Link Prediction in Metabolic Networks using Topology-based Mixture Models

Akira Ninagawa
akiraninagawa@cs25.scitec.kobe-u.ac.jp

Koji Eguchi
eguchi@port.kobe-u.ac.jp

Department of Computer Science
and Systems Engineering,
Kobe University, Kobe, Japan

Background

- A large variety of information can be represented as networks, and available data for such networks have increased recently.
- For instance: social networks, ecological webs, communication networks, metabolic networks, and protein interaction networks.
- Our focus is on biological metabolic networks.
- A metabolic network represents the process of converting the food that was taken from outside the body into energies and chemical compounds necessary for living.

Background

➢*Complex Network analysis* or *link mining* has become crucial tools in a wide variety of fields.

➢In the area of the *complex network analysis*, macroscopic properties have been pursued, such as scale-free property.

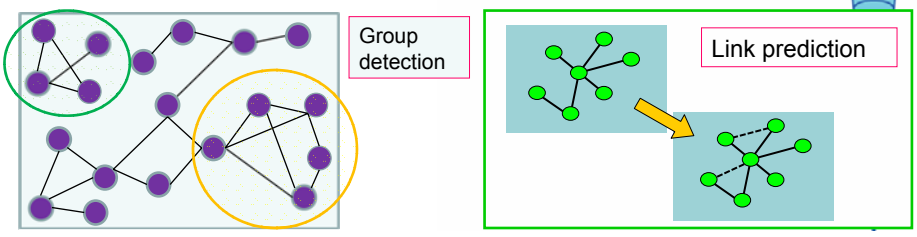
➢In the area of the *link mining*, some particular tasks have been addressed from microscopic points of view, such as group detection (*aka*, network clustering) and link prediction in a network.

Background

➢Group detection or network clustering is the task of classifying vertices in a network into underlying groups in an unsupervised manner.

➢Link prediction is the task of predicting the existence of an unobserved link between two vertices.

➢We focus on these two tasks, especially link prediction.



Background

Information used for link prediction

- Link prediction using only vertex attributes
  - It requires sufficient expertise on the target domain.
  - High precision may be achieved at the expense of less flexibility.
- Link prediction using only network topology information
  - It requires less or no expertise on the target domain.
  - It can be widely used, although precision might be lower.
  - There are not many researches in this category, but it is promising.

Variations of topology-based link prediction methods

- By measuring similarity / proximity between vertices
- By computing likelihood of edges using observed subgraph

Research objectives

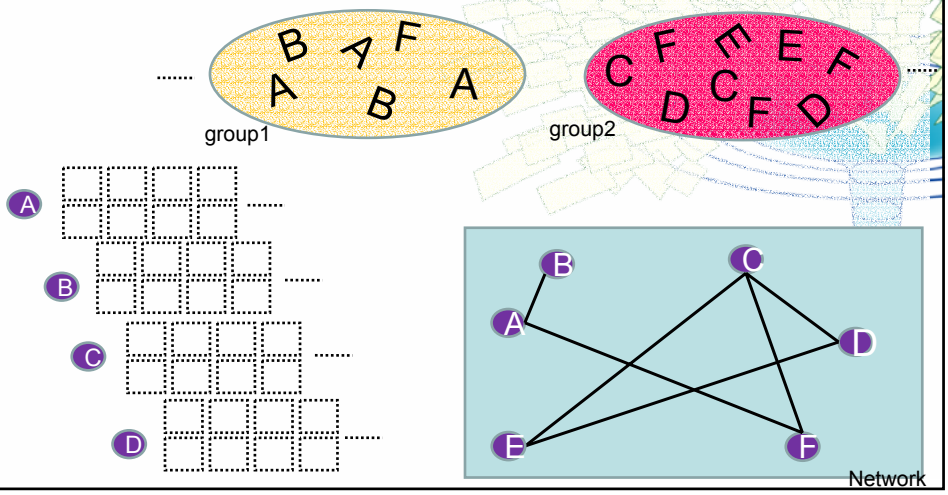
- We propose a link prediction method only based on network topology.
- We compute likelihood of unobserved edges using a hierarchical Bayesian mixture model (our proposed method), which is estimated from observed subgraph in the network.

To our knowledge, there are no researches on such a topology-based mixture model for link prediction, either in biological or non-biological domains.

- We demonstrate the improvement brought by our method in the link prediction task, compared with some existing methods.

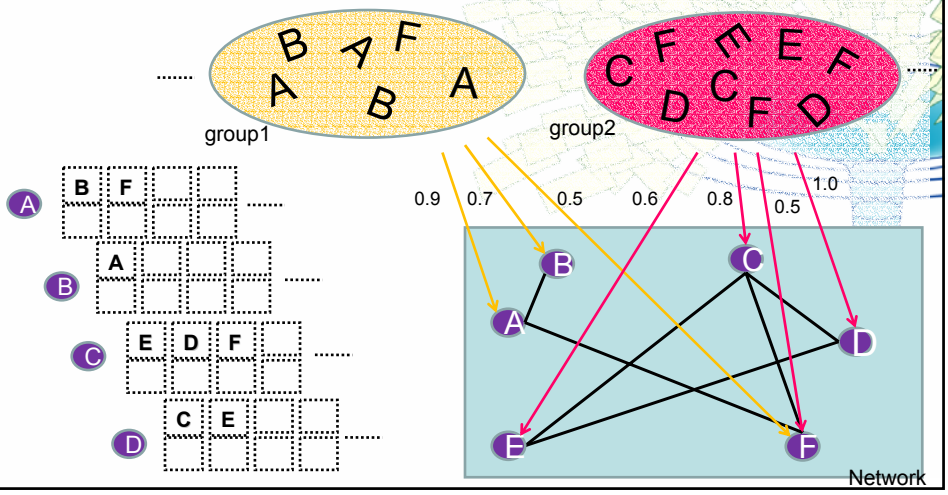
Proposed method

➢The adjacent vertices distribution of each vertex is assumed to be a mixture of underlying groups in the network; and each group is assumed to be a multinomial distribution of vertices.



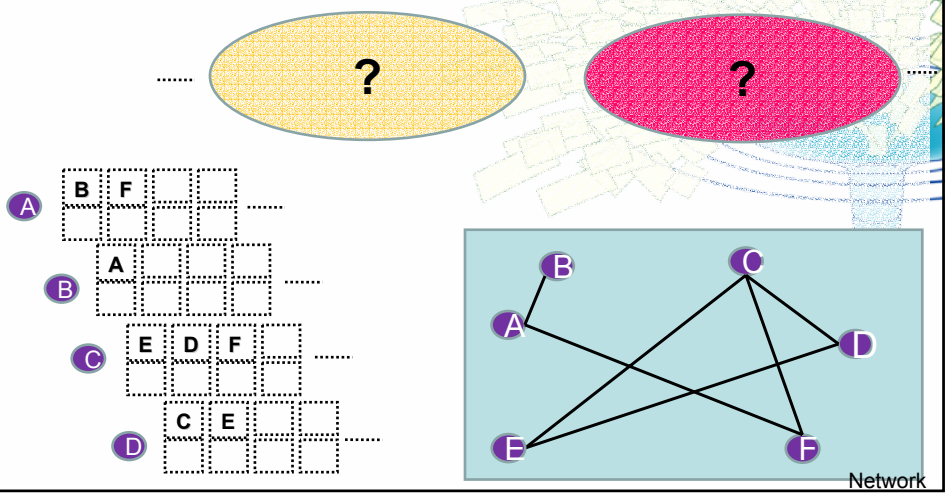
Proposed method

➢The adjacent vertices distribution of each vertex is assumed to be a mixture of underlying groups in the network; and each group is assumed to be a multinomial distribution of vertices.



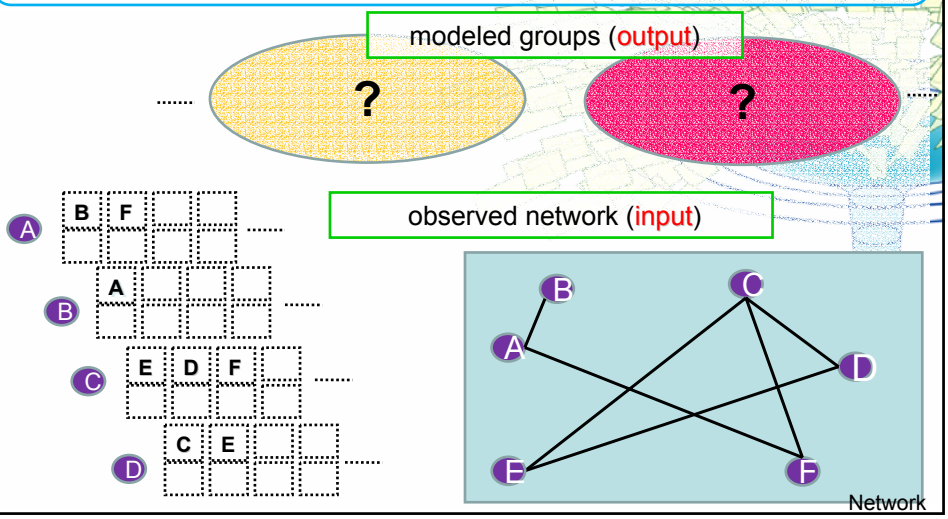
Proposed method

➢We further introduced conjugate (Dirichlet) distributions for the per-vertex group distribution and the per-group vertex distribution.
➢Underlying groups are estimated such as by Gibbs Sampling method.



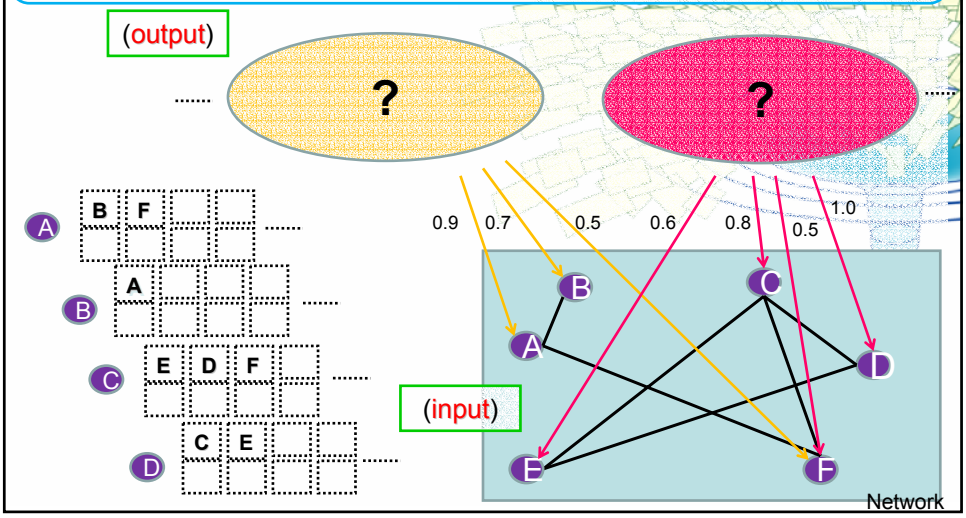
Proposed method

➢We further introduced conjugate (Dirichlet) distributions for the per-vertex group distribution and the per-group vertex distribution.
➢Underlying groups are estimated such as by Gibbs Sampling method.



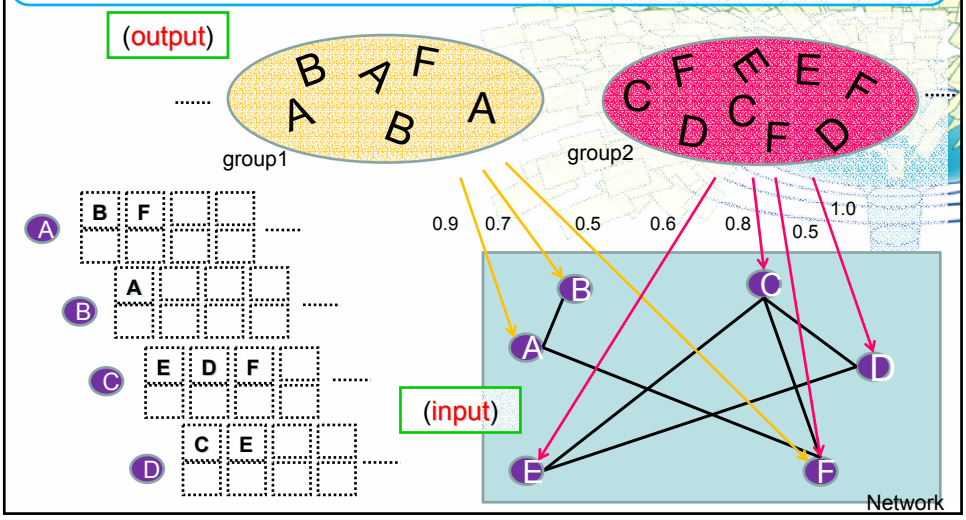
Proposed method

➢We further introduced conjugate (Dirichlet) distributions for the per-vertex group distribution and the per-group vertex distribution.
➢Underlying groups are estimated such as by Gibbs Sampling method.



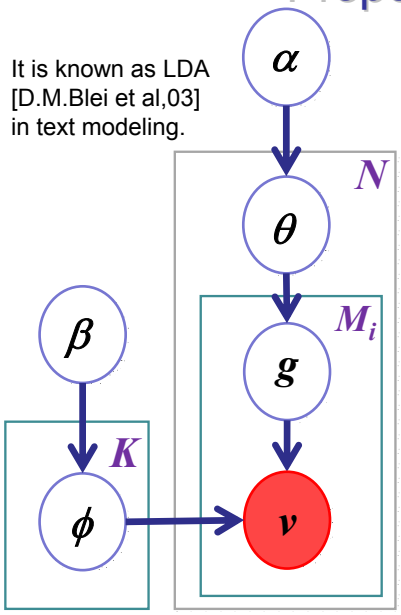
Proposed method

➢We further introduced conjugate (Dirichlet) distributions for the per-vertex group distribution and the per-group vertex distribution.
➢Underlying groups are estimated such as by Gibbs Sampling method.



Proposed method

It is known as LDA [D.M.Blei et al,03] in text modeling.



A hierarchical Bayesian model of a network

α : a hyperparameter of Dirichlet distributions for the per-vertex group distributions

β : a hyperparameter of Dirichlet distributions for the per-group vertex distributions

θ : per-vertex group (multinomial) distributions

ϕ : per-group vertex (multinomial) distributions

g : group

v : vertices in observed network

N : the number of vertices in a target network

M_i : the number of vertices adjacent from a vertex v_i , that is, the degree of vertex v_i

K : the number of groups

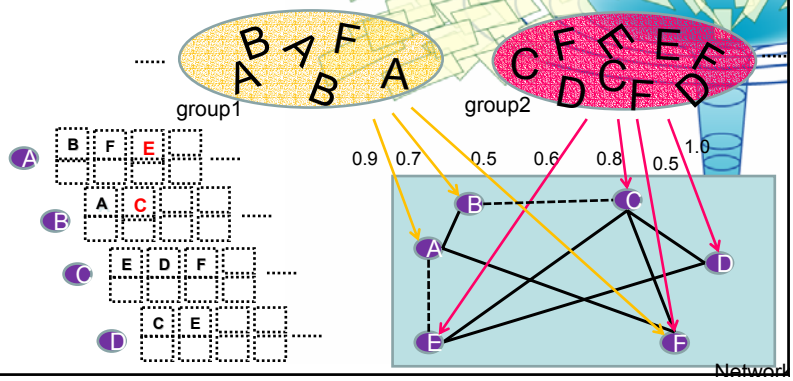
Unknown parameters can be estimated using Gibbs Sampling [T.L.Griffitds et al, 04].

Proposed method

➢Then, we predict unobserved links from the estimated models of groups.

① We compute the likelihood of generating a link between a pair of vertices, using the models.

② We rank all unobserved links in descending order of the likelihood, and then evaluate the ranking.



Experiments

We predict links in a metabolic network

experimental settings

- The data we used in our experiments is the metabolic network of "S.Cerevisiae" extracted from KEGG/PATHWAY database [Y.Yamanishi et al, 05].
- The number of vertices is 668, the number of links is 782, and the proportion of the links to all vertex pairs is 0.0125.
- We used 80% of all the vertex pairs as training data, 10% as development data and the remainder as test data.
- We conducted experiments on the task of link prediction using 50 sets of training data, development data and test data that are randomly sampled.
- We compared the proposed method with the three existing methods.
- We determined the following four parameters: hyperparameter α and β of Dirichlet prior distributions, the number of latent groups K, and the number of iterations for Gibbs Sampling, so that development-set log-likelihood is maximized.

Experiments

Three existing methods for comparison

•Jaccard's Coefficient:

It is based on the idea that a pair of vertices, each of which has smaller degree is more important than others.

$$Jaccard(X,Y) = \frac{|a_X| \bigcap |a_Y|}{|a_X| \bigcup |a_Y|}$$

(where a_X indicates X's adjacent vertices.)

•Adamic/Adar:

This measure assigns different weight to each common adjacent vertex.
A larger weight is assigned to a vertex of smaller degree.

$$Adamic/Adar(X,Y) = \sum_{k \in |a_X| \cap |a_Y|} \frac{1}{\log |a_k|}$$

•Katz:

- The Katz value is determined according to both the number of paths and the length of each path.
- γ is a weight parameter fixed somewhere from 0 to 1.
- $paths^{\langle l \rangle}_{X,Y}$ indicates the number of paths from vertex X to vertex Y of which length is l.

$$Katz(X,Y) = \sum_{l=1}^{\infty} \gamma^l \cdot \left| paths^{\langle l \rangle}_{X,Y} \right|$$

Experiments

We used Mean Average Precision to evaluate the result

Mean Average Precision (MAP)

$$\frac{1}{|data|}\sum_{d\in data}\left\{\frac{1}{|true_d|}\sum_{r\in rank_d} prec(r)\right\}$$

- ***data*** indicates test data set (in this case, $|data| = 50$).
- $true_d$ indicates the whole set of links that appear in the test data ***d***.
- $rank_d$ indicates the list of predicted links with ranks, corresponding to test data ***d***.
- $prec(r)$ indicates the precision of the r-th rank in the predicted link list.
- The precision is defined as the proportion of true links out of r top-ranked predicted links.

Experiments

Experimental results

	MAP (%)	MP@10 (%)
Adamic/Adar	0.03014	0.7273
Jaccard	0.00236	0.1818
Katz	3.587	9.636

	MAP (%)	MP@10 (%)
proposed (K = 80)	21.44	43.40
proposed (K = 90)	21.25	35.60
proposed (K = 100)	22.05	42.40

•MAP is what was defined in the previous slide.

•MP@10 is the precision of 10 top-ranked predicted links, averaged over 50 sets of test data.

•The link prediction performance of our method is more than 18 points higher than that of the other three methods, in terms of MAP.

•According to mean of precision at the 10th rank (MP@10), our methods remarkably outperformed the baselines, as well.

Future work

- To apply to larger-scale networks
- To consider multiple type of links or multiple types of vertices

  We are planning to investigate dependence between different types of vertices, which can be classified such as using EC number.

- To improve the inference algorithm

  We are planning to modify Gibbs Sampling algorithm to better suit our model.

Thank you for your attention

The process of generating a network is formalized as follows:

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

Given hyperparameters α and β, the full joint distribution over all variables and parameter is as follows:

$$P(E, Z, \theta, \phi \mid \alpha, \beta) = P(\phi \mid \beta) \prod_{i=1}^{N} P(\theta_i \mid \alpha) P(z_i \mid \theta_i) P(e_i \mid z_i, \phi)$$

Hyperparameter:

The hyperparameter is the parameter influencing the entire probability model and defining the prior probability.

Dirichlet distribution:

- The Dirichlet distribution is a distribution over multinomial parameters.
- Multinomial β distribution
- Multinomial conjugate prior distribution

Gibbs Sampling:

Assuming the conditional probability distribution fixing variables except one variable, we sample many times with it.

About our network data:

- Each vertex represents an enzyme, and each link represents that two enzymes are observed to act consecutively as catalysts.
- The reason using the biological network is that our methods may be useful to identify the unknown parts of the metabolic network.

Proposed method

Gibbs Sampling

➢Gibbs Sampling is used as the algorithm to solve approximately since exact estimations using a hierarchical bayesian model are generally difficult.

➢Gibbs Sampling is one of statistical tools called Markov chain Monte Carlo methods(MCMC).

➢MCMC is the method to simulate Markov chains which is invariant distribution.

➢In Gibbs Sampling, using the random numbers and conditional distribution, a sample sequence is generated by repeatedly updating variables obtained from the observed distribution. And the requested invariant distribution is obtained.

➢Generated sample sequences are used to calculate expectation and marginal probability.

A hierarchical Bayesian model

In the case using bayesian model not having hierarchies, a posteriori distribution of parameter when data are given can be obtained using Bayes' theorem by following:

$$\Pr(\lambda \mid D) = \frac{\Pr(D \mid \lambda)\Pr(\lambda)}{\Pr(D)}$$

So the parameter λ is generated according to $\Pr(\lambda)$,
and given the parameter, data D are generated according to $\Pr(D|\lambda)$.

A hierarchical Bayesian model is stratified this model.

The hyperparameter μ is generated according to $\Pr(\mu)$, and given the parameter, the parameter λ is generated according to $\Pr(\lambda \mid \mu)$.
Additionally, given the parameter, data D are generated according to $\Pr(D|\lambda)$.
So the posteriori distribution is obtained according to this model by following:

$$\Pr(\lambda, \mu \mid D) \propto \Pr(D \mid \lambda)\Pr(\lambda \mid \mu)\Pr(\mu)$$

- When data are divided into some hierarchies a good result may be obtained if the factor common to each hierarchy is modeled by the prior distribution of hyperparameter.