Journal of Machine Learning Research 17 (2016) 1-21 Submitted 1/16; Revised 7/16; Published 11/16

# Nonparametric Network Models for Link Prediction

**Sinead A. Williamson** SINEAD.WILLIAMSON@MCCOMBS.UTEXAS.EDU
*Department of Statistics and Data Science/*
*Department of Information, Risk and Operations Management,*
*University of Austin at Texas*

**Editor:** Edo Airoldi

## Abstract

Many data sets can be represented as a sequence of interactions between entities—for example communications between individuals in a social network, protein-protein interactions or DNA-protein interactions in a biological context, or vehicles' journeys between cities. In these contexts, there is often interest in making predictions about future interactions, such as who will message whom.

A popular approach to network modeling in a Bayesian context is to assume that the observed interactions can be explained in terms of some latent structure. For example, traffic patterns might be explained by the size and importance of cities, and social network interactions might be explained by the social groups and interests of individuals. Unfortunately, while elucidating this structure can be useful, it often does not directly translate into an effective predictive tool. Further, many existing approaches are not appropriate for sparse networks, a class that includes many interesting real-world situations.

In this paper, we develop models for sparse networks that combine structure elucidation with predictive performance. We use a Bayesian nonparametric approach, which allows us to predict interactions with entities outside our training set, and allows the both the latent dimensionality of the model and the number of nodes in the network to grow in expectation as we see more data. We demonstrate that we can capture latent structure while maintaining predictive power, and discuss possible extensions.

**Keywords:** Dirichlet process, networks, Bayesian nonparametrics, Gibbs sampling, hierarchical modeling

## 1. Introduction

We are often interested in characterizing and predicting the interactions between objects, be they individuals within an organization, proteins within a cell, or transportation hubs within a region. We can represent these objects as nodes in a network, with the non-zero edges of the network describing the interactions between nodes. For example, we can represent a social network as a binary network, where each node corresponds to an individual, and an edge between nodes corresponds to a friendship between individuals. Patterns of email communication can be modeled using an integer-valued network, with integer-valued edges representing the number of emails sent from one individual to another. Interactions between proteins can be represented using a real-valued network, where the nodes correspond to proteins and the edges correspond to interaction strength.

A number of statistical models for such networks have been proposed. Many of these models fall under the *stochastic blockmodel* (SB) framework (Holland et al., 1983; Wang and Wong, 1987; Snijders and Nowicki, 1997), where each node is assumed to belong to one of $K$ latent groups, and the interaction between two nodes depends only on their group assignments. This basic model can be extended by allowing the number of latent groups to be unbounded, as in the infinite relational model (IRM, Kemp et al., 2006), or by allowing each node to exhibit membership in multiple latent groups, as in the mixed membership stochastic blockmodel (MMSB, Airoldi et al., 2008). One thing that these models have in common is that they treat nodes as exchangeable, and assume that there exists a fixed, stationary network between these nodes. Each node is represented by the totality of its interactions with other nodes, and we use this information to cluster (or, in the case of the MMSB, co-cluster) the data into distinct groups.

In this paper we follow a different approach: we treat the interactions, rather than the nodes, as data points, and construct an exchangeable sequence of directed binary links. Each link corresponds to a single interaction—such as "friending" or "liking" in a social network, or sending a single email—and is characterized in terms of an ordered pair of nodes. We may observe multiple links between two nodes; this corresponds to repeated interactions (for example, sending multiple emails).

This approach has a number of advantages. Unlike the stochastic blockmodel family, the approach described in this paper allows us to model sparse graphs, where the number of non-zero entries grows as $O(M)$, where $M$ is the number of nodes. Sparsity is a property of many real-life networks, which tend to exhibit small-world behavior (Caron and Fox, 2015; Orbanz and Roy, 2014).

Another advantage is that our model is explicitly designed for the prediction task. In many scenarios, we might be interested in what the next interaction will be: who will email whom, for example. Stochastic blockmodels aim to model a fully observed network, where the absence of an observed edge is interpreted as an explicitly observed zero. In this setting, any predictions must directly contradict these observed zeros. While it is possible to explicitly mark edges as "missing", we can only do this for a small subset of unobserved edges—if we assume all zero edges in a stochastic blockmodel are in fact unobserved, the maximum likelihood network will have all the edges equal to one. Conversely, by constructing an integer-valued network via an exchangeable sequence of links, we frame our problem in a manner that directly provides a predictive distribution over the location of the next link, and allows us to continuously update our posterior predictive distribution in the face of new data. Further, by choosing to place a nonparametric distribution over the sequence of links, we can easily incorporate previously unseen nodes, without any prior knowledge of the number of such nodes.

A further advantage is seen in the computational complexity of the model. Under a stochastic blockmodel with $M$ nodes and $K$ clusters, the computational cost of evaluating the likelihood of the $i$th node belonging to the $k$th cluster grows as $O(M)$, meaning that the overall computational cost of inferring the cluster allocations, without resorting to approximations, scales as $O(M^2K)$. Conversely, under the proposed model, the cluster likelihood for a link involves only the two nodes associated with the interaction, yielding $O(N)$ computational complexity where $N$ is the number of links. If $N$ grows significantly slower than $M^2$—as is the case if we assume sparsity—this will lead to computational savings as our data set grows.

This paper begins by examining existing Bayesian network models in Section 2, before going on to present our model in Section 3. We begin by introducing the Dirichlet network distribution in Section 3.1, before proceeding to describe how a mixture of such distributions can create a flexible network model with interesting latent structure in Section 3.2. While the focus of this paper is on integer-valued networks, we also consider extensions to the binary case in Section 3.3. While the primary method we consider does not exhibit exchangeable links, we can sample an auxiliary integer-value network and leverage its exchangeability to obtain predictive distributions. In Section 4, we describe an MCMC sampler for the mixture of Dirichlet network distributions, before presenting experimental results in Section 5.

### 1.1 Notation

We will use the notation $Z$ to represent an $M \times M$ network, with elements $z_{sr} \in \mathbb{N}$ indicating the relationship between nodes $s$ and $r$. If $z_{sr} \in \{0, 1\}$, then a non-zero value indicates the presence of a relationship. If $z_{sr}$ is allowed to take on arbitrary non-negative integer values, we take this to indicate the number of interactions (for example, emails in a social network, packages in a computer network) between nodes $s$ and $r$. Unless otherwise specified, we will assume $Z$ to be a directed network, where $Z \neq Z^T$.

It will sometimes be more convenient to represent the matrix $Z$ as a sequence of interactions $Y = y_1, y_2, \ldots$, where each interaction $y_i$ consists of an ordered pair of nodes. We can reconstruct the matrix Z by setting $z_{sr} = \sum_i \mathbb{I}(y_i = (s, r))$, where $\mathbb{I}(\cdot)$ represents an indicator function, that returns one iff the statement it refers to is true.

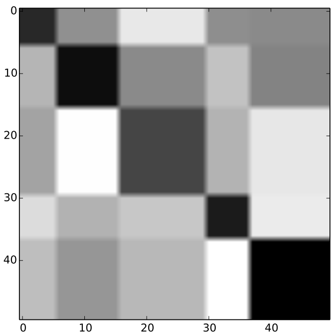
## 2. Related Work

A number of models have been proposed for modeling interactions within a network. In a Bayesian context, most of these models fall under the general category of stochastic blockmodels (SBs), a class of clustering-based models that generate dense networks. We discuss stochastic blockmodels in Section 2.1.

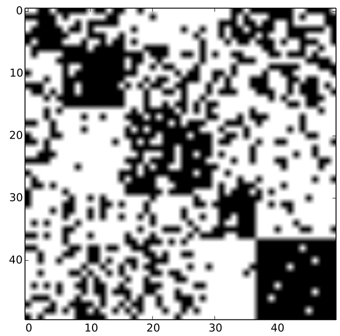
Recently, there has been growing interest in models for sparse networks, where the number of edges grows linearly (rather than quadratically) with the number of nodes. We discuss relevant work in this area in Section 2.2.

### 2.1 Stochastic Blockmodels and Related Models

The basic stochastic blockmodel (Holland et al., 1983; Wang and Wong, 1987) posits that each node belongs to one of $K$ latent clusters. For each pair $(i, j)$ of clusters there is a cluster-specific distribution over interactions governed by some parameter $\theta_{i,j}$; conditioned on their cluster allocations $c_s$ and $c_r$, interactions between nodes $s$ and $r$ are i.i.d. samples from the distribution parametrized by $\theta_{c_s,c_r}$. Typically, this is a Bernoulli distribution, giving rise to a binary matrix; however Mariadassou et al. (2010) show that the basic idea can be extended to real- or integer-valued networks by using different choices of distribution (for example, Gaussian or Poisson).




(a) Latent structure of the stochastic blockmodel.




(b) Sample from the stochastic blockmodel.

Figure 1: A stochastic blockmodel with 5 clusters.

To position SBs in a Bayesian context (Snijders and Nowicki, 1997), we can place appropriate conjugate priors on the cluster parameters $\theta_{i,j}$ and the cluster proportions. For example, a Bayesian version of the standard Bernoulli stochastic blockmodel takes the form

$$\begin{aligned}\pi &\sim \text{Dirichlet}(\phi)\\ \theta_{i,j} &\sim \text{Beta}(\alpha,\beta),\quad i,j\in\{1,\dots,K\}\\ c_s &\sim \text{Discrete}(\pi),\quad s\in\{1,\dots,M\}\\ z_{sr} &\sim \text{Bernoulli}(\theta_{c_s,c_r}).\end{aligned}$$

Figure 1a shows the underlying partitioning of the nodes, with the color of each partition indicating the corresponding value of $\theta_{c_s,c_r}$ in a binary SB. Figure 1b shows an instantiation of a network generated from this structure.

The basic stochastic blockmodel can be extended in a number of ways. The infinite relational model (IRM, Kemp et al., 2006) allows an unbounded number of latent clusters by placing a Dirichlet process prior over the cluster probabilities; this removes the need to prespecify a latent dimensionality and allows the number of clusters to grow (in expectation) with the number of nodes. A more general class of models is obtained if we associate each pair of nodes with a location in some metric space, and place a continuous or piecewise-continuous parameter function over this space (Lloyd et al., 2012); this class of model includes the SB and the IRM.

The SB and the IRM both assume that each node belongs to a single cluster. The mixed-membership stochastic blockmodel (MMSB, Airoldi et al., 2008) relaxes this assumption by associating each node with a *distribution* over latent clusters. To generate the interaction between nodes $s$ and $r$, each node selects a cluster from their individual distributions over clusters; the interaction is then generated according to the distribution associated with this pair. This extension allows the model to capture the fact that individuals may perform multiple "roles" leading to different patterns of interaction: Ian may be friends with Jamel because they both play tennis, and friends with Keira because they both study computer science.

While SBs are well-suited to community detection, they are less appropriate for the task of predicting unseen interactions, i.e., asking questions such as "who will Ian interact with next?". Stochastic blockmodels, and related models, explicitly model the entire network, with the likelihood for a data point's cluster allocation(s) depending on its interactions with all $M$ nodes in the network. Unless explicitly marked as missing, zeros in the network indicate the observed absence of a relationship, and affect the likelihood. As a result, predictions about the locations of unseen interactions must directly contradict the zeros present in the training set. This cannot be avoided by marking all zeros as unobserved, since in this case the maximum likelihood network is maximally connected.

Further, if we explicitly model the absence of interactions, the model likelihood is changed if we discover the existence of an $(M+1)$st node who hasn't yet interacted with anyone—since we must explicitly cluster this node and include its interactions in other nodes' likelihood terms. Therefore, if we want to allow prediction of links to or from individuals not included in our training set, we must know in advance the number of such individuals. This is not realistic in many settings: we do not know how many people will join a social network in the future.

Another consequence of modeling both non-zero- and zero-valued edges is that the computational cost of evaluating the conditional probability that node $i$ belongs to cluster $k$ scales linearly with the number of nodes, since

$$P(c_s = k | c_{-s}, Z, \pi, \{\theta_{i,j}\}) \propto \pi_k \prod_{r=1}^{M} P(z_{sr}|\theta_{k,c_r}) P(z_{rs}|\theta_{c_r,k}).$$

Therefore, resampling the cluster allocations of all $M$ cluster allocations scales quadratically with $M$ (and linearly with $K$). As $M$ grows, Gibbs sampling (Snijders and Nowicki, 1997) quickly becomes computationally infeasible. Variational methods (Mariadassou et al., 2010; Airoldi et al., 2008) generally give faster inference (albeit at the cost of lower estimate quality); however they still scale quadratically in the number of nodes. A number of approximate methods have been proposed that reduce the computational cost by approximating the full likelihood (Amini et al., 2013; Ho et al., 2012); however as with variational methods these approaches are no longer asymptotically guaranteed to sample from the true model.

## 2.2 Statistical Models for Sparse Networks

A limitation of stochastic blockmodel-type approaches—and indeed the more general class of models discussed in Lloyd et al. (2012)—is that they yield dense models almost surely (Orbanz and Roy, 2014). In other words, the number of non-zero entries in the resulting network grows as $O(M^2)$. This follows from the structure shown in Figure 1a: each partition is a simple Erdös-Rényi $G(n,p)$ model, with all possible relationships being equally likely, and the number of non-zero relationships growing in expectation with the number of pairs of nodes in that partition.

This contrasts with the sparse nature of many real-life social networks, where the number of non-zero entries grows as $O(M)$, since the number of interactions a person makes does not grow proportionately with the size of the network. In general, an individual will only interact with a small subset of the total population. Caron and Fox (2015) have shown that several real-world networks, including the ENRON data set explored in this paper, have a very high probability of exhibiting this form of sparsity.

Caron and Fox (2015) presented a construction for sparse networks based on a Poisson process. An integer-valued network is represented as a discrete measure $Z = \sum_{n=1}^{\infty} z_n \delta_{s_n, r_n}$ on $\mathbb{R}^2$, where each atom's size $z_n$ indicates the edge value, and the location $(s_n, r_n)$ specifies a pair of nodes. The atoms are distributed according to a Poisson process, with base measure given by the outer product of two generalized gamma process (GGP)-distributed random measures (Brix, 1999), i.e.

$$\begin{aligned} W &\sim \mathrm{GGP}(\rho, \lambda) \\ Z &\sim \mathrm{PP}(W \times W). \end{aligned} \tag{1}$$

This construction can also be used to generate a binary network, by thresholding the integer-valued network $N$. Caron and Fox (2015) demonstrate that the construction in Equation 1 can be used to generate networks that are sparse in terms of the number of edges, and exhibit power-law degree distribution. These are both properties that are commonly found in real world networks.

A link between the sparse binary models of Caron and Fox (2015) and the stochastic blockmodel family has recently been made explicit by Veitch and Roy (2015). Under their "graphex" construction for random binary networks, a candidate set of nodes, and associated node-specific parameters, is selected via a Poisson process on $\mathbb{R}^2$. As with the SB, the probability of a link between two nodes is governed by the nodes' parameter values via an appropriate link function, meaning that the SB is a member of this graphex-based class of models. However, different choices of link models can yield sparse graphs including the binary model of Caron and Fox (2015).

## 3. Nonparametric Models for Networks

As we saw in Section 2, stochastic blockmodels assume a fixed, fully observed network, where zero-valued entries are taken to represent the observed absence of an interaction, and model the network by clustering these nodes. We take a different approach: We model a network as a sequence of observed interactions, and aim to predict the locations of future interactions by explicitly clustering the interactions, rather than the nodes.

To do so, we consider distributions over a sequence of links connecting a set of nodes. Each link, therefore, is associated with an (ordered) pair of nodes sampled from some distribution over such pairs; we may have multiple links associated with a given pair. To allow the network to expand over time, and to facilitate out-of-sample prediction, we let this set of nodes be countably infinite and use a Bayesian nonparametric distribution to assign probabilities to potential pairs.

The main focus of this paper is on integer-valued networks—we will use the running example of an email network—where there can be multiple links between the same pair of nodes. While not explored in as much depth, we also suggest modifications that allow us to model binary networks in Section 3.3.

### 3.1 Dirichlet Network Distributions

A simple way of constructing an integer-valued network with an unbounded number of nodes is to place a probability distribution $G$ over a countably infinite number of actors. We can represent such a network as a sequence of (sender, receiver) pairs; each pair might, for example, correspond to a single email from a sender to a receiver, or a single journey between two cities. The value of a (directed) edge from a "sender" $s$ to a "receiver" $r$ is the number of times we have seen the pair $(s, r)$. We call each individual pair in the sequence a link; the value of an edge between two nodes is the number of links between them.

To generate such a pair, we simply sample a sender and a receiver according to $G$. Let $N$ be the total number of links in our network—that is, the total sum of the edge values. An appropriate prior over $G$ might be the Dirichlet process, so that

$$\begin{aligned} G &\sim \mathrm{DP}(\tau, \Theta) \\ s_n, r_n &\overset{i.i.d.}{\sim} G, \quad n = 1, \dots, N \\ z_{ij}^{(N)} &= \sum_{n=1}^{N} \mathbb{I}(s_n = i, r_n = j). \end{aligned} \tag{2}$$

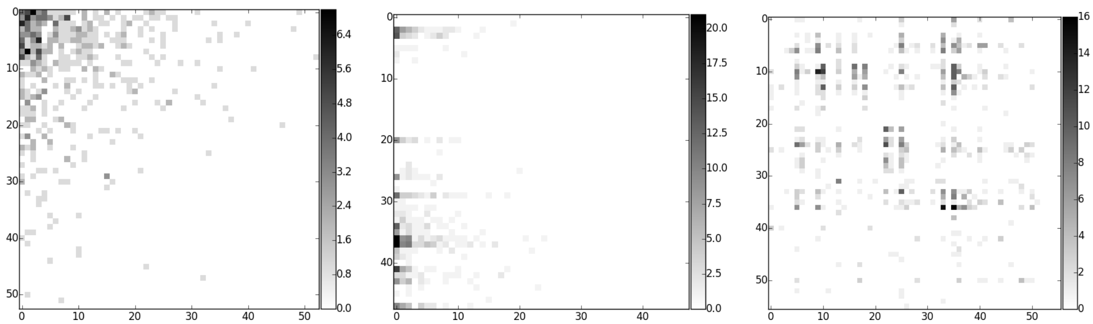
In other words, we generate a sequence of links by sampling with replacement from the distribution implied by the product measure $G \times G$. We will refer to this construction as a symmetric Dirichlet network distribution (DND). Figure 2a shows a network constructed in this manner.

This model is related to the sparse network model proposed by Caron and Fox (2015) and described in Section 2.2. As we described in Equation 1, Caron and Fox generate a directed, integer-valued graph $Z$ by sampling interactions according to a Poisson process, with rate given by the product measure $W \times W$ where $W \sim GGP$. If $W$ has finite total mass $W(\Omega)$, then this can equivalently be described as:

$$\begin{aligned} W \sim& \mathrm{GGP}(\rho, \lambda) \\ N \sim& \mathrm{Poisson}(W(\Omega)^2) \\ s_n, r_n \overset{i.i.d.}{\sim}& \frac{W}{W(\Omega)}, n = 1, \dots, N \\ Z =& \sum_{n=1}^{N} \delta_{(s_n, r_n)}. \end{aligned} \tag{3}$$

If we choose a standard gamma process as the random measure $W$ in Equation 3 then, conditioned on the total number of links $N$, we recover the symmetric nonparametric model described in Equation 2. In this paper, we focus on the gamma process/Dirichlet process case in order to achieve simple inference strategies; however the models proposed in this section can easily be extended to use a normalized generalized gamma process (Lijoi et al., 2008), or some other random probability measure such as a Pitman-Yor process (Pitman and Yor, 1997), in place of the Dirichlet process.

This basic model assumes that the probability of a node being the "sender" of a link is the same as the probability of being a "receiver". In practice, this is often not a reasonable



(a) Sample from a symmetric Dirichlet network distribution ($\tau = 10, N = 500$).

(b) Sample from an asymmetric Dirichlet network distribution ($\gamma = 50, \tau = 10, N = 500$).

(c) Sample from an asymmetric mixture of Dirichlet network distributions ($\gamma = 20, \tau = 5, \alpha = 5, N = 1000$).

Figure 2: Samples from Dirichlet network distributions and mixtures of Dirichlet network distributions.

assumption. For example, in a university's email network, administrators may send out a large number of group emails—that is, they often operate in the "sender" role—but receive a relatively small number of emails. When modeling human migration patterns, the United States had over 13 times as many immigrants as emigrants in 2015, whereas Micronesia had around twice as many emigrants as immigrants (United Nations, Department of Economic and Social Affairs, Population Division, 2015). We can capture this form of asymmetry by replacing the single distribution $G$ over nodes (Equation 2) with a pair of distributions, $A$ and $B$. To ensure the two distributions have the same support (meaning that any node can have both incoming and outgoing links), we couple these two distributions via a shared, discrete base measure $H$. The generalized process becomes

$$\begin{aligned} H &:= \sum_{i=1}^{\infty} h_i \delta_{\theta_i} \quad \sim \mathrm{DP}(\gamma, \Theta) \\ A &:= \sum_{i=1}^{\infty} a_i \delta_{\theta_i} \quad \sim \mathrm{DP}(\tau, H) \\ B &:= \sum_{i=1}^{\infty} b_i \delta_{\theta_i} \quad \sim \mathrm{DP}(\tau, H) \end{aligned} \qquad \begin{aligned} s_n &\sim A,\ n = 1, \dots, N \\ r_n &\sim B,\ n = 1, \dots, N \\ z_{ij}^{(N)} &= \sum_{n=1}^{N} \mathbb{I}(s_n = i, r_n = j). \end{aligned} \tag{4}$$

Figure 2b shows a network constructed in this manner; note the nodes that "send" a large number of links are not necessarily those that "receiver" a large number of links. The concentration parameter $\tau$ governs how similar the two distributions are to the base measure, and hence to each other.

On their own, like the sparse models explored by Caron and Fox (2015), the symmetric and asymmetric Dirichlet network distributions allow very little internal structure. While we do see some preferential attachment due to the discrete nature of the underlying random measures, there is no clustering structure—if we know that an email was sent by a given sender, this tells us nothing about the receiver. This makes it a poor model for real-world networks, where we observe cliques of users who are strongly interconnected, or groups that interact with other groups in characterizable manners. While we cannot capture such behavior with the basic symmetric or asymmetric Dirichlet network distributions described above, we can use them as a component of more complex and flexible models.

### 3.2 Mixtures of Dirichlet Network Distributions

Rather than use a single Dirichlet network distribution over integer-valued networks, as described by Equations 2 and 4, we can use a *mixture* of such distributions, which we will refer to as a mixture of Dirichlet network distributions, or MDND. By ensuring both sender and receiver belong to a common mixture component, we break the independence between sender and receiver, allowing us to identify communication patterns that cannot be captured using the DND or the related Caron and Fox (2015) model.. To allow links between the subgraphs associated with each mixture component, we couple the networks using a shared, discrete base measure $H$. Concretely, in the asymmetric case, let

$$\begin{aligned} D :=&(d_k, k \in \mathbb{N}) &&\sim \mathrm{GEM}(\alpha) \\ H :=&\sum_{i=1}^{\infty} h_i \delta_{\theta_i} &&\sim \mathrm{DP}(\gamma, \Theta) \\ A_k :=&\sum_{i=1}^{\infty} a_{k,i} \delta_{\theta_i} &&\sim \mathrm{DP}(\tau, H),\ k=1,2,\dots \\ B_k :=&\sum_{i=1}^{\infty} b_{k,i} \delta_{\theta_i} &&\sim \mathrm{DP}(\tau, H) \end{aligned} \qquad \begin{aligned} c_n &\sim D,\ n=1,\dots,N \\ s_n &\sim A_{c_n} \\ r_n &\sim B_{c_n} \\ z_{ij}^{(N)} &= \sum_{n=1}^{N} \mathbb{I}(s_n = i, r_n = j), \end{aligned} \tag{5}$$

where GEM($\alpha$) is the distribution over the size-biased atom sizes of a Dirichlet process with concentration parameter $\alpha$. Figure 2c shows a network constructed in this manner. A symmetric version of the MDND is recovered if we replace the sender- and receiver-specific distributions $A_k$ and $B_k$ with a shared distribution $G_k \sim \mathrm{DP}(\tau, H)$, and is appropriate when we believe the distribution over edges originating from a given node is similar to the distribution over edges ending at that node. For the remainder of this paper, we will focus on the asymmetric setting.

We can verbalize the generative process of the asymmetric MDND as follows. To generate the $n$th 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 the cluster-specific distributions $A_{c_n}$ and $B_{c_n}$. The concentration parameters $\alpha$, $\tau$ and $\gamma$ can be manipulated to obtain differing network properties. The parameter $\alpha$ controls the number of clusters, with the total number of clusters used to model $N$ links growing approximately as $O(\gamma \log N)$. The parameter $\tau$ controls the degree of similarity between the clusters: As $\tau$ decreases, the overlap between clusters will tend to decrease. Increasing $\gamma$ increases the overall number of nodes represented in the network.

Since the pairs are sampled i.i.d. given the random measures, the resulting sequence is exchangeable, meaning the construction is appropriate for sequences of links where there is no specific ordering of the links, or where the order is believed to be irrelevant. This has useful implications for inference: It means we can easily construct a Gibbs sampler, as described in Section 4. However, in many networks the order in which links are formed does carry information. In Section 3.3, we discuss an extension for explicitly ordered binary links, and in Section 6, we will discuss possible extensions of the integer-valued network model to explicitly ordered links.

### 3.3 Extension: Binary-valued Networks

Many real-world networks exhibit binary, rather than real-valued, edges. One way of capturing this behavior is to threshold the integer-valued edges generated by the DND or the MDND. The most straightforward version of this is simply to sample a network $Z = (z_{ij})$ according to the DND or the MDND, and then generate a binary network $Y = (y_{ij})$ by letting $y_{ij} = 1$ iff $z_{ij} > 0$. If we place a negative binomial distribution over $N = \sum_{i,j} z_{ij}$, so that $N \sim \mathrm{NB}(\alpha, 1/(1+\beta))$ for some $\beta > 0$, we can represent this thresholded model as

$$\begin{aligned}
\Gamma := \sum_{k=1}^{\infty} \mu_k \delta_{\theta_k} &\sim \mathrm{GP}(\alpha D_0, \beta) & s_n^{(k)} &\sim G_k, \quad n = 1, \dots, N_k \\
G_0 &\sim \mathrm{DP}(\gamma, H_0) & r_n^{(k)} &\sim G_k \\
G_k &\sim \mathrm{DP}(\tau, G_0), \quad k = 1, 2, \dots & z_{ij}^{(N)} &= \sum_{k=1}^{\infty} \sum_{n=1}^{N_k} \mathbb{I}(s_n^{(k)} = i, r_n^{(k)} = j) \\
N_k &\sim \mathrm{Poisson}(\mu_k) & y_{ij}^{(N)} &= \mathbb{I}(z_{ij}^{(N)} > 0), \\
N &:= \sum_k N_k & &
\end{aligned} \tag{6}$$

where GP indicates a gamma process. This approach is a form of *restricted exchangeable distribution*, as described by Williamson et al. (2013). As in the unrestricted, integer-valued network, the distribution over edge events is exchangeable. We note that this truncation technique is the same as that used by Caron and Fox (2015) to generate binary networks, and indeed if $Z$ is the symmetric version of the DND given by Equation 3, and if we obtain a symmetric network by mirroring the link counts, then this construction corresponds to the binary network described in Caron and Fox (2015), itself a special case of the class of binary network models described by Veitch and Roy (2015).

If our observations are explicitly ordered, and we believe that ordering to be important, we can modify the DND or the MDND to sample *without* replacement from the set of possible links, giving a non-exchangeable model. This form of non-exchangeability mimics behavior found in many naturally-occurring networks. For example, in a social network, a user will add their close friends first, and then over time add more distant acquaintances. In integer-valued networks exchangeability can represent the fact that close friends will communicate both early and often, but in an exchangeable binary setting all relationships appear identical.

The resulting binary network model is mathematically equivalent to a *censored* DND or MDND, where we only observe the first instance of a link between nodes $i$ and $j$:

$$
\begin{aligned}
D :=& (d_k, k \in \mathbb{N}) && \sim \mathrm{GEM}(\alpha) \\
H :=& \sum_{i=1}^{\infty} h_i \delta_{\theta_i} && \sim \mathrm{DP}(\gamma, \Theta) \\
A_k :=& \sum_{i=1}^{\infty} a_{k,i} \delta_{\theta_i} && \sim \mathrm{DP}(\tau, H),\ k = 1, 2, \dots \\
B_k :=& \sum_{i=1}^{\infty} b_{k,i} \delta_{\theta_i} && \sim \mathrm{DP}(\tau, H)
\end{aligned}
\qquad
\begin{aligned}
c_t \sim& D, \quad t = 1, 2, \dots \\
s_t \sim& A_{c_t} \\
r_t \sim& B_{c_t} \\
y_{ij}^{(t)} =& \mathbb{I}\left( \sum_{t'=1}^{t} \mathbb{I}(s_{t'} = i, r_{t'} = j) > 0 \right).
\end{aligned}
$$

Due to the finite probability of sampling an existing $(s, r)$ pair, $Y^{(t+1)}$ may be the same as $Y^{(t)}$; instead we would likely work with the corresponding non-repeating sequence $(Z^{(n)}, n \in \mathbb{N})$, where $Z^{(n)} = \min_{t'} \left( Y^{(t')} : \sum_{t=1}^{t'} \mathbb{I}(Y^t \neq Y^{t-1}) = n \right)$. While this censored model is no longer exchangeable, we can make use of the underlying exchangeable sequence of interactions to make predictions.

### 3.4 Relationship to Other Models

As we described in Section 3.1, the Dirichlet network distribution is strongly related to the sparse network models of Caron and Fox (2015)—in fact, conditioned on the total number of links, the integer-valued model of Caron and Fox (2015) is a special case of the symmetric DND, and the truncated model of Equation 6 describes the binary model of Caron and Fox (2015) as a special case. However, the DND on its own lacks the flexibility to model structured networks, where nodes tend to belong to locally connected sub-networks, and where knowing who sent a message tells us something about the intended recipient. The mixture of Dirichlet network distributions allows us to capture multiple sub-networks, while allowing interaction between sub-networks via a common Dirichlet process-distributed base measure $H$.

A related integer-valued network model is described by Crane and Dempsey (2016). This model is explicitly designed for multi-way interactions, such as collaborations or actors co-starring in movies, but can be modified to give two-way interactions. The number of "roles" in an interaction is sampled from an appropriate distribution, and for each role, nodes are sampled from a (single) Pitman-Yor process. The two-way interaction setting corresponds to a Pitman-Yor variant of the symmetric Dirichlet network distribution obtained by replacing the Dirichlet process in Equation 2 with a Pitman-Yor process; as such it is unable to capture the clustering behavior obtained using the mixture of Dirichlet network distributions.

The models described in this paper also bear some similarity to stochastic blockmodels, which were described in Section 2. The main difference between the stochastic blockmodel family and the models proposed in this paper is that, under the blockmodel paradigm, nodes are clustered into a (potentially infinite, in the case of the IRM) number of clusters. Conversely, the MDND directly clusters links, rather than nodes, and represents a network as a (potentially infinite) sequence of pairs of nodes. This creates a natural framework for questions of prediction. For the basic symmetric DND described in Section 3.1, the predictive distribution over the next pair of nodes is available in analytic form via an urn representation. For the mixture models proposed in Sections 3.2 and 3.3, the predictive distribution depends on the values of latent variables, but we can easily sample from this distribution, as we will see in Section 4.

While the MDND does not explicitly cluster the nodes, we can obtain a similar mixed-membership interpretation to that found in the MMSB. We can think of each cluster of links representing a latent topic of conversation between nodes. Each topic of conversation is described by distributions over the nodes likely to take part in such a conversation. If we condition on the fact that the $s$th node is sending an email, we can use these distributions to infer the probability of that email belonging to a given discussion. Since the hierarchical construction of Equation 5 ensures that the topics of conversation have overlapping participants, the node will be associated with a conditional distribution over an unbounded number of conversations.

In Section 2.2, we discussed how the graphex construction for binary exchangeable networks by Veitch and Roy (2015) allows us to represent the stochastic blockmodel and the sparse binary model of Caron and Fox (2015) using a common framework. While the integer-valued MDND, and the non-exchangeable binary network described in Section 3.3, do not fall under the graphex framework, it suggests an alternative way to describe the exchangeable binary network obtained by thresholding a MDND (as described in Equation 6).

## 4. Fully Nonparametric Inference via an Urn Scheme

In the simple symmetric network model of Equation 2, we can directly evaluate the predictive distribution over the $n$th link, given the previous $n-1$ links, via a straightforward extension of the Pölya urn sampler for the Chinese restaurant process Neal (1998), where the probability of seeing a link between two nodes is proportional to the product of those nodes' degrees (excluding the link in question):

$$P(y_n = (s,r)|y_1,\dots,y_{n-1}) = \begin{cases} \frac{m_i(m_j+\mathbb{I}(i=j))}{(2n-2+\tau)(2n-1+\tau)} & \text{if } m_i \neq 0, m_j \neq 0 \\ \frac{m_i\tau}{(2n-2+\tau)(2n-1+\tau)} & \text{if } m_i \neq 0, m_j = 0 \\ \frac{m_j\tau}{(2n-2+\tau)(2n-1+\tau)} & \text{if } m_i = 0, m_j \neq 0 \\ \frac{\tau^{1+\mathbb{I}(i=j)}}{(2n-2+\tau)(2n-1+\tau)} & \text{if } m_i = 0, m_j = 0, \end{cases}$$

where $m_i = \sum_{n'=1}^{n-1}(\mathbb{I}(s_{n'}=i)+\mathbb{I}(r_{n'}=i))$ is the sum of the links to or from node $i$.

The MDND is based on a mixture of coupled hierarchical Dirichlet processes, allowing us to construct a collapsed Gibbs sampler by modifying the direct assignment sampler for the hierarchical Dirichlet process introduced by Teh et al. (2006). Recall that associated with each cluster $k$ we have a sender-specific distribution $A_k$ and a receiver-specific distribution $B_k$.[1] Let $\eta_k = \sum_{i=1}^N I_{c_i=k}$ be the number of links associated with cluster $k$; let $m_{k,i}^{(1)}$ be the number of edges associated with cluster $k$ that originate from node $i$ (that is, edges where node $i$ is the "sender"); and let $m_{k,i}^{(2)}$ be the number of edges associated with cluster $k$ that end at node $i$ (that is, edges where node $i$ is the "receiver"). We also introduce auxiliary

1. In this section, we focus on the asymmetric MDND, where there are separate distributions over senders and receivers; extension to the symmetric case is straightforward.

count variables $\rho^{(1)}_{k,i}$ and $\rho^{(2)}_{k,i}$[2] and a probability vector $(\beta_1, \dots, \beta_J, \beta_u) \sim \text{Dir}(\rho^{(\cdot)}_{\cdot 1}, \dots, \rho^{(\cdot)}_{\cdot J}, \gamma)$, where $\rho^{\cdot}_{\cdot i} = \sum_k \rho^{(1)}_{k,i} + \rho^{(2)}_{k,i}$; here $\beta_1, \dots, \beta_J$ correspond to the atoms of $H$ that are associated with represented nodes, and $\beta_u = \sum_{j=J+1}^{\infty} h_j$.

The distribution over the cluster assignment of the $n$th link, given $\beta$ and all other $N-1$ links, is given by:

$$P(c_n = k | s_n, r_n, c^{\neg n}, \beta) \propto \begin{cases} \eta_k^{\neg n}(m^{(1)\neg n}_{k,s_n} + \tau\beta_{s_n})(m^{(2)\neg n}_{k,r_n} + \tau\beta_{r_n}) & \text{if } \eta_k^{\neg n} > 0 \\ \alpha\tau^2\beta_{s_n}\beta_{r_n} & \text{if } \eta_k^{\neg n} = 0. \end{cases} \tag{7}$$

where we use $c^{\neg n}$ to indicate the sequence $(c_{n'} : n' \neq n)$ and $m^{(1)\neg n}_{k,i} = \sum_{n' \neq n} I_{c_{n'}=k, s_{n'}=i}$, i.e., the $\neg n$ notation is used to exclude the value associated with the current observation.

Following Fox et al. (2007) we can sample the "dish counts" $\rho^{(1)}_{k,i}$ ($\rho^{(2)}_{k,i}$) by simulating the partitioning of $m^{(1)}_{k,i}$ ($m^{(2)}_{k,i}$) according to a Chinese restaurant process with parameter $\tau\beta_k$.

Conditioned on the cluster assignments for the first $N$ links, and the probability vector $(\beta_1, \dots, \beta_J, \beta_u)$, we can evaluate the predictive distribution over the $N+1$st link as:

$$P(y_{N+1} = (s,r) | c_{1:N}, y_{1:N}, \beta) = \begin{cases} \sum_{k=1}^{K_+} \frac{\eta_k}{N+\alpha} \frac{m^{(1)}_{k,s} + \tau\beta_s}{\eta_k + \tau} \frac{m^{(2)}_{k,r} + \tau\beta_r}{\eta_k + \tau} + \frac{\alpha}{N+\alpha}\beta_s\beta_r & \text{if } s, r \leq J \\ \sum_{k=1}^{K_+} \frac{\eta_k}{N+\alpha} \frac{m^{(1)}_{k,s} + \tau\beta_s}{\eta_k + \tau}\beta_u + \frac{\alpha}{N+\alpha}\beta_s\beta_u & \text{if } s \leq J, r > J \\ \sum_{k=1}^{K_+} \frac{\eta_k}{N+\alpha}\beta_u \frac{m^{(2)}_{k,r} + \tau\beta_r}{\eta_k + \tau} + \frac{\alpha}{N+\alpha}\beta_u\beta_r & \text{if } r \leq J, s > J \\ \beta_u^2 & \text{if } r, s > J. \end{cases}$$

To improve mixing, we augmented the sampler with split/merge moves, proposed using the Restricted Gibbs method of Jain and Neal (2004).

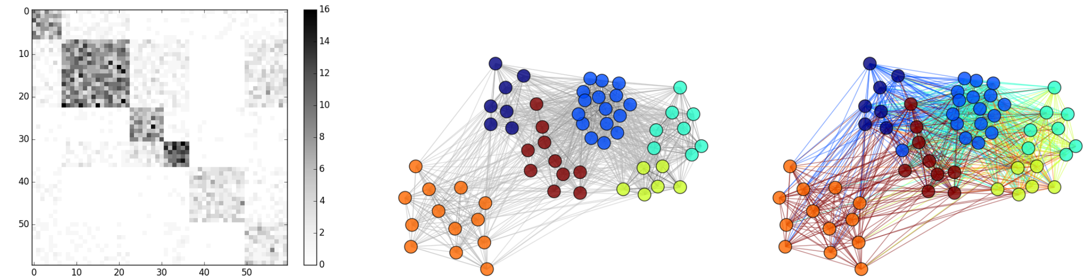
## 5. Experimental Evaluation

We begin by demonstrating the ability of the MDND to recover latent network structure in a two synthetically-generated data sets, before performing a quantitative analysis on three real-world networks.

### 5.1 Synthetic Data

In Figure 3, we show the performance of the MDND evaluated on a 60-node network generated according to a stochastic blockmodel. A distribution over cluster memberships was drawn from a Dirichlet$(5, 5, 5, 5, 5, 5)$ distribution. Intra-cluster link parameters $\theta_{i,i}$ were distributed according to $\theta_{i,i} \sim \text{Gamma}(5, 1)$, and inter-cluster link parameters $\theta_{i,j}$ were distributed according to $\theta_{i,j} \sim \text{Gamma}(0.5, 1)$. Link counts $z_{sr}$ for each pair $(s, r)$ were Poisson-distributed given the appropriate parameter $\theta_{c_s,c_r}$, where $c_s$ is the cluster of

2. In the Chinese franchise interpretation of Teh et al. (2006), $\rho^{(1)}_{k,i}$ and $\rho^{(2)}_{k,i}$ correspond to the number of tables in a given restaurant serving dish $i$.




(a) 60-node network generated according to a stochastic blockmodel; color indicates number of links.

(b) 60-node network generated according to a stochastic blockmodel; nodes are colored according to ground-truth labels.

(c) Clusterings recovered using MDND. Edges are colored according to their most common label. Nodes are colored according to the most common label out of their incoming and outgoing edges.

Figure 3: Structure recovery: stochastic blockmodel.

node $s$. The raw matrix of link counts is shown in Figure 3a, and a visualization of the matrix is shown in Figure 3b, where the edge color indicates the number of nodes and the node color indicates the ground-truth cluster labels.[3]

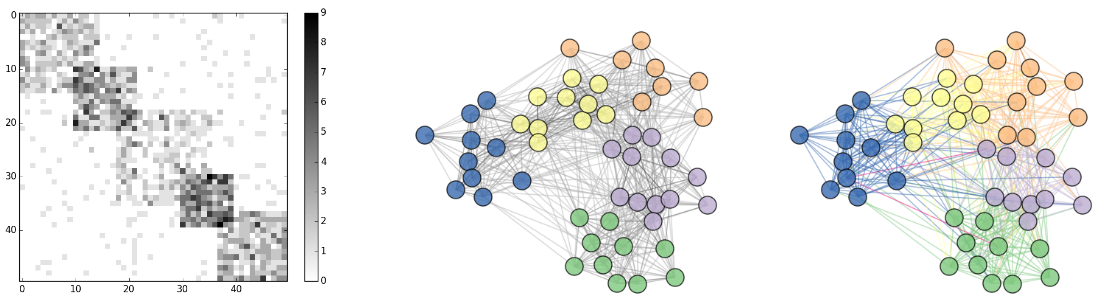
Figure 3c shows the structure recovered by the MDND. Each edge in the MDND representation has multiple cluster labels, one per link; the edges are colored according to their most common cluster label. The nodes are colored according to the most common cluster label amongst their incoming and outgoing edges. As we can see from Figure 3c, all but one node is most commonly associated with its ground-truth label.

Figure 4 evaluates the MDND on a 50-node with manually constructed overlapping blocks. The network shown in Figure 4a was generated from 5 equiprobable clusters; each cluster puts 95% of its probability mass on one of five overlapping blocks, and the remaining mass is uniformly distributed over the entire population. Figure 4b shows a visualization of this matrix, where the edge color indicates the number of nodes and the node color indicates most common ground-truth label amongst the incoming and outgoing edges, as in Figure 3b. Figure 4c shows the structure recovered by the MDND; as before, the edges are colored according to their most common cluster label, and the nodes are colored according to the most common cluster label amongst their incoming and outgoing edges. Again, we have very high agreement between the ground-truth labels and the recovered clustering.

### 5.2 Real Network Data

We compared the mixture of Dirichlet network distributions to the infinite relational model, the mixed membership stochastic blockmodel, and several baselines, in three real-world scenarios: A small network representing character interactions in Shakespeare's 'Macbeth';

3. The visualization layouts in this section (Figures 3b, 3c, 4b and 4c) were generated using the Python package NetworkX with a Fruchterman-Reingold force-directed algorithm.




(a) 50-node network generated from manually constructed overlapping blocks; color indicates number of links.

(b) 50-node network generated from manually constructed overlapping blocks; nodes are colored according to ground-truth labels.

(c) Clusterings recovered using MDND. Edges are colored according to their most common label. Nodes are colored according to the most common label out of their incoming and outgoing edges.

Figure 4: Structure recovery: Overlapping blocks.

a medium-sized network representing political interactions; and a large network representing email interactions. In Section 5.2.1 we describe the comparison methods, and in Section 5.2.2 we describe the three networks and present our results.

#### 5.2.1 Comparison Methods

We compare the mixture of Dirichlet network distributions to a single symmetric Dirichlet network distribution; to integer-valued variants of the mixed-membership stochastic blockmodel (MMSB, Airoldi et al., 2008) and the infinite relational model (IRM, Kemp et al., 2006); and to two baseline methods.

1. **Symmetric Dirichlet network distribution**. We modeled the data using a single symmetric DND as described in Section 3.1, with Dirichlet process concentration parameter $\tau = 1$.

2. **Infinite relational model.** Kemp et al. (2006) describe a variant of the IRM appropriate for integer-valued data. Each pair of clusters $(i, j)$ is associated with a positive real-valued parameter $\theta_{ij}$, and the $N$ links are assigned to clusters according to a multinomial distribution parametrized by the $\theta_{ij}$. Inference in this model is performed using existing C code released by the authors of Kemp et al. (2006). This code was not able to handle the number of nodes present in the Enron data sets.

3. **Mixed-membership stochastic blockmodel.** While the MMSB is designed for binary-valued networks, it can trivially be extended to integer-valued networks by replacing the Bernoulli distributions with Poisson distributions, and placing gamma priors on the Poisson parameters. While, to the best of our knowledge, this extensions has not yet been explored in the literature, it is a natural, and easily implementable,

extension. We perform inference in this model, with $K = 50$ clusters, using Gibbs sampling, via an existing R package (Chang, 2012) that was modified to replace the beta/Bernoulli pairs with gamma/Poisson pairs. Since inference in this model was significantly slower than inference in the IRM and the DNM, we only compared with the gamma/Poisson MMSB on our smallest data set.

4. **Baseline 1: equiprobable links.** Our first baseline assumed that all (sender, receiver) pairs are equally likely, provided the sender and receiver are different—so if we have $M$ nodes, the probability of a given link is $1/M(M-1)$.

5. **Baseline 2: Dirichlet-multinomial distribution over links**. Here we assumed an $M(M-1)$-dimensional Dirichlet prior over the potential links, and looked at the conditional distribution given the $N$ observed links,

$$P((s_{N+1}, r_{N+1}) = (i,j)) = \frac{\sum_{n=1}^{N} \mathbb{I}(s_n = i, r_n = j) + \alpha_{ij}}{M(M-1)+N}.$$

   We note that this corresponds to an IRM where each pair of nodes is in its own cluster.

#### 5.2.2 Evaluation on Three Real Networks

We compared the mixture of Dirichlet network distributions to the comparison methods described above, on three real data sets. We describe the data sets below; Table 1 summarizes the networks' statistics.

1. **Macbeth**. This data set represents the implied social network in Shakespeare's 'Macbeth'. We constructed a directed network where each link indicates an uninterrupted block of speech from the speaker to all other characters presently on stage. To evaluate predictive performance, we split the play into 5 contiguous subsets, each containing (approximately) $N/5$ consecutive edges. We used these subsets to generate a 5-fold split into training data (4 of the 5 subsets) and a test set (the remaining one subset), and evaluated the joint predictive likelihood on the test set.

2. **Militarized disputes**. Next, we evaluated our model on a network representing militarized disputes between 188 countries (Maoz, 2005). The data set contains 8650 disputes between 1816 and 1976. We split this data set into 10 subsets, and used these subsets to generate 10 train/test splits, where 9 subsets were used for training and 1 for evaluation.

3. **ENRON email network**. Finally, we looked at subsets of the ENRON email data set (Klimmt and Yang, 2004). We looked at five training sets, corresponding to the total set of emails sent and received in each of the first 5 months of 2000. For each data set, we evaluated predictive performance on the first 1000 emails sent in the subsequent month.

Table 2 shows the log predictive likelihoods obtained on the above data sets. In each case, we condition on the latent structure obtained on the training set, and consider the joint conditional distribution over the test set. For the MMSB and the IRM, our training

| Network | Number of links ($N$) | Number of nodes ($M$) | $N/M^2$ |
|---|---|---|---|
| Macbeth | 2153 | 39 | 1.42 |
| Military Disputes | 8650 | 188 | 0.245 |
| Enron (average) | 13994 | 3883.8 | 9.14e-4 |
| Enron (range) | 8692–20464 | 3006–4652 | 7.55e-4–1.03e-3 |

Table 1: Network statistics. Note that the Macbeth and Military Disputes data sets were split into equally sized subsets, while the Enron data set was split according to month. We therefore report both the average and the range of the Enron monthly statistics.

| | Macbeth | Military Disputes | ENRON |
|---|---|---|---|
| Symmetric DND | $-2900.56 \pm 193.06$ | $-5019.35 \pm 38.70$ | $-15030.77 \pm 208.71$ |
| **MDND** | $\mathbf{-1769.74 \pm 71.60}$ | $\mathbf{-5000.74 \pm 54.64}$ | $\mathbf{-9053.68 \pm 235.02}$ |
| IRM | $-1941.50 \pm 102.69$ | $-6984.25 \pm 16.49$ | - |
| MMSB | $-3077.22 \pm 65.42$ | - | - |
| Baseline 1 | $-2723.33 \pm 0$ | $-5748.78 \pm 40.59$ | $-16509.75 \pm 125.47$ |
| Baseline 2 | $-2462.94 \pm 70.05$ | $-5433.24 \pm 38.76$ | $-15800.01 \pm 138.84$ |

Table 2: Test set log likelihood (mean $\pm$ standard error).

set explicitly included those nodes with interactions present in the test set but not in the training set. We note that this is an unrealistic setting that gives an advantage to the MMSB and the IRM: In most situations, the number of new nodes is unknown. However, without including at least the correct number of unseen nodes, we would be unable to obtain an estimate for the predictive performance on the data sets used.

For the IRM and the MMSB, we already have cluster assignments for each pair of nodes, and can use the cluster parameters to directly obtain a probability distribution over pairs of nodes. We use this probability distribution to directly calculate the joint predictive log likelihood. For the MDND, we do not yet have cluster assignments for the test set links. It is analytically intractable to sum over all possible test set cluster assignments. Instead, we estimate the predictive log likelihood using our Gibbs sampler to generate 100 samples of the test cluster assignments, conditioned on the training set assignments and the test set links. We then use the harmonic mean of the likelihoods given these cluster assignments as an estimator for the overall joint predictive likelihoods.

The MMSB code was unable to run on the Disputes and ENRON data sets, and the IRM code was unable to run on the ENRON data set. By looking at the ratio $M/N^2$ in Table 1, we see that the Disputes data set is much sparser than Macbeth, and the ENRON data set is much sparser than Macbeth. This means that the blockmodel-based approaches, which scale quadraticaly with the number of nodes, are unable to run. We note that our experiments were based on of existing implementations of the MMSB and IRM; while different implementations may be able to process the entire data set, it is likely to be much slower than the MDND, due to the density of the network.

We see that, in each case, the MDND out-performs the comparison methods—even though we have included more information in the MMSB and the IRM by making use of the number of unseen nodes.

## 6. Discussion and Future Work

We have presented a new Bayesian nonparametric model, the mixture of Dirichlet network distributions, for integer-valued networks where the number of nodes is unbounded and grows in expectation with the number of binary links. This model allows us to capture sparse networks with latent structure. Existing network models focus either on latent structure—capturing the fact that each node will have a different pattern over which nodes it connects with—or on capturing sparsity; this is, to our knowledge, the first model that combines these two goals. Further, unlike most existing Bayesian network models, this model is explicitly designed for prediction. We can use the mixture of Dirichlet network distributions to obtain an explicit predictive distribution over the nodes associated with an as-yet unseen observation, even if we have not observed these nodes in our training set; we have shown good predictive and qualitative performance on a variety of data sets.

The mixture of Dirichlet network distributions is based on a simpler network model that we refer to as a Dirichlet network distribution. In the symmetric setting—where a common distribution is used for both senders and receivers—this corresponds to a special case of the integer-valued network models of Caron and Fox (2015) and Crane and Dempsey (2016). While these models can be used to obtain desirable properties such as network sparsity and power law degree distribution, they are unable to capture community-type structure in the network. By using a mixture of these networks, we can capture multiple modalities of interaction between nodes; by using a nonparametric hierarchical framework we ensure that both the number of nodes is unbounded, and that nodes can interact as part of multiple clusters. The MDND therefore increases the modeling flexibility of this class of models, while retaining desirable sparsity properties.

The mixture of Dirichlet network distributions is an exchangeable model: It is invariant to permutations of the order in which we observe links. While this is computationally appealing and leads to a straightforward predictive distribution, it does not allow us to capture network dynamics in integer-valued networks. In practice, such dynamics may be important: an individual's level of activity within a topic may vary over time, and the overall popularity of topics may change. A number of authors have found that adding temporal dynamics to network models improves performance (Ishiguro et al., 2010; Xing et al., 2010; Xu and Hero III, 2013). In the case of Dirichlet network distributions, similar temporal dynamics could be incorporated by replacing some or all of the component Dirichlet processes with *dependent* Dirichlet processes (MacEachern, 2000; Lin et al., 2010; Ren et al., 2008); we intend to explore this in a future work.

In addition to the base model for integer-valued networks, we also discussed extensions to binary networks. The methods considered involve truncating the exchangeable integer-valued network; while it is possible to obtain exchangeable binary networks related to those considered by Caron and Fox (2015) and Veitch and Roy (2015), we argued that a dynamic truncation, yielding a non-exchangeable model, is more appropriate for a temporally expanding binary network. Unfortunately, inference in such truncated models is trickier than the integer-valued case. While we can analytically sample the censored observations as auxiliary variables and recover the integer network, the number of censored observations grows according to a coupon-collector problem with the number of observed links, making this approach infeasible for large data sets. An interesting avenue for future research is to develop scalable inference methods for this setting.

## Acknowledgments

Sinead Williamson is supported by NSF grant 1447721.

## References

E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

A.A. Amini, A. Chen, P.J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

A. Brix. Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31(4):929–953, 1999.

F. Caron and E.B. Fox. Sparse graphs using exchangeable random measures. arXiv:1401.1137 [stat.ME], 2015.

J. Chang. *lda: Collapsed Gibbs sampling methods for topic models.*, 2012. URL `http://CRAN.R-project.org/package=lda`. R package version 1.3.2.

H. Crane and W. Dempsey. Edge exchangeable models for network data. arXiv:1603.04571 [math.ST], 2016.

E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky. The sticky HDP-HMM: Bayesian nonparametric hidden Markov models with persistent states. Technical Report P-2777, Massachusetts Institute of Technology, 2007.

Q. Ho, J. Yin, and E.P. Xing. On triangular versus edge representations—towards scalable modeling of networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2132–2140, 2012.

P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

K. Ishiguro, T. Iwata, N. Ueda, and J.B. Tenenbaum. Dynamic infinite relational model for time-varying relational data analysis. In *Advances in Neural Information Processing Systems (NIPS)*, pages 919–927, 2010.

S. Jain and R.M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182, 2004.

C. Kemp, J.B. Tenenbaum, T.L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *National Conference on Artificial Intelligence (AAAI)*, pages 381–388, 2006.

B. Klimmt and Y. Yang. Introducing the Enron corpus. In *Conference on Email and Anti-Spam (CEAS)*, 2004.

A. Lijoi, I. Prünster, and S.G. Walker. Investigating nonparametric priors with Gibbs structure. *Statistica Sinica*, 18(4):1653–1668, 2008.

D. Lin, E. Grimson, and J.W. Fisher III. Construction of dependent Dirichlet processes based on Poisson processes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1396–1404, 2010.

J. Lloyd, P. Orbanz, Z. Ghahramani, and D.M. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1007–1015, 2012.

S.N. MacEachern. Dependent Dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*, 2000.

Z. Maoz. Dyadic militarized interstate dispute dataset, version 2.0. http://psfaculty.ucdavis.edu/zmaoz/dyadmid.html, 2005.

M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: A variational approach. *Annals of Applied Statistics*, 4(2):715–742, 2010.

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. Technical Report 9815, Dept. of Statistics, University of Toronto, 1998.

P. Orbanz and D. Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 437–461, 2014.

J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

L. Ren, D.B. Dunson, and L. Carin. The dynamic hierarchical Dirichlet process. In *International Conference on Machine Learning (ICML)*, pages 824–831, 2008.

T.A.B. Snijders and T. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

United Nations, Department of Economic and Social Affairs, Population Division. Trends in international migrant stock: The 2015 revision. United Nations database, POP/DB/MIG/Stock/Rev.2015, 2015.

V. Veitch and D. Roy. The class of random graphs arising from exchangeable random measures. arXiv:1512.03099 [math.ST], 2015.

Y.J. Wang and G.Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

S.A. Williamson, S.N. MacEachern, and E.P. Xing. Restricting nonparametric distributions. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2598–2606, 2013.

E.P. Xing, W. Fu, and L. Song. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 4(2):535–566, 2010.

K.S. Xu and A.O. Hero III. Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 201–210. Springer, 2013.