Bayesian nonparametric models for prediction in networks

Sinead Williamson

Department of Statistics and Data Sciences
McCombs School of Business

THE UNIVERSITY OF TEXAS AT AUSTIN

Why model networks?

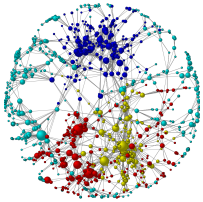
THE UNIVERSITY OF TEXAS AT AUSTIN

Many datasets take the form of networks or graphs...

- Social networks have binary (is friend, follows) or integer (retweets, shares) edges.

- Email networks have integer (number of emails) edges.

- Biological networks may have binary, integer or real-valued edges.



Modeling networks: Prediction vs Description

THE UNIVERSITY OF TEXAS AT AUSTIN

There are a number of reasons we might want to model networks...

- **Network recovery**: We may only have a noisy version of the underlying network.
- **Description/characterization**: We may wish to find a latent explanation of the network structure – e.g. community detection.
- **Anomaly detection**: We may wish to detect unusual nodes or sub-graphs - for example to detect spammers.
- **Influence modeling**: We may wish to identify "trend-setters" in a social network.
- **Link prediction**: We may wish to predict future links
  - Who will friend who next?
  - Who will email who next?
  - Where will the next interaction occur?
  - Where will there be future network traffic

Existing approaches

THE UNIVERSITY OF TEXAS AT AUSTIN

- Most existing Bayesian approaches have focused on **community detection**.
- Each node is modeled as belonging to one or more groups of communities.
- A node's community membership governs its relationships with other nodes.
- Intuition: If I am in the group "faculty", I am likely to send emails to the groups "students" and "faculty", but not to the group "sports clubs".

- The basic model of this type is the **stochastic blockmodel (SB)** [Wang and Wong, 1987, Snijders and Nowicki, 1997] :
  - We sample a global distribution over K community memberships, $\theta \sim \text{Dir}(\alpha)$.
  - For each node i, we sample a community membership $x_i \overset{i.i.d.}{\sim} \theta$.
  - For each pair of nodes i, j we have a set of parameters $B(x_i, x_j)$.
  - We sample an edge $z_{i \to j} \sim f(B(x_i, x_j))$.

- Infinite-dimensional variants such as the **Infinite Relational Model** [Kemp et al., 2006] replace the finite Dirichlet distribution with a Dirichlet process or other nonparametric prior, allowing an unbounded number of communities.

Mixed membership stochastic blockmodels

THE UNIVERSITY OF TEXAS AT AUSTIN

- A limitation of the Stochastic Blockmodel is that nodes can only belong to one cluster or community.
- Sometimes I might wear different "hats" – for example, I might be a member of the "faculty" group and also the "roller skating" group.
- I might interact with one person wearing my "faculty hat", but with another person wearing my "roller skating hat".
- The **mixed membership stochastic blockmodel** (MMSB) [Airoldi et al., 2008] incorporates mixed-membership behavior into the model.
  - For each node i, we sample a distribution over groups, $\pi_i \sim \text{Dir}(\alpha)$
  - For each pair of nodes, we draw two group membership indicators, $x_{i \to j} \sim \pi_i$ and $x_{i \leftarrow j} \sim \pi_j$
  - We then sample the edge based on these two group memberships, $z_{i \to j} \sim f(B(x_{i \to j}, x_{i \leftarrow j}))$

Limitations

THE UNIVERSITY OF TEXAS AT AUSTIN

- Stochastic blockmodel-based approaches are generally good at finding interesting community structure.
- Mixed membership variants allow even more flexibility (overlapping communities).
- They can be used to construct models for identifying influential or anomolous nodes.
- However, they generally perform poorly at prediction of future edges.
  - Zero edges explicitly instantiated.
  - Edges or nodes to be queried must be explicitly noted as "missing"
  - The number of nodes is fixed... so can't predict out-of-sample edges.
- Can only model dense networks [Orbanz and Roy, 2014]
  - Reasonable for small networks where everyone knows each other.
  - Unrealistic for large networks where each node only interacts with a small subset of all nodes.
- Typically scale poorly, since all N^2 edges are explicitly modeled.

A new nonparametric model for prediction

THE UNIVERSITY OF TEXAS AT AUSTIN

- Stochastic blockmodels cluster nodes based on the entirety of their interactions with other nodes.
  - Likelihood for each node depends on N interactions.
  - Number of nodes known a priori... even in nonparametric extensions.
- Rather than cluster the nodes using an (ad)mixture model, we can instead cluster the edges, based on their end-points.
- More concretely, we cluster *links* – binary interactions that sum to give the edge value.
  - Likelihood for each link depends on two endpoints.
  - Computationally advantageous if network sum (i.e. number of links) scales slower than N^2.
  - Directly modeling sequence of links gives us a simple predictive distribution.
  - Number of nodes unbounded... easy to predict out-of-sample.
- We are going to focus on **integer valued networks**, such as email networks, traffic networks, citation networks.
- Later, we will discuss ways of extending to binary networks.

A simple network model

THE UNIVERSITY OF TEXAS AT AUSTIN

- Let's first look at a simple example, before constructing a more realistic model.
- We will represent our integer-valued network as an exchangeable sequence of pairs, or links, (s_n, r_n), between a "sender" s_n and a "receiver" r_n.
- The value of an edge we sum over all links, $z_{i\to j} = \sum_{n=1}^{N} \mathbb{I}(s_n = i, r_n = j)$
- We place two coupled nonparametric distributions over the senders s_n and the receivers r_n:

$$H := \sum_{i=1}^{\infty} h_i \delta_{\theta_i} \sim \mathsf{DP}(\gamma, \Theta)$$

DP-distributed base measure ensures common support between senders and receivers

$$A := \sum_{i=1}^{\infty} a_i \delta_{\theta_i} \sim \mathsf{DP}(\tau, H) \qquad B := \sum_{i=1}^{\infty} b_i \delta_{\theta_i} \sim \mathsf{DP}(\tau, H)$$

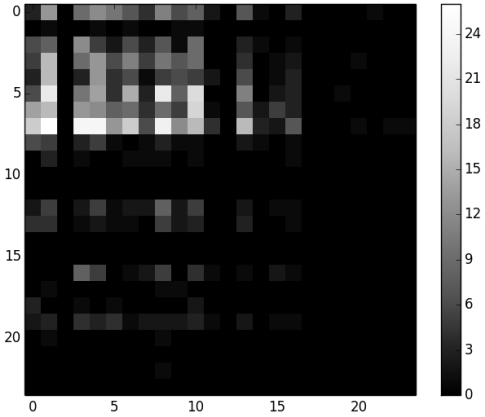
Separate DPs for senders and receivers

$$s_n \sim A \qquad r_n \sim B$$

Senders and receivers sampled from their respective distributions.

- We can obtain a symmetric graph by using a single Dirichlet process.
- We can obtain a bipartite graph by using a continuous basemeasure in lieu of A.

Structure of the simple model



- Unbounded number of nodes due to infinite-dimensional support.
- No real structure beyond a preferential attachment-like behavior.

Relationship to other models

THE UNIVERSITY OF TEXAS AT AUSTIN

- This is related to an integer-valued network model by [Caron and Fox, 2014]
- They sample edges according to a Poisson process with a discrete, symmetric, nonparametric base measure:

$$W \sim \mathsf{GGP}(\rho, \lambda) \qquad Z \sim \mathsf{PP}(W \times W),$$

where GGP indicates a generalized gamma process

Relationship to other models

- This is related to an integer-valued network model by [Caron and Fox, 2014]
- They sample edges according to a Poisson process with a discrete, symmetric, nonparametric base measure:

$$W \sim \text{GGP}(\rho, \lambda) \qquad Z \sim \text{PP}(W \times W),$$

  where GGP indicates a generalized gamma process
- This can be equivalently written as a symmetric version of our model, with a Poisson number of links:

$$W \sim \text{GGP}(\rho, \lambda) \qquad N \sim \text{Poisson}(W(\Omega)^2)$$

$$s_n, r_n \overset{i.i.d.}{\sim} \frac{W}{W(\Omega)}, n = 1, \ldots, N \qquad Z = \sum_{n=1}^{N} \delta_{(s_n, r_n)}$$

Relationship to other models

- This is related to an integer-valued network model by [Caron and Fox, 2014]
- They sample edges according to a Poisson process with a discrete, symmetric, nonparametric base measure:

$$W \sim \text{GGP}(\rho, \lambda) \qquad Z \sim \text{PP}(W \times W),$$

  where GGP indicates a generalized gamma process
- This can be equivalently written as a symmetric version of our model, with a Poisson number of links:

$$W \sim \text{GGP}(\rho, \lambda) \qquad N \sim \text{Poisson}(W(\Omega)^2)$$

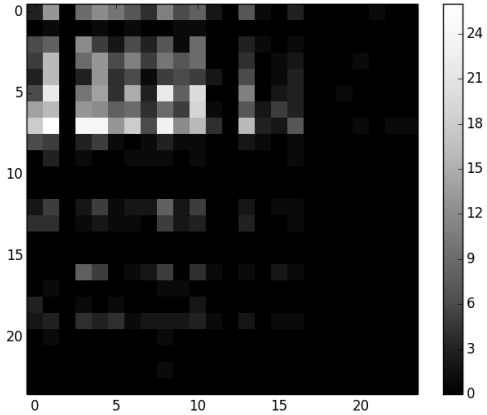
$$s_n, r_n \overset{i.i.d.}{\sim} \frac{W}{W(\Omega)}, n = 1, \ldots, N \qquad Z = \sum_{n=1}^{N} \delta_{(s_n, r_n)}$$

- Caron and Fox show that the resulting network is sparse in terms of the number of edges, and exhibits power-law degree distribution.
- We could obtain these guarantees by replacing the DPs with normalized generalized gamma processes.
  - Interesting question - do we get the same properties with Pitman-Yor processes?

Adding structure

THE UNIVERSITY OF TEXAS AT AUSTIN

These are certainly nice properties... but real networks don't look like this!



- Real networks have more complex structure than this.
- We see clustering and the formation of cliques... here we only have one cluster.

Adding structure

THE UNIVERSITY OF TEXAS AT AUSTIN

- Stochastic blockmodels/MMSB obtain this structure using a mixture/admixture of Erdös-Renyi-like network behaviors.
  - Each *individual* belongs to one (or more) latent clusters or communities.
  - Conditioned on the cluster indicators, links are exchangeable.
- We can follow a similar approach: A mixture of simple models!
  - Each component is a simple network.
  - Different components put high probability on different nodes.
- Intuition: Emails clustered by type of person they are to/from.
  - An *email* might belong to a faculty-to-student cluster.
  - This cluster assigns high probability to senders being faculty and receivers being students.
  - An *individual* might have high probability under several clusters.

Adding structure

THE UNIVERSITY OF TEXAS AT AUSTIN

More concretely,

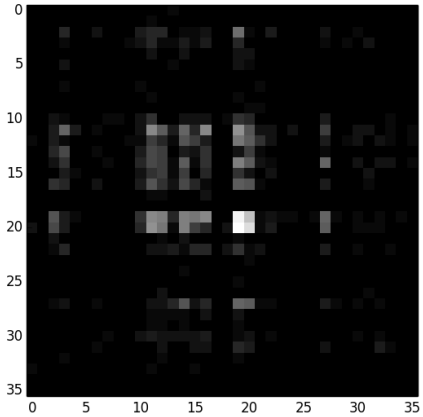
$$
\begin{aligned}
D := (d_k, k \in \mathbb{N}) &\sim \text{GEM}(\alpha) && \text{distribution over clusters} \\
H := \sum_{i=1}^{\infty} h_i \delta_{\theta_i} &\sim \text{DP}(\gamma, \Theta) && \text{shared distribution over nodes} \\
A_k := \sum_{i=1}^{\infty} a_{k,i} \delta_{\theta_i} &\sim \text{DP}(\tau, H) && \text{per-cluster distribution over sender nodes} \\
B_k := \sum_{i=1}^{\infty} b_{k,i} \delta_{\theta_i} &\sim \text{DP}(\tau, H) && \text{per-cluster distribution over receiver nodes} \\
c_n &\sim D && \text{pick a cluster for the } n\text{th link} \\
s_n &\sim A_{c_n} && \text{sample a sender...} \\
r_n &\sim B_{c_n} && \text{and a receiver}
\end{aligned}
$$

- To generate the nth link, we first select a cluster c_n.
- We then select a "sender" s_n and a "receiver" r_n—identifying a link (s_n, r_n)—according to group-specific distributions A_{c_n} and B_{c_n}.

Adding structure

THE UNIVERSITY OF TEXAS AT AUSTIN

Now we have some block structure!



- α controls the number of groups.
- τ controls the degree of similarity/degree of overlap between the groups.
- γ and τ control the total number of nodes and the sparsity of the resulting network.

Desirable properties

THE UNIVERSITY OF TEXAS AT AUSTIN

- Explicit distribution over exchangeable sequences of links
  - Straightforward Gibbs sampling (see later).
  - Easy to obtain predictive distribution over next link.
- MMSB-like structure.
  - Clusters can overlap, individuals can have high probability in multiple groups.
  - Automatically incorporates the fact that some people make more links than others.
- Unbounded number of nodes.
  - In a social network we want to predict linking to someone we haven't seen before.
- If we use a normalized generalized gamma process (and perhaps Pitman-Yor process?), we can get graph sparsity and power law behavior.
- Only models non-zero edges.
  - Inference scales linearly with the number of links.
  - In sparse, realistic networks, this grows slower than N^2.
  - e.g. Enron dataset has 500K emails between 37K individuals... number of emails $\approx N^2/2700$

Comparison with MMSB

THE UNIVERSITY OF TEXAS AT AUSTIN

- If we use an MMSB w/ Poisson likelihood (to generate integer-valued networks), the likelihood of a given edge is

$$P(Z_{ij} = z_{ij}) = \sum_{k=1}^{K} \sum_{\ell=1}^{K} \pi_{i \to j,k} \pi_{i \leftarrow j,\ell} \text{Poisson}(z_{ij}; B_{k\ell})$$

- We can transform this into a predictive model by using a Poisson *process* likelihood.
- Then, the probability that the *next* link being from node i to node j is

$$P((S_n, R_n) = (i, j)) \propto \sum_{k=1}^{K} \sum_{\ell=1}^{K} \pi_{i \to j,k} \pi_{i \leftarrow j,\ell} B_{k\ell}$$

- By comparison, the likelihood under our model is

$$P((S_n, R_n) = (i, j)) = \sum_{k=1}^{\infty} d_k a_{k,i} b_{k,j}$$

Inference

THE UNIVERSITY OF TEXAS AT AUSTIN

- At its heart, this is simply a Dirichlet process mixture of Dirichlet processes (with a little extra structure to ensure overlapping support of clusters).
- This is a fully nonparametric, rather than semiparametric model: The atoms of the component Dirichlet processes directly correspond to nodes.
  - We normally use a DP with some sort of parametric kernel... skipping this step makes computing the likelihood trivial.
  - We don't need to assign data to parametrized clusters within each group.
- We can directly extend existing Gibbs samplers for the Dirichlet process and Hierarchical Dirichlet process.
- We have developed exact urn-based schemes and asymptotically exact (weak limit) parametric schemes.
- Currently working on a variational inference algorithm to allow faster inference.

Qualitative evaluation

THE UNIVERSITY OF TEXAS AT AUSTIN

- To test whether our algorithm is finding reasonable structure, we evaluated the clusters found in Shakespearean plays.
- Each speech gives a link from the speaker to the other characters on stage at that time.

Raw text

ACT I
SCENE I. A desert place.
Thunder and lightning. Enter three Witches
First Witch
When shall we three meet again
In thunder, lightning, or in rain?
Second Witch
When the hurlyburly's done,
When the battle's lost and won.
Third Witch
That will be ere the set of sun.

Network

(First Witch, Second Witch)
(First Witch, Third Witch)
(Second Witch, First Witch)
(Second Witch, Third Witch)
(Third Witch, First Witch)
(Third Witch, Second Witch)

Qualitative evaluation

THE UNIVERSITY OF TEXAS AT AUSTIN

We can examine which characters appear as senders and receivers in each cluster (top 6 clusters shown):

- *(Ross, Banquo, Lennox, Donalbain, Malcolm)*
- *(First Murderer, Second Murderer, Third Murderer, Fleance, Macbeth, Banquo)*
- *(Third Witch, Second Witch, First Witch, Macbeth, Hecate)*
- *(Malcolm, Macduff, Siward, Angus, Caithness)*
- *(Scottish Doctor, Macbeth, Lady Macbeth, Servant 2, Gentlewoman)*
- *(Son of Macduff, Lady Macduff, Messenger 2, Ross)*

Similar results on other plays.

Evaluation: Empirical performance

THE UNIVERSITY OF TEXAS AT AUSTIN

- To evaluate predictive performance, we ran our model on a subset of the ENRON email dataset.
- We picked a week of email, and predicted the next day's emails.
- We compared against a Poisson variant of MMSB, as described earlier.
- Since MMSB cannot do out-of-sample prediction, we included all individuals from the 8-day period in our MMSB training set.

	MMSB (K=10)	MMSB (K=5)	Our model
01/3/2000	-9499.4	-10306.23	**-8648.16**
08/3/2000	**-5963.84**	-8083.43	-6976.93
15/3/2000	-9367.49	-10373.46	**-9143.96**
22/3/2000	**-5430.44**	-6968.4	-5981.72
29/3/2000	-17920.82	-19559.72	**-14695.02**
average	-9636.40	011958.25	**-9089.16**

- Note: Our model is significantly quicker than the MMSB on this data – running more than a week and more than K=10 was computationally infeasible.
- The MMSB result is somewhat artificial as we included the "new" users in our training set.

Extension: Binary networks

THE UNIVERSITY OF TEXAS AT AUSTIN

- The models presented are only appropriate for integer-valued networks.
- However they can be modified to allow binary networks:
  - We can create a **finitely exchangeable** binary network by thresholding the models described here.
  - We can perform inference by sampling the thresholded links as auxiliary variables.
- Below is a social network between dolphins, modeled using the finitely exchangeable variant (clusters agree with previous analyses):
  - *(Shmuddel, SN4, SM9, Fork, Stripes, TSN103)*
  - *(Haecksel, Cross, MN60, Vau, Jonah)*
  - *(Web, Jet, Zig, Feather, Wave)*
  - *(TR77, Bumper, TR99, SN96, Patchback)*
- This model loses interpretability as a predictive model, because we are no longer modeling a sequence of links.

Extension: Binary networks

THE UNIVERSITY OF TEXAS AT AUSTIN

- Another possible option is to create a **non-exchangeable** binary network.
- We sample the links sequentially, and threshold any repeated links.
- Intuition: The early links are likelier to have a higher number of thresholded counts.
- Non-exchangeable nature makes sense – we are likely to see frequently reinforced links early on.
- We can augment our model back to the integer model by sampling the number of thresholded counts.
  - Works well for small or sparse models.
  - If the network is large and dense this can quickly become infeasible.
- Alternative approach: modify approaches for restricted nonparametric processes [Williamson et al., 2013] to Metropolis-Hastings sample the latent measures.
- Work in progress.

Further extensions

THE UNIVERSITY OF TEXAS AT AUSTIN

- The simplicity of this model makes it easily modifiable.
- We are currently looking at incorporating **side information** about the links – specifically, text associated with emails.
- Our mixture model approach gives a natural framework for incorporating such information in a heterogeneous mixture model.
- Goal: suggest recipients based on text; inprove link prediction.

- Another interesting future direction is incorporating temporal dynamics.
- We could replace one or multiple Dirichlet processes with *dependent* Dirichlet processes to capture temporal evolution.
- This could capture both evolution in group dynamics, and deletion of edges.

Summary

THE UNIVERSITY OF TEXAS AT AUSTIN

- Existing Bayesian network models are geared towards characterization, often at the expense of predictive power.
- By performing clustering on links rather than on nodes, we can obtain explicit predictive models.
- Nonparametric mixture model allows individuals to belong to multiple groups, maintaining good characterization performance.
- Since we typically have fewer links than pairs of nodes, we often have faster inference than competing methods.

Questions?

References

Airoldi, E., Blei, D., Fienberg, S., and Xing, E. (2008).
Mixed membership stochastic blockmodels.
Journal of Machine Learning Research, 9:1901–2014.

Caron, F. and Fox, E. (2014).
Bayesian nonparametric models of sparse and exchangeable random graphs.
arXiv: 1401.1137.

Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T., and Ueda, N. (2006).
Learning systems of concepts with an infinite relational model.
In *AAAI*.

Orbanz, P. and Roy, D. (2014).
Bayesian models of graphs, arrays and other exchangeable random structures.
IEEE Transactions on Pattern Analysis and Machine Intelligence.

Snijders, T. and Nowicki, T. (1997).
Estimation and prediction for stochastic blockmodels for graphs with latent block structure.
Journal of Classification, 14:75–100.

Wang, Y. and Wong, G. (1987).
Stochastic blockmodels for directed graphs.
Journal of the American Statistical Association, 82:8–19.

Williamson, S., MacEachern, S., and Xing, E. (2013).
Restricting nonparametric distributions.
In *NIPS*.