Latent Variable Models for Network Data Over Time

James Foulds, Arthur Asuncion, Carter Butts, Christopher DuBois, Padhraic Smyth
Department of Computer Science, University of California, Irvine

Scalable Methods for the Analysis of Network-Based Data

1. Abstract

Real-world relational data sets, such as social networks, often involve measurements over time. We propose a Bayesian nonparametric latent feature model for such data, where the latent features for each actor in the network evolve according to a Markov process, extending recent work on similar models for static networks. We show how the number of features and their trajectories for each actor can be inferred simultaneously and demonstrate the utility of this model on prediction tasks using synthetic and real-world data.

2. Introduction

Latent variable models are a common approach to modeling social network data. In this style of model, actors are assumed to be represented by vectors of latent (i.e. unobserved) variables that, along with any observed covariates, determine the network structure. By inferring such latent variables from observed networks, it is possible to make predictions on unseen relationships, and sometimes to obtain a sociological explanation for network phenomena.

In this work, we extend the non-parametric latent feature relational model of Miller et al. (2009) to longitudinal networks, i.e. networks that change over time.

3. Prior Work: the Non-Parametric Latent Feature Relational Model (Miller et al., 2009)

- Each actor *i* is represented by a binary vector of features $\boldsymbol{Z}_i$
- Intuitively, these features may correspond to recreational interests, club memberships, social cliques, employment …
- The number of features *K* can be learned automatically due to the non-parametric Indian Buffet Process prior on $\boldsymbol{Z}$
- Edge probabilities are conditionally independent given the features
- Probability of edge between actor i and actor j is

$$\Pr(y_{ij}=1)=\sigma(Z_i W Z_j^T)$$

Logistic function

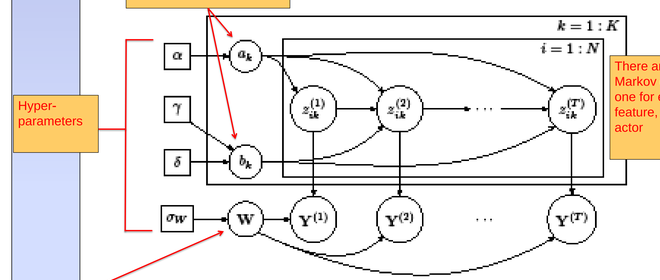
Weight matrix $\boldsymbol{W}$ specifies how the features interact

4. A Dynamic Relational Infinite Feature Model (DRIFT)

- A model for longitudinal (time-varying) networks.
- The model assumes the latent features of Miller et al. (2009).
- Actors change their features over time, using the Markov dynamics of the infinite Factorial Hidden Markov Model (Van Gael et al., 2009).
- In turn, the edge probabilities of the network change over time.
- The number of features can be determined automatically by the Bayesian MCMC inference algorithm, and therefore does not need to be specified in advance.

Transition probabilities for each feature

$$z_{ik}^{(t)} \sim \text{Bernoulli}(a_k^{1-z_{ik}^{(t-1)}} b_k^{z_{ik}^{(t-1)}})$$



Hyper-parameters

There are NxK Markov chains, one for each feature, for each actor

Real-valued weight matrix $\boldsymbol{W}$ specifies how the features affect edge probabilities

Networks at each timestep generated as in the Latent Feature Relational Model

$$y_{ij}^{(t)} \sim \text{Bernoulli}(\sigma(\mathbf{z}_i^{(t)} \mathbf{W} \mathbf{z}_j^{(t)\mathbf{T}}))$$

Figure 1. Graphical model for the finite version of DRIFT. The full model is defined to be the limit of this model as the number of features *K* approaches infinity. This "infinite" construction allows us to infer the number of "active" features from the data.

5. Markov Chain Monte Carlo Inference Algorithm

- Adaptively determine the number of features using the slice sampling trick based on the stick-breaking construction of the Indian Buffet Process.
- Blocked Gibbs sampling on the other variables.
  - Forward-backward dynamic programming on each actor's feature chain.
  - Metropolis-Hastings updates for $\boldsymbol{W}$s.

5. Experimental Analysis: Synthetic Data

Recovering the latent features on synthetic data.

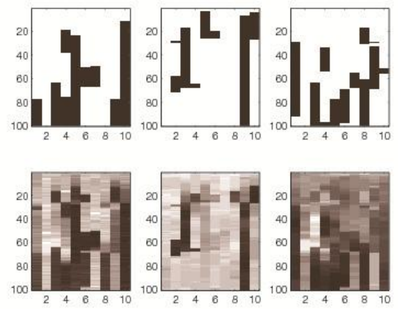
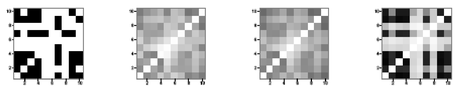


Figure 2. Ground truth features (top) and features learned by DRIFT (bottom). Each image represents one feature, with rows corresponding to timesteps and columns corresponding to actors.

Predicting the network at the next timestep



(a) True $\mathbf{Y}$ (b) Baseline (c) LFRM (all) (d) DRIFT

Figure 3. Predictive density for the next timestep after the training data according to DRIFT, LFRM and a simple baseline method, on synthetic data.

6. Experimental Analysis: Enron Company Email Data

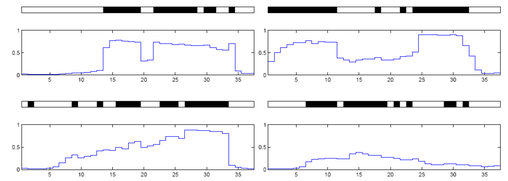


Figure 4. Presence and absence of an edge (black means that an edge is present) (top) and probability of an edge predicted by DRIFT, for four pairs of actors.

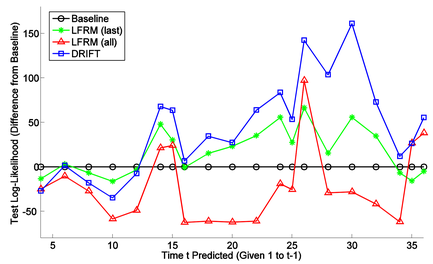
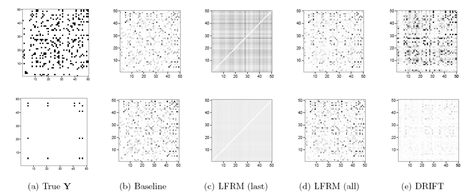


Figure 5. Predicting one timestep into the future.



(a) True $\mathbf{Y}$ (b) Baseline (c) LFRM (last) (d) LFRM (all) (e) DRIFT

Figure 6. Held out graph at time t = 30 (top row) and t=36 (bottom row), and posterior predictive distributions for each method.

7. Quantitative Experimental Results

Synthetic Dataset	Naive	Baseline	LFRM (last/current)	LFRM (all)	DRIFT
Forecast LL	-31.6	-32.6	-28.4	-31.6	**−11.6**
Missing Data LL	-575	-490	-533	-478	**−219**
Forecast AUC	N/A	0.608	0.779	0.596	**0.939**
Missing Data AUC	N/A	0.689	0.675	0.691	**0.925**
Enron Dataset	**Naive**	**Baseline**	**LFRM (last/current)**	**LFRM (all)**	**DRIFT**
Forecast LL	-141	-108	-119	-98.3	**−83.5**
Missing Data LL	-1610	-1020	-1410	-981	**−639**
Forecast AUC	N/A	0.874	0.777	0.891	**0.910**
Missing Data AUC	N/A	0.921	0.803	0.933	**0.979**

8. Discussion

We have proposed a latent feature model for network data over time and showed how to perform Bayesian inference on the model using an MCMC algorithm. Empirical results suggest that the proposed dynamic model can outperform static and baseline methods on both synthetic and real data.

References

K.T. Miller, T.L. Griffiths, and M.I. Jordan. Non-parametric latent feature models for link prediction. NIPS, 2009.
J. Van Gael, Y.W. Teh, and Z. Ghahramani. The infinite factorial hidden Markov model. NIPS, 2009.