Nonparametric Latent Feature Models for Link Prediction

Kurt T. Miller
tadayuki@cs.berkeley.edu

Thomas L. Griffiths
tom_griffiths@berkeley.edu

Michael I. Jordan
jordan@cs.berkeley.edu

University of California, Berkeley

Introduction

As the availability and importance of relational data—such as the friendships summarized on a social networking website—increases, it becomes increasingly important to have good models for such data. The kinds of latent structure that have been considered for use in predicting links in such networks have been relatively limited. In particular, the machine learning community has focused on latent class models, adapting Bayesian nonparametric methods to jointly infer how many latent classes there are while learning which entities belong to each class. We pursue a similar approach with a richer kind of latent variable—latent features—using a Bayesian nonparametric approach to simultaneously infer the number of features at the same time we learn which entities have each feature. Our model combines these inferred features with known covariates in order to perform link prediction. We demonstrate that the greater expressiveness of this approach allows us to improve performance on three datasets.

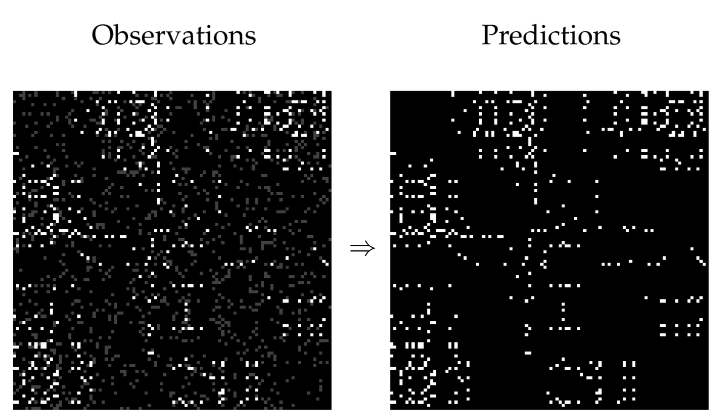
Link Prediction

Assume we observe a partial set of relationships or links between pairs of entities in a network. Link prediction is the task of predicting whether or not unobserved relationships hold.

In all figures on this poster, relations are stored in an $N \times N$ matrix, where each row and column corresponds to an entity. Each entry in the matrix is:

- *White* means the relationship holds from the the row entity to the column entity.
- *Black* means the relationship does not hold.
- *Gray* means the relationship is unobserved.

In link prediction, we wish to go from observations to predictions:



In generative models, a common approach is to assume that there is some underlying latent variable Z_i for each entity i and parameters W such that for $y_{ij} = Y(i,j) \in \{0,1\}$ representing whether or not there is a link from i to j,

$$\Pr(Y|Z,W) = \prod_{i,j} \Pr(y_{i,j}|Z_i, Z_j, W).$$

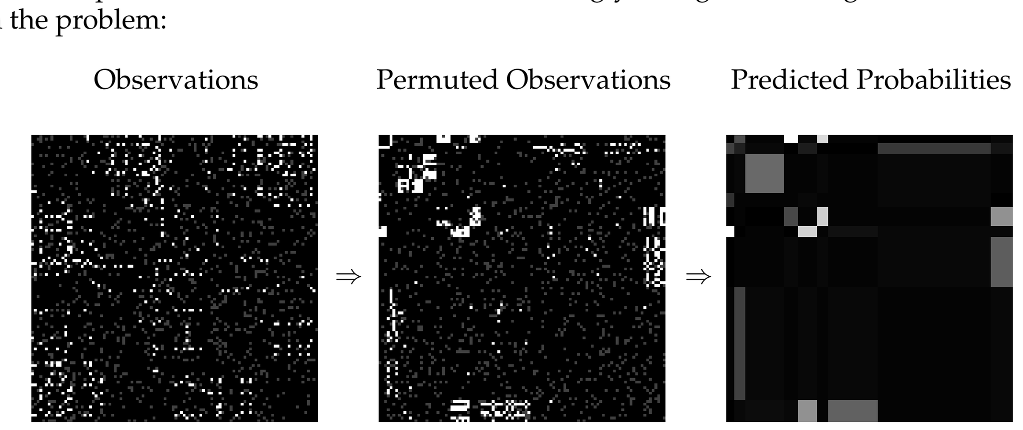
Our goal

An expressive, generative model for latent feature based link prediction.

Latent Class Overview

In recent years, several Bayesian models have been developed for link prediction in relational data. Models such as the Infinite Relational Model (IRM) (Kemp et al., 2006) and the Mixed-Membership Stochastic Blockmodel (MMSB) (Airoldi et al., 2006) assume that there exists a set of latent classes that each entity we observe belongs to and that each entity either belongs to a single class or has a distribution over the classes. Conditioned on the latent class membership of each node, all relations are assumed to be generated independently.

Why should class based approaches work? If we assign each entity to an appropriate class and permute the rows and columns accordingly, it might reveal significant structure in the problem:



Can latent class models capture all link behavior?

Technically yes, since we can place each entity in its own class. Practically, it can sometimes make more sense to look at other representations.

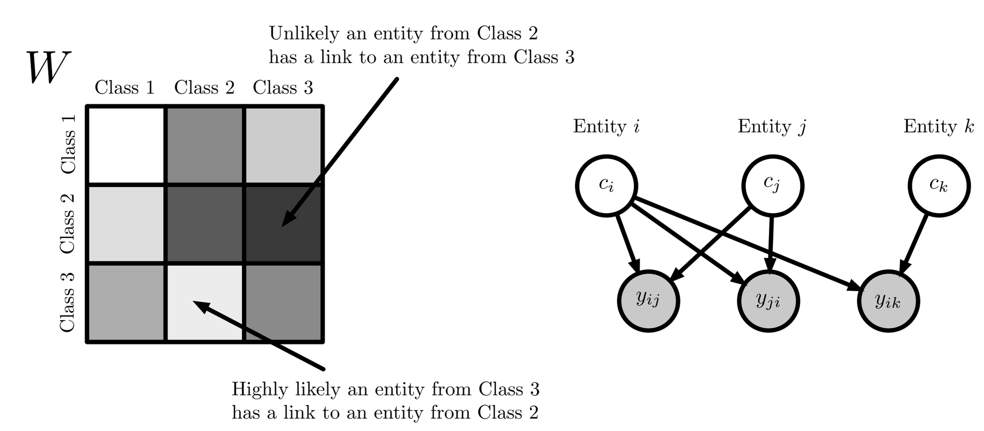
For example, in a high school social network, we might learn classes such as "high school student," "male," "athlete," and "musician." Restricting people to belong to one of these classes or to have a mixed membership ignores the shared information in these classes. By allowing ourselves to have a more featural description, we have greater flexibility.

Latent Class Approaches

We wish to move beyond latent class models to latent feature models, but before we do that, we review two of the main relevant latent class models. In all these approaches, we assume there are K latent classes (where K need not be fixed a priori).

The Infinite Relational Model (IRM)

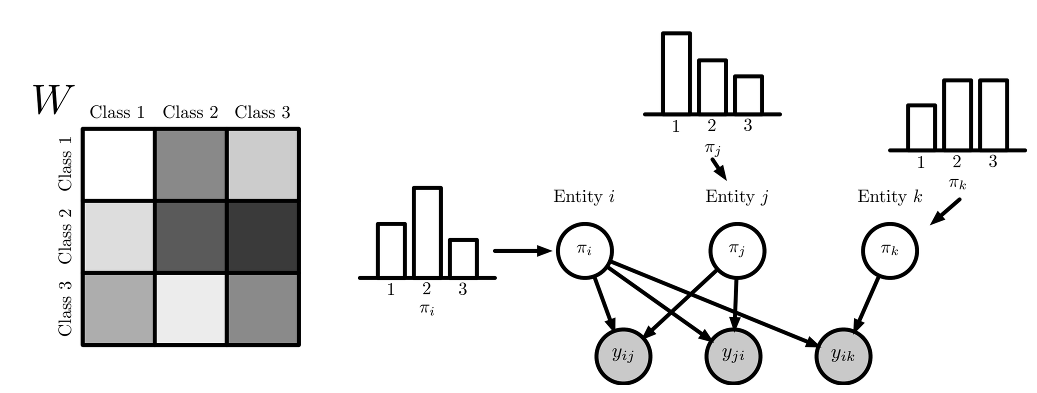
- Each entity i is assigned a class z_i. Draw these class assignments from the Chinese Restaurant Process.
- Generate a matrix W where for each pair of represented classes c_1 and c_2, draw $W(c_1, c_2) \in [0,1]$ where $W(c_1, c_2)$ is the probability of an entity in class c_1 having a link to an entity in class c_2.
- To generate whether or not there is a link from entity i to entity j, $Y(i,j) \in \{0,1\}$, draw $$Y(i,j) \sim \text{Bernoulli}(W(z_i, z_j)).$$



The Mixed-Membership Stochastic Blockmodel (MMSB)

- Each entity i has a mixed membership π_i, where π_i is a multinomial distribution over the classes drawn from a Dirichlet distribution.
- Generate a matrix W where for each pair of represented classes c_1 and c_2, draw $W(c_1, c_2) \in [0,1]$ where $W(c_1, c_2)$ is the probability of an entity in class c_1 having a link to an entity in class c_2. Remember, though, that entities now have mixed membership.
- To generate whether or not there is a link from entity i to entity j, $Y(i,j) \in \{0,1\}$, draw $$Y(i,j) \sim \text{Bernoulli}(\pi_i^\top W \pi_j).$$

To do this, each entity draws a class to use in relation $Y(i,j)$, and then $Y(i,j)$ is drawn from the corresponding part of W.



The Indian Buffet Process/Beta Process

Before describing our latent feature model, we explain the Bayesian nonparametric prior we will use to model the latent features.

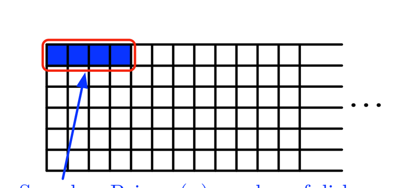
The *Indian Buffet process* (IBP) is a generative process that defines a prior on sparse binary matrices (Griffiths and Ghahramani, 2006). It is the limit of a beta-Bernoulli model on $N \times K$ matrices as $K \to \infty$ where N is the number of objects in our model and K is the fixed, finite number of features. To generate a matrix Z in the finite context, we sample

$$\pi_k \sim \text{Beta}(\alpha/K, 1) \qquad k \in \{1,\ldots,K\}$$
$$z_{ik} \sim \text{Bernoulli}(\pi_k) \qquad i \in \{1,\ldots,N\}, k \in \{1,\ldots,K\}$$

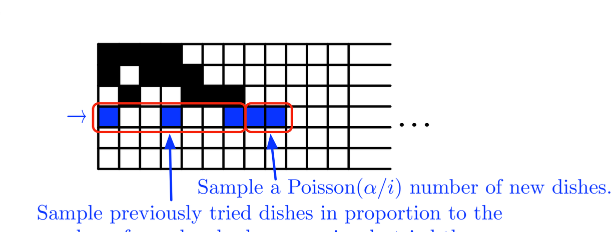
where α is a parameter. Conditioned on π_k all entries of the k^{th} column are independent Bernoulli samples.

The IBP is a culinary metaphor that describes how to generate matrices from this distribution when $K \to \infty$. Each row corresponds to a diner and each column corresponds to a dish at a buffet. An entry of one at (i,j) means the i^{th} diner tried the j^{th} dish. This matrix is filled in as follows:

- The first customer



- The i^{th} customer



The IBP is an infinitely exchangeable process, which means that the probability of Z is the same when you permute all the rows (up to a particular notion of equivalence classes). By De Finetti's Theorem, there must be some underlying stochastic process that when conditioned upon makes all observations independent. It was shown in (Thibaux and Jordan, 2007) that this underlying stochastic process is the *Beta Process*.

The Nonparametric Latent Feature Relational Model

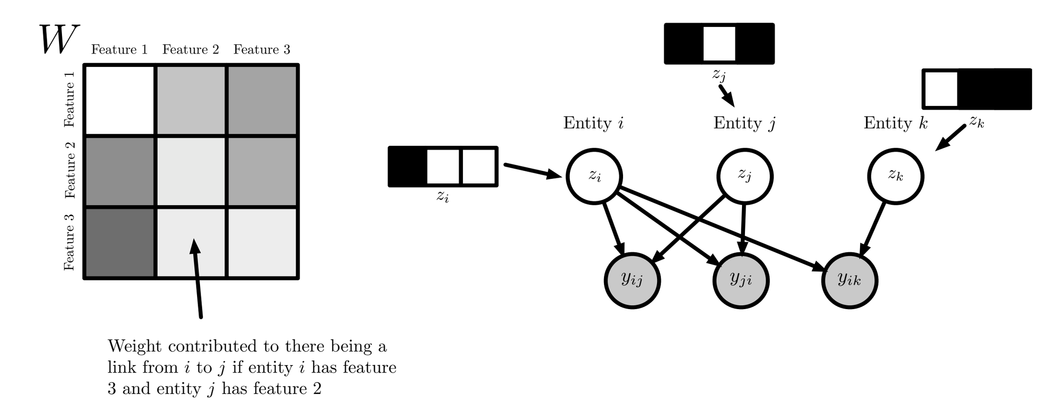
Why do we restrict ourselves to latent classes? Our model lets the latent variables be a binary vector indicating which of a set of features each entity has. This allows us to capture the factorial nature of many networks. We refer to this model as the *Nonparametric Latent Feature Relational Model* (NLFRM).

Basic Model

In the very basic version of the NLFRM, links are generated as follows:

- Every entity i is assigned a binary vector Z_i. Draw the latent feature matrix Z from the Indian Buffet Process.
- Generate a matrix W. Instead of restricting the entries to being in $[0,1]$, we let W be real-valued.
- The probability of there being a link from i to j in Y is then $$\Pr(y_{ij} = 1|Z_i, Z_j, W) = \sigma\left(Z_i^\top W Z_j\right)$$ where $\sigma(\cdot)$ is either the logistic or probit function in order to map real numbers to $[0,1]$ numbers.

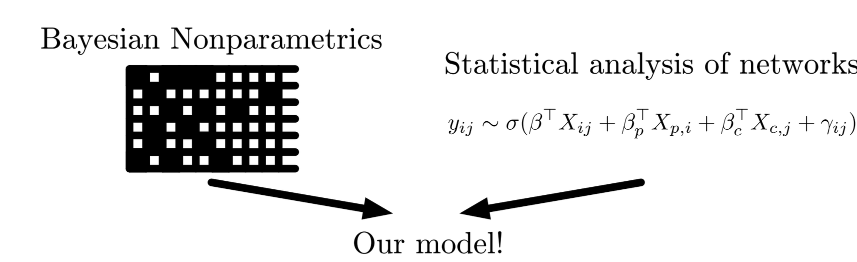
  If we wish to predict non-binary quantities, we can change $\sigma(\cdot)$ accordingly.



Concretely, the basic generative model is therefore

- $Z \sim \text{IBP}(\alpha)$.
- $W_{ij} \sim \mathcal{N}(0, \sigma_w^2)$ for all i,j.
- $Y \sim \sigma\left(Z^\top W Z\right)$.

Full Model



In our full model, we combine the power of Bayesian nonparametrics with logit models used for the statistical analysis of networks. Specifically, we are often given X, a set of known covariates. Assume we are given:

- X_{ij} - Covariates that influence y_{ij}.
- $X_{p,i}$ - Covariates of entity i when it is the parent of a link.
- $X_{c,i}$ - Covariates of entity i when it is the child of a link.

Then by introducing additional normally distributed parameters $a, b, c, \beta, \beta_p, \beta_c$, our full model is:

$$\Pr(y_{ij} = 1|Z_i, Z_j, W, X, \beta, a, b, c) = \sigma\left(Z_i W Z_j^\top + \beta^\top X_{ij} + (\beta_p^\top X_{p,i} + a_i) + (\beta_c^\top X_{c,j} + b_j) + c\right).$$

Furthermore, we might also observe multiple relations $Y^1, Y^2, \ldots, Y^m$ that we wish to simultaneously predict. In this case, we use a single set of features Z, but different weight vectors W^i for each relation.

Variations

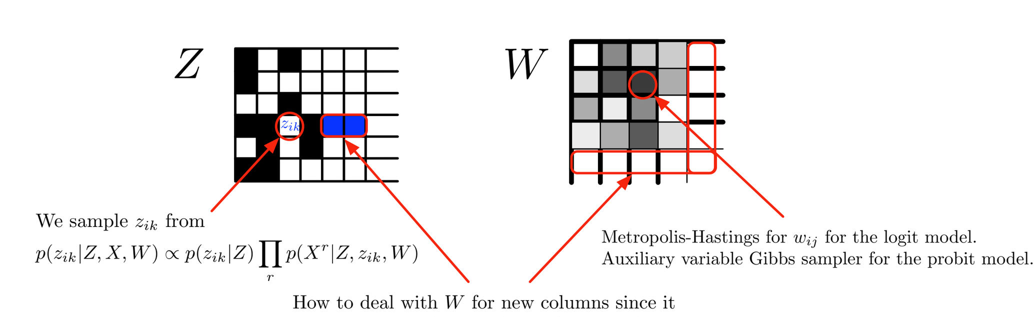
The above model works for general graphs or relations. Sometimes, we wish to work with undirected graphs or symmetric relations. Here there are two possibilities:

- In the most general case, let W be symmetric.
- If we assume relationships only depend on shared features, let W be diagonal.

Inference

The NLFRM is amenable to approximate posterior inference via Markov Chain Monte Carlo (MCMC). We focus on inference for the nonparametric component since inference for the parametric component has been addressed in the literature.

Sampling the non-zero columns of Z is straightforward. However, since we have a nonparametric model, we must sample from features and their weights that have not been observed yet. Since we have a non-conjugate model, we must deal with new weights intelligently.

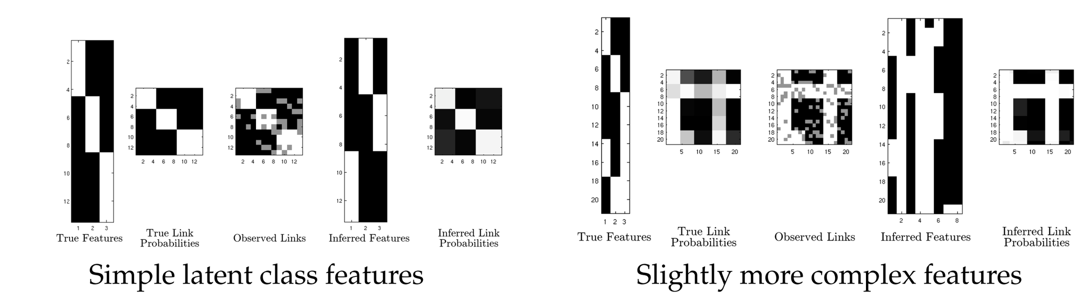


Results

We establish expectations of what the NLFRM can and cannot do with synthetic data before comparing it with latent class approached on three real datasets.

Synthetic data

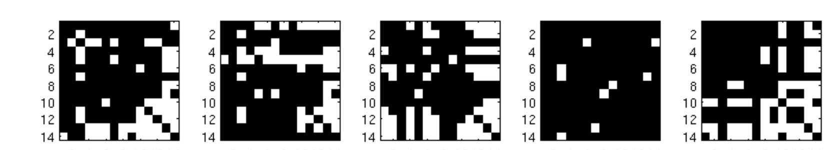
We generated very simple observations from two known sets of features to see how well we could explain the data and how well we could recover the features.



As can be seen, in both cases, we explain the data well, but except in the simplest cases, we cannot expect features to necessarily be interpretable or agree with known features.

Real data - Country Relations and Tribal Kinship Relations

We also tested our model on the Alywarra tribal kinship data and the country relations data used in (Kemp et al., 2006). The countries data set consisted of 56 different relations, five of which are shown below with relations such as "energy consumed," "illiterates," and "freedom of opposition."



The kinship data set consists of 26 familial relations amongst an Australian tribe. An example of one of these relations is seen on the far left.

On both of these datasets, we train on 80% of the data and test on the held-out 20%, measuring performance based on the Area Under the ROC Curve (AUC). We compare against the IRM and the MMSB, two latent class models.

	Countries single	Countries global	Alyawarra single	Alyawarra global
NLFRM w/ IRM	0.8521 ± 0.0035	**0.8772** ± 0.0075	0.9346 ± 0.0013	**0.9183** ± 0.0108
NLFRM rand	**0.8529** ± 0.0037	0.7067 ± 0.0534	**0.9443** ± 0.0018	0.7127 ± 0.030
IRM	0.8423 ± 0.0034	0.8500 ± 0.0033	0.9310 ± 0.0023	0.8943 ± 0.0300
MMSB	0.8212 ± 0.0032	0.8643 ± 0.0077	0.9005 ± 0.0022	0.9143 ± 0.0097

AUC on the countries and kinship datasets. Bold identifies the best performance.

Since there are many local optima when using the IBP, initialization was very important. We report results using a principled initialization taking advantage of the fact that the IRM is a special case of the NLFRM and a completely random initialization. In addition, we compare results when we train and test on each relation independently and when we learn a global latent representation for all relations.

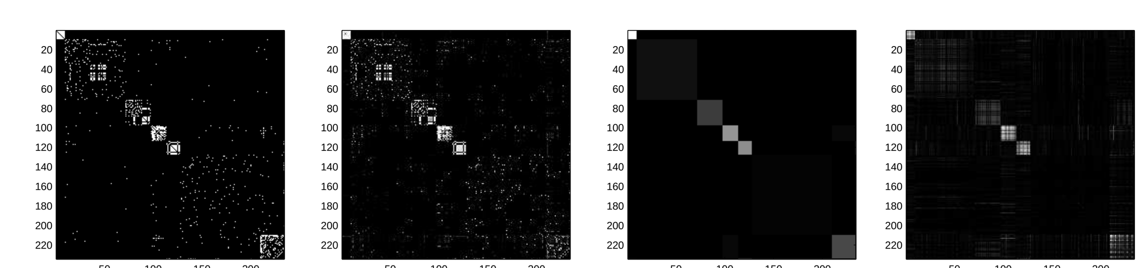
Real data - NIPS Coauthorship

To highlight the expressiveness of the NLFRM, we used the coauthorship data from NIPS 1-17 compiled in (Globerson et al., 2007). We took the 234 authors who had published most, and trained on 80% of the data and tested on the held-out 20%.

NLFRM w/ IRM	NLFRM rand	IRM	MMSB
0.9509	0.9466	0.8906	0.8705

AUC on the NIPS coauthorship data. All were within ±0.013

In addition to quantitatively being better, we are also qualitatively better since we have a much more expressive model.



(a) True relations (b) NLFRM predictions (c) IRM predictions (d) MMSB predictions

Predictions for all algorithms on the NIPS coauthorship dataset. In (a), a white entry means two people wrote a paper together. In (b-d), the lighter an entry, the more likely that algorithm predicted the corresponding people would interact.

References

E. M. Airoldi, D. M. Blei, E. P. Xing, and S. E. Fienberg. Mixed membership stochastic block models for relational data, with applications to protein-protein interactions. In *Proceedings of International Biometric Society-ENAR Annual Meetings*, 2006.

A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean embedding of co-occurence data. *The Journal of Machine Learning Research*, 8:2265–2295, 2007.

T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems (NIPS) 18*. Cambridge, MA: MIT Press, 2006.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the American Association for Artificial Intelligence (AAAI)*, 2006.

R. Thibaux and M. Jordan. Hierarchical beta processes and the indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2007.