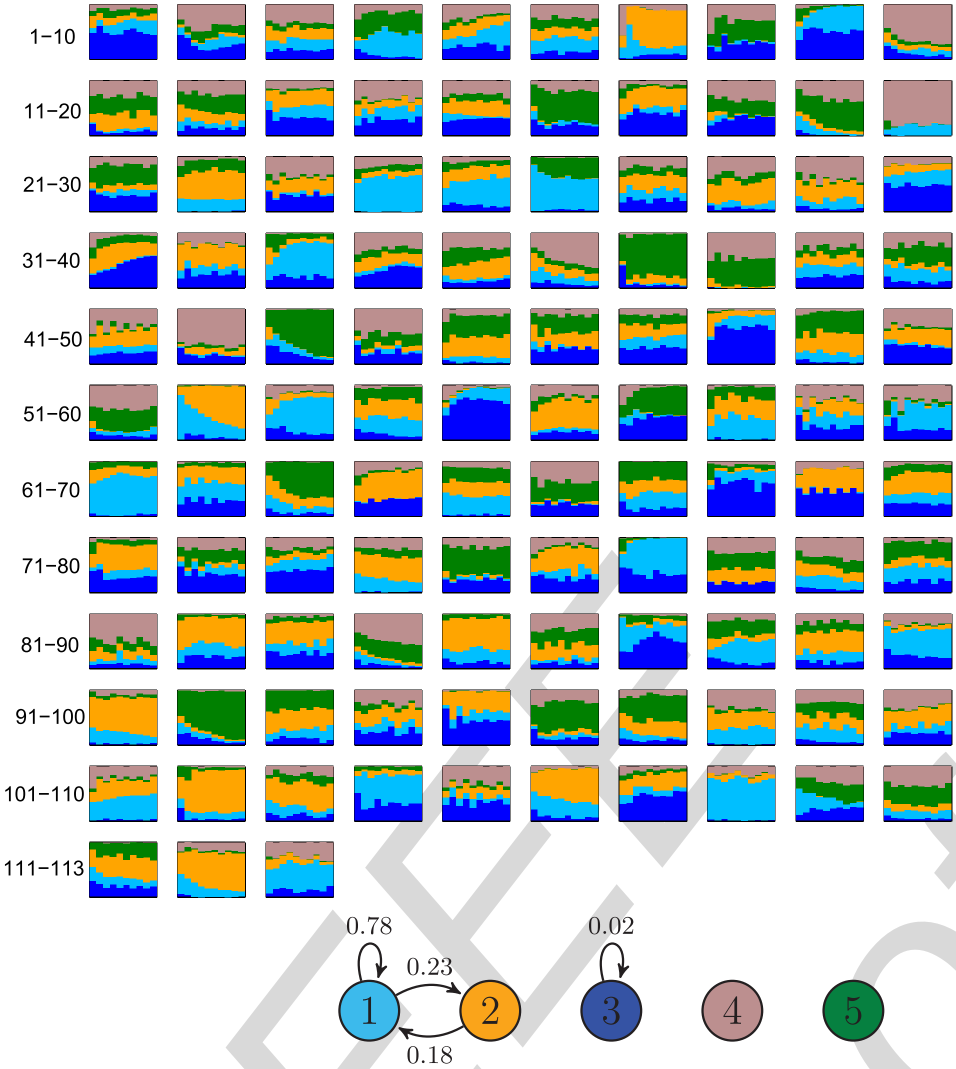


Fig. 11. MTI model's performance on the hypertext 2009 dynamic contact network. Numbers on the left side: orders of nodes. Each bar chart: dynamic behavior of one node's mixed-membership distribution, where the x-axis stands for the ten time stamps. Different colors are interpreted as the communities we have discovered, and their role-compatibility is represented below the bar chart.

D. Sampson Monastery Dataset

The Sampson Monastery dataset is used here to extend the study. There are 18 monks in this dataset, and their social linkage data is collected at three different time points with various interactions. Here, we especially focus on the like-specification. In the like-specification data, each monk selects three monks as his closest friends. In our settings, we mark the selected interactions as 1, otherwise 0. Thus, an $18 \times 18 \times 3$ social network dataset is constructed, with each row having three elements valued at 1.

According to the previous studies in [8] and [23], the monks are divided into four communities: *young Turks*, *loyal opposition*, *outcasts*, and an *interstitial group*.

Fig. 9 shows the detailed results of the MTI model. As three communities have been detected, we put all the results in a two-simplex, in which we denote the communities as A, B, and C. For trajectory convenience, we also color the nodes according to which special group they belong. The results show that these groups behave significantly differently. The *loyal opposition* group lies closer to C, and the *interstitial group* tends to belong to A. Both of their mixed-membership distributions are stable across time. The *outcasts* and *young Turks* groups lie much closer to B.

We also show the role-compatibility matrix in Fig. 10 for comparison. Compared with the results given in [8], our results have larger compatibility values for the same role. Also, the first role's value in our model is 0 versus 0.6 that is reported in [8].

E. Hypertext 2009 Dynamic Contact Network

This dataset [42] is collected from the ACM Hypertext 2009 conference. 113 conference attendees volunteered to wear radio badges that recorded their face-to-face contacts during the conference. The original data is composed of records such as (t, i, j), where t is the communication time and i, j are the attendees' ID. By adaptively partitioning the whole time period into ten parts and noting the interaction data as 1 if communicated during the time stamps, we obtain a $113 \times 113 \times 10$ binary matrix. Fig. 11 shows the dynamic behavior of the nodes' mixed-membership distributions and the corresponding role-compatibility matrix.

The results show that almost half of all the mixed-membership distributions fluctuate during these time stamps. This phenomenon coincides with our common knowledge that people at academic conferences tend to communicate causally. Thus, people's roles may change during different time stamps.

The learned value of the role-compatibility matrix is about the sky blue community, whose intrarole-compatibility value is 0.6932. It has a small probability of interaction with other communities. The other community's compatibility value is almost 0. This might be the reason for sparsity in the interaction data.

Here we specially mention node 108. In the record, this person is always the first to communicate with others on each of the three days. His/her mixed-membership distribution is mainly composed of the sky blue community 1, which indicates he/she could be an organizer of this conference. The other nodes with mixed-membership distribution dominated by community 1, such as nodes 24, 53, 61, all were engaged actively with others according to the record.

Another interesting phenomenon is that the nodes containing the orange community 2 interact with community 1 at a probability of 0.2. This might be an indication that most of the attendees communicated with the organizers for various reasons.

VII. Conclusion

Modeling complex networking behaviors in a dynamic setting is crucial for widespread applications, including social media, social networks, online business, and market dynamic analysis. This challenges the existing learning systems that have limited power to address the dynamics. In this paper, we have provided a generalized and flexible framework to improve the popular MMSB by allowing a network to have infinite types of communities with relationships that change across time periods. By incorporating a time-sticky factor into the mixed-membership distributions, we have realistically modeled the time-correlation among latent labels. Both Gibbs sampling and adapted slice-efficient sampling have been used to infer the desired target distribution. Quantitative analysis on the MCMC's convergence behavior, including the convergence test, autocorrelation function, and so forth, has been provided to demonstrate the inference performance. The results of the experiments verify that our proposed DIM3 is effective in constructing the dynamic mixed-membership distribution and role-compatibility matrix.

Possible future work includes a systematic application of DIM3 to various large real-world social networks. In particular, we are also interested in adapting our model to many atypical applications, for example, where sequences of networks have nonbinary and directional measurements. We will also study many other flexible frameworks for modeling persistence of memberships across time. Lastly, we will perform an extensive study into patterns of joint dynamics of $\{\boldsymbol{\pi}_i^t\}$ to extract meaningful latent information from them. This is done in a setting where the number of components between $\boldsymbol{\pi}_i^t$ and $\boldsymbol{\pi}_i^{t+1}$ may differ.

Recent developments [46], [47] in the large-scale learning of latent space modeling give us more insights for possible future work. These improvements include parsimonious link modeling [46] that reduces the parameter size from $\mathcal{O}(n^2K^2)$ to $\mathcal{O}(n^2K)$, the utilization of the stochastic variational inference method [48], and a triangular representation of networks [49], [47], which could reduce the parameter size to $\mathcal{O}(nK^2)$. Through these, we are hoping to enlarge our model's scalability to millions of nodes and hundreds of communities.

To describe the time dependency, the dependent Dirichlet process (DDP) [50] provides an alternative. Among the various constructions of the DDP [51]–[55], we may construct the DDP by projecting the gamma process into different subspaces and normalizing them individually, through which the overlapping spaces reflect the correlation. Lin *et al.* [56] discuss the intrinsic relationship between the Poisson process, gamma process and Dirichlet process and uses three operations namely *superposition*, *subsampling*, and *point transition* to evolve from one Dirichlet process to another, with an elegant and solid theory support. Subsequent literatures including [57]–[59] extend this paper from different perspectives.

References

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

[2] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *Proc. 21st Nat. Conf. Artif. Intell. (AAAI)*, Jul. 2006, pp. 381–388. [Online]. Available: http://www.aaai.org/Library/AAAI/2006/aaai06-061.php

[3] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin, "Detecting communities and their evolutions in dynamic social networks—A Bayesian approach," *Mach. Learn.*, vol. 82, no. 2, pp. 157–189, 2011.

[4] K. Ishiguro, T. Iwata, N. Ueda, and J. B. Tenenbaum, "Dynamic infinite relational model for time-varying relational data analysis," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2010, pp. 919–927.

[5] 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, Jan. 2008.

[6] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for systems with state persistence," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 312–319.

[7] Q. Ho, L. Song, and E. P. Xing, "Evolving cluster mixed-membership blockmodel for time-evolving networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 342–350.

[8] E. Xing, W. Fu, and L. Song, "A state-space mixed membership blockmodel for dynamic network tomography," *Ann. Appl. Statist.*, vol. 4, no. 2, pp. 535–566, 2010.

[9] C. Heaukulani and Z. Ghahramani, "Dynamic probabilistic models for latent feature propagation in social networks," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 275–283.

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

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

[12] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "A sticky HDP-HMM with application to speaker diarization," *Ann. Appl. Statist.*, vol. 5, no. 2A, pp. 1020–1056, 2011.

[13] E. Fox, E. B. Sudderth, M. I. Jordan, and A. Willsky, "Bayesian nonparametric inference of switching dynamic linear models," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1569–1585, Apr. 2011.

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

[15] T. L. Griffiths and Z. Ghahramani, “The Indian buffet process: An introduction and review,” *J. Mach. Learn. Res.*, vol. 12, pp. 1185–1224, Jul. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2021026.2021039

[16] Y. W. Teh, D. Görür, and Z. Ghahramani, “Stick-breaking construction for the Indian buffet process,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2007, pp. 556–563.

[17] J. Zhu, “Max-margin nonparametric latent feature models for link prediction,” in *Proc. 29th Int. Conf. Mach. Learn. (ICML)*, 2012, pp. 719–726.

[18] T. Jebara, “Maximum entropy discrimination,” in *Machine Learning*. New York, NY, USA: Springer-Verlag, 2004, pp. 61–98.

[19] K. Palla, D. A. Knowles, and Z. Ghahramani, “An infinite latent attribute model for network data,” in *Proc. 29th Int. Conf. Mach. Learn.*, Edinburgh, Scotland, Jun. 2012, pp. 1607–1614.

[20] P.-S. Koutsourelakis and T. Eliassi-Rad, “Finding mixed-memberships in social networks,” in *Proc. AAAI Spring Symp. Social Inf. Process.*, 2008, pp. 48–53.

[21] Q. Ho, A. P. Parikh, and E. P. Xing, “A multiscale community blockmodel for network exploration,” *J. Amer. Statist. Assoc.*, vol. 107, no. 499, pp. 916–934, 2012. [Online]. Available: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2012.682530

[22] 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, p. 7, Jan. 2010.

[23] D. I. Kim, M. Hughes, and E. Sudderth, “The nonparametric metadata dependent relational model,” in *Proc. 29th Annu. Int. Conf. Mach. Learn.*, Jun. 2012, pp. 1559–1566.

[24] P. Sarkar and A. W. Moore, “Dynamic social network analysis using latent space models,” *ACM SIGKDD Explorations Newslett.*, vol. 7, no. 2, pp. 31–40, Dec. 2005.

[25] P. Sarkar, S. M. Siddiqi, and G. J. Gordon, “A latent space approach to dynamic embedding of co-occurrence data,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2007, pp. 420–427.

[26] J. R. Foulds, C. DuBois, A. U. Asuncion, C. T. Butts, and P. Smyth, “A dynamic relational infinite feature model for longitudinal social networks,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 287–295.

[27] W. Fu, L. Song, and E. P. Xing, “Dynamic mixed membership blockmodel for evolving networks,” in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 329–336.

[28] M. Kalli, J. E. Griffin, and S. G. Walker, “Slice sampling mixture models,” *Statist. Comput.*, vol. 21, no. 1, pp. 93–105, Jan. 2011.

[29] S. G. Walker, “Sampling the Dirichlet mixture model with slices,” *Commun. Statist. Simul. Comput.*, vol. 36, no. 1, pp. 45–54, 2007.

[30] C. E. Rasmussen, “The infinite Gaussian mixture model,” in *Advances in Neural Information Processing Systems*, vol. 12. Cambridge, MA, USA: MIT Press, 2000, pp. 554–560.

[31] 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.

[32] M. Plummer, N. Best, K. Cowles, and K. Vines, “CODA: Convergence diagnosis and output analysis for MCMC,” *R News*, vol. 6, no. 1, pp. 7–11, 2006.

[33] A. Gelman and D. B. Rubin, “Inference from iterative simulation using multiple sequences,” *Statist. Sci.*, vol. 7, no. 4, pp. 457–472, 1992.

[34] J. Geweke, “Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments,” in *Bayesian Statistics*, Oxford, U.K.: Oxford Univ. Press, 1992, pp. 169–193.

[35] P. Heidelberger and P. D. Welch, “A spectral method for confidence interval generation and run length control in simulations,” *Commun. ACM*, vol. 24, no. 4, pp. 233–245, Apr. 1981.

[36] B. Kapferer, *Strategy and Transaction in an African Factory: African Workers and Indian Management in a Zambian Town*. Manchester, U.K.: Manchester Univ. Press, 1972.

[37] S. F. Sampson, “Crisis in a cloister,” Ph.D. dissertation, Dept. Sociology, Cornell Univ., Ithaca, NY, USA, 1969.

[38] R. L. Breiger, S. A. Boorman, and P. Arabie, “An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling,” *J. Math. Psychol.*, vol. 12, no. 3, pp. 328–383, Aug. 1975.

[39] B. Klimt and Y. Yang, “The Enron corpus: A new dataset for email classification research,” in *Machine Learning: ECML*. Berlin, Germany: Springer-Verlag, 2004, pp. 217–226.

[40] S. Asur, S. Parthasarathy, and D. Ucar, “An event-based framework for characterizing the evolutionary behavior of interaction graphs,” *ACM Trans. Knowl. Discovery Data*, vol. 3, no. 4, p. 16, Nov. 2009.

[41] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng, “Analyzing communities and their evolutions in dynamic social networks,” *ACM Trans. Knowl. Discovery Data*, vol. 3, no. 2, p. 8, Apr. 2009.

[42] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, “What’s in a crowd? Analysis of face-to-face behavioral networks,” *J. Theoretical Biol.*, vol. 271, no. 1, pp. 166–180, 2011.

[43] T. M. Newcomb, *The Acquaintance Process*. New York, NY, USA: Holt, Rinehart & Winston, 1961.

[44] S. C. Freeman and L. C. Freeman, “The networkers network: A study of the impact of a new communications medium on sociometric structure,” School Social Sci., Univ. California, San Francisco, CA, USA, Tech. Rep. 46, 1979.

[45] J. S. Coleman, *Introduction to Mathematical Sociology*. New York, NY, USA: MacMillan, 1964.

[46] P. Gopalan, S. Gerrish, M. Freedman, D. M. Blei, and D. M. Mimno, “Scalable inference of overlapping communities,” in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2012, pp. 2249–2257.

[47] J. Yin, Q. Ho, and E. Xing, “A scalable approach to probabilistic latent space inference of large-scale networks,” in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2013, pp. 422–430.

[48] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, “Stochastic variational inference,” *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1303–1347, 2013.

[49] D. R. Hunter, S. M. Goodreau, and M. S. Handcock, “Goodness of fit of social network models,” *J. Amer. Statist. Assoc.*, vol. 103, no. 481, pp. 248–258, 2008.

[50] S. N. MacEachern, “Dependent nonparametric processes,” in *Proc. Sec. Bayesian Statist. Sci.*, Alexandria, VA, USA, 1999, pp. 50–55.

[51] F. Caron, M. Davy, and A. Doucet, “Generalized Polya urn for time-varying Dirichlet process mixtures,” in *Proc. Uncertainty Artif. Intell.*, 2007, pp. 33–40.

[52] Y. Chung and D. B. Dunson, “The local Dirichlet process,” *Ann. Inst. Statist. Math.*, vol. 63, no. 1, pp. 59–80, 2011.

[53] D. B. Dunson, “Bayesian dynamic modeling of latent trait distributions,” *Biostatistics*, vol. 7, no. 4, pp. 551–568, 2006.

[54] N. Bouguila and D. Ziou, “A Dirichlet process mixture of generalized Dirichlet distributions for proportional data modeling,” *IEEE Trans. Neural Netw.*, vol. 21, no. 1, pp. 107–122, Jan. 2010.

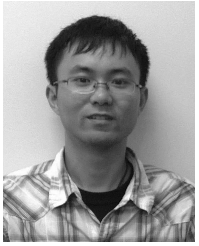
[55] V. Rao and Y. W. Teh, “Spatial normalized gamma processes,” in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2009, pp. 1554–1562.

[56] D. Lin, E. Grimson, and J. W. Fisher, III, “Construction of dependent Dirichlet processes based on Poisson processes,” in *Advances in Neural Information Processing Systems Foundation*. Cambridge, MA, USA: MIT Press, 2010.

[57] D. Lin and J. W. Fisher, “Coupling nonparametric mixtures via latent Dirichlet processes,” in *Advances in Neural Information Processing Systems*, vol. 25. Cambridge, MA, USA: MIT Press, 2012, pp. 55–63.

[58] C. Chen, N. Ding, and W. L. Buntine, “Dependent hierarchical normalized random measures for dynamic topic modeling,” in *Proc. 29th Int. Conf. Mach. Learn.*, Jun. 2012, pp. 895–902.

[59] C. Chen, V. Rao, W. Buntine, and Y. W. Teh, “Dependent normalized random measures,” in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 969–977.



Xuhui Fan received the bachelor’s degree in mathematical statistics from the University of Science and Technology of China, Hefei, China, in 2010. He is currently pursuing the Ph.D. degree with the University of Technology at Sydney, Sydney, NSW, Australia.

His current research interests include statistical machine learning.

Longbing Cao (SM'06) received the Ph.D. degrees in pattern recognition and intelligent systems and computing sciences.

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

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 at Sydney (UTS), Sydney, 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.

Dynamic Infinite Mixed-Membership Stochastic Blockmodel

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

***Abstract*— Directional and pairwise measurements are often used to model interactions in a social network setting. The *mixed-membership stochastic blockmodel* (MMSB) was a seminal work in this area, and its ability has been extended. However, models such as MMSB face particular challenges in modeling dynamic networks, for example, with the unknown number of communities. Accordingly, this paper proposes a *dynamic infinite mixed-membership stochastic blockmodel*, a generalized framework that extends the existing work to potentially infinite communities inside a network in dynamic settings (i.e., networks are observed over time). Additional model parameters are introduced to reflect the degree of persistence among one's memberships at consecutive time stamps. Under this framework, two specific models, namely *mixture time variant* and *mixture time invariant* models, are proposed to depict two different time correlation structures. Two effective posterior sampling strategies and their results are presented, respectively, using synthetic and real-world data.**

***Index Terms*— Bayesian nonparametric, dynamic, Gibbs sampling, Markov Chain Monte Carlo (MCMC) inference, mixed-membership stochastic blockmodel (MMSB), slice sampling.**

I. Introduction

NETWORKING applications with dynamic settings (i.e., networks observed over time) are widely seen in real-world environments, such as link prediction and community detection in social networks, social media interactions, capital market movements, and recommender systems. A deep understanding of such dynamic network mechanisms relies on latent relation analysis and latent variable modeling of dynamic network interactions and structures. This presents both challenges and opportunities to existing learning theories. The intricacy associated with the time-varying attributes makes learning and inference a difficult task, but at the same time, one can explore the evolutionary behavior of a network structure more realistically in this time-varying setting. The various dynamic characteristics of such a network can therefore be revealed in real applications.

A number of researchers have recently attempted to address this issue. Some notable earlier examples include *stochastic blockmodel* [1] and its infinite community case *infinite relational model* (IRM) [2] where the aim is to partition a network of nodes into different groups on the basis of their pairwise and directional binary interactions. It was extended in [3] to infer the evolving community's behavior over time. Their work assumes that a fixed number of K communities exist to which one node can potentially belong. However, in many applications, an accurate estimate of K beforehand may be impractical and its value may also vary during time stamps.

A *dynamic* IRM [4] is an alternative way to address the same problem, where K can be inferred from data itself. However, just as described in [2], its drawback is that this model assumes each node i must belong to only one single community. Therefore, an interaction between nodes i and j can only be determined by their community indicators. This approach can be inflexible in many scenarios, such as the monastery example depicted in [5], where one monk can belong to different communities. To this end, Airoldi *et al.* [5] introduce the concept of mixed-membership, where they assume each node i might belong to multiple communities. The membership indicators of one's interaction are no longer a fixed value of a special community. Instead, they are sampled from the nodes' mixed-membership distributions.

The aforementioned work addresses some aspects (such as infinite, dynamic, mixed-membership, and data-driven inference) of relational modeling. An emergent need is to effectively unify these models to provide a flexible and generalized framework which can encapsulate the advantages of most of this paper and address multiple aspects of complexities in one model. This is certainly not an easy thing to do because of the need to understand the relations among aspects and to build a seamless approach to aggregate the challenges. Accordingly, we propose a *dynamic infinite mixed-membership stochastic blockmodel* (DIM3).

DIM3 has the following features: 1) it allows a network to have an infinite number of latent communities; 2) it allows mixed-membership associated with each node; 3) the model adapts to dynamic settings and the number of communities varies with the time; and 4) it is apparent that in many social networking applications, a node's membership may become consistent (i.e., unchanged) over consecutive time stamps. For example, a person's opinion of a peer is more likely to be consistent in two consecutive time stamps.

Manuscript received July 31, 2013; revised August 28, 2014 and November 4, 2014; accepted November 5, 2014. This work was supported by the Australian Research Council Discovery Project under Grant DP130102691.

X. Fan and L. Cao are with the Faculty of Engineering and Information Technology, Advanced Analytics Institute, University of Technology, Sydney, NSW 2007, Australia (e-mail: xuhui.fan@student.uts.edu.au; long-bing.cao@uts.edu.au).

R. Y. D. Xu is with the Faculty of Engineering and Information Technology, University of Technology, Sydney, NSW 2007, Australia (e-mail: yida.xu@uts.edu.au).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author.

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/TNNLS.2014.2369374

To model this persistence, we devise two different implementations. The first is to have a single mixed-membership distribution for each node at different time intervals. The persistence factor is dependent on the statistics of each node's interactions with the rest of the nodes. The second implementation is to allow a set of mixed-membership distributions to associate with each node, and they are time-invariant. The number of elements in the set varies nonparametrically, as reported in [6]. The persistence factor is dependent on the value of the membership indicator at the previous time stamp.

Consequently, two effective sampling algorithms are designed for our proposed models, using either the Gibbs or slice sampling technique for efficient model inference. Their convergence behavior and mixing rate are analyzed and displayed in the first part of the experiment. In the experimental analysis, we show that we can assess nodes' positions in the network and their developing trends, predict unknown links according to the current structure, understand the network structure and identify change points. The techniques proposed can be used for forecasting the political tendencies of senators [7], predicting the function of a protein in biology [8], and tracking authors' community cooperation in academic circles [9].

The rest of the article is organized as follows. Section II introduces the preliminary knowledge for our work, including a brief introduction to mixed-membership stochastic blockmodel (MMSB) and Dirichlet processes. Section III details our main framework and explains how it can incorporate infinite communities in a dynamic setting. The related work is reviewed in Section IV. The inference schemes for the two proposed models are detailed in Section V. In Section VI, we show the experimental results of the proposed models by using both synthetic and real-world social network data. The conclusion is given in Section VII.

II. Preliminary Knowledge

A. Notations

For notational clarity, we first define the key terms and their meanings, as shown in Table I.

B. Introduction to MMSB and Bayesian Nonparametrics

1) Mixed-Membership Stochastic Blockmodel: MMSB [5] aims to model each node's individual mixed-membership distribution. In MMSB, each interaction e_{ij} corresponds to two membership indicators: s_{ij} from the sender i and r_{ij} to the receiver j (w.l.o.g. (Without Loss Of Generality), we assume $s_{ij} = k$, $r_{ij} = l$). The interaction's value is determined by the compatibility of two corresponding communities k and l. Fig. 1 shows the graphical model, and the detailed generative process can be described as:

1) $\forall \{k, l\} \in \mathcal{N} > 0$, draw the communities' compatibility values $W_{k,l} \sim Beta(\lambda_1, \lambda_2)$;
2) $\forall i \in \{1, \cdots, n\}$, draw node i's mixed-membership distribution $\pi_i \sim Dirichlet(\beta)$;
3) $\forall \{i, j\} \in \{1, \cdots, n\}^2$, for interaction 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 interaction $e_{ij} \sim$ Bernoulli $(W_{s_{ij},r_{ij}})$.

TABLE I

Notations for DIM3

n	number of nodes
K	number of discovered communities
T	number of whole time stamps
t	the specific time stamp
e_{ij}^t	directional, binary interactions at time t
$\boldsymbol{\beta}$	a stick-breaking representation to denote the "significance" of all existing communities at all times
γ, α	concentration parameters for HDP
κ	a sticky parameter representing the time-persistence effect
s_{ij}^t	sender's (from i to j) membership indicator at time t
r_{ij}^t	receiver's (from j to i) membership indicator at time t
Z	all the membership indicators, i.e. $Z = \{s_{ij}^t, r_{ij}^t\}_{i,j,t}$
$z_{i\cdot}^t$	node i's membership indicators at time t, i.e. $\{s_{ij}^t, r_{ji}^t\}_{j=1}^n$
$\boldsymbol{m}_{ik}^t$	in the Chinese Restaurant Franchise analogy, the number of tables having dish k at restaurant i and time t
π_i^t	mixed-membership distribution for node i at time t, it generates $s_{\boldsymbol{i}1}^t, \cdots, s_{\boldsymbol{i}n}^t, r_{1\boldsymbol{i}}^t, \cdots, r_{n\boldsymbol{i}}^t$
π_{ik}^t	the "significance" of community k for node i at time t
W	role-compatibility matrix
$W_{k,l}$	compatibilities between communities k and l
$n_{k,l}^t$	number of links from communities k to l at time t i.e. $n_{k,l}^t = \#\{ij : s_{ij}^t = k, r_{ij}^t = l.\}$
$n_{k,l}^{t,1}$	part of $m_{k,l}$ where the corresponding $e_{ij}^t = 1$ at time t, i.e. $n_{k,l}^{t,1} = \sum_{s_{ij}^t=k, r_{ij}^t=l} e_{ij}^t$
$n_{k,l}^{t,0}$	part of $m_{k,l}$ where the corresponding $e_{ij}^t = 0$ at time t, i.e. $n_{k,l}^{t,0} = n_{k,l}^t - n_{k,l}^{t,1}$
N_{ik}^t	number of times that a node i has participated in community k (either sending or receiving message) at time t, i.e. $N_{ik}^t = \#\{j : s_{\boldsymbol{i}j}^t = k\} + \#\{j : r_{j\boldsymbol{i}}^t = k\}$

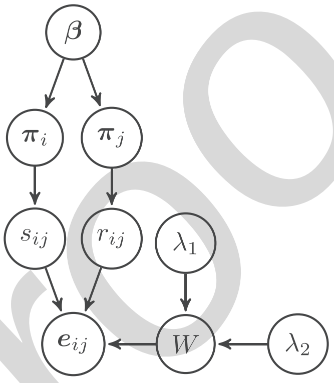


Fig. 1. MMSB model.

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

W is the communities' compatibility matrix as described previously. The prior $P(W)$ is elementwise beta distributed, which is a conjugate to the Bernoulli distribution $P(e_{ij}|.)$. Therefore, a marginal distribution of $P(e_{ij})$, that is, $\int_W p(e_{ij}|W)p(W)d(W)$ can be obtained on the basis of data analysis, and hence there is no need to explicitly sample the values of W.

2) Bayesian Nonparametrics: In the dynamic setting, the Bayesian nonparametric method is a perfect tool for allowing the communities' numbers to vary across time periods. In our case, we use variants of the hierarchical Dirichlet

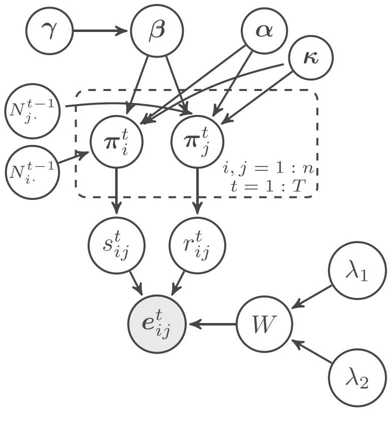


Fig. 2. MTV model.

process (HDP) [10] to model the mixed-membership distribution $\{\pi_i\}_{i=1}^n$, where $\forall i \in \{1,\dots,n\}, \pi_i \sim \mathrm{DP}(\alpha, \boldsymbol{\beta})$ and $\boldsymbol{\beta}$ is generated from a stick-breaking construction $\boldsymbol{\beta} = \sum_{k=1}^{\infty} \boldsymbol{\beta}_k \delta_k, \boldsymbol{\beta}_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l), \beta'_l \sim Beta(1,\gamma))$ [11].

III. Dynamic Infinite Mixed-Membership Stochastic Blockmodel

A. General Settings

In DIM3, we allow each node's membership indicators to change across time periods. Additionally, it is imperative that these indicators should contain the time-persistence property with past values, through which the reality of social behavior can be reflected. Here, we use the strategy of incorporating a sticky parameter κ into the mixed-membership distributions to overcome this issue [6], [12]. Different detailed designs are proposed for the mixture time variant (MTV) and mixture time invariant (MTI) models; however, the common idea is that the current mixed-membership distributions are influenced by the corresponding distributions at the previous time.

Once the current mixed-membership distributions have been selected, the interaction data is generated in the same way as MMSB. Thus, this paper is focused on the details of mixed-membership distribution constructions following the main route of the HDP [10]. Also, we should note that the intermediate variable $\boldsymbol{\beta}$ is identical for both models, representing the significance of all the communities across time periods, and its construction is the same as the stick-breaking construction as described in Section II-B2.

B. Mixture Time Variant (MTV) Model

Fig. 2 shows the graphical model of the MTV model. Here we only show all the variables involved for time t, and omit those for the other time points, where the structure is identical at any other time $\tau \neq t$.

Let us focus on the mixed-membership distribution's construction in the MTV model, which is

$$\boldsymbol{\pi}_i^t \sim DP\left(\alpha+\kappa, \frac{\alpha\boldsymbol{\beta} + \frac{\kappa}{2n}\cdot\sum_k N_{ik}^{t-1}\boldsymbol{\delta}_k}{\alpha+\kappa}\right) \quad (1)$$

$$s_{ij}^t \sim \boldsymbol{\pi}_i^t, r_{ij}^t \sim \boldsymbol{\pi}_j^t \quad \forall i,j \in \mathcal{N}, t \geq 1. \quad (2)$$

The mixed-membership distribution $\{\boldsymbol{\pi}_i^t\}_{1:n}^{1:T}$ is sampled from the Dirichlet process with a concentration parameter $(\alpha+\kappa)$ and a base measure $(\alpha\boldsymbol{\beta} + \frac{\kappa}{2n}\sum_k N_{ik}^{t-1}\boldsymbol{\delta}_k/\alpha+\kappa)$. There will be $N \times T$ of these distributions. They jointly describe each node's activities.

In the base measure, the introduced sticky parameter κ stands for each node's time influence on its mixed-membership distribution. In other words, we assume that each node's mixed-membership distribution at time t will be largely influenced by its activities at time $t-1$. This is reflected in the hidden label's multinomial distribution whereby the previous explicit activities will occupy a fixed proportion $\kappa/\alpha+\kappa$ of the current distribution. The larger the value of κ, the more weight the activities at $t-1$ will have at time t.

As our method is largely based on the HDP framework, we use the popular Chinese Restaurant Franchise (CRF) [6], [10] analogy to explain our model. Using the CRF analogy, the mixed-membership distribution associated with a node i at time t can be seen as a restaurant $\boldsymbol{\pi}_i^t$, with its dishes representing the communities. If a customer s_{ij}^t(or r_{ji}^t) eats the dish k at the ith restaurant at time t, then $s_{ij}^t(r_{ji}^t) = k$. For all $t > 1$, the restaurant $\boldsymbol{\pi}_i^t$ will have its own specials on the dishes served, representing the sticky configuration in the graphical model. In contrast to the sticky HDP–hidden Markov model (HMM) [6] approach, which places emphasis on one dish only, we allow multiple specials in our work, where the weight of each special dish is adjusted according to the number of dishes served at this restaurant at time $t-1$, that is, $(\kappa/2n)\sum_k N_{ik}^{t-1}\boldsymbol{\delta}_k$. Therefore, we can ensure that the special dishes are served persistently across time in the same restaurant.

C. Mixture Time Invariant (MTI) Model

We show the MTI model in Fig. 3. Here we only show the interaction e_{ij}^1 and omit the other interactions, whose structure is directly derived.

The $\boldsymbol{\beta}$ in the MTI model is identical to that in the MTV model, and we sample the mixed-membership distribution and membership indicators as follows:

$$\boldsymbol{\pi}_i^{(k)} \sim \mathrm{DP}\left(\alpha+\kappa, \frac{\alpha\boldsymbol{\beta}+\kappa\boldsymbol{\delta}_k}{\alpha+\kappa}\right) \quad \forall i,k \in \mathcal{N} \quad (3)$$

$$s_{ij}^t \sim \boldsymbol{\pi}_i^{\left(s_{ij}^{t-1}\right)}, r_{ij}^t \sim \boldsymbol{\pi}_j^{\left(r_{ij}^{t-1}\right)} \quad \forall i,j \in \mathcal{N}, t \geq 1. \quad (4)$$

We assign uninformative priors on sampling the initial membership indicators $\{s_{ij}^0, r_{ij}^0\}_{i,j}$, that is, $\{s_{ij}^0, r_{ij}^0\}_{i,j}$ are sampled from a multinomial distribution, with each category having an

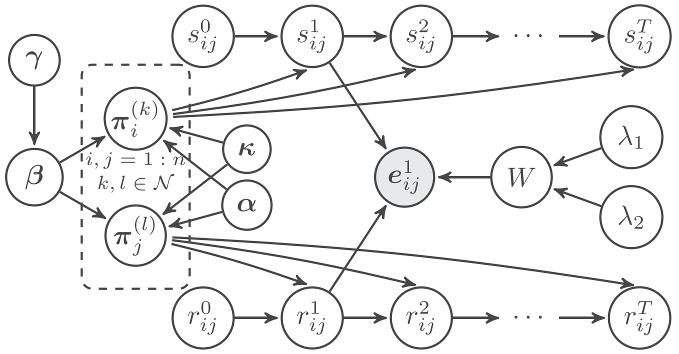


Fig. 3. MTI model.

equalized success probability. The dimension of this multinomial distribution is automatically adjusted according to the current number of communities in the model.

On each node's membership distribution, our MTI model is essentially a Sticky HDP–HMM [6], [12], [13]. In this model, each node has a variable number of mixed-membership distributions associated with it, which may be infinite. At time $t \geq 2$, its membership indicator s_{ij}^t (or r_{ij}^t) is generated from $\boldsymbol{\pi}_i^{(s_{ij}^{t-1})}$ (or $\boldsymbol{\pi}_j^{(r_{ij}^{t-1})}$). To encourage persistence, each π_{ik} is generated from the corresponding $\boldsymbol{\beta}$, where κ is added to $\boldsymbol{\beta}$'s kth component [6], [12], [13].

Returning to the CRF [10] analogy, we have $N \times \infty$ matrix, where its (i, k)th element refers to $\pi_i^{(k)}$, which can be seen as the weights of eating each of the available dishes. A customer s_{ij}^t (or r_{ji}^t) can therefore only travel between restaurants located at the ith row of the matrix. When $\pi_i^{(k)}$'s kth component is more likely to be larger, it means that the dish k is a special dish for restaurant k. Therefore, a customer at restaurant k at time $t-1$ is more likely to eat the same dish (i.e., kth dish), and hence to stay at restaurant k at time t.

D. Discussion and Comparison

Here, we discuss the difference between the two models in the design of the time-persistence property. The MTV model allows the mixed-membership distribution itself to change over time stamps. However, there is only a single (but different) distribution for each node at each individual time stamp. The membership indicator of a node at time t is dependent on the statistics of all membership indicators of the same node at $t-1$ and $t+1$. With a larger value of the sticky parameter κ, the current mixed-membership distribution tends to be more similar to that of the previous time stamp.

In contrast, the MTI model requires the mixed-membership distributions to stay invariant over time. However, there may be an infinite number of possible distributions associated with each node, due to a HDP prior, often only a few distributions will be discovered. In this case, the membership indicator at the current time is dependent and more likely to have the same value as it has in the previous time stamp.

IV. Related Work

We here provide a detailed review of some of the current state-of-the-art in relational learning and at the same time, distinguish our paper from existing ones. In general, we categorize the relational learning models into two major frameworks: the *latent feature model* (LFM) and *latent class model* (LCM). Both frameworks assume that a node's interaction is a Bernoulli draw, which is parameterized by an entry from the role-compatibility matrix. Their main difference is hence in the way the entry is indexed. For LCM, it is assumed that the indices for each pair of nodes are derived from the two associated hidden class labels; in case of LFM, it is assumed that the indices are, however, determined from a set of latent features associated with the pair of nodes.

A representative work for LFM is the *latent feature relational model* (LFRM) [14], which uses a latent feature matrix and a corresponding link generative function to define the model. To account for the variable number of features associated with each node, it uses the Indian Buffet Process [15], [16] as a prior. The *max-margin latent feature relational model* (Med-LFRM) [17] uses the *maximum entropy discrimination* (MED) [18] technique to minimize the hinge loss which measures the quality of link prediction. The *infinite latent attribute* (ILA) model [19] uses a Dirichlet process to construct a substructure within each feature, and all the features are used through the LFRM model.

On the LCM front, the classical approach is the MMSB which enables each node to be associated with multiple membership indicators, and an interaction is formed using one of these indicators. Several variants are subsequently proposed from MMSB, with examples including [20] which extends the MMSB into the infinite communities case [21], which uses the nested Chinese Restaurant Process [22] to build a communities' hierarchical structure, and [23] which incorporates the node's attribute information into its membership indicator construction in MMSB.

Like any data modeling problem, interaction data may also change over time; therefore, dynamic extensions are found in both the LCM and LFM frameworks. Examples such as [24] and [25] describe the time dependency by using Gaussian linear motion models. The *dynamic relational infinite feature model* (DRIFT) [26], which employs an independent Markov dynamic transition matrix to correlate consecutive time interaction data, is a natural extension of the LFRM. *Latent feature propagation* (LFP) [9] directly integrates observed interactions, rather than the latent feature matrix, in the current time to model the distribution of latent features at the next time stamp. On the dynamic setting of MMSB, Xing *et al.* [8] and Fu *et al.* [27] place a parameter (the mean)-dependent Gaussian distribution to consider the time correlation, whereas Ho *et al.* [7] consider hierarchical communities modeling that evolves. However, as both of these two models require predefinition of the number of communities, additional techniques, such as cross-validation, are necessary when choosing the number of communities. Furthermore, their implicit description of the time dependency may not be sufficiently intuitive.

V. Inference

Two sampling schemes are implemented to complete the inference on the MTV model: standard Gibbs sampling and slice-efficient sampling, which both target the same posterior distribution.

A. Gibbs Sampling for the MTV Model

The Gibbs sampling scheme is largely based on [10]. The variables of interest are: $\boldsymbol{\beta}$, Z and auxiliary variables $\hat{\boldsymbol{m}}$, where $\hat{\boldsymbol{m}}$ refers to the number of tables having dish k as in [6] and [10] without counting the tables that are generated from the sticky portion, that is, κN_{ik}^{t-1}. Note that we do not sample $\{\boldsymbol{\pi}_i^t\}_{1:n}^{1:T}$, as it gets integrated out.

1) Sampling $\boldsymbol{\beta}$: $\boldsymbol{\beta}$ is the prior for all $\{\boldsymbol{\pi}_i^t\}s$, which can be viewed as the ratios between the community components for all communities. Its posterior distribution is obtained through the auxiliary variable $\hat{\boldsymbol{m}}$

$$(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\beta}_\mu) \sim \mathrm{Dir}(\hat{\boldsymbol{m}}_{\cdot 1}, \cdots, \hat{\boldsymbol{m}}_{\cdot K}, \gamma) \tag{5}$$

where its detail can be found in [10].

2) Sampling $\{s_{ij}^t\}_{n\times n}^{1:T}$, $\{r_{ij}^t\}_{n\times n}^{1:T}$: Each observation e_{ij}^t is sampled from a fixed Bernoulli distribution, where the Bernoulli's parameter is contained within the role-compatibility matrix W whose rows and columns are indexed by a pair of corresponding membership indicators $\{s_{ij}^t, r_{ij}^t\}$. W.l.o.g., $\forall k, l \in \{1, \cdots, K+1\}$, the joint posterior probability of $(s_{ij}^t = k, r_{ij}^t = l)$ is

$$\begin{aligned}
&\Pr\left(s_{ij}^t = k, r_{ij}^t = l | Z\backslash\{s_{ij}^t, r_{ij}^t\}, e, \boldsymbol{\beta}, \alpha, \lambda_1, \lambda_2, \kappa\right)\\
&\propto \Pr\left(s_{ij}^t = k | \{s_{ij_0}^t\}_{j_0\neq j}, \{r_{j_0 i}^t\}_{j_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t-1}\right)\\
&\quad\cdot \prod_{l=1}^{2n} \Pr\left(z_{il}^{t+1} | z_{i\cdot}^t / s_{ij}^t, s_{ij}^t = k, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t+1}\right)\\
&\quad\cdot \Pr\left(r_{ij}^t = l | \{r_{i_0 j}^t\}_{i_0\neq i}, \{s_{ji_0}\}_{i_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t-1}\right)\\
&\quad\cdot \prod_{l=1}^{2n} \Pr\left(z_{jl}^{t+1} | z_{j\cdot}^t / r_{ij}^t, r_{ij}^t = l, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t+1}\right)\\
&\quad\cdot \Pr\left(e_{ij}^t | E\backslash\{e_{ij}^t\}, s_{ij}^t = k, r_{ij}^t = l, \mathbf{Z}\backslash\ \{s_{ij}^t, r_{ij}^t\}, \lambda_1, \lambda_2\right).
\end{aligned} \tag{6}$$

The first two terms of (6)

$$\begin{aligned}
&\Pr\left(s_{ij}^t = k | \{s_{ij_0}^t\}_{j_0\neq j}, \{r_{j_0 i}^t\}_{j_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t-1}\right)\\
&\quad\cdot \prod_{l=1}^{2n} \Pr\left(z_{il}^{t+1} | z_{i\cdot}^t / s_{ij}^t, s_{ij}^t = k, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t+1}\right)\\
&\propto \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k + N_{ik}^{t+1} + \kappa N_{ik}^{t,-s_{ij}^t} + \kappa\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + N_{ik}^{t+1} + \kappa N_{ik}^{t,-s_{ij}^t}\right)} \cdot \frac{\Gamma\left(\alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t,-s_{ij}^t}\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t,-s_{ij}^t} + \kappa\right)}\\
&\quad\cdot \begin{cases} \alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t-1} + N_{ik}^{t,-s_{ij}^t}, & k \in \{1, \ldots, K\};\\ \alpha\boldsymbol{\beta}_\mu, & k = K+1 \end{cases}
\end{aligned} \tag{7}$$

where $N_{ik}^0 = 0$, $N_{ik}^{T+1} = 0$, $\forall i \in \{1, \ldots, n\}, k \in \{1, \ldots, K\}$.

The following two terms of (6) are:

$$\begin{aligned}
&\Pr\left(r_{ij}^t = l | \{r_{i_0 j}^t\}_{i_0\neq i}, \{s_{ji_0}^t\}_{i_0=1}^n, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t-1}\right)\\
&\quad\cdot \prod_{l=1}^{2n} \Pr\left(z_{jl}^{t+1} | z_{j\cdot}^t / r_{ij}^t, r_{ij}^t = l, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t+1}\right)\\
&\propto \frac{\Gamma\left(\alpha\boldsymbol{\beta}_l + N_{jl}^{t+1} + \kappa N_{jl}^{t,-r_{ij}^t} + \kappa\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_l + N_{jl}^{t+1} + \kappa N_{jl}^{t,-r_{ij}^t}\right)} \cdot \frac{\Gamma\left(\alpha\boldsymbol{\beta}_l + \kappa N_{jl}^{t,-r_{ij}^t}\right)}{\Gamma\left(\alpha\boldsymbol{\beta}_l + \kappa N_{jl}^{t,-r_{ij}^t} + \kappa\right)}\\
&\quad\cdot \begin{cases} \alpha\boldsymbol{\beta}_l + \kappa N_{jl}^{t-1} + N_{jl}^{t,-r_{ij}^t}, & l \in \{1, \ldots, K\}\\ \alpha\boldsymbol{\beta}_\mu, & l = K+1. \end{cases}
\end{aligned} \tag{8}$$

The last term, that is, the likelihood term, is calculated as

$$\begin{aligned}
&\Pr\left(e_{ij}^t | E\backslash\{e_{ij}^t\}, s_{ij}^t = k, r_{ij}^t = l, \mathbf{Z}\backslash\{s_{ij}^t, r_{ij}^t\}, \lambda_1, \lambda_2\right)\\
&= \begin{cases} \dfrac{n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^t} + \lambda_1}{n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} + \lambda_1 + \lambda_2}, & \boldsymbol{e}_{ij}^t = 1\\[2ex] \dfrac{n_{k,l}^{t,0,-\boldsymbol{e}_{ij}^t} + \lambda_2}{n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} + \lambda_1 + \lambda_2}, & \boldsymbol{e}_{ij}^t = 0 \end{cases}
\end{aligned} \tag{9}$$

where $n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} = n_{k,l}^t - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l) = \sum_{i'j'} \mathbf{1}(s_{i'j'}^t = k, r_{i'j'}^t = l) - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l)$, $n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^t} = n_{k,l}^{1,t} - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l)\boldsymbol{e}_{ij}^t = \sum_{i'j':s_{i'j'}^t=k, r_{i'j'}^t=l} \boldsymbol{e}_{i'j'}^t - \mathbf{1}(s_{ij}^t = k, r_{ij}^t = l)\boldsymbol{e}_{ij}^t$, and $n_{k,l}^{t,0,-\boldsymbol{e}_{ij}^t} = n_{k,l}^{t,-\boldsymbol{e}_{ij}^t} - n_{k,l}^{t,1,-\boldsymbol{e}_{ij}^t}$.

The detailed derivation of (7)–(9) is given in Assuming the current sample of $\{s_{ij}^t, r_{ij}^t\}$ has values ranging between $1 \ldots K$, we let the undiscovered (i.e., new) community be indexed by $K+1$. Then, to sample a pair (s_{ij}^t, r_{ij}^t) in question, we need to calculate all $(K+1)^2$ combinations of values for the pair.

3) Sampling $\hat{\boldsymbol{m}}$: Using the restaurant-table-dish analogy, we denote $\boldsymbol{m}_{ik}^t$ as the number of tables having dish k, $\forall i, k, t$. This is related to the variable $\hat{\boldsymbol{m}}$ used in sampling $\boldsymbol{\beta}$; it also includes the counts of the unsticky portion, that is, $\alpha\boldsymbol{\beta}_k$.

The sampling of $\boldsymbol{m}_{ik}^t$ incorporates a similar strategy as in [6] and [10], which is independently distributed from

$$\Pr\left(\boldsymbol{m}_{ik}^t = m | \alpha, \boldsymbol{\beta}_k, N_{ik}^{t-1}, \kappa\right) \propto S\left(N_{ik}^t, m\right)\left(\alpha\boldsymbol{\beta}_k + \kappa N_{ik}^{t-1}\right)^m \tag{10}$$

where $S(\cdot, \cdot)$ is the Stirling number of the first kind.

For each node, the ratio of generating new tables is the result of two factors: 1) a Dirichlet prior with parameter $\{\alpha, \boldsymbol{\beta}\}$ and 2) the sticky configuration from membership indicators at $t-1$, that is, κN_{ik}^{t-1}.

To sample $\boldsymbol{\beta}$, we need to only include tables generated from the unsticky portion, that is, $\hat{\boldsymbol{m}}$, where each $\hat{\boldsymbol{m}}_{ik}^t$ can be obtained from a single binomial raw

$$\hat{\boldsymbol{m}}_{ik}^{t} \sim \text{Binomial}\left(\boldsymbol{m}_{ik}^{t}, \frac{\alpha\boldsymbol{\beta}_k}{\frac{\kappa}{2n}N_{ik}^{t-1} + \alpha\boldsymbol{\beta}_k}\right). \tag{11}$$

$$\hat{\boldsymbol{m}}_k = \sum_{i,t} \hat{\boldsymbol{m}}_{ik}^{t}. \tag{12}$$

B. Adapted Slice-Efficient Sampling for the MTV Model

We also incorporate the slice-efficient sampling [28], [29] to our model. The original sampling scheme was designed to sample the Dirichlet process mixture model. To adapt it to our framework, which is based on a HDP prior and also has pairwise membership indicators, we use the auxiliary variables $U = \{u_{ij,s}^{t}, u_{ij,r}^{t}\}$ for each of the latent membership pairs $\{s_{ij}^{t}, r_{ij}^{t}\}$. With Us, we are able to limit the number of components in which $\boldsymbol{\pi}_i$ needs to be considered, which is otherwise infinite.

Under the slice-efficient sampling framework, the variables of interest are now extended to: $\boldsymbol{\pi}_i^t, \{u_{ij,r}^{t}, u_{ij,s}^{t}\}, \{s_{ij}^{t}, r_{ij}^{t}\}, \boldsymbol{\beta}, \boldsymbol{m}$:

1) Sampling $\boldsymbol{\pi}^t$: For each node $i = 1, \ldots, N$; $t = 1, \ldots, T$: we generate $\boldsymbol{\pi}_i^{'t}$ using the stick-breaking process [11], where each kth component is generated using $\boldsymbol{\pi}_{ik}^{'t} \sim \text{beta}(\boldsymbol{\pi}_{ik}^{'t}; a_{ik}^{t}, b_{ik}^{t})$ where

$$\begin{aligned} a_{ik}^{t} &= \alpha\boldsymbol{\beta}_k + N_{ik}^{t} + \kappa N_{ik}^{t-1} \\ b_{ik}^{t} &= \alpha\left(1 - \sum_{l=1}^{k}\boldsymbol{\beta}_l\right) + N_{i,k_0>k}^{t} + \kappa N_{i,k^0>k}^{t-1} \end{aligned} \tag{13}$$

where $\boldsymbol{\pi}_k^{t} = \boldsymbol{\pi}_k^{'t}\prod_{i=1}^{k-1}(1 - \boldsymbol{\pi}_i^{'t})$.

2) Sampling $u_{ij,s}^{t}, u_{ij,r}^{t}, s_{ij}^{t}, r_{ij}^{t}$: We use $u_{ij,s}^{t} \sim U(0, \boldsymbol{\pi}_{is_{ij}^{t}}^{t})$, $u_{ij,r}^{t} \sim U(0, \boldsymbol{\pi}_{jr_{ij}^{t}}^{t})$. The hidden label subsequently obtained is then independently sampled from the finite candidates

$$\begin{aligned} &P\big(s_{ij}^{t} = k, r_{ij}^{t} = l | Z, e_{ij}^{t}, \boldsymbol{\beta}, \alpha, \kappa, N, \boldsymbol{\pi}, u_{ij,s}^{t}, u_{ij,r}^{t})\big) \\ &\quad\propto \mathbf{1}\big(\boldsymbol{\pi}_{ik}^{t} > u_{ij,s}^{t}\big) \cdot \mathbf{1}\big(\boldsymbol{\pi}_{jl}^{t} > u_{ij,r}^{t}\big) \\ &\qquad\cdot \prod_{l=1}^{2n} \Pr\left(z_{il}^{t+1} | z_{i\cdot}^{t}/s_{ij}^{t}, s_{ij}^{t} = k, \boldsymbol{\beta}, \alpha, \kappa, N_i^{t+1}\right) \\ &\qquad\cdot \prod_{l=1}^{2n} \Pr\left(z_{jl}^{t+1} | z_{j\cdot}^{t}/r_{ij}^{t}, r_{ij}^{t} = l, \boldsymbol{\beta}, \alpha, \kappa, N_j^{t+1}\right) \\ &\qquad\cdot \Pr\left(e_{ij}^{t} | E\backslash\{e_{ij}^{t}\}, s_{ij}^{t} = k, r_{ij}^{t} = l, \boldsymbol{Z}\backslash\{s_{ij}^{t}, r_{ij}^{t}\}, \lambda_1, \lambda_2\right). \end{aligned} \tag{14}$$

We refer the reader to (7)–(9) for the detailed calculation of each term in (14).

3) Sampling $\boldsymbol{\beta}$: An obvious choice for the proposal distribution of $\boldsymbol{\beta}$ used in M-H is its prior $p(\boldsymbol{\beta}|\gamma) = \text{stick} - \text{breaking}(\gamma)$. However, this proposal may be noninformative, which results in a low acceptance rate. We sample $\boldsymbol{\beta}^*$ conditioned on an auxiliary variable $\hat{\boldsymbol{m}}$: $(\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*, \boldsymbol{\beta}_{K+1}^*) \sim Dir(\hat{\boldsymbol{m}}_1, \ldots, \hat{\boldsymbol{m}}_K, \gamma)$, to increase the M-H's acceptance rate, where $\hat{\boldsymbol{m}}$ are sampled in accordance with the method proposed in Section V-A3 [(10)–(12)]. However, instead of sampling $\boldsymbol{\beta}$ directly from $\boldsymbol{m}$ as described in Section V-A3, we only use it for our proposal distribution, as we explicitly sample $\{\pi_i\}_{i=1}^{n}$. The acceptance ratio is hence (τ indexes the iteration time)

$$A(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(\tau)}) = \min(1, a) \tag{15}$$

$$a = \frac{\prod_{t,i}\left[\prod_{d=1}^{K+1}\Gamma\big(\alpha\boldsymbol{\beta}_d^{(\tau)}\big) \cdot \big[\pi_{id}^{t}\big]^{\alpha\boldsymbol{\beta}_d^*}\right]}{\prod_{t,i}\left[\prod_{d=1}^{K+1}\Gamma\big(\alpha\boldsymbol{\beta}_d^{*}\big) \cdot \big[\pi_{id}^{t}\big]^{\alpha\boldsymbol{\beta}_d^{(\tau)}}\right]} \cdot \frac{\prod_{d=1}^{K}\big[\boldsymbol{\beta}_d^{(\tau)}\big]^{\hat{\boldsymbol{m}}_d-\gamma}}{\prod_{d=1}^{K}\big[\boldsymbol{\beta}_d^{*}\big]^{\hat{\boldsymbol{m}}_d-\gamma}}. \tag{16}$$

C. Hyperparameter Sampling

The hyperparameters involved in the MTV model are γ, α, and κ. However, it is impossible to compute their posterior individually. Therefore, we place three prior distributions on some combination of the variables. A vague gamma prior $\mathcal{G}(1, 1)$ is placed on both $\gamma, (\alpha + \kappa)$. A beta prior is placed on the ratio $\kappa/\alpha + \kappa$.

To sample γ value, since $\log(\gamma)$'s posterior distribution is log-concave, we use the adaptive rejection sampling (ARS) method [30].

To sample $(\alpha + \kappa)$, we use the auxiliary variable sampling [10], and this needs the auxiliary variable $\boldsymbol{m}$ in (10), as proposed in [10].

To sample $\kappa/(\alpha + \kappa)$, we place a vague beta prior $\mathcal{B}(1, 1)$ on it, with a likelihood of $\{\boldsymbol{m}_{ik}^{t} - \hat{\boldsymbol{m}}_{ik}^{t}, \forall i, k, t > 1\}$ in (11). The posterior is in an analytical form that can be sampled, owing to its conjugate property.

D. Gibbs Sampling for the MTI Model

The variables of interest are: $\boldsymbol{\beta}$, Z and auxiliary variables $\hat{\boldsymbol{m}}$, where $\hat{\boldsymbol{m}}$ refers to the number of tables having dish k as used in [6] and [10] without counting the tables generated from the sticky portion, that is, κN_{ik}^{t-1}. As the hyperparameters in the MTI model are quite similar to those in [12], we do not present the hyperparameters here. Interested readers can refer to [6], [12], and [13] for the detailed implementation.

1) Sampling $\boldsymbol{\beta}$: $\boldsymbol{\beta}$'s sampling is the same as (1).

2) Sampling s_{ij}^{t}, r_{ij}^{t}: The posterior probability of s_{ij}^{t}, r_{ij}^{t} is

$$\begin{aligned} &\Pr\big(s_{ij} = k, r_{ij} = l | \alpha, \boldsymbol{\beta}, \kappa, \{N_{\cdot\cdot}^{(i)}\}, \{N_{\cdot\cdot}^{(j)}\}, \boldsymbol{e}, \lambda_1, \lambda_2, Z\big) \\ &\quad\propto \Pr\big(s_{ij}^{t} = k | \alpha, \boldsymbol{\beta}, \kappa, N_{s_{ij}^{t-1}\cdot}^{(i)}, s_{ij}^{t-1}\big) \\ &\Pr\big(r_{ij}^{t} = l | \alpha, \boldsymbol{\beta}, \kappa, N_{r_{ij}^{t-1}\cdot}^{(j)}, r_{ij}^{t-1}\big) \\ &\qquad\cdot \Pr\big(\boldsymbol{e}_{ij}^{t} | \boldsymbol{e}/\{\boldsymbol{e}_{ij}^{t}\}, s_{ij}^{t} = k, r_{ij}^{t} = l, Z/\{s_{ij}^{t}, r_{ij}^{t}\}, \lambda_1, \lambda_2\big). \end{aligned} \tag{17}$$

The first term of (17) is

$$\begin{aligned} &\Pr\big(s_{ij}^{t} = k | \alpha, \boldsymbol{\beta}, \kappa, N_{s_{ij}^{t-1}\cdot}^{(i)}, s_{ij}^{t-1}\big) \\ &\propto \big(\alpha\boldsymbol{\beta}_k + N_{s_{ij}^{t-1}k}^{(i)} + \kappa\delta\big(s_{ij}^{t-1}, k\big)\big) \\ &\quad\cdot\left(\frac{\alpha\boldsymbol{\beta}_{s_{ij}^{t+1}} + N_{ks_{ij}^{t+1}}^{(i)} + k\delta\big(k, s_{ij}^{t+1}\big) + \delta\big(k, s_{ij}^{t-1}\big)\delta\big(k, s_{ij}^{t+1}\big)}{\alpha + N_{k\cdot}^{(i)} + \kappa + \delta\big(s_{ij}^{t-1}, k\big)}\right). \end{aligned} \tag{18}$$

$$\begin{vmatrix} 0.95 & 0.05 & 0 \\ 0.05 & 0.95 & 0.05 \\ 0.05 & 0 & 0.95 \end{vmatrix} \qquad \begin{vmatrix} 0.95 & 0.2 & 0 \\ 0.05 & 0.95 & 0.05 \\ 0.2 & 0 & 0.95 \end{vmatrix} \qquad \begin{vmatrix} 0.05 & 0.95 & 0 \\ 0.05 & 0.05 & 0.95 \\ 0.95 & 0 & 0.05 \end{vmatrix} \qquad \begin{vmatrix} 0.05 & 0.95 & 0 \\ 0.2 & 0.05 & 0.95 \\ 0.95 & 0 & 0.2 \end{vmatrix}$$

Fig. 4. Four cases of the compatibility matrix. Left (Case 1): large diagonal values and small nondiagonal values. Left-middle (Case 2): large diagonal values and mediate nondiagonal values. Right-middle (Case 3): large nondiagonal values and small diagonal values. Right (Case 4): small diagonal values and mediate nondiagonal values.

The second term of (17) is

$$\Pr\left(r_{ij}^{t} = l|\alpha, \boldsymbol{\beta}, \kappa, N^{(j)}_{r_{ij}^{t-1}\cdot}, r_{ij}^{t-1}\right) \propto \left(\alpha\boldsymbol{\beta}_l + N^{(j)}_{r_{ij}^{t-1}l} + \kappa\delta\left(r_{ij}^{t-1}, l\right)\right) \cdot \left(\frac{\alpha\boldsymbol{\beta}_{r_{ij}^{t+1}} + N^{(i)}_{lr_{ij}^{t+1}} + l\delta\left(l, r_{ij}^{t+1}\right) + \delta\left(l, r_{ij}^{t-1}\right)\delta\left(l, r_{ij}^{t+1}\right)}{\alpha + N^{(i)}_{l\cdot} + \kappa + \delta\left(r_{ij}^{t-1}, l\right)}\right). \quad (19)$$

The likelihood of $\Pr(\boldsymbol{e}_{ij}^{t}|\boldsymbol{e}/\{\boldsymbol{e}_{ij}^{t}\}, s_{ij}^{t} = k, r_{ij}^{t} = l, Z/\{s_{ij}^{t}, r_{ij}^{t}\}, \lambda_1, \lambda_2)$ is the same as (9).

3) Sampling $\hat{\boldsymbol{m}}$: $\hat{\boldsymbol{m}}$ is similar to that in the MTV model; however, it differs in the incorporation of κ

$$\Pr\left(\boldsymbol{m}_{qk}^{(i)} = m|\alpha, \boldsymbol{\beta}_k, \kappa, N_{qk}^{(i)}\right) \propto S\left(N_{qk}^{(i)}, m\right)\left(\alpha\boldsymbol{\beta}_k + \kappa\right) \quad (20)$$

$$\hat{\boldsymbol{m}}_{qk}^{(i)} \sim \text{Binomial}\left(\boldsymbol{m}_{qk}^{(i)}, \frac{\alpha\boldsymbol{\beta}_k}{\kappa + \alpha\boldsymbol{\beta}_k}\right) \quad (21)$$

$$\hat{\boldsymbol{m}}_{\cdot k} = \sum_{q,i} \hat{\boldsymbol{m}}_{qk}^{(i)}. \quad (22)$$

E. Inference Discussions

Both the Gibbs sampling and slice-efficient sampling are two feasible ways to accomplish our task. They have different advantages and disadvantages.

As mentioned previously, Gibbs sampling in our MTV model integrates out the mixed-membership distribution $\{\boldsymbol{\pi}_i^t\}$. It is the marginal approach [31]. The property of community exchangeability makes it simple to implement. However, theoretically, the obtained samples mix slowly as the sampling of each label is dependent on other labels.

Slice-efficient sampling is a conditional approach [28] whereas the membership indicators are independently sampled from $\{\boldsymbol{\pi}_i^t\}$. In each iteration, given $\{\boldsymbol{\pi}_i^t\}$ and the role-compatibility matrix W, we can parallelize the process of sampling membership indicators, which may help to improve the computation, especially when the number of nodes (N) becomes larger, and the number of communities (k) becomes smaller.

VI. Experiments

The performance of our DIM3 model is validated by experiments on both synthetic and real-world datasets. On the synthetic datasets, we implement the finite-communities cases of our models as baseline algorithms, namely as the f-MTV and f-MTI model. On the real-world datasets, we individually implement three benchmark models: MMSB, IRM, and LFRM to the best of our understanding. Also, we compare DRIFT with our models on real-world datasets, and the source code is provided by [26].

A. Synthetic Datasets

For the synthetic data generation, the variables are generated by following [7]. We use $N = 20, T = 3$, and hence E is a $20 \times 20 \times 3$ asymmetric and binary matrix. The parameters are set up in a way so that 20 nodes are equally partitioned into four groups. The ground-truth of the mixed-membership distributions for each of the groups are $[0.8, 0.2, 0; 0, 0.8, 0.2; 0.1, 0.05, 0.85; 0.4, 0.4, 0.2]$.

We consider four different cases to fully assess DIM3 against the ground-truth; all lie in the three-role-compatibility matrix.

The detailed results of the role-compatibility matrix on these four cases are shown in Fig. 4.

1) Markov Chain Monte Carlo Analysis: The convergence behavior is tested in terms of two quantities: the cluster number K, that is, the number of different values Z can take, and the deviance D of the estimated density [28], [31], which is defined as

$$D = -2\sum_{i,j,t} \log\left(\sum_{k,l} \frac{N_{ik}^{t} \cdot N_{jl}^{t}}{4n^2T} p(e_{ij}^{t}|Z, \lambda_1, \lambda_2)\right). \quad (23)$$

In our Markov Chain Monte Carlo (MCMC) stationary analysis, we run five independent Markov chains and discard the first half of the Markov chains as a burn-in. With the random partition of three initial classes as the starting point, 20 000 iterations are conducted in our samplings.

The simulated chains satisfy the standard convergence criteria, when the test was implemented using the CODA package [32]. In Gelman and Rubin's diagnostics [33], the value of the proportional scale reduction factor is 1.09 (with upper C.I. 1.27) for k, 1.03 (with upper C.I. 1.09) for D in the Gibbs sampling, and 1.02 (with upper C.I. 1.06) for k, 1.02 (with upper C.I. 1.02) for D in slice sampling. Geweke's convergence diagnostics [34] are also employed, with the proportion of the first 10% and last 50% of the chain for comparison. The corresponding z-scores are calculated in the interval $[-2.09, 0.85]$ for five chains. In addition, the stationarity and half-width tests of the Heidelberg and Welch Diagnostic [35] are both passed in all cases, with the p-value higher than 0.05. On the basis of all these statistics, the Markov chain's stationarity can be safely ensured in our case.

The efficiency of the algorithms can be measured by estimating the integrated autocorrelation time τ for K and D. 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 is.

Referenece [28] used an estimator $\widehat{\tau}$ as

$$\widehat{\tau} = \frac{1}{2} + \sum_{l=1}^{C-1} \widehat{\rho}_l \quad (24)$$

TABLE II

INTEGRATED AUTOCORRELATION TIMES ESTIMATOR $\widehat{\tau}$ FOR K AND D

		K					D				
Sampling	γ \ α	0.1	0.3	0.5	1	2	0.1	0.3	0.5	1	2
MTV-g	0.1	177.2	93.65	26.91	50.21	11.24	358.8	148.3	23.94	84.75	4.31
	0.3	260.5	54.00	9.18	5.31	6.56	389.5	315.0	3.11	26.32	4.78
	0.5	1.83	8.33	7.54	3.95	5.24	2.88	79.34	90.93	3.17	3.82
	1.0	5.57	6.45	3.44	3.64	4.56	3.19	2.78	1.76	8.14	5.74
	2.0	4.30	2.87	3.35	2.98	3.28	95.48	1.91	3.29	8.74	6.55
MTV-s	0.1	248.6	90.63	161.3	9.58	17.69	8.67	59.90	57.57	1.87	3.70
	0.3	120.6	66.23	44.35	11.40	7.28	29.05	20.64	30.01	45.57	3.40
	0.5	18.99	27.27	6.08	8.76	10.40	39.66	3.87	5.30	3.17	5.83
	1.0	5.79	9.19	11.85	8.46	7.25	40.51	4.85	3.12	6.88	10.51
	2.0	3.17	8.41	5.35	5.48	5.05	25.54	34.82	4.61	35.61	12.68

where $\widehat{\rho_l}$ is the estimated autocorrelation at lag l and C is a cutoff point, which is defined as $C := \min\{l : |\widehat{\rho_l}| < 2/\sqrt{M}\}$, and M is the number of iterations.

We test the sampling efficiency of the MTV-g and MTV-s models on Case 1 with the same setting as [31]. Of the whole 20 000 iterations, the first half of the samples is discarded as a burn-in and the remainder are thinned $1/20$. We manually try different values of the hyperparameters γ and α and show the integrated autocorrelation time estimator in Table II. Although some outliers exist, we can see that there is a general trend that, with a fixed α value, the autocorrelation function decreases when the γ value increases. This same phenomenon happens on α while γ is fixed. This result confirms our empirical knowledge. The larger value of γ, α will help to discover more clusters, followed by a smaller autocorrelation function.

On the other hand, we confirm that MTV-g and MTV-s models do not show much difference in the mixing rate of the Markov Chain, as shown in Table II. As mentioned in the previous section, slice sampling provides a mixed-membership distribution-independent sampling scheme, which enjoys the time efficiency of parallel computing in one iteration. For large-scale datasets, it is a feasible solution. In Gibbs sampling, parallel computing is impossible as the sampling variables are in a dependent sequence.

Fig. 5 shows the trace plot of the training log-likelihood against the iterations on Case 1. As we can see, the sampler in the MTI model converges to the high training log-likelihood region faster than the MTV model. Also, the MTI model reaches a higher training log-likelihood than the MTV model.

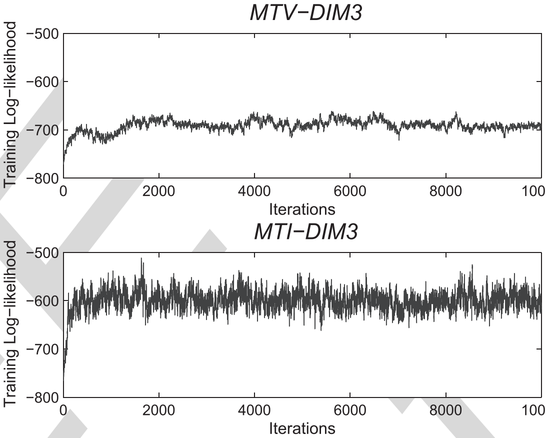


Fig. 5. Top: training log-likelihood trace plot on the MTV-g model. Bottom: training log-likelihood trace plot on the MTI-g model.

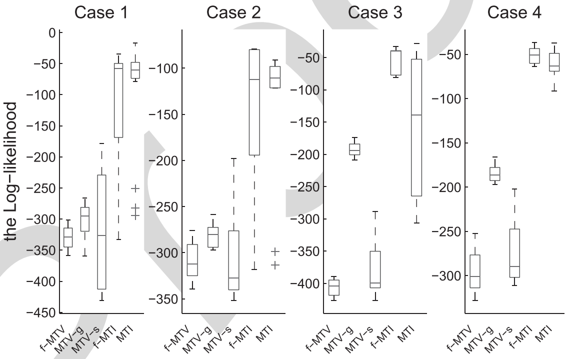


Fig. 6. Log-likelihood performance on all the four cases.

2) Further Performance: We will compare the models in terms of the log-likelihood (Fig. 6); the average l_2 distance between the mixed-membership distributions and its ground-truth; and the l_2 distance between the posterior role-compatibility matrix and its ground-truth (Table III).

From the log-likelihood comparison shown in Fig. 6, we can see that the MTI model performs better than the MTV model in general. On the average l_2 distance to the ground-truth performance, the MTI model also performs better. The superiority of the MTI model's performance over that of the MTV model is within our expectation, as the MTI model describes the membership indicator's time consistency more accurately (i.e., integrating the sticky parameter κ on the specific membership indicator, rather than the mixed-membership distribution). Also, the hidden Markov property enables the MTI model to categorize membership indicators into the same mixed-membership distributions on the basis of its previous value. This seems to be a more effective method than the time-based grouping in the MTV model. However, in situations where there are dramatic changes amongst the membership distributions over time, the MTI model will not respond well. The MTV model is much more effective and robust under these settings as the distribution consistency is a more robust modeling strategy. In addition, the assumption

TABLE III

AVERAGE l_2 DISTANCE TO THE GROUND-TRUTH

Cases	Role-compatibility matrix						Mixed-memberships					
	f-MTV	MTV-g	MTV-s	f-MTI	MTI	MMSB	f-MTV	MTV-g	MTV-s	f-MTI	MTI	MMSB
1	0.239	0.243	0.259	0.114	**0.086**	0.271	0.366	0.384	0.403	0.199	**0.191**	0.411
2	0.206	0.225	0.240	**0.195**	0.204	0.285	0.355	0.355	0.319	**0.207**	0.227	0.398
3	0.134	0.201	0.246	0.117	**0.087**	0.280	0.278	0.289	0.589	0.208	**0.187**	0.329
4	**0.195**	0.214	0.267	0.220	0.219	0.246	0.258	0.285	0.277	0.192	**0.182**	0.310
large size	0.220	0.239	0.215	0.142	**0.059**	0.237	0.243	0.316	0.246	0.147	**0.120**	0.296

TABLE IV

RUNNING TIME (SECONDS PER ITERATION)

N.	f-MTV	MTV-g	MTV-s	f-MTI	MTI	p-MTV-s[1]
20	0.20	0.28	0.23	0.15	0.31	0.29
50	1.03	1.52	1.29	0.95	1.91	1.79
100	3.69	5.76	4.81	3.74	7.49	5.06
200	15.61	24.17	19.87	15.82	30.19	21.64
500	106.96	154.45	119.82	105.61	202.09	132.43
1000	493.44	888.86	642.28	597.29	1102.90	393.24

[1] p-MTV-s denotes the parallel implementation of the MTV-s inference.

that there exist different membership distributions at each time instance makes it possible to parallelize the MTV model to some extent, making it suitable for dealing with large-scale problems.

Here, we compare the computational complexity (running time) of the models in one iteration, with K discovered communities, and show the results in Table IV. We discuss the MTV-g and MTV-s models as an instance. In the MTV-g model, the number of variables to be sampled is $(2K + 2n^2T)$, whereas a total of $(2K + 4n^2T + nT)$ variables are sampled in the MTV-s model. However, the posterior calculation of Z in the MTV-s model can be directly obtained from the mixed-membership distribution, while we need to calculate the ratio for each of Z in the MTV-g model. Also, the U value at each time can be sampled in one operation as its independency in the MTV-s model. The result shows that the MTV-s model runs faster than the MTV-g model, which is in accordance with our assumption.

We also tried a parallel implementation of the slice variables $\{u^t_{ij,s}, u^t_{ij,r}\}_{i,j,t}$'s in the MTV-s model. During each iteration, these slice variables are partitioned into four parts (as our machine has four cores) and are sampled independently, while other variables are still sampled in a sequence. Its corresponding running time is shown in the last column of Table IV. It shows that the parallel design costs even more time when the dataset size is small ($N \leq 500$). This may be due to the time spent on transferring the variables. However, it needs less time when the dataset size becomes larger ($N > 500$). This verifies that our parallel slice sampling method is a promising approach in achieving scalability.

3) Larger Data Size Results: We also conduct the experiments with a larger synthetic dataset ($N = 100, T = 20$). With the same construction as previous ones, we increase the role number to 5 and set the role-compatibility matrix as shown in Fig. 7.

$$\begin{vmatrix} 0.95 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.1 & 0.9 & 0.1 & 0.1 & 0.1 \\ 0.05 & 0.1 & 0.9 & 0.05 & 0 \\ 0.05 & 0 & 0.05 & 0.9 & 0.15 \\ 0 & 0.05 & 0.1 & 0.05 & 0.9 \end{vmatrix}$$

Fig. 7. Larger dataset's role-compatibility matrix.

We set five groups in this network, with the group sizes as $[35, 20, 20, 20, 5]$ and the mixed-membership distributions for each of the groups as $[0.8, 0.1, 0, 0.05, 0.05; 0.02, 0.85, 0.05, 0.03, 0.05; 0.1, 0, 0.9, 0, 0;\ 0.05, 0.1, 0, 0.85, 0; 0, 0.2, 0, 0.4, 0.4]$. The detailed results are also given in Table III. As we can see, our MTI model still achieves the best performance of all the models.

B. Real-World Datasets Performance

We select ten real-world datasets for benchmark testing. Their detailed information, including the number of nodes, the number of edges, edge types, and time intervals, is given in Table VII. Following a general test on the training log-likelihood of the training data and area under the ROC (Receiver Operating Characteristic) curve (AUC) of the test data, we elaborate the results on three selected datasets in the following.

We use a fivefold cross validation method to certify our model's performance on the real-world datasets. The hyper-parameters γ, κ, α are sampled according to the sampling strategy mentioned in Section V. Each experiment is run ten times and we report their mean and standard deviation in Tables V and VI.

In these two tables, the bold type denotes the best value in each row. As we can see, our MTI model performs best in eight of the ten datasets on the training log-likelihood and six of the ten datasets on the AUC value. In the remaining datasets, although our MTI model's performance is still quite competitive, the DRIFT model has the best values, possibly because, in these datasets, all associated communities from both nodes are considered in generating the link between these two nodes [14]. The MTV models still do not perform well enough, for the reason previously given. The IRM's results are the worst, which reflects that the simple structure (i.e., each node occupies only one class) may not be enough to capture the full structure in relational learning.

C. Kapferer Tailor Shop

The Kapferer Tailor Shop data [1] records interactions in a tailor shop at two time points. In this time period, the employees in the shop negotiate for higher wages. The dataset

TABLE V

TRAINING LOG-LIKELIHOOD PERFORMANCE (95% CONFIDENCE INTERVAL = MEAN ∓1.96× STANDARD DEVIATION)

Dataset	MTV-g	MTV-s	MTI	MMSB	IRM	LFRM	DRIFT
Kapferer	-673.7 ∓ 15.9	-698.9 ∓ 15.2	$\mathbf{-501.5 \mp 0.0}$	-618.4 ∓ 59.8	-658.6 ∓ 70.3	-865.1 ∓ 70.1	-783.2 ∓ 92.3
Sampson	-347.6 ∓ 23.4	-350.4 ∓ 22.2	$\mathbf{-242.0 \mp 0.0}$	-353.0 ∓ 16.3	-366.8 ∓ 0.6	-332.2 ∓ 16.9	-275.2 ∓ 52.0
Student-net	-1054.4 ∓ 48.5	-1059.3 ∓ 46.2	$\mathbf{-594.3 \mp 0.0}$	-881.4 ∓ 29.9	-1201.2 ∓ 1.6	-1069.6 ∓ 42.2	-905.8 ∓ 46.3
Enron	-2274.2 ∓ 25.6	-2154.4 ∓ 43.3	$\mathbf{-1335.7 \mp 17.1}$	-1512.5 ∓ 6.5	-2264.8 ∓ 26.2	-1742.9 ∓ 36.0	-1492.3 ∓ 13.2
Senator	-897.3 ∓ 16.2	-887.4 ∓ 43.2	$\mathbf{-657.4 \mp 12.3}$	-713.2 ∓ 64.2	-843.6 ∓ 23.5	-673.2 ∓ 43.6	-678.6 ∓ 48.5
DBLP-link	-1923.9 ∓ 19.4	-2124.6 ∓ 26.4	$\mathbf{-1049.6 \mp 7.5}$	-2082.0 ∓ 12.0	-2953.1 ∓ 4.9	-1746.5 ∓ 15.4	-1426.1 ∓ 46.2
Hypertext	-5276.7 ∓ 9.6	-5281.4 ∓ 10.3	$\mathbf{-2923.2 \mp 0.0}$	-4083.5 ∓ 77.8	-5432.7 ∓ 19.6	-3747.5 ∓ 94.3	-3942.3 ∓ 48.5
Newcomb	-1075.0 ∓ 47.6	-1098.1 ∓ 48.0	-876.7 ∓ 0.0	-1835.2 ∓ 14.2	-1965.9 ∓ 1.8	-1203.0 ∓ 14.7	$\mathbf{-789.3 \mp 63.2}$
Freeman	-658.5 ∓ 19.6	-664.1 ∓ 19.2	$\mathbf{-405.2 \mp 0.0}$	-673.5 ∓ 73.9	-728.9 ∓ 66.9	-917.2 ∓ 35.7	-794.2 ∓ 66.2
Coleman	-1500.8 ∓ 63.7	-1532.8 ∓ 64.2	-1003.9 ∓ 0.0	-1302.8 ∓ 130.2	-689.5 ∓ 3.2	-606.7 ∓ 65.1	$\mathbf{-546.1 \mp 26.9}$

TABLE VI

AUC PERFORMANCE (95% CONFIDENCE INTERVAL = MEAN ∓1.96× STANDARD DEVIATION)

Dataset	MTV-g	MTV-s	MTI	MMSB	IRM	LFRM	DRIFT
Kapferer	0.816 ∓ 0.074	0.816 ∓ 0.011	$\mathbf{0.928 \mp 0.000}$	0.893 ∓ 0.001	0.751 ∓ 0.016	0.891 ∓ 0.034	0.905 ∓ 0.013
Sampson	0.804 ∓ 0.000	0.821 ∓ 0.098	$\mathbf{0.927 \mp 0.000}$	0.836 ∓ 0.002	0.738 ∓ 0.005	0.841 ∓ 0.012	0.855 ∓ 0.029
Student-net	0.867 ∓ 0.030	0.877 ∓ 0.095	0.934 ∓ 0.000	0.938 ∓ 0.001	0.809 ∓ 0.004	0.862 ∓ 0.076	$\mathbf{0.949 \mp 0.015}$
Enron	0.834 ∓ 0.097	0.853 ∓ 0.143	0.920 ∓ 0.001	0.907 ∓ 0.013	0.820 ∓ 0.082	0.894 ∓ 0.073	$\mathbf{0.956 \mp 0.079}$
Senator	0.849 ∓ 0.129	0.839 ∓ 0.046	$\mathbf{0.931 \mp 0.001}$	0.880 ∓ 0.022	0.829 ∓ 0.064	0.892 ∓ 0.056	0.925 ∓ 0.076
DBLP-link	0.831 ∓ 0.046	0.816 ∓ 0.017	$\mathbf{0.926 \mp 0.000}$	0.918 ∓ 0.000	0.817 ∓ 0.010	0.891 ∓ 0.062	0.891 ∓ 0.034
Hypertext	0.861 ∓ 0.029	0.843 ∓ 0.027	$\mathbf{0.901 \mp 0.023}$	0.844 ∓ 0.008	0.788 ∓ 0.015	0.853 ∓ 0.042	0.871 ∓ 0.010
Newcomb	0.814 ∓ 0.049	0.795 ∓ 0.090	0.931 ∓ 0.000	0.836 ∓ 0.001	0.765 ∓ 0.013	0.879 ∓ 0.041	$\mathbf{0.960 \mp 0.027}$
Freeman	0.875 ∓ 0.133	0.862 ∓ 0.041	$\mathbf{0.915 \mp 0.000}$	0.867 ∓ 0.001	0.790 ∓ 0.008	0.883 ∓ 0.026	0.897 ∓ 0.022
Coleman	0.891 ∓ 0.067	0.872 ∓ 0.052	0.928 ∓ 0.000	0.928 ∓ 0.001	0.888 ∓ 0.004	0.929 ∓ 0.018	$\mathbf{0.945 \mp 0.052}$

TABLE VII

DATASET INFORMATION

Dataset	Nodes	Edge	Time	Link Type
Kapferer [36]	39	256	2	friends
Sampson [37], [38]	18	168	3	like
Student-net	50	351	3	friends
Enron [39]	151	1980	12	email
Senator [7]	100	5786	8	vote
DBLPlink [40], [41]	100	5706	10	citation
Hypertext [42]	113	7264	10	contact
Newcomb [43]	17	1020	15	contact
Freeman [44]	32	357	2	friends
Coleman [45]	73	506	2	co-work

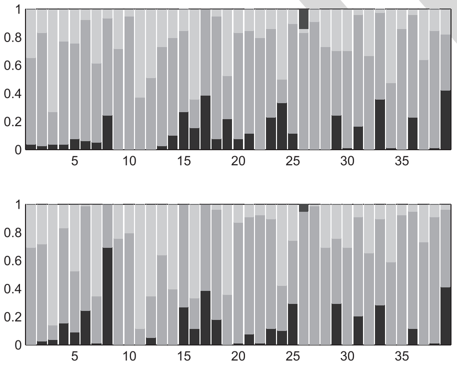


Fig. 8. MTI model's performance on Kapferer Tailor Shop dataset. The x-axis stands for the nodes, while the y-axis represents the mixed-membership distribution. Different colors represent various communities we discovered. Top bar chart: all the employees' mixed-membership distributions in Time 1. Bottom bar chart: all the employees' mixed-membership distributions in Time 2.

is of particular interest because two strikes occur after each time point, with the first failing and the second successful.

We mainly use the work–assistance interaction matrix in the dataset. The employees have eight occupations: head tailor (19), cutter (16), line 1 tailor (1-3, 5-7, 9, 11-14, 21, 24), button machiner (25-26), line 3 tailor (8, 15, 20, 22-23, 27-28), ironer (29, 33, 39), cotton boy (30-32, 34-38), and line 2 tailor (4, 10, 17-18).

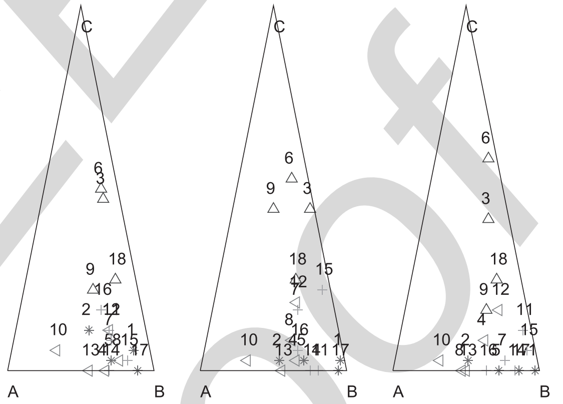


Fig. 9. Nodes' mixed-membership distribution of the MTI model on Sampson Monastery dataset. Left to right: time 1–3. Blue: *loyal opposition*. Red: *outcasts*. Green: *young Turks*. Magenta: *interstitial group*.

$$\begin{vmatrix} 0.09 & 0 & 0.0 \\ 0.05 & 0.99 & 0.02 \\ 0.01 & 0 & 0.96 \end{vmatrix} \quad \begin{vmatrix} 0.01 & 0 & 0.03 \\ 0.02 & 0.78 & 0 \\ 0.02 & 0 & 0.67 \end{vmatrix}$$

Fig. 10. Role-compatibility matrix. Left: MTV-g. Right: MTI.

In Fig. 8, we can see that the yellow communities at Time 2 are larger than those at Time 1, which means that people tend to have another community at Time 2, rather than being mostly dominated by one large group at Time 1. This larger yellow community may be the result of the first failed strike, after which employees start to shift to the minor (yellow) community for a successful strike.

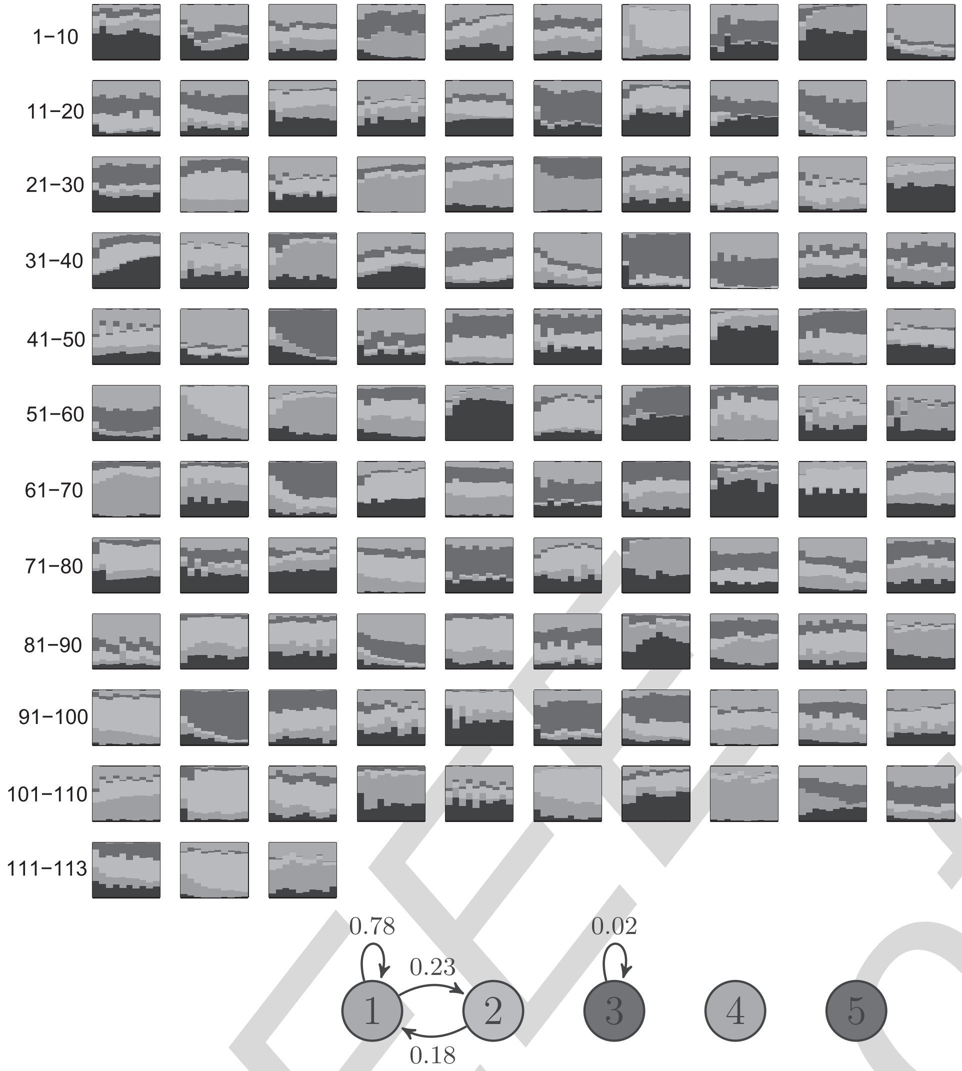


Fig. 11. MTI model's performance on the hypertext 2009 dynamic contact network. Numbers on the left side: orders of nodes. Each bar chart: dynamic behavior of one node's mixed-membership distribution, where the x-axis stands for the ten time stamps. Different colors are interpreted as the communities we have discovered, and their role-compatibility is represented below the bar chart.

D. Sampson Monastery Dataset

The Sampson Monastery dataset is used here to extend the study. There are 18 monks in this dataset, and their social linkage data is collected at three different time points with various interactions. Here, we especially focus on the like-specification. In the like-specification data, each monk selects three monks as his closest friends. In our settings, we mark the selected interactions as 1, otherwise 0. Thus, an $18 \times 18 \times 3$ social network dataset is constructed, with each row having three elements valued at 1.

According to the previous studies in [8] and [23], the monks are divided into four communities: *young Turks*, *loyal opposition*, *outcasts*, and an *interstitial group*.

Fig. 9 shows the detailed results of the MTI model. As three communities have been detected, we put all the results in a two-simplex, in which we denote the communities as A, B, and C. For trajectory convenience, we also color the nodes according to which special group they belong. The results show that these groups behave significantly differently. The *loyal opposition* group lies closer to C, and the *interstitial group* tends to belong to A. Both of their mixed-membership distributions are stable across time. The *outcasts* and *young Turks* groups lie much closer to B.

We also show the role-compatibility matrix in Fig. 10 for comparison. Compared with the results given in [8], our results have larger compatibility values for the same role. Also, the first role's value in our model is 0 versus 0.6 that is reported in [8].

E. Hypertext 2009 Dynamic Contact Network

This dataset [42] is collected from the ACM Hypertext 2009 conference. 113 conference attendees volunteered to wear radio badges that recorded their face-to-face contacts during the conference. The original data is composed of records such as (t, i, j), where t is the communication time and i, j are the attendees' ID. By adaptively partitioning the whole time period into ten parts and noting the interaction data as 1 if communicated during the time stamps, we obtain a $113 \times 113 \times 10$ binary matrix. Fig. 11 shows the dynamic behavior of the nodes' mixed-membership distributions and the corresponding role-compatibility matrix.

The results show that almost half of all the mixed-membership distributions fluctuate during these time stamps. This phenomenon coincides with our common knowledge that people at academic conferences tend to communicate causally. Thus, people's roles may change during different time stamps.

The learned value of the role-compatibility matrix is about the sky blue community, whose intrarole-compatibility value is 0.6932. It has a small probability of interaction with other communities. The other community's compatibility value is almost 0. This might be the reason for sparsity in the interaction data.

Here we specially mention node 108. In the record, this person is always the first to communicate with others on each of the three days. His/her mixed-membership distribution is mainly composed of the sky blue community 1, which indicates he/she could be an organizer of this conference. The other nodes with mixed-membership distribution dominated by community 1, such as nodes 24, 53, 61, all were engaged actively with others according to the record.

Another interesting phenomenon is that the nodes containing the orange community 2 interact with community 1 at a probability of 0.2. This might be an indication that most of the attendees communicated with the organizers for various reasons.

VII. Conclusion

Modeling complex networking behaviors in a dynamic setting is crucial for widespread applications, including social media, social networks, online business, and market dynamic analysis. This challenges the existing learning systems that have limited power to address the dynamics. In this paper, we have provided a generalized and flexible framework to improve the popular MMSB by allowing a network to have infinite types of communities with relationships that change across time periods. By incorporating a time-sticky factor into the mixed-membership distributions, we have realistically modeled the time-correlation among latent labels. Both Gibbs sampling and adapted slice-efficient sampling have been used to infer the desired target distribution. Quantitative analysis on the MCMC's convergence behavior, including the convergence test, autocorrelation function, and so forth, has been provided to demonstrate the inference performance. The results of the experiments verify that our proposed DIM3 is effective in constructing the dynamic mixed-membership distribution and role-compatibility matrix.

Possible future work includes a systematic application of DIM3 to various large real-world social networks. In particular, we are also interested in adapting our model to many atypical applications, for example, where sequences of networks have nonbinary and directional measurements. We will also study many other flexible frameworks for modeling persistence of memberships across time. Lastly, we will perform an extensive study into patterns of joint dynamics of $\{\boldsymbol{\pi}_i^t\}$ to extract meaningful latent information from them. This is done in a setting where the number of components between $\boldsymbol{\pi}_i^t$ and $\boldsymbol{\pi}_i^{t+1}$ may differ.

Recent developments [46], [47] in the large-scale learning of latent space modeling give us more insights for possible future work. These improvements include parsimonious link modeling [46] that reduces the parameter size from $\mathcal{O}(n^2K^2)$ to $\mathcal{O}(n^2K)$, the utilization of the stochastic variational inference method [48], and a triangular representation of networks [49], [47], which could reduce the parameter size to $\mathcal{O}(nK^2)$. Through these, we are hoping to enlarge our model's scalability to millions of nodes and hundreds of communities.

To describe the time dependency, the dependent Dirichlet process (DDP) [50] provides an alternative. Among the various constructions of the DDP [51]–[55], we may construct the DDP by projecting the gamma process into different subspaces and normalizing them individually, through which the overlapping spaces reflect the correlation. Lin *et al.* [56] discuss the intrinsic relationship between the Poisson process, gamma process and Dirichlet process and uses three operations namely *superposition*, *subsampling*, and *point transition* to evolve from one Dirichlet process to another, with an elegant and solid theory support. Subsequent literatures including [57]–[59] extend this paper from different perspectives.

References

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

[2] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *Proc. 21st Nat. Conf. Artif. Intell. (AAAI)*, Jul. 2006, pp. 381–388. [Online]. Available: http://www.aaai.org/Library/AAAI/2006/aaai06-061.php

[3] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin, "Detecting communities and their evolutions in dynamic social networks—A Bayesian approach," *Mach. Learn.*, vol. 82, no. 2, pp. 157–189, 2011.

[4] K. Ishiguro, T. Iwata, N. Ueda, and J. B. Tenenbaum, "Dynamic infinite relational model for time-varying relational data analysis," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2010, pp. 919–927.

[5] 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, Jan. 2008.

[6] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for systems with state persistence," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 312–319.

[7] Q. Ho, L. Song, and E. P. Xing, "Evolving cluster mixed-membership blockmodel for time-evolving networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 342–350.

[8] E. Xing, W. Fu, and L. Song, "A state-space mixed membership blockmodel for dynamic network tomography," *Ann. Appl. Statist.*, vol. 4, no. 2, pp. 535–566, 2010.

[9] C. Heaukulani and Z. Ghahramani, "Dynamic probabilistic models for latent feature propagation in social networks," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 275–283.

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

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

[12] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "A sticky HDP-HMM with application to speaker diarization," *Ann. Appl. Statist.*, vol. 5, no. 2A, pp. 1020–1056, 2011.

[13] E. Fox, E. B. Sudderth, M. I. Jordan, and A. Willsky, "Bayesian nonparametric inference of switching dynamic linear models," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1569–1585, Apr. 2011.

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

[15] T. L. Griffiths and Z. Ghahramani, "The Indian buffet process: An introduction and review," *J. Mach. Learn. Res.*, vol. 12, pp. 1185–1224, Jul. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2021026.2021039

[16] Y. W. Teh, D. Görür, and Z. Ghahramani, "Stick-breaking construction for the Indian buffet process," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2007, pp. 556–563.

[17] J. Zhu, "Max-margin nonparametric latent feature models for link prediction," in *Proc. 29th Int. Conf. Mach. Learn. (ICML)*, 2012, pp. 719–726.

[18] T. Jebara, "Maximum entropy discrimination," in *Machine Learning*. New York, NY, USA: Springer-Verlag, 2004, pp. 61–98.

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

[20] P.-S. Koutsourelakis and T. Eliassi-Rad, "Finding mixed-memberships in social networks," in *Proc. AAAI Spring Symp. Social Inf. Process.*, 2008, pp. 48–53.

[21] Q. Ho, A. P. Parikh, and E. P. Xing, "A multiscale community blockmodel for network exploration," *J. Amer. Statist. Assoc.*, vol. 107, no. 499, pp. 916–934, 2012. [Online]. Available: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2012.682530

[22] 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, p. 7, Jan. 2010.

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

[24] P. Sarkar and A. W. Moore, "Dynamic social network analysis using latent space models," *ACM SIGKDD Explorations Newslett.*, vol. 7, no. 2, pp. 31–40, Dec. 2005.

[25] P. Sarkar, S. M. Siddiqi, and G. J. Gordon, "A latent space approach to dynamic embedding of co-occurrence data," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2007, pp. 420–427.

[26] J. R. Foulds, C. DuBois, A. U. Asuncion, C. T. Butts, and P. Smyth, "A dynamic relational infinite feature model for longitudinal social networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 287–295.

[27] W. Fu, L. Song, and E. P. Xing, "Dynamic mixed membership blockmodel for evolving networks," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 329–336.

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

[29] S. G. Walker, "Sampling the Dirichlet mixture model with slices," *Commun. Statist. Simul. Comput.*, vol. 36, no. 1, pp. 45–54, 2007.

[30] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Advances in Neural Information Processing Systems*, vol. 12. Cambridge, MA, USA: MIT Press, 2000, pp. 554–560.

[31] 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.

[32] M. Plummer, N. Best, K. Cowles, and K. Vines, "CODA: Convergence diagnosis and output analysis for MCMC," *R News*, vol. 6, no. 1, pp. 7–11, 2006.

[33] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statist. Sci.*, vol. 7, no. 4, pp. 457–472, 1992.

[34] J. Geweke, "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," in *Bayesian Statistics*, Oxford, U.K.: Oxford Univ. Press, 1992, pp. 169–193.

[35] P. Heidelberger and P. D. Welch, "A spectral method for confidence interval generation and run length control in simulations," *Commun. ACM*, vol. 24, no. 4, pp. 233–245, Apr. 1981.

[36] B. Kapferer, *Strategy and Transaction in an African Factory: African Workers and Indian Management in a Zambian Town*. Manchester, U.K.: Manchester Univ. Press, 1972.

[37] S. F. Sampson, "Crisis in a cloister," Ph.D. dissertation, Dept. Sociology, Cornell Univ., Ithaca, NY, USA, 1969.

[38] R. L. Breiger, S. A. Boorman, and P. Arabie, "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling," *J. Math. Psychol.*, vol. 12, no. 3, pp. 328–383, Aug. 1975.

[39] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *Machine Learning: ECML*. Berlin, Germany: Springer-Verlag, 2004, pp. 217–226.

[40] S. Asur, S. Parthasarathy, and D. Ucar, "An event-based framework for characterizing the evolutionary behavior of interaction graphs," *ACM Trans. Knowl. Discovery Data*, vol. 3, no. 4, p. 16, Nov. 2009.

[41] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng, "Analyzing communities and their evolutions in dynamic social networks," *ACM Trans. Knowl. Discovery Data*, vol. 3, no. 2, p. 8, Apr. 2009.

[42] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, "What's in a crowd? Analysis of face-to-face behavioral networks," *J. Theoretical Biol.*, vol. 271, no. 1, pp. 166–180, 2011.

[43] T. M. Newcomb, *The Acquaintance Process*. New York, NY, USA: Holt, Rinehart & Winston, 1961.

[44] S. C. Freeman and L. C. Freeman, "The networkers network: A study of the impact of a new communications medium on sociometric structure," School Social Sci., Univ. California, San Francisco, CA, USA, Tech. Rep. 46, 1979.

[45] J. S. Coleman, *Introduction to Mathematical Sociology*. New York, NY, USA: MacMillan, 1964.

[46] P. Gopalan, S. Gerrish, M. Freedman, D. M. Blei, and D. M. Mimno, "Scalable inference of overlapping communities," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2012, pp. 2249–2257.

[47] J. Yin, Q. Ho, and E. Xing, "A scalable approach to probabilistic latent space inference of large-scale networks," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2013, pp. 422–430.

[48] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1303–1347, 2013.

[49] D. R. Hunter, S. M. Goodreau, and M. S. Handcock, "Goodness of fit of social network models," *J. Amer. Statist. Assoc.*, vol. 103, no. 481, pp. 248–258, 2008.

[50] S. N. MacEachern, "Dependent nonparametric processes," in *Proc. Sec. Bayesian Statist. Sci.*, Alexandria, VA, USA, 1999, pp. 50–55.

[51] F. Caron, M. Davy, and A. Doucet, "Generalized Polya urn for time-varying Dirichlet process mixtures," in *Proc. Uncertainty Artif. Intell.*, 2007, pp. 33–40.

[52] Y. Chung and D. B. Dunson, "The local Dirichlet process," *Ann. Inst. Statist. Math.*, vol. 63, no. 1, pp. 59–80, 2011.

[53] D. B. Dunson, "Bayesian dynamic modeling of latent trait distributions," *Biostatistics*, vol. 7, no. 4, pp. 551–568, 2006.

[54] N. Bouguila and D. Ziou, "A Dirichlet process mixture of generalized Dirichlet distributions for proportional data modeling," *IEEE Trans. Neural Netw.*, vol. 21, no. 1, pp. 107–122, Jan. 2010.

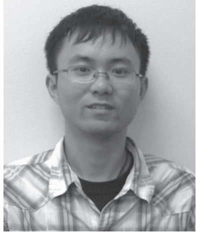
[55] V. Rao and Y. W. Teh, "Spatial normalized gamma processes," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2009, pp. 1554–1562.

[56] D. Lin, E. Grimson, and J. W. Fisher, III, "Construction of dependent Dirichlet processes based on Poisson processes," in *Advances in Neural Information Processing Systems Foundation*. Cambridge, MA, USA: MIT Press, 2010.

[57] D. Lin and J. W. Fisher, "Coupling nonparametric mixtures via latent Dirichlet processes," in *Advances in Neural Information Processing Systems*, vol. 25. Cambridge, MA, USA: MIT Press, 2012, pp. 55–63.

[58] C. Chen, N. Ding, and W. L. Buntine, "Dependent hierarchical normalized random measures for dynamic topic modeling," in *Proc. 29th Int. Conf. Mach. Learn.*, Jun. 2012, pp. 895–902.

[59] C. Chen, V. Rao, W. Buntine, and Y. W. Teh, "Dependent normalized random measures," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 969–977.



Xuhui Fan received the bachelor's degree in mathematical statistics from the University of Science and Technology of China, Hefei, China, in 2010. He is currently pursuing the Ph.D. degree with the University of Technology at Sydney, Sydney, NSW, Australia.

His current research interests include statistical machine learning.

Longbing Cao (SM'06) received the Ph.D. degrees in pattern recognition and intelligent systems and computing sciences.

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

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 at Sydney (UTS), Sydney, 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.