Bayesian inference on random simple graphs with power law degree distributions

Juho Lee [1] **Creighton Heaukulani** [2] **Zoubin Ghahramani** [2 3] **Lancelot F. James** [4] **Seungjin Choi** [1]

Abstract

We present a model for random simple graphs with power law (i.e., heavy-tailed) degree distributions. To attain this behavior, the edge probabilities in the graph are constructed from Bertoin–Fujita–Roynette–Yor (BFRY) random variables, which have been recently utilized in Bayesian statistics for the construction of power law models in several applications. Our construction readily extends to capture the structure of latent factors, similarly to stochastic blockmodels, while maintaining its power law degree distribution. The BFRY random variables are well approximated by gamma random variables in a variational Bayesian inference routine, which we apply to several network datasets for which power law degree distributions are a natural assumption. By learning the parameters of the BFRY distribution via probabilistic inference, we are able to automatically select the appropriate power law behavior from the data. In order to further scale our inference procedure, we adopt stochastic gradient ascent routines where the gradients are computed on minibatches (i.e., subsets) of the edges in the graph.

1. Introduction

In statistical applications, random graphs serve as Bayesian models for network data, that is, data consisting of objects and the observed linkages between them. Here we will focus on models for random *simple graphs* (that is, graphs with edges that take binary values), which are appropriate for applications where we observe either the presence or absence of links between objects in the network. For example, in social networks, nodes may represent individuals and a link (i.e., a nonzero value of an edge) could represent friendship. In a protein-protein interaction network, nodes may represent proteins and links could represent an observed physical or chemical interaction between proteins. Many domains involving network data (including social and protein-protein interaction networks) have been shown to exhibit power law, i.e., heavy-tailed, degree distributions (Barabási & Albert, 1999). Models for random graphs with power law degree distributions, also called *scale-free* random graphs, have therefore become one of the most actively studied areas of graph theory and network science (Bollobás et al., 2001; Albert & Barabási, 2002; Dorogovtsev & Mendes, 2002). In this paper we present a model for simple, scale-free random graphs, which we apply as a probabilistic model for several network datasets.

The model we present in this paper is a special case of the *generalized random graph* defined by Britton et al. (2006), and studied further by van der Hofstad (2016, Ch. 6), which outlines a framework for defining scale-free random graphs, but does not provide practical constructions, much less algorithms for performing statistical inference on the model components given data. Here we provide one such practical construction, along with a *variational inference* routine (Jordan et al., 1999) for efficient posterior inference. What's more, our construction readily generalizes to include the structure of latent factors/clusters, as captured by the popular *stochastic blockmodels* (Nowicki & Snijders, 2001; Airoldi et al., 2009), while maintaining power law behavior in the graph.

Applying Bayesian inference algorithms on network datasets is a challenge because likelihood computations, in general, scale with the number of edges in the graph, which is $O(n^2)$ in a network with n nodes. To help overcome these difficulties, we follow Hoffman et al. (2013) and develop a *stochastic variational inference* algorithm in which we approximate many likelihood computations on only subsets of the data, called *minibatches*. In the case of a network dataset, the minibatches are comprised of subsets of edges in the graph.

We apply this inference procedure to several network

[1]Pohang University of Science and Technology, Pohang, South Korea [2]University of Cambridge, Cambridge, UK [3]Uber AI Labs, San Francisco, CA, USA [4]Hong Kong University of Science and Technology, Hong Kong. Correspondence to: Juho Lee <stonecold@postech.ac.kr>, Seungjin Choi <seungjin@postech.ac.kr>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

datasets that are commonly observed to possess power law structure. Our experiments show that accurately capturing this power law structure improves performance on tasks predicting missing edges in the networks.

2. Bayesian models for simple graphs

We represent a simple graph with n nodes by an adjacency matrix $X := (X_{i,j})_{i,j\leq n}$, where $X_{i,j} = 1$ if there is a link between nodes i and j and $X_{i,j} = 0$ otherwise. Here we will only consider undirected graphs, in which case X represents a symmetric matrix. Furthermore, we do not allow self links, so the diagonal entries in X are meaningless. Most probabilistic models for simple graphs take the entries in X to be conditionally independent Bernoulli random variables; in particular, for every $i, j \leq n$, let $p_{i,j}$ be the (random) probability of a link between nodes i and j, and let $X_{i,j} \mid p_{i,j} \sim \text{Bernoulli}(p_{i,j})$. For every simple graph $x := (x_{i,j})_{i,j\leq n}$, we may then write the likelihood for the parameters $p := (p_{i,j})_{i,j\geq 1}$ given X as

$$P(X = x \mid p) = \prod_{i<j\leq n} p_{i,j}^{x_{i,j}}(1 - p_{i,j})^{1-x_{i,j}}, \qquad (1)$$

where in our case it should be clear that the product is only over $i, j \leq n$ such that $i < j$ and $i \neq j$. Random simple graphs date back to the Erdös–Rényi model, which may be reviewed, along with the more general theory of random graphs, in the text by Bollobás (1998). A random graph is called scale-free when the fraction of nodes in the network having k connections to other nodes behaves like $k^{-\tau}$ for large values of k and some exponent $\tau > 1$. More precisely, let $D_{n,i} := \sum_{j\neq i} X_{i,j}$ denote the (random) degree of node i, for every $i \leq n$. Then X is (asymptotically) scale-free when, for every node $i \leq n$,

$$\mathbb{P}\{D_{n,i} = k\} \sim ck^{-\tau}, \quad \text{as } n \to \infty, \qquad (2)$$

for some constant $c > 0$, a power law exponent $\tau > 1$, and k sufficiently large. Here the notation $A \sim B$ denotes that the ratio $A/B \to 1$ in the specified limit.

In order to model scale-free random graphs, Britton et al. (2006) suggested reparameterizing the model in Eq. (1) by a sequence of odds ratios $r_{i,j} := p_{i,j}/(1 - p_{i,j})$, for every $i < j \leq n$, which factorize as $r_{i,j} = U_iU_j$, for some $U := (U_1, \ldots, U_n)$. The node-specific factors U_i are then modeled as $U_i := W_i/\sqrt{L}$ for some sequence of nonnegative random variables $W := (W_1, \ldots, W_n)$ and where $L := \sum_{i=1}^n W_i$. In a series of results, (Britton et al., 2006, Thms. 3.1 & 3.2) and (van der Hofstad, 2016, Cor. 6.11 & Thm. 6.13) assert conditions on the random variables W so that the limiting distribution of the degrees $D_{n,i}$ is a mixed Poisson distribution. We will further detail these previous results in Section 4.

The distribution of W_i is interpreted here as a prior distribution for the degree $D_{n,i}$ of node i, and if its distribution has heavy tails, then so will the distribution of $D_{n,i}$. Conversely, if the distribution of W_i does not have heavy tails, then neither will the distribution of the degrees $D_{n,i}$. We explore this alternative in Section 7.

Previous authors did not suggest any particular choices for the distribution of W_i, and so we elect to model them with BFRY random variables (Bertoin et al., 2006; Devroye & James, 2014), which have a heavy-tailed distribution and have recently played a role in the construction of several power law models in Bayesian statistics. Other heavy tailed distributions, such as those exhibited by log normal random variables, may also be used to model the W_i, and these options may be explored. One benefit of the BFRY distribution is that the thickness of its tails, and thus the power law behavior of the resulting graph, may be straightforwardly controlled by the discount parameter α.

3. A generalized random graph

Consider the model from the previous section, parameterized by the odds ratios $r := (r_{i,j} : i < j \leq n)$. Define

$$G(r) := \prod_{i<j\leq n}(1 + r_{i,j}) = \prod_{i<j\leq n}(1 + U_iU_j), \qquad (3)$$

and note that the conditional likelihood in Eq. (1) may be rewritten in terms of the degrees $D_{n,i}$ as

$$P(X = x \mid r) = G(r)^{-1} \prod_{i<j\leq n}(U_iU_j)^{x_{i,j}} \qquad (4)$$

$$= G(r)^{-1} \prod_{i\leq n} U_i^{D_{n,i}}. \qquad (5)$$

The random simple graph X is called a *generalized random graph*, and we will henceforth write $X \mid r \sim \text{GRG}(n, r)$.

Let $\alpha \in (0, 1)$, which we call the *discount parameter*, and let $C_1, C_2, \ldots$ be a sequence of positive values satisfying

$$\lim_{n\to\infty} C_n = \infty \quad \text{and} \quad \lim_{n\to\infty} C_n^\alpha/n = 0. \qquad (6)$$

Let the weights $W_1, \ldots, W_n$ be i.i.d. with density

$$f_n(w) \propto w^{-\alpha-1}(1 - e^{-w})\mathbb{1}_{\{0\leq w\leq C_n\}}. \qquad (7)$$

(These are truncated BFRY random variables and will be discussed, along with a method for simulation, in Section 3.1.) Then the corresponding generalized random graph has an (asymptotic) power law degree distribution with power law exponent $\tau = 1 + \alpha$. We summarize this construction in the following theorem:

Theorem 3.1. *For every n, let $W_1, \ldots, W_n$ be i.i.d. with density f_n and let $(D_{n,i})_{i\leq n}$ be the degrees of the generalized random graph $X \mid r \sim \text{GRG}(n, r)$, where*

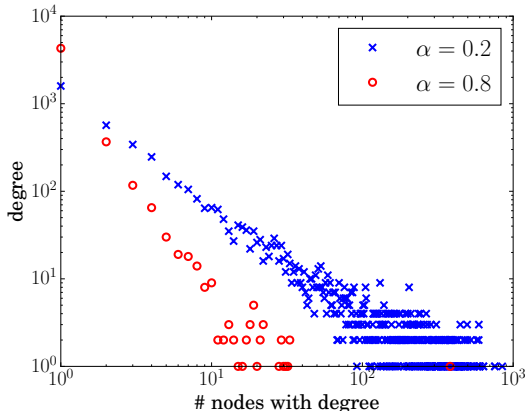

Figure 1. The number of nodes with various degrees for two simulated graphs with $n = 3000$ nodes and differing values for α.

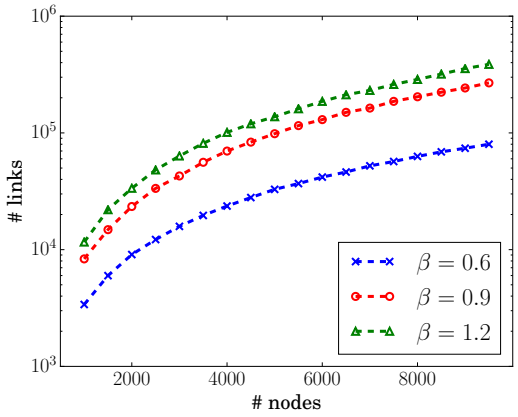

Figure 2. The average number of links in simulated graphs with varying sparsity parameter β.

$r := (r_{i,j})_{i<j\leq n}$ *is the sequence of odds ratios*

$$r_{i,j} = W_i W_j / L, \qquad i < j \leq n, \tag{8}$$

and $L := \sum_i W_i$. *Then the following hold:*

1. *For* $y \gg 1$, $\mathbb{P}\{D_{n,i} = y\} \sim cy^{-1-\alpha}$, *for every node* i *and for some constant* c, *as* $n \to \infty$.
2. *For any* m, *the collection* $D_{n,1}, \ldots, D_{n,m}$ *are asymptotically independent, as* $n \to \infty$.

This construction is closely related to the model described by van der Hofstad (2016, Thm. 6.13), and the proof of Theorem 3.1, which is provided in the supplementary material, follows analogously to the results by Britton et al. (2006, Thms. 3.1 & 3.2). Note that the power law exponent $\tau = 1 + \alpha$ of the graph (as described by Eq. (2)) is determined by the parameter $\alpha \in (0, 1)$, and takes values in $(1, 2)$. While power law exponents in $(2, 3)$ has often been suggested in the past, it has more recently been shown that exponents within the $(1, 2)$ range of our model is more appropriate in many domains (van der Hofstad, 2016, Ch. 1); (Crane & Dempsey, 2015).

3.1. Truncated BFRY random variables

A random variable W with density function f_n given by Eq. (7) is a ratio of gamma and beta random variables, upper truncated at C_n. In particular let

$$g \sim \text{gamma}(1 - \alpha, 1) \quad \text{and} \quad b \sim \text{beta}(\alpha, 1), \tag{9}$$

be independent, then the ratio $Z := g/b$ has density $p(z) \propto z^{-\alpha-1}(1 - e^{-z})$ on $(0, \infty)$ (by construction), which is known as the Bertoin-Fujita-Roynette-Yor (BFRY) distribution (Bertoin et al., 2006; Devroye & James, 2014) and has been used in the construction of power law models in some recent applications in machine learning (James et al., 2015; Lee et al., 2016). The random variable W is then obtained by upper truncating the random variable Z at C_n. By our requirements on the sequence C_n (c.f. Eq. (6)), the density function f_n of W approaches the density function of the BFRY random variable Z as $n \to \infty$, that is,

$$\lim_{n\to\infty} f_n(w) = \frac{\alpha}{\Gamma(1-\alpha)} w^{-\alpha-1}(1 - e^{-w}), \tag{10}$$

which is heavy-tailed with infinite moments. It is straightforward to simulate these truncated BFRY random variables by repeatedly simulating g and b as in Eq. (9), accepting $W := g/b$ as a sample when $W < C_n$.

The truncation of W at C_n produces a random variable with finite mean (for $n < \infty$), which is essential when constructing the generalized random graph and motivates the construction by van der Hofstad (2016, Thm. 6.13) alluded to earlier; see Section 4. For simplicity, one could take $C_n = n$, but the flexibility to set this parameter allows us to control other properties of the model. For example, in the next section we show how to vary this truncation level to control the sparsity of the graph.

3.2. Controlling power law and sparsity in the graph

The discount parameter $\alpha \in (0, 1)$ controls the power law behavior of the graph, where decreasing α results in heavier tails in the degree distribution of the nodes in the graph. We can visualize this behavior by simulating graphs at different values of α. In Section 3, we set $C_n = n$ and show the number of nodes of varying degrees in two simulated graphs, one with $\alpha = 0.2$ and one with $\alpha = 0.8$.

The degree distribution of the nodes in a graph of course affects the sparsity of the graph; to characterize this relationship, we can upper bound the expected number of links in the graph as follows:

Theorem 3.2. *Let* E_n *be the number of positive edges in the graph. Then* $\mathbb{E}[E_n] = O(nC_n^{1-\alpha})$.

The derivation of this result is provided in the supplementary material. While varying α can thus control the sparsity of the graph in addition to the power law behavior, we often want to decouple these behaviors, in which case we could parameterize the truncation level as $C_n = n^\beta$, for some *sparsity parameter* $\beta > 0$. Note the restriction $\alpha < \min\{1, 1/\beta\}$ must be enforced in order to ensure that the conditions in Eq. (6) are satisfied. In this case, the bound in Theorem 3.2 becomes $\mathbb{E}[E_n] = O(n^{1+\beta(1-\alpha)})$. The interpretation here is that increasing the upper bound C_n increases the likelihood that any particular node will link to others, but does not affect the (asymptotic) power law characterized by Theorem 3.1. In Section 3.2, we display the average number of positive edges in graphs that were simulated with fixed $\alpha = 0.3$ and varying values of the sparsity parameter β. We note that in simulations, we encountered numerical issues in $\beta > 1.4$ regimes.

4. Related work

Referring to the construction for generalized random graphs in Section 2, Britton et al. (2006, Thm. 3.1) shows that when the weights W_i have finite first and second moments, then the limiting distribution of the degree $D_{n,i}$ is a mixed Poisson distribution. Most such distributions are light-tailed, however, in which case the degrees will not exhibit power law behavior. Britton et al. (2006, Thm. 3.2) therefore provides an alternative construction in which W_i may have infinite moments (so that it may exhibit a heavy tail), which results in a graph with a power law exponent of $\tau = 2$. Finally, van der Hofstad (2016, Thm. 6.13) suggests yet another construction where the W_i are upper truncated to be of order $o(n)$, where n is the number of nodes in the graph. The resulting random variables therefore have finite moments, yet exhibit a heavy tail, and the resulting random graph has a heavy tailed degree distribution with an arbitrary power law exponent. None of these results suggest a particular choice for the distribution of W_i, however, and so we have elected to use BFRY random variables (which are heavy tailed) that are upper truncated (so that they have finite moments). We note that the requirements on our truncation level (c.f. Eq. (6)) is less strict than the $o(n)$ criterion of the van der Hofstad (2016, Thm. 6.13) construction.

The reader may consult the surveys by Bollobás & Riordan (2003); Albert & Barabási (2002); Dorogovtsev & Mendes (2002) for a background on scale-free random graphs, which is too large to review here. While these models are numerous, the following recent pieces of work in the Bayesian statistics and machine learning communities may be of interest to the reader: Caron & Fox (2014); Veitch & Roy (2015); Crane & Dempsey (2016); Cai & Broderick (2015). This collection of work discusses power law degree distributions, albeit in some cases in multi-graphs (i.e., graphs with nonnegative integer-valued edges) and in some cases the power law behavior is not characterized, only numerically observed in simulations. Many of these models can be seen to invoke their power law properties from the *Pitman–Yor process* (Pitman & Yor, 1997) (or related stochastic processes), where the extent of this behavior is controlled by the discount parameter $\alpha \in (0, 1)$ of the Pitman–Yor model, which, like the BFRY distribution, is related to a stable subordinator of index α.

5. Incorporating latent factors

Latent factor models for relational data assume that a set of latent clusters underlie the network. For example, in a social network, the latent factors could be the unobserved hobbies or interests of individuals, which determine the observed friendships in the network. Bayesian models for latent factors in relational data are widespread, with some of the most popular based on *stochastic blockmodels*, where models for unsupervised learning, or clustering, are used to infer the latent factors (Nowicki & Snijders, 2001; Kemp et al., 2006; Airoldi et al., 2009; Miller et al., 2009). In this section, we present extensions of the generalized random graph that incorporate latent factors by scaling the odds ratios, while maintaining their power law degree distribution.

We will first provide a general result showing how to incorporate random scaling variables into the model, followed by specific examples that model these scaling variables with latent clusters. Let the odds ratios in the generalized random graph be given by $r_{i,j} = A_{i,j} U_i U_j$ for some $A_{i,j} \geq 0$. Note that $p_{i,j} \to 1$ as $A_{i,j} \to \infty$ and $p_{i,j} \to 0$ as $A_{i,j} \to 0$, and so the edge-specific weight $A_{i,j}$ simply scales the link probability. The random graph $X \mid r \sim \mathrm{GRG}(n, r)$ then has the likelihood

$$P(X = x \mid r) = G(r)^{-1} \prod_{i<j\leq n} A_{i,j}^{x_{i,j}} \prod_{i\leq n} U_i^{D_{n,i}}, \quad (11)$$

where the normalization term $G(r)$ in Eq. (3) is now

$$G(r) := \prod_{i<j\leq n} (1 + A_{i,j} U_i U_j) \quad (12)$$

$$= \sum_x \prod_{i<j\leq n} A_{i,j}^{x_{i,j}} \prod_{i\leq n} U_i^{D_{n,i}}, \quad (13)$$

where the final equality follows simply because $\sum_x P(X = x \mid r) = 1$. So constructed, the odds ratios r will influence the link probabilities in the generalized random graph, but will not affect the power law behavior of the degree distributions (under some assumptions on the random variables $A_{i,j}$). We summarize this construction in the following theorem, the proof for which is provided in the supplementary material:

Theorem 5.1. *Let $(W_i)_{i\le n}$ be i.i.d. random variables with density function $f_n(w)$ (in Eq. (7)). Let $(A_{i,j})_{i<j\le n}$ be a collection of uniformly bounded random variables, where, for every $i \le n$, the collection $(A_{i,j})_{j>i}$ is exchangeable. Let $(D_{n,i})_{i\le n}$ be the degrees of the random graph $X \mid r \sim \mathrm{GRG}(n, r)$, where $r := (r_{i,j})_{i<j\le n}$ is the sequence of odds ratios*

$$r_{i,j} = A_{i,j} W_i W_j / L, \qquad i < j \le n, \tag{14}$$

and where $L := \sum_i W_i$. Then the degrees $(D_{n,i})_{i\le n}$ satisfy statements (1) and (2) in Theorem 3.1.

For example, we may construct *stochastic blockmodels*, such as those introduced by Nowicki & Snijders (2001), as follows: For every $i \le n$, let Z_i be a random variable taking values in $\{1, \dots, K\}$, indicating which one (and only one) of K different factors to associate with node i. We want the latent cluster assignments for two nodes i and j to influence their link probability, which we could capture with a set of parameters $\theta_{k,\ell}$, for $k, \ell = 1, \dots, K$. Then the parameter θ_{Z_i,Z_j} could represent, or influence, the probability of a link between nodes i and j. Taking a Bayesian approach, the indicator variables Z_i may be modeled with a Dirichlet-categorical conjugate distribution and their values may be inferred via probabilistic inference. An example of such a model could be summarized as follows: Let

$$\begin{aligned} Z_i &\sim \mathrm{categorical}(\pi), && i \le n, && (15)\\ \pi &\sim \mathrm{Dirichlet}(c/K), && \text{where } c > 0, && (16)\\ \theta_{\ell,k} &\sim \mathrm{gamma}(a_\theta, b_\theta), && \ell, k \le K, && (17)\\ A_{i,j} &= \theta_{Z_i,Z_j}, && i < j \le n, && (18) \end{aligned}$$

and construct the random graph X as in Theorem 5.1. Kemp et al. (2006) developed a nonparametric extension of a similar model that in a sense takes the limit $K \to \infty$, allowing an appropriate number of clusters to be automatically inferred from the data. In this case, the marginal law of the indicator variables $Z_1, \dots, Z_n$ is given by a Chinese restaurant process (with concentration parameter c).

Several generalizations of the stochastic blockmodel allow the clusters underlying the network to overlap, leading to *mixed membership stochastic blockmodels* (Airoldi et al., 2009) or the related *latent feature relational models* (Miller et al., 2009). To capture this structure, we may generalize the indicators Z_i to now represent a binary K-vector with entry $Z_{i,k} = 1$ indicating node i is associated with cluster k, now called a *feature*, and $Z_{i,k} = 0$ otherwise. One example of such a model could be summarized as follows:

$$\begin{aligned} Z_{i,k} &\sim \mathrm{Bernoulli}(p_k), && i \le n, k \le K, && (19)\\ p_k &\sim \mathrm{beta}(c, c\gamma/K), && k \le K, \text{and } c, \gamma > 0, && (20)\\ \theta_{\ell,k} &\sim \mathrm{gamma}(a_\theta, b_\theta), && \ell, k = 1, 2, \dots, && (21)\\ A_{i,j} &= \sum_{k,\ell} \theta_{k,\ell} Z_{i,k} Z_{j,\ell}, && i < j \le n, && (22) \end{aligned}$$

and construct the random graph X as in Theorem 5.1. Miller et al. (2009) derived a nonparametric extension of this model that in a sense takes the limit $K \to \infty$, in which case the marginal law of the vectors $Z_1, \dots, Z_n$ is that of an *Indian buffet process* (with mass parameter γ and concentration parameter c) (Ghahramani et al., 2007).

6. Variational inference

We derive a variational Bayesian inference algorithm (Jordan et al., 1999) that approximates the (optimal state of the) posterior distribution of the model components, given a network dataset. We approximate the required gradients in this procedure with stochastic gradient ascent (Bottou, 2010; Hoffman et al., 2013), computed on minibatches (i.e., subsets) of edges in the graph.

6.1. The variational lower bound

In variational inference, we approximate the posterior distribution on the latent variables $W := (W_1, \dots, W_n)$ with a variational distribution $q(W; \theta)$, the parameters θ of which are fit to maximize the following lower bound on the marginal likelihood

$$\log p(X) \ge \mathbb{E}_{q(W;\theta)}\Big[\log \frac{p(X \mid W; \alpha) p(W; \alpha)}{q(W; \theta)}\Big], \tag{23}$$

where $p(X \mid W)$ is the likelihood function computed as in Eq. (5), and $p(W; \alpha)$ is the prior on W represented by the density function in Eq. (7). The (non random) discount parameter α is inferred by corresponding gradient ascent updates maximizing the likelihood of the model, which is described in Section 6.4.

We specify a mean field variational distribution $q(W; \theta) = \prod_{i=1}^n q(W_i; \theta_i)$. We considered several approximations for the marginals $q(W_i; \theta_i)$ including truncated BFRY and truncated gamma distributions, however, in our experiments we found that the following *rectified gamma distribution* performed well:

$$\begin{aligned} W_i &=_q \min\{W_i', C_n\}, && (24)\\ W_i' &\sim \mathrm{gamma}(\theta_{i,\mathrm{shp}}, \theta_{i,\mathrm{rte}}), && (25) \end{aligned}$$

independently for every $i \le n$, where $\theta_{i,\mathrm{shp}}$ and $\theta_{i,\mathrm{rte}}$ denote the shape and rate parameters of the gamma distribution, respectively, and the notation $=_q$ emphasizes that this formula holds under the variational distribution q.

6.2. Stochastic gradient ascent

We maximize the lower bound on the right hand side of Eq. (23) by stochastic gradient ascent, where on the t-th step of the algorithm, we make the following updates to the

parameters in parallel

$$\theta_i^{(t+1)} \leftarrow \theta_i^{(t)} + \rho_t \nabla_{\theta_i} \mathbb{E}_{q(W;\theta^{(t)})}[\mathcal{L}(X, W; \theta^{(t)})], \quad (26)$$

for $i \leq n$ and some sequence $(\rho_t)_{t\geq 1}$ of positive numbers satisfying the Robbins–Monro criterion (Robbins & Monro, 1951) $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$, and where

$$\mathcal{L}(X, W; \theta) := \log p(X, W; \alpha) - \log q(W; \theta) \quad (27)$$

$$= \sum_{(i,j)\in\mathcal{E}} \log p(X_{i,j} \mid W) + \sum_{i=1}^{n} \log p(W_i; \alpha) - \sum_{i=1}^{n} \log q(W_i; \theta_i), \quad (28)$$

where $\mathcal{E}$ denotes the observed edges (both links and non-links) in the dataset. We cannot evaluate the expectation (with respect to the rectified gamma distributions $q(W; \theta)$) analytically, and so we elect to use a particular Monte Carlo approximation of this gradient detailed by Knowles (2015), which was developed for gamma variational distributions and easily applies to the rectified gamma case.

Briefly, for every $i \leq n$, create the collection of S Monte Carlo samples from the variational distribution as follows: Independently for $s \leq S$, let $z_i^{(s)} \sim \text{Uniform}(0, 1)$, and set $W_i^{(s)} = \psi(z_i^{(s)}; \theta_i)$, where $\psi(z; \theta) := \min\{F_\theta^{-1}(z), C_n\}$ and $F_\theta^{-1}(x)$ is the inverse of the cumulative distribution function for a gamma random variable. For convenience, we recall that

$$F_{a,b}(x) = \int_0^x \frac{b^a}{\Gamma(a)} t^{a-1} e^{-bt} \mathrm{d}t. \quad (29)$$

For every $k \leq n$, the gradient with respect to the parameters θ_k is then approximated by

$$\nabla_{\theta_k} \mathbb{E}_{q(W;\theta)}[\mathcal{L}(X, W; \theta)] \approx \frac{1}{S} \sum_s \nabla_{W_k} \mathcal{L}(X, W^{(s)}; \theta) \nabla_{\theta_k} \psi(z_k^{(s)}; \theta_k), \quad (30)$$

where $W^{(s)} := (W_1^{(s)}, \ldots, W_n^{(s)})$. This estimator is unbiased and has low enough variance that often a single sample suffices for the approximation (Salimans & Knowles, 2013; Kingma & Welling, 2014). The gradient of ψ is nonzero only when $\{F_{\theta_k}^{-1}(z_k^{(s)}) < C_n\}$, in which case we may immediately obtain the partial derivative with respect to the rate parameter; in particular, we have

$$\nabla_{\theta_{k,\text{rte}}} \psi(z_k^{(s)}; \theta_k) = \begin{cases} \frac{z_k^{(s)}}{\theta_{k,\text{rte}}}, & \text{if } F_{\theta_k}^{-1}(z_k^{(s)}) < C_n, \\ 0, & \text{otherwise.} \end{cases} \quad (31)$$

The partial derivative with respect to the shape parameter $\nabla_{\theta_{k,\text{shp}}} \psi(z_k^{(s)}; \theta_k)$ does not have a closed form solution and must be approximated. Different approximation routines are suggested by Knowles (2015) for different regimes of the shape parameter $\theta_{k,\text{shp}}$, and we found these approximations to be accurate and efficient in our experiments.

6.3. Minibatches of edges in the graph

Computing the n required gradients in Eq. (26) may be done in parallel, and this computation, whether performed analytically or with automatic differentiation methods, scales with the number of edges in the graph. This can be prohibitive for many network datasets, and we therefore introduce a further approximation where this gradient is evaluated on subsets (a.k.a. *minibatches*) of the dataset, a technique from stochastic gradient ascent (Bottou, 2010) adopted in the context of variational Bayesian inference by Hoffman et al. (2013). In the case of a network dataset, we may select minibatches that are subsets of the observed edges in the graph. In particular, write the gradient of Eq. (28) with respect to the variable W_k (which is required by Eq. (30)) as

$$\nabla_{W_k} \mathcal{L}(W^{(s)}; \theta) = \sum_{(i,j)\in\mathcal{E}} g_{(i,j)}(X, W^{(s)}; k), \quad (32)$$

where $g_{(i,j)}(X, W; k) := \nabla_{W_k}[\log p(X_{i,j} \mid W) + |\mathcal{E}|^{-1} \log p(W; \alpha) - |\mathcal{E}|^{-1} \log q(W; \theta)]$ is the gradient that ignores all but one edge of the graph. We may therefore compute the unbiased estimate of this gradient

$$\nabla_{W_k} \mathcal{L}(W^{(s)}; \theta) \approx \frac{|\mathcal{E}|}{|\mathcal{B}|} \sum_{(i,j)\in\mathcal{B}} g_{(i,j)}(X, W^{(s)}; k), \quad (33)$$

on a minibatch $\mathcal{B} \subseteq \mathcal{E}$ of the observed edges.

6.4. Inference on the parameters α and β

Without good prior knowledge of how to set the discount parameter α and the sparsity parameter β controlling the power law and sparsity behaviors of the graph, respectively, we infer their values from the data. First consider the discount parameter, which we infer with gradient ascent. After every update to the latent variables W, we fix them to their mean under the distribution q, i.e., $\hat{W} := (\hat{W}_1, \ldots, \hat{W}_n)$ where $\hat{W}_i = \mathbb{E}_{q(W_i;\theta_i)}[W_i]$, and take a step in the direction of the gradient

$$\nabla_\alpha \log p(\hat{W}; \alpha) = \sum_{i=1}^{n} \nabla_\alpha \log p(\hat{W}_i; \alpha) \quad (34)$$

$$= \sum_{i=1}^{n} \Big[-\frac{\nabla_\alpha Z_{\alpha,\beta}}{Z_{\alpha,\beta}} - \log(\hat{W}_i) \Big], \quad (35)$$

which is straightforward to derive from the density function in Eq. (7), and where the normalization term

$$Z_{\alpha,\beta} := \int_0^{C_n} w^{-\alpha-1}(1 - e^{-w}) \, \mathrm{d}w \quad (36)$$

is a function of α and β, if we let $C_n = n^\beta$ as suggested in Section 3.2. We do not have a closed form solution for

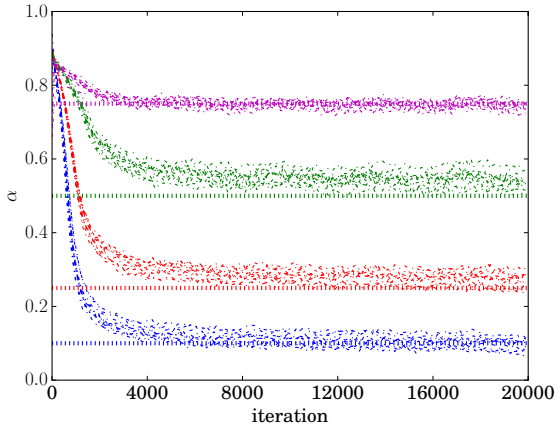

Figure 3. Trace plots of the discount parameter α during 10 different inference runs, each time simulating a dataset from the model with either $\alpha \in \{0.1, 0.3, 0.5, 0.7\}$ and intializing α randomly.

Table 1. Comparision between the BFRY model and the Gamma baseline model when α is known. The test log-likelihoods were averaged over the last 4,000 of 20,000 gradient descent updates.

true α	model	max test log-likel	avg test log-likel
$\alpha = 0.3$	BFRY	**-57323.19 ± 91.62**	**-57675.72 ± 31.71**
	Gamma	-71341.90 ± 116.82	-71841.66 ± 47.38
$\alpha = 0.5$	BFRY	**-21077.62 ± 79.64**	**-21289.75 ± 34.23**
	Gamma	-24430.38 ± 73.06	-24701.06 ± 11.31
$\alpha = 0.7$	BFRY	**-7894.67 ± 41.84**	**-8027.42 ± 51.08**
	Gamma	-8511.48 ± 22.45	-8601.50 ± 15.42

this term when $C_n < \infty$, and, unfortunately, inference on model parameters where the likelihood is difficult to evaluate is a challenging problem; for example, see the approaches taken by Murray et al. (2006) on such problems, which those authors call *doubly intractable distributions*. Accurate inference for α is important in our model, because it controls the power law behavior of the graph. In our experiments, we approximate the gradient in Eq. (35) for (fixed β) by approximating $Z_{\alpha,\beta}$ (via Eq. (36)) and $\nabla_\alpha Z_{\alpha,\beta} = \int_0^{C_n} -w^{-\alpha-1}(1-e^{-w}) \log w \, \mathrm{d}w$, with line integrals. In the Section 7, we demonstrate that this approximation works well in various regimes of α, with slight overestimation for moderate values.

Similar approaches to infer β may be derived with finite difference approximations; we did not find these approaches successful in our experiments, however, and so we instead select β by cross validation.

7. Experiments

We first demonstrate how the inference procedure in Section 6.4 can correctly differentiate between various regimes of α. We ran an experiment where for each value $\alpha \in \{0.1, 0.3, 0.5, 0.7\}$, we simulated 10 datasets from the model with $n = 1,000$ nodes, while fixing $\beta = 1.0$. For each simulated dataset, we ran an instance of the inference routine with α randomly initialized. In Fig. 3, we show the trace plots of alpha during each instance of the inference routine. For comparison, the true values of α are also shown as horizontal dashed lines. We can see that the inference routine can correctly distinguish between these different regimes of α, with slight overestimation in the moderate α regime. Interestingly, despite random initializations of $\alpha \in (0, 1)$, the algorithm always immediately inflates α to around 0.9, and then slowly decreases this value during inference, regardless of what value of α generated the data.

Table 2. Comparision between the BFRY model and the Gamma baseline model on the air traffic, blogs, and social network datasets. The test log-likelihoods were averaged over the last 4,000 of 20,000 gradient descent updates.

dataset	model	max test log-likel	avg test log-likel
500Air	BFRY	**-1628.51 ± 10.46**	**-1654.20 ± 6.79**
	Gamma	-1842.10 ± 3.97	-1870.35 ± 0.28
polblogs	BFRY	**-474.67 ± 32.20**	**-503.20 ± 37.85**
	Gamma	-555.24 ± 18.27	-596.78 ± 0.78
Fb107	BFRY	**-18098.38 ± 20.50**	**-18209.94 ± 12.86**
	Gamma	-18403.66 ± 31.76	-18568.05 ± 2.79
openfl	BFRY	**-16561.13 ± 137.89**	**-16947.70 ± 177.21**
	Gamma	-17475.52 ± 31.97	-17746.79 ± 6.65

We next demonstrate that accurately capturing power law structures in datasets will improve predictive performance. While fixing $\beta = 1.0$, we simulate three network datasets with 5,000 nodes from our model with discount parameters $\alpha = 0.3, 0.5$, and 0.7, respectively, which therefore exhibit increasingly lighter-tailed degree distributions. The generated graphs have 117,300, 32,925, and 9,460 links, respectively. To establish a baseline model that does not exhibit power law degree distributions but is otherwise comparable to our model, we implement the generalized random graph where the node-specific weights are constructed from the gamma random variables $W_i \sim \mathrm{gamma}(\theta, 1)$, for some positive parameter θ, i.i.d. for every node $i \leq n$. Note that the parameter θ controls the sparsity of the generated graph; larger values of θ imply denser graphs. It follows analogously to Theorem 5.1 that

$$\mathbb{P}\{D_{n,i} = k\} \sim \frac{k^{\theta-1}}{2^{k+\theta}}, \tag{37}$$

for $k \gg 1$, as $n \to \infty$. This model therefore does not exhibit power law behavior, as desired. We refer to this model as "Gamma" and the power law graph model as "BFRY".

We ran an experiment holding out 20% of the edges in the

Table 3. Inferred hyperparameters in the experiments.

	true $\alpha = 0.3$	true $\alpha = 0.5$	true $\alpha = 0.7$	500Air	polblogs	Fb107	openfl
BFRY – α	0.33 ± 0.00	0.53 ± 0.00	0.68 ± 0.00	0.23 ± 0.03	0.64 ± 0.06	0.00 ± 0.00	0.67 ± 0.21
Gamma – θ	5.29 ± 0.01	1.42 ± 0.00	0.51 ± 0.00	5.10 ± 0.01	0.66 ± 0.00	33.58 ± 0.01	0.47 ± 0.00
BFRY – β	–	–	–	1.08 ± 0.16	1.40 ± 0.00	0.80 ± 0.0	1.28 ± 0.10

simulated graphs as test sets, training the two models on the remaining 80% of the edges. We used a mini-batch size of 5,000 edges (note that the training dataset corresponds to almost 10 million observed edges). We ran each inference procedure for 20,000 steps of stochastic gradient ascent updates, using Adam (Kingma & Ba, 2015) to adjust the learning rates at each step. We repeated each experiment 5 times, each time holding out a different test set and using a different random initialization. Again, for this experiment we fixed $\beta = 1$. In Table 1 we report a mean log-likelihood metric for the test datasets, where the metric for each run is obtained by averaging the test log-likelihoods across the states for the last 4,000 steps of the inference procedure; the displayed intervals are at ± 1 standard deviation about the metric, from across the 5 repeats. We also report a max log-likelihood metric, which simply records the maximum test log-likelihood across the last 4,000 steps of the inference procedure, instead of the average. The best performing method is highlighted in bold (which in each case was the BFRY model).

In each case, we see that the BFRY model achieves higher test log-likelihood metrics than the Gamma model, as expected, implying that accurately capturing a power law degree distribution improves predictive performance (when power law behavior is truly present in the network). In Table 3, we report the inferred values for α, which were reasonably accurate, though we see slight overestimation for some regimes, as seen in the demonstration earlier. For the baseline Gamma model, we optimized the hyperparameter θ using gradient ascent maximizing the evidence lower bound of the model (c.f. Eq. (23)), and the inferred values are also reported in Table 3.

Next, we ran similar experiments on the following network datasets, each of which are expected to exhibit power law degree distributions:

- 'USTop500Airports': 500 nodes, 2,980 links
- 'openflights': 7,976 nodes, 15,243 links
- 'polblogs': 1,490 nodes, 9,517 links
- 'Facebook107': 1,034 nodes, 26,749 links

Where appropriate, we saved only the upper triangular parts of the adjacency matrices. The 'USTop500Airports' dataset contains the (undirected, unweighted) flight connections between the 500 busiest US airports. The similar, though much larger, 'openflights' dataset contains the flight connections between non-US airports. Scale-free networks have been proposed for such *traffic networks*, detailed for these datasets by Colizza et al. (2007). The 'polblogs' dataset contains the links between political blogs (judged by hyperlinks between the front webpages of the blogs) in the period leading up to the 2004 US presidential election, which is observed to exhibit power law degree distributions by Adamic & Glance (2005). The 'Facebook107' dataset contains "friendships" between users of a Facebook app, collected by Leskovec & McAuley (2012); social networks are widely studied for their power law degree distributions.

For both the Gamma and BFRY models, we ran our variational inference procedure for 20,000 steps on each dataset. As before, we repeated the experiment 5 times for each network, each time holding out a different 20% of the edges in the network as a testing set. We selected the value of β from among the grid $\{0.6, 0.9, 1.0, 1.2, 1.4\}$ with 5-fold cross validation on the training set. We set the minibatch size to be equal to the number of nodes in the graph; for example, we used minibatches of 1,490 edges for the polblog dataset. The evaluation metrics on the test datasets are summarized in Table 2, and the inferred hyperparameter values are reported in Table 3. We see that the BFRY model once again outperforms the Gamma baseline model, according to the test log-likelihood metrics.

Probabilistic inference on α by the BFRY model provides some of the most interesting analyses here. With $\alpha \approx 0.00$ (underflowing our machine's precision), the Facebook107 social network has the degree distribution with the heaviest tails, followed by the USTop500Airports traffic network with $\alpha \approx 0.23$, the polblog citation network with $\alpha \approx 0.64$, and the openflights network has the lightest tailed degree distribution with $\alpha \approx 0.67$.

8. Future work

Future work could focus on implementing the latent factor modeling generalizations presented in Section 5, which are natural assumptions in many domains where networks are expected to exhibit power law degree distributions. Alternative approaches to inference on the sparsity parameter β should also be explored, since controlling the sparsity in the graph was important for good predictive performance.

Acknowledgements

The authors thank Remco van der Hofstad for helpful advice and anonymous reviewers for helpful feedback. J. Lee and S. Choi were partly supported by an Institute for Information & Communications Technology Promotion (IITP) grant, funded by the Korean government (MSIP) (No.2014-0-00147, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Center)) and Naver, Inc. C. Heaukulani undertook this work in part while a visiting researcher at the Hong Kong University of Science and Technology, who along with L. F. James was funded by grant rgc-hkust 601712 of the Hong Kong Special Administrative Region.

References

Adamic, L. A. and Glance, N. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pp. 36–43, 2005. URL http://www.cise.ufl.edu/research/sparse/matrices/Newman/polblogs.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems*, 2009.

Albert, R. and Barabási, A-L. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

Barabási, A. and Albert, R. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Bertoin, J., Fujita, T., Roynette, B., and Yor, M. On a particular class of self-decomposable random variables: the durations of bessel excursions straddling independent exponential times. *Probability and Mathematical Statistics*, 26:315–366, 2006.

Bollobás, B. *Random graphs*. Springer, 1998.

Bollobás, B. and Riordan, O. M. Mathematical results on scale-free random graphs. *Handbook of graphs and networks: from the genome to the internet*, pp. 1–34, 2003.

Bollobás, B., Riordan, O., Spencer, J., and Tusnády, G. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3):279–290, 2001.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, 2010.

Britton, T., Deijfen, M., and Martin-Löf, A. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124(6):1377–1397, 2006.

Cai, D. and Broderick, T. Completely random measures for modeling power laws in sparse graphs. In *NIPS 2015 Workshop on Networks in the Social and Information Sciences*, 2015.

Caron, F. and Fox, E. B. Sparse graphs using exchangeable random measures. *arXiv preprint arXiv:1401.1137*, 2014.

Colizza, V., Pastor-Satorras, R., and Vespignani, A. Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, 3(4):276–282, 2007. URL https://sites.google.com/site/cxnets/usairtransportationnetwork.

Crane, H. and Dempsey, W. Atypical scaling behavior persists in real world interaction networks. *arXiv preprint arXiv:1509.08184*, 2015.

Crane, H. and Dempsey, W. Edge exchangeable models for network data. *arXiv preprint arXiv:1603.04571*, 2016.

Devroye, L. and James, L. F. On simulation and properties of the stable law. *Statistical methods & applications*, 23 (3):307–343, 2014.

Dorogovtsev, S. N. and Mendes, J. F. F. Evolution of networks. *Advances in physics*, 51(4):1079–1187, 2002.

Ghahramani, Z., Griffiths, T. L., and Sollich, P. Bayesian nonparametric latent feature models. *Bayesian Statistics*, 8:201–226, 2007. See also the discussion and rejoinder.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

James, L. F., Orbanz, P., and Teh, Y. W. Scaled subordinators and generalizations of the Indian buffet process. *arXiv preprint arXiv:1510.07309*, 2015.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

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

Kingma, D. P. and Ba, J. Adam: a method for stochastic optimization. In *ICLR*, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *ICLR*, 2014.

Knowles, D. A. Stochastic gradient variational Bayes for gamma approximating distributions. *arXiv preprint arXiv:1509.01631*, 2015.

Lee, J., James, L. F., and Choi, S. Finite-dimensional BFRY priors and variational Bayesian inference for power law models. In *Advances In Neural Information Processing Systems*, pp. 3162–3170, 2016.

Leskovec, J. and McAuley, J. J. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems 25*, 2012. URL https://snap.stanford.edu/data/egonets-Facebook.html.

Miller, K., Jordan, M. I., and Griffiths, T. L. Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*, 2009.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. Mcmc for doubly-intractable distributions. In *UAI*, 2006.

Nowicki, K. and Snijders, T. A. B. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

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

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, 1951.

Salimans, T. and Knowles, D. A. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

van der Hofstad, R. *Random graphs and complex networks: Volume 1*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016. URL http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf.

Veitch, V. and Roy, D. M. The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*, 2015.