Vol.2009-BIO-17 No.9
2009/5/25

大規模ネットワーク構造の確率的グループモデルに基づくリンク予測

蜷川　陽[†1]　江口浩二[†1]

近年，複雑ネットワークのモデリングは生物学や社会学などの分野において重要な課題となっている．このような課題に対してこれまで多くの研究が行われてきたが，その多くは対象となるネットワークに関する明示的な事前知識を要求するものであった．一方，最近，明示的な事前知識を要求しない混合多項分布を用いた手法が提案され，社会ネットワークなどにおける頂点グループの検出に有効であることが示されている．本稿ではグループ検出とは異なる課題として，複雑ネットワークにおけるリンク予測に焦点を当てる．この目的のもと，混合多項分布に事前分布を仮定したベイズ混合多項分布を用いて，これをギブスサンプリング法によって推定する．代謝ネットワークと共著ネットワークのそれぞれから抽出した 50 通りのデータセットで実験を行い，提案手法によるリンク予測性能が従来手法と比較して有意に改善することを示す．

Link Prediction using Probabilistic Group Models of Network Structure

Akira Ninagawa[†1] and Koji Eguchi[†1]

Modeling of complex networks is a crucial task such as in biology and social sciences. A large number of researches have been conducted for such a problem; however, most of them require explicit, specific prior knowledge on target networks. On the other hand, a few recent works on multinomial mixture models presented that those models do not require such explicit prior knowledge and turned out to be effective for the task of group detection of vertices such as in social networks. This paper focuses on another task, link prediction in such complex networks, using a Bayesian multinomial mixture model, which assumes unobservable prior distributions over multinomial mixtures based on network structure and are estimated using Bayesian inference via Gibbs sampling. We demonstrate that, using this method, link prediction performance was significantly improved compared to conventional methods through experiments using 50 data sets extracted from a metabolic network or a co-authorship network.

1. Introduction

Recently, network analysis has become an increasingly important tool to exploit structural properties of a complex system in a wide variety of fields. In the fields of biology and pharmacology, analysis of biological networks, such as metabolic networks and protein-protein interaction networks, has been actively investigated and considered as a promising approach for hypothesis generation.[8] Social network analysis has also attracted considerable attention of sociologists, computer scientists, and even the ordinary people.[14),18] Complex networks in other fields have been researched as well, such as networks of the Internet like the World Wide Web, and ecological chain networks. Network analysis is not a new research subject in those fields; however, finding and understanding common properties in such real complex networks is a trend in the last decade.[2),18] Very recently, Newman et al.[12] investigated a simple multinomial mixture model for exploratory analysis of networks. One of the advantages of their model is that prior knowledge on target networks is mostly not required, while it is usually required in other conventional methods of network analysis. The task considered in their study was group detection in several social networks and a dependency network of words.

Link mining has also been studied, on the other hand, in the research community of data mining where addressing specific tasks are more emphasized rather than finding general properties in networks. The various task of link mining includes such as group detection, link prediction, entity classification, entity ranking, and subgraph discovery.[5] This paper focuses on the task of link prediction, which is the problem of predicting the existence of an unobserved link between two entities, based on other observed links and sometimes based on attributes of the entities as well. Link prediction is one of the crucial tasks, especially for biological networks. For instance, it is known that there exist a number of missing links in an assembled pathway of metabolic networks, and to predict such links is a promising task. Two types of features can be used to address the task of link prediction: one is observed link structure of a targeted network and the other is object attribute corresponding to each vertex.[5] In the paper, we take the former approach that does not necessarily depend on target networks.

†1 神戸大学大学院工学研究科情報知能学専攻
Department of Computer Science and Systems Engineering, Kobe University

This paper is motivated by the question of how well the multinomial mixture modeling approach based on observed link structure works for a practical task, link prediction in real-world network data, not using prior knowledge as possible or not using object attributes. For this objective, this paper investigates a Bayesian multinomial mixture model, which assumes unobservable prior distributions over multinomial mixtures based on link structure and is estimated using Bayesian inference, such as via Gibbs sampling. It is an extension of Newman et al.'s multinomial mixture model[12] that was mentioned previously. Newman et al.'s model achieves group detection (a.k.a. network clustering or community discovery), which classifies each vertex in a network into underlying groups in an unsupervised manner. Differently from other conventional methods, this model achieves "soft clustering" of network vertices, such that the probability indicating membership of multiple groups is computed for each vertex, on the basis of observation of patterns or behaviors of connections between vertices. Introducing unobservable prior distributions to the multinomial mixtures allows robustly and accurately capturing the patterns of connections in the network, as sometimes done in topic modeling.[4] Using such discovered underlying groups, we address the task of link prediction in complex networks. We demonstrate, through experiments with a metabolic network and a co-authorship network, that our method is effective in terms of prediction performance.

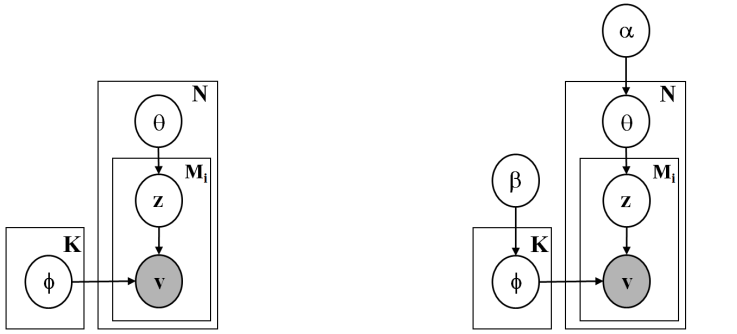
2. Related Work

A large number of researches have been conducted for modeling and analysis of complex networks, such as biological networks[8] and social networks[14]. Most of the existing methods required explicit, specific prior knowledge on targeted networks. However, very recently, Newman et al.[12] used a simple multinomial mixture model that does not require such explicit prior knowledge for the task of group detection of entities in social networks and a dependency network of words. Their model is based on the idea that each vertex's adjacent vertices are represented as a mixture of latent groups, where each latent group is represented as a multinomial distribution over vertices. They demonstrated that the model was effective to detect groups for both "assortative" networks in which vertices have most of their connections within the same group, and "disassortative" networks in which vertices have most of their connections outside their group, not requiring the prior knowledge on whether a target network is assortative or disassortative. The model used in that paper requires estimating every multinomial parameter from an observed adjacency matrix. Such kind of multinomial mixture models are known, in general, to have risks of overfitting and not modeling new entities.[4]

Zhang et al.[21),22] used a multinomial mixture model or a Gaussian mixture model with unobservable prior distributions for the task of group detection in coauthor networks. Using unobservable prior distributions is a good way to address the problems above. Their focus is rather on representing or profiling of observed entities, assuming explicit prior knowledge that a target network is assortative. That assumption is effective typically in coauthor networks; however, the motivation is different from that of Newman et al.[12] mentioned previously, in the sense of not assuming explicit prior knowledge as possible.

While the task considered in those papers above[12),21),22] was to detect groups of entities in networks, this paper is focused on the task of link prediction in unfamiliar, real-world network data. Moreover, this paper is motivated by the question of how well multinomial mixture modeling approach works based on observed network structure for the link prediction task. Link prediction is the task of predicting the existence of an unobserved link between two entities.[5] This task is sometimes viewed as a binary classification: for any two potentially linked entities, predict whether an indicator variable of this link is 1 or 0; other times the task is viewed as ranking according to similarity or affinity between the two entities. The latter is more general because it can also be interpreted as a binary classification when the ranking list is split into two parts, considering the upper and lower parts to be positive and negative, respectively. This paper evaluates link prediction from the view of similarity ranking.

This paper is also related to statistical topic models[4),7], which are based on the idea that each document is represented as a mixture of latent topics, where each latent topic is a probability distribution over words. Hofmann[7] proposed Probabilistic Latent Semantic Indexing (PLSI) that represents per-document multinomial topic distributions and per-topic multinomial word distributions in order to capture underlying topics in a set of documents. Blei et al.[4] extendeded it and developed Latent Dirichlet Allocation (LDA), by introducing Dirichlet priors on the multinomial distributions. Those established techniques can be applied to our research, since PLSI corresponds to Newman et al.'s model[12] that represents per-vertex group mixtures and per-group multinomial vertex distribution to capture underlying group in a target network. LDA corresponds to the Zhang's network model[22] and the model we use in this paper. However, applying those models to link prediction in real-world networks has not been investigated, to our knowledge.



(a) Newman's multinomial mixture model (b) Bayesian multinomial mixture model

Fig. 1 Graphical model representations.

3. Methodology

3.1 A Generative Model

Before presenting our methodology, we introduce some technical terms and notations. We start with a network G that consists of a set of vertices or entities $\mathbf{v} = \{v_i\}$ $(i = 1, .., N)$ and a set of edges or links $\mathbf{E} = \{\mathbf{e}_i\}$ $(i = 1, .., N)$, in which $\mathbf{e}_i = \{e_{ij}\}$ $(j = 1, .., M_i)$ indicates a set of all edges from vertex v_i to others. $\mathbf{E}$ is essentially equivalent to the adjacency matrix of the network. We assume that network G is comprised of a set of underlying groups $\mathbf{g} = \{g_k\}$ $(k = 1, .., K)$, each of which group is defined as a distribution over vertices. Let z_{ij} to be the group assigned to vertex v_i's adjacent vertex v_j. Therefore, $z_{ij} = g_k$ represents that group g_k is assigned to vertex v_j adjacent from vertex v_i. Moreover, $\mathbf{Z} = \{\mathbf{z}_i\}$ $(i = 1, .., N)$ can be defined where $\mathbf{z}_i = \{z_{ij}\}$ $(j = 1, .., M_i)$. We then consider a probabilistic mixture model, where each vertex is represented as a mixture of the groups. $P(\mathbf{z}_i|\theta_i)$ indicates per-vertex mixture distribution over groups; in other words, the probability of sampling a group that an arbitrary vertex adjacent from vertex v_i belongs to. Moreover, $P(\mathbf{E}|\mathbf{Z}, \phi_k)$ indicates per-group multinomial distribution over edges; in other words, the probability of sampling an edge having a vertex that belongs to group g_k. Parameters θ_i and ϕ_k are sampled from Dirichlet distributions specified by given hyperparameters α and β, respectively. We denote $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ as the entire sets $\{\theta_i\}$ $(i = 1, .., N)$ and $\{\phi_k\}$ $(k = 1, .., K)$, respectively. The probabilistic mixture model above is a simple hierarchical Bayesian model[3] in the sense that parameters θ_i and ϕ_k are sampled from the respective conjugate prior distributions. This model is referred to as Bayesian multinomial mixture model, in this paper. The graphical model representation of the Bayesian multinomial mixture model is shown in Fig. 1(b). In the graphical model representation, dependencies between variables or parameters are represented, where shaded circles indicate observed variables while white circles latent variables or unknown parameters. Each plate represents repeated i.i.d. sampling and the number at a corner of the plate indicates the number of times of the sampling. N indicates the number of vertices in a target network, K the number of groups, and M_i the number of vertices adjacent from vertex v_i, that is, the degree of vertex v_i. In contrast, the graphical model representation of Newman's multinomial mixture model[12] is shown in Fig. 1(a), where no prior distributions are introduced and thus robust, accurate estimation of model parameters is hard to achieved.[4]

The Bayesian multinomial mixture model above is a "generative" model of network, and the process of generating a network is formalized as follows:

(1) For all v_i vertices sample $\theta_i \sim \mathit{Dirichlet}(\alpha)$
(2) For all g_k groups sample $\phi_k \sim \mathit{Dirichlet}(\beta)$
(3) For each of the M_i vertices v_j adjacent from vertex v_i:
  (a) Sample a group $z_{ij} \sim \mathit{Multinomial}(\theta_i)$
  (b) Sample a vertex $v_j \sim \mathit{Multinomial}(\phi_{z_{ij}})$

where (v_i, v_j) corresponds to an edge e_{ij}. Given hyperparameters α and β, the full joint distribution over all variables and parameters is as follows:

$$p(\mathbf{E}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\alpha, \beta) = p(\boldsymbol{\phi}|\beta) \prod_{i=1}^{N} p(\theta_i|\alpha) P(\mathbf{z}_i|\theta_i) P(\mathbf{e}_i|\mathbf{z}_i, \boldsymbol{\phi}) \tag{1}$$

This can be transformed into the following equation:

$$p(\mathbf{E}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\alpha, \beta) = \prod_{i=1}^{N} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_{ik}^{\alpha-1+n_{i\cdot k}} \times \prod_{k=1}^{K} \frac{\Gamma(N\beta)}{\Gamma(\beta)^N} \prod_{j=1}^{N} \phi_{kj}^{\beta-1+n_{\cdot jk}} \tag{2}$$

where n_{ijk} indicates the count that group g_k is assigned to vertex v_i's adjacent vertex v_j, and '$\cdot$' means a corresponding index is marginalized. In other words, $n_{\cdot jk} = \sum_i n_{ijk}$ and $n_{i\cdot k} = \sum_j n_{ijk}$. N and K indicate the number of vertices and the number of underlying groups in a target network, respectively.

3.2 Estimation

Given the observed edges $\mathbf{E} = \{e_{ij}\}$, the task of Bayesian inference is to compute the posterior distribution over the latent group assignment variable $\mathbf{z} = \{z_{ij}\}$, the per-vertex distribution over groups $\boldsymbol{\theta} = \{\theta_i\}$ and per-group distribution over edges $\boldsymbol{\phi} = \{\phi_k\}$. We use Gibbs sampling for the task of Bayesian inference. Gibbs sampling inference uses the marginalized distribution over $\mathbf{E}$ and $\mathbf{Z}$, as follows[6]:

$$P(\mathbf{E}, \mathbf{Z}|\alpha, \beta) = \prod_{i=1}^{N} \frac{\Gamma(K\alpha)}{\Gamma(K\alpha + n_{i\cdot\cdot})} \prod_{k=1}^{K} \frac{\Gamma(\alpha + n_{i\cdot k})}{\Gamma(\alpha)} \times \prod_{k=1}^{K} \frac{\Gamma(N\beta)}{\Gamma(N\beta + n_{\cdot\cdot k})} \prod_{j=1}^{N} \frac{\Gamma(\beta + n_{\cdot jk})}{\Gamma(\beta)} \tag{3}$$

Given the current state of all except one group assignment to an edge e_{ij}, the conditional probability of z_{ij} is given by:

$$P(z_{ij} = k|\mathbf{Z}^{\neg ij}, \mathbf{E}, \alpha, \beta) = \frac{(\alpha + n_{i\cdot k}^{\neg ij})(\beta + n_{\cdot jk}^{\neg ij})(N\beta + n_{\cdot\cdot k}^{\neg ij})^{-1}}{\sum_{k'=1}^{K}(\alpha + n_{i\cdot k'}^{\neg ij})(\beta + n_{\cdot jk'}^{\neg ij})(N\beta + n_{\cdot\cdot k'}^{\neg ij})^{-1}} \tag{4}$$

where $n^{\neg ij}$ corresponds to variables or counts excluding e_{ij} and z_{ij}. The conditional probability specified by Equation (4) can be used to carry out the Gibbs sampling inference.

3.3 Link Prediction

We first estimate the unknown parameters of the Bayesian multinomial mixture model using observed links in a target network; and then rank vertex pairs according to the (log-)likelihood of generating each vertex pair from the estimated model. We refer to the set of vertex pairs to be ranked as "test set". The test-set log-likelihood is defined as follows:

$$\log P(\mathbf{E}_{test}|\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{e_{ij} \in \mathbf{E}_{test}, i<j} \log(P(e_{ij}|\theta_i, \boldsymbol{\phi})P(e_{ji}|\theta_j, \boldsymbol{\phi})) \tag{5}$$

where $\mathbf{E}_{test} = \{e_{ij}\}$ is the entire set of edges in test set. The probability $P(e_{ij}|\theta_i, \boldsymbol{\phi})$ can be obtained by the distribution $P(\mathbf{e}_i|\theta_i, \boldsymbol{\phi})$, as follows, where $\mathbf{e}_i$ is a set of all edges from vertex v_i to others in the test set.

$$P(\mathbf{e}_i|\theta_i, \boldsymbol{\phi}) = \prod_{j=1}^{M_j} \sum_{k=1}^{K} P(e_{ij}|z_{ij} = g_k, \phi_k)P(z_{ij} = g_k|\theta_i) \tag{6}$$

$$= \prod_{h=1}^{N} \left\{ \sum_{k=1}^{K} P(e_{ih}|z_{ih} = g_k, \phi_k)P(z_{ih} = g_k|\theta_i) \right\}^{n_{ih\cdot}} \tag{7}$$

where $n_{ih\cdot} = \sum_k n_{ihk}$ and n_{ihk} indicates the count that group g_k is assigned to vertex v_i's adjacent vertex v_h. $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are estimated via Gibbs sampling inference. θ_{ik} and ϕ_{kj} are obtained by the following equations, according to Griffiths et al.[6]:

$$\theta_{ik} = \frac{n_{i\cdot k}^{\neg ij} + \alpha}{\sum_{k'=1}^{K} n_{i\cdot k'}^{\neg ij} + K\alpha} \tag{8}$$

$$\phi_{kj} = \frac{n_{\cdot jk}^{\neg ij} + \beta}{\sum_{j'=1}^{N} n_{\cdot j'k}^{\neg ij} + N\beta} \tag{9}$$

4. Experiments

In this section, we evaluate through experiments the Bayesian multinomial mixture model described in Section 3 on the task of link prediction in real-world network data, and compare it with several existing methods based only on the network structure. We used network data of a metabolic network or a co-authorship network for the experiments.

4.1 Existing Methods

First of all, we explain five existing methods from earlier works to compare with the proposed method. Those methods are well accepted and well investigated.[9),11)] Each measure described below indicates similarity or affinity between a pair of vertices, i.e., how similar a pair of entities is, according to link structure of a target network. Ranking of vertex pairs is determined according to the similarity. By evaluating the ranking, the performance of link prediction can be measured.

Note that all the measures are defined only using observed links. Hereafter, we denote $\mathbf{a}_i$ as a set of vertices adjacent from vertex v_i.

(1) Common Neighbors[13]:

$$Common = |\mathbf{a}_i \cap \mathbf{a}_j| \tag{10}$$

Common Neighbors is a measure based on the idea that a pair of vertices are likely to be adjacent when these vertices share a number of common adjacent vertices.

(2) Jaccard[17]:

$$Jaccard = \frac{|\mathbf{a}_i \cap \mathbf{a}_j|}{|\mathbf{a}_i \cup \mathbf{a}_j|} \tag{11}$$

Jaccard's coefficient is a standard measure of similarity in the field of information retrieval. It is based on the idea that a pair of vertices each of which has smaller degree is more im-

portant than others. The value of *Jaccard* increases when each of a pair of vertices has a few adjacent vertices and those adjacent vertices are common.

(3) Adamic-Adar[1]:

$$Adamic\text{-}Adar = \sum_{k \in \mathbf{a}_i \cap \mathbf{a}_j} \frac{1}{\log |\mathbf{a}_k|} \tag{12}$$

The Adamic-Adar measure assigns different weight to each common adjacent vertex. A larger weight is assigned to a vertex of smaller degree.

(4) Preferential Attachment[13]:

$$Preferential = |\mathbf{a}_i| \cdot |\mathbf{a}_j| \tag{13}$$

Preferential Attachment is different from the other above measures slightly. This measure is based on a model for generating scale-free networks, in which a vertex with a larger degree tends to connect to other vertices.

(5) Katz[10]:

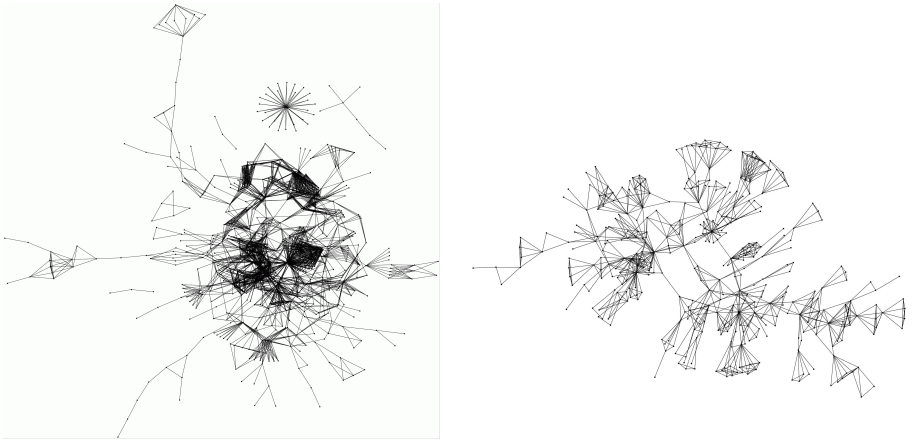
$$Katz_\mu = \sum_{\ell=1}^{\infty} \mu^\ell |paths_{ij}^{(\ell)}| \tag{14}$$

$Katz_\mu$ is defined as a measure on the basis of all the paths between a pair of vertices. The value of $Katz_\mu$ is determined according to both the number of paths between a pair of vertices and the length of each path. The notation $paths_{ij}^{(\ell)}$ in Equation (14) indicates the number of paths from vertex v_i to vertex v_j of which length is ℓ. Therefore, shorter length paths are more emphasized. For a large number of ℓ, the corresponding set of paths exponentially grows. Therefore, we imposed the constraint that the paths of which length satisfies $\ell \leq 3$ were only used, in the computation with Equation (14). We fixed the weight parameter $\mu = 0.05$, according to earlier works.[9),11)]

4.2 Experimental Settings

4.2.1 The Network Data

The network used in our experiments is a metabolic network and a co-authorship network. Metabolic networks, in general, represent the process of converting the food that was taken from outside the body into energies and chemical compounds necessary for living. In such metabolism, various enzymes serve as catalysts in the chemical reaction. In the metabolic network, each vertex represents an enzyme observed to act as a catalyst, and each edge represents that two enzymes were observed to act consecutively as catalysts. The data used in our experiments is the metabolic pathway of "S.Cerevisiae" that were constructed by Yamanishi et al.[19] by extracting from KEGG/PATHWAY database[20]. The co-authorship network is the data of scientists working in the area of network science, and was used in 15). We only used the largest connected component of this network data. The overview of the network data is shown in Fig 2(a) and (b). The property of the data is shown in Table 1. The degree distributions of these two data are shonw in Fig. 3. As shown in these figures, scale-free property is observed in these networks.[2)]



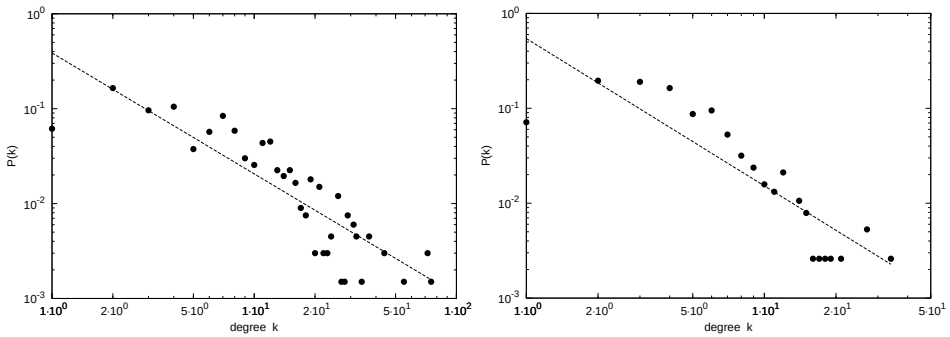
(a) the metabolic network (b) the co-authorship network

Fig. 2 Structure of networks used in experiments.

Table 1 The data of a metabolic network and a co-authorship network used in experiments.

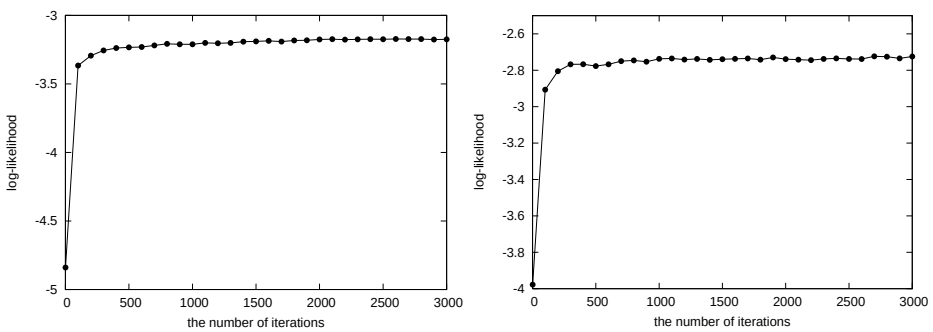
	the metaboloc network	the co-authorship network
The number of entities	668	379
The number of links	2782	914
Links/all entity pairs	0.0125	0.0126
Average shortest path length	5.711	6.042
Clustering coefficient[18)]	0.3367	0.7412
Average degree of entities	8.342	4.823

For each of these network data, we used 80% of all the vertex pairs as training data, 10% as development data and the remainder as test data. We estimated the unknown parameters of the mixture model using the training data, varying hyperparameters α and β; and determined opti-



(a) the metabolic network (b) the co-authorship network

Fig. 3 The degree distribution of entities.



(a) the metabolic network (b) the co-authorship network

Fig. 4 Test-set log-likelihood according to the number of iterations.

mal values of the hyperparameters so that log-likelihood of the development data are maximized. After determining the hyperparameters, we merged the development data and the training data; and using them, we estimated the unknown parameters of the mixture model, again. Therefore, 90% of the whole network was used for estimating the model, finally. When we split the training data, development data and test data, we removed the vertices only appearing in the development data or the test data but not appearing in the training data, since those isolated vertices are not able to be predicted using the model estimated with the training data. We conducted experiments on the task of link prediction using 50 sets of training data, development data and test data that were randomly sampled from the entire set of vertices to ensure the fixed proportion mentioned previously. Using each of the data sets, we compared the proposed method with the five existing methods.

4.2.2 Parameter Estimation

It is necessary in our experiments to determine the following three parameters: hyperparameters α and β of Dirichlet prior distributions and the number of latent groups K for the Bayesian multinomial mixture model.

For the number of latent groups K, we used 10 values in the range of 10 to 100 with an interval of 10. For each K value, we determined the two hyperparameters α and β with each training data set so that development-set log-likelihood is maximized; and then obtained the average value of α, as well as β, over those determined with 50 sets of training/development data.

With the determined hyperparameters, we estimated the unknown parameters of the mixture model using both the development data and the training data; and obtained test-set log-likelihood using Eq. (5). The test-set log-likelihood (or the development-set log-likelihood) means the negative logarithm of perplexity with respect to the test data (or the development data). The perplexity is a well-accepted criterion to measure accuracy of statistical models, such as language models.[16)]

We also investigated how the test-set log-likelihood can be improved according to the number of iterations. As shown in Fig. 4, the log-likelihood rises sharply by around 300 iterations, and it gradually converges afterward. According to the result, it can be said that the log-likelihood at around 1000 iterations is reasonable. We therefore fixed the number of iterations to be 1000 in our experiments below.

4.3 Evaluation on Link Prediction Task

We used mean average precision (MAP) as an evaluation metric of the task of link prediction. MAP is well accepted for evaluation of information retrieval task, and it is known to be easily understandable and stable to evaluate ranking. MAP is defined as follows:

$$\frac{1}{|\mathbf{data}|}\sum_{d\in\mathbf{data}}\left\{\frac{1}{|\mathbf{true}_d|}\sum_{r\in\mathbf{rank}_d} prec(r)\right\} \tag{15}$$

where **data** denotes a set of test data ($|\mathbf{data}| = 50$), $\mathbf{true}_d$ indicates the entire set of "true" links in test data d (i.e., all appeared links in d), and $\mathbf{rank}_d$ indicates the entire set of links predicted by

a method using the training data corresponding to d. The notation $prec(r)$ indicates precision at rank r in the ranking of predicted links, where precision is defined as the proportion of predicted true links out of r top-ranked predicted links. Here, the link prediction ranking is achieved according to test-set log-likelihood in the case of our method, and according to a similarity measure in the case of the other existing methods.

4.4 Experimental Results

We carried out experiments with our method and the five existing methods using 50 sets of training data and test data and then calculated MAP. The result of each method is shown in Table 2. MAP and another variation MP@10 are shown in this table. We computed MAP values by imposing the constraint that the link prediction ranking is cut off at the rank of 1000. MP@10 indicates the mean of precision at the rank of 10.

According to Table 2, the link prediction performance of our method is more than 17 points higher than that of the other five methods, in terms of MAP, in the case of the metabolic network. Its percentage improvement (i.e., the ratio of degree of improvement to baseline performance) was 498% higher, compared with Katz measure as the baseline. The improvement obtained by our method was statistically significant over either Common Neighbors, Jaccard, Adamic-Adar, Preferential Attachment, or Katz, where $p < 0.01$ with the two-sided Wilcoxon signed-rank test. According to the other evaluation metrics, our methods remarkably outperformed the baselines, as well.

As for the case of the co-authorship network, the proposed method also works well; however, its link prediction performance was less than that of CommonNeighbors and Katz. As you can see in Table 1, the clustering coefficient[18] of the co-authorship network was much larger than that of the metabolic network. This means that the edge connectivity in the co-authorship network is dense. In such a case, methods based on local structure like CommonNeighbors or Katz seem to work quite well.

We demonstrate the Recall-Precision curves of our proposed method under the condition with the optimal topic numbers to compare with the five existing methods in Fig. 5(a) for the metabolic network and Fig. 5(b) for the co-authorship network.

Table 2 Evaluation results on link prediction task in the metabolic network.

	MAP (%)	MP@10 (%)
CommonNeighbors	2.884	9.273
PreferentialAttachment	0.1488	2.545
Adamic/Adar	0.03015	0.7273
Jaccard	0.002364	0.1818
Katz	3.587	9.636
proposed($K = 80$)	21.44	43.40
proposed($K = 90$)	21.25	35.60
proposed($K = 100$)	22.05	42.40

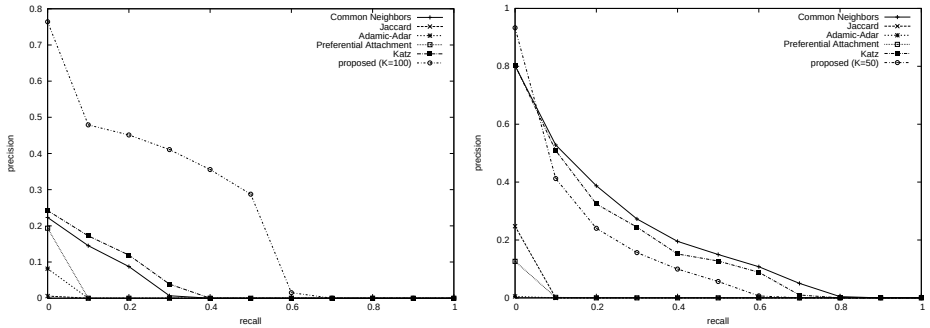
Table 3 Evaluation results on link prediction task in the co-authorship network.

	MAP (%)	MP@10 (%)
CommonNeighbors	30.41	92.73
PreferentialAttachment	0.2843	2.727
Adamic/Adar	0.08760	0.5455
Jaccard	0.5067	8.545
Katz	25.68	84.73
proposed($K = 40$)	12.77	47.82
proposed($K = 50$)	13.55	48.73
proposed($K = 60$)	12.88	49.09

5. Conclusions

In this paper, we proposed a method to predict unobserved links from observed link information, by modeling underlying groups in the network only on the basis of patterns or behaviors of connections in the network. The model is a simple hierarchical Bayesian model, which assumes unobservable prior distributions over Newman et al.'s multinomial mixture model[12] and is estimated using Bayesian inference via Gibbs sampling.

Conventional structure-based link prediction methods, such as Common Neighbors, Jaccard, Adamic-Adar, Preferential Attachment and Katz measures, are often based on local structure (for instance, based on counts of vertices commonly adjacent from a pair of vertices in the network). On the contrary, multinomial mixture models of network can capture patterns of vertex connectivity from observation in the entire network. We demonstrated, through our experiments using a metabolic network or a co-authorship network, that our method works well in the task of link prediction, compared with five conventional methods based on link structure. Especially for the metabolic network, the improvement was statistically significant than all the five conventional methods. As future works, we plan to perform experiments with larger networks.



(b) the metabolic network (b) the co-authorship network

Fig. 5 Recall-precision curves of the proposed method and five existing methods.

Acknowledgments This work was supported in part by the Grant-in-Aid for Scientific Research (B) (#20300038) and Scientific Research on Priority Areas “Info-plosion” (#19024055) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1) Lada Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.
2) Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct. 1999.
3) ChristopherM. Bishop. *Pattern Recognition and Machine Learning*. Springer-Velag, 2006.
4) DavidM. Blei, AndrewY. Ng, and MichaelI. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
5) Lise Getoor and Christopher P. Diehl. Link mining: A survey. *SIGKDD Explorations*, 7(2):3–12, May 2005.
6) ThomasL. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228–5235, 2004.
7) Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Anual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, USA, 1999.
8) BjörnH. Junker and Falk Schreiber. *Analysis of Biological Networks*. Wiley-Interscience, 2008.
9) Hisashi Kashima and Naoki Abe. A parameterized probabilistic model of network evolution for supervised link prediction. *icdm*, 0:340–349, 2006.
10) Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953.
11) David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the 12th ACM International Conference on Information and Knowledge Management*, pages 556–559, New Orleans, Louisiana, USA, Nov. 2003.
12) M.E.J.Newman and E.A.Leicht. Mixture models and exploratory analisis in networks. *proceedings of the National Academy of Sciences of the United States of America*, 104(23):9564–9569, 2007.
13) M.E.J. Newman. Clustering and preferential attachment in grouwing networks. *Physical Review Letters E*, 64, Apr 2001.
14) M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, May 2003.
15) M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104, 2006.
16) Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Pentice Hall, 1993.
17) Gerard Salton. *Introduction to Modern Information Retrieval (McGraw-Hill Computer Science Series)*. McGraw-Hill, September 1983.
18) DuncanJ. Watts and StevenH. Strogatz. Collective dynamics of ’small-world’ networks. *Nature*, (393):440–442, Jun. 1998.
19) Yoshihiro Yamanishi, Jean-Philippe Vert, and Minoru Kanehisa. Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, 21(1):468–477, 2005.
20) Y.Yamanishi, J.P. vert, and M.Kanehisa. Protein network inference from multiple genomic data: a supervised approach. In *BIOINFOMATICS*, volume20, pages i363–i370, 2004.
21) Haizheng Zhang, C.Lee Giles, HenryC. Foley, and John Yen. Probabilistic community discovery using hierarchical latent gaussian mixture model. In *Proccedings of the 22th Association for the Advancement of Artificial Intelligence Conference (AAAI 2007)*, pages 663–668, Vancouver, Canada, Jul. 2007.
22) Haizheng Zhang, Baojun Qiu, C.Lee Giles, HenryC. Foley, and John Yen. An LDA-based community structure discovery approach for large-scale social networks. In *ISI*, pages 200–207, 2007.