Integrating Community and Role Detection in Information Networks

Ting Chen*† Lu-An Tang‡ Yizhou Sun† Zhengzhang Chen‡ Haifeng Chen‡
Guofei Jiang‡

Abstract

Community detection and role detection in information networks have received wide attention recently, where the former aims to detect the groups of nodes that are closely connected to each other and the latter aims to discover the underlying roles of nodes in the network. Traditional studies treat these two problems as orthogonal issues and propose algorithms for these two tasks separately. In this paper, we propose to integrate communities and roles in a unified model and detect both of them simultaneously for information networks. Intuitively, (1) correctly detecting the communities in a network will lead to the success of detecting roles of nodes, such as opinion leaders and followers in social networks; and (2) correctly identifying the roles of the nodes will lead to a better network modeling and thus a better detection of communities. A novel probabilistic network model, the Mixed Membership Community and Role model (MMCR), is then proposed, which models the latent community and role of each node at the same time, and the probability of links are defined accordingly. By testing our model on synthetic networks and two real-world networks, we demonstrate that our approach leads to better performance for both community detection and role detection. Moreover, our model has a better interpretation for link generation in networks according to the link prediction task.

1 Introduction

Discovering community structures in information networks has been a hot topic in the past few years [1, 2, 3, 4, 5]. The task of community detection is to find clusters of nodes that are closely connected within the same cluster and loosely connected between different clusters. Community detection is useful for understanding the underlying network structures. For example, detecting the communities of a scientific collaboration network can help reveal the underlying subfields of research. Meanwhile, researchers recently have shown increasing interests on automatically discovering different roles for nodes in information networks [6, 7, 8], which aims to identify different functions served by different nodes. The concept of roles, such as bridge, core/periphery and so on, has been found important in social and network theory [9, 10]. For example, it is reported that 1% of Twitter users who span structural holes control 25% of the information diffusion on Twitter [9, 8].

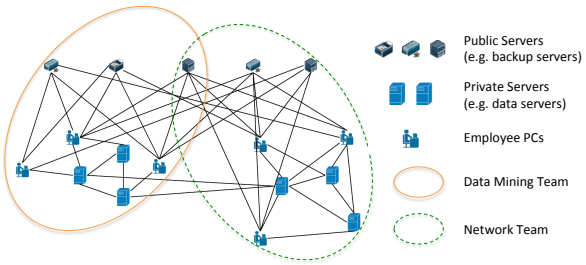


Figure 1: Enterprise machine communication network. Nodes denote machines and links denote their communications.

Traditional studies treat the two tasks as orthogonal problems: the existing community detection algorithms usually ignore the roles of nodes [2, 3, 5]; and most existing role discovery algorithms detect roles without taking community structures into account [4, 6, 7]. In real information networks, however, communities and roles are tightly coupled and cannot be separated, as shown in Example 1.

Example 1. (Enterprise Network). Figure 1 shows a machine communication network inside a modern enterprise, which is referred to enterprise network thereafter. There are two communities (i.e., the network team and the data mining team)[1] and three roles (i.e., public servers, private servers, and PCs) in this enterprise network. In some real scenarios, the community and role information is inaccessible or only partially

*Part of the work is done during first author's internship at NEC Labs America.

†College of Computer and Information Science, Northeastern University. {tingchen, yzsun}@ccs.neu.edu

‡NEC Labs America. {ltang, zchen, haifeng, gfj}@nec-labs.com

[1]Public servers are commonly shared among communities, but some might have stronger affiliations in certain communities due to reasons such as locality.

available. Intuitively, community detection and role detection can mutually enhance each other. On one hand, roles can be useful to model and detect communities. For example, private servers usually densely connect to other machines in the same community, and revealing private servers will better model a community and detect other machines in the community. On the other hand, community structure is important to detect roles. For example, in order to distinguish private servers from public servers, we can rely on their connection behavior across different communities: private servers mainly link to nodes within the same community (like local hubs), while public servers could link to nodes in different communities (like bridges).

Based on the above example, it is clear that nodes in the same community could have different linking patterns due to their roles, and the functions of some roles are dependent on communities. Hence, it is important to model the community and role together, and both tasks can benefit each other: (1) by identifying roles, the model can overcome the drawbacks of current network (generative) models, which treat nodes within a community equally [4, 5]; (2) by identifying communities, the model can clearly find the roles that are dependent on communities [8].

In this paper, we propose a coherent generative framework called MMCR, which can naturally integrate both community and role detection. The intuition behind our generative framework is: in order to capture both community and role structures simultaneously, every node is associated with not only community membership, but also role membership; when two nodes attempt to interact (i.e., form an edge between them), both community and role memberships should have impact on determining the link generation probability.

The contributions of this paper are summarized below:

- Studying a novel problem of integrating community and role detection in information networks.
- Proposing a unified probabilistic generative model that defines the link generation probability based on both community and role labels of nodes, and a Gibbs sampling based inference algorithm is proposed.
- Evaluating the proposed method on three synthetic networks and two real information networks, and demonstrating the effectiveness of our proposed model on both community/role detection and link prediction tasks.

2 Preliminaries and Problem Definition

Information networks are ubiquitous nowadays, examples include World Wide Web, scientific collaboration networks, and enterprise networks (as shown in Example 1). Consider an information network $G = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the node set and $\mathcal{E}$ is the link set. Network G is represented using an adjacency matrix E, where $E_{ij} = 1$ indicates a link from node i to node j, and $E_{ij} = 0$ indicates there is no link between the two nodes. We assume the network is undirected in this paper, and the proposed model can be extended to the directed ones naturally.

2.1 Communities and Roles in Information Networks. A community is usually considered as a cluster of nodes that are closely connected within the same cluster and loosely connected between different clusters [1, 2]. As in Example 1, machines from the same team form a community, as they are closely connected within the same team, and loosely connected with machines in other teams. In many scenarios, nodes in the network can belong to multiple communities simultaneously, thus we follow [4, 5] and adopt a mixed membership to denote community affiliations of a node. More specifically, a probabilistic membership vector π_i with the dimension as the number of communities is assigned to each node i, and π_{ik} indicates the probability of node i belonging to the k-th community.

Roles can be considered as the functions played by nodes in the network. The concept of roles has been explored in several social and network theories, such as bridges, core, and periphery [9, 10]. In real applications, roles are in general with different meanings. As seen in Example 1, three types of machines correspond to three roles. In this paper, we mainly consider roles that are tightly coupled with communities in the context of information/social network: local hubs inside the community (e.g., local servers), global hubs or bridges nodes connecting different communities (e.g., public servers), and periphery nodes (e.g., PCs). Since a node can have multiple roles in many real networks, similar to community modeling, we use a mixed membership for role affiliations of a node. Specifically, a probabilistic vector θ_i with the dimension the same as the the number of roles is assigned to each node i, and θ_{ik} is the probability of the node i serving as the k-th role.

2.2 Mixed Membership Stochastic Blockmodel. Mixed Membership Stochastic Blockmodel (MMSB) [4] is a well-known probabilistic generative model for discovering latent groups in networks. In this paper, we further extend MMSB model by integrating roles together with communities in generating links. In MMSB, each node is associated with a soft group membership vector π_i, which is further drawn from a Dirichlet distribution, and the link generation probability given the group labels is determined by

a group-group interaction probability matrix S. The generative process of links can then be described as:
For each node i:
- Draw a group membership distribution vector $\pi_i \sim$ Dirichlet(α)

For each node pair (i, j):
- Draw node i's latent group $Z_{ij} \sim$ Multinomial(π_i)
- Draw node j's latent group $Z_{ji} \sim$ Multinomial(π_j)
- Draw the link $E_{ij} \sim$ Bernoulli($S_{Z_{ij},Z_{ji}}$)

There are some variations of MMSB. For example, motivated by the assortative property of communities, assortative MMSB (aMMSB) constrain the group-group interaction matrix S to be almost diagonal (with very small interaction probability across groups). Although MMSB and aMMSB can be used to detect communities or roles by modeling the pair-wise interaction of nodes, it is clear that neither MMSB nor aMMSB can directly model both community and role simultaneously since each node only has one membership vector.

2.3 The Integrated Community and Role Detection Problem. Given an information network G, the goal is to integrate communities and roles in a unified model, and automatically infer the community and role memberships π and θ for nodes in the network.

3 The Mixed Membership Community and Role (MMCR) Model

We now introduce our proposed mixed membership community and role detection model (MMCR) in detail.

3.1 Motivation. In order to integrate community and role detection, we propose to model them in a unified generative model. We have observed in Example 1 that, even nodes in the same community can have different linking probabilities due to their roles, and the linking patterns of roles are dependent on communities. Thus, we think that a unified model should attribute the link formation to both community and role memberships of nodes, instead of only one type of latent groups as in MMSB.

3.2 The Generative Model. Let π_i and θ_i be the clustering and role membership vectors for node i, respectively, for a pair of nodes (i, j), their community and role assignments $(Z^c_{ij}, Z^c_{ji}, Z^r_{ij}, Z^r_{ji})$ are drawn according to the multinomial distribution parametrized by their membership distribution vectors. Then a link is formed according to a Bernoulli distribution, whose parameter (denoted as B) is dependent on both community and role assignments $(Z^c_{ij}, Z^c_{ji}, Z^r_{ij}, Z^r_{ji})$, which will be specified below.

Intuitively, (1) when the two nodes are from different communities, they tend to have small interaction probability, unless they are roles like global hubs or bridges, which we call role-based background connection; and (2) when two nodes are in the same community, they usually interact with each other via higher probability, but the quantity is still dependent on the role memberships of the two nodes, which we call role-based within community connection. For the first type of connections, we use B_0 matrix to denote all the role-role interaction probablities in the background; and for the second type of connections, we use $B_k (k > 0)$ matrix to denote all the role-role interaction probabilites in community k. Formally, we define δ-function according to two community assignments as:
(3.1)

$$\delta(a,b) = \begin{cases} k, & \text{if } a = b = k, k > 0 \\ & \text{(i.e. two nodes are in the } k\text{-th community)} \\ 0, & \text{otherwise} \\ & \text{(i.e. two nodes are in different communities)} \end{cases}$$

We then use $B_{\delta(Z^c_{ij},Z^c_{ji}),Z^r_{ij},Z^r_{ji}}$ to denote the connection probability between a pair of nodes with the community and role assignments. It is not difficult to see that when two nodes are in different communities, i.e., $\delta(Z^c_{ij}, Z^c_{ji}) = 0$, the interaction probability is then $B_{0,Z^r_{ij},Z^r_{ji}}$. In this case, only their role assignments affect their interaction probability. When two nodes are in the same k-th community, i.e., $\delta(Z^c_{ij}, Z^c_{ji}) = k$, the interaction probability is then $B_{k,Z^r_{ij},Z^r_{ji}}$, which indicates that the interaction probability depends on both the community and their roles.

To incorporate some prior knowledge we have about roles, such as that a local hub should have higher linking probability in its community, we put priors to interaction parameter B, as well as other parameters. The generative process of the proposed model (MMCR) can be summarized as follows:
For each entry (k, p, q) in B (k can take 0 here):
- Draw $B_{k,p,q} \sim$ Beta($\xi^1_{k,p,q}, \xi^2_{k,p,q}$)

For each node i:
- Draw a community membership distribution vector $\pi_i \sim$ Dirichlet(α^c)
- Draw a role membership distribution vector $\theta_i \sim$ Dirichlet(α^r)

For each node pair (i, j):
- Draw node i's community $Z^c_{ij} \sim$ Multinomial(π_i)
- Draw node j's community $Z^c_{ji} \sim$ Multinomial(π_j)
- Draw node i's role $Z^r_{ij} \sim$ Multinomial(θ_i)
- Draw node j's role $Z^r_{ji} \sim$ Multinomial(θ_j)
- Draw link $E_{ij} \sim$ Bernoulli($B_{\delta(Z^c_{ij},Z^c_{ji}),Z^r_{ij},Z^r_{ji}}$)

It is worth noting that the proposed MMCR generalizes both MMSB and aMMSB. If we set the number of communities to one, we obtain MMSB; if we set the number of roles to be one, we recover aMMSB.

4 The Inference Algorithm

Given the network data, we need to infer the posterior distribution of the variables in the model, e.g., the community and role membership distribution vectors π, θ. Due to the complicated integrals over hidden states in the posterior inference, exact inference is intractable [11], thus we adopt Gibbs sampling inference [12].

4.1 The Conditional Distribution. We use the collapsed Gibbs sampling [12] for the learning, in which the continuous Dirichlet membership variables θ and π are integrated out. Only the membership assignments of a pair of nodes are sampled at a time according to their conditional distribution. The conditional distribution $P(Z^c_{ij} = a, Z^c_{ji} = b, Z^r_{ij} = p, Z^r_{ji} = q|E_{ij}, Z_{-ij}, \alpha^c, \alpha^r, \xi^1, \xi^2)$, which is the community and role membership assignments of a pair of node i, j given the link observation E_{ij} and the current assignments of the rest node pairs $Z_{-ij} = \{Z^c_{-ij}, Z^r_{-ij}\}$, is derived as follows (the detailed derivation can be found in supplementary material):

(4.2)
$$P(Z^c_{ij} = a, Z^c_{ji} = b, Z^r_{ij} = p, Z^r_{ji} = q|E_{ij}, Z_{-ij}, \alpha^r, \alpha^c, \xi^1, \xi^2)$$
$$\propto \frac{(n^{-ij}_{\delta(a,b)pq+} + \xi^1)^{E_{ij}} (n^{-ij}_{\delta(a,b)pq-} + \xi^2)^{1-E_{ij}}}{n^{-ij}_{\delta(a,b)pq+} + n^{-ij}_{\delta(a,b)pq-} + \xi^1 + \xi^2}$$
$$(h^{-ij}_{ia} + \alpha^c)(h^{-ij}_{jb} + \alpha^c)(m^{-ij}_{ip} + \alpha^r)(m^{-ij}_{jq} + \alpha^r)$$

It is worth noting that this conditional distribution is proportional to two parts: (1) the rate of link or non-link given the community and role assignments of the two nodes, and (2) the ratio (after normalization) of community and role membership assignments of both nodes. Both parts are calculated by excluding current community and role assignments of the pair of nodes.

4.2 The Sampling Procedure and Parameter Estimation. Having obtained the conditional distribution, the collapsed Gibbs sampling algorithm is straightforward. One can initialize the Markov chain by some random community and role membership assignments for all node pairs, and then run the chain by sequentially re-sampling assignments of each pair of nodes conditioned on the rests according to E.q. (4.2). Once the assignments of a pair of nodes are updated, the counters n, m, h in E.q. (4.2) are also updated intermediately. After enough number of iterations, the Markov chain approaches the equilibrium distribution, and then the subsequent samples of the community and role assignments can be collected to estimate the posterior distribution of variables, such as the role membership distribution vector θ_i,community membership distribution vector π_i, and the role-role interaction matrices B.

The community membership of node i is also Dirichlet distributed, and its mean at a-th dimension is:

$$\pi_{ia} = \frac{h_{ia} + \alpha^c}{\sum_{a=1}^{K^c} h_{ia} + K^c \alpha^c} \tag{4.3}$$

The role membership of node i is Dirichlet distributed with mean at p-th dimension given by:

$$\theta_{ip} = \frac{m_{ip} + \alpha^r}{\sum_{p=1}^{K^r} m_{ip} + K^r \alpha^r}, \tag{4.4}$$

Finally the interaction tensor B is Beta distributed, the mean of each entry can be estimated by:

$$B_{kpq} = \frac{n_{kpq+} + \xi^1}{n_{kpq+} + n_{kpq-} + \xi^1 + \xi^2}. \tag{4.5}$$

We also note that the generative models of all MMSB, aMMSB and MMCR require computations that are square to the number of nodes in the network. Stochastic inferences techniques such as stochastic variational inference[5], or sub-sampling the non-existing links can be adopted to speed up the inferences. But in the experiments below, we simply follow the exact inferences process as described here.

5 Experiments

In this section, we conduct two types of experiments in real information networks: (1) community and role detection, and (2) link prediction. The former evaluates the performance of MMCR on detecting communities and roles in real networks, by comparing metrics like accuracy when ground-truth is available, as well as by case studies. The latter is designed to verify if our network generative model MMCR is reasonable, as a superior generative model is expected to better predict previously unseen data. We also conduct some experiments on synthetic community and role data to verify the generative model.

5.1 Data Sets. Experiments are mainly conducted on three synthetic networks, and two real-world networks: the enterprise machine communication network and Enron employee email communication network. For synthetic networks description, we leave it to the synthetic experiments subsection later. Here we briefly introduce the two real data sets:

Enterprise Network. This dataset is similar to Example 1. The data set was collected inside a department consisting of three different groups over a month. It contains 73 nodes and 694 links, and has three different roles (PC, Private Server, and Public Server) and three different communities. It is worth mentioning that the number of nodes in different communities and roles are imbalanced (the communities contain 11, 16, 22 nodes respectively, the other 27 nodes are public servers that are commonly shared among communities).

Enron Network. Enron employees' email communication network is created using original Enron email data set [13], which contains email communication records not only between Enron employees but also with people outside the company. To study the communities and roles inside Enron company, we constrain the communication network among those Enron employees. The links are binarized, indicating whether two employees have communication history. Finally, there are 155 nodes, and 3572 links (counting both directions). Unlike Enterprise network, we do not have the ground-truth communities and roles available in Enron network. However, we do have some meta-information about the network which enables a case study, including the titles and group affiliations of the Enron employees.

5.2 Evaluation Metrics. For the first task of community and community detection, we consider two metrics: Normalized Mutual Information and accuracy. NMI is widely used for evaluation of clustering results as well as community detection [5], thus it is used in our experiments when algorithms are run with no seed/label is provided. However, when a few labels on nodes are provided, the community detection task then becomes semi-supervised, thus accuracy is used for evaluation. Given two clusters C_1 and C_2, NMI is defined as follows:

$$\text{NMI}(C_1, C_2) = \frac{\text{MI}(C_1, C_2)}{[\text{H}(C_1) + \text{H}(C_2)]/2} \tag{5.6}$$

Where $\text{MI}(C_1, C_2)$ is the mutual information of the two clusters, and $\text{H}(C_i)$ is the entropy of the cluster C_i.

For the second task of link prediction, we utilize three metrics: perplexity, AUC and AP. Perplexity is the exponential of the average negative log-likelihood of the held-out node pairs, which measures the generalization of the network generative model [5] (the calculation of perplexity is given in appendix). AUC and AP are widely used in link prediction. AUC is area under ROC (Receiver Operating Characteristic) Curve, it reflects the probability of a positive instance (a link, in our case) can be ranked higher than the negative instance (a non-link in our case). There is evidence showing for imbalanced data (such as in link prediction, there are much more non-links than links), AUC is less discriminative for distinguishing the performance of different models [14]. Thus we include the metric of Average Precision (i.e. the area under the Precision Recall Curve), abbreviated as AP, which might be a more discriminative metric. The calculation of AUC and AP can be found in [14].

5.3 Experimental Settings Baselines. On both community/role detection and link prediction tasks, we mainly compare our method with MMSB [4] and aMMSB [5], as they are both state-of-the-art network generative models closely related to community and role detection. Since none of these methods can integrate both tasks, thus they are applied to each of the community and role detection tasks at a time. To be more clear, we treat the network as the input for MMSB/aMMSB, and the output latent groups of MMSB/aMMSB as potential communities or roles. Noted that if the community or role labels on nodes are provided (in the inference they will also be fixed), MMSB/aMMSB can be guided to detect such desired groups, but without labels, their discovered latent groups for community detection and role detection are the same.

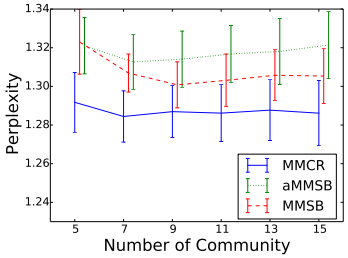


Figure 2: Perplexities under different number of communities on Enron network. The smaller the better.

Number of roles and communities. For enterprise network, we set the numbers of communities and roles according to the ground truth. For Enron network, the number of communities is estimated from data, while the number of roles is set to three since it can give us most meaningful roles as explained in Section 2. The number of communities is chosen by the perplexity on 10% hold-out node pairs. The perplexities for MMCR, aMMSB and MMSB are shown in Figure 2. According to the smallest perplexities, for the following experiments, we choose the number of communities to be 7 for both MMCR and aMMSB, but 9 for MMSB.

For the first task of community and role detection, there are community and role labels available in Enterprise network, which enables us to evaluate the results directly. We conduct the experiments in two settings: one is completely unsupervised, the other is with a few seeds/labels provided (some nodes are selected and labeled in the input). We run each model 20 times, and choose the top 10 according to their likelihoods, and then calculate their means and standard deviations. However, in Enron network, we do not have the community and role ground truth, thus a case study is provided.

For the second task of link prediction, we split the edges in both networks into 10 folds randomly, and for each fold, we sample the number of non-links at network

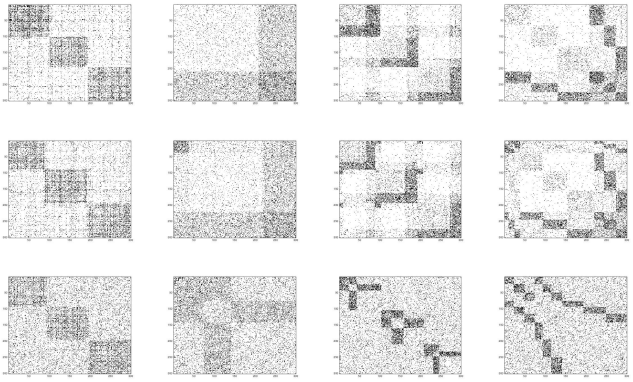


Figure 3: Synthetic networks. Each row is for one synthetic network. Each column is a type of organization of the same adjacency matrix, and the layout of nodes is determined by their communities/roles. The matrices in the first column are organized by community; the ones in the second column are organized by role; for the third column, they are organized first by community, and by role for nodes within the same community; finally for the last column, they are organized first by role, then by community for nodes within the same role.

sparsity[2], which resembles a 10-fold cross-validation.

For hyper-parameter α in all three models, we set it to $1/K$ (here K is the number of groups), and the (ξ^1, ξ^2) for B is set to $(15, 1)$ for diagonals, and $(1, 15)$ for non-diagonals. For enterprise network, due to the limited size of data, as well as the imbalanced communities, we use a constraint version of MMCR, where all three communities share the same role-role interaction matrix, that is to use a binarized version of δ-function.

5.4 Synthetic Experiments and Analysis. We create three synthetic networks using the generative process of MMCR described above. The first synthetic network has three communities, and each community has a core/periphery structure. Within the same community, the core nodes are densely connected with both core and periphery nodes, while the periphery nodes are less densely link to other periphery nodes. The second synthetic network has not only core/periphery roles, but also has an additional bridge role. The bridge nodes have higher probability of linking to nodes in the other communities. The last synthetic network contains roles that are neither core/periphery nor bridge structures. In this network, we create roles inside each community that form a hierarchy.

[2]Since both two real networks are sparse (there are far more non-links than links), it is more accurate to use test sets with the ratio of links and non-links that is the same with real networks.

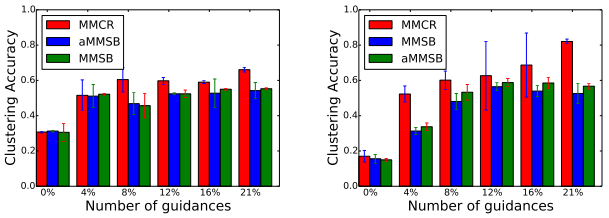
Table 1: NMI results on synthetic networks.

	Role			Community		
	Net1	Net2	Net3	Net1	Net2	Net3
MMCR	**0.76**	**0.69**	**0.74**	**1.0**	**0.97**	**0.90**
MMSB	0.58	0.41	0.37	0.28	0.28	0.32
aMMSB	0.72	0.55	0.18	0.23	0.21	0.32

To demonstrate the structures of the three synthetic networks, their adjacency matrices are shown in Figure 3. Each row contains four differently organized adjacency matrices of the same synthetic network. The figures in the first column are adjacency matrices where nodes are organized only according to their major community memberships[3], hence the mass of the matrix is concentrated on its diagonal. However, as we organize the nodes in the adjacency matrices according to their roles, the adjacency matrices become differently (as shown in the second column). The core/periphery structures in the first and second networks, as well as bridge role in the second network, are shown clearly, so is the hierarchical structure in the third network. In the third column, we organize nodes in the adjacency matrix first according to their communities, and then for those in the same community, we organize them according to their roles; in the last column, the sorting is "reversed", nodes are organized first according to their roles and then according to their communities. We are surprising to see that the four different organization of adjacency matrices yield quite different perspective of the networks, thus if we do consider the community and role simultaneously, it is likely we will have a wrong understanding about the network structure.

To assess the recovery of community and role structure, we compare our model with two closely related generative models, MMSB [4] and aMMSB [5], which can be seen as special cases of our model. For simplicity, the number of communities and roles are set to their true numbers. Following the convention [5], we utilize the NMI (Normalized Mutual Information) as evaluation metric (definition of NMI will be introduced in the next sub-section). The experiments are repeated for five times, the mean NMI are shown in Table 1 . From the results we find that our model can well recover the communities and roles simultaneously, while MMSB and aMMSB have difficult time recovering them. We suspect the reason that MMSB and aMMSB cannot perform well is that, as shown in the Figure 1, the same adjacency matrix can be comparatively well organized both by community and role, which create some confusion for the models that do not explicitly model both community and role. Since in our model, we attribute

[3]Since nodes' membership vector is a distribution, we assign each node to its highest probability community and role.



(a) Community detection (b) Role detection

Figure 4: Performance of community and role detection for Enterprise network.

the generative process of links to both community and role memberships of nodes, the MMCR model can simultaneously detect both community and role structures more accurately.

5.5 Detection Results Analysis in Enterprise Network. Figure 4 shows both the NMI and accuracy results on both community and role detection tasks in Enterprise network. When no label on nodes is provided (0% in x-axis), we do not know the match between the detected results and the ground truth, thus we evaluate algorithms via NMI; when some labels are provided (>0% in x-axis), we use accuracy to evaluate the agreement of predicted results and ground truth. Firstly, we notice that when no seed/label is provided, our model performs comparatively to the baseline models, which might be due to the skewed role positioning and imbalance of communities in Enterprise network. Secondly, with only a few seeds/labels provided, our model can easily improve its performance, and it outperforms the baselines by more than 10% in both community and role detection when only about 10% of seeds/labels is provided (for all models); this further suggests that our model can better capture the underlying network structure than both MMSB and aMMSB. Thirdly, we observe that by integrating community and role detection together as in our model, both tasks can be enhanced.

Figure 5 shows both MMCR discovered and ground-truth communities and roles in the Enterprise network[4]. In the ground truth (Figure 5(b) and Figure 5(d)), it can be seen that the communities and roles scatter across the network, their structures are unclear, mixed and hard to detect. Nonetheless, our model can still successfully recover three different roles, and reasonably discover three communities in the Enterprise network.

5.6 Detection Case Study in Enron Network. Since there are no ground-truth communities and roles in the network, in order to evaluate the effectiveness of communities and roles found by MMCR, we try to interpret the discovered communities and roles by comparing them with the titles and group affiliations of Enron employees. For example, there are more than ten different titles of Enron employees in the data, such as VPs, Counsel, Attorney, etc.. From the titles we can see some indicators of hierarchy, such as senior managers (including VPs and Dir. managers, etc.), middle managers (including directors, managers, etc.), and other regular employees. There are more than ten different group affiliations within Enron data, many of them can be grouped by their functionalities (such as Legal, Financial, etc.), and regions (Midwest Region Trading, West Region Trading, etc.).

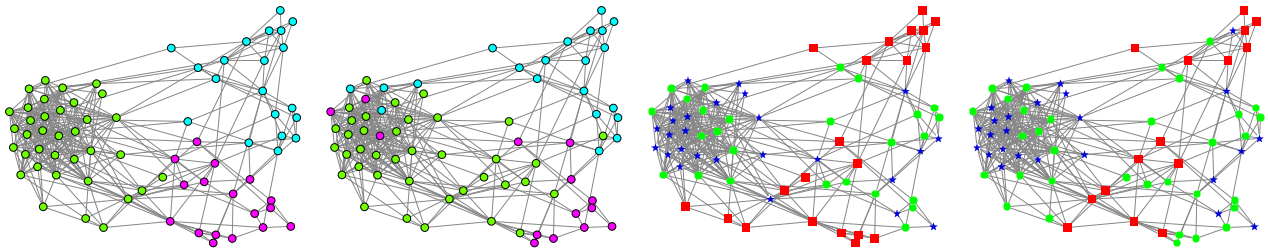
The MMCR discovered communities and roles are shown in Figure 6. Figure 6(a) shows the discovered communities, since there are many different teams in Enron data, we summarize each community with the most frequently-appeared one or two team names. They are: (1) Legal, (2) Financial, (3) NE / SE Region Trading, (4) Northeast and ERCOT Region Trading, (5) Midwest Region Trading, (6) West Power Real Time Scheduling and Trading, (7) West Region Trading and Origination. The found communities generally fit the nature of functionality and region. Figure 6(b) shows the detected roles. We find three types of roles, and they are dependent on the simultaneously detected communities: nodes belong to the first role (colored blue) are the bridge nodes, or structural hole spanners that connect to many different communities, they are popular and mainly distributed in the center of the graph, also we find most of them (more than 80%) are senior managers in the company. Nodes in the second role (colored red) are local hubs, they are mainly distributed in the community centers, although not as popular as the first role in general, they have dense connections inside their own communities; also we find many of them (more than 50%)[5] are senior and middle mangers. The last role (colored green) is periphery, they are distributed at the marginal areas of the network, and they are less popular than the other two roles; and we find many of them (more than 50%)[6] are regular employees. From the correspondence between obtained communities/roles and affiliations/titles in the data, we can see that the proposed MMCR can find meaningful structures in the Enron network.

5.7 Link Prediction Results Analysis. As mentioned before, one of the ways to access whether or not

[4]The discovered results are obtained with 12% seeds/labels, and memberships are assigned with the highest likelihood.

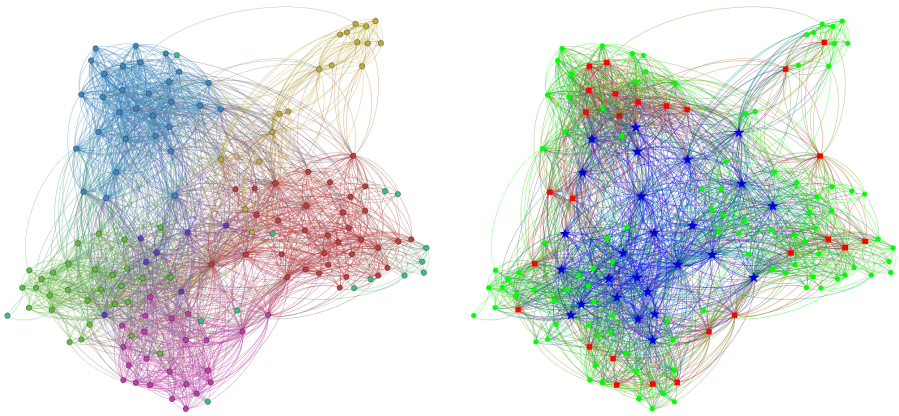
[5]There are active regular employees act like local hubs, such as Attorney.

[6]There are also some middle managers act like regular employees, we suspect those are less active or much specialized, so their email communications may seem like regular employees.



(a) Discovered communities (b) True communities (c) Discovered roles (d) True roles

Figure 5: Discovered communities and roles in the Enterprise network by MMCR. In (a) and (b), three communities of group private servers and PCs are shown, nodes in the same community are colored the same (for public servers in (b), we color it using its most closed group). In (c) and (d), we have three types of roles: stars are public servers, squares are group private servers, and circles are PCs. Figures are best viewed in color.



(a) Discovered communities (b) Discovered roles

Figure 6: Discovered communities roles in Enron network by MMCR. In (a), seven communities are found. In (b), three roles are found: blue star nodes are bridge nodes, red square nodes are local hub nodes, circle green nodes are periphery nodes. Figures are best viewed in color.

a generative model is reasonable is to evaluate the likelihood of held-out data. If a generative model can better predict previously unseen data than other models do, this indicates that the former generative model may be closer to the real model that generates the data. As in our network generative model, the held-out link likelihood estimation or link prediction task provides another angle of evaluation for generative models.

Table 2 shows the link prediction mean results of three methods over both Enterprise and Enron networks (all three models' the standard deviations of both perplexity and AUC are around 0.01, and of AP are around 0.03). We find our method outperforms the other two methods in all three different metrics. It is worth noting that the link prediction task is highly imbalanced (most candidate node pairs are non-links), thus the results of perplexity and AUC may be dominated by the large quantity of negative links and become less discriminative compared to AP [14], which suits better for imbalanced data. The superior performance of our model in link prediction task suggests that by combining both community and role, the model can better capture the network structure, thus making better prediction on non-observed links and non-links.

6 Related Work

Our work is related to both community detection and role discovery. Many of traditional community detection algorithms are based on modularity optimization [2]. And recently there are some community detection algorithms proposed based on statistical inference, which are shown more flexible and accurate [4, 3, 5]. A survey of community detection algorithms can be found in [1]. For our task of integrating community and role detection, it is difficult to directly adopt these community detection algorithms, due to the following reasons: (1) the interaction between roles might not be assortative (which is assumed for community [5]), some roles

Table 2: Link prediction results. For perplexity, the smaller the better. For AUC and AP, the larger the better.

	Enterprise	Enron
	Perplexity	
MMCR	**1.345**	**1.284**
aMMSB	1.357	1.313
MMSB	1.358	1.306

	Enterprise		Enron	
	AUC	AP	AUC	AP
MMCR	**0.902**	**0.743**	**0.919**	**0.701**
aMMSB	0.834	0.378	0.904	0.654
MMSB	0.833	0.375	0.908	0.670

such as bridges that tend to connect across communities. (2) some role structures (such as core/periphery) are embedded in community structures, so (flat) community structures alone cannot simultaneously capture both communities and roles.

The research of role in social and network theory can be found in [9, 10], where role structures such as structural hole, core/periphery are studied. In data mining fields, there is some work that automatically discovers roles in social and information networks [6, 15, 8, 7]. However, they usually ignore the community structures, and in most of role discovery work [6, 15, 16, 17], the role is defined on structural feature space, but we consider role that are tightly coupled with community, and directly connect them with pairwise interaction between nodes. We also notice the work [18] on detecting community and role with a generative model, in which links are generated purely according to role assignments of nodes. Different from theirs, we assign a role distribution to each node, and regard links as generated according to both community and role assignments. Also, most of these methods are difficult to incorporate explicit guidance on communities and roles of nodes provided by users.

7 Conclusions

In this paper, we study a novel problem of integrating community and role detection in a coherent framework. We propose a generative model that can simultaneously model both communities and roles. Empirical studies on three synthetic networks and two real information networks show that the proposed MMCR model can effectively discover the communities and roles in the networks and outperform other baselines, including two state-of-the-art network generative models MMSB and aMMSB. And superior performance of our model on link prediction task also shows our algorithm can better model the underlying structures of information networks.

Acknowledgment

This work is partially supported by NSF CAREER #1453800, Northeastern TIER 1, and Yahoo! ACE Award.

References

[1] Santo Fortunato. Community detection in graphs. *Physics Reports*, 2010.

[2] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004.

[3] Brian Ball, Brian Karrer, and Mark EJ Newman. An efficient and principled method for detecting communities in networks. *Physical Review E*, 2011.

[4] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. In *NIPS*, 2009.

[5] Prem Gopalan, Sean Gerrish, Michael Freedman, David M Blei, and David M Mimno. Scalable inference of overlapping communities. In *NIPS*, 2012.

[6] Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, and Lei Li. RolX: structural role extraction & mining in large graphs. In *SIGKDD*, 2012.

[7] M Puck Rombach, Mason A Porter, James H Fowler, and Peter J Mucha. Core-periphery structure in networks. *SIAM Journal on Applied Mathematics*, 2014.

[8] Tiancheng Lou and Jie Tang. Mining structural hole spanners through information diffusion in social networks. In *WWW*, 2013.

[9] Ronald S Burt. Structural holes: The social structure of competition. *Explorations in Economic Sociology*, 1993.

[10] Stephen P Borgatti and Martin G Everett. Models of core/periphery structures. *Social Networks*, 2000.

[11] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[12] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *PNAS*, 101, 2004.

[13] Bryan Klimt and Yiming Yang. Introducing the enron corpus. In *CEAS*, 2004.

[14] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML*, 2006.

[15] Sean Gilpin, Tina Eliassi-Rad, and Ian Davidson. Guided learning for role discovery (glrd): framework, algorithms, and applications. In *SIGKDD*, 2013.

[16] Yiye Ruan and Srinivasan Parthasarathy. Simultaneous detection of communities and roles from large networks. In *COSN*, 2014.

[17] Yu Han and Jie Tang. Probabilistic community and role model for social networks. In *SIGKDD, 2015*.

[18] Gianni Costa and Riccardo Ortale. A bayesian hierarchical approach for exploratory analysis of communities and roles in social networks. In *ASONAM, 2012*.