2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Infinite Motif Stochastic Blockmodel for Role Discovery in Networks

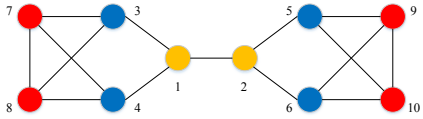
Yulong Pei[1], Jianpeng Zhang[2], George Fletcher[1], Mykola Pechenizkiy[1]
[1]Eindhoven University of Technology, Eindhoven, the Netherlands
[2]National Digital Switching System Engineering & Technological R&D Center, Zhengzhou, China
Emails: {y.pei.1,g.h.l.fletcher,m.pechenizkiy}@tue.nl, zjp@ndsc.com.cn

***Abstract*—Role/block discovery is an essential task in network analytics so it has attracted significant attention recently. Previous studies on role discovery either relied on first or second-order structural information to group nodes but neglected the higher-order information or required the number of roles/blocks as the input which may be unknown in practice. To overcome these limitations, in this paper we propose a novel generative model, infinite motif stochastic blockmodel (IMM), for role discovery in networks. IMM takes advantage of high-order motifs in the generative process and it is a nonparametric Bayesian model which can automatically infer the number of roles. To validate the effectiveness of IMM, we conduct experiments on synthetic and real-world networks. The obtained results demonstrate IMM outperforms other blockmodels in role discovery task.**

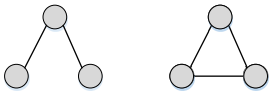
I. Introduction

Role/block discovery plays an essential role in network analytics. Role discovery can be defined as the process that takes a graph and picks out sets of nodes with similar structural patterns [1]. Discovering roles in networks can shed light on numerous graph mining tasks. Therefore, the problem of role discovery has attracted significant attentions recently. Previous work on role discovery can be categorized into two types: graph-based methods and feature-based methods [1], [2]. The most representative graph-based method is stochastic blockmodel [3], [4], [5]. Feature-based methods consist of two steps: feature extraction and role assignment, e.g., RolX [6]. However, previous studies either relied on first or second-order structural information to group nodes but neglected the higher-order information or required the number of roles/blocks as the input but failed to infer it automatically.

Therefore, two challenges remain in role discovery task. The first challenge is how to capture the high-order graph structures. Since role discovery analyzes networks from the global perspective, the first and second-order structural patterns, i.e., edges and shared neighbors, may fail in identifying roles. Another issue in using first and second-order structural patterns is they are incapable of dealing with the sparsity of networks. However, in real-world networks, both the direct edges and shared neighbors are often sparse. The second challenge is how to determine the number of roles/blocks. In the literature, there are two types of methods to select the right number of blocks/roles: (1) nonparametric models and (2) model selection methods. Nonparametric models, e.g., IRM [5], automatically choose an appropriate number of blocks using stochastic process as the prior that can generate an infinite number of clusters. Most widely used model selection methods include Bayesian information criterion (BIC) [7] and minimum description length (MDL) [8]. For instance, MMSB [4] uses BIC to choose the number of blocks and RolX [6] selects the number of roles using MDL. To illustrate these challenges, a motivating example is shown in Fig. 1.



(a) Borgatti-Everett network [9]. All nodes have the same degree and different colors denote different roles/blocks.



(b) Two types of motifs we consider in this paper.

Fig. 1. The motivating example on the Borgatti-Everett network.

TABLE I. Numbers of different types of motifs on the Borgatti-Everett network.

Node ID	1	2	3	4	5	6	7	8	9	10
Type 1	9	9	6	6	6	6	3	3	3	3
Type 2	0	0	1	1	1	1	2	2	2	2

A motivating example (Borgatti-Everett Network). Fig. 1(a) shows the Borgatti-Everett network [9] which consists of ten nodes. For simplicity, we focus on undirected graphs and only consider two types of motifs shown in Fig. 1(b) in this study. Based on the social theory in roles (a.k.a. positions) [7], these nodes can be grouped into three roles (in different colors). These three roles can be interpreted as star (Node 1 and 2 in yellow), periphery (Node 3 - 6 in blue) and clique (Node 7 - 10 in red). Traditional methods based on first and second-order structures have problem clustering these nodes into right roles because: (1) all ten nodes have the same degree but they are in different roles, and (2) some nodes (e.g., Node 7 and 9) have no shared neighbors but they belong to the same role. However, motif-based method can solve this problem

ASONAM '19, August 27-30, 2019, Vancouver, Canada

ACM ISBN 978-1-4503-6868-1/19/08/$15.00
https://doi.org/10.1145/3341161.3342921

effectively. We count the numbers of two different types of motifs (shown in Fig. 1(b)) for nodes in the Borgatti-Everett network and the statistics is shown in Table I. It is clear that these nodes can be clustered into three groups ($\{1, 2\}$, $\{3, 4, 5, 6\}$ and $\{7, 8, 9, 10\}$) since they have exactly the same motif distributions inside each role group. Thus, the failure of traditional methods in role discovery encourages a new model which can take high-order motifs into consideration. Another issue is that in practice, we may not know the number of roles in advance. Thus, it is meaningful to determine the right number of roles automatically by the model itself.

To tackle these two challenges motivated by the above example, we propose a novel generative model named Infinite Motif Stochastic BlockModel (IMM). On the one hand, IMM is a high-order model which takes advantage of the motifs in the generative process. On the other hand, it is a nonparametric Bayesian model which can automatically infer the number of roles from the data. To validate the effectiveness of IMM, we conduct experiments in role discovery on both synthetic and real-world networks. We evaluate discovered roles quantitatively on synthetic networks and visualize the results on real-world networks as a case study.

The contributions of our work are summarized as follows:

- We propose a novel generative model, infinite motif stochastic blockmodel (IMM), to discover roles in networks. IMM is a nonparametric Bayesian model to generate higher-order motif information.
- We derive Gibbs sampling algorithm for model inference to learn the latent variables in IMM.
- The conducted experimental results on both synthetic and real-world networks demonstrate the effectiveness of IMM in role discovery.

II. Notations and Backgrounds

We first summarize some notations in Table II and briefly review the related models, then present the problem statement of the motif-based nonparametric model for role discovery.

Stochastic blockmodel (SBM) [3] is the original generative model to detect blocks in networks. SBM partitions nodes in hard clustering and is not flexible to incorporate prior knowledge. To solve these problems, mixed membership stochastic blockmodel (MMSB) [4] has been proposed where a role distribution for each node will be inferred and the graphical representation of MMSB is shown in Fig. 2(a). Formally, the generative process of MMSB is:

$$\begin{aligned} \theta|\alpha &\sim Dirichlet(\eta) \\ z|\theta &\sim Multinomial(\theta) \\ B_{ij}|\beta^1,\beta^2 &\sim Beta(\beta^1,\beta^2) \\ r_{ij}|z,B &\sim Bernoulli(B(z_i,z_j)) \end{aligned} \tag{1}$$

where $a|b \sim Distribution(b)$ means that sampling the variable a from the distribution with parameter b.

However, MMSB requires the number of roles/blocks as the input which may be difficult to obtain in advance in practice. Infinite relational model (IRM) [5] has been proposed using the Chinese restaurant process (CRP) [11] as the prior. As a discrete-time stochastic process, CRP can generate an infinite number of clusters so IRM can infer the right number of roles based on the observed edges. The graphical representation of IRM is shown in Fig. 2(b) and the formal generative process is:

$$\begin{aligned} z|\alpha &\sim CRP(\alpha) \\ B_{ij}|\beta^1,\beta^2 &\sim Beta(\beta^1,\beta^2) \\ r_{ij}|z,B &\sim Bernoulli(B(z_i,z_j)) \end{aligned} \tag{2}$$

Note that in both MMSB and IRM, the essence is to infer the latent variables, i.e., roles, based on the observed edges between nodes.

Although IRM can infer the role number automatically, it only models the first-order patterns, i.e., edges, in networks. Mixed membership triangular model (MMTM) [10] solved this problem by modeling higher-order motifs and MMTM aims to infer roles from the observed motifs in networks. The graphical representation of MMTM is shown in Fig. 2(c) and the formal generative process is:

$$\begin{aligned} \theta|\alpha &\sim Dirichlet(\alpha) \\ z|\theta &\sim Multinomial(\theta) \\ B_{uvw}|\beta &\sim Dirichlet(\beta) \\ m_{uvw}|z,B &\sim Multinomial(B(z_u,z_v,z_w)) \end{aligned} \tag{3}$$

Problem Statement Given a network G represented by a motif sequence $M = \{(e_u, e_v, e_w, t)\}$ where each motif is a 4-tuple, $e_u, e_v, e_w \in E$, and $t \in R$ is the type of motif (in this work there are two motif types shown in Fig. 1(b)), the motif-based nonparametric model for role discovery aims to assign each node to a role and infer the number of roles simultaneously.

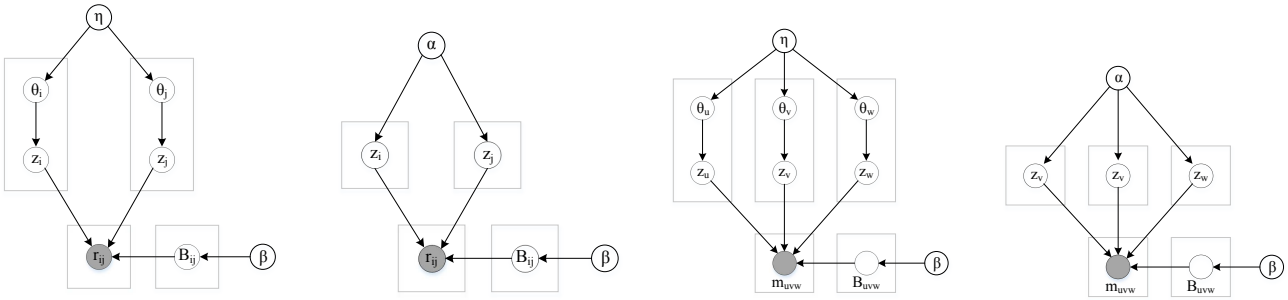
TABLE II. Summary of the notations.

Symbol	Description
N	Number of nodes.
$n_{i,k}$	Number of times that node i being assigned to role k
m_{uvw}	Motif type for node u, v and w
B	Role compatible matrix
z_i	Role assignment for node i
Z_{-i}	Role assignment for all nodes except node i
θ	Role interaction matrix
α	Hyperparameter of CRP
η	Hyperparameter of Dirichlet distribution
β	Hyperparameter of Beta distribution
Distribution	**Description**
$Dirichlet(\eta)$	Dirichlet distribution with parameter η
$Multinomial(\theta)$	Multinomial distribution with parameter θ
$Beta(\beta^1,\beta^2)$	Beta distribution with parameter β^1 and β^2
$Bernoulli(\omega)$	Bernoulli distribution with parameter ω
$CRP(\alpha)$	Chinese restaurant process with parameter α

Algorithm 1 Gibbs Sampling Algorithm for IMM

Input: A network G represented by a motif sequence $M = \{(e_u, e_v, e_w, type)\}$, hyperparameter for CRP prior α, and hyperparameter for Dirichlet prior β
Output: Role assignment z and role interaction pattern B

```
Initialize z for each node
while not converge do
    Update role assignment statistics n_{i,k}
    Sample role for each node by assigning it into existing roles or a new role according to Eq. (7)
end while
```



(a) MMSB [4] (b) IRM [5] (c) MMTM [10] (d) IMM (our model)

Fig. 2. Graphical representations of different stochastic blockmodels.

III. Infinite Motif Stochastic Blockmodel

Infinite Motif Stochastic Blockmodel (IMM) takes motifs as the input which can effectively capture the high-order structural patterns of networks. It also utilizes CRP as the prior which can infer the role numbers automatically. Now we introduce the generative process of IMM and the inference algorithm in detail.

A. The Generative Model

The graphical representation of IMM is shown in Fig. 2(d) and the formal generative process is:

$$\begin{aligned} z|\alpha &\sim CRP(\alpha) \\ B_{ijk}|\beta &\sim Dirichlet(\beta) \\ m_{uvw}|z, B &\sim Multinomial(B(z_u, z_v, z_w)) \end{aligned} \tag{4}$$

In detail, in the first line of Eq. (4), IMM employs the CRP as the prior for the latent variables, i.e., roles of nodes, so it can generate an infinite number of roles based on the input networks. Similar to MMSB, in the second line the Dirichlet distribution is used as the prior to generate the role interaction B_{ijk} for three nodes which form a predefined motif. In the last line, the observed motif is generated from a multinomial distribution with the role interaction as parameter.

The posterior of Z can be computed as:

$$P(Z|M, \alpha, \beta) \propto P(M|Z)P(Z|\alpha), \tag{5}$$

where the probability $P(M|Z)$ of generating M is:

$$P(M|Z) \propto \int P(M|Z, B, \beta)P(B|\beta)dB. \tag{6}$$

B. The Inference Algorithm

The aim of inference algorithm for IMM is to infer the latent variable z and B based on the observed motif sequence. In this work, we use Gibbs sampling algorithm to approximate the conditional probability of role assignment since it is a moderately efficient method for Bayesian models. Due to the page limit, we omit the detailed process but list the conditional distribution for sampling as follows:

$$P(z_i = k|M, Z_{-i}, \alpha, \beta) \propto \tag{7}$$
$$\begin{cases} \frac{\alpha}{N-1+\alpha} \prod_{u,v,w} \Gamma(m_{uvw} + \beta) & \text{for existing cluster} \\ \frac{n_{i,k}}{N-1+\alpha} \prod_{u,v,w} \Gamma(m_{uvw} + \beta) & \text{for new cluster} \end{cases}$$

where $\Gamma(n) = (n-1)!$ is the Gamma function. The Gibbs sampling algorithm for IMM inference is shown in Algorithm 1.

IV. Experiments

A. Experimental Setup

We evaluate our model on role discovery using synthetic and real-world networks. For synthetic data, we generate two networks using SBM [3] and they are visualized in Fig. 3. Each network consists of 100 nodes and these nodes belong to 4 roles. For real-world networks, we use Zachary karate and Les Misérables network. IMM is compared to some baseline models including MMSB, MMTM and IRM. We use the same (hyper)parameters in all models: CRP parameter $\alpha = 5$, Dirichlet parameter $\eta = 4$ and Beta parameter $\beta = 7$. We leave the hyperparameter selection as our future work.

B. Evaluation Metrics

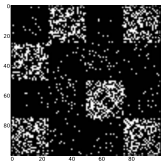
To evaluate the experimental results, we use purity and normalized mutual information (NMI) as the evaluation metrics. These metrics are widely used in the evaluation of clustering methods with ground-truth labels.

Purity measures the extent to which each cluster contained data points from primarily one class. The purity of a clustering is obtained by the weighted sum of individual cluster purity values which defined as:

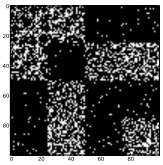
$$Purity = \frac{1}{N} \sum_{i=1}^{k} max_j |c_i \cap t_j|, \tag{8}$$

where N is number of objects, k is number of clusters, c_i is a cluster in C, and t_j is the classification which has the max count for cluster c_i. NMI evaluates the clustering quality based on information theory, and is defined by normalization on the mutual information between the cluster assignments and the pre-existing input labeling of the classes:

$$NMI(\mathcal{C}, \mathcal{D}) = \frac{2 * \mathcal{I}(\mathcal{C}, \mathcal{D})}{\mathcal{H}(\mathcal{C}) + \mathcal{H}(\mathcal{D})}, \tag{9}$$



(a) Diagonal block structures.



(b) Non-diagonal block structures.

Fig. 3. Visualization of two synthetic networks.

TABLE III. EXPERIMENTAL RESULTS ON THE SYNTHETIC NETWORKS.

datasets	SYN 1		SYN 2	
metrics	Purity	NMI	Purity	NMI
MMSB [4]	0.45	0.2603	0.33	0.1681
IRM [5]	0.45	0.2452	0.44	0.1758
MMTM [10]	0.60	0.3964	0.48	**0.2453**
IMM (our model)	**0.65**	**0.4204**	**0.51**	0.2242

where obtained cluster $\mathcal{C}$ and ground-truth cluster $\mathcal{D}$. The mutual information $\mathcal{I}(\mathcal{C},\mathcal{D})$ is defined as $\mathcal{I}(\mathcal{C},\mathcal{D}) = \mathcal{H}(\mathcal{C}) - \mathcal{H}(\mathcal{C}|\mathcal{D})$ and $\mathcal{H}(\cdot)$ is the entropy.

C. Results

1) Synthetic networks: The results of role discovery on two synthetic networks are shown in Table III. From these results, some conclusions can be drawn:

- IMM outperform all other models in detecting roles which indicates the effectiveness of our model.
- Motif-based models perform better than edge-based models which demonstrate that high-order motifs can better capture the global role information.
- Nonparametric models (IMM and IRM) can effectively detect the number of roles which is more meaningful in practice when the role number is unknown.

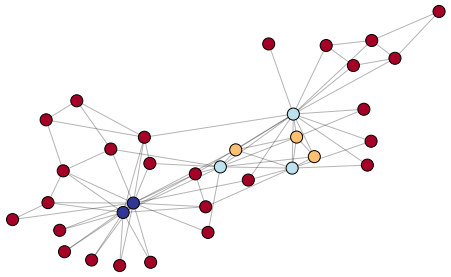
2) Real-world networks: The visualization of roles on two real-world networks are shown in Fig. 4(a) and 4(b). These results, demonstrate the effectiveness of IMM in identifying roles. In detail,

- *Zachary network*: The two blue nodes are stars for the left community, yellow and light blue are stars for the right community[1], and red nodes are peripheries.
- *Les Misérables network*: The blue nodes are stars including dark and sky blue nodes, red nodes are cliques, orange nodes are peripheries, and yellow nodes are bridges to link stars and followers.

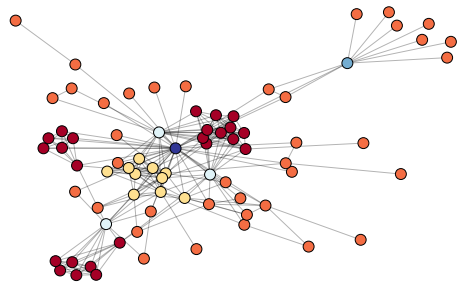
V. CONCLUDING REMARKS

In this work we proposed a novel generative model, infinite motif stochastic blockmodel (IMM), for role discovery. IMM is advantageous in two aspects: (1) it models higher-order motifs to infer the roles which can effectively capture the global structural information of networks, and (2) it is a nonparametric Bayesian model to infer the number of roles automatically which is more suitable in real-world network analytics. We evaluated IMM in role discovery compared to state-of-the-art blockmodels and the results indicate the effectiveness of IMM. In future work we will explore scalable inference algorithm for IMM, e.g., collapsed variational Bayesian inference method, and test our model on larger-scale networks. We will also extend our method to different types of networks, e.g., dynamic networks.

[1]We have prior knowledge on Zachary karate network that it consists of two communities.



(a) Roles on Zachary network.



(b) Roles on Les Misérables network.

Fig. 4. Visualization of roles on two real-world networks.

REFERENCES

[1] R. A. Rossi and N. K. Ahmed, "Role discovery in networks," *IEEE TKDE*, vol. 27, no. 4, pp. 1112–1131, 2015.

[2] Y. Pei, J. Zhang, G. Fletcher, and M. Pechenizkiy, "Dynmf: role analytics in dynamic social networks," in *IJCAI*. AAAI Press, 2018, pp. 3818–3824.

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

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

[5] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *AAAI*, vol. 3, 2006, p. 5.

[6] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li, "Rolx: structural role extraction & mining in large graphs," in *KDD*. ACM, 2012, pp. 1231–1239.

[7] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[8] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[9] S. P. Borgatti and M. G. Everett, "Notions of position in social network analysis," *Sociological methodology*, pp. 1–35, 1992.

[10] J. Yin, Q. Ho, and E. P. Xing, "A scalable approach to probabilistic latent space inference of large-scale networks," in *NIPS*, 2013, pp. 422–430.

[11] J. Pitman *et al.*, "Combinatorial stochastic processes," Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for , Tech. Rep., 2002.