Sampling and Inference for Beta Neutral-to-the-Left Models of Sparse Networks

Benjamin Bloem-Reddy
Department of Statistics
University of Oxford

Adam Foster
Department of Statistics
University of Oxford

Emile Mathieu
Department of Statistics
University of Oxford

Yee Whye Teh
Department of Statistics
University of Oxford

Abstract

Empirical evidence suggests that heavy-tailed degree distributions occurring in many real networks are well-approximated by power laws with exponents η that may take values either less than and greater than two. Models based on various forms of exchangeability are able to capture power laws with $\eta < 2$, and admit tractable inference algorithms; we draw on previous results to show that $\eta > 2$ cannot be generated by the forms of exchangeability used in existing random graph models. Preferential attachment models generate power law exponents greater than two, but have been of limited use as statistical models due to the inherent difficulty of performing inference in non-exchangeable models. Motivated by this gap, we design and implement inference algorithms for a recently proposed class of models that generates η of all possible values. We show that although they are not exchangeable, these models have probabilistic structure amenable to inference. Our methods make a large class of previously intractable models useful for statistical inference.

1 INTRODUCTION

Sparsity and heavy-tailed degree distributions are believed to occur in many real networks (Newman, 2005; Clauset, Shalizi, and Newman, 2009). Sparsity has been well-studied and is an intuitive concept: The typical social network user interacts with only a vanishing fraction of all users as the network grows. Heavy-tailed degree distributions and the mechanisms that generate them are not as well understood. However, empirical evidence indicates that heavy-tailed distributions are expressed in a wide range of settings, including network degrees (Clauset,

Shalizi, and Newman, 2009). Power law degree distributions, in which the proportion of vertices with degree d is $\propto d^{-\eta}$, are often used as models for real degree distributions, and serve as a useful analytic tool for characterizing the asymptotic properties of random network models.

Many statistical network models make the assumption of exchangeability over vertices, appealing to the Aldous–Hoover theorem (Hoover, 1979; Aldous, 1981) for theoretical justification. Noting that networks sampled from these models cannot be sparse, Orbanz and Roy (2015) posed a question paraphrased as, "Can a probabilistic model for random graphs produce sparse networks and have some useful notion of probabilistic symmetry?" A generation of models answered in the affirmative by incorporating other notions of exchangeability: In an exchangeable point process representation of a network (Caron and Fox, 2017; Veitch and Roy, 2015; Borgs et al., 2016), or as an exchangeable sequence of edges (Crane and Dempsey, 2017; Cai, Campbell, and Broderick, 2016; Williamson, 2016). Under certain parameterizations, these models generate sparse networks. They are able to generate asymptotic power law degree distributions, providing a better fit to real network data than their vertex exchangeable counterparts. However, the power law exponent of the degree distribution in both model classes is constrained to the interval $\eta \in (1, 2)$. That interval is not an artifact of particular model specifications. Rather, it is a basic property resulting from the fact that the average vertex degree is asymptotically unbounded; vertex degrees grow, on average, linearly in the number of edges. For some data, this property may be undesirable; such properties ideally would be inferred from a model able to capture a larger range of power law behavior.

In a largely disjoint literature, so-called preferential attachment (PA) models have been studied primarily for their ability to generate power law degree distributions from a simple size-biased reinforcement mechanism, and for their analytical tractability (e.g., Barabási and Albert, 1999; Berger et al., 2014; Peköz, Röllin, and Ross, 2017).

PA models have power law exponents $\eta > 2$. As we explain in Section 2, the exponent range is tied to PA models' non-exchangeability—a property that has made them, until now, of limited use as statistical models. In particular, if the history of the network is unobserved, the order of the edges must be inferred or marginalized; even for networks of modest size, such inference over permutations is generally intractable.

Recently, Bloem-Reddy and Orbanz (2017) introduced a class of models that can generate random graphs with power law degree distributions of any exponent $\eta \in (1, \infty)$. For reasons discussed below, we propose naming them Beta Neutral-to-the-Left (BNTL) models. BNTL models generalize many known models that have a size-biased reinforcement mechanism, including a sub-class of edge exchangeable models based on the Pitman–Yor process, and variations of the PA model. The cost of the additional flexibility is exchangeability; BNTL models depend on the times at which new vertices arrive and are not exchangeable in any known sense. However, as we show in Section 3, BNTL models have probabilistic structure—namely, left-neutrality—that may be exploited for efficient computation, making a large class of previously intractable models useful for statistical inference.

Bloem-Reddy and Orbanz (2017) established the asymptotic properties of BNTL models; statistical modeling and inference were left unstudied. Our contributions are:

- We identify left-neutrality as the key property that yields tractable inference schemes.

- We categorize and give solutions to the BNTL inference problem based on what data are available: We design schemes for maximum likelihood estimation when vertex arrival times are observed, and for Bayesian inference when an unlabeled network is observed.

- We implement these schemes on real networks of various sizes, from modest ($\sim 10^2$ vertices) to massive ($\sim 10^6$ vertices).

2 POWER LAWS IN RANDOM GRAPH MODELS

This section provides some context, and collects and interprets various results from random graph models with asymptotic degree distributions exhibiting power law tails. Although none of the results here are new, to our knowledge they have not been coherently synthesized in the literature. Technical details are omitted; they may be found in the references given throughout the section. We focus our attention on edge exchangeable and PA models because they are most similar to the BNTL framework.

A graph G is a set of vertices, $\mathcal{V}(G)$, and of edges,[1] $\mathcal{E}(G)$, between them. A multigraph allows for multiple edges between vertices; we consider each edge to be distinct, rather than as one integer-valued edge. We consider only multigraphs and henceforth refer to them as graphs. A sequence of growing graphs $G_1, G_2, \ldots$ is a stochastic process $\mathbf{G}$, indexed by the number of edges, n. Hence, G_n may be interpreted as G_{n-1} with an additional edge, either between two vertices in $\mathcal{V}(G_{n-1})$, between to new vertices, or between one old and one new vertex. We assume that the edges are labeled according to the order in which they appear, though this assumption is not necessary for edge exchangeable models (discussed below). As such, G_n may be viewed simply as a sequence of edges $\mathbf{E}_n := (E_1, \ldots, E_n)$ or, even more simply, as a sequence of *ends of edges*, denoted $\mathbf{Z}_{2n} = (Z_1, \ldots, Z_{2n})$. We denote by $G(\mathbf{Z}_{2n})$ the labeled graph with n edges constructed from $\mathbf{Z}_{2n}$. (For convenience, we will use the subscript n for all sequences when there is no risk of confusion.)

For a graph $G(\mathbf{Z}_n)$, $K_n := |\mathcal{V}(G(\mathbf{Z}_n))|$ is the number of vertices (i.e., the number of unique values in $\mathbf{Z}_n$); the degree of vertex j, $d_{j,n} := \sum_{i=1}^{n} \mathbb{1}\{Z_i = j\}$, is equal to the number of ends of edges connected to it. Let $m_n(d)$ denote the number of vertices with degree d. The asymptotic degree distribution of $G_1, G_2, \ldots$ is said to have power law tail with exponent $\eta > 1$ if

$$ p_n(d) = \frac{m_n(d)}{K_n} \xrightarrow[n \to \infty]{p} p_d \overset{d \uparrow \infty}{\sim} L(d)d^{-\eta} \quad \text{for large } d , $$

such that $\sum_{d \geq 1} p_d = 1$, for some slowly varying function $L(d)$: $\lim_{x \to \infty} L(rx)/L(x) = 1$ for all $r > 0$ (Bingham, Goldie, and Teugels, 1989). For power law tails, we state the following fact (see Appendix A).

Fact. As $n \to \infty$, if the expected average degree is unbounded, then $\eta \in (1, 2)$; if it is bounded, $\eta \in (2, \infty)$.

Edge exchangeable models (Crane and Dempsey, 2017; Cai, Campbell, and Broderick, 2016). Let G_n be specified by its sequence of edges $\mathbf{E}_n$ (not necessarily ends of edges), which is assumed to be exchangeable: Its distribution is invariant under all permutations of the order of the edges for all n, i.e., the labels carry no information about their distribution. As a consequence of the law of large numbers for exchangeable sequences, the counts of all non-zero multi-edges grow linearly in n and thus so do the vertex degrees. That is, $d_{j,n} = \Theta(n)$. Furthermore, $K_n = o(n)$. The average degree is unbounded, implying that if the degree distribution tail follows a power law, then $\eta \in (1, 2)$.

---

[1] We treat all graphs as undirected; extension to directed graphs is straightforward.

As an example, consider sampling $\mathbf{Z}$ from the the Pitman–Yor process ($\mathcal{PYP}$) (Ishwaran and James, 2001) with parameters $\tau \in (0, 1), \theta > -\tau$,

$$\mathbb{P}[Z_{n+1} \in \cdot | \mathbf{Z}_n] = \frac{\theta + K_n \tau}{n + \theta} \delta_{K_n+1}(\cdot) \tag{1}$$
$$+ \frac{n - K_n \tau}{\theta + n} \sum_{j=1}^{K_n} \frac{d_{j,n} - \tau}{n - K_n \tau} \delta_j(\cdot) \,.$$

It can be shown that the asymptotic degree distribution has power law tail (Pitman, 2006),

$$n^{-\tau} m_d(n) \xrightarrow[n\to\infty]{\text{a.s.}} p_d \overset{d\uparrow\infty}{\sim} d^{-(1+\tau)} \,,$$

which implies that $\eta_\tau = 1 + \tau \in (1, 2)$.

The predictive rule (1) demonstrates why the expected average degree is unbounded. The probability that Z_{n+1} corresponds to a new vertex is $\frac{\theta + K_n \tau}{n + \theta}$, which is arbitrarily close to zero as $n \to \infty$. For large n, the expected interarrival time between new vertices becomes arbitrarily large, and edges pile up on the existing vertices. Intuitively, vertex j takes part in a constant fraction of all interactions as n grows. This property is shared by all edge exchangeable models; an analogous property holds for exchangeable point process models (see Appendix B).

Preferential attachment models. Although the $\mathcal{PYP}$ has the same size-biased reinforcement mechanism common to all PA models, typically it is not considered to be part of the same class as the PA models in the probability literature, of which Barabási and Albert (1999) provide the prototypical example. However, the difference between them amounts to how frequently new vertices appear (Bloem-Reddy and Orbanz, 2017). For our purposes, this is best illustrated with a simple PA model, the Yule–Simon (YS) model (Simon, 1955). For $\beta \in (0, 1)$, $\mathbf{Z}$ is generated via the predictive rule

$$\mathbb{P}[Z_{n+1} \in \cdot | \mathbf{Z}_n] = \beta \delta_{K_n+1}(\cdot) + (1 - \beta) \sum_{j=1}^{K_n} \frac{d_{j,n}}{n} \delta_j(\cdot). \tag{2}$$

The YS model is known to generate power law degree distributions with $\eta_\beta = 1 + \frac{1}{1-\beta} \in (2, \infty)$ (Simon, 1955). Different versions of PA exhibit a range of possible η's, but it is generally the case that $\eta_{\text{PA}} > 2$, and this is tied to their lack of exchangeability. The average rate at which new vertices arrive is constant in n; hence, $K_n = \Omega(n)$, implying bounded expected average degree. The "edge pileup" phenomenon of exchangeable models does not occur: $d_{j,n} = o(n)$. In the YS model, $d_{j,n} = \Theta(n^{1-\beta})$.

3 BETA NTL MODELS

BNTL models were introduced under the name (α, T)-models by Bloem-Reddy and Orbanz (2017), who studied their distributional and asymptotic properties. We briefly review the definition of BNTL models and describe the properties that make them amenable to inference.

In the predictive distributions (1)-(2), the probability that Z_{n+1} is a new vertex is independent of the degrees $d_{j,n}$, which allows the sampling of $\mathbf{Z}$ to be separated into two parts: A sequence $T_1 < T_2 < \dots$ of *arrival times* of new vertices, and size-biased reinforcement at all steps not associated with an arrival time. As such, a BNTL model is parameterized by a scalar "discount parameter" $\alpha \in (-\infty, 1)$ and a probability distribution Λ on strictly increasing integer-valued sequences, which specifies the law of the arrival times $T_1, T_2, \dots$. A sequence $\mathbf{Z}$ is said to have law BNTL(α, Λ) if, for a random arrival time sequence $\mathbf{T} = (T_1, T_2, \dots) \sim \Lambda$, $\mathbf{Z}$ is sampled as

$$\mathbb{P}[Z_{n+1} \in \cdot | \mathbf{Z}_n, \mathbf{T}] = \mathbb{1}\{n + 1 = T_{K_n+1}\} \delta_{K_n+1}(\cdot)$$
$$+ \mathbb{1}\{n + 1 < T_{K_n+1}\} \sum_{j=1}^{K_n} \frac{d_{j,n} - \alpha}{n - K_n \alpha} \delta_j(\cdot) \,. \tag{3}$$

In practice, it may be simpler to specify the distribution of *interarrival* times $\Delta_j = T_j - T_{j-1}$, and use their partial sums to construct $\mathbf{T}$; we discuss this in more detail in Section 4. The similarity of (3) to (1)-(2) is not coincidental. The $\mathcal{PYP}$ and the YS model each correspond to particular parameterizations of the BNTL model: The YS model corresponds to i.i.d. $\Delta_j \sim \text{Geom}(\beta)$; the arrival time distribution induced by the $\mathcal{PYP}$ in (1) also has known form (see (11)).

For a given $\mathbf{T}$, the probability of any $G(\mathbf{Z}_n)$ is

$$\mathbb{P}[G(\mathbf{Z}_n)|\mathbf{T}] = \mathbb{P}[G(\mathbf{Z}_n)|\mathbf{T}_{K_n+1}, K_n] \tag{4}$$
$$= \frac{\Gamma(d_{1,n} - \alpha)}{\Gamma(n - K_n \alpha)} \prod_{j=2}^{K_n} \frac{\Gamma(T_j - j\alpha)\Gamma(d_{j,n} - \alpha)}{\Gamma(T_j - 1 - (j-1)\alpha)\Gamma(1 - \alpha)} \,.$$

A crucial property that makes BNTL models amenable to inference is that conditioned on $\mathbf{T}$, the joint probability (4) factorizes over the vertices; each term is expressed in terms of its arrival time, T_j, and its degree, $d_{j,n}$. Note that given $\mathbf{T}$, the degree sequence $\mathbf{d}_{K_n} := (d_{1,n}, \dots, d_{K_n,n})$ is a sufficient statistic for α. Furthermore, the distribution of the arrival times (and therefore K_n) is independent of the degrees. The factorization becomes explicitly useful in the Gibbs sampling updates and in the maximum likelihood estimating equations in Section 4.

Sampling representation. Like their exchangeable counterpart the $\mathcal{PYP}$, BNTL models have a sampling representation in terms of products of independent beta random

variables: (3) is an urn sequence corresponding to the following (Bloem-Reddy and Orbanz, 2017):

- $\mathbf{T} \sim \Lambda$.
- $\Psi_j | T_j \sim \mathrm{Beta}(1 - \alpha, T_j - 1 - (j - 1)\alpha)$ for $j \geq 1$.
- $P_{j,K_n} = \Psi_j \prod_{\ell=j+1}^{K_n} (1 - \Psi_\ell)$
- $Z_n \sim \begin{cases} \delta_{K_n}(\cdot) & \text{for } n = T_{K_n} \\ \mathrm{Categorical}(P_{j,K_n}) & \text{o.w.} \end{cases}$

(By convention, $\mathrm{Beta}(a, 0)$ is a point mass on 1, so $\Psi_1 = 1$.) The last two items specify that when there are k vertices in the graph, Z_n is sampled from a categorical distribution over those vertices, each with probability $P_{j,k}$. After the subsequent arrival time, T_{k+1}, when there are $k + 1$ vertices, the probability that $Z_n = j$ is

$$P_{j,k+1} = \begin{cases} P_{j,k}(1 - \Psi_{k+1}), & j \in \{1, \ldots, k\} \\ \Psi_{k+1}, & j = k + 1 \end{cases}.$$

That is, the vector of probabilities $\mathbf{P}_k = (P_{1,k}, \ldots, P_{k,k})$ grows in length as each new vertex arrives, and each of the previous entries is scaled by $(1 - \Psi_{\mathrm{new}})$.

Neutrality. The recursive scaling of $P_{j,k}$ is the essence of a neutral-to-the-left (NTL) sequence. A random vector $\mathbf{X} = (X_1, \ldots, X_k) \in \mathbb{R}_+^k$, is NTL if the increments,

$$R_j := \frac{X_j}{\sum_{i=1}^{j} X_i}, \tag{5}$$

form a sequence of mutually independent random variables; a non-decreasing stochastic process Z defined on $\mathbb{R}$ is NTL if the vector of increments $\frac{Z(t_j) - Z(t_{j-1})}{Z(t_j)}$ is NTL for any finite partition $-\infty \leq t_1 < \ldots < t_k \leq \infty$ of $\mathbb{R}$ (Doksum, 1974). A bit of algebra shows that $\mathbf{P}_k$ is NTL: The corresponding sequence of increments is $R_j = \Psi_j$, for all k. Intuitively, this must be the case due to the recursive scaling construction. Together with the beta random variables in the sampling representation, left-neutrality characterizes these models; hence the name.

Neutral-to-the-right (NTR) processes are better known than NTL processes, and appear throughout the Bayesian statistics literature, both explicitly (Walker and Muliere, 1997; James, 2006) and implicitly in the form of the stick-breaking constructions of the Dirichlet Process and the $\mathcal{PYP}$ (e.g., Ishwaran and James, 2001). The properties of right- and left-neutrality are, as their names suggest, symmetric opposites: A NTR vector in reverse order is NTL, and vice versa.

The independence properties that make NTR stick-breaking constructions useful for modeling and inference purposes transfer in large part to NTL models. The Ψ_j's are conditionally independent given the T_j's; along with

the parameters of the beta distribution, this independence induces the factorized form in (4). In the exchangeable random partitions literature, a model with joint probability that factorizes over the blocks and the probability of having K_n blocks is known as *Gibbs-type* (Gnedin and Pitman, 2006).

Sparsity and power law tails in BNTL models. The asymptotic behavior of BNTL models is controlled primarily by the arrival times, $\mathbf{T}$. In order to obtain sparse graphs, K_n must be $\omega(n^{1/2})$. If $\mathbf{T}$ are the arrival times from an exchangeable sequence Z, then K_n is at most $\Theta(n^\delta)$, for some $\delta \in (0, 1)$, in which case $\eta = 1 + \delta$ (Pitman, 2006); thus, sparse graphs generated this way have $\eta \in (3/2, 2)$. For the $\mathcal{PYP}$, $\delta = \tau$. Alternatively, for $\mathbf{T}$ sampled such that the mean interarrival time,

$$\bar{\Delta}_{K_n} := \frac{1}{K_n - 1} \sum_{j=2}^{K_n} \Delta_j, \tag{6}$$

converges to some finite μ, then $K_n = \Theta(n)$ and $\eta = 1 + \frac{\mu - \alpha}{\mu - 1} > 2$. Furthermore, vertex degrees grow as $d_{j,n} = \Theta(n^{\frac{\mu-1}{\mu-\alpha}})$ (Bloem-Reddy and Orbanz, 2017). Thus, depending on the specification of the arrival time distribution, BNTL models can achieve any $\eta \in (1, \infty)$.

Microclustering in BNTL partitions. The sequence $\mathbf{Z}_n$ can be transformed into an arrival-ordered partition $\Pi(\mathbf{Z}_n) := \{B_{1,n}, \ldots, B_{K_n,n}\}$ of $[n] := \{1, \ldots, n\}$ by grouping $\mathbf{Z}_n$ into blocks $B_{j,n} := \{i \in [n] : Z_i = j\}$. There is a bijective mapping between $\Pi(\mathbf{Z}_n)$ and $G(\mathbf{Z}_n)$ for all n (Bloem-Reddy and Orbanz, 2017), which puts blocks of the partition in correspondence with vertices of the graph. Hence, properties of $G(\mathbf{Z}_n)$ translate into properties of $\Pi(\mathbf{Z}_n)$. In particular, the growth rate of vertex degrees translates to the growth rate of blocks sizes. Recent work (Betancourt et al., 2016; Di Benedetto, Caron, and Teh, 2017) has explored the so-called microclustering property, which is defined as block sizes that grow sub-linearly in n. The $\eta > 2$ range of BNTL models corresponds precisely with this property. Although we do not make explicit statements about partition-valued data, statements about graphs are easily translated into statements about partitions via the correspondence between blocks and vertices. In particular, the inference algorithms in Section 4 are valid for partition-valued data.

4 INFERENCE

Although PA models exhibit a range of power laws not captured by exchangeable models, they face a significant barrier to use as statistical models due to their inherent lack of exchangeability. At a high level, applying a non-exchangeable model to data for which the order is unknown requires inference over permutations of the data.

This is, in general, a prohibitively difficult problem even for modest n. However, using the probabilistic structure of BNTL models, we design a Gibbs sampling algorithm that overcomes this difficulty for networks with thousands of vertices (Section 4.1). If the ordered edge sequence is observed, maximum likelihood estimation scales to networks with millions of vertices (Section 4.2).

Given the hierarchical nature of BNTL models, inference may be performed at a number of levels. In the simplest case, suppose the data are a sequence of edge-ends, $\mathbf{Z}_n$. From this sequence the arrival times and the arrival-ordered graph can be perfectly reconstructed, and inferring the parameters ϕ of the arrival distribution Λ^ϕ and the parameters $\mathbf{\Psi}_{K_n} := (\Psi_j)_{j=1}^{K_n}$ is straightforward: Simple maximum likelihood estimators exist for $\mathbf{\Psi}_{K_n}$ (see Appendix C), and for the parameters of many arrival time distributions of interest, or equally simple MCMC samplers may be constructed for Bayesian inference.

More challenging are the situations in which some aspect of the data is not perfectly observed. For graph-valued observations, the following table summarizes the range of possibilities, in order of increasing difficulty of inference:

Observation	Unobserved variables
End of edge sequence $\mathbf{Z}_n$	$\alpha, \phi, \mathbf{\Psi}_{K_n}$
Vertex arrival-ordered graph	$\alpha, \phi, \mathbf{\Psi}_{K_n}, \mathbf{T}_{K_n}$
Unlabeled graph	$\alpha, \phi, \mathbf{\Psi}_{K_n}, \mathbf{T}_{K_n}, \sigma[K_n]$

The last row presents a significant challenge. In particular, the unobserved variables include a permutation σ mapping the arrival-ordered sequence to some arbitrary ordering of the vertices (by which the vertices are uniquely identified). For a graph with K_n vertices, there are $K_n!$ possible permutations. Conditioned on a sequence of arrival times, some permutations have zero posterior probability, making the problem space both high-dimensional and constrained. Despite these difficulties, the inference problem is much simpler than that of a generic non-exchangeable model for a sequence of n data points: Even in sparse graphs, typically $K_n \ll n$ and thus the dimension of the problem is exponentially smaller. Furthermore, the form of (4) yields simple conditional distributions for Gibbs sampling.

4.1 GIBBS SAMPLING

In this section, we build from the simplest inference problem to the hardest, progressing through the table in the previous section. The full sampler infers the posterior distributions of the parameters $\mathbf{\Psi}_{K_n}$ and α, of the arrival times $\mathbf{T}_{K_n}$, of the parameters of the arrival time distribution, and of the permutation of the vertices. In order to maintain the structure of the factorization over vertices

in (4), we assume that the arrival time distribution has a Markov factorization (with a slight abuse of notation):

$$\mathbf{\Lambda}^\phi(\mathbf{T}_k) = \delta_{T_1}(1) \prod_{j=2}^k \Lambda_j^\phi(\Delta_j | T_{j-1}) , \qquad (7)$$

with ϕ representing any parameters. Examples are i.i.d. interarrivals such that $\Lambda_j^\phi(\Delta_j | T_{j-1}) = p_\phi(\Delta_j)$; and interarrivals that depend on the previous interarrivals through their sum and the number of previous arrivals, such as the interarrival sequence generated by exchangeable Gibbs-type sequences (De Blasi et al., 2015).

Suppose we observe a sequence of edge-ends $\mathbf{Z}_n$. Denote the partial sums of the ordered degree sequence as $\bar{d}_{j,n} = \sum_{i=1}^j d_{j,n}$. For any fixed α and ϕ,

$$p_{\alpha,\phi}(\mathbf{Z}_n, \mathbf{\Psi}_{K_n}) = \qquad (8)$$
$$\prod_{j=2}^{K_n} \frac{\Psi_j^{d_{j,n}-\alpha-1}(1-\Psi_j)^{\bar{d}_{j-1,n}-(j-1)\alpha-1}}{B(1-\alpha, T_j - 1 - (j-1)\alpha)} \Lambda_\phi(T_j \mid T_{j-1})$$
$$\times \Lambda_\phi(T_{K_n+1} > n \mid T_{K_n}) ,$$

where $B(a,b)$ is the beta function, and $\Lambda_\phi(T_{K_n+1} > n \mid T_{K_n})$ is the censored probability of vertex $K_n + 1$'s unobserved arrival time. Note that marginalizing $\mathbf{\Psi}_{K_n}$ recovers (4).

Updates for $\mathbf{\Psi}_{K_n}$. From (8) it is clear that

$$\Psi_j \mid \mathbf{Z}_n, \mathbf{\Psi}_{\setminus j} \sim \text{Beta}(d_{j,n} - \alpha, \bar{d}_{j-1,n} - (j-1)\alpha) , \qquad (9)$$

where $\mathbf{\Psi}_{\setminus j}$ is shorthand for the sequence $\mathbf{\Psi}_{K_n}$ with Ψ_j excluded. That is, given the arrival-ordered block sizes, the $\mathbf{\Psi}_{K_n}$ are independent of each other and of the arrival times, and the beta distribution is the conjugate prior for the BNTL sampling process. To understand this, consider a second scenario in which a graph is observed with vertices labeled in order of arrival (though not their time of arrival). The data consist of an ordered sequence of degrees, $\mathbf{d}_{K_n} = (d_1, \ldots, d_{K_n})$, which corresponds to more than one possible edge-end sequence $\mathbf{Z}_n$. The model places equal probability on each sequence that gives rise to the same arrival-ordered degree sequence $\mathbf{d}_{K_n}$ and the same arrival times; summing over these sequences yields

$$p_{\alpha,\phi}(\mathbf{d}_{K_n}, \mathbf{\Psi}_{K_n} \mid \mathbf{T}_{K_n}, K_n) \qquad (10)$$
$$= \prod_{j=2}^{K_n} \frac{\Psi_j^{d_j-\alpha-1}(1-\Psi_j)^{\bar{d}_{j-1}-(j-1)\alpha-1}}{B(1-\alpha, T_j - 1 - (j-1)\alpha)} \binom{\bar{d}_j - T_j}{d_j - 1} .$$

The binomial coefficients count the number of sequences $\mathbf{Z}_n$ that yield $\mathbf{d}_{K_n}$, given $\mathbf{T}_{K_n}$ (Griffiths and Spanò, 2007). (10) is a product of binomial likelihoods with beta priors. Hence, the conjugacy derived in (9).

Updates for α. In both observation scenarios, generic MCMC methods such as slice sampling (Neal, 2003) can be used to sample from the full conditional distribution of α. We use slice sampling in the experiments in Section 5.

Updates for ϕ. Many models of i.i.d. interarrival times will yield conjugate updates for ϕ. For other models, generic MCMC methods can be used. In the experiments in Section 5, we consider three interarrival models: i.i.d. Geom(β) and i.i.d. Pois$_+(\lambda)$, which is the Poisson distribution shifted to the positive integers; and the interarrival distribution induced by the $\mathcal{PYP}$, which is (Griffiths and Spanò, 2007)

$$\Lambda_{j+1}^{\theta,\tau}(\Delta_{j+1} = s \mid T_j) \qquad (11)$$
$$= (\theta + j\tau)\frac{\Gamma(\theta + T_j)\Gamma(T_j + s - 1 - j\tau)}{\Gamma(\theta + T_j + s)\Gamma(T_j - j\tau)} .$$

In the former two cases, conjugate updates are performed (conditioned on $\mathbf{T}_{K_n}$); in the latter case, we perform univariate slice sampling for each of θ and τ.

Updates for $\mathbf{T}_{K_n}$. The assumed Markov structure of the arrival times induces a simple conditional distribution for T_j that is supported on the set $S_j = \{T_{j-1} + 1, \ldots, T_{j-1} + M_j\}$, where $M_j = \min\{T_{j+1} - T_{j-1} - 1, \bar{d}_{j-1} - T_{j-1} + 1\}$. The support set enforces the constraints that $\bar{d}_{j-1} \geq T_j - 1$, and that $T_{j-1} < T_j < T_{j+1}$. Conditioning on T_{j-1} and T_{j+1}, updating T_j is equivalent to updating Δ_j and Δ_{j+1}; for $j = 2, \ldots, K_n - 1$,

$$p_{\alpha,\phi}(\Delta_j = s, \Delta_{j+1} = T_{j+1} - T_{j-1} - s \mid \mathbf{T}_{\backslash j}, \mathbf{d}_n)$$
$$\propto \frac{\Lambda_{j+1}^{\phi}(T_{j+1} - T_{j-1} - s \mid T_{j-1} + s)\Lambda_j^{\phi}(s \mid T_{j-1})}{B(1 - \alpha, T_{j-1} + s - 1 - (j-1)\alpha)}$$
$$\times \binom{\bar{d}_j - T_{j-1} - s}{d_j - 1} .$$

For $j = K_n$,

$$p_{\alpha,\phi}(\Delta_{K_n} = s \mid \mathbf{T}_{\backslash K_n}, \mathbf{d})$$
$$\propto \Lambda_{K_n}^{\phi}(s \mid T_{K_n - 1})\binom{n - T_{K_n - 1} - s}{d_{K_n} - 1}$$
$$\times \frac{\Lambda_{K_n+1}^{\phi}(\Delta_{K_n+1} > n - T_{K_n - 1} - s \mid T_{K_n - 1} - s)}{B(1 - \alpha, T_{K_n - 1} + s - 1 - (K_n - 1)\alpha)} ,$$

and $M_{K_n} = \min\{n - T_{K_n - 1} - 1, \bar{d}_{j-1} - T_{j-1} + 1\}$.

For i.i.d. interarrivals with distribution p_ϕ, the updates are particularly easy to compute because

$$\Lambda_{j+1}^{\phi}(T_{j+1} - T_{j-1} - s \mid T_{j-1} - s)\Lambda_j^{\phi}(s \mid T_{j-1})$$
$$= p_\phi(T_{j+1} - T_{j-1} - s)p_\phi(s) . \qquad (12)$$

$p_\phi(s)$ can be computed for each $s \in \{1, \ldots, M_j\}$; each term multiplied by the corresponding term in $s \in$

$\{M_j, \ldots, 1\}$ yields (12). In the case of Geom(β) interarrivals, the distribution is uniform on $s \in \{1, \ldots, M_j\}$:

$$\Lambda_{j+1}^{\phi}(T_{j+1} - T_{j-1} - s \mid T_{j-1} - s)\Lambda_j^{\phi}(s \mid T_{j-1}) \quad (13)$$
$$= \beta(1 - \beta)^{T_{j+1} - T_{j-1} - s - 1}\beta(1 - \beta)^{s-1} \propto 1 .$$

Updates for $\sigma[K_n]$. Given a sample of $\mathbf{T}_{K_n}$, the order of the vertices can be updated via a series of adjacent swap proposals. Let σ_j be the identity of the j-th vertex in the current sampling iteration. A sampling update of σ proposes swapping $\sigma_j \leftrightarrow \sigma_{j+1}$ with probability proportional to the value of (10), with $\mathbf{\Psi}_{K_n}$ marginalized and with d_j and d_{j+1} swapped. Due to the factorization over vertices, all but the j-th and $j + 1$-st terms are the same; as a result, swap proposals are inexpensive to compute (for compactness, '$-$' indicates all other variables):

$$p_{\alpha,\phi}(\sigma_j \leftrightarrow \sigma_{j+1}|-) \propto \frac{\Gamma(\bar{d}_{j-1} + d_{j+1} - T_j + 1)}{\Gamma(\bar{d}_{j+1} - d_j - T_{j+1} + 2)}$$
$$p_{\alpha,\phi}(\sigma_j \not\leftrightarrow \sigma_{j+1}|-) \propto \frac{\Gamma(\bar{d}_{j-1} + d_j - T_j + 1)}{\Gamma(\bar{d}_j - T_{j+1} + 2)} .$$

The simplicity of swap proposals enables many swaps to be sampled in a short amount of computational time, helping to overcome the high dimensionality of the sample space. We note that in general, local proposals of all possible permutations of $m > 1$ consecutive vertices are possible and $m > 2$ would likely enhance exploration of the state space; here we consider only $m = 2$.

Computational complexity. The slice sampling updates for α, which require evaluation of (4), are of complexity $\mathcal{O}(K_n)$, as are the permutation swap proposals. Updates for the arrival parameter(s) ϕ depend on the model, but as they depend only on the K_n arrival times, they are at most $\mathcal{O}(K_n)$. The most expensive update is that of $\mathbf{T}_{K_n}$, which is $\mathcal{O}(n)$, though the constant hidden in $\mathcal{O}$ may vary greatly across arrival models.

4.2 MAXIMUM LIKELIHOOD FOR PARAMETERS IN EDGE SEQUENCES

Suppose the edge-end sequence $\mathbf{Z}_n$ is observed. For arrival time distribution Λ^ϕ, ϕ and α can be estimated by maximum likelihood (ML). The likelihood admits the factorization

$$p_{\alpha,\phi}(\mathbf{Z}_n) = p_\alpha(\mathbf{Z}_n|\mathbf{T}_{K_n})\Lambda_\phi(\mathbf{T}_{K_n}) , \qquad (14)$$

with the practical implication that the estimating equations for α and ϕ can be solved separately. In particular,

$$\hat{\alpha} = \arg\max_{\alpha \in (-\infty, 1)} \log p_\alpha(\mathbf{Z}_n|\mathbf{T}_{K_n}) , \qquad (15)$$

where $p_\alpha(\mathbf{Z}_n \mid \mathbf{T}_{K_n})$ is as in (4).

Table 1: Results of Gibbs sampling experiments on synthetic data ($\alpha^* = 0.75$). The top four rows show results from each of four different BNTL models fit to a synthetic graph with 500 edges generated by the coupled $\mathcal{PYP}$ BNTL model; the bottom four rows show the same BNTL models fit to a synthetic graph with Geom(0.25)-distributed interarrivals.

Gen. arrival distn.	K_n	Inference model	$\|\hat{\alpha} - \alpha^*\|$	$\|\hat{\mathbf{S}} - \mathbf{S}^*\|$	Pred. log-lik.	Runtime (sec.)	ESS
$\mathcal{PYP}(1.0, 0.75)$	260	$(\tau, \mathcal{PYP}(\theta, \tau))$	**0.046 ± 0.002**	**28.5 ± 0.7**	**-2637.0 ± 0.1**	297.6 ± 0.2	0.80 ± 0.09
$\mathcal{PYP}(1.0, 0.75)$	260	$(\alpha, \mathcal{PYP}(\theta, \tau))$	0.045 ± 0.003	33.4 ± 1.0	-2638.4 ± 0.2	313.6 ± 0.4	0.77 ± 0.07
$\mathcal{PYP}(1.0, 0.75)$	260	$(\alpha, \text{Geom}(\beta))$	0.049 ± 0.004	66.8 ± 1.2	-2660.5 ± 0.7	90.5 ± 0.1	0.78 ± 0.09
$\mathcal{PYP}(1.0, 0.75)$	260	$(\alpha, \text{Pois}_+(\lambda))$	0.054 ± 0.004	68.0 ± 0.7	-2902.5 ± 1.4	112.5 ± 0.1	0.79 ± 0.07
Geom(0.25)	251	$(\tau, \mathcal{PYP}(\theta, \tau))$	0.086 ± 0.002	56.6 ± 1.3	-2386.8 ± 0.1	295.4 ± 0.6	0.83 ± 0.06
Geom(0.25)	251	$(\alpha, \mathcal{PYP}(\theta, \tau))$	0.078 ± 0.003	54.2 ± 2.0	-2387.5 ± 0.5	312.7 ± 0.3	0.66 ± 0.09
Geom(0.25)	251	$(\alpha, \text{Geom}(\beta))$	**0.043 ± 0.003**	24.8 ± 0.8	**-2382.6 ± 0.2**	87.2 ± 0.1	0.92 ± 0.04
Geom(0.25)	251	$(\alpha, \text{Pois}_+(\lambda))$	**0.041 ± 0.003**	**21.0 ± 0.5**	-2562.2 ± 0.2	109.5 ± 0.1	0.91 ± 0.05

Closed-form MLEs are known for many i.i.d. interarrival distributions. For the Geom(β) and Pois$_+(\lambda)$ distributions used in Section 5, $\hat{\beta} = \frac{K_n - 1}{n - K_n}$ and $\hat{\lambda} = \frac{n - K_n}{K_n - 1}$. MLEs for θ and τ in $\mathcal{PYP}$-induced interarrivals can be found by numerically optimizing the product over arrival times of (11). See Appendix D for details. Maximum a posteriori (MAP) estimates are straightforward to compute by placing priors on the model parameters and including the relevant prior probabilities in (14)-(15).

4.3 RELATED WORK

There is relatively little previous work on statistical inference for non-exchangeable models of network data. Bloem-Reddy and Orbanz (2018) develop sequential Monte Carlo methods for non-exchangeable models; those methods are feasible only for networks with hundreds of vertices. See references therein for related ideas based on importance sampling. Where BNTL models overlap with edge exchangeable models, there exist inference algorithms that do not account for arrival times. Namely, if $\mathbf{Z}_n$ is assumed to be an exchangeable sequence of edge-ends, then the sampling and estimation algorithms for Gibbs-type partitions can be used. For example, Gibbs sampling methods for the $\mathcal{PYP}$ are derived in Ishwaran and James (2001). Crane and Dempsey (2017) give maximum likelihood estimating equations. However, neither method infers arrival times, and the inference techniques do not extend to the wider class of non-exchangeable BNTL models.

Wan et al. (2017) propose MLEs for the parameters of a class of PA models when the edge sequence is observed. A MLE of the parameter α in a slightly different PA model was proposed by Gao and van der Vaart (2017) for observed edge sequences. The PA model considered there has random initial degrees, rather than random arrival times, but the initial degrees play a similar role to the arrival times. Those authors find that conditioned on

the initial degrees, the degree sequence at step n, $\mathbf{d}_{K_n}$, is sufficient for α, and that the MLE is asymptotically normal. Based on the similarities of the models and the corresponding log-likelihoods, it is plausible that similar properties hold for BNTL models.

5 EXPERIMENTS

We apply the inference methods developed in Section 4 to data. The first set of experiments is in the unlabeled network setting, in which the posterior distribution over vertex ordering must be inferred along with the model parameters. In a second set of experiments, we consider graphs with edges labeled in order of appearance, and demonstrate that maximum likelihood and MAP estimation scale to networks with millions of nodes.[2]

5.1 BAYESIAN INFERENCE

We first apply the Gibbs sampler from Section 4.1 to synthetic data, which allows us to study the effects of model misspecification on parameter estimation, and to demonstrate the feasibility of inference over the vertex order and the arrival times. We generated two synthetic graphs, each with 1,000 edges: One from a $\mathcal{PYP}(\theta, \tau)$ sequence (1) in which τ is forced to be equal to the BNTL parameter α, which corresponds to the edge exchangeable Hollywood model of Crane and Dempsey (2017); and one from a BNTL model with i.i.d. Geom(β)-distributed interarrival times. We set $\theta = 1.0$, $\beta = 0.25$, and in both cases, $\alpha = 0.75$.

For each of the graphs, we held out the final 500 edges for prediction, and we fit four different BNTL models to the first 500 edges whose order we treated as unknown: One with $\mathcal{PYP}(\theta, \tau)$-induced arrivals and $\alpha = \tau$ (the "coupled $\mathcal{PYP}$" model), which is the same as the generative

---

[2]Julia code is available at `https://github.com/emilemathieu/NTL.jl`.

Table 2: Scaling performance of the Gibbs sampler.

	100 edges	1,000 edges	10,000 edges
$|\hat{\alpha} - \alpha^*|$	0.12 ± 0.01	0.03 ± 0.00	0.01 ± 0.00
$|\hat{\beta} - \beta^*|$	0.02 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
$|\hat{\mathbf{S}} - \mathbf{S}^*|$	10.3 ± 0.4	33.9 ± 0.9	343 ± 1.6
ESS	0.90 ± 0.04	0.85 ± 0.05	0.75 ± 0.08
Runtime (s)	21 ± 0.0	213 ± 0.4	2267 ± 2

model of the first synthetic dataset; one with $\mathcal{PYP}(\theta, \tau)$-induced arrivals and α allowed to vary separately from τ (the "uncoupled $\mathcal{PYP}$" model); and two i.i.d. interarrival models, with $\text{Geom}(\beta)$- and $\text{Pois}_+(\lambda)$-distributed interarrivals. We ran 125,000 Gibbs sampling iterations, including a burn-in of 25,000, and collected one in every 1,000 iterations for a total of 1,000 samples. To assess performance, we calculated the average absolute error (relative to the true value) of MCMC samples of α, and of $\mathbf{S} := \frac{1}{K_n - 1}\sum_{j>1}(\bar{d}_{j-1} - T_j)$. The latter statistic captures how well the sampler recovers the vertex permutation and the arrival times. We also calculated the predictive log-likelihood of a further 500 edges. Average runtimes[3] and effective sample size (ESS) factors, based on the log of the normalized L_1 distance between the sampled degree sequence and the true degree sequence, are also shown.

Table 1 summarizes the results, averaged over 10 repetitions. The top four rows show the results of fitting four BNTL inference models to the coupled $\mathcal{PYP}$ dataset. Unsurprisingly, the inference models with arrivals induced by the $\mathcal{PYP}$ achieve the lowest errors in α and $\mathbf{S}$, and highest predictive log-likelihood. The bottom four rows show the same four inference models fit to the $\text{Geom}(0.25)$ BNTL dataset; the i.i.d. interarrival models achieve lower errors, and the $\text{Geom}(\beta)$ inference model attains the highest predictive log-likelihood. Although the $\text{Pois}_+(\lambda)$ inference model attains low errors in α and $\mathbf{S}$, the low variance of the Poisson distribution compared to the Geometric distribution means that it attains low predictive probability due to the relatively frequent occurrence of large interarrivals.

As discussed in Section 4.1, the most expensive Gibbs update is that of the arrival time sequence. As such, the $\text{Geom}(\beta)$ interarrival inference model benefits greatly from (13), which implies that computation of Λ_j^ϕ is not required. The i.i.d. interarrival models each have conjugate updates for their parameters, whereas the $\mathcal{PYP}$ interarrival models require slice sampling for $\phi = (\theta, \tau)$. These differences are reflected in the runtimes shown in Table 1. Finally, all four inference models exhibit good ESS factors, indicating that the sampler is exploring per-

mutation space beyond simply swapping vertices of the same degree.

Scaling in n. In order to study how sampling and computational efficiency scale with the size of the network, we generated a single BNTL network of 10,000 edges with i.i.d. $\text{Geom}(0.25)$-distributed arrival times, and performed Gibbs sampling using the subgraphs formed by the first n edges, with $n \in \{100, 1{,}000, 10{,}000\}$. Table 2 shows the results of 10 repetitions, each of 150,000 Gibbs iterations; samples were collected once every 1,000 iterations after a burn-in period of 75,000 iterations. Parameter estimation is increasingly accurate for increasing n without major decrease in ESS, indicating that the sampler is taking advantage of the increased statistical signal in the bigger network. Runtimes increase at a rate linear in n.

5.2 MAXIMUM LIKELIHOOD ESTIMATION ON EDGE SEQUENCES

For observed edge sequences, maximum likelihood estimation scales to networks with millions of vertices and tens of millions of edges. To demonstrate, we compute MLEs on a collection of temporal network datasets available from the Stanford Network Analysis Project (SNAP) (Leskovec and Krevl, 2014).

For each of the datasets listed in Table 3, we fit MLEs of α and of the parameters of three different interarrival models: coupled $\mathcal{PYP}(\theta, \alpha)$; uncoupled $\mathcal{PYP}(\theta, \tau)$; and $\text{Geom}(\beta)$. Table 4 displays the MLEs of the model parameters and the plug-in estimates of the asymptotic power law degree exponent, η. (The asymptotic degree distribution of the uncoupled $\mathcal{PYP}$ model is unknown.) Note that due to the factorization of the likelihood in (14), $\hat{\alpha}$ is the same for any model in which α is not coupled to the arrival distribution. In order to assess model fitness, we fit MLEs for the same BNTL models to the first 80% of the edges in each network and calculated the predictive log-likelihood based on the MLEs of the remaining 20%; this is also shown in Table 4. For context, the arrival time sequence of each dataset is plotted in Figure 1. Unsurprisingly, whether or not the arrival times are approximately linear in n largely determines which BNTL model fits

Table 3: SNAP temporal network datasets.

Dataset	# of vertices	# of edges
Ask Ubuntu	159,316	964,437
UCI social network	1,899	20,296
EU email	986	332,334
Math Overflow	24,818	506,550
Stack Overflow	2,601,977	63,497,050
Super User	194,085	1,443,339
Wikipedia talk pages	1,140,149	7,833,140

---

[3]All Gibbs sampling experiments were run on a quad-core (3.1 GHz) Dell desktop running Linux.

Table 4: MLEs on full datasets, and predictive log-likelihood for final 20% of edges based on MLEs fit to the first 80%, for three different BNTL models. Note that the uncoupled $\mathcal{PYP}(\theta, \tau)$ and $\text{Geom}(\beta)$ interarrival models have the same $\hat{\alpha}$ due to the factorization in (14).

| Dataset | Coupled $\mathcal{PYP}(\theta, \alpha)$ | | | $\hat{\alpha}$ | Uncoupled $\mathcal{PYP}(\theta, \tau)$ | | $\text{Geom}(\beta)$ | | |
	$(\hat{\theta}, \hat{\alpha})$	$\hat{\eta}$	Pred. l-l.		$(\hat{\theta}, \hat{\tau})$	Pred. l-l.	$\hat{\beta}$	$\hat{\eta}$	Pred. l-l.
Ask Ubuntu	(18080, 0.25)	1.25	-3.707e6	-2.54	(-0.99, 0.99)	-3.678e6	0.083	2.32	**-3.678e6**
UCI social network	(320.4, 4.4e-11)	–	-1.600e5	-4.98	(5.50, 0.52)	**-1.595e6**	0.016	2.10	-1.596e5
EU email	(113.6, 2.5e-14)	–	**-8.06e5**	-1.86	(113.6, 9.2e-10)	**-8.06e5**	0.001	2.00	-8.07e5
Math Overflow	(2575, 0.15)	1.15	-1.685e6	-6.62	(-0.97, 0.997)	-1.670e6	0.025	2.19	**-1.670e6**
Stack Overflow	(297600, 0.11)	1.11	-3.358e8	-8.94	(-1.0, 1.0)	-3.333e8	0.020	2.21	**-3.333e8**
Super User	(20640, 0.24)	1.24	-5.855e6	-4.19	(-0.996, 1.0)	**-5.775e6**	0.067	2.37	-5.775e6
Wikipedia talk pages	(14870, 0.54)	1.54	-3.074e7	-0.25	(-1.0, 1.0)	**-3.066e7**	0.073	2.10	-3.066e7

best. The two densest networks, the EU email and UCI social networks, exhibit arrival times that are sub-linear in n; as such, the $\mathcal{PYP}$ models fit best. Note that the coupled $\mathcal{PYP}$ model estimates $\hat{\alpha} \approx 0$, indicating the lack of a power law tail in the degree distribution. In the rest of the networks, the arrival times appear approximately linear in n. The $\text{Geom}(\beta)$ and uncoupled $\mathcal{PYP}$ model fit best. However, we note that in these cases the MLEs for the uncoupled $\mathcal{PYP}$ model are at the boundaries of the parameter range ($\hat{\theta} \approx -1, \hat{\tau} \approx 1$). This illustrates that although the uncoupled $\mathcal{PYP}$ model is more flexible than the coupled version, the underlying arrival time model cannot capture linear arrival time sequences without driving the parameters to the boundaries.

6 DISCUSSION

BNTL models are a useful tool to reason about asymptotic properties of a network. For example, the exponent of the asymptotic power law tail is a function of model parameters, which can be estimated from finite-size networks without dealing with the large fluctuations of heavy-tailed degree distributions in finite samples. Furthermore, the ability to capture the full range of power law exponents and sparsity levels within the same model class allows for model fitness comparisons using the same set of techniques, as in Section 5. We have designed a set of inference algorithms for these models; in doing so, we have made a large class of previously intractable models useful for statistical inference.

Future research directions. The full Gibbs sampler scales reasonably well to networks with thousands of vertices; in order to scale to larger networks, further work is needed. One possible approach is via Metropolis–Hastings with cheap joint proposals of the arrival times and the permutation, which may be able to take larger steps in sample space. A different direction is variational inference, though permutations pose a significant challenge in that context; recent work (Linderman et al., 2018) is a step in that direction.
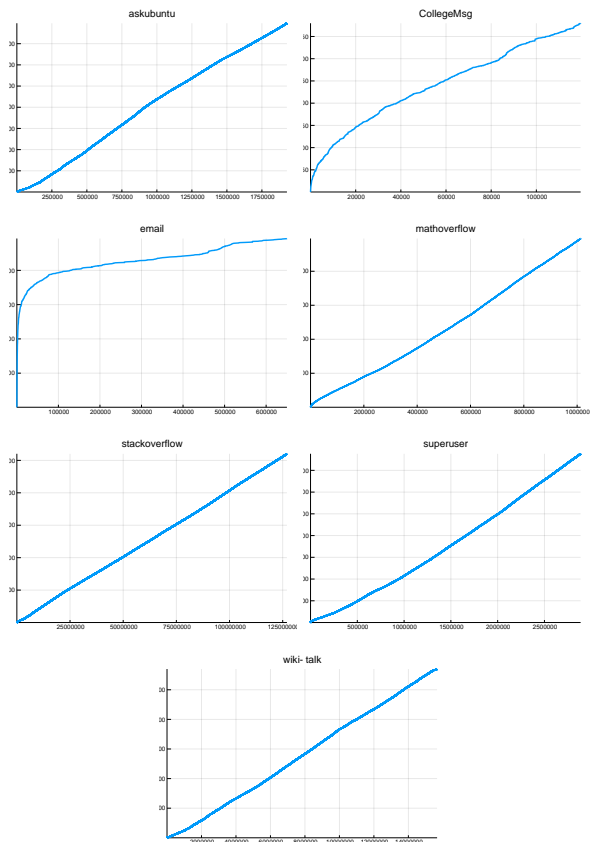
Figure 1: Arrival time sequences for SNAP data. Vertical axis is number of vertices, K_n; horizontal axis is n.

References

Aldous, D. J. (1981). "Representations for partially exchangeable arrays of random variables". In: *Journal of Multivariate Analysis* 11.4, pp. 581–598.

Barabási, A.-L. and R. Albert (1999). "Emergence of scaling in random networks". In: *Science* 186.5439, pp. 509–512.

Berger, N. et al. (2014). "Asymptotic behavior and distributional limits of preferential attachment graphs". In: *Ann. Probab.* 42.1, pp. 1–40.

Betancourt, B. et al. (2016). "Flexible Models for Microclustering with Application to Entity Resolution". In: *NIPS 29*, pp. 1417–1425.

Bingham, N. H., C. M. Goldie, and J. L. Teugels (1989). *Regular Variation*. Vol. 27. Encyclopedia of Mathematics and its Applications. Cambridge University Press.

Bloem-Reddy, B. and P. Orbanz (2017). "Preferential Attachment and Vertex Arrival Times". In: arXiv: 1710.02159 [math.PR].

– (2018). "Random Walk Models of Network Formation and Sequential Monte Carlo Methods for Graphs". In: *Journal of the Royal Statistical Society: Series B*. To appear.

Borgs, C. et al. (2016). "Sparse exchangeable graphs and their limits via graphon processes". In: arXiv: 1601.07134 [math.PR].

Cai, D., T. Campbell, and T. Broderick (2016). "Edge-exchangeable graphs and sparsity". In: *NIPS 29*, pp. 4242–4250.

Caron, F. and E. B. Fox (2017). "Sparse graphs using exchangeable random measures". In: *Journal of the Royal Statistical Society: Series B* 79.5, pp. 1–44.

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). "Power-Law Distributions in Empirical Data". In: *SIAM Review* 51.4, pp. 661–703.

Crane, H. and W. Dempsey (2017). "Edge exchangeable models for interaction networks". In: *Journal of the American Statistical Association*.

De Blasi, P. et al. (2015). "Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2, pp. 212–229.

Di Benedetto, G., F. Caron, and Y. W. Teh (2017). "Non-exchangeable random partition models for microclustering". In: arXiv: 1711.07287 [stat.ME].

Doksum, K. (1974). "Tailfree and Neutral Random Probabilities and Their Posterior Distributions". In: *Ann. Probab.* 2.2, pp. 183–201.

Gao, F. and A. van der Vaart (2017). "On the asymptotic normality of estimating the affine preferential attachment network models with random initial degrees". In: *Stochastic Processes and their Applications*.

Gnedin, A. and J. Pitman (2006). "Exchangeable Gibbs partitions and Stirling triangles". In: *Journal of Mathematical Sciences* 138.3, pp. 5674–5685. ISSN: 1573-8795.

Griffiths, R. C. and D. Spanò (2007). "Record Indices and Age-Ordered Frequencies in Exchangeable Gibbs Partitions". In: *Electron. J. Probab.* 12, pp. 1101–1130.

Hoover, D. N. (1979). *Relations on probability spaces and arrays of random variables*. Tech. rep. Institute of Advanced Study, Princeton.

Ishwaran, H. and L. F. James (2001). "Gibbs Sampling Methods for Stick-Breaking Priors". In: *Journal of the American Statistical Association* 96.453, pp. 161–173.

James, L. F. (2006). "Poisson calculus for spatial neutral to the right processes". In: *Ann. Statist.* 34.1, pp. 416–440.

Leskovec, J. and A. Krevl (2014). *SNAP Datasets: Stanford Large Network Dataset Collection*. http://snap.stanford.edu/data.

Linderman, S. W. et al. (2018). "Reparameterizing the Birkhoff Polytope for Variational Permutation Inference". In: *AISTATS 21*.

Neal, R. M. (2003). "Slice Sampling". In: *Ann. Statist.* 31.3, pp. 705–767.

Newman, M. E. J. (2005). "Power laws, Pareto distributions and Zipf's law". In: *Contemporary physics* 46.5, pp. 323–351.

Orbanz, P. and D. M. Roy (2015). "Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2, pp. 437–461.

Peköz, E. A., A. Röllin, and N. Ross (2017). "Joint degree distributions of preferential attachment random graphs". In: *Advances in Applied Probability* 49.2, pp. 368–387.

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Vol. 1875. Ecole d'Eté de Probabilités de Saint-Flour XXXII. Springer-Verlag Berlin Heidelberg.

Simon, H. A. (1955). "On a class of skew distribution functions". In: *Biometrika* 42.3-4, pp. 425–440.

Veitch, V. and D. M. Roy (2015). "The Class of Random Graphs Arising from Exchangeable Random Measures". In: arXiv: 1512.03099 [math.ST].

Walker, S. G. and P. Muliere (1997). "Beta–Stacy processes and a generalization of the Pólya-urn scheme". In: *Ann. Statist.* 25.4, pp. 1762–1780.

Wan, P. et al. (2017). "Fitting the linear preferential attachment model". In: *Electron. J. Statist.* 11.2, pp. 3738–3780.

Williamson, S. A. (2016). "Nonparametric Network Models for Link Prediction". In: *Journal of Machine Learning Research* 17.202, pp. 1–21.