Random Walk Models, Preferential Attachment, and Sequential Monte Carlo Methods for Analysis of Network Data

Benjamin Bloem-Reddy

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Random Walk Models, Preferential Attachment, and Sequential Monte Carlo Methods for Analysis of Network Data

Benjamin Bloem-Reddy

Networks arise in nearly every branch of science, from biology and physics to sociology and economics. A signature of many network datasets is strong local dependence, which gives rise to phenomena such as sparsity, power law degree distributions, clustering, and structural heterogeneity. Statistical models of networks require a careful balance of flexibility to faithfully capture that dependence, and simplicity, to make analysis and inference tractable. In this dissertation, we introduce a class of models that insert one network edge at a time via a random walk, permitting the location of new edges to depend explicitly on the structure of the existing network, while remaining probabilistically and computationally tractable. Connections to graph kernels are made through the probability generating function of the random walk length distribution. The limiting degree distribution is shown to exhibit power law behavior, and the properties of the limiting degree sequence are studied analytically with martingale methods. In the second part of the dissertation, we develop a class of particle Markov chain Monte Carlo algorithms to perform inference for a large class of sequential random graph models, even when the observation consists only of a single graph. Using these methods, we derive a particle Gibbs sampler for random walk models. Fit to synthetic data, the sampler accurately recovers the model parameters; fit to real data, the model offers insight into the typical length scale of dependence in the network, and provides a new measure of vertex centrality.

The arrival times of new vertices are the key to obtaining results for both theory and inference. In the third part, we undertake a careful study of the relationship between the arrival times, sparsity, and heavy tailed degree distributions in preferential attachment-type models of partitions and graphs. A number of constructive representations of the limiting degrees are obtained, and connections are made to exchangeable Gibbs partitions as well as to recent results on the limiting degrees of preferential attachment graphs.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This dissertation owes much to excellent mentors and colleagues at Columbia. I have learned a great deal from my advisor, Peter Orbanz, and I am grateful for his guidance and generosity over the past four years. I am particularly thankful for his patience as I made progress on this research, and for his thoughtful feedback on early drafts of papers and presentations. My time at Columbia has been enhanced by the Machine Learning Reading Group led by Dave Blei. I am grateful to the other participants for the weekly presentations and discussions, which proved stimulating and useful; in particular, Ari Pakman's presentation of Particle MCMC methods was a pedagogical masterpiece, and was the catalyst for much of the work in Chapter 4. My understanding of stochastic dependence in probability models and inference techniques is greatly influenced by long conversations with Rajesh Ranganath.

I would like to thank the committee members: John Cunningham, for his mentorship and collaboration; Dave Blei, for his generosity and collegiality; and Dan Roy and Harry Crane for their willingness to serve on the committee and their thoughtful feedback on earlier drafts.

The research presented in this dissertation was conducted during a time of rapid progress in this little corner of the field. I am grateful to Victor Veitch, Dan Roy, and Harry Crane for

discussing their work with me. I (and the field) benefited from the workshop on Networks, Random Graphs and Statistics at Columbia in May 2016, and from the workshops at the Isaac Newton Institute in July 2016 as part of the Theoretical Foundations for Statistical Network Analysis program.

It is a true privilege to be given the time and resources required to write this dissertation. I am grateful for the support from the Department of Statistics, financial and otherwise. For the care and thoughtfulness that keeps the department functioning, I am indebted to the inimitable Dood Kalicharan.

On a personal note, I am thankful to my family for their support and encouragement. This dissertation was possible because of them. It is my happy fortune that Leonora arrived before I started writing; her presence (and smiles) helped me keep proper perspective. Finally, and most importantly, I am thankful to Martha for a countably infinite number of things, her unwavering support and grace, and our long walks through Riverside Park among them.

Chapter 1

Introduction

Network data consist of a set of entities and the *interactions* between them. They arise in nearly every branch of science, from biology and physics to sociology and economics. Their ubiquity has prompted an ever-growing body of literature; empirical studies and probability models of random graphs date back to at least the first half of the 20th century, while the mathematical foundations of graph theory date back to Euler in 1735 (Kolaczyk, 2009; Goldenberg, Zheng, Fienberg, and Airoldi, 2010). Recent years have witnessed continued growth throughout varied applied fields, and a rapid expansion of methodological research in probability, statistics, and machine learning. As observed by Kolaczyk (2009), two of the main forces behind this growth are

> "(i) an increasing tendency towards a systems-level perspective in the sciences, away from the reductionism that characterized much of the previous century, and (ii) an accompanying facility for high-throughput data collection, storage, and management."

The systems-level perspective is widespread throughout science, as is the proliferation of data, and much of modern statistics and machine learning research focuses on addressing the challenges that arise in these settings.

Network data in particular present a unique challenge to statisticians. In contrast to many classical statistical problems, the patterns of the interactions are of primary interest rather than a nuisance for which to control. Moreover, there is ample evidence that the interactions are highly dependent. An oft-cited example is the high degree of transitivity in social networks (e.g. Holland and Leinhardt, 1971; Newman, 2009): because Paul and Alfred are friends, and Alfred and Edgar are friends, it is likely that Paul and Edgar are friends. Other widely observed phenomena, discussed in more detail below, cannot occur with non-negligible probability without a high degree of dependence between the interactions.

We consider a network to be represented as a sequence of growing graphs; a **statistical network model** is a family of probability distributions $\mathcal{P} = \{P_\theta : \theta \in \mathcal{T}\}$ on networks, parameterized by θ. An observation consists of a single network and is explained as either a network drawn from the model or a subset of such a network.[1] When designing a statistical network model, we are guided in part by the following objectives:

(i) **Faithfulness to salient properties of real data.** A model should aim to "establish a link with any theoretical knowledge about the system and with previous experimental work," and "[t]here should be consistency with known limiting behavior" (Cox and Hinkley, 1974, p. 5). In other words, the family $\mathcal{P}$ should include distributions that as-

[1]We assume that observations are *coherent* with their larger counterparts. That is, the manner in which they are generated, such as subsampling, respects the probabilistic structure of the data generating process. This is not always the case, and is an important consideration. See Shalizi and Rinaldo (2013), Orbanz and Roy (2015), Crane and Dempsey (2015a), and Veitch and Roy (2015).

sign non-negligible probability to networks with the characteristics deemed important by theory and experiment, and by reasoning about limiting properties.

What are the salient properties of networks that should be captured by statistical models? Two of the most widely studied are asymptotic properties and therefore cannot be observed directly. Nonetheless, there is strong theoretical and empirical evidence that networks are sparse and often have power law degree distributions (e.g. M. Faloutsos, P. Faloutsos, and C. Faloutsos, 1999; Mitzenmacher, 2003; Leskovec, Kleinberg, and C. Faloutsos, 2007; Clauset, Shalizi, and Newman, 2009; Newman, 2009, and references therein). Other properties include so-called "small worlds" (Travers and Milgram, 1969; Watts and Strogatz, 1998), hubs (Newman, 2009), and the aforementioned transitivity, or clustering, property. These empirical properties, though not well understood theoretically, should guide the development of models and methods for analyzing network data.

(ii) **Tractable analysis.** Whether or not a model succeeds in capturing relevant aspects of data is determined in part by analyzing the theoretical properties of the distributions in $\mathcal{P}$. Furthermore, successful models may provide conceptual insight into the underlying system that generated the data.

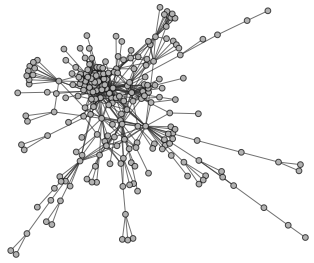
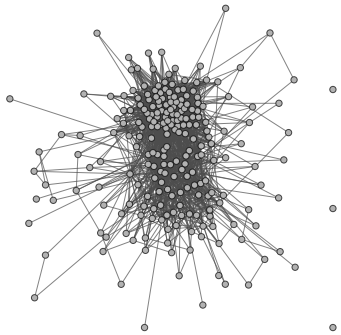
Each $P_\theta \in \mathcal{P}$ represents a network as a set of vertices and edges, possibly with labels or covariates, regarded as a system of random variables; P_θ is their joint distribution. Informally, for structure like the properties described in (i) to occur with non-negligible probability, the variables must be dependent. As a general rule, more dependence in a system of random variables leads to more complicated analysis. Alternately, as-

sumptions constraining dependence typically lead to more tractable analysis. In the network literature, a prototypical example is the basic preferential attachment (PA) model (de Solla Price, 1965; Barabási and Albert, 1999) and its more bespoke variations (e.g. Aiello, Chung, and Lu, 2001; Bianconi and Barabási, 2001; Cooper and Frieze, 2003; Borgs, Chayes, Daskalakis, and Roch, 2007). PA models use a simple mechanism for generating networks, in which a vertex participates in a new interaction with probability proportional to the number of its previous interactions. The intentional simplicity of PA models makes them amenable to probabilistic analysis: Power law degree distributions (Bollobás, Riordan, Spencer, and Tusnády, 2001), local weak limits (Berger, Borgs, Chayes, and Saberi, 2014), and the asymptotic distributional properties of the scaled degrees (Móri, 2005; Peköz, Ross, and Röllin, 2014) are notable examples. However, that simplicity constrains the flexibility needed to capture network properties other than degree statistics.

(iii) **Tractable inference procedures.** Estimating and performing inference on the model parameters, and using the results for secondary tasks such as prediction, should be feasible in a reasonable amount of computational time. In general, a model with less dependence between its constituent random variables requires less computation. As a result, the statistics and machine learning literature has largely focused on models in which the edges are conditionally independent. As we discuss in detail in this dissertation, restrictions on dependence between edges can lead to model misspecification. However, computation easily becomes intractable if too much dependence is present in the model, and some balance must be struck.

These objectives are often in tension with each other. On one hand, the model should allow for sufficient dependence to successfully capture the structural phenomena that arise in real networks; on the other, increased dependence in the model leads to difficult analysis and more complex inference procedures. The trade-offs required to balance these objectives are a central theme of this dissertation, and we will revisit them throughout.

The work in this dissertation is motivated by the question of how statistical network models handle stochastic dependence within a network. Many models constrain dependence either explicitly (e.g. Holland and Leinhardt, 1981; Bollobás, Janson, and Riordan, 2007), or as a consequence of other assumptions like exchangeability (e.g. Lloyd, Orbanz, Ghahramani, and Roy, 2012; Caron and Fox, 2015; Veitch and Roy, 2015; Borgs, Chayes, Cohn, and Holden, 2016; Crane and Dempsey, 2016; Cai, Campbell, and Broderick, 2016). As an example, the following networks are (a) a protein-protein interaction network (see Section 3.3), and (b) its reconstruction sampled from a graphon model (discussed in more detail in Section 2.3.1) fitted to (a) (Lloyd, Orbanz, Ghahramani, and Roy, 2012):

(a)  (b) 

The data set (a) contains pendants (several degree-1 vertices linked to a single vertex), isolated chains of edges, hubs, etc. For these to arise at random requires local dependence between edges on different length-scales, and they are conspicuously absent from (b). That is not a shortcoming of the method used to fit the model, but inherent to graphon models, since

they constrain edges to be conditionally independent given certain vertex-wise information: whether or not an edge is present in a particular realization of the network does not depend on the presence or absence of other edges.

The constraints on dependence are imposed for good reasons. At a basic level, they are an attempt to answer the following question, which captures the tension between the objectives above:

> *How do we simplify the probabilistic structure of P_θ, yet capture the relevant properties of the network?*

This type of question (and its inherent vagueness) is common throughout statistics and machine learning, where dependent, highly structured data from diverse fields such as text analysis (Srivastava and Sahami, 2009; Blei, 2012), image processing (Krizhevsky, Sutskever, and Hinton, 2012), neuroscience (Helmstaedter, 2015), genetics (Kohane, 2011; Libbrecht and Noble, 2015), medicine (Hripcsak and Albers, 2012; P. B. Jensen, L. J. Jensen, and Brunak, 2012), and recommendation systems (Salakhutdinov and Mnih, 2008) present similar challenges. Although it greatly simplifies analysis and inference when the network is broken into simpler components, structure in the data may be discarded in the process. If the discarded structure is relevant to statistical analysis, its absence constitutes a form of model misspecification.

This dissertation focuses on addressing these issues, and on better understanding the trade-offs involved when modeling structured, dependent data. To do so, we introduce a class of models that, rather than constraining dependence, express it on different length-scales. That dependence generates a range of interesting structures that are faithful to

various aspects of real networks. We also derive a number of theoretical properties and develop tractable inference procedures, both of which shed light on the interplay of the modeling objectives discussed above. A particular sequence of random variables emerges as crucial to theoretical analysis and inference. We undertake a detailed study of their properties and their influence on sparsity and power law degree distributions, and make connections to a number of models for partitions and graphs in the probability and statistics literature.

Organization

Chapter 2: Background. We briefly review previous work upon which this dissertation builds. The preliminaries contained in this chapter serve as a reference for later results, introduce notation, and provide a formal framework for the subsequent chapters. We discuss exchangeability and stochastic dependence, survey some existing models of network data, and review some results from spectral graph theory and sequential Monte Carlo methods.

Chapter 3: Random walk models of networks. We introduce a simple class of models that construct a network as a sequence of edges; each new edge depends on the entire existing edge structure via a random walk. The typical length of the random walk controls the typical length-scale of dependence in the network. Certain special properties of random walks on graphs are used to demonstrate that, for a particular subclass of models, the degree distributions behave much like those of preferential attachment models. In particular, they can exhibit power law behavior.

Using the methods developed in Chapter 4, the model is applied to data, demonstrating

how different parameters and latent variables can be usefully interpreted. Comparisons are made with other network models, and we discuss potential issues with comparisons between models that treat the dependence in the network differently.

Chapter 4: Inference methods for sequential models. Estimation and inference procedures are developed. When the entire edge sequence is observed, maximum likelihood can be used; we derive estimating equations for the entire class of models introduced in Chapter 3. When the edge sequence is only partially observed, including the case when only the final network is observed, the latent sequence must be imputed; we build on the work of Andrieu, Doucet, and Holenstein (2010) and develop Markov chain Monte Carlo (MCMC) methods for a wide class of sequential models satisfying a Markov property and a monotonicity property. MCMC sampling for the fully observed case is also demonstrated.

Chapter 5: Nested urn models of partitions and graphs. Crucial to the results of Chapters 3 and 4 is the sequence of times at which new vertices enter the graph. Chapter 5 studies the properties of preferential attachment-type partitions and graphs that arise by randomizing that sequence with different distributions. We prove almost sure convergence of the scaled degree sequence to a random limit, and that the asymptotic rate at which new vertices appear determines the proper scaling of the degree sequence. We make connections to exchangeable Gibbs partitions (Gnedin and Pitman, 2006) and neutral-to-the-left processes, and derive a number of constructive representations of the limit objects.

Chapter 2

Background

In this chapter, we review the previous work upon which the subsequent chapters build. Exchangeable random sequences and partitions are reviewed in Section 2.2. Those objects provide a conceptual backdrop for the study of networks, and the methods used to analyze them are adapted to more complicated situations in Chapter 5. In Section 2.3, we give an overview of the literature on network models, with a focus on the strengths and weaknesses of various models in the framework outline in Chapter 1. Section 2.4 presents some results from spectral graph theory that are relevant to the developments in Chapters 3 and 4, and Section 2.5 reviews the Sequential Monte Carlo and Markov Chain Monte Carlo methods that are necessary for the inference methods in Chapter 4.

2.1 Notation

Throughout this dissertation, we assume that random variables are defined on a common, abstract probability space $(\Omega, \mathcal{A}, \mathbb{P})$. As is standard, all random variables X are measurable

mappings into their state space, i.e. for an S-valued random variable X, $X : \Omega \to S$. Random variables are written uppercase, and their realized values are lowercase. Parameters are typically Greek letters, e.g. Θ denotes a random parameter and θ a particular realization. We use $\mathbb{N}$ to denote $\{0, 1, 2, \ldots\}$, and $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$. For a process indexed by $t \in \mathbb{N}$, $\mathcal{A}_t$ denotes the σ-algebra generated by the process up to and including t.

2.2 Exchangeable random sequences and the Pólya urn

Urn models are among the most well-studied probability models, forms of which appear in the work of Laplace, Bernoulli, and others (Mahmoud, 2008). The version of Eggenberger and Pólya (1923), often called the Pólya urn, is a basic urn model dating back at least to the work of Markov from 1905 to 1907, and Paul and Tatyana Ehrenfest in 1907 (Mahmoud, 2008). The simplest version of the Pólya urn starts with an urn containing one black ball and one white ball. At each step t, a ball is drawn uniformly at random from the urn, and returned to the urn along with an additional ball of the same color. Despite its simplicity, the urn and its related probability distributions exhibit a number of remarkable properties. A full treatment is beyond the scope of this chapter, but there are numerous good references, among them Johnson and Kotz (1977) and Mahmoud (2008). The basic model has been generalized in too many directions to cite here; some notable results include those of Blackwell and MacQueen (1973), Hoppe (1984), Pitman (1996), Janson (2006), and Bacallado, Favaro, and Trippa (2013).

We focus here on aspects of the basic Pólya urn that lend insight into more complicated models studied in the subsequent chapters of this work. First, consider the urn that begins

with some number $b_0 \geq 1$ black balls and $w_0 \geq 1$ white balls. Define $X_s = \mathbb{1}\{\text{draw } s \text{ is black}\}$, and $B_t := \sum_{s=1}^{t} X_t$ and $W_t := t - B_t$. Then

$$\mathbb{P}(X_1, X_2, \ldots, X_t) = \frac{\prod_{b=0}^{B_t-1}(b_0 + b)\prod_{w=0}^{W_t-1}(w_0 + w)}{\prod_{s=0}^{t-1}(b_0 + w_0 + s)} \tag{2.1}$$

$$= \frac{\Gamma(B_t + b_0)\Gamma(W_t + w_0)\Gamma(b_0 + w_0)}{\Gamma(b_0)\Gamma(w_0)\Gamma(t + b_0 + w_0)} \, . \tag{2.2}$$

Much can be learned from the form of these equations. The analysis now proceeds along two separate lines, rejoining in Theorem 2.3.

2.2.1 Predictive martingales

Given a sequence $X_1, \ldots, X_t$ and fixed $p_b, p_w \in \mathbb{N}$, consider the predictive probability of any particular further sequence of length $p := p_b + p_w$, $X_{t+1}, \ldots, X_{t+p}$, such that p_b of the elements are black, i.e. $B_{t+p} - B_t = p_b$:

$$\begin{aligned}\mathbb{P}(X_{t+1}, \ldots, X_{t+p} \mid X_1, \ldots, X_t) \quad &= \frac{\Gamma(B_t + p_b + b_0)\Gamma(W_t + p_w + w_0)}{\Gamma(B_t + b_0)\Gamma(W_t + w_0)} \frac{\Gamma(t + b_0 + w_0)}{\Gamma(t + p + b_0 + w_0)} \\ &:= Z_t(p_b, p_w) \, .\end{aligned}$$

Because $B_0 = W_0 = 0$, $Z_t(p_b, p_w)$ is well defined for all $t \geq 0$, if $p_b > -b_0$ and $p_w > -w_0$, though its interpretation as a predictive probability breaks down for negative or non-integer p_b, p_w. A different interpretation, valid for all $\mathbb{R}$-valued $p_b > -b_0$ and $p_w > -w_0$, is as the likelihood ratio for urns with different starting conditions: The numerator is the probability of seeing B_t black balls when the urn has $b_0 + p_b$ black balls and $w_0 + p_w$ white balls at $t = 0$; the denominator is the same probability when the urn starts with b_0 black balls and

w_0 white balls. This interpretation will prove useful in Chapter 5.

Observe that by Stirling's formula (e.g. Tricomi and Erdélyi, 1951),

$$Z_t(p_b, p_w) = \frac{(B_t + b_0)^{p_b}}{(t + b_0 + w_0)^{p_b}} \frac{(W_t + w_0)^{p_w}}{(t + b_0 + w_0)^{p_w}} (1 + O(t^{-1})) \; . \tag{2.3}$$

Therefore, if the limit $\lim_{t\to\infty} Z_t(p_b, p_w)$ exists, its expectation can be used to compute the joint (p_b, p_w)-th moment of the limiting proportions of black balls and white balls, ξ_b and ξ_w. Since p_b and p_w are arbitrary, the moments completely characterize the limiting distribution of (ξ_b, ξ_w), if they exist. The next proposition shows that the moments indeed exist, and how to compute them. Although the results are not new, the proof techniques can be adapted to more complicated situations, and they are used to establish many of the theoretical results in Chapters 3 and 5.

PROPOSITION 2.1. *For any fixed* $\mathbb{R}$*-valued* $p_b > -b_0/2$ *and* $p_w > -w_0/2$,

$$(t + b_0 + w_0)^{-1}(B_t + b_0) \xrightarrow{t\to\infty} \xi \quad \textit{almost surely,}$$

and

$$Z_t(p_b, p_w) \xrightarrow{t\to\infty} \xi^{p_b}(1 - \xi)^{p_w} \quad \textit{almost surely.}$$

Furthermore, ξ *is a* $\mathrm{Beta}(b_0, w_0)$ *random variable.*

PROOF. $Z_t(p_b, p_w)$ is clearly non-negative by definition. For any $t \in \mathbb{N}_+$, let $\mathcal{A}_t$ denote the

σ-algebra generated by the first t draws from the urn. Then

$$
\begin{aligned}
&\mathbb{E}[Z_{t+1}(p_b, p_w) \mid \mathcal{A}_t] \\
&\quad = \frac{\Gamma(B_t + p_b + b_0)\Gamma(W_t + p_w + w_0)}{\Gamma(B_t + b_0)\Gamma(W_t + w_0)} \frac{\Gamma(t + 1 + b_0 + w_0)}{\Gamma(t + 1 + p + b_0 + w_0)} \quad \cdots \\
&\qquad \times \mathbb{E}\left[1 + \frac{p_b \cdot \mathbb{1}\{X_{t+1} = 1\}}{B_t + b_0} + \frac{p_w \cdot \mathbb{1}\{X_{t+1} = 0\}}{W_t + w_0} \;\middle|\; \mathcal{A}_t\right] \\
&\quad = \frac{\Gamma(B_t + p_b + b_0)\Gamma(W_t + p_w + w_0)}{\Gamma(B_t + b_0)\Gamma(W_t + w_0)} \frac{\Gamma(t + 1 + b_0 + w_0)}{\Gamma(t + 1 + p + b_0 + w_0)} \Big(1 + \frac{p}{t + b_0 + w_0}\Big) \\
&\quad = Z_t(p_b, p_w) \; .
\end{aligned}
$$

This shows that $(Z_t(p_b, p_w), \mathcal{A}_t)_{t \geq 0}$ is a non-negative martingale with finite expectation for any $p_b > b_0$ and $p_w > w_0$. By the Martingale Convergence Theorem (e.g. Çinlar, 2011, Chapter V.4), $Z_t(p_b, p_w)$ converges almost surely. The factors of $Z_t(p_b, p_w)$ can be shown in the same way to have almost sure limits,

$$
(t + b_0 + w_0)^{-1}(B_t + b_0) \xrightarrow{\text{a.s.}} \xi_b \; ,
$$

and likewise for ξ_w. Based on this and on the asymptotics in (2.3), it must be that the limit of $Z_t(p_b, p_w)$ is $\xi_b^{p_b} \xi_w^{p_w}$.

By basic properties of the Gamma function (see, e.g. Hofstad, 2016, Chapter 8, p. 306),

$$
Z_t(p_b, p_w)^2 \leq \Big(\prod_{i \in \{b,w\}} \frac{\Gamma(2p_i + 1)}{\Gamma(p_i + 1)^2} \Big) Z_t(2p_b, 2p_w).
$$

But $Z_t(2p_b, 2p_w)$ has finite expectation for $2p_b > -b_0$ and $2p_w > -w_0$, so $Z_t(p_b, p_w)$ is bounded in L_2 and therefore converges in L_2 and also in L_1, for any $p_b > -b_0/2$ and

$p_w > -w_0/2$. From (2.3),

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{(B_t + b_0)^{p_b}}{(t + b_0 + w_0)^{p_b}} \frac{(W_t + w_0)^{p_w}}{(t + b_0 + w_0)^{p_w}}\right] = \mathbb{E}[\xi_b^{p_b} \xi_w^{p_w}] = \lim_{t\to\infty} \mathbb{E}[Z_t(p_b, p_w)] = Z_0(p_b, p_w)$$

$$= \frac{\Gamma(b_0 + p_b)\Gamma(w_0 + p_w)}{\Gamma(b_0)\Gamma(w_0)} \frac{\Gamma(b_0 + w_0)}{\Gamma(b_0 + w_0 + p)}$$

$$= \int_0^1 \xi^{p_b}(1 - \xi)^{p_w} d\nu_{\beta(b_0,w_0)}(\xi) \,, \tag{2.4}$$

where $d\nu_{\beta(b_0,w_0)}(\xi)$ is the density of a random variable Ξ with distirbution Beta(b_0, w_0), which identifies the distribution of $\xi_b = 1 - \xi_w = \xi$ as Beta(b_0, w_0). □

Remark. The proposition can be extended easily to urns with k colors, which converges in the limit to a Dirichlet random variable. The beauty of this technique is that it allows one to show convergence, obtain the rate of convergence, and compute the moments with a single family of martingales. Related martingales were used to prove results for extensions of the basic two-color Pólya urn in Freedman (1965) and Gouet (1989, 1993), and for the k color urn in Blackwell and Kendall (1964) and Gouet (1997). The technique was also used in Móri (2005) to analyze the degrees of preferential attachment trees. ◁

For the purposes of this chapter, the primary significance of Proposition 2.1 is that both predictive distributions and the **empirical distributions**, i.e. the limiting proportions, converge to the same random variable with a Beta distribution. This is not a coincidence; it is a special case of the more general result in Section 2.2.3.

2.2.2 Conditionally i.i.d. and mixed i.i.d. sequences

For the second line of analysis, we show that the urn sequence $X_1, X_2, \ldots$ is a mixture of **independent and identically distributed**, or i.i.d., random variables, a result due to de Finetti (1930). In order to do so, we recognize (2.2) as a ratio of Beta functions, leading to the identity[1]

$$\mathbb{P}(X_1, X_2, \ldots, X_t) = \int_0^1 \xi^{B_t}(1-\xi)^{t-B_t} d\nu_{\beta(b_0,w_0)}(\xi) \tag{2.5}$$

$$= \int_0^1 \Big(\prod_{s=1}^{t} \xi^{X_s}(1-\xi)^{1-X_s}\Big) d\nu_{\beta(b_0,w_0)}(\xi)\ , \tag{2.6}$$

which shows that the urn sequence $X_1, X_2, \ldots$ is a mixture of i.i.d. sequences, and it is equivalent in distribution to a conditionally i.i.d. sequence:

$$\Xi \sim \text{Beta}(b_0, w_0) \tag{2.7}$$

$$X_1, X_2, \ldots \mid \Xi = \xi \overset{\text{iid}}{\sim} \text{Bernoulli}(\xi)\ .$$

An equivalent sampling scheme is the following two-color **paintbox**:

(1) Sample $\Xi \sim \text{Beta}(b_0, w_0)$

(2) Partition the unit interval into subintervals $I_b = [0, \Xi)$ and $I_w = [\Xi, 1]$

(3) Sample $U_1, U_2, \ldots \overset{\text{iid}}{\sim} \text{Uniform}[0, 1]$

(4) Paint each ball s according to the color of U_s. That is, set $X_s = \mathbb{1}\{U_s \in I_b\}$.

[1]Not coincidentally, (2.2) is also known as the Beta-Bernoulli distribution; it is the posterior predictive distribution of Bernoulli observations with its conjugate prior, the Beta distribution.

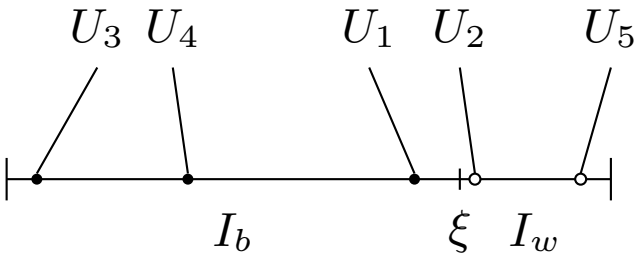


Figure 2.1: Sampling a conditionally i.i.d. binary sequence: Given $\Xi = \xi$, $X_s = \mathbb{1}\{U_s < \xi\}$.

Figure 2.1 shows a schematic of this process corresponding to the sequence $1, 0, 1, 1, 0$.

All of these results extend to more general sequences of random variables, as we discuss in the next two sections.

2.2.3 de Finetti's theorem

The convergence of the empirical distributions, the identity (2.6), and the conditional i.i.d. sampling representation (2.7) are special cases of more general results on random sequences whose distributions are invariant under finite permutations of $\mathbb{N}_+$. More precisely, a finite permutation of $\mathbb{N}_+$ is a bijective transformation $t \mapsto j_t$ such that $t = j_t$ for all but finitely many t.

Definition 2.2. A finite or infinite random sequence $X := (X_1, X_2, \ldots)$ in a measurable space $(S, \mathcal{S})$ is said to be **exchangeable** if its distribution is invariant to any finite permutation of the elements of X. That is,

$$(X_{j_1}, \ldots, X_{j_m}) \stackrel{d}{=} (X_1, \ldots, X_m) \,,$$

where $\stackrel{d}{=}$ denotes equality in distribution, for any collection $j_1, \ldots, j_m$ of distinct elements of the index set of X.

The concept of exchangeability is central to Bayesian statistical analysis, and we will return to it frequently throughout this work. Inspection of (2.1) shows that the urn sequence $X_1, X_2, \ldots, X_t$ is exchangeable for any t: Its distribution depends only on the sum $\sum_{s=1}^{t} X_s$, which is (trivially) unchanged by permutations of $X_1, \ldots, X_t$. This is just one example of an exchangeable binary sequence; remarkably, *any* exchangeable sequence of random variables has a unique representation like (2.6) and (2.7), a result known as **de Finetti's theorem**.

At a high level, de Finetti's theorem states that any infinite random sequence with values in S is exchangeable if and only if it is conditionally i.i.d. given a random variable Ξ with values in $\mathcal{M}(S)$, the space of probability measures on S. It also states that both the empirical distributions and the predictive distributions $\hat{P}_t := \mathbb{P}(X_{t+1} \mid X_1, \ldots, X_t)$ converge weakly to Ξ.

THEOREM 2.3 (de Finetti (1930, 1937), Hewitt and Savage (1955)). *Let $X = (X_1, X_2, \ldots)$ be an infinite sequence of random variables with values in a measurable Borel space $(S, \mathcal{S})$. X is exchangeable if and only if there is a random probability measure Ξ on S such that the elements $X_1, X_2, \ldots$ are conditionally i.i.d. with distribution Ξ. Moreover, for any sequence of sets $A_1, A_2, \ldots \in \mathcal{S}$,*

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2 \ldots) = \int_{\mathcal{M}(S)} \prod_{s=1}^{\infty} \xi(A_s) \nu(d\xi) \; ,$$

where ν, which uniquely determines the law of X, is the law of Ξ. Furthermore, both the empirical distributions and the predictive distributions converge weakly to Ξ, that is for

every $A \in \mathcal{S}$*,*

$$\frac{1}{t}\sum_{s=1}^{t} \mathbb{1}\{X_s \in A\} \xrightarrow{t\to\infty} \Xi(A) \quad \textit{almost surely,}$$

and

$$\hat{P}_t(A) = \mathbb{P}(X_{t+1} \in A \mid X_1, \ldots, X_t) \xrightarrow{t\to\infty} \Xi(A) \quad \textit{almost surely.}$$

This is now a standard result, proof of which can be found in numerous places. See, e.g. Kallenberg (2005, Chapter 1.1) for a few different approaches and an illuminating discussion of the equivalence for infinite sequences of exchangeability and another probabilistic symmetry, contractability. The latter result, on the convergence of the predictive distributions, is somewhat less standard; see Fortini, Ladelli, and Regazzini (2000), and Fortini and Petrone (2012).

2.2.4 Exchangeable random partitions and Kingman's paintbox

Exchangeable models of partitions have been studied extensively in the probability literature (e.g. Kingman, 1978a,b; Pitman, 1996, 2006), and have been used as statistical models in a range of applications, most notably forming the basis of clustering and mixture models (Pitman, 2006; Hjort, Holmes, Müller, and Walker, 2010; De Blasi et al., 2015). A **partition** of $\mathbb{N}_+$

$$\pi = (A_1, A_2, \ldots)$$

divides $\mathbb{N}_+$ into a possibly infinite number of non-overlapping subsets $A_j \subset \mathbb{N}_+$, called **blocks**. We assume the blocks to be ordered by their least elements. Denote by $\widetilde{\mathcal{P}}_t$ the space of partitions of $[t] := \{1, 2, \ldots, t\}$ ordered by their least elements. A finite partition is regarded as the **restriction** to the first t elements of a partition of $\mathbb{N}_+$: For any $t < t' \in \mathbb{N}_+$, we obtain π_t from $\pi_{t'}$ by deleting the elements $t+1, \ldots, t'$ and removing any empty blocks.

For example, a partition of [12] might be $\pi_{12} = (\{1, 2, 4, 5, 12\}, \{3, 7, 8\}, \{6, 10\}, \{9, 11\})$, and its unique restriction to [8] is $\pi_8 = (\{1, 2, 4, 5\}, \{3, 7, 8\}, \{6\})$. A block with one element is called a **singleton**.

A *random* partition Π of $\mathbb{N}_+$ is a partition-valued random variable. In this dissertation, we only consider distributions on Π for which the finite-dimensional distributions are **coherent**. More precisely, let $\mathcal{T}_j^{t+1}$ be an operator that acts on partitions by inserting $t+1$ into the j-th block, i.e. for $\Pi_t := \{A_{1,t}, \ldots, A_{k,t}\}$,

$$\mathcal{T}_j^{t+1}\Pi_t := \begin{cases} \{A_{1,t}, \ldots, A_{j,t} \cup (t+1), \ldots, A_{k,t}\} & \text{for } 1 \leq j \leq k \\ \{A_{1,t}, \ldots, A_{k,t}, 1\} & \text{for } j \geq k+1 \end{cases} .$$

Coherence requires that

$$\mathbb{P}(\Pi_t = \{A_1, \ldots, A_k\}) = \sum_{j=1}^{k+1} \mathbb{P}(\mathcal{T}_j^{t+1}\Pi_t) \, . \tag{2.8}$$

We call the sequence $(\Pi_t)_{t \geq 1}$ a **partition process**.

Π is said to be exchangeable if the probability distribution of Π_t is invariant under the natural action of the symmetric group of permutations of $[t]$, for each $t \in \mathbb{N}_+$ (Pitman, 2006). Exchangeability implies that the distribution of Π_t depends only on the sizes of the

blocks, that is

$$\mathbb{P}(\Pi_t = \{A_1, A_2, \ldots, A_k\}) = p(|A_1|, \ldots, |A_k|) \ ,$$

for some *symmetric* function p on k-tuples of non-negative integers that sum to t (Pitman, 1995). Letting t and k vary defines a function $p : \cup_{k=1}^{\infty} \mathbb{N}^k \to [0, 1]$, called the **Exchangeable Probability Partition Function** (EPPF).

A random partition Π of $\mathbb{N}_+$ can be formed from a random sequence $X_1, X_2, \ldots$ by assigning the index of each element of the sequence to a block of Π. The most natural assignment is by equivalence, that is $(X_t = X_{t'}) \iff (t, t' \in A_j)$ for some j. Clearly, if $X_1, X_2, \ldots$ is exchangeable, then so is Π. Kingman (1978a,b) showed that every exchangeable random partition has the following **paintbox** representation, depicted in Figure 2.2:

(1) Sample a random sequence $\mathbf{C} = (C_1, C_2, \ldots)$ of scalars $C_j \in [0, 1]$, which satisfy $C_1 \geq C_2 \geq \ldots$ and $\sum_j C_j \leq 1$. Define $W_k := \sum_{j=1}^{k} C_j$.

(2) Partition the unit interval into sub-intervals $I_j = [w_{j-1}, w_j)$ and $I_\infty = (1 - w_\infty, 1]$.

(3) Sample $U_1, U_2, \ldots \overset{\text{iid}}{\sim} \text{Uniform}[0, 1]$.

(4) Form Π by assigning $t \in \mathbb{N}_+$ to block A_j if $U_t \in I_j$. Assign to its own block every t for which $U_t \in I_\infty$.

The name paintbox comes from viewing each C_j as representing a different color; the observations $\{t \ : \ U_t \in I_j\}$ are painted with the color C_j. The interval I_∞, if it has non-zero measure, represents a "continuum of colors" (Kingman, 1978b), also known as dust (Bertoin, 2006); each $U_t \in I_\infty$ forms a singleton block of a distinct color. An extension of the paintbox

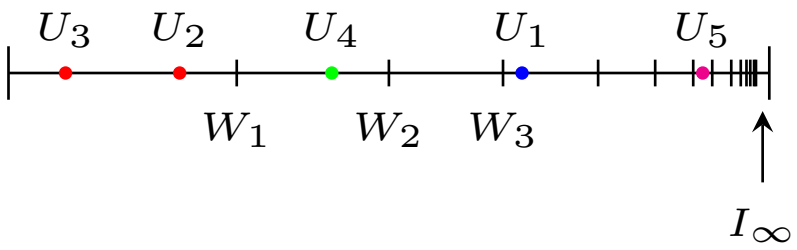


Figure 2.2: Sampling from a paintbox distribution with random sequence $\mathbf{C} = (C_1, C_2, \ldots)$. Two numbers $t, t' \in \mathbb{N}_+$ are assigned to the same block A_j of the partition Π if the uniform variables U_t and $U_{t'}$ are in the same interval I_j.

sampling scheme will play a role in Chapter 5.

2.3 Models of network data

The most common representation of a network is a graph. A graph $G = (\mathbf{V}, \mathbf{E})$ consists of a set $\mathbf{V}$ of **vertices**,[2] and a set $\mathbf{E}$ of **edges**. We denote by $|\mathbf{V}|$ the cardinality of $\mathbf{V}$, i.e. the number of vertices in G, and likewise for $\mathbf{E}$. As an illustrative example, consider an online social network. Users are represented as vertices and interactions between users form the edges. A **simple graph** consists of $\{0, 1\}$-valued undirected edges; in the social network example, an edge might represent whether or not two users are friends. Typically, self-edges are not allowed in simple graphs. In a **multigraph**, edges can take values in $\mathbb{N}$, representing, say, the number of interactions between two users. In a more general form, a **weighted graph**, edges might be $\mathbb{R}$-valued, and in all of these cases the edges might be directed. Furthermore, there may be covariate information attached to the vertices and the edges; for example, the physical locations of two users and the time at which an interaction between them occurred. A relatively newer area of research is on multidimensional or multilayer

[2]The network science literature typically uses the word **nodes**, but we use the language of graph theory in this dissertation.

networks, with different layers representing different types of interactions. See Kivelä et al. (2014) for a recent review. It is important to note that in all of these representations, there is often some loss of information, possibly because it was not observed or not recorded, or because it was discarded during the transformation of a network into its representation as a graph.

In this dissertation, only undirected simple graphs and multigraphs are considered. (Directed versions of results are typically straightforward adaptations of the undirected versions.) When discussing a property that applies to both simple and multigraphs, or when the type of graph is clear from the context, we will simply use the term graph. When a network is represented as a graph, a statistical network model is formulated as a probability model on graphs, known as a **random graph model**.

The literature on models of random graphs is far too extensive to cover here; we restrict our attention to models with properties that motivate or enhance our understanding of the models proposed in Chapter 3. For broader views of the literature, see the surveys of Goldenberg, Zheng, Fienberg, and Airoldi (2010), Hunter, Krivitsky, and Schweinberger (2012), and Orbanz and Roy (2015) and the text by Kolaczyk (2009), which focus on statistical models; the texts of Durrett (2006) and Hofstad (2016) give thorough overviews of the extensive probabilistic literature, particularly for dynamic or evolutionary models. Newman (2009) takes a network science perspective. This section is a high-level overview, delving only deeply enough to provide insight relevant to this dissertation. Models are grouped roughly according to whether or not they exhibit some form of exchangeability.

We view $(G_t)_{t\in\mathbb{T}}$ as a stochastic process on $(\mathcal{G}_t)_{t\in\mathbb{T}}$, the space of (possibly labeled) graphs indexed by a totally ordered (by the relation $\leq$) set $\mathbb{T}$. Generally, we assume that

$G_s \subseteq G_t$ for $s \leq t$. For the models studied in detail in this dissertation, $\mathbb{T} = \mathbb{N}_+$; some of the models discussed in this section have $\mathbb{T} = \mathbb{R}_+$. In either case, $(G_t)_{t>0}$ is a well-defined stochastic process. If a model is defined on labeled graphs, then we assume that $(G_t)_{t>0}$ is a set of labeled graphs, unless explicitly stated otherwise. For example, a multigraph constructed from a sequence of interactions $X_1, X_2, \ldots$ may require that each multi-edge $e_{ij} \in \mathbb{N}$ have a vector of labels $\ell_{ij} \in \mathbb{N}_+^{|e_{ij}|}$ specifying which elements of the interaction sequence are represented by e_{ij}. $(G_t)_{t>0}$ is the labeled multigraph sequence $(\mathbf{V}_t, \mathbf{E}_t, \boldsymbol{\ell}_t)_{t>0}$, with label set $\boldsymbol{\ell}_t$.

A random graph model is a family of distributions $\mathcal{P} = \{P_\theta \,:\, \theta \in \mathcal{T}\}$ on such processes. A basic property of a random graph model is its index set, which typically corresponds to the size of the graph. Models from the statistics literature traditionally have been indexed by the number of vertices, i.e. $\mathbb{T} = \mathbb{N}_+$, where t is the number of vertices of G_t. (This convention likely can be traced to the literature on social networks, where much of network analysis has roots (e.g. Holland and Leinhardt, 1971, 1976, 1977).) More recent models index the process by the number of edges (e.g. Crane and Dempsey, 2015a, 2016; Williamson, 2016; Cai, Campbell, and Broderick, 2016, Chapter 3 of this dissertation). As those papers discuss, the treatment of interactions as the data points, or statistical units, resolves some of the issues that arise when viewing G_t as a partial observation of a larger graph. To paraphrase, if vertices are the statistical unit, it is implicitly assumed that G_t contains all interactions between its vertices; a graph can only grow via the observation of new vertices, but not via additional interactions. This is plainly unrealistic in many modeling situations, and there are implications for subsampling and for prediction (see Orbanz and Roy, 2015; Crane and Dempsey, 2015a, 2016; Williamson, 2016). Not every model has an index set

that corresponds to a deterministic graph quantity. For example, the index set of the recent models of Caron and Fox (2015), Veitch and Roy (2015), and Borgs, Chayes, Cohn, and Holden (2016) is $\mathbb{T} = \mathbb{R}_+$, and corresponds to the expected number of edges, i.e. $\mathbb{E}[G_s] = s$.

A basic representation of a graph is as an **adjacency matrix** A, with elements $[A]_{ij} = w(v_i, v_j)$, where $w(v_i, v_j)$ is the weight of an edge between vertex i and vertex j. For simple graphs, the range of w is $\{0, 1\}$; for multigraphs it is $\mathbb{N}$. The **degree** $\deg(v)$ of a vertex v is the sum of the weight of all of its edges, that is $\deg(v) := \sum_u [A]_{vu}$. The **volume** of a graph is the sum of its degrees, $\mathrm{vol}(G) := \sum_{v \in \mathbf{V}} \deg(v)$.

Of the many statistics used to quantify graph properties, much attention has been paid recently to the following two:

- **Sparsity.** Roughly, the average degree grows more slowly than the number of edges needed for a complete graph on the same number of vertices. Let $N_e(t)$ and $N_v(t)$ denote the number of edges and vertices, respectively, in G_t. For $1 \leq \varepsilon < 2$, we call a graph sequence $(G_t)_{t>0}$ ε-sparse if

$$\limsup_{t \to \infty} \frac{N_e(t)}{N_v(t)^\varepsilon} = c_\varepsilon > 0 \,. \tag{2.9}$$

  If $\varepsilon \geq 2$, the network is called *dense*.

- **Power law degree distribution.** Let $m_d(t)$ be the number of vertices of degree d in G_t. The **degree distribution** is the normalized histogram vector $N_v(t)^{-1}(m_{1,t}, m_{2,t}, \dots)$ of the vertex degrees in G_t. A graph sequence $(G_t)_{t>0}$ exhibits a power law degree

distribution with exponent $\eta > 1$ if

$$p_d(t) := \frac{m_d(t)}{N_v(t)} \overset{t\uparrow\infty}{\sim} L(d)d^{-\eta} \quad \text{for all large } d \text{ as} \quad t \to \infty\,, \tag{2.10}$$

for some slowly varying function $L(d)$, that is, $\lim_{x\to\infty} L(rx)/L(x) = 1$ for all $r > 0$ (e.g. Feller, 1971; Bingham, Goldie, and Teugels, 1989), and where $a(t) \overset{t\uparrow\infty}{\sim} b(t)$ indicates $\lim_{t\to\infty} a(t)/b(t) \to 1$. Note that $L(d)$ controls the shape of the distribution, but not the upper tail; hence, variations in finite sample behavior and the lower tail are captured by $L(d)$.

On the basis of theory and extrapolation of empirical evidence, it is widely believed that real networks are often sparse and exhibit power law degree distributions (e.g. Barabási and Albert, 1999; M. Faloutsos, P. Faloutsos, and C. Faloutsos, 1999; Leskovec, Kleinberg, and C. Faloutsos, 2007; Clauset, Shalizi, and Newman, 2009; Newman, 2009). Many real networks are estimated to exhibit $\eta > 2$ (Chung and Lu, 2006; Clauset, Shalizi, and Newman, 2009), though there are examples with $1 < \eta < 2$ (Clauset, Shalizi, and Newman, 2009; Crane and Dempsey, 2015b). As we discuss in Section 5.5, for models of edge sequences generated with a PA-type mechanism, exchangeable and non-exchangeable models appear to be complementary in terms of the levels of sparsity and the range of possible power law exponents appearing in the degree distribution.

2.3.1 Models based on probabilistic symmetry

Vertex-exchangeable models. Of the models in the statistics and machine learning literature, many fall in this category, also known as graphon models. Examples include

the stochastic blockmodel (SBM) when class structure is unobserved (e.g. Nowicki and Snijders, 2001; Airoldi, Blei, Fienberg, and Xing, 2008) and its Bayesian nonparametric version (Kemp, Tenenbaum, T. L. Griffiths, Yamada, and Ueda, 2006; Xu, Tresp, Yu, and Kriegel, 2006), and others (Hoff, 2008; Miller, Jordan, and T. L. Griffiths, 2009; Roy and Teh, 2009; Leskovec, Chakrabarti, Kleinberg, C. Faloutsos, and Ghahramani, 2010; Lloyd, Orbanz, Ghahramani, and Roy, 2012; Zhou, 2015).

As the name implies, the graph process is indexed by the number of vertices. A graph is **vertex-exchangeable** if its distribution is invariant to permutations of the vertex labels. More precisely, let A represent the adjacency matrix of a random graph G on vertex set $[t]$, so that the matrix entry $[A]_{ij} = 1$ if there is an edge between vertices i and j, and $[A]_{ij} = 0$ otherwise. Then A is vertex-exchangeable if

$$([A]_{ij}) \stackrel{d}{=} \left([A]_{\sigma(i)\sigma(j)}\right) \tag{2.11}$$

for all permutations σ of $[t]$. The Aldous–Hoover theorem (Hoover, 1979; Aldous, 1981; Kallenberg, 2005), which yields a de Finetti-style representation for exchangeable arrays (up to equivalence classes), provides an attractive theoretical foundation for such models.

All vertex-exchangeable models of simple graphs can be represented by a random function on the unit square known as a graphon, $W : [0,1]^2 \to [0,1]$, from which $G(t, W)$, with adjacency matrix A, is generated as follows:

$$W \sim \nu$$

$$U_i \overset{\text{iid}}{\sim} \text{Uniform}[0,1], \quad i \leq t$$

$$[A]_{ij} \mid W, U_i, U_j \overset{\text{ind}}{\sim} \text{Bernoulli}(W(U_i, U_j)), \quad i < j \leq t \,. \tag{2.12}$$

W thus defines a distribution over graphs, $P_W(G_t) := \mathbb{P}(G_t \mid W)$. Since the collection of random variables (U_i) are i.i.d. and independent of W, P_W is invariant under any permutation of the vertex set. Importantly, the edges $[A]_{ij}$ are conditionally independent given W and (U_i).

These properties still obtain if we let $t \to \infty$. The Aldous–Hoover theorem states that W plays a role analogous to Ξ in Theorem 2.3: W is the weak limit of the empirical distributions (up to equivalence classes defined by weak isomorphisms of W (Diaconis and Janson, 2007; Lovász, 2013)), and given W, $G_1, G_2, \ldots$ can be sampled by the conditionally independent edge process described by (2.12). Furthermore, any exchangeable infinite random graph can be represented as a mixture of $G(\infty, W)$. See Diaconis and Janson (2007), Lovász (2013), and Orbanz and Roy (2015) for details.

However, as discussed in the introduction, graphon models constrain the dependence between edges and may be misspecified for many network analysis problems. Various symptoms of this problem have been noted (e.g. Borgs, Chayes, Cohn, and Zhao, 2014; Orbanz and Roy, 2015). For example, the generated graph is dense or empty, and unless it is k-partite or disconnected, the distance between any two vertices in an infinite vertex-exchangeable graph is almost surely 1 or 2. Although sparsity has been the most widely noted issue, the more fundamental culprit is the conditional independence of the edges: A graphon can encode any fixed pattern on some number t of vertices, but this pattern then occurs on every possible subgraph of size t with fixed probability (Diaconis and Janson, 2007).

Edge-exchangeable models. Graphs formed from an exchangeable sequence of edges are known as **edge-exchangeable** (Crane and Dempsey, 2015a, 2016; Williamson, 2016; Cai, Campbell, and Broderick, 2016). The models introduced in Chapters 3 and 4 are also formed from a sequence of edges. However, the sequence is not exchangeable. The difference is important; in the exchangeable case, much of the content Section 2.2 on exchangeable sequences applies to edge-exchangeable graphs directly or can be refined to apply. In particular, for a non-trivial graph, the edges are derived from a random partition of $\mathbb{N}_+$ and are conditionally i.i.d. given a random probability measure analogous to Kingman's paintbox representation. A number of earlier models such as the configuration model (see, e.g. Bender and Canfield, 1978; Bollobás, 1980; Chung and Lu, 2002, 2003; Newman, 2009; Chatterjee, Diaconis, and Sly, 2011; Riordan, 2012) are similar (though not equivalent) to edge-exchangeable graphs in which the random probability measure on edges factorizes into a product of measures on vertices.

A number of edge-exchangeable models that generate sparse graphs exhibiting power law degree distributions have been proposed (Crane and Dempsey, 2015a, 2016; Cai, Campbell, and Broderick, 2016). Despite addressing some of the shortcomings of vertex-exchangeable models, edge-exchangeable models still restrict dependence between edges in a way that may be misspecified for many network analysis problems.

$\mathbb{R}_+^2$-exchangeable models. A class of models proposed by Caron and Fox (2015) and studied in more depth by Veitch and Roy (2015), Borgs, Chayes, Cohn, and Holden (2016), and Janson (2016) treats a random graph as the partial realization of a point process in $\mathbb{R}_+^2$, the distribution of which is invariant under measure-preserving transformations of $\mathbb{R}_+^2$. The

index set is $\mathbb{R}$; the size of a graph is a stochastic function of t. It is not clear if inference for general models in this class is computationally tractable, and there are open questions as to the interpretation of some aspects of the model. Veitch and Roy (2016) makes progress on some of these questions.

One exception is the following special case for generating multigraphs studied by Caron and Fox (2015):

(1) Sample a purely atomic random measure $\Xi_v = (\Theta, \mathbf{C})$ on $\mathbb{R}_+^2$, with atom weights $\mathbf{C} = (C_1, C_2, \ldots)$.

(2) Construct a measure on pairs of vertices (i, j) as the direct product $\Xi_e = \Xi_v \times \Xi_v$. Denote by $\Xi_e(t)$ the restriction of Ξ_e to $[0, t] \times [0, t]$, with mass $M(t)$, and $\Xi_e^*(t) := \Xi_e(t)/M(t)$, such that the probability of sampling an edge e_{ij} is $C_i C_j / M(t)$.

(3) Sample a total number of edges for the network, $N_e \sim \text{Poisson}(M(t))$.

(4) Draw N_e i.i.d. samples from $e_i \sim \Xi_e^*(t)$.

(5) Form the multigraph from the edges $(e_i)_{i=1}^{N_e}$.

The final two steps make clear that conditioning on $\Xi_e(t)$ yields an i.i.d. sequence of edges. Thus, an edge-exchangeable model with directing measure $\Xi_e^*(t)$ generates graphs that are equal in distribution to graphs sampled as above, when both have N_e edges. This equivalence, along with the factorizable form of Ξ_e enabled an MCMC sampling scheme based on degree statistics to be used in Caron and Fox (2015). See also Herlau, Schmidt, and Mørup (2016) for a block-structured version.

2.3.2 Sequential models of network formation

The non-exchangeable random graph models considered in this dissertation form graphs from a sequence of edges and vertices. To date, the vast majority of such models have been some type of reinforcement process. Reinforcement processes have appeared in numerous models of self-organizing systems in which some large-scale property is attained from small-scale interactions. See Pemantle (2007) for a review. Importantly, limiting properties are not explicitly modeled by the small-scale interactions; rather, there is an "emergence of macro-structure" (Arthur, Ermoliev, and Kaniovski, 1987). The most well-known example is the Pólya urn, from which the limiting proportions emerge. Reinforcement processes also have been successful at modeling power law distributions in a range of settings; see Mitzenmacher (2003) for some examples. As discussed in the introduction, power law degree distributions have been observed in many real networks. As such, it is natural to consider models based on reinforcement processes. We review some previous work here, and undertake a closer study in Chapter 5.

Preferential attachment models. Motivated by the power law degree distributions observed in many real networks, Barabási and Albert (1999) (BA) proposed a model based on a simple reinforcement process that they called **preferential attachment** (PA). The BA model starts with an arbitrary configuration of m_0 non-isolated vertices and proceeds as follows: at each step $t+1$, attach a new vertex with ℓ edges to the graph G_t; for each of the ℓ new edges, sample an existing vertex $v \in \mathbf{V}(G_t)$ for attachment with probability proportional to the degree of v. Vertices with higher degree are more likely to have new

edges attached, i.e. "the rich get richer." BA random graphs were shown to have a power law degree distribution with exponent $\gamma_{BA} = 3$ (Bollobás, Riordan, Spencer, and Tusnády, 2001). Many variations of the basic preferential attachment scheme have been formulated and studied; recent work has explored representations of their limits. See, e.g. Móri (2005), Athreya, Ghosh, and Sethuraman (2008), Peköz, Ross, and Röllin (2014), and James (2015).

Although modeling local dependence in real networks was not the intention of the BA model, it is instructive to consider how it falls short as such a model. It captures the overall scaling of the degree distribution, but the reinforcement rule in the BA model is too simplistic to model real networks. At each step t, the degree sequence of G_t is sufficient to predict the distribution of G_{t+1}. The mechanism for inserting edges is insensitive to rearrangements of the existing edges as long as the degree sequence is maintained, and therefore all dependence between edges arises via the degrees. While the degrees may capture some of the dependence, it is unrealistic to assume that real network structure arises only from the degree sequence. In this respect, PA models have much in common with factorizable edge-exchangeable models. We explore the connection in Chapter 5.

Other sequential models. There are a number of other models that are motivated by an observed property of real networks, such as small-world models (Watts and Strogatz, 1998). Other sequential models for network evolution have also been proposed, such as copying models (Chung, Lu, Dewey, and Galas, 2003). In general, however, such models either are not flexible enough to be good models of network data or have proved inferentially intractable, or both. Notable exceptions are methods proposed in the evolutionary biology literature (Wiuf, Brameier, Hagberg, and Stumpf, 2006; Thorne and Stumpf, 2012; Wang,

Jasra, and De Iorio, 2014), which we discuss further in Section 4.5.

2.4 Random walks and spectral graph theory

Spectral graph theory is a field of math that focuses on the eigenvalue spectrum of various matrices that are used to represent graphs. Although considered a field of pure mathematics, with connections to differential geometry and algebra, the study of graph spectra has applications in physics, chemistry, and other areas of science. Closer to the subject area of this dissertation, graph spectra have been used to study the mixing properties of Markov chains on discrete state spaces (e.g. Aldous, 1983; Boyd, Diaconis, Parrilo, and Xiao, 2009; Levin, Peres, and Wilmer, 2009).

In this section, the **random walk** matrix $\mathbf{W}$ and the **normalized graph Laplacian** matrix $\mathbf{L}$ are the objects of focus. The standard reference is Chung (1997), from which this section borrows heavily. Denote by $D = \text{diag}(\deg(v_1), \ldots, \deg(v_n))$ the diagonal degree matrix of a graph on n vertices. For a graph with adjacency matrix A, $\mathbf{W}$ and $\mathbf{L}$ are defined as

$$\mathbf{W} := D^{-1}A = \mathbb{I} - D^{-1/2}\mathbf{L}D^{1/2} \tag{2.13}$$

$$\mathbf{L} := D^{1/2}(\mathbb{I} - \mathbf{W})D^{-1/2} = \mathbb{I} - D^{-1/2}AD^{-1/2}\,, \tag{2.14}$$

where $\mathbb{I}$ is the identity matrix of dimension equal to the number of vertices in G. By convention, $[D^{-1}]_{ii} = 0$ for any isolated vertices v_i.

$\mathbf{W}$ is called the random walk matrix because it is the transition matrix for random walks: $[\mathbf{W}]_{ij} = w(v_i, v_j)/\deg(v_i)$. For a random walk started on v_i, the probability of

reaching v_j after k steps is $[\mathbf{W}^k]_{ij}$.

$\mathbf{L}$ derives its name from its interpretation as the discrete analogue to the Laplacian operator ∇^2 on continuous spaces. Consider a function $g : \mathbf{V} \to \mathbb{R}$ such that $g(u)$ is the value of g on vertex u, and g is a column vector in $\mathbb{R}^n$. Denote by $\mathbf{1}$ the vector of all ones. An alternative way of defining $\mathbf{L}$ is as an operator on the space of functions $g : \mathbf{V} \to \mathbb{R}$. $\mathbf{L}$ is the unique operator that satisfies

$$\begin{aligned}\langle g, \mathbf{L}g\rangle &= -\frac{1}{2}\sum_{u \bullet\!-\!\bullet v} w(u,v)\Big(\frac{g(u)}{\sqrt{\deg(u)}} - \frac{g(v)}{\sqrt{\deg(v)}}\Big)^2 \\ &= -\frac{1}{2}\sum_{u \bullet\!-\!\bullet v} w(u,v)(f(u) - f(v))^2 \quad \text{for} \quad f = D^{-1/2}g\end{aligned}$$

where $\langle g_1, g_2\rangle = g_1' g_2 = \sum_u g_1(u)g_2(u)$ is the inner product on $\mathbb{R}^n$ and $u \bullet\!-\!\bullet v$ indicates that $w(u,v) > 0$; the sum is over all edges, i.e. over all unordered pairs $u \bullet\!-\!\bullet v$. This representation makes clear the analogy to ∇^2: $\mathbf{L}$ quantifies the "smoothness" of g through its local variations, just as $\langle f, \nabla^2 f\rangle_\Omega$ quantifies the local variations of a function f on some continuous space Ω. The intuition is made precise in the following:

THEOREM 2.4 (Spectral properties of $\mathbf{L}$ and $\mathbf{W}$). *$\mathbf{L}$ is a symmetric, positive semidefinite matrix with eigenvalues $(\sigma_i)_{i=1}^n$ and eigenvectors $(\boldsymbol{\psi}_i)_{i=1}^n$, which have the following properties:*

(i) The eigenvalues satisfy $0 \le \sigma_i \le 2$, and the number of eigenvalues equal to zero is the number of connected components of G.

(ii) $\boldsymbol{\psi}_1 = D^{1/2}\mathbf{1}/\sqrt{vol(G)}$ is an eigenvector with the eigenvalue $\sigma_1 = 0$.

(iii) Denote by $\mathcal{H}_i$ the subspace generated by the first i harmonic eigenfunctions *$\widetilde{\boldsymbol{\psi}}_i = D^{-1/2}\boldsymbol{\psi}_i$.*

Then σ_{i+1} satisfies

$$\sigma_{i+1} = \inf_{f \perp \mathcal{H}_i} \frac{\sum_{u \bullet\!\!-\!\!\bullet v} w(u,v)(f(u)-f(v))^2}{\sum_{u \in \mathbf{V}} f(u)^2 \deg(u)}, \tag{2.15}$$

and $\widetilde{\boldsymbol{\psi}}_{i+1}$ is the f that achieves the infimum.

(iv) The right eigenvectors of $\mathbf{W}$ are the harmonic eigenvectors $\left(\widetilde{\boldsymbol{\psi}}_i\right)_{i=1}^n$, with associated eigenvalues $(1-\sigma_i)_{i=1}^n$, that is $\mathbf{W}\widetilde{\boldsymbol{\psi}}_i = (1-\sigma_i)\widetilde{\boldsymbol{\psi}}_i$.

(v) The left eigenvectors of $\mathbf{W}$ are $\widehat{\boldsymbol{\psi}}_i = D^{1/2}\boldsymbol{\psi}_i = D\widetilde{\boldsymbol{\psi}}_i$, that is $\widehat{\boldsymbol{\psi}}_i'\mathbf{W} = (1-\sigma_i)\widehat{\boldsymbol{\psi}}_i'$.

(vi) If G is connected and non-bipartite, then $\sigma_n < 2$ and the unique stationary distribution of a Markov chain on G with transition matrix $\mathbf{W}$ is $\mathbb{S} = \widehat{\boldsymbol{\psi}}_1/||\widehat{\boldsymbol{\psi}}_1|| = D\mathbf{1}/vol(G)$, in which case $\mathbb{S}'\mathbf{W} = \mathbb{S}'$.

(vii) Let G be a simple graph. Let $\mathcal{V}$, if it exists, be a collection of mutually non-adjacent (resp. adjacent) vertices such that each vertex in $\mathcal{V}$ has the same closed neighborhood of size d. Then there are $|\mathcal{V}|-1$ eigenvalues of 1 (resp. $\frac{d+1}{d}$) associated with eigenvectors such that $\boldsymbol{\psi}(v) \neq 0$ iff $v \in \mathcal{V}$.

PROOF. For properties (i)-(iii), (vi), see Chung (1997), Chapter 1. Properties (iv)-(v) follow from the relationship of $\mathbf{L}$ and $\mathbf{W}$ (2.14). (vii), known as the **twin vertex** property, is due to Butler (2008, 2016). Butler (2015) has a version of (vii) applicable to general weighted graphs, and to twin subgraphs, but the statements are not as succinct. □

At a high level, the interpretation of properties (ii)-(v) is that the eigenvectors encode structure of decreasing smoothness as i increases. The first-order harmonic eigenvector is

constant, the second-order harmonic eigenvector is the next smoothest, and so on, with the n-th order harmonic eigenvector the "wiggliest". Heuristically, smoother functions are correlated over longer distances and are less sensitive to local perturbations; the low-order harmonics can be thought to encode longer range, weaker dependence, while the opposite is true of the high-order harmonics. Interestingly, property (vii) says that the middle of the eigenvalue spectrum consists of eigenvectors that are non-zero only on vertices that look the same to the rest of the graph. The eigenvectors of $\mathbf{L}$ are obtained from those of $\mathbf{W}$ by a simple change of basis transformation, and therefore encode the same structure.

The eigenvectors of $\mathbf{L}$ are used throughout statistics and machine learning. An entire sub-field on spectral clustering algorithms relies on the eigensystem of L and similar matrices. See von Luxburg (2007) for a thorough review. A family of kernels defined on graphs regularize the higher-order eigenfunctions by damping the eigenvalue spectrum (Smola and Kondor, 2003; Belkin, Matveeva, and Niyogi, 2004). Kirichenko and Zanten (2015) use a Bayesian approach to Laplacian regularization to estimate smooth functions on graphs.

The properties of $\mathbf{L}$ and $\mathbf{W}$ will be important to the theoretical results in Chapter 3, and to deriving inference algorithms in Chapter 4.

2.5 SMC and particle MCMC methods

Sequential Monte Carlo (SMC) and Markov chain Monte Carlo (MCMC) methods provide the basic building blocks for the inference procedures we develop in Chapter 4. We briefly review the basic framework of SMC, and give an overview of recent work incorporating SMC into MCMC.

2.5.1 Basic SMC

We briefly recall SMC algorithms. The canonical application is a state space model. Observed is a sequence $x_{1:T} = (x_1, \ldots, x_T)$. The model explains the sequence using an unobserved sequence of latent states $Z_{1:T}$, and defines three quantities:

- A Markov kernel $q_\theta^t(\bullet \mid Z_{t-1})$ that models transitions between latent states.
- An emission distribution p_θ^t that explains each observation as $x_t \sim p_\theta^t(\bullet \mid Z_t)$.
- A vector θ collecting the model parameters.

In the simplest case, θ is fixed. The inference target is then the posterior distribution $\mathcal{L}_\theta(Z_{1:T} \mid x_{1:T})$ of the latent state sequence.

A SMC algorithm (see Algorithm 2.1) generates some number $N \in \mathbb{N}$ of state sequences $Z_{1:T}^1, \ldots, Z_{1:T}^N$, and then approximates the posterior as a sample average over these sequences, weighted by their respective likelihoods. The imputed states $Z_{1:t}^i$ are called **particles**. Since the sequence of latent states is Markov, particles can be generated sequentially as $Z_t \sim q_\theta^t(Z_t|Z_{t-1})$. In cases where sampling from q_θ^t is not tractable, q_θ^t it is additionally approximated by a simpler proposal kernel r_θ^t. The likelihood, and the accuracy of approximation of q_θ^t by r_θ^t, are taken into account by computing normalized weights

$$w_t^i := \frac{\tilde{w}_t^i}{\sum_{j=1}^N \tilde{w}_t^j} \qquad \text{where} \qquad \tilde{w}_t^i = p_\theta^t(x_t|Z_t^i) \cdot \frac{q_\theta^t(Z_t^i \mid Z_{t-1}^i)}{r_\theta^t(Z_t^i \mid Z_{t-1}^i)} \,. \tag{2.16}$$

In step t, SMC generates particles Z_t^i by first resampling from the previous particles $(Z_{1:t-1}^i)_i$ with probability proportional to their weights, e.g. by sampling from the multinomial dis-

tribution

$$A_t^i \sim \mathrm{MN}(N, (w_{t-1}^i)_{i=1}^N) \ .$$

The next set $(Z_{1:t}^i)_i$ of particles is then generated by sampling $Z_t^i \sim r_\theta^t(\bullet \mid Z_{t-1}^{A_t^i})$ and setting $Z_{1:t}^i = (Z_{1:t-1}^{A_t^i}, Z_t^i)$. Resampling a final time after the T-th step yields a complete array $(Z_{1:T}^i)_i$, with which the posterior is approximated as the average

$$\mathcal{L}_\theta(dz_{1:T} \mid x_{1:T}) \approx \frac{1}{N} \sum_{i=1}^N \delta_{Z_{1:T}^i}(dz_{1:T}) \ . \tag{2.17}$$

Alternatively, the final resampling step can be omitted, in which case the posterior is approximated with a weighted average. Either approximation is asymptotically unbiased as $N \to \infty$. See Doucet and Johansen (2011) for a thorough review.

Algorithm 2.1 (SMC sampling)**.**

- Initialize $Z_1^i \overset{\text{iid}}{\sim} q_\theta^1(\bullet)$ and $\tilde{w}_1^i = q_\theta(Z_1^i) p_\theta^1(x_1 \mid Z_1^i)$ for each $i \leq N$.
- For $t = 2, \ldots, T-1$, iterate:
  - Resample indices $A_t^i \sim \mathrm{MN}(N, (w_{t-1}^i)_i)$.
  - Draw $Z_t^i \sim r_\theta^t(\bullet \mid Z_{t-1}^{A_t^i})$ for each i.
  - Compute weights as in (2.16) and normalize to obtain w_t^i.
- Resample N complete sequences $Z_{1:T}^i = Z^i \sim \mathrm{MN}(N, (w_{T-1}^i)_i)$ and set $w_T^i = 1/N$.

2.5.2 Pseudo-marginal and particle MCMC methods

In addition to generating an asymptotically unbiased approximation of the posterior, SMC yields an unbiased approximation of the marginal likelihood $p_\theta(x_{1:T})$ of the data. Define

$$\hat{p}_\theta(x_{1:T}) := \hat{p}_\theta(x_1) \prod_{t=1}^{T} \hat{p}_\theta(x_t \mid x_{t-1}) \quad \text{where} \quad \hat{p}_\theta(x_t \mid x_{t-1}) = \frac{1}{N} \sum_{i=1}^{N} \tilde{w}_t^i \,. \tag{2.18}$$

Then $\mathbb{E}[\hat{p}_\theta(x_{1:T})] = p_\theta(x_{1:T})$. Proof of unbiasedness is given under very general conditions in Del Moral (2004, Chapter 7); a more accessible proof is in Pitt, Silva, Giordani, and Kohn (2010), in the context of a version of SMC called the auxiliary particle filter.

$\hat{p}_\theta(x_{1:T})$ is useful for parameter inference. Consider a Metropolis–Hastings (MH) sampler targeting the joint posterior distribution of $(\Theta, Z_{1:T}) \mid x_{1:T}$, where Θ has prior $P_{[\Theta]}$. Given a state $(\Theta, Z_{1:T})$, a new proposal is generated by first proposing a new parameter $\Theta^* \sim \tilde{q}(\bullet \mid \Theta)$, then generating a sample $Z^*_{1:T} \mid \Theta^*, x_{1:T}$ with Algorithm 2.1. The proposal is accepted with probability $\min\{1, A_{MH}\}$, with MH acceptance ratio

$$\begin{aligned} A_{MH} &= \frac{P_{[\Theta]}(\Theta^*) p_{\Theta^*}(Z^*_{1:T}) p_{\Theta^*}(x_{1:T} \mid Z^*_{1:T})}{\tilde{q}(\Theta^* \mid \Theta) p_{\Theta^*}(Z^*_{1:T} \mid x_{1:T})} \frac{\tilde{q}(\Theta \mid \Theta^*) p_\Theta(Z_{1:T} \mid x_{1:T})}{P_{[\Theta]}(\Theta) p_\Theta(Z_{1:T}) p_\Theta(x_{1:T} \mid Z_{1:T})} \\ &= \frac{P_{[\Theta]}(\Theta^*) p_{\Theta^*}(x_{1:T})}{P_{[\Theta]}(\Theta) p_\Theta(x_{1:T})} \frac{\tilde{q}(\Theta \mid \Theta^*)}{\tilde{q}(\Theta^* \mid \Theta)} \,. \end{aligned} \tag{2.19}$$

The marginal likelihood $p_\Theta(x_{1:T})$ is typically intractable (hence a reason for using SMC), but using any unbiased positive estimate $\hat{p}_\Theta(x_{1:T})$ in A_{MH} is enough to generate an ergodic Markov chain that converges to the correct posterior distribution (Andrieu and Roberts, 2009; Andrieu, Doucet, and Holenstein, 2010).[3] This general technique is known as the

[3] In fact, unbiasedness is stronger than necessary. If the estimator is biased, such that $\mathbb{E}[\hat{p}_\theta(x_{1:T})] = b p_\theta(x_{1:T})$, it is sufficient that $b > 0$ is independent of the value of θ (Andrieu and Roberts, 2009).

pseudo-marginal method (Beaumont, 2003; Andrieu and Roberts, 2009). When SMC is used to generate proposals and estimate the marginal likelihood, it is known as particle MCMC (Andrieu, Doucet, and Holenstein, 2010). We further develop particle MCMC in the context of sequential models of random graphs in Chapter 4.

Chapter 3

Random walk models of networks

The class of models studied in the following generate a graph by inserting one edge at a time: Start with a single edge (with its two terminal vertices). For each new edge, select a vertex V in the current graph at random.

- With a fixed probability $\alpha \in (0, 1]$, add a new vertex and connect it to V.
- Otherwise, start a random walk at V, and connect its terminal vertex to V.

In contrast to the models discussed in Section 2.3, the location of newly inserted edges depends on those of previously inserted ones, through the random walk. The dependence results in some unusual properties. One is that, with two scalar parameters, the model generates a perhaps unexpected range of different graph structures (see Figure 3.1 in Section 3.1 for examples). The random walk step can be regarded as a model of interactions modulated by the existing graph:

- Long walks can connect vertices far apart in the graph. In this sense, the (expected) length of the walk is a range scale.

- Where convenient, the walk may be interpreted explicitly, as a process on the network (e.g. users in a social network forming connections through other users). More generally, it biases connections towards vertices reachable along multiple paths.

Thus, random walk models can explain a range of graph structures as the outcome of interactions on a certain length scale. Importantly, they are also statistically and analytically tractable.

The approach taken here is to make a modeling assumption on the network formation process (as e.g. preferential attachment models do); whether that assumption is considered adequate or not must depend on the problem at hand. Conceptually, however, the random walk models studied in the following allow slightly more intricate forms of stochastic dependence than the models discussed above, without sacrificing applicability.

Chapter overview. Random walk models are defined in Section 3.1. They can be chosen to generate either multigraphs or simple graphs, which in either case are sparse. A quantity that plays a key role in both theory and inference is the history of graph, i.e. the order Σ in which edges are inserted. In dynamic networks (e.g. Durrett, 2006; Kolaczyk, 2009), where a graph is observed over time, Σ is an observed variable. If only the final graph is observed, Σ is latent. Section 3.2 establishes some theoretical properties:

- Under suitable conditions, the limiting degree distribution can be characterized (Theorem 3.3). The generated graphs can exhibit power law properties.

- Conditionally on the order in which vertices are inserted, the scaled degree sequence converges to an almost sure limit. Each limiting relative degree can be generated

marginally by a sampling scheme reminiscent of "stick-breaking". See Theorem 3.4. Results of this type are known for preferential attachment models, for which the vertex order is deterministic, and hence need not be conditioned upon. See Chapter 5 for more general results of this type.

- If the length of the random walk tends to infinity, the model converges to a generalization of the preferential attachment model (Proposition 3.5).

Using the inference methods developed in Chapter 4, we fit the model to data. In Section 3.3, we discuss empirical results and applications to data:

- The role of the random walk parameter can be understood in more detail by relating it to the mixing time of a simple random walk on the data set; see Section 3.3.1.
- Comparisons between the random walk model and other network models are given in Section 3.3.2; we also discuss issues raised by such how comparisons are performed.
- The latent order Σ can be related to measures of vertex centrality (Section 3.3.3).

3.1 Model definition

A **random walk** on a connected graph g started at vertex v is a random sequence of vertices $(v, V_1, \ldots, V_k)$, such that neighbors in the sequence are connected in g. Figuratively, a walker repeatedly moves along an edge to a randomly selected neighbor, until k edges have been crossed. Edges and vertices may be visited multiple times. The random walk is **simple** if the next vertex is selected with uniform probability from the neighbors of the current one. The distribution of the terminal vertex V_k of a simple random walk of length k on

g is denoted $\mathrm{SRW}(g, v, k)$. The model considered in the following is, like many sequential network models, most easily specified in terms of a generative algorithm:

Algorithm 3.1 (Random walk model on multigraphs)**.**

- Fix $\beta \in (0, 1]$, a distribution P on $\mathbb{N}$, and a graph G_1 with a single edge connecting two vertices.
- For $t = 2, \ldots, T$, generate G_t from G_{t-1} as follows:
  (1) Select a vertex V of G_{t-1} at random (see below).
  (2) With probability β, attach a new vertex to V.
  (3) Else, draw $K \sim P$ and connect V to $V' \sim \mathrm{SRW}(G_{t-1}, V, K)$.

The vertex V in (1) is either sampled uniformly from the current vertex set, or from the **size-biased** (or **degree-biased**) distribution $\mathbb{S}(v) = \deg(v) / \sum_{v' \in \mathbf{V}(g)} \deg(v')$, where $\deg(v)$ is the degree of vertex v in g.

Algorithm 3.1 generates multigraphs, since the two vertices selected by the random walk may already be connected (resulting in a multi-edge), or may coincide (resulting in a self-loop). To generate simple graphs instead, step (3) above is replaced by:

(3') Else, draw $K \sim P$ and $V' \sim \mathrm{SRW}(G_{t-1}, V, K)$. Connect V to V' if they are distinct and not connected; else, attach a new vertex to V.

Either sampling scheme defines the law of a sequence $(G_1, G_2, \ldots)$ of graphs. We denote this law $\mathbf{RW}_{\mathrm{U}}(\beta, P)$ if the vertex V in (1) is chosen uniformly, or $\mathbf{RW}_{\mathrm{SB}}(\beta, P)$ in the size-biased case. Each such law is defined both on multigraphs and on simple graphs, depending on which sampling scheme is used.

We generally choose the length of the random walk as a Poisson variable (although most theoretical results in the next section hold for any choice of P): Denote by $\mathrm{Poisson}_+(\lambda)$ the Poisson distribution with parameter λ, shifted by 1, i.e. the law of $K + 1$, where K

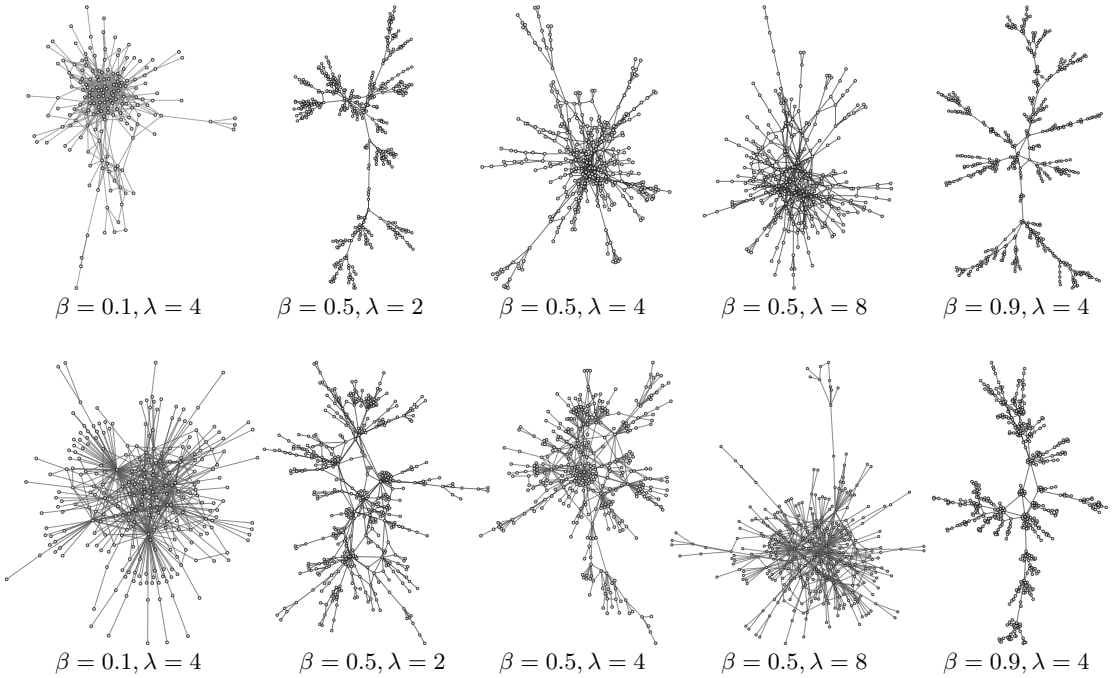


Figure 3.1: Examples of simple graphs generated by a random walk model: $\mathbf{RW}_{\mathrm{U}}(\beta, \mathrm{Poisson}_+(\lambda))$ distribution (top row), and by a $\mathbf{RW}_{\mathrm{SB}}(\beta, \mathrm{Poisson}_+(\lambda))$ distribution (bottom row).

is Poisson. We write $\mathbf{RW}(\beta, \lambda)$ for $\mathbf{RW}(\beta, \mathrm{Poisson}_+(\lambda))$ where convenient. Examples of graphs generated by this distribution are shown in Figure 3.1.

3.2 Model properties

We first observe that graphs generated by the model are sparse by construction. Let $(G_1, \ldots, G_t, \ldots)$ be a graph sequence generated by any **RW** model on simple or multi-graphs, with parameter β, where G_1 has a single edge connecting two vertices. Denote by $\mathbf{V}(G_t)$ and $\mathbf{E}(G_t)$ the set of vertices and of edges, respectively, in G_t. For any **RW** model, each vertex is added with one edge and there is a constant positive probability at each step of adding a new vertex, which leads to the following:

OBSERVATION. *Graphs generated by the random walk model are sparse: In a sequence* $(G_1, G_2, \ldots) \sim \mathbf{RW}(\beta, P)$*, the number of edges grows as* $|\mathbf{E}(G_t)| = \Theta(|\mathbf{V}(G_t)|)$.

3.2.1 Mixed random walks and the graph Laplacian

Most results in the following involve the law of a random walk. For a walk of fixed length, this law is determined by the graph's Laplacian matrix (Chung, 1997). If the length is randomized, one has to mix against its distribution; for suitable choice of the length distribution, the mixed law of the random walk is still available in closed form.

Consider an undirected graph G with n vertices, adjacency matrix A and degree matrix $D = \mathrm{diag}(\deg(v_1), \ldots, \deg(v_n))$. The probability that a simple random walk started at a vertex u terminates at v after exactly k steps is the entry (u, v) of the matrix $\mathbf{W}^k = (D^{-1}A)^k$. If the length of the walk is a random variable K with law P, the marginal probability of reaching v from u is

$$\mathbb{P}\left(V_{end} = v \mid V_0 = u,\, G\right) \quad = \quad \Big[\sum_{k\in\mathbb{N}} (D^{-1}A)^k P\left(K = k\right)\Big]_{uv} = \Big[\mathbb{E}_P\left(D^{-1}A\right)^K\Big]_{uv} . \quad (3.1)$$

A useful way to represent the information in $\mathbf{W}$ is as the matrix

$$\mathbf{L} := D^{1/2}(\mathbb{I}_n - D^{-1}A)D^{-1/2} \qquad \text{with } \mathbb{I}_n := \text{identity matrix} , \quad (3.2)$$

known as the **normalized graph Laplacian** (Chung, 1997), the properties of which were reviewed in Section 2.4. The law of a random walk of random length is obtained as follows:

PROPOSITION 3.1. *Let g be a connected, undirected simple graph or multigraph on n vertices, and let K have law P with support $\mathbb{N}_+$, such that P has probability generating function $H_P(z) := \mathbb{E}_P[z^K]$ for $|z| \leq 1$. Let $(\sigma_i)_{i=1}^n$ be the eigenvalues, and $(\boldsymbol{\psi}_i)_{i=1}^n$ the eigenvectors, of $\mathbf{L}$, and define*

$$\mathbf{K}_P := H_P(\mathbb{I}_n - \mathbf{L}) = \sum_{i=1}^{n} H_P(1-\sigma_i)\boldsymbol{\psi}_i\boldsymbol{\psi}_i' \,. \tag{3.3}$$

Then for a simple random walk $(V_0, \ldots, V_K)$ of random length K on g,

$$\mathbb{P}\left(V_{end} = v \mid V_0 = u,\, g\right) = \left[D^{-1/2}\mathbf{K}^P D^{1/2}\right]_{uv} . \tag{3.4}$$

As a consequence of Proposition 3.1, there are the following important special cases:

PROPOSITION 3.2. *Let g be a connected, undirected graph on n vertices, and define*

$$\mathbf{K}_{\mathrm{Poisson}_+(\lambda)} = \mathbf{K}^\lambda := e^{-\lambda\mathbf{L}} \qquad \textit{and} \qquad \mathbf{K}_{\mathrm{NB}_+(r,p)} = \mathbf{K}^{r,p} := \left(\mathbb{I}_n + \tfrac{p}{1-p}\mathbf{L}\right)^{-r} . \tag{3.5}$$

Then for a simple random walk $(V_0, \ldots, V_K)$ of random length K on g,

$$\mathbb{P}\left(V_{end} = v \mid V_0 = u,\, g\right) = \left[D^{-1/2}(\mathbb{I}_n - \mathbf{L})\mathbf{K}D^{1/2}\right]_{uv} , \tag{3.6}$$

with $\mathbf{K} = \mathbf{K}^\lambda$ if K has law $\mathrm{Poisson}_+(\lambda)$, or $\mathbf{K} = \mathbf{K}^{r,p}$ if K has law $\mathrm{NB}_+(r,p)$.

For a Poisson length, the result derives from the relationship

$$\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k t^k}{k!}B^k = e^{-\lambda(\mathbb{I}_n - tB)} \qquad \text{for any } B \in \mathbb{R}^{n\times n} \tag{3.7}$$

of the Poisson convolution semigroup to the matrix exponential function. It extends to the negative binomial distribution, since the latter is obtained from the Poisson by mixing: Let $\mathrm{NB}_+(r,p)$ denote the negative binomial distribution, again shifted to positive integers, with parameter r and p. The random walk model satisfies

$$\mathbf{RW}_{\mathrm{U}}(\beta, \mathrm{NB}_+(r,p)) = \int \mathbf{RW}_{\mathrm{U}}(\beta, \mathrm{Poisson}_+(\lambda))\ \gamma(\lambda \mid r, \tfrac{p}{1-p}) d\lambda \tag{3.8}$$

where $\gamma(\bullet \mid a, b)$ denotes the gamma density with parameters a and b. The same holds for $\mathbf{RW}_{\mathrm{SB}}$. The matrices (3.5) arise in other contexts, in particular in the machine learning literature: The **heat kernel** $\mathbf{K}^{\lambda}$ and the **regularized Laplacian kernel** $\mathbf{K}^{r,p}$ have applications in collaborative filtering, semi-supervised learning and manifold learning (Kondor and Lafferty, 2002; Lafferty and Lebanon, 2005; Fouss, Yen, Pirotte, and Saerens, 2006). Both act as a smoothing operators on functions defined on the vertex set, by damping the eigenvalue spectrum (Smola and Kondor, 2003).

3.2.2 Asymptotic degree properties

A property of network models intensely studied in the theoretical literature is the behavior of vertex degrees as the graph grows large. For preferential attachment graphs, limiting degree distributions can be determined analytically (Durrett, 2006; Hofstad, 2016). Our next results describe analogous properties for random walk models. The proofs use invariance of the degree-biased distribution under the operators $\mathbf{K}^{\lambda}$ and $\mathbf{K}^{r,p}$ to reduce to proof techniques developed for preferential attachment, despite the presence of the random walk.

Denote by $\mathbf{V}_{d,t}$ the subset of vertices in G_t with degree d, and let $m_{d,t} := |\mathbf{V}_{d,t}|$. The

degree distribution is the normalized histogram vector $N_t^{-1}(m_{1,t}, m_{2,t}, \dots)$ of the vertex degrees in G_t. Let $p_d(t)$ be the probability that a vertex sampled uniformly at random from $\mathbf{V}(G_t)$ has degree d. When the average probability (over vertices) of inserting a self-loop—i.e., of a random walk ending where it started—vanishes for large t, as is the case for $\lambda \to \infty$, it can be shown that as $t \to \infty$, $\mathbf{RW}_{\text{SB}}$ random multigraphs have degree distribution of the form

$$p(d) = \rho \frac{\Gamma(d)\Gamma(1+\rho)}{\Gamma(d+1+\rho)} \quad \text{where} \quad \rho := \left(1 + \frac{\beta}{2-\beta}\right) . \tag{3.9}$$

This distribution over $\mathbb{N}_+$ is known as the **Yule–Simon distribution** (e.g. Durrett, 2006). For large d, $p(d)$ scales as a power law in d with exponent $\gamma \in (2, 3]$, where

$$\gamma := 2 + \frac{\beta}{2-\beta} = 1 + \rho . \tag{3.10}$$

For general λ that may have non-vanishing average self-loop insertion probabilities, simulations indicate the same behavior.

THEOREM 3.3 (Degree distribution). *Let a sequence of multigraphs $(G_1, G_2, \dots)$ have law $\mathbf{RW}_{\text{SB}}(\beta, P)$, for some distribution P on $\mathbb{N}$, such that for all $d \in \mathbb{N}_+$*

$$\frac{1}{m_{d,t}} \sum_{v \in \mathbf{V}_{d,t}} P(V_{end} = v | V_0 = v, G_t) = o(1) \qquad \text{as } t \to \infty . \tag{3.11}$$

Then the scaled number of vertices in G_t with degree d follows a power law in d with exponent

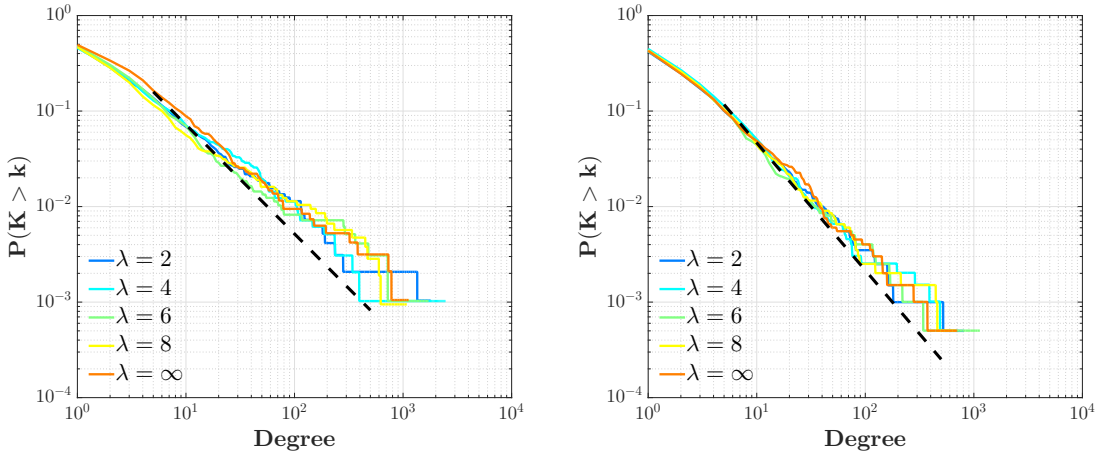


Figure 3.2: Simulated degree distributions for multigraphs with $T = 4000$ edges: left, $\beta = 0.25$; right, $\beta = 0.5$. In both cases, the distributions for finite λ appear to be the same as for $\lambda \to \infty$.

γ *as in* (3.10)*. In particular, for all* $d \in \mathbb{N}_+$*,*

$$\frac{m_{d,t}}{\beta t} \longrightarrow p(d) \qquad \textit{in probability as } t \to \infty \,. \tag{3.12}$$

Condition (3.11) controls the number of self-loops in the graph, by requiring that the probability of a random walk ending where it starts vanishes for $t \to \infty$, separately for each degree d. Although we do not have a method for checking when the condition is satisfied for finite λ, simulations indicate that the degree distribution behaves as in the $\lambda \to \infty$ case, where Condition (3.11) is known to be satisfied. Figure 3.2 shows simulated degree distributions for multigraphs with 4000 edges, for $\beta \in \{0.25, 0.5\}$. In both cases, the distributions for finite λ appear to be the same as for $\lambda \to \infty$, which is a generalization of the basic preferential attachment model (see Section 3.2.3).

Denote by $\mathbf{deg}_t := (\deg_t(v_1), \deg_t(v_2), \dots)$ the **degree sequence** of a sequential random graph model, where v_j is the j-th vertex to appear in the graph sequence. For the

preferential attachment model, the limiting degree sequence $\mathbf{deg}_\infty$ can be studied analytically, and has received considerable attention in the probability literature (see e.g. Durrett, 2006; Hofstad, 2016). The PA degree sequence is closely related to a Pólya urn, and like the limit of an urn, $\mathbf{deg}_\infty$ is itself a random variable: Informally, edges created early have sufficiently strong influence on later edges that randomness does not average out asymptotically. The joint law of $\mathbf{deg}_\infty$ can be obtained explicitly (Móri, 2005; Peköz, Ross, and Röllin, 2014; Hofstad, 2016), and determines the local weak limit (Berger, Borgs, Chayes, and Saberi, 2014). In some cases, it also admits constructive representations: Using only sequences of independent elementary random variables, one can generate a random sequence whose joint distributions are identical to those of the limiting degree sequence (Móri, 2005; James, 2015). Such representations are closely related to "stick-breaking constructions" of random partitions (e.g. Pitman, 2006).

The next result similarly describes the limiting degree sequence $\mathbf{deg}_\infty$ of the $\mathbf{RW}_{\text{SB}}$ model. The relevant technical tool is to condition on the order in which edges occur: For a graph sequence $(G_1, G_2, \dots)$ be a graph sequence generated by the model, denote by S_j the time index at which vertex v_j is inserted in the model (that is, G_{S_j} is the first graph in sequence that contains v_j), and let $S_{1:r} := (S_1, \dots, S_r)$. Let Z_α be a positive stable random variable with index α, i.e. characterized by the Laplace transform $\mathbb{E}[e^{-tZ_\alpha}] = e^{-t^\alpha}$, and $f_\alpha(z)$ its density. Define $Z_{\alpha,\theta}$, for $\theta > -\alpha$, as a random variable with the polynomially tilted density $f_{\alpha,\theta}(z) \propto z^{-\theta} f_\alpha(z)$. The variable $M_{\alpha,\theta} := Z_{\alpha,\theta}^{-\alpha}$ is said to have **generalized Mittag-Leffler distribution** with parameters α and θ (Pitman, 2006; James, 2015).

THEOREM 3.4 (Degree sequence). *Let a sequence of multigraphs $(G_1, G_2, \dots)$ have law $\mathbf{RW}_{\text{SB}}(\beta, P)$, for some distribution P on $\mathbb{N}$. Conditionally on $(S_1, S_2, \dots)$, the scaled degree*

sequence converges jointly to a random limit: For each $r \in \mathbb{N}_+$,

$$t^{-1/\rho}(\deg_t(v_1), \deg_t(v_2), \ldots, \deg_t(v_r)) \,\big|\, S_{1:r} \quad \xrightarrow{t\to\infty} \quad (\xi_1, \xi_2, \ldots, \xi_r) \,\big|\, S_{1:r} \tag{3.13}$$

almost surely. Each limiting conditional law $\mathcal{L}(\xi_j \mid S_j)$*, for* $j \geq 1$*, can be represented constructively: Let* $M_j \sim \text{Mittag-Leffler}\left(\rho^{-1}, S_j - 1\right)$*,* $B_j \sim \text{Beta}\left(1, \rho(S_j - 1)\right)$*, and* $\psi_j \sim \text{Beta}\left(S_j, S_{j+1} - S_j\right)$ *be conditionally independent given* S_j*. Then for* $1 \leq i < j$

$$\xi_j | S_j \stackrel{d}{=} M_j \cdot B_j | S_j \stackrel{d}{=} \xi_i \prod_{k=i}^{j-1} \psi_k^{1/\rho} \,|S_{i:j} \,. \tag{3.14}$$

The marginal law $\mathcal{L}(\xi_j)$ *is uniquely characterized by the moments*

$$\begin{aligned} \mathbb{E}\big[\xi_j^k\big] &= \frac{\Gamma(k+1)\Gamma(j-1)}{\Gamma(j-1+\frac{k}{\rho})} \beta^{\frac{k}{\rho}} \, {}_2F_1\big(1+\tfrac{k}{\rho}, \tfrac{k}{\rho}; j-1+\tfrac{k}{\rho}; 1-\beta\big) \\ &= \Gamma(k+1)\beta^{\frac{k}{\rho}} j^{-\frac{k}{\rho}} \cdot (1 + O(j^{-1})) \qquad \text{as } j \to \infty \,, \end{aligned} \tag{3.15}$$

where ${}_2F_1(a, b; c; z)$ *is the ordinary hypergeometric function.*

For any fixed T, one can fix an arbitrary enumeration of the edge set of G_T. The sequence $G_{1:T}$ then determines a (random) permutation Σ of the edge set $\{1, \ldots, T\}$, specifying the order in which edges where inserted, and $G_{1:T}$ can be represented equivalently as the pair (Σ, G_T). We will find in Chapter 4 that Σ also plays an important role in inference, as a latent variable. Σ completely determines $(S_1, \ldots, S_T)$ (though not vice versa), and the result above remains valid if one conditions on Σ rather than $S_{1:T}$. We study the properties of $S_{1:T}$ in PA-type models in Chapter 5, where we find that they determine the edge density

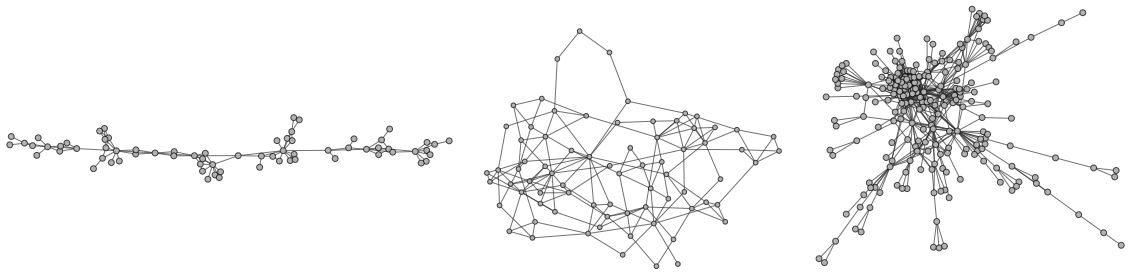
and the properties of the asymptotic degree sequence.

The simple constructive representation of the limiting distribution (3.14) seems only to exist marginally for the random walk model, not for the joint law $\mathcal{L}(\xi_1, \ldots, \xi_r \,|\, S_{1:r})$, in contrast to PA-type models, where a recursive analogue of (3.14) holds even jointly (Peköz, Ross, and Röllin, 2014; James, 2015). Chapter 5 also establishes simple constructive representations for the marginal and joint laws of closely related models.

3.2.3 Relationship between RW and preferential attachment models

If the random walk step (3) in Algorithm 3.1 is omitted—that is, for $\beta = 1$—the $\mathbf{RW}_{\text{SB}}$ model becomes reminiscent of the well-known preferential attachment model (Barabási and Albert, 1999). There are varying definitions of the PA model, but in essence, it inserts a new vertex into the graph at every step, and connects it to m vertices sampled from the size-biased distribution on the current vertex set; m is a model parameter.

A generalization of this model, due to Aiello, Chung, and Lu (2002), is defined as follows: Fix $\beta \in (0, 1]$ and let G_1 be a graph with a two vertices connected by a single



(i) (ii) (iii)

Figure 3.3: Network data examples: (i) the largest connected component of the NIPS co-authorship network, 2002-03 (Globerson, Chechik, Pereira, and Tishby, 2007); (ii) San Juan Sur family ties (Loomis, Morales, Clifford, and Leonard, 1953); (iii) protein-protein interactome (Butland et al., 2005).

edge. In each step t, construct G_t from G_{t-1} as follows: Select a vertex V according to the size-biased distribution on G_{t-1}; with probability β, attach a new vertex; otherwise, connect V to a vertex V' sampled independently from the same size-biased distribution. This model is denoted $\mathbf{ACL}(\beta)$ in the following. It can be regarded as a natural extension of the Yule–Simon model from sequences to graphs.

PROPOSITION 3.5. *The limit in distribution* $\mathbf{RW}_{\text{SB}}(\beta, \infty) := \lim_{\lambda\to\infty} \mathbf{RW}_{\text{SB}}(\beta, \lambda)$ *exists for every* β*, and* $\mathbf{RW}_{\text{SB}}(\beta, \infty) = \mathbf{ACL}(\beta)$ *if both models start with the same seed graph.*

Remark (Generalization of Proposition 3.5). The proof of Proposition 3.5 relies on the fact that if K is almost surely infinite, only the lowest order eigenvector, corresponding to $\sigma_1 = 0$, contributes to the mixed random walk probability (3.6). In fact, the same can be shown to hold for any distribution P (with parameters ϕ) that has as a limiting case a point mass at $K = \infty$:

$$\lim_{\phi} P(K < \infty) = 0 \,.$$

If this is the case, then $H_P(z) = \delta_1(z)$, in which case Proposition 3.1 shows that only the lowest order eigenvector contributes to the mixed random walk probability. Appendix A.5 shows that this implies distributional equivalence to the $\mathbf{ACL}(\beta)$ model. ◁

3.3 Experimental evaluation

Using inference methods developed in Chapter 4 for fitting the **RW** model to data observed only at G_T, this section evaluates properties of the model on real-world data and compares

its performance to other network models. We also discuss the interpretation of the random walk parameter as a length scale, and of the latent order Σ as a measure of vertex centrality.

We consider three network datasets, shown in Figure 3.3, that exhibit a range of characteristics. The first is the largest connected component (LCC) of the NIPS co-authorship network in 2002-03, extracted from the data used in Globerson, Chechik, Pereira, and Tishby (2007). As shown in Figure 3.3, it has a global chain structure connecting highly localized communities. The second dataset represents ties between families in San Juan Sur (SJS), a community in rural Costa Rica (Loomis, Morales, Clifford, and Leonard, 1953; Batagelj and Mrvar, 2006), and is chosen here as an example of a network with small diameter. The third is the protein-protein interactome (PPI) of Butland et al. (2005), which exhibits features such as chains, pendants, and heterogeneously distributed hubs. Summary statistics are given in Table 3.1. None of these data sets is particularly large: Sampler diagnostics show graphs of this size suffice to reliably recover model parameters under the random walk model.

Using Algorithm 4.3, the posterior distributions of β and λ for both the uniform and the size-biased $\mathbf{RW}(\beta, \text{Poisson}_+(\lambda))$ model can be sampled. Kernel-smoothed posterior distributions under all three datasets are shown in Figure 3.4. Although posteriors in the uniform and size-biased case are very similar, these models generate graphs with different

Table 3.1: Summary statistics for data used in experiments.

Dataset	*Vertices*	*Edges*	*Clustering coeff.*	*Diameter*	*L_2-mixing time of r.w.*
NIPS	70	114	0.70	14	≥ 62.4
SJS	75	155	0.32	7	≥ 6.1
PPI	230	695	0.32	11	≥ 9.2

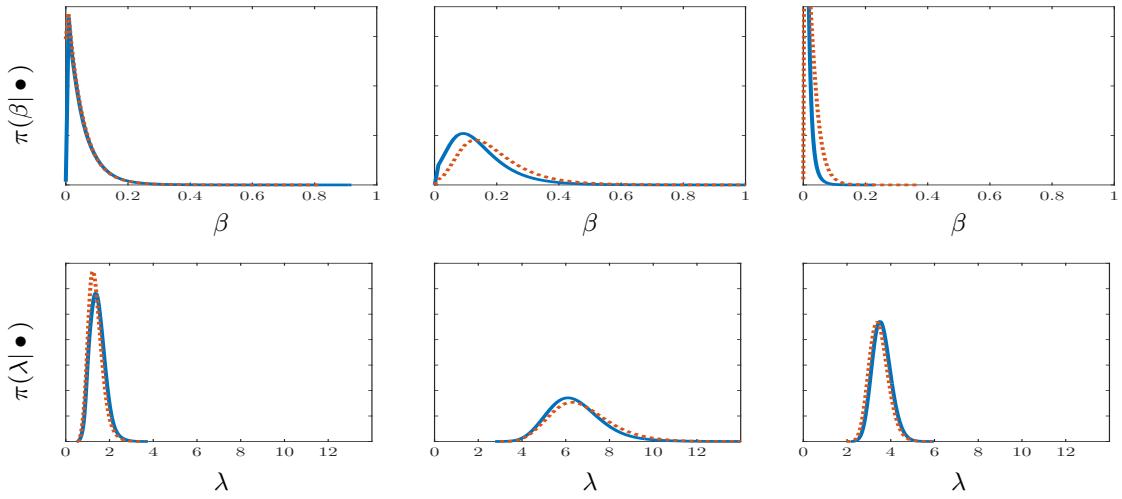


Figure 3.4: Kernel-smoothed estimates of the posterior distributions of β and λ, under the models $\mathbf{RW}_{\mathrm{U}}$ (blue/solid) and $\mathbf{RW}_{\mathrm{SB}}$ (orange/dotted). *Left column*: NIPS data. *Middle*: SJS. *Right*: PPI. Posteriors are based on 1000 samples each (lag 40, after 1000 burn-in iterations; 100 samples each are drawn from 10 chains).

characteristics for identical parameter values (see Section 3.3.2).

3.3.1 Length scale

The distance of vertices that can form connections under the model is governed by the parameter λ of the random walk, which can hence be interpreted as a form of length-scale. This scale can be compared to the mixing time of a random walk on the observed graph (listed, in L_2 norm, in the table above for each data set): If λ is significantly smaller, the placement of edges inserted by the random walk strongly depends on the graph structure; if λ is large relative to mixing time, dependence between edges is weak. Based on Figure 3.4, we observe the following:

- Concentration of the λ-posterior on small values in $[1, 2]$ for the NIPS data thus indicates predominantly short-range dependence in the data and, since the mixing time is much larger, strong dependence between edges. Concentration of the β-posterior

near the origin means connections are mainly formed through existing connections.

- In the SJS network, the posterior peaks near $\lambda = 6$, and hence near the lower bound on the mixing time, with β again small. Thus, the principal mechanism for inserting edges under the model is again the random walk, but a random walk of typical length has almost mixed. Hence, the local connections it creates do not strongly influence each other.

- The PPI network exhibits an intermediate scale of dependence, with most posterior mass in the range $\lambda \in [3, 5]$. Although structures like pendants and chains indicate strong local dependence, there are also denser, more homogeneously connected regions that indicate weaker dependence with longer range.

Narratives explaining these effects are not hard to come by—for example, short-range dependence in the NIPS data indicate new collaborations are often formed through existing co-authors; the weak dependence between edges in the SJS data suggests connections can be formed by some process external to the observed graph, as may be expected for a community concentrated in a small geographic area—but warrant caution, as all data storytelling does.

3.3.2 Model fitness

Assessing model fitness on network data is difficult: For other types of data, cross validation is often the tool of choice, and indeed, cross-validated link prediction (where links and non-links are deleted uniformly and independently at random) is widely used for model evaluation in the machine learning literature. Even if link prediction is considered the rel-

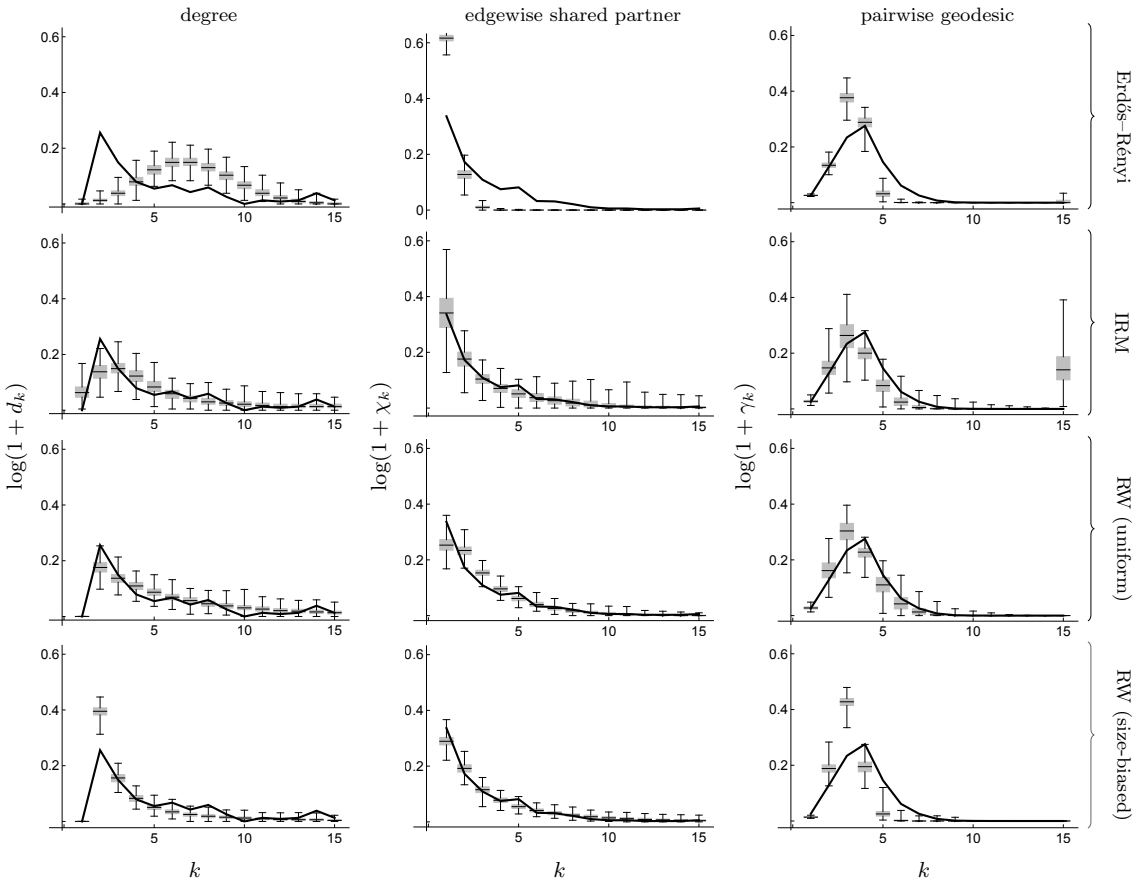


Figure 3.5: Estimated PPDs for the PPI data set of three statistics: Normalized degree d_k, normalized edgewise shared partner statistic χ_k, and normalized pairwise geodesic γ_k. Results are shown for four models: Erdős–Rényi (top row), IRM (second row), $\mathbf{RW}_{\mathrm{U}}$ (third row), and $\mathbf{RW}_{\mathrm{SB}}$ (bottom row). The black line represents the distribution from the PPI data.

evant statistic of interest, however, cross-validating network data requires subsampling the observed network. Any choice of such a subsampling algorithm implies strong assumptions on how the data was generated; it may also favor one model over another.

We therefore use a protocol developed by Hunter, Goodreau, and Handcock (2008), who compare models to data by fixing a set of network statistics. The fitted model is evaluated by comparing the chosen statistics of the data to those of networks simulated from the model. The selected statistics specify which properties are considered important

in assessing fit. To capture a range of structures, the following count statistics are proposed in Hunter, Goodreau, and Handcock (2008), which we also adopt here:

- Normalized degree statistics (ND), d_k, the number of vertices of degree k, divided by the total number of vertices.
- Normalized edgewise shared partner statistics (NESP), χ_k, the number of unordered pairs $\{i, j\}$ such that i and j are connected and have exactly k common neighbors, divided by the total number of edges.
- Normalized pairwise geodesic statistics (NPG), γ_k, the number of unordered pairs $\{i, j\}$ with distance k in the graph, divided by the number of dyads.

In a Bayesian setting, executing the protocol amounts to performing posterior predictive checks (Box, 1980; Gelman, Meng, and Stern, 1996) via the following procedure:

(1) Sample the model parameters from the posterior, $\theta \sim \pi(\theta|G_T)$.

(2) Simulate a graph of the same size as the data, $G^{(s)} \sim P(G|\theta)$.

(3) Calculate the statistic(s) of interest for the simulated graph, $f(G^{(s)})$.

Table 3.2: Summary of goodness-of-fit: total variation distance of PPDs to the empirical distribution of the PPI and NIPS data sets. Smaller values indicate a better fit.

	PPI data			*NIPS data*		
Model	*Degree*	*ESP*	*Geodesic*	*Degree*	*ESP*	*Geodesic*
EPM	$0.49 \pm .03$	$0.31 \pm .07$	$0.65 \pm .04$	$0.57 \pm .06$	$0.43 \pm .15$	$0.72 \pm .05$
IRM	$0.30 \pm .04$	$0.15 \pm .07$	$0.25 \pm .07$	$0.29 \pm .08$	$0.46 \pm .10$	$0.36 \pm .13$
ER	$0.57 \pm .02$	$0.45 \pm .03$	$0.23 \pm .02$	$\mathbf{0.26 \pm .06}$	$0.69 \pm .06$	$0.41 \pm .06$
ACL	$0.28 \pm .02$	$\mathbf{0.09 \pm .02}$	$0.34 \pm .03$	$0.42 \pm .05$	$0.51 \pm .06$	$0.50 \pm .04$
RW-U	$\mathbf{0.23 \pm .03}$	$0.17 \pm .02$	$\mathbf{0.16 \pm .08}$	$\mathbf{0.26 \pm .04}$	$\mathbf{0.33 \pm .05}$	$\mathbf{0.22 \pm .08}$
RW-SB	$0.27 \pm .02$	$0.11 \pm .02$	$0.34 \pm .04$	$0.42 \pm .05$	$0.39 \pm .06$	$0.45 \pm .07$

The value $f(G^{(s)})$ is then a sample from the posterior predictive distribution (PPD) of the statistic f under the model. If the PPD places little mass on the observed value of the statistic, this indicates the model does not explain those properties of the data measured by the given statistic.

For each of the datasets and statistics above, PPDs are estimated for the following models: The Erdős–Rényi model (ER); the infinite relational model (IRM) with a Chinese Restaurant Process prior on the number of blocks (Kemp, Tenenbaum, T. L. Griffiths, Yamada, and Ueda, 2006; Xu, Tresp, Yu, and Kriegel, 2006); the infinite edge partition model with an underlying hierarchical gamma process (EPM) (Zhou, 2015); the ACL model described in Section 3.2.3; and the $\mathbf{RW}_{\mathrm{U}}$ and $\mathbf{RW}_{\mathrm{SB}}$ models. Table 3.2 lists the total variation distance between the empirical distribution of each statistic on the observed graph and on graphs generated from the respective models. Standard errors are computed over 1000 samples. Smaller values indicate a better fit.

For the ER model, the IRM, and the two random walk models, PPD estimates are shown in more detail in Figure 3.5, on a logarithmic scale. Some comments:

- In terms of the protocol of Hunter, Goodreau, and Handcock (2008), the uniform

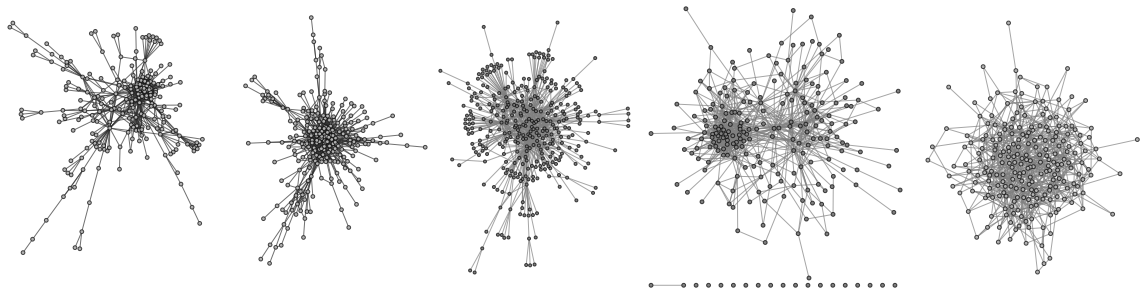


Figure 3.6: Reconstructions of the PPI network (i), sampled from the posterior mean of (ii) the $\mathbf{RW}_{\mathrm{U}}$ model, (iii) the $\mathbf{RW}_{\mathrm{SB}}$ model, (iv) the IRM, and (v) the Erdős–Rényi model.

random walk model provides the best fit on both data sets.

- Of the models used for comparison, the IRM does similarly well, although it does place significant posterior mass on d_0 and γ_∞ (since it tends to generate isolated vertices).

- In the case of the IRM, good fit comes at the price of model complexity: The IRM is a prior on stochastic blockmodels with an infinite number of classes, a finite number of which are invoked to explain a graph of finite size. For the PPI data set, for example, the IRM posterior sharply concentrates at 9 classes, which amounts to 53 scalar parameters, compared to 2 parameters of the RW model.

The numerical results can be illustrated by sampling reconstructions of the input network from models fitted to the data. Figure 3.6 compares reconstructions of the PPI data set generated by the random walk models to those generated by the IRM and the ER model. Such visual network comparisons should be treated with great caution; nonetheless, the comparison underscores that for some types of data, including the data set depicted here, the random walk model provides an arguably better structural fit.

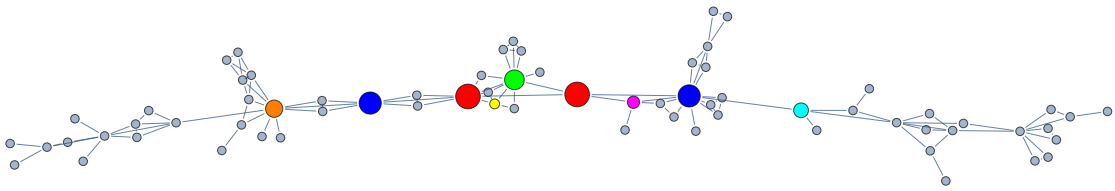


Figure 3.7: NIPS authors sampled earliest in latent sequence under the $\mathbf{RW}_{\mathrm{U}}$ model. Colored vertices correspond to those in Table 3.3.

Table 3.3: Measures of vertex centrality for the NIPS graph. Figure 3.7 maps these vertices in the graph.

Arrival Order (median order)			*Degree (degree)*	*Btwn. Cent. (# shortest paths)*	*Info. Cent. (info. cent.)*	*Katz Cent. (Katz cent.)*
●	(r)	1 (6)	14 (4)	2 (2376)	1 (0.4478)	6 (0.2380)
●	(l)	1 (6)	4 (8)	3 (2340)	2 (0.4443)	2 (0.3740)
●	(r)	2 (7)	1 (12)	1 (2797)	3 (0.4374)	3 (0.2571)
●	(l)	2 (7)	8 (6)	5 (2116)	5 (0.4178)	4 (0.2539)
●		3 (8)	2 (10)	11 (728)	4 (0.4259)	1 (0.3789)
●		4 (10)	2 (10)	4 (2196)	11 (0.3877)	5 (0.2458)
●		5 (12)	14 (4)	6 (1870)	17 (0.3645)	35 (0.0667)
●		6 (13)	14 (4)	16 (173)	6 (0.4040)	19 (0.1291)
●		7 (14)	14 (4)	23 (28)	7 (0.3956)	7 (0.2264)

3.3.3 Latent arrival order and vertex centrality

Various network statistics attempt to measure the importance of individual vertices for interactions within the network, a property known as **centrality**. This idea can be formalized in a variety of ways. A simple measure of centrality is simply the degree of a vertex. Other examples include Eigenvalue Centrality, Katz Centrality, and Information Centrality (e.g. Kolaczyk, 2009; Newman, 2009), which each measure different properties of the graph. Under a random walk model, the number of random walks passing through a vertex in the process of network formation provides an obvious measure of that vertex's importance to the formation of the network. A simpler proxy is the order in which vertices are inserted. If only the final graph G_T is observed, this is the order induced by the latent vertex arrival times $S_{1:T}$ in Theorem 3.4, which is in turn determined by the latent edge order Σ, as given by the posterior distribution of the bridge $G_{1:T}$ generated by the inference algorithm (see Chapter 4). The median of the inferred arrival time of vertices in Σ under the $\mathbf{RW}_{\mathrm{U}}$ model for the NIPS dataset, for those nine authors with earliest median appearance, is listed in

Table 3.3; other measures of centrality are also listed. For each measure, the vertex's rank is listed, with the numerical value in parentheses. Since one would generally expect that vertices inserted into the graph earlier tend to have larger degree, it is interesting to note that arrival order seems to correlate more closely with rank of Betweenness Centrality and of Information Centrality than with vertex degree.

3.4 Discussion

We have introduced a class of network models that can account for dependence of new links on existing links, but are nonetheless tractable. In light of our results, there are two reasons for tractability:

- There exists a latent variable—the edge insertion order Σ—conditioning on which greatly simplifies the distribution of $G_{1:T} \mid G_T$.
- The effects of the parameters β and λ are sufficiently distinct that, as we show in Chapter 4, inference is possible—a single graph generated by the model carries sufficient signal for parameter values to be recovered with high accuracy.

Conditioning on history information, either on the order in which vertices are inserted, or on entire history Σ, is instrumental both for theoretical results and for inference. As mentioned in the introduction, a qualitatively similar effect—the joint distribution of the network graph simplifies conditionally on a suitable latent variable—is observed in other network models, such as graphon models and the configuration model. In Chapter 5, we undertake a careful study of Σ and $S_1, S_2, \ldots$ in preferential attachment-type models.

Open questions. There are open questions beyond the scope of this dissertation. Some of these are obvious, such as applications to dynamic network problems; generalizations to disconnected graphs (e.g. using a disconnected seed graph); or which choices of the distribution P of the random walk length in Theorem 3.3 satisfy condition (3.11).

A question we consider intriguing is whether the random walk itself results in power law behavior. Intuitively, the probability to reach a vertex by random walk increases with degree; that suggests the random walk might result in an effect similar to preferential attachment. For any $\mathbf{RW}(\beta, \text{Poisson}_+(\lambda))$ model, the probability (3.1) can be written as

$$\mathbb{P}(V_{end} = v | V_0 = u,\, G) = \left[\frac{d_v}{\text{vol}(G)} + \sqrt{\frac{d_v}{d_u}} \sum_{i=2}^{n} (1 - \sigma_i) e^{-\lambda \sigma_i} \boldsymbol{\psi}_i(u) \boldsymbol{\psi}_i'(v) \right],$$

where $(\boldsymbol{\psi}_i)$ are the eigenvectors, and (σ_i) the eigenvalues, of the normalized graph Laplacian $\mathbf{L}$. Thus, the probability to terminate at v is a mixture of a preferential attachment term (proportional to d_v), and a term depending on λ and on the structure of G at various scales. Our power law results in Section 3.2 stem from size-biased selection from the vertex set, not the random walk. This mechanism is not used in the $\mathbf{RW}_{\text{U}}$ model on multigraphs, which empirically nonetheless exhibits a heavy-tailed degree distribution. That seems to suggest an affirmative answer, but at present, we have no proof.

Chapter 4

Inference methods for sequential models

We consider the problem of performing estimation and inference of model parameters for sequential models of networks generally, and random walk models in particular. If the entire history Σ of a multigraph is observed, model parameters can be estimated by maximum likelihood or traditional MCMC.

The primary source of difficulty presents itself when the edge order Σ is unobserved. If only the final graph is observed, inference is still possible, by treating the history Σ as a latent variable and imputing it using a sampling algorithm. This problem is of broader interest, since it permits the application of sequential or dynamic network models, including the PA model and many others, to a single observed graph. Furthermore, the random walk mechanism presents inferential challenges: Given G_t, the probability of G_{t+1} does not depend on a simple statistic of G_t, as in the PA model, but on the entire structure of

G_t. Using the results of Section 3.2.1, we derive collapsed sampling updates that improve sampling efficiency.

Chapter overview. In Section 4.1 we assume a $\mathbf{RW}(\beta, P)$ model, and that the history of the network is observed, in which case the Markov structure of the $\mathbf{RW}(\beta, P)$ model likelihood admits tractable maximum likelihood estimators. The estimators are obtainable in closed form for β, and via estimating equations for the parameters of P. We derive the quantities necessary for solving the estimating equations via numerical optimization.

In Section 4.2 we consider the general problem of performing inference on a sequential model when the history is only partially observed, and in particular when only G_T is observed. Given the sequential specification of such models, sequential Monte Carlo (SMC) techniques are the natural building block for inference algorithms. For joint inference of the latent history and the parameters of the model, we extend the particle MCMC methods in Andrieu, Doucet, and Holenstein (2010) for models satisfying a Markov property and a monotonicity property. In Section 4.3, we apply the results of Sections 3.2 and 4.2 to derive a collapsed particle Gibbs sampler for the $\mathbf{RW}(\beta, \text{Poisson}_+(\lambda)$ model, and demonstrate empirically that it accurately recovers the parameters from a single G_T generated from the model. We also demonstrate an example of MCMC for the case when Σ is observed, in Section 4.4. Proofs of all theoretical results are given in Appendix C.

4.1 Maximum likelihood estimation for fully observed sequential models

We consider parameter estimation for multigraphs in the "dynamic" case, where the entire history of a graph is observed. Let $(G_1, \ldots, G_t)$ be a multigraph sequence generated by any **RW** model, with parameter β, where for simplicity G_1 has a single edge connecting two vertices.[1] Denote by (v_s, v'_s) the pair of vertices connected by the edge inserted in step $s \leq t$. The edge was generated by step (2) of the sampling scheme if and only if one of the vertices has degree 0 in G_{s-1}, and hence if the indicator variable $B_s := \mathbb{1}\{\min\{\deg_{s-1}(v_s), \deg_{s-1}(v'_s)\} = 0\}$ takes value 1. Since the model creates vertices by independent β-coin flips, the number N_t of vertices in G_t is $(N_t - 2) \sim \text{Binomial}(t-1, \beta)$, and therefore

$$\hat{\beta}_t = \frac{\sum_{s=2}^{t} B_s}{t-1} = \frac{N_t - 2}{t-1} \tag{4.1}$$

is a maximum likelihood estimator for β. The sequence's probability of occurrence under a $\mathbf{RW}(\beta, P)$ model is

$$L_t(\beta, \phi) := C_{N_t,t}(\beta) \prod_{s=2}^{t} \mathbb{P}\{\text{edge}_s = (v_s, v'_s)\}^{1-B_s} , \tag{4.2}$$

where $C_{N_t,t}(\beta)$ denotes the probability of $N_t - 2$ under a $\text{Binomial}(t-1, \beta)$ distribution, and ϕ denotes the parameter(s) of the random walk length distribution, P. The product's

[1] This section applies to arbitrary fixed seed graphs, as long as estimators are modified appropriately.

factors are given by Proposition 3.1. In particular, define

$$Q_s^\phi := D_s^{-1/2}\mathbf{K}_s^\phi D_s^{1/2}\,, \tag{4.3}$$

with the kernel $\mathbf{K}_s^\phi$ induced by P as in (3.3). Denote the relative degree of a vertex v in G_s by $\bar{d}_s(v) := \deg_s(v)/\sum_{v'\in\mathbf{V}G_s}\deg_s(v')$. Then for a $\mathbf{RW}_{\mathrm{SB}}(\beta, P)$ model, $\hat{\beta}_t$ as in (4.1) and

$$\hat{\phi}_t := \arg\max_\phi \prod_{s=2}^{t}\left([Q_{s-1}^\phi]_{v_s v'_s}\bar{d}_{s-1}(v_s) + [Q_{s-1}^\phi]_{v'_s v_s}\bar{d}_{s-1}(v'_s)\right)^{1-B_s} \tag{4.4}$$

are maximum likelihood estimators of the parameters β and ϕ. For a $\mathbf{RW}_{\mathrm{U}}(\beta, P)$ model, $\hat{\phi}_t$ is obtained by replacing $\bar{d}_{s-1}(v_s)$ with $1/N_{s-1}$. Neither estimate depends on the other, and $\hat{\phi}_t$ can be computed by straightforward numerical optimization using the derivatives of $\ell_t(\beta,\phi) := \log L_t(\beta,\phi)$,

$$\nabla_\phi \ell_t(\beta,\phi) = \sum_{s=2}^{t}(1-B_s)\frac{\left([\dot{Q}_{s-1}^\phi]_{v_s v'_s}\bar{d}_{s-1}(v_s) + [\dot{Q}_{s-1}^\phi]_{v'_s v_s}\bar{d}_{s-1}(v'_s)\right)}{\left([Q_{s-1}^\phi]_{v_s v'_s}\bar{d}_{s-1}(v_s) + [Q_{s-1}^\phi]_{v'_s v_s}\bar{d}_{s-1}(v'_s)\right)} \tag{4.5}$$

$$\begin{aligned}\nabla_\phi^2 \ell_t(\beta,\phi) = \sum_{s=2}^{t}(1-B_s)\Bigg(&\frac{\left([\ddot{Q}_{s-1}^\phi]_{v_s v'_s}\bar{d}_{s-1}(v_s) + [\ddot{Q}_{s-1}^\phi]_{v'_s v_s}\bar{d}_{s-1}(v'_s)\right)}{\left([Q_{s-1}^\phi]_{v_s v'_s}\bar{d}_{s-1}(v_s) + [Q_{s-1}^\phi]_{v'_s v_s}\bar{d}_{s-1}(v'_s)\right)}\\ &- \frac{\left([\dot{Q}_{s-1}^\phi]_{v_s v'_s}\bar{d}_{s-1}(v_s) + [\dot{Q}_{s-1}^\phi]_{v'_s v_s}\bar{d}_{s-1}(v'_s)\right)^2}{\left([Q_{s-1}^\phi]_{v_s v'_s}\bar{d}_{s-1}(v_s) + [Q_{s-1}^\phi]_{v'_s v_s}\bar{d}_{s-1}(v'_s)\right)^2}\Bigg)\,,\end{aligned} \tag{4.6}$$

where $\dot{Q}^\phi$ and $\ddot{Q}^\phi$ denote the first and second derivatives, respectively, of Q^ϕ with respect to ϕ.

For a $\mathbf{RW}_{\mathrm{SB}}(\beta, \mathrm{Poisson}_+(\lambda))$ model, for example, define

$$Q_s^\lambda := D_s^{-1/2}(\mathbb{I}_{N_s} - \mathbf{L}_s)\mathbf{K}_s^\lambda D_s^{1/2}\,, \tag{4.7}$$

with the heat kernel $\mathbf{K}_s^\lambda = e^{-\lambda \mathbf{L}}$, as in (3.5). The derivatives are

$$\dot{Q}_s^\lambda = -D_s^{-1/2}(\mathbb{I}_{N_s} - \mathbf{L}_s)\mathbf{L}_s\mathbf{K}_s^\lambda D_s^{1/2}$$

$$\ddot{Q}_s^\lambda = D_s^{-1/2}(\mathbb{I}_{N_s} - \mathbf{L}_s)\mathbf{L}_s^2\mathbf{K}_s^\lambda D_s^{1/2} \ .$$

Similarly, tractable ML estimators can be obtained in the negative binomial case, by substituting $\mathbf{K}^{r,p}$ for $\mathbf{K}^\lambda$ in (4.7).

We note that $\hat{\beta}_t$ behaves as a classical maximum likelihood estimator. $\hat{\phi}_t$ may behave differently; in particular, we observe the following:

- The variance of $\hat{\phi}_t$ is connected to β: The expected effective sample size is

$$\mathbb{E}[ESS_t^\phi] = \mathbb{E}[t - 1 - (N_t - 2)] = (1 - \beta)t + 1 \ . \tag{4.8}$$

  Thus, for higher values of β, which corresponds to sparser graphs, $\hat{\phi}_t$ may have high variance, regardless of the properties of P or the structure of $G_{1:T}$.

- G_s enters the likelihood through Q_s^ϕ. Thus, the accuracy and variance of $\hat{\phi}_t$ depends strongly on the structure of each G_s. If, for example, the typical K_s is small compared to the mixing time of a random walk on G_s, the signal relevant to estimating ϕ will be relatively strong. If, on the other hand, the typical K_s is on the order of the mixing time of G_s, the signal will be lower and $\hat{\phi}_t$ will have high variance. Furthermore, it may be difficult to recover ϕ corresponding to typical K much larger than the mixing time.

For simple graphs, maximum likelihood estimation is still possible in principle, but is more

complicated since the effects of β and λ are no longer independent: In step (3') in Section 3.1, the probability of generating an additional vertex, which increments the count N_t, depends on the outcome of the random walk. The definition of step (3') is one of several possible ways to translate the multigraph case to simple graphs, but dependence of β and λ is not an artifact of a specific definition: Rather, it is due to the fact that a simple graph contains no observable evidence of a random walk connecting two previously connected vertices.

4.2 Particle methods for partially observed sequential models

This section develops Markov chain sampling methods for a commonly encountered problem: A sequential network model is assumed, but only the final graph G_T generated by the model is observed, as opposed to the entire history $G_{1:T} := (G_1, \ldots, G_T)$ of the generative process. The methods developed here are easily adapted to the case when the history is partially observed, say at $G_{t_1^*}, G_{t_2^*}, \ldots$. Then the subsequences $G_{1:t_1^*}$, $G_{t_1^*:t_2^*}, \ldots$, can be handled in the same manner as $G_{1:T}$.

The methods developed here assume a Bayesian setup, with a prior distribution $\mathcal{L}(\theta)$ on the model parameter θ, where $\mathcal{L}(\bullet)$ generically denotes the law of a random variable. They sample the posterior distribution $\mathcal{L}(\theta|G_T)$, and are applicable to sequential network models satisfying two properties:

(P1) The sequence $(G_1, \ldots, G_T)$ of graphs generated by the models forms a Markov chain on the set of finite graphs, that is, $G_{t+1} \perp\!\!\!\perp_{G_t} (G_1, \ldots, G_{t-1})$ for $t < T$.

(P2) The sequence is strictly increasing, in the sense that $G_t \subsetneq G_{t+1}$ almost surely.

These hold for the random walk models, but also for many other network formation models, such as preferential attachment graphs, fitness models, and vertex copying models (e.g. Newman, 2009; Goldenberg, Zheng, Fienberg, and Airoldi, 2010). Despite the attention these models have received in the literature, little work exists on inference (see Section 4.5 for references).

If only a single graph G_T is observed, inference requires the unobserved history to be integrated out. The result is a likelihood of the form

$$p_\theta(G_T \mid G_1) = \int p_\theta(G_T, G_{2:(T-1)} \mid G_1)\, dG_{2:(T-1)} \,, \tag{4.9}$$

where θ is the vector of model parameters. Since the variables G_t take values in large combinatorial sets, the integral amounts to a combinatorial sum that is typically intractable. The strategy is to approximate the integral with a sampler that imputes the unobserved graph sequence, noting that only valid sequences which lead to G_T will have non-zero likelihood (4.9). The sequential nature of the models makes SMC and particle methods the natural tools of choice. Building on the methods reviewed in Section 2.5, we develop tractable algorithms that enable inference on data with partially observed or unobserved history.

4.2.1 SMC algorithms for graph bridges

Consider a sequential network model satisfying properties (P1) and (P2) above, with model parameters collected in a vector θ. For now, θ is fixed, and the objective is to reconstruct the graph sequence $G_{1:T}$ from its observed final graph G_T and a fixed initial graph G_1, i.e.

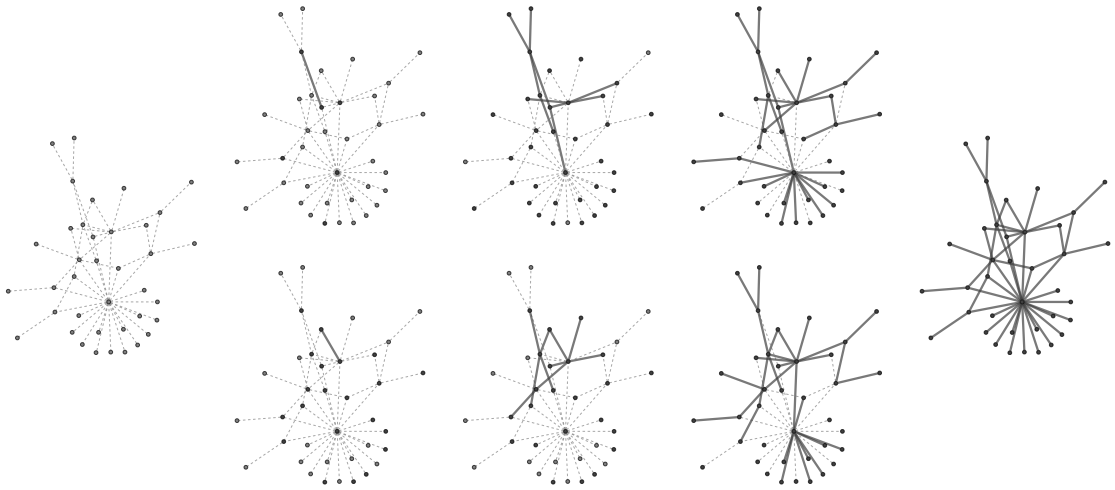


Figure 4.1: Two graph bridges generated by Algorithm 4.1: A graph G_{50} is drawn from a $\mathbf{RW}_{\mathrm{U}}(\beta, P)$ model, and two graph bridges $G_{1:50}$ are sampled conditionally on the fixed input G_{50}. Shown are the graphs G_1, G_{10} and G_{30} of each bridge.

the relevant posterior distribution is $\mathcal{L}_\theta(G_{1:T} \mid G_T, G_1)$. By the Markov property (P1), each step in the process is governed by a Markov kernel

$$q_\theta^t(g \mid g') := P(G_t = g \mid G_{t-1} = g') \; .$$

The task of a sampler is hence to impute the conditional sequence $G_{2:(T-1)} \mid G_1, G_T$, which is a stochastic process conditioned on its initial and terminal point, also known as a **bridge**. Compared to the SMC sampling algorithm for state space models, the unobserved sequence $G_2, \ldots, G_{T-1}$ takes the role of the hidden states $z_{1:T}$, whereas G_1 and the single observed graph G_T replaces the observed sequence $x_{1:T}$. The emission likelihood $p_\theta^t(x_t|z_t)$ is replaced by a bridge likelihood, of a graph G_t given G_1 and G_T.

The relevant likelihood functions, however, are themselves intractable: If g_s is a fixed

graph at index s in the sequence, the probability of observing g_t at a later point $t > s$ is

$$p_\theta^{t,s}(g_t|g_s) := \int \Big(\prod_{i=s}^{t-1} q_\theta^{i+1}(g_{i+1}|g_i)\Big) dg_{s+1} \cdots dg_{t-1} \qquad \text{whenever } t > s \ .$$

For a candidate graph G_t generated by the sampler, the likelihood under observation G_T is the **bridge likelihood**

$$L_\theta^t(G_t) := p_\theta^{T,t}(G_T|G_t) \ .$$

The bridge likelihood is intractable unless $t = T - 1$; indeed, for $t = 1$, this is precisely the integral (4.9) that we set out to approximate.

A sequence $G_{1:T}$ satisfying (P2) is equivalent to an enumeration of the T edges of G_T in order of occurrence. Let σ be a permutation of $\{1, \ldots, T\}$, i.e. an ordered list containing each element of the set exactly once, and

$$\begin{aligned} \sigma_t(G_T) := &\text{ graph obtained by deleting all edges of } G_T \text{ not listed in } \sigma_t, \\ &\text{and any resulting isolated vertices.} \end{aligned}$$

A random sequence $G_{1:T}$ can then be specified as a pair (Σ, G_T), for a *random* permutation Σ of the edges in G_T, with $G_t = \Sigma_t(G_T)$. Given G_T, not every permutation σ is a valid candidate for the unobserved order: σ must describe a sequence of steps of non-zero probability from the initial graph to G_T, and we hence have to require

$$L_\theta^t(\sigma_t(G_T)) > 0 \qquad \text{for all } t \leq T \ . \tag{4.10}$$

If so, we call σ **feasible** for G_T. For G_T given and $G_t = \Sigma_t(G_T)$, we use G_t and Σ_t

interchangeably. The target distribution of the SMC bridge sampler at step t is then

$$\gamma_\theta(t) := L_\theta^t(\Sigma_t)P(\Sigma_t) = L_\theta^t(\Sigma_t)\prod_{s=1}^{t-1} q_\theta^{s+1}(\Sigma_{s+1} \mid \Sigma_s)\ , \tag{4.11}$$

which satisfies the recursion

$$\gamma_\theta(t) = \begin{cases} \frac{L_\theta^t(\Sigma_t)}{L_\theta^{t-1}(\Sigma_{t-1})} q_\theta^t(\Sigma_t \mid \Sigma_{t-1})\gamma_\theta(t-1) & \text{if } \Sigma_t \text{ is feasible} \\ 0 & \text{otherwise} \end{cases} .$$

The intractable part is the bridge likelihood ratio $L_\theta^t / L_\theta^{t-1}$. Define

$$h_t(\Sigma_t) := \begin{cases} \mathbb{1}\{L_\theta^t(\Sigma_t) > 0\} & \text{for all } t < T-1 \\ L_\theta^t(\Sigma_t) & \text{for } t = T-1 \end{cases} .$$

As Proposition 4.1 below shows, the apparently crude approximation $L_\theta^t \approx h_t$ still produces asymptotically unbiased samples from the posterior $\mathcal{L}_\theta(G_{1:T} \mid G_T, G_1)$; this is based on methodology developed in Del Moral and Murray (2015) for bridges in continuous state spaces. If Σ_t is feasible, substituting h_t into (4.11) yields the surrogate recursion

$$\gamma_\theta(t) = \frac{q_\theta^t(\Sigma_t \mid \Sigma_{t-1})}{r_\theta^t(\Sigma_t \mid \Sigma_{t-1})}\gamma_\theta(t-1) \qquad \text{for} \quad t < T-1\ .$$

The SMC proposal kernel r_θ^t is hence chosen as the truncation of q_θ^t to feasible permutations,

$$r_\theta^t(\Sigma_t \mid \Sigma_{t-1}) := \frac{\mathbb{1}\{L_\theta^t(\Sigma_t) > 0\} q_\theta^t(\Sigma_t \mid \Sigma_{t-1})}{\tau_\theta^t(\Sigma_{t-1})}\ , \tag{4.12}$$

with

$$\tau_\theta^t := \sum_{\sigma_t : L_\theta^t(\sigma_t) > 0} q_\theta^t(\sigma_t \mid \Sigma_{t-1}) \, .$$

For particles $G_t^i = \Sigma_t^i(G_T)$, the unnormalized SMC weights (2.16) are

$$\tilde{w}_t^i = \begin{cases} \tau_\theta^t(\Sigma_{t-1}^i) & \text{if } t < T-1 \\ q_\theta^t(\Sigma \mid \Sigma_{T-1}^i) \, \tau_\theta^{T-1}(\Sigma_{T-2}^i) & \text{if } t = T-1 \end{cases} . \tag{4.13}$$

The sampling algorithm then generates a graph bridge as follows:

Algorithm 4.1 (Bridge sampling)**.**

- Initialize $G_1^i := G_1$, $\Sigma_1^i \sim \text{Uniform}\{1, \dots, T\}$, and $w_1^i := 1/N$ for each $i \leq N$.
- For $t = 2, \dots, T-1$, iterate:
  - Resample indices $A_t^i \sim \text{MN}(N, (w_{t-1}^i)_i)$.
  - Draw $\Sigma_t^i \sim r_\theta^t(\bullet \mid \Sigma_{t-1}^{A_t^i})$ as in (4.12) for each i.
  - Compute weights as in (4.13) and normalize to obtain w_t^i.
- Resample N complete sequences $G_{1:T}^i = \Sigma^i \sim \text{MN}(N, (w_{T-1}^i)_i)$.

See Figure 4.1 for an illustration. Computation of $h_t(\Sigma_t)$ and $\tau_\theta^t(\Sigma_{t-1})$ is simplified by the constraints on Σ: Given Σ_{t-1}, the requirement that Σ_t must again be a restriction of Σ implies $\Sigma_t = (\Sigma_{t-1}, e_t)$, for some $e_t \in \{1, \dots, T\} \setminus \Sigma_{t-1}$. Within this set, the e_t for which Σ_t is feasible are simply those edges in G_T connected to $\Sigma_{t-1}(G_T)$.

PROPOSITION 4.1. *Let q_θ^t, for $t = 1, \dots, T$, be the Markov kernels defining a sequential network model that satisfies conditions (P1) and (P2). Given an observation G_T, Algorithm 4.1 produces samples that are asymptotically unbiased as $N \to \infty$. That is, for any*

bounded function f on graph sequences,

$$\frac{1}{N}\sum_{i=1}^{N} f(G^i_{1:T}) \xrightarrow{p} \mathbb{E}[f(G_{1:T}) \mid G_T, G_1] \quad \text{as } N \to \infty\,, \tag{4.14}$$

where the expectation is evaluated with respect to the model posterior $\mathcal{L}_\theta(G_{1:T} \mid G_T, G_1)$.

Finally, the particle MCMC methods of the next section will require an unbiased estimate of the bridge likelihood, $L^1_\theta(G_1)$. Define the estimator

$$\hat{L}^1_\theta := \prod_{t=2}^{T-1}\left[\left(\sum_{i=1}^{N}\frac{\tilde{w}^i_t}{N}\right)\left(\frac{\sum_{i=1}^{N} h_{t-1}(G_T \mid G^i_{t-1})\tilde{w}^i_{t-1}}{\sum_{i=1}^{N}\tilde{w}^i_{t-1}}\right)\right]\,. \tag{4.15}$$

Then there is the following:

PROPOSITION 4.2. *Let q^t_θ, for $t = 1, \ldots, T$, be the Markov kernels defining a sequential network model that satisfies conditions (P1) and (P2), and let r^t_θ be the corresponding proposal kernels. Then given an observation G_T and a fixed G_1, $\hat{L}^1_\theta$ as in* (4.15) *is unbiased, that is $\mathbb{E}[\hat{L}^1_\theta] = L^1_\theta(G_1) = p_\theta(G_T \mid G_1)$, for any $N \geq 1$.*

Remark. If estimates exhibit high variance, it is straightforward to modify Algorithm 4.1 to use adaptive resampling (see Del Moral and Murray, 2015), and to replace multinomial resampling above by residual or stratified resampling (e.g. Doucet and Johansen, 2011). If a given model admits a more bespoke approximation h_t to the bridge likelihood, this approximation can be substituted for h_t, following Del Moral and Murray (2015). In this case, some of the equations above require (elementary) modifications; see the proof of a more general version of Proposition 4.2 in Appendix B.2. ◁

4.2.2 Parameter inference

Algorithm 4.1 generates a history of a graph under a model with fixed parameter vector θ. For parameter inference, the parameters are treated as a random variable Θ, with prior distribution $P_{[\Theta]}$, and the task is to generate samples from the joint posterior $\mathcal{L}(\Theta, G_{1:T}|G_1, G_T)$. The sample space of the sampler is thus extended by the domain of Θ. The bridge likelihood $L^1_\Theta(G_1)$ is a marginal likelihood that naturally leads to pseudo-marginal methods (Lin, Liu, and Sloan, 2000; Beaumont, 2003; Andrieu and Roberts, 2009) and particle MCMC (Andrieu, Doucet, and Holenstein, 2010), which uses SMC to compute an unbiased estimate of $L^1_\Theta(G_1)$. Substituting into the Metropolis–Hastings acceptance ratio of a proposal $\tilde{\Theta}$ from a (yet to be specified) proposal distribution $\tilde{q}$ yields

$$\mathbb{P}\{\text{ accept } \tilde{\Theta} \} = \frac{\hat{L}^1_{\tilde{\Theta}} \cdot P_{[\Theta]}(\tilde{\Theta})}{\hat{L}^1_{\Theta} \cdot P_{[\Theta]}(\Theta)} \cdot \frac{\tilde{q}(\Theta \mid \tilde{\Theta})}{\tilde{q}(\tilde{\Theta} \mid \Theta)} \,. \tag{4.16}$$

Using (4.15) and (4.16), Algorithm 4.2 defines a particle marginal Metropolis–Hastings (PMMH) sampler.

Algorithm 4.2.

- Initialize $\Theta^0 \sim P_{[\Theta]}$.
- For $j = 1, \ldots, J$ iterate:
  (1) Draw candidate a value $\tilde{\Theta} \sim \tilde{q}(\bullet \mid \Theta^{j-1})$.
  (2) Run Algorithm 4.1 with parameter $\tilde{\Theta}$ to compute $\hat{L}^1_{\tilde{\Theta}}$ as in (4.15).
  (3) Accept $\tilde{\Theta}$ with probability (4.16) and set $\Theta^j := \tilde{\Theta}$; else set $\Theta^j := \Theta^{j-1}$.
  (4) If $\tilde{\Theta}$ is accepted, select a single graph sequence $G^j_{1:T}$ by resampling from the particles output by Algorithm 4.1; else set $G^j_{1:T} = G^{j-1}_{1:T}$.
- Output the sequence $(\Theta^1, G^1_{1:T}), \ldots, (\Theta^J, G^J_{1:T})$.

The algorithm asymptotically samples the joint posterior $\mathcal{L}(\Theta, G_{1:T} \mid G_1, G_T)$, or the marginal posterior $\mathcal{L}(\Theta \mid G_1, G_T)$ if step (4) is omitted:

PROPOSITION 4.3. *If the proposal density $\tilde{q}(\bullet \mid \bullet)$ is chosen such that the Metropolis–Hastings sampler defined by* (4.16) *is irreducible and aperiodic, Algorithm 4.2 is a PMMH sampler. The marginal distributions $\mathcal{L}(\Theta^j, G^j_{1:T})$ of its output sequence satisfy*

$$\|\mathcal{L}(\Theta^j, G^j_{1:T}) - \mathcal{L}(\bullet \mid G_T, G_1)\|_{\mathrm{TV}} \xrightarrow{j\to\infty} 0 \,. \tag{4.17}$$

This is true regardless of the sample size N generated by Algorithm 4.1 in step (2).

Although the result holds asymptotically independently of N, a larger value of N will generally speed up convergence and reduce variance.

4.3 Particle Gibbs for RW(β, λ) models

The computational cost of Algorithm 4.2 stems mostly from two sources: Each rejection in (3) requires an additional execution of steps (1) and (2) in order to produce a sample distinct from the previous one, and the cost of each such execution may be high. Details depend on the network model, but in most cases, (2) is the most expensive step. Rejections can be addressed by turning the MH algorithm into a Gibbs sampler, which eliminates the acceptance step, and does not require the choice of a proposal kernel $\tilde{q}$. The resulting algorithm constitutes a particle Gibbs (PG) sampler (Andrieu, Doucet, and Holenstein, 2010). Such an algorithm is described below, now specifically for the random walk model. The algorithm uses two sequences of auxiliary variables B_t and K_t, each corresponding to

one model parameter. At each step, some of these variables, as well as the model parameters themselves, can be integrated out, which simplifies sampling significantly.

For a $\mathbf{RW}(\beta, \text{Poisson}_+(\lambda))$ model, the parameter takes the form $\Theta = (\beta, \lambda)$, and we fix a $\text{Beta}(a_\beta, b_\beta)$ prior for β, denoted $P_{[\beta]}$, and a $\text{Gamma}(a_\lambda, b_\lambda)$ prior, denoted $P_{[\lambda]}$, for λ. For compactness, we let $\vartheta := (a_\beta, b_\beta, a_\lambda, b_\lambda)$. To sample a sequence $G_1, \ldots, G_T$ from the model given a pair (β, λ), one can generate two i.i.d. sequences, $\mathbf{B} = (B_1, \ldots, B_T)$ of $\text{Bernoulli}(\beta)$ variables, and $\mathbf{K} = (K_1, \ldots, K_T)$ of $\text{Poisson}_+(\lambda)$ variables. In step t of Algorithm 3.1, a new edge is inserted if $B_t = 1$; otherwise, two vertices are connected by random walk of length K_t. Since $G_{1:T}$, $\mathbf{B}$ and $\mathbf{K}$ are dependent random variables, the kernel q^t_ϑ in Algorithm 4.1 is a function of G_{t-1}, B_t and K_t. As shown in Appendix C, the entries B_t and K_t, along with the model parameters β and λ, can be integrated out of the kernel. $\mathbf{B}$ and $\mathbf{K}$ can therefore be sampled separately from the SMC steps, rather than inside Algorithm 4.1, which improves exploration of the sample space.

In its jth iteration, the algorithm updates $\mathbf{B}$ and $\mathbf{K}$ by looping over the indices $t \leq T$ of $G_{1:T}$. Since Gibbs samplers condition on every update immediately, vectors maintained by the sampler are of the form

$$\mathbf{B}^j_{-t} := (B_1^{j+1}, \ldots, B_{t-1}^{j+1}, B_t^j, \ldots, B_T^j) \quad \text{and} \quad \mathbf{K}^j_{-t} := (K_1^{j+1}, \ldots, K_{t-1}^{j+1}, K_t^j, \ldots, K_T^j)$$

Marginalizing out β, K_t, and λ yields the posterior predictive distribution of B_t^{j+1}; similarly, the predictive distribution of K_t^{j+1} marginalizes out λ, B_t^{j+1}, and β. Both distributions can be obtained in closed form, despite their dependence on $G^j_{1:T}$ (see Appendix C for details). We abuse notation and let the index $j = 0$ refer to the predictive distribution under the prior,

i.e. we write $\mathcal{L}(\mathbf{B}^1|\mathbf{B}^0, G^0_{1:T}) := \int \mathcal{L}(\mathbf{B}^1 \mid \beta) P_{[\beta]}(d\beta)$, and similarly for $\mathcal{L}(\mathbf{K}^1 \mid \mathbf{K}^0, G^0_{1:T})$.

Algorithm 4.3 (Particle Gibbs sampling for $\mathbf{RW}(\beta, \text{Poisson}_+(\lambda))$ models.)**.**

- For $j = 0, \ldots, J$:
  - (1) Draw $(\mathbf{B}^{j+1}, \mathbf{K}^{j+1}) \sim \mathcal{L}(\mathbf{B}^{j+1}, \mathbf{K}^{j+1} \mid \mathbf{B}^j, \mathbf{K}^j, G^j_{1:T})$.
  - (2) Using Algorithm 4.1, with $q^t_\vartheta(\bullet \mid G^j_{t-1}, \mathbf{B}^{j+1}_{-t}, \mathbf{K}^{j+1}_{-t})$ and $N \geq 2$, update $G^{j+1}_{1:T}$ and $a^{j+1}_{2:T}$. At each step t, substitute the previous iteration's bridge $G^j_{1:t}$ for the particle with index a^j_t (see below).
  - (3) Draw $\beta^{j+1} \mid \mathbf{B}^{j+1}$ and $\lambda^{j+1} \mid \mathbf{K}^{j+1}$ from their conjugate posteriors.
- Output the sequence $(G^1_{1:T}, \beta^1, \lambda^1), \ldots, (G^J_{1:T}, \beta^J, \lambda^J)$.

The algorithm is used for inference in the random walk model in all experiments reported in Section 3.3. It constitutes a blocked Gibbs sampler with blocks (β, λ), $(\mathbf{B}, \mathbf{K})$, and $G_{1:T}$. Due to the marginalization described above, the parameter values β^j and λ^j generated in (3) are not used in the execution of the sampler; they only serve as output. Step (2), which seeds Algorithm 4.1 with a bridge $G^j_{1:T}$ generated during the previous iteration, is a **conditional SMC** step that biases the sampler towards $G^j_{1:T}$ and is necessary to make Algorithm 4.3 a valid Gibbs sampler. Although the resampled indices a_t have no effect on the sampled sequence $G_{1:T}$, the Gibbs sampler is defined on an augmented space that includes the random variables generated during the SMC steps, and so requires that we include them in the conditional SMC update. See Andrieu, Doucet, and Holenstein (2010) for more on conditional SMC.

Proposition 4.4. *Algorithm 4.3 is a valid particle Gibbs sampler. For any $N \geq 2$, it generates a sequence $(G^j_{1:T}, \beta^j, \lambda^j)_j$ whose marginal laws converge as*

$$\|\mathcal{L}(G^j_{1:T}, \beta^j, \lambda^j) - \mathcal{L}(G_{1:T}, \beta, \lambda \mid G_T, G_1)\|_{\text{TV}} \xrightarrow{j\to\infty} 0\,. \tag{4.18}$$

to the model posterior $\mathcal{L}(G_{1:T}, \beta, \lambda \mid G_T, G_1)$.

4.3.1 Variance reduction and practical considerations in PG sampling

A well-documented problem encountered by SMC methods is that of *path degeneracy*, when the particles concentrate on only a few paths. Each time the particles are resampled, fewer and fewer distinct paths are propagated. On the other hand, not resampling results in *weight degeneracy*, where the weights concentrate on only a few particles. Both issues lead to SMC estimates with high finite-sample variance (see, e.g. Doucet and Johansen, 2011). As noted by a number of authors, PG samplers may be "sticky" due to the issues inherited from SMC: by conditioning on the previous latent trajectory $G_{1:T}^{j-1}$ and repeatedly resampling the particles, the conditional SMC update may make only a few changes, resulting in $G_{1:T}^{j}$ being nearly the same as $G_{1:T}^{j-1}$.

A number of methods have been developed to address this problem. In traditional SMC, *adaptive resampling*, where the particles are resampled only if the effective number of particles falls below a fixed threshold, is typically employed in practice (Doucet and Johansen, 2011). Backward simulation (Lindsten and Schön, 2013) is a set of smoothing techniques for SMC that can improve performance, and has been adapted to particle MCMC (Whiteley, 2010; Lindsten and Schön, 2012; Chopin and Singh, 2015). A related method, particle Gibbs with ancestor sampling (PGAS) (Lindsten, Jordan, and Schön, 2014), has proven successful in a range of applications.

PGAS adds a step to Algorithm 4.1: after the particles have been propagated to step t, each particle's ancestor is resampled from the set of particles at step $t-1$ from which it could have been propagated. Although the particles still degenerate to one path, the

ancestor sampling step causes the degenerate path to be different than $G_{1:T}^{j-1}$. In our own experiments with the $\mathbf{RW}(\beta, \text{Poisson}_+(\lambda))$ model, we found that although PGAS improved performance, gains typically were not sufficient to justify the additional computation. The culprit is the high-dimensional discrete state space. In order for ancestor sampling to work well, each particle i at step t must have many possible ancestors from step $t-1$, i.e. there must be many particles j for which $q_\theta^t(G_t^i \mid G_{t-1}^j) > 0$. For unconstrained models with a continuous state space, the condition is trivially satisfied for all i and j. In a discrete state space, however, the condition is often not satisfied for $i \neq j$. That the model constrains the graphs to be connected for all t further constrains the possible transitions. As such, we observe that a model allowing for multiple connected components (with a mechanism for merging components) may benefit more from PGAS.

In our experiments, we found that breaking the conditional SMC update into sub-blocks $G_{1:b_1}, G_{b_1:b_2}, \ldots, G_{b_n:T}$, as suggested in Andrieu, Doucet, and Holenstein (2010), improves performance by reducing the necessary number of particles. It also reduces the memory requirements. Finally, although adaptive resampling based on effective sample size within in the conditional SMC update does not satisfy the assumptions necessary for the theoretical guarantees of Proposition 4.4, we found that it improves exploration of the space of latent trajectories $G_{1:T}$ without degrading the quality or accuracy of parameter sampling. Huggins and Roy (2015) introduce a novel form of effective sample size, called ∞-ESS, and apply it to particle Gibbs samplers with adaptive resampling; this seems to be a promising direction for future particle MCMC implementations.

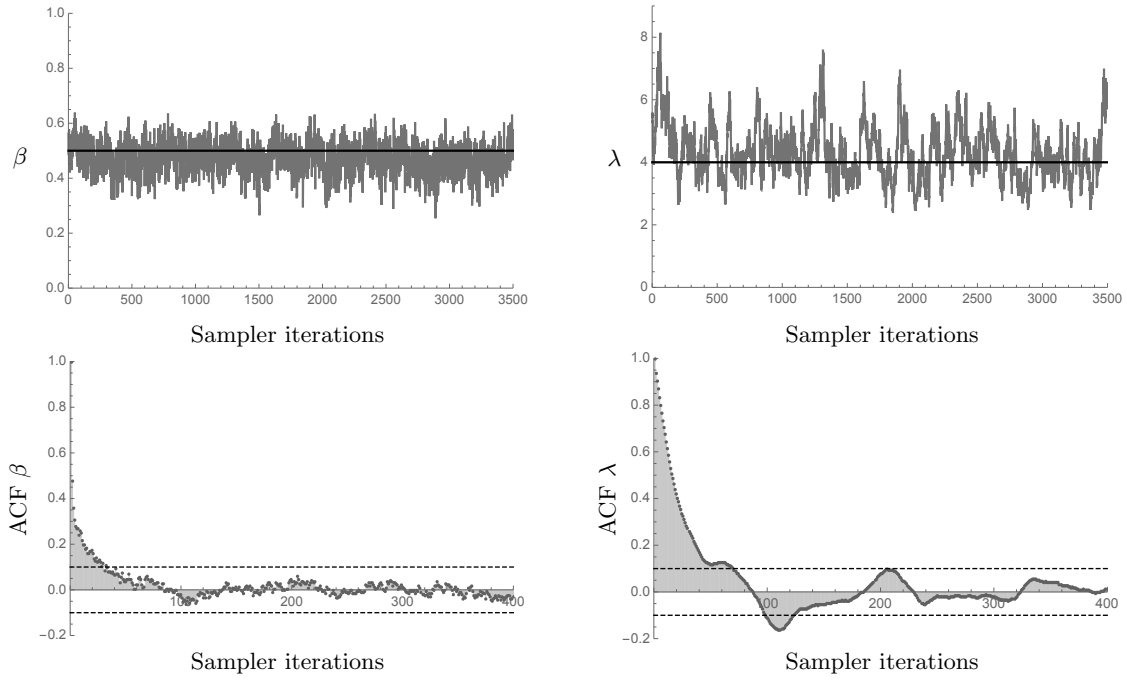


Figure 4.2: Top: Posterior sample traces for a PG sampler fit to a synthetic graph G_T generated with $\beta = 0.5$ and $\lambda = 4$ (solid black lines). Bottom: The corresponding autocorrelation functions.

4.3.2 Sampler diagnostics on synthetic data

To assess sampler performance, we test the sampler's ability to recover parameter values for graphs generated by the model. For each such experiment, a single graph G_T with $T = 250$ edges is generated for fixed values of β and λ. The joint posterior distribution of β and λ given G_T is then estimated using Algorithm 4.3, run with a moderate number ($N = 100$) of particles. Regardless of the input parameter value, a uniform prior is chosen for β, and a Gamma$(1, 4)$ prior for λ. For the sake of brevity, we report results only for the $\mathbf{RW}_{\mathrm{U}}(\beta, \mathrm{Poisson}_+(\lambda))$ model, on simple graphs (which pose a harder challenge for inference than multigraphs, since edge multiplicities provide additional information about the history of the graph). Figure 4.2 shows example traces of the samplers and the corresponding

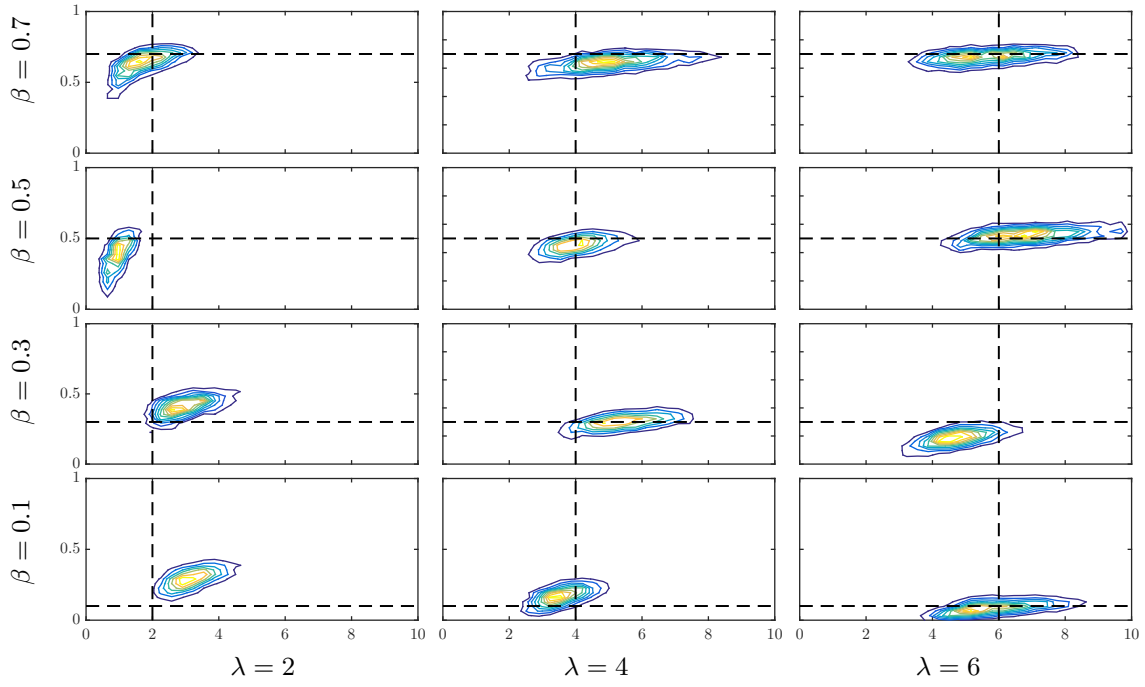


Figure 4.3: Joint posterior distributions, given graphs generated from an $\mathbf{RW}_{\mathrm{U}}$ model for different parameter settings. β is on the vertical axis, λ on the horizontal axis, and generating parameter values are shown by dashed lines.

autocorrelation functions. The samplers mix and converge quickly, and as one would expect, only λ displays any noticeable autocorrelation. Posteriors for various input parameters are shown in Figure 4.3. Clearly, recovery of model parameters from a single graph is possible. The effect of the model's modification to generate simple graphs—the use of step (3') rather than (3) in Algorithm 3.1—is apparent for $\lambda = 2$: For small values of λ, a significant proportion of random walks ends at their starting point, resulting in the insertion of a new vertex. Since the model otherwise inserts new vertices as an effect of β, the two parameters are correlated in the posterior, which is clearly visible in Figure 4.3, especially for intermediate values of β. For applications to real data, see Section 3.3.

4.4 MCMC sampling for fully observed sequential models

When a Bayesian approach to inference is preferred, it is straightforward to construct MCMC samplers for fully observed sequential models. In particular, Algorithms 4.2 and 4.3 are modified by excluding the SMC update, and making the following modifications:

- For a Metropolis–Hastings sampler, step (2) of Algorithm 4.2 is replaced by an evaluation of the model likelihood $L_T(\beta, \lambda)$, as in (4.2), and the acceptance ratio (4.16) is modified accordingly.

- For a Gibbs sampler, step (1) of Algorithm 4.3 is modified to condition on the observed $G_{1:T}$, rather than the previous iteration's sample $G^j_{1:T}$.

As an example, we generated a sequence $G_1, \ldots, G_T$, with $T = 400$ edges, from a $\mathbf{RW}_{\mathrm{U}}(\beta, \lambda)$ model with $\beta = 0.2$ and $\lambda = 4$. We fitted a $\mathbf{RW}_{\mathrm{U}}(\beta, \lambda)$ model to the sequence using a Gibbs sampler as in Algorithm 4.3 and the modification specified above. Typical sampler traces are shown in Figure 4.4. The traces show that the sampler converges quickly. The marginalizations described in Section 4.3 allow good exploration of the sample space; autocorrelation between samples typically becomes negligible after 30-40 Gibbs updates.

4.5 Discussion

The methods developed here demonstrate that tractable and reliable inference is possible for sequential models of data even when the data consists of one observation at the end of the sequence. That inference is possible when the latent order is relevant to the model is somewhat surprising. Indeed, it might be argued that a primary reason for assuming

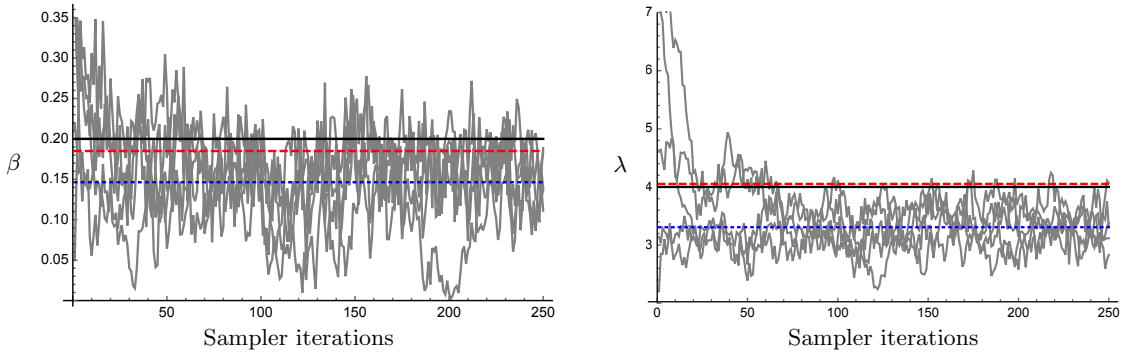


Figure 4.4: Posterior sample traces (gray lines) for a Gibbs sampler fit to a synthetic graph sequence $G_{1:T}$ with $T = 400$ edges, generated with $\beta = 0.2$ and $\lambda = 4$ (solid black lines). The maximum a posteriori estimates from 50,000 samples are displayed (dotted blue lines), along with the maximum likelihood estimates (dashed red lines) based on the variables $B_{2:T}$ and $K_{2:T}$ generated with $G_{1:T}$.

exchangeability is to avoid the issue of imputing an order. The key is that the observed structure retains signs of the latent order; the dependence between edges and vertices that makes inference difficult is also necessary to perform inference at all.

Computational complexity. The inference algorithms in Section 4.2 work well for moderately sized graphs (of up to a few hundred vertices). The dominant source of computational complexity is sampling the latent order Σ, which even for the simplest models scales at least quadratically in the number of edges in G_T. The computational role of Σ is similar to the latent vertex positions in inference algorithms for graphon models (which have comparable input size limitations) (Lloyd, Orbanz, Ghahramani, and Roy, 2012).

Although improvements in computational efficiency or approximations are certainly possible, we do not expect the methods to scale to very large graphs. For example, one might replace in the inference algorithms the mixed random walk probability matrix Q by a low-rank approximation. However, the best low-rank approximation is not a right stochastic

matrix (like Q), or even guaranteed to have non-negative entries. Constrained low-rank approximations are possible, e.g. Ho and Van Dooren (2008) and Deng and Huang (2015), but the complexity of sampling the latent order remains high. Very large graphs are not the subject of this work. It is crucial that, under the random walk model, graphs of feasible size do provide sufficient data for reliable inference. For more complicated models, that might not be the case.

When the sequence $G_{1:T}$ is fully observed, the inference methods in Section 4.1 and Section 4.4 may scale to larger graphs, and more complicated models may be tractable. In that case, we expect that computational improvements via approximations would be significant.

Extensions and generalizations. One obvious extension of the models studied here is the use of different random walk length distributions, P. Much of the sampling efficiency for the Poisson case derived from marginalizing the random walk length out of the SMC kernel at each step. Proposition 3.1 provides a method for general P, and thus designing efficient particle MCMC samplers for general P seems to be straightforward. Additional marginalization, of the random walk parameter λ, was possible because of the Poisson-Gamma conjugacy relationship and the resulting closed-form posterior predictive distribution. Other candidates for P that would admit similar marginalizations are Binomial (with a Beta prior), and categorical (with a Dirichlet prior).

Finally, although the particle methods in Section 4.2 were developed in the context of sequential network data, they can be used for any type of composite data for which a sequential model satisfies the Markov property (P1) and the monotonicity property (P2). For example, inference can be performed on non-exchangeable models of partition structures or

permutations.

Related work. Existing work on the use of sampling algorithms for inference in sequential network models mainly follows one of two general approaches: Importance sampling (Wiuf, Brameier, Hagberg, and Stumpf, 2006), or approximate Bayesian computation (ABC) schemes (Thorne and Stumpf, 2012). ABC relies on heuristic approximations of the likelihood, and therefore is not guaranteed to sample from the correct target distribution. The basic approach used above—conduct inference by imputing graph sequences generated by a suitable sampler—seems to be due to Wiuf, Brameier, Hagberg, and Stumpf (2006), based on earlier work by R. C. Griffiths and Tavaré (1994a,b) on ancestral inference for coalescent models. The work of Wang, Jasra, and De Iorio (2014) is related to ours in that it employs particle MCMC, for a particular sequential model. All these methods are, in short, applicable if the sequential model in question is very simple. Otherwise, they suffer (1) from the high variance of estimates that is a hallmark of importance sampling (Doucet and Johansen, 2011), and (2) infeasible computational cost. The algorithm in Wang, Jasra, and De Iorio, 2014, for example, samples backwards through the sequence generated by the model, and for each reverse step $G_t \to G_{t-1}$ requires computing the probability of every possible forward step $G'_{t-1} \to G_t$ under the given model; even for the (still rather simple) random walk model, that is no longer practical.

A separate body of work applies SMC to a related problem, inference in probabilistic graphical models (e.g. Naesseth, Lindsten, and Schön, 2014). Here, models lack a sequential structure, which is addressed algorithmically by constructing the dependence structure clique by clique via artificial intermediate distributions; not surprisingly, the design of good

intermediate distributions turns out to be crucial. In a sequential graph model, these intermediate distributions are given by the model.

Chapter 5

Nested urn models of partitions and graphs

Crucial to the results in Section 3.2.2 and to the inference algorithms in Chapter 4 are the random sequence of times at which new vertices appear, $S_{1:\infty}$. Furthermore, the random walk model generates a sequence of edges that is not exchangeable; the lack of exchangeability stems from the random walk, and from the constant probability of a new vertex appearing. It is natural, therefore, to consider what effect the constant probability has on the properties of the resulting graph sequences. In this chapter, we study simple preferential attachment-type models, and randomize $S_{1:\infty}$ with different distributions. Due to their relative simplicity, we begin by analyzing partitions, and turn to graphs in Section 5.4.

5.1 Partitions from nested urn sequences

The building block of models studied in this chapter is a generalization of the basic Pólya urn, the properties of which were reviewed in Section 2.2. The basic Pólya urn scheme can be extended by adding new colors to the urn and augmenting the sampling probabilities. More precisely, let $C_1, C_2, \ldots, C_t := (C_t)_{t\geq 1}$ be a sequence of random variables with values in a measurable space $(\Omega_c, \mathcal{C})$, and let ν be a non-atomic probability distribution on Ω_c. In what follows, we will refer to C_t as the color of the t-th ball drawn from the urn. We denote the number of balls of the j-th color (in order of appearance) after t steps by $n_j(t)$.

Of central importance to this chapter are the steps in $C_1, C_2, \ldots$ when a new color first appears. We call such steps **arrival times**,[1] denoted $1 = s_1, s_2, s_3, \ldots$, with inter-arrival times $\delta_j := s_j - s_{j-1}$. Note that $\delta_j > 0$ for all $j \in \mathbb{N}_+$. For urns with a finite number r of colors, we use the convention $s_j = \infty$ for $j > r$.

Given a sequence of arrival times, $s_{1:\infty} := (s_1 < s_2 < \ldots)$, a ball of a new (random) color is added on steps $t = s_j$; for the remainder of the steps, the urn is updated as the usual multicolor Pólya urn, resulting in the following predictive distribution:

$$\mathbb{P}(C_{t+1} \in \bullet \mid C_{1:t}, s_{1:\infty}) = \mathbb{1}_{t+1}(s_{k(t)+1})\nu(\bullet) + (1 - \mathbb{1}_{t+1}(s_{k(t)+1})) \sum_{j=1}^{k(t)} \frac{n_j - \alpha}{t - \alpha k(t)} \delta_{C_j^*}(\bullet) \,. \tag{5.1}$$

Although it is a straightforward modification, explicitly separating the arrival times from the Pólya urn process allows us to study a number of seemingly different urn models within

[1] Other names used in the literature include increments (Nacu, 2006) and record indices (e.g. R. C. Griffiths and Spanò, 2007).

the same framework.

Algorithm 5.1 (Multicolor Pólya urn scheme with fixed arrival times)**.**
Fix $\alpha \in (-\infty, 1)$ and $s_{1:\infty} := (1 = s_1 < s_2 < \dots)$.

- Begin with one ball of color $C_1^* \sim \nu$, and set $C_1 = C_1^*$, $n_{C_1^*}(1) = 1$. For $t \geq 2$:
  - If $t = s_j$ for some j, add a ball of new, distinct color $C_j^* \sim \nu$ to the urn and set $C_t = C_j^*$.
  - Else, draw a ball of color C_i^* with probability proportional to $n_{C_i^*}(t-1) - \alpha$, and replace it along with an additional ball of the same color. Set $C_t = C_i^*$, $n_{C_i^*}(t) = n_{C_i^*}(t-1) + 1$, and $n_{C_j^*}(t) = n_{C_j^*}(t-1)$ for $j \neq i$.

A partition Π_t of $[t] := \{1, \dots, t\}$ is constructed by grouping balls of the same color, i.e. $A_j = \{t : C_t = C_j^*\}$, with the blocks in order of their least elements. Note that the least element of block A_j is s_j. Analysis of Algorithm 5.1 is simplified by observing that it is equivalent to the following:

Algorithm 5.2 (Nested Pólya urn scheme with fixed arrival times)**.**
Fix $\alpha \in (-\infty, 1)$ and $s_{1:\infty} \in \mathbb{N}_+^\infty$.

- Begin with $s_2 - 1$ balls of color $C_1^* \sim \nu$ in urn $u = 1$.
- For steps $t \geq s_2$, proceed as follows:
  (1) If $t = s_j$ for some j, create a new urn with 1 ball of a new, distinct color $C_j^* \sim \nu$ and $(n_{C_i^*}(t-1))_{i<j}$ balls of the first $j-1$ colors. Set $C_t = C_j^*$.
  (2) Else, starting with urn $u = \max\{j : s_j \leq t\}$ recursively do:
    - Draw a ball of color C_i^* from urn u with probability proportional to $n_{C_i^*}(t-1) - \alpha$; replace it and add a ball of the same color. If the ball was some color other than C_u^*, repeat with urn $u - 1$; else set $C_t = C_u^*$ and $n_{C_u^*}(t) = n_{C_u^*}(t-1) + 1$, and go to step $t + 1$.

We call a partition process $(\Pi_t)_{t \geq 1}$ generated by Algorithm 5.1 or Algorithm 5.2 a **nested Pólya partition** (nPP) process on $\widetilde{\mathcal{P}}$, the space of partitions ordered by their least elements. We denote its law as $\mathbf{nPP}(\alpha \mid s_{1:\infty})$, which is a regular conditional probability distribution on $\mathbb{N}_+^\infty \times \widetilde{\mathcal{P}}$.

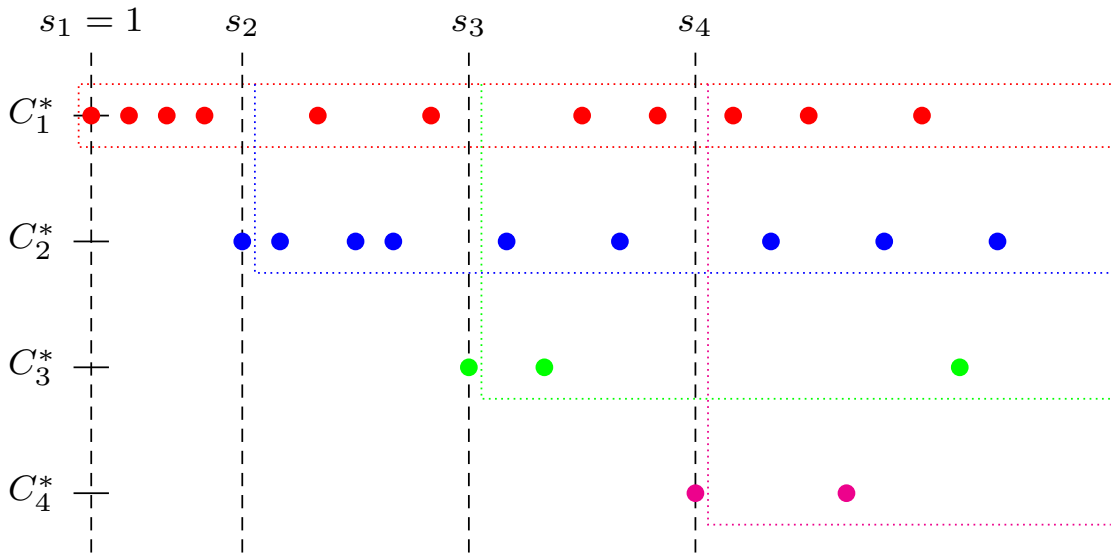


Figure 5.1: A nested exchangeable sequence. Dotted boxes contain observations that are exchangeable: All observations of C_1^* are (trivially) exchangeable with each other, all observations of C_1^* and C_2^* that occur at $t > s_2$ are exchangeable, and so on.

We define the **conditional partition probability function** (CPPF) of Π_t with $k(t)$ blocks to be

$$\mathbb{P}(\Pi_t \mid s_{1:k(t)}) = \frac{1}{\Gamma(t - \alpha k(t))} \prod_{j=1}^{k(t)} \frac{\Gamma(s_j - \alpha j)}{\Gamma(s_j - 1 - \alpha(j-1) + \mathbb{1}_1(j))} \frac{\Gamma(n_j(t) - \alpha)}{\Gamma(1-\alpha)} \,. \tag{5.2}$$

In contrast to exchangeable random partitions, the distributions of which are invariant to all permutations σ of $[t]$, the CPPF of the nested Pólya partition is invariant to a subset of all permutations of $[t]$. Let $\mathcal{S}_t$ be the set of all permutations of $[t]$, and $\mathcal{S}_t^\delta \subset \mathcal{S}_t$ is the subset of permutations that, when applied to Π_t, leave (5.2) unchanged. Further, let $\mathcal{C}_t^*$ be the set of unique colors observed up to and including step t. A sufficient condition for $\sigma \in \mathcal{S}_t^\delta$ is that it satisfies both of the following properties:

$$C_{\sigma(s_j)} = C_{s_j} \quad \text{for all} \quad j \quad \text{s.t.} \quad s_j \leq t$$

$$C_{\sigma(s)} \in \mathcal{C}_s^* \quad \text{for all} \quad s \leq t \; .$$

The restrictions induce a *nested exchangeability* structure in the sequence $C_1, C_2, \ldots$. Let $(t_{k,i})_{i \geq 1}$ be the indices of the occurrences of the colors $C_1^*, \ldots, C_k^*$ that occur after s_k. Then the subsequence $(C_{t_{k,i}})_{i \geq 1}$ forms an exchangeable sequence in i, and (5.2) is invariant to all σ that act only on $(C_{t_{k,i}})_{i \geq 1}$. See Figure 5.1 for an illustration.

In what follows, we use the conventions that a Beta(a, ∞) distribution is a point mass at 0, and a Beta$(a, 0)$ distribution is a point mass at 1.

In light of the nested exchangeability structure of the nPP, and of the properties of binary exchangeable sequences, in particular the two-color paintbox scheme from Section 2.2.4, a nested paintbox scheme is natural. The following algorithm defines such a sampling scheme.

Algorithm 5.3 (Nested Pólya Paintbox with fixed arrival times)**.**
Fix $\alpha \in (-\infty, 1)$ and $s_{1:\infty} \in \mathbb{N}_+^\infty$.

(1) Let $\psi_1, \psi_2, \ldots$ be independent beta random variables sampled as:

$$\psi_j \sim \text{Beta}(1-\alpha, s_j - 1 - (j-1)\alpha) \; . \tag{5.3}$$

(2) Partition the unit interval into sub-intervals as:

$$\phi_j = \psi_j \cdot \prod_{i=j+1}^{\infty} (1-\psi_i) \; , \quad W_j = \sum_{i=1}^{j} \phi_i \; , \quad \text{and} \quad I_j = [W_{j-1}, W_j) \; . \tag{5.4}$$

(3) For $t = 1, 2, \ldots$ sample $U_t \overset{\text{iid}}{\sim} \text{Uniform}[0,1]$, and define

$$\tilde{U}_t := \begin{cases} W_{r-1} + \phi_r U_t & \text{if } t = s_r \\ W_r U_t & \text{if } s_r < t < s_{r+1} \end{cases} \; . \tag{5.5}$$

(4) Construct the partition Π_s by the rule:

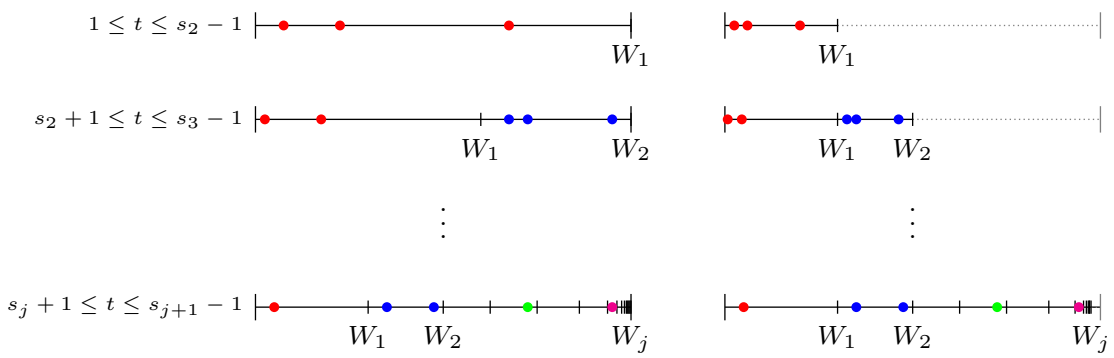
$$t \in A_j \quad \text{iff} \quad \tilde{U}_t \in I_j, \quad t \leq s \; .$$

Although the partition of the unit interval in Algorithm 5.3 potentially involves an infinite number of random variables, each round of sampling can be conducted using a finite number of variables. Observe that

$$\begin{aligned} \mathbb{P}(U_t \in I_k) = \frac{\phi_k}{W_r} &= \frac{\psi_k \prod_{i=k+1}^{\infty}(1-\psi_i)}{\prod_{m=r+1}^{\infty}(1-\psi_m)} \quad \text{for} \quad s_r < t < s_{r+1} \\ &= \psi_k \prod_{i=k+1}^{r} (1-\psi_i) \qquad (5.6) \\ &= \frac{\psi_k \prod_{i=2}^{k}(1-\psi_i)^{-1}}{\sum_{\ell=2}^{r} \psi_\ell \prod_{m=1}^{\ell}(1-\psi_m)^{-1}} \; . \qquad (5.7) \end{aligned}$$

See Figure 5.2 for an illustration.

Proposition 5.1. *Fix $\alpha \in (-\infty, 1)$ and $s_{1:\infty}$. Let $(\Pi_t)_{t \geq 1}$ have law* $\mathbf{nPP}(\alpha \mid s_{1:\infty})$.



Sampling according to (5.6). Sampling according to (5.7).

Figure 5.2: The nested Pólya paintbox sampling scheme. On the left, $W_j = 1$ in each sampling round j, but the intervals I_k for $k \le j$ change. On the right, each interval I_k is constant across each sampling round j, but W_j changes.

Further, let $(\Pi_t^{\psi})_{t\ge 1}$ *be generated by Algorithm 5.3. Then*

$$\Pi_t \stackrel{d}{=} \Pi_t^{\psi}, \quad \textit{for all} \quad t \ge 1 \,.$$

Proof. Each urn in Algorithm 5.2 is a two-color Pólya urn; therefore, as $t \to \infty$, the proportion of color c_j balls in each urn $j \ge 1$ converges to a random limit $\psi_1 = 1$ and $\psi_j \sim \text{Beta}(1-\alpha, s_j - 1 - (j-1)\alpha)$. The sequence of balls for each urn is exchangeable, and by de Finetti's theorem (Theorem 2.3), each urn sequence is conditionally i.i.d. given ψ_j. For $s_j < t < s_{j+1}$, (5.6) is the probability that the t-th ball is of color c_k, which establishes the equality in distribution of the urn scheme and the nested paintbox scheme.

Alternatively, define $\bar{n}_j(t) := \sum_{i=1}^{j} n_i(t)$. Then

$$\mathbb{E}_{(\psi_j)_j}[\mathbb{P}(\Pi_t, \psi_{1:k(t)} \mid s_{1:k(t)})]$$

$$= \prod_{j=1}^{k(t)} \frac{\Gamma(s_j - j\alpha)}{\Gamma(1-\alpha)\Gamma(s_j - 1 - (j-1)\alpha + \mathbb{1}_1(j))} \int_0^1 \psi_j^{n_j(t)-\alpha} (1-\psi_j)^{\bar{n}_{j-1}(t)-(j-1)\alpha} d\psi_j \tag{5.8}$$

$$= \prod_{j=1}^{k(t)} \frac{\Gamma(s_j - j\alpha)}{\Gamma(1-\alpha)\Gamma(s_j - 1 - (j-1)\alpha + \mathbb{1}_1(j))} \frac{\Gamma(n_j(t)-\alpha)\Gamma(\bar{n}_{j-1}(t)-(j-1)\alpha)}{\Gamma(\bar{n}_j(t)-j\alpha)} \tag{5.9}$$

$$= \frac{1}{\Gamma(t-\alpha k(t))} \prod_{j=1}^{k(t)} \frac{\Gamma(s_j - \alpha j)}{\Gamma(s_j - 1 - (j-1)\alpha + \mathbb{1}_1(j))} \frac{\Gamma(n_j(t)-\alpha)}{\Gamma(1-\alpha)} \tag{5.10}$$

$$= \mathbb{P}(\Pi_t \mid s_{1:k(t)}) \, . \tag{5.11}$$

□

Remark. We make the following observations in relation to previous work:

- A version of Algorithm 5.3, with sampling rounds of fixed length ℓ, was used by Berger, Borgs, Chayes, and Saberi (2005, 2014) to analyze the Benjamini–Schramm limit (Benjamini and Schramm, 2001) of PA graphs. See also Durrett (2006).

- In R. C. Griffiths and Spanò (2007), the representation (5.4) was derived (with a different proof) for the limiting proportions of block sizes in infinite EGPs, i.e. when $S_{1:\infty}$ are the random arrival times generated by an EGP; it was noted there that conditionally on $S_{1:\infty} = s_{1:\infty}$, the elements of the sequence

$$\tilde{\xi} := \left(\frac{\xi_j}{\sum_{i=1}^{j} \xi_i} \right)_{j>1} = (\psi_j)_{j>1}$$

  are mutually independent. $(\xi_j)_{j\geq 1}$ is therefore a *neutral-to-the-left* (NTL) process, and R. C. Griffiths and Spanò (2007) show that the NTL property characterizes the class

of EGPs. That characterization holds only in the case that $(\xi_j)_{j\geq 1}$ are the limiting proportions; the NTL property holds for nPPs in general.

- As noted in R. C. Griffiths and Spanò (2007), (5.4) is a reversed version of a neutral-to-the-right (NTR) process called the Beta–Stacy process, as developed in Walker and Muliere (1997). It is interesting to observe that the nested urn process in Algorithm 5.2 is a reversed version of the urn process formulated in Walker and Muliere (1997).

◁

In the basic Pólya urn, the proportions of white and black balls converge to a well-defined random limit; we might ask whether the scaled counts converge in the nested Pólya urn, and if so, what is the appropriate scaling and what are the limit objects? Algorithm 5.3 describes a constructive representation of the limiting objects with fixed $s_{1:\infty}$. The scaled count sequence requires further analysis.

Fix a sequence of arrival times $s_{1:\infty}$, and encode a finite sequence of shifts by a vector $\mathbf{p} = (p_1, p_2, \dots) \in \mathbb{N}^\infty$. That is, $p_j = 0$ for all $j > k^*$, for some k^*. Let the partial sums of the shifts be denoted by $\tilde{p}_j = \sum_{i=1}^{j} p_i$. Consider the partition process $\Pi_t^{\mathbf{p}}$ where the arrival times are shifted such that $s_j^{\mathbf{p}} = s_j + \tilde{p}_{j-1}$; the additional steps due to p_j are allocated to block j.

For $t \geq s_{k^*}$, compare the unshifted partition process Π_t, with block counts $(n_j(t))_{j=1}^{k(t)}$, to the shifted partition process $\Pi_t^{\mathbf{p}}$ with block counts $(n_j(t) + p_{j+1})_{j=1}^{k(t)}$. The ratio of the CPPFs is

$$Z_{\mathbf{p}}(t) := \frac{\mathbb{P}(\Pi_t^{\mathbf{p}} \mid s_{1:\infty} + \tilde{p}_{1:\infty})}{\mathbb{P}(\Pi_t \mid s_{1:\infty})}$$

$$= \frac{\Gamma(t-\alpha k(t))}{\Gamma(t-\alpha k(t)+\tilde{p}_{k^*})} \left[\prod_{j=1}^{k^*} \frac{\Gamma(n_j(t)-\alpha+p_j)}{\Gamma(n_j(t)-\alpha)}\right] \cdots$$
$$\times \left[\prod_{j=1}^{k(t)-1} \frac{\Gamma(s_{j+1}-1-\alpha j)\Gamma(s_{j+1}-\alpha(j+1)+\tilde{p}_j)}{\Gamma(s_{j+1}-1-\alpha j+\tilde{p}_j)\Gamma(s_{j+1}-\alpha(j+1))}\right] . \tag{5.12}$$

The likelihood ratio (5.12) may also be interpreted as the probability of the perturbed process conditioned on the unperturbed process. Note that $Z_{\mathbf{p}}(t)$ is well-defined even for $\mathbb{R}$-valued p_i, but its interpretation is no longer clear. As in the basic Pólya urn in Section 2.2, it is a martingale.

PROPOSITION 5.2. *Fix a vector of shifts* $\mathbf{p} = (p_j)_{j\geq 1}$, *with* $p_j > -(1-\alpha)$ *for* $j > 1$, *such that* $p_j = 0$ *for all* $j > k^*$. *Then* $Z_{\mathbf{p}}(t)$ *is a nonnegative martingale with respect to* $(\mathcal{A}_t)_{t\geq 1}$, *the filtration generated by the partition process, for* $t \geq s_{k^*}$. *Furthermore, if* $p_1, \ldots, p_r > -(1-\alpha)/2$, *then* $Z_{\mathbf{p}}(t)$ *converges in* L_2 *and therefore in* L_1.

PROOF. Observe that $Z_{\mathbf{p}}(t)$ is nonnegative by construction, and

$$\mathbb{E}\left[\frac{\Gamma(t+1-\alpha k(t+1))}{\Gamma(t+1-\alpha k(t+1)+\tilde{p}_{k^*})} \prod_{j=1}^{k^*-1} \frac{\Gamma(n_j(t+1)-\alpha+p_j)}{\Gamma(n_j(t+1)-\alpha)} \,\middle|\, \mathcal{A}_t, s_{1:\infty}\right]$$
$$= \frac{\Gamma(t-\alpha k(t))}{\Gamma(t-\alpha k(t)+\tilde{p}_{k^*})} \prod_{j=1}^{k^*-1} \frac{\Gamma(n_j(t)-\alpha+p_j)}{\Gamma(n_j(t)-\alpha)} \cdots$$
$$\times \left(1 + \mathbb{1}\{t+1 = s_{k(t)+1}\} \frac{\Gamma(t-\alpha k(t)+\tilde{p}_{k^*})\Gamma(t+1-\alpha(k(t)+1))}{\Gamma(t-\alpha k(t))\Gamma(t+1-\alpha(k(t)+1)+\tilde{p}_{k^*})}\right) .$$

The indicator is non-zero only for arrival times; a rearrangement shows that $Z_{\mathbf{p}}(t)$ is therefore a martingale. Furthermore, by the properties of the Gamma function, for a finite

constant $C_{\mathbf{p}}$ (see Hofstad (2016), Chapter 8.7 for proof of a similar inequality),

$$Z_{\mathbf{p}}(t)^2 \leq C_{\mathbf{p}} Z_{2\mathbf{p}}(t) \; .$$

Now, $\mathbb{E}[Z_{\mathbf{p}}(t) \mid s_{1:\infty}] = \mathbb{E}[Z_{\mathbf{p}}(s_{k^*}) \mid s_{1:\infty}] < \infty$ for all $p_i > -(1-\alpha)$, so $\mathbb{E}[Z_{2\mathbf{p}}(t)]$ is finite when each $p_i > -(1-\alpha)/2$. This implies that $Z_{\mathbf{p}}(t)$ is an L_2-bounded martingale; it therefore converges in L_2 and also in L_1. □

$Z_{\mathbf{p}}(t)$ is our main tool for establishing asymptotic properties of the urn sequence. The next theorem states that the joint moments of the ordered block size sequence converge to a random limit.

THEOREM 5.3 (Limiting count sequence with fixed arrival times). *Fix $\alpha \in (-\infty, 1)$ and $s_{1:\infty} \in \mathbb{N}_+^\infty$. Let $(\Pi_t)_{t\geq 1}$ be a nested Pólya partition with law $\mathbf{nPP}(\alpha \mid s_{1:\infty})$. Assume*

$$\lim_{j\to\infty} \frac{s_j}{j} = \mu_\delta \quad \textit{for some} \quad \mu_\delta \in \mathbb{R}_{>0} \cup \{\infty\} \; .$$

Then the scaled count sequence converges jointly to a random limit: For each $r \in \mathbb{N}_+$,

$$\lim_{t\to\infty} M_{\mathbf{p}}^*(s_{1:\infty}) t^{-\gamma} \mathbb{E}[n_1(t)^{p_1} \cdots n_r(t)^{p_r}] = \mathbb{E}[\xi_1^{p_1} \cdots \xi_r^{p_r}] \quad \textit{where} \quad \gamma = 1 - \frac{1-\alpha}{\mu_\delta - \alpha} \; , \tag{5.13}$$

where $M_{\mathbf{p}}^*(s_{1:\infty})$ is a constant that depends on $\mathbf{p}$. The limiting law $\mathcal{L}(\xi_1, \ldots, \xi_r)$ is characterized by the joint moments

$$\mathbb{E}[\xi_1^{p_1}, \ldots, \xi_r^{p_r}] = \prod_{j=1}^{r} \frac{\Gamma(1-\alpha+p_j)}{\Gamma(1-\alpha)} \frac{\Gamma(s_j - j\alpha + \tilde{p}_{j-1})}{\Gamma(s_j - j\alpha + \tilde{p}_j)} \quad \textit{where} \quad \tilde{p}_j = \sum_{i=1}^{j} p_i \; . \tag{5.14}$$

PROOF. Define the ratio

$$R_{j,p_j}(t) := \frac{\Gamma(n_j(t) - \alpha + p_{j+1})}{\Gamma(n_i(t) - \alpha)} \; . \tag{5.15}$$

For large $n_i(t)$, $R_{i,p_i}(t) \overset{t\uparrow\infty}{\sim} n_t(t)^{p_i}$, where $a(t) \overset{t\uparrow\infty}{\sim} b(t)$ indicates $\lim_{t\to\infty} a(t)/b(t) \to 1$. We will show that $R_{i,p_i}(t)$ converges to $\xi_i^{p_i}$ when properly scaled. Because ξ_i is positive, these moments uniquely determine the distribution of ξ_i. Define

$$c_{\mathbf{p}}(t) := \prod_{s=1}^{t-1} \frac{s - \alpha k(s)}{s - \alpha k(s) + \tilde{p}(s)\mathbb{1}\{(s+1) \notin s_{1:\infty}\}} \quad \text{where} \quad \tilde{p}(s) := \sum_{i>1} p_i \mathbb{1}\{s_i \leq s\} \; . \tag{5.16}$$

Then $Z_{\mathbf{p}}(t)$ can be rewritten as

$$Z_{\mathbf{p}}(t) := c_{\mathbf{p}}(t) \prod_{i=1}^{r} R_{j_i,p_i}(t) \; . \tag{5.17}$$

Based on the asymptotic behavior of $R_{i,p_i}(t)$, the proper scaling of the block sizes will depend on the behavior of $c_{\mathbf{p}}(t)$ for large t. Note that

$$c_{\mathbf{p}}(t) = \left(\prod_{s=1}^{t-1} \frac{s - \alpha k(s)}{s - \alpha k(s) + \tilde{p}(s_m)} \right) \left(\prod_{m=1}^{k(t)-1} \frac{s_{m+1} - 1 - \alpha m + \tilde{p}(s)}{s_{m+1} - 1 - \alpha m} \right) := r_1(t) r_2(t) \; . \tag{5.18}$$

Now, by Stirling's approximation (Gradshteyn and Ryzhik, 2015, p. 8.327),

$$\begin{aligned} \ln r_1(t) &= \sum_{s=1}^{t-1} \ln(s - \alpha k(s)) - \ln(s - \alpha k(s) + \tilde{p}(s)) \\ &= \sum_{s=1}^{t-1} -\ln\left(1 + \frac{\tilde{p}(s)}{s - \alpha k(s)}\right) = C_1 - \sum_{s=1}^{t-1} \frac{\tilde{p}}{s - \alpha k(s)} + O(s^{-2}) \; , \end{aligned}$$

for some constant C_1 that captures the non-vanishing error in the approximation. By assumption, $k(s) \overset{t\uparrow\infty}{\sim} 1 + (s-1)/\mu_\delta$. Therefore, for large t,

$$\ln r_1(t) = C_1 - \frac{\tilde{p}}{1-\alpha/\mu_\delta} \ln(t-1) + O(t^{-1}) \,,$$

and $r_1(t) \overset{t\uparrow\infty}{\sim} t^{-\tilde{p}/(1-\alpha/\mu_\delta)}$. For the second term of $c_{\mathbf{p}}(t)$,

$$\ln r_2(t) = \sum_{m=1}^{k(t)-1} \ln\left(1 + \frac{\tilde{p}(s_m)}{s_{m+1}-1-\alpha m}\right) = C_2 + \sum_{m=1}^{k(t)-1} \frac{\tilde{p}}{s_{m+1}-1-\alpha m} + O(s_{m+1}^{-2}) \,.$$

By assumption, $s_j \overset{t\uparrow\infty}{\sim} 1 + \mu_\delta(j-1)$; for large t,

$$\ln r_2(t) = C_2 + \frac{\tilde{p}}{\mu_\delta - \alpha} \ln(t-1) + O(t^{-1}) \,,$$

and $r_2(t) \overset{t\uparrow\infty}{\sim} t^{\tilde{p}/(\mu_\delta-\alpha)}$. Therefore,

$$c_{\mathbf{p}}(t) = M^*_{\mathbf{p}}(s_{1:\infty}) t^{-\tilde{p}\frac{\mu_\delta-1}{\mu_\delta-\alpha}} (1 + O(t^{-1})) \,, \tag{5.19}$$

where $M^*_{\mathbf{p}}(s_{1:\infty})$ captures the approximation constants C_1 and C_2. Note that if $\mu_\delta = \infty$, $c_{\mathbf{p}}(t)$ scales as $t^{-\tilde{p}}$. With this scaling, denoting $\mathbf{p}_j$ as the restriction of $\mathbf{p}$ to its first j components, followed by all zeros, then

$$\lim_{t\to\infty} \mathbb{E}[Z_{\mathbf{p}}(t) \mid s_{1:\infty}] = \lim_{t\to\infty} M^*_{\mathbf{p}} t^{-\tilde{p}\frac{\mu_\delta-1}{\mu_\delta-\alpha}} \mathbb{E}[n_1(t)^{p_2}, \ldots, n(t)_r^{p_{r+1}} \mid s_{1:\infty}] \tag{5.20}$$

$$= \mathbb{E}[Z_{\mathbf{p}_{r-1}}(s_r) \mid s_{1:\infty}] \frac{\Gamma(1-\alpha+p_r)}{\Gamma(1-\alpha)} \frac{\Gamma(s_r - \alpha r + \tilde{p}_{r-1})}{\Gamma(s_r - \alpha r + \tilde{p}_r)} \,. \tag{5.21}$$

Iterating yields the stated moments. □

Random arrival times. By randomizing the arrival times, we can further characterize the limiting random variables ξ_j conditional on the first j arrival times via their **left-conditional** laws $\mathcal{L}(\xi_j \mid S_{1:j})$. Let Λ be the law of $S_{1:\infty}$ such that $\Lambda(S_1 = 1) = 1$. We will assume throughout that Λ admits a Markov decomposition into a sequence of probability distributions for the inter-arrival times Δ_j. That is,

$$\Lambda(S_{1:\infty}) = \prod_{j \geq 1} \Lambda^{(j-1)}(\Delta_j \mid S_{j-1}) \,, \tag{5.22}$$

for some sequence $(\Lambda^{(j)})_{j \geq 1}$ of discrete probability measures on $\mathbb{N}_+$. The decomposition amounts to requiring that the j-th arrival time S_j depends on the previous arrival times only through S_{j-1}. That is, $S_1, S_2, \ldots$ is an increasing Markov process on $\mathbb{N}_+$. Furthermore, they do not depend on the sequence $C_1, C_2, \ldots$, which means they can be sampled independently of the urn process. Furthermore, in order to avoid partitions consisting entirely of singletons, we require that $\Lambda^{(j)}(\Delta_{j+1} > 1) > 0$ for $j > 1$.

As with the exchangeable random partitions in Section 2.2.4, the restriction Π_{t-1} of a nested partition process with fixed arrival times is obtained from Π_t by deletion of the t-th element. We require the same property of Λ-random nested partition processes. Furthermore, we require their finite-dimensional distributions to be coherent. We re-state that requirement from (2.8):

$$\mathbb{P}(\Pi_t = \{A_1, \ldots, A_k\}) = \sum_{j=1}^{k(t)+1} \mathbb{P}(\mathcal{T}_j^{t+1} \Pi_t) \,. \tag{5.23}$$

The coherence requirement and the Pólya urn process determine the prediction rule

$$\mathbb{P}(\mathcal{T}_j^{t+1}\Pi_t \mid \Pi_t) = \begin{cases} (1-\lambda^{(k)}(t+1))\frac{n_j(t)-\alpha}{t-\alpha k} & \text{for } 1 \le j \le k \\ \lambda^{(k)}(t+1) & \text{for } j = k+1 \end{cases} , \tag{5.24}$$

for a sequence $(\lambda^{(k)})_{k\ge 1}$ of **sequential arrival time probabilities** induced by Λ. For $\Lambda^{(j)}$ on $\mathbb{N}_+$ with probability generating function

$$H_{\Lambda^{(j)}}(z) = \sum_{n=1}^{\infty} q_n z^n ,$$

the sequential arrival time probabilities are

$$\lambda^{(k)}(t) = \Lambda^{(k)}(\Delta_{k+1} = t - S_k \mid \Delta_{k+1} \ge t - s_k, S_k = s_k) \tag{5.25}$$

$$= \frac{q_{t-s_k}}{1 - \sum_{n=1}^{t-s_k-1} q_n} . \tag{5.26}$$

For Λ satisfying (5.22) and the coherence property (5.23), there is the disintegration

$$\mathbf{nPP}(\alpha, \Lambda) = \int_{\mathbb{N}_+^\infty} \mathbf{nPP}(\alpha \mid \delta_{1:\infty})\Lambda(d\delta_{1:\infty}) , \tag{5.27}$$

and the following theorem.

THEOREM 5.4 (Limiting count sequence for Λ-random arrival times). *Fix $\alpha \in (-\infty, 1)$ and assume that Λ satisfies* (5.22) *and that*

$$\frac{S_j}{j} \xrightarrow[j\to\infty]{\Lambda\text{-a.s.}} \mu_\delta \quad \textit{for some} \quad \mu_\delta \in \mathbb{R}_{>0} \cup \{\infty\} ,$$

and

$$\frac{1}{j}\sum_{m=1}^{j} \mathbb{1}\{\Delta_m \le n\} \xrightarrow[j\to\infty]{\Lambda\text{-a.s.}} \Lambda(\Delta \le n) \quad \textit{for all} \quad n \in \mathbb{N}_+ \,.$$

Generate $(\Pi_t)_{t\ge 1} \sim \mathbf{nPP}(\alpha, \Lambda)$. *Then for each* $r \in \mathbb{N}_+$, *conditionally on the first* r *arrival times,*

$$\lim_{t\to\infty} \tilde{M}^*_{\mathbf{p}}(\Lambda) t^{-\gamma} \mathbb{E}[n_1(t)^{p_1} \cdots n_r(t)^{p_r} \mid S_{1:r}] = \mathbb{E}[\xi_1^{p_1} \cdots \xi_r^{p_r} \mid S_{1:r}] \quad \textit{where} \quad \gamma = 1 - \frac{1-\alpha}{\mu_\delta - \alpha} \,, \tag{5.28}$$

for some constant $\tilde{M}^*_{\mathbf{p}}(\Lambda)$. *For each* $j \in \mathbb{N}_+$, *the marginal limiting left-conditional law* $\mathcal{L}(\xi_j \mid S_{1:j})$ *is uniquely characterized by its moments for* $p > -(1-\alpha)/2$,

$$\mathbb{E}[\xi_j^p \mid S_{1:j}] = \frac{\Gamma(1-\alpha+p)}{\Gamma(1-\alpha)} \prod_{m=1}^{j-1} \prod_{s=S_m}^{S_{m+1}-1} \frac{s-\alpha m}{s-\alpha m + p(1-\lambda^{(m)}(s+1))} \,. \tag{5.29}$$

Proof. As in the proof of Theorem 5.3, we rely on martingale techniques. Define (as before), for fixed $\alpha \in (-\infty, 1)$ and $p_i \in \mathbb{R}_{>-(1-\alpha)}$, the ratio

$$R_{i,p_i}(t) := \frac{\Gamma(n_i(t) - \alpha + p_i)}{\Gamma(n_i(t) - \alpha)} \,.$$

Let $\mathbf{p} := (p_i)$ be a vector of shifts. Further, let $k(t)$ be the number of blocks in the partition at step t. Define

$$\tilde{c}_{\mathbf{p}}(t) := \prod_{s=1}^{t-1} \frac{s-\alpha k(s)}{s-\alpha k(s) + \tilde{p}(1-\lambda^{(k(s)+1)}(s+1))} \quad \text{where} \quad \tilde{p} := \sum_{i=1}^{r} p_i \,. \tag{5.30}$$

Then

$$\tilde{Z}_{\mathbf{p}}(t) := \tilde{c}_{\mathbf{p}}(t) \prod_{i=1}^{r} R_{j_i,p_i}(t) \tag{5.31}$$

is a nonnegative martingale for $t \geq S_{j_r}$, with respect to $(\mathcal{A}_t)_{t\geq 1}$, the filtration generated by the partition process. As in Theorem 5.3, a straightforward adaptation of the proof of Proposition 2.1 shows that $\tilde{Z}_{\mathbf{j},\mathbf{p}}(t)$ converges in L_2 and therefore in L_1 for $p_1, \ldots, p_r > -(1-\alpha)/2$.

As in the fixed arrival time case, the asymptotics of $\tilde{c}_{\mathbf{p}}(t)$ determine the scaling in (5.28). In particular (abusing notation by absorbing all constants into C),

$$\begin{aligned}
\ln \tilde{c}_{\mathbf{p}}(t) &= -\sum_{m=1}^{k(t)-1} \sum_{s=S_m}^{S_{m+1}-1} \ln\left(1 + \frac{\tilde{p}}{s-\alpha m} - \frac{\lambda^{(m+1)}(s+1)}{s-\alpha m}\right) &(5.32)\\
&= C - \sum_{m=1}^{k(t)-1} \sum_{s=S_m}^{S_{m+1}-1} \frac{\tilde{p}}{s-\alpha m} - \frac{\lambda^{(m+1)}(s+1)}{s-\alpha m} \\
&= C - \frac{\tilde{p}}{1-\alpha/\mu_\delta} \ln(t-1) + \tilde{p} \sum_{n=1}^{\infty} \sum_{m=1}^{k(t)-1} \frac{\mathbb{1}\{\Delta_m \geq n\} \frac{q_n}{1-\sum_{i=1}^{n-1} q_i}}{m(\mu_\delta - \alpha)} \\
&= C - \frac{\tilde{p}}{1-\alpha/\mu_\delta} \ln(t-1) + \tilde{p} \sum_{n=1}^{\infty} \sum_{m=1}^{k(t)-1} \frac{(1-\sum_{i=1}^{n-1} q_i) \frac{q_n}{1-\sum_{i=1}^{n-1} q_i}}{m(\mu_\delta - \alpha)} \\
&= C - \frac{\tilde{p}}{1-\alpha/\mu_\delta} \ln(t-1) + \tilde{p} \sum_{n=1}^{\infty} \sum_{m=1}^{k(t)-1} \frac{q_n}{m(\mu_\delta - \alpha)} \\
&= C - \frac{\tilde{p}}{1-\alpha/\mu_\delta} \ln(t-1) + \frac{\tilde{p}}{\mu_\delta - \alpha} \ln(t-1)\,. &(5.33)
\end{aligned}$$

Therefore, $\tilde{c}_{\mathbf{p}}(t)$ scales as $\tilde{M}^*_{\mathbf{p}}(s_{1:\infty}) t^{-\tilde{p}\frac{\mu_\delta - 1}{\mu_\delta - \alpha}}$. (5.29) follows from $\mathbb{E}[\xi_j^p \mid S_{1:j}] = \mathbb{E}[\tilde{Z}_{j,p}(S_j) \mid S_{1:j}]$. □

The moments of the limiting joint left-conditional law can be calculated iteratively. For

a vector $\mathbf{p}$, define $\mathbf{p}_j$ to be the restriction to the first j components, and likewise for the sum $\tilde{p}_j$. Define $\tilde{p}_0 = 0$.

COROLLARY 5.5. *Let $\tilde{c}_{\mathbf{j},\mathbf{p}}(t)$ be as in (5.30). For $(\Pi_t)_{t\geq 1} \sim \mathbf{nPP}(\alpha, \Lambda)$ as in Theorem 5.4, the mixed moments of the limiting joint left-conditional law $\mathcal{L}(\xi_1, \ldots, \xi_j \mid S_{1:r})$ are*

$$\mathbb{E}[\xi_1^{p_1} \cdots \xi_r^{p_r} \mid S_{1:r}] = \prod_{i=1}^{r} \frac{\Gamma(1-\alpha+p_i)}{\Gamma(1-\alpha)} \frac{\tilde{c}_{\tilde{p}_i}(S_i)}{\tilde{c}_{\tilde{p}_{i-1}}(S_i)} \,, \tag{5.34}$$

with $\tilde{c}_{\tilde{p}}(t)$ as in (5.30).

PROOF. With $\mathbf{j} = (1, \ldots, r)$,

$$\begin{aligned}\mathbb{E}[\xi_1^{p_1} \cdots \xi_r^{p_r} \mid S_{1:r}] &= \mathbb{E}[\tilde{Z}_{\mathbf{j},\mathbf{p}}(S_r) \mid S_{1:r}] \\ &= \mathbb{E}[\tilde{Z}_{\mathbf{j}_{r-1},\mathbf{p}_{r-1}}(S_r) \mid S_{1:(r-1)}] \frac{\Gamma(1-\alpha+p_r)}{\Gamma(1-\alpha)} \frac{\tilde{c}_{\mathbf{j},\mathbf{p}}(S_r)}{\tilde{c}_{\mathbf{j}_{r-1},\mathbf{p}_{r-1}}(S_r)} \,.\end{aligned}$$

Iterating for $r-1, \ldots, 1$ yields the result. □

Specific families of distributions are studied in detail in the following sections, yielding constructive representations of the limiting left-conditional laws.

5.2 Exchangeable Gibbs partition processes

An exchangeable random partition is said to be of *Gibbs-type* if, for some (non-random) nonnegative weights $u = (u_j)_{j\geq 1}$ and $v = (v_{t,k})_{t\geq 1, k\geq 1}$, the EPPF has the form

$$p(|A_1|, \ldots, |A_k|) = v_{t,k} \prod_{j=1}^{k} u_{|A_j|} \,. \tag{5.35}$$

The Gibbs-type family of models was studied by Gnedin and Pitman (2006) and forms the basis for many Bayesian nonparametric statistical models. See De Blasi et al. (2015) for a detailed review. Gnedin and Pitman (2006) showed that the weights define an exchangeable Gibbs-type partition (EGP) if and only if[2]

$$\begin{aligned} u_j &= (1-\alpha)\dots(j-1+1-\alpha) \\ &= \frac{\Gamma(j+1-\alpha)}{\Gamma(1-\alpha)} \quad \text{for} \quad -\infty < \alpha < 1 \,, \end{aligned} \tag{5.36}$$

and the sequence v satisfies $v_{1,1} = 1$ and the backward recursion

$$v_{t,k} = (t - \alpha k) v_{t+1,k} + v_{t+1,k+1} \,. \tag{5.37}$$

A sequential construction can be deduced: Given $\Pi_t = \{A_1, A_2, \dots, A_k\}$, the probability that the next observation assigned to a new block $k+1$ is

$$\lambda^{(k)}_{\alpha,v}(t+1) := \frac{v_{t+1,k+1}}{v_{t,k}} \,. \tag{5.38}$$

The probability that it is assigned to existing block j is

$$\frac{n_j(t) - \alpha}{t - \alpha k} \left(1 - \lambda^{(k)}_{\alpha,v}(t+1)\right) \,. \tag{5.39}$$

These predictive probabilities uniquely characterize the law of an EGP, which we denote as $\mathbf{EGP}(\alpha, v)$.

[2]The case $\alpha = -\infty$ is also well-defined but it corresponds to a partition with a single block. Likewise, $\alpha = 1$ corresponds to all singleton blocks.

Conditioned on being in an existing block, the new observation is allocated with probability identical to that of the urn processes in the previous section. Furthermore, at each t the probability of an arrival time depends on the existing process only through t and the number of previous arrivals. Denote by $\Lambda_{\mathbf{EGP}(\alpha,v)}$ the law of the arrival times induced by a $\mathbf{EGP}(\alpha, v)$ process. Then

$$\mathbf{EGP}(\alpha, v) = \int_{\mathbb{N}_+^\infty} \mathbf{nPP}(\alpha \mid \delta_{1:\infty}) \Lambda_{\mathbf{EGP}(\alpha,v)}(d\delta_{1:\infty}) \; . \tag{5.40}$$

EGPs are a special case of Theorem 5.4, with $\Lambda_{\mathbf{EGP}(\alpha,v)}$. Furthermore, $\mu_\delta = \infty$, implying a scaling factor of t^{-1}, as is well known for exchangeable sequences (see Section 2.2).

The moments of the limiting left-conditional laws (5.29) for EGPs have a particularly simple form. Combining (5.37) and (5.38),

$$\mathbb{E}_{\alpha,v}[\xi_j^p \mid S_{1:j}] = \frac{\Gamma(1-\alpha+p)}{\Gamma(1-\alpha)} \prod_{m=1}^{j-1} \prod_{s=S_m}^{S_{m+1}-1} \left(1 + p\frac{v_{s+1,m}}{v_{s,m}}\right)^{-1} \quad \text{for} \quad p > -(1-\alpha)/2 \; . \tag{5.41}$$

Using the identity (R. C. Griffiths and Spanò, 2007, Lemma 4.1)

$$\mathbb{E}_{\alpha,v}[\xi_j^p \mid S_j] = \frac{\Gamma(1-\alpha+p)}{\Gamma(1-\alpha)} \frac{v_{S_j+p,j}}{v_{S_j,j}} \; , \tag{5.42}$$

we deduce that

$$\mathbb{E}_{\alpha,v}\left[\prod_{m=1}^{j-1} \prod_{s=S_m}^{S_{m+1}-1} \left(1 + p\frac{v_{s+1,m}}{v_{s,m}}\right)^{-1} \,\middle|\, S_j\right] = \frac{v_{S_j+p,j}}{v_{S_j,j}} \; , \tag{5.43}$$

where the expectation is taken with respect to $\Lambda_{\alpha,v}$, over the first $j-1$ arrival times.

Ewens–Pitman EGPs. Perhaps the most distinguished member of the class of EGPs is the so-called *Ewens–Pitman* two-parameter family, denoted as $\mathbf{EGP}(\alpha,\theta)$. Ewens (1972) introduced the one-parameter Ewens sampling formula (ESF), which corresponds to $\mathbf{EGP}(0,\theta)$, in the context of the sampling theory of alleles. See the recent survey Crane (2016) for a thorough overview of the ESF's many appearances throughout probability and statistics. Pitman (1995) studied the two-parameter model. See Pitman (1996, 2006) for more details, including earlier related work.

Ewens–Pitman EGPs have weights

$$v_{t,k}(\alpha,\theta) = \alpha^k \frac{\Gamma(k+\theta/\alpha)}{\Gamma(\theta/\alpha)} \frac{\Gamma(\theta)}{\Gamma(t+\theta)} \quad \text{with} \quad -\infty < \alpha < 1 \quad \text{and} \quad \theta \geq -\alpha \; . \tag{5.44}$$

The special form of $v_{t,k}(\alpha,\theta)$ as a ratio v_k/C_t enables constructive representations of the limiting left-conditional laws.

PROPOSITION 5.6. *Let $(\Pi_t)_{t\geq 1}$ have law $\mathbf{EGP}(\alpha,\theta)$, for some $-\infty < \alpha < 1$, $\theta > -\alpha$. Then (5.28) holds with $\gamma = 1$. Furthermore, for any $j \in \mathbb{N}_+$,*

$$\xi_j \mid S_j \stackrel{d}{=} \psi_j \quad \textit{where} \quad \psi_j \sim \mathrm{Beta}(1-\alpha, S_j+\theta-1+\alpha) \; , \tag{5.45}$$

and for any $k > j \in \mathbb{N}_+$,

$$\xi_k \mid S_{j:k} \stackrel{d}{=} \xi'_j \prod_{i=j}^{k-1} B_i \quad \textit{where} \quad B_i \stackrel{ind}{\sim} \mathrm{Beta}(S_i+\theta, \Delta_{i+1}) \quad \textit{and} \quad \xi'_j \stackrel{d}{=} \xi_j \mid S_j \; . \tag{5.46}$$

The members of the Ewens–Pitman family are the only EGPs such that ξ_j depends on the previous inter-arrival times only through their sum, as in (5.45)*, that is,*

$$\xi_j \mid S_{1:j} \stackrel{d}{=} \xi_j \mid S_j \ . \tag{5.47}$$

PROOF. Note that in the case of $\alpha \in (-\infty, 0)$, S_j will be finite only for $j \leq \theta/|\alpha|$. In that case, for $j > \theta/|\alpha|$, $\psi_j = \xi_j = 0$. For all parameter values, $\mu_\delta = \infty$, so Theorem 5.4 holds with $\gamma = 1$.

For non-zero ξ_j, $\frac{v_{s+1,k}}{v_{s,k}} = (s+\theta)^{-1}$, which induces a left-conditional limiting law that depends only on S_j,

$$\begin{aligned} \mathbb{E}_{\alpha,\theta}[\xi_j^p \mid S_{1:j}] &= \frac{\Gamma(1-\alpha+p)}{\Gamma(1-\alpha)} \frac{\Gamma(S_j+\theta)}{\Gamma(S_j+\theta+p)} \\ &= \mathbb{E}_{\alpha,\theta}[\xi_j^p \mid S_j] \ . \end{aligned} \tag{5.48}$$

The moments are recognizable as those of a Beta$(1-\alpha, S_j+\theta-1+\alpha)$ random variable. (5.46) also can be verified by checking the moments. Each B_i serves to "shift" the second term of the product in (5.48):

$$\mathbb{E}_{\alpha,\theta}[\xi_j^p \mid S_j]\mathbb{E}_{\alpha,\theta}[B_j^p \mid S_{j:j+1}] = \frac{\Gamma(1-\alpha+p)\Gamma(S_j+\theta)}{\Gamma(1-\alpha)\Gamma(S_j+\theta+p)} \frac{\Gamma(S_j+\theta+p)\Gamma(S_{j+1}+\theta)}{\Gamma(S_j+\theta)\Gamma(S_{j+1}+\theta+p)} \tag{5.49}$$

$$= \mathbb{E}_{\alpha,\theta}[\xi_{j+1}^p \mid S_{j+1}] \ . \tag{5.50}$$

Iterating gives the result for general $k > j$.

Kerov (2006) showed that the only EGPs with v-weights representable as ratios v_k/C_t

are the members of the Ewens–Pitman two-parameter family (see also Gnedin and Pitman, 2006), which implies the final claim, due to the form of (5.41). □

The simplicity of the marginal left-conditional laws does not carry over entirely to the joint left-conditional laws. The additional dependence induced by conditioning ξ_1 on $S_{2:r}$, and ξ_2 on $S_{3:r}$, and so on, complicates the representations.

Let X be a positive random variable such that its Mellin transform is

$$\mathbb{E}[X^p] = \frac{\Gamma(a_1+\beta_1 p)\cdots\Gamma(a_K+\beta_K p)}{\Gamma(a_1)\cdots\Gamma(a_K)} \frac{\Gamma(b_1)\cdots\Gamma(b_N)}{\Gamma(b_1+\beta_1 p)\cdots\Gamma(b_N+\beta_N p)} \,. \tag{5.51}$$

Then X is said to have moments of *Gamma type* (Janson, 2010). When all $\beta_k, \beta_n = \pm 1$, X is said to have **G distribution**, denoted $\mathbf{G}(a_1,\ldots,a_K;b_1,\ldots,b_N)$ (Dufresne, 2010). **G** random variables appear in the joint limiting left-conditional laws.

PROPOSITION 5.7. *Let $(\Pi_t)_{t\geq 1}$ have law $\mathbf{EGP}(\alpha,\theta)$, for some $0<\alpha<1$, $\theta>-\alpha$. Then the joint limiting left-conditional law, conditional on $S_{1:r}$, has the following constructive representation: For $j=1,2,\ldots,$ let ψ_j be independent* $\mathrm{Beta}(1-\alpha, S_j+\theta-1+\alpha)$ *random variables, and let G_j be independent $\mathbf{G}(S_j+\theta; S_j+\theta-1+\alpha)$ random variables, with Mellin transform as in* (5.51)*. Set*

$$\tilde{\xi}_j = \psi_j \prod_{i=j+1}^{r} G_i(1-\psi_i) \quad \textit{for each} \quad 1\leq j\leq r \,. \tag{5.52}$$

Then for each $r\in\mathbb{N}_+$,

$$(\xi_1,\ldots,\xi_r) \mid S_{1:r} \stackrel{d}{=} (\tilde{\xi}_1,\ldots,\tilde{\xi}_r) \,.$$

PROOF. Using (5.34), for any $p_1, \ldots, p_r > -(1-\alpha)/2$

$$\mathbb{E}_{\alpha,\theta}[\xi_1^{p_1} \cdots \xi_r^{p_r} \mid S_{1:r}] = \prod_{j=1}^{r} \frac{\Gamma(1-\alpha+p_j)}{\Gamma(1-\alpha)} \frac{\Gamma(S_j+\theta+\tilde{p}_{j-1})}{\Gamma(S_j+\theta+\tilde{p}_j)} \,. \tag{5.53}$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}[\tilde{\xi}_1^{p_1} \cdots \tilde{\xi}_r^{p_r}] = \mathbb{E}\Big[\psi_1^{p_1} \prod_{j=2}^{r} \psi_j^{p_j}(1-\psi_j)^{\tilde{p}_{j-1}} G_j^{\tilde{p}_{j-1}}\Big] = \mathbb{E}[\psi_1^{p_1}] \prod_{j=2}^{r} \mathbb{E}[\psi_j^{p_j}(1-\psi_j)^{\tilde{p}_{j-1}}]\mathbb{E}[G_j^{\tilde{p}_{j-1}}] \\
&= \frac{\Gamma(1-\alpha+p_1)}{\Gamma(1-\alpha)} \frac{\Gamma(S_1+\theta)}{\Gamma(S_1+\theta+p_1)} \times \cdots \\
&\quad \prod_{j=2}^{r} \frac{\Gamma(1-\alpha+p_j)\Gamma(S_j+\theta-1+\alpha+\tilde{p}_{j-1})\Gamma(S_j+\theta)}{\Gamma(1-\alpha)\Gamma(S_j+\theta-1+\alpha)\Gamma(S_j+\theta+\tilde{p}_j)} \frac{\Gamma(S_j+\theta+\tilde{p}_{j-1})\Gamma(S_j+\theta-1+\alpha)}{\Gamma(S_j+\theta)\Gamma(S_j+\theta-1+\alpha+\tilde{p}_{j-1})} \\
&= \prod_{j=1}^{r} \frac{\Gamma(1-\alpha+p_j)}{\Gamma(1-\alpha)} \frac{\Gamma(S_j+\theta+\tilde{p}_{j-1})}{\Gamma(S_j+\theta+\tilde{p}_j)} \,,
\end{aligned}$$

which establishes the claim. □

5.3 Yule–Simon partition processes

The models in this section get their name from their similarity to the pure birth point process proposed in Yule (1925) to model the evolution of the number of species within a genus, and on the distribution of sizes of genera. Simon (1955) proposed the sequential model described below as a simple model for text, with the aim of constructing a distribution of word frequencies that exhibited power law behavior. Simon, recognizing the connection to Yule's work, proposed naming the resulting distribution after Yule, but Simon's name has historically been attached, as well. In fact, variations on this model have appeared in various branches of science; a fascinating review of its different incarnations is Simkin and

Roychowdhury (2011).

The Yule–Simon model differs from EGPs in that at each t, the probability of a new block is constant:

$$\lambda_{\beta}^{(k)}(t+1) = \beta \quad \text{for all} \quad k, t \in \mathbb{N}_+ \, . \tag{5.54}$$

The original Yule–Simon model allocated observations to an existing block with probability proportional to the block size, and is therefore an example of a nested Pólya urn with $\alpha = 0$.[3] We use the general Pólya urn rule and denote the law of Yule–Simon partitions as $\mathbf{YS}(\beta, \alpha)$.

The model gives rise to an eponymous distribution, which characterizes the asymptotic *block size distribution*, i.e. the probability p_d that a block sampled uniformly at random is of size d. Simon (1955) showed that the distribution is

$$p_d = \rho \frac{\Gamma(d)\Gamma(1+\rho)}{\Gamma(d+1+\rho)} \quad \text{for} \quad \rho > 0. \tag{5.55}$$

In the context of $\mathbf{YS}(\beta)$ partitions, $\rho = 1/(1-\beta)$. We defer further discussion of the block size distribution until Section 5.4, when it will be used to study the degree distribution of random graphs formed from random partitions.

The Yule–Simon model is perhaps the simplest nested Pólya partition model with i.i.d. inter-arrival times. We refer to such models as τ-i.i.d. , where τ is the distribution of the

[3]Simon (1955) actually makes a weaker assumption, that the probability an observation is allocated to an existing block of size k is proportional to the total number of observations in all blocks of size k. Although it is satisfied by the Pólya urn mechanism with $\alpha = 0$, the assumption leaves open the possibility of non-uniform allocation among blocks of size k.

arrival times. The sequential construction, in which a Bernoulli(β) random variable at each step determines whether or not to create a new block, is converted to a nested Pólya urn process by setting $\Delta_1 = 1$ and sampling

$$\Delta_j \overset{\text{iid}}{\sim} \text{Geometric}(\beta) \quad \text{for} \quad j \geq 2 \, .$$

The memoryless property, unique to the Geometric distribution of all discrete probability distributions, yields simple representations of the left-conditional laws, particularly when $\alpha = 0$, similar to those in the Ewens–Pitman $\mathbf{EGP}(\alpha, \theta)$ case.

PROPOSITION 5.8. *Let* $(\Pi_t)_{t \geq 1}$ *have law* $\mathbf{YS}(\beta, \alpha)$. *Then* (5.28) *holds with* $\gamma = \frac{1-\beta}{1-\beta\alpha}$. *Let* $\rho = 1/(1-\beta)$, *and define*

$$\begin{aligned} M_j &\sim \text{Mittag-Leffler}(\rho^{-1}, S_j - 1 - \alpha(j-1)) \\ \psi_j &\sim \text{Beta}(1-\alpha, \rho(S_j - 1 - \alpha(j-1)) + \alpha) \\ V_j &\sim \text{Beta}(S_j - \alpha j, \alpha) \\ B_j &\sim \text{Beta}(S_j - \alpha(j-1), \Delta_{j+1} - \alpha) \end{aligned}$$

For any $j \in \mathbb{N}_+$,

$$\xi_j \mid S_{1:j} \overset{d}{=} \psi_j M_j \prod_{m=1}^{j-1} V_m^{1/\rho} \, , \tag{5.56}$$

and for any $k > j \in \mathbb{N}_+$,

$$\xi_k \mid S_{1:k} \stackrel{d}{=} \xi'_j \prod_{i=j}^{k-1} V_i^{1/\rho} B_i^{1/\rho} \quad \textit{where} \quad \xi'_j \stackrel{d}{=} \xi_j \mid S_{1:j} \ . \tag{5.57}$$

The members of the $\mathbf{YS}(\beta, 0) = \mathbf{YS}(\beta)$ *family are the only* $\mathbf{nPP}(\alpha, \Lambda_\tau^{\text{i.i.d.}})$ *partitions such that* ξ_j *depends on the previous inter-arrival times only through their sum. That is,*

$$\xi_j \mid S_{1:j} \stackrel{d}{=} \xi_j \mid S_j \ .$$

Remark. In the case of $\alpha = 0$, the distributional identities simplify considerably. In particular, (5.56) becomes

$$\xi_j \mid S_{1:j} \stackrel{d}{=} \psi_j M_j \ .$$

◁

Proof. The almost sure limit of $(S_j - 1)/(j-1)$ is $\mu_\delta = 1/\beta$, which yields $\gamma = \frac{1-\beta}{1-\beta\alpha}$. (5.29) implies

$$\begin{aligned}
\mathbb{E}_{\beta,\alpha}[\xi_j^p \mid \Delta_{1:j}] &= \frac{\Gamma(1-\alpha+p)\Gamma(S_j - \alpha(j-1))}{\Gamma(1-\alpha)\Gamma(S_j - \alpha(j-1) + p/\rho)} \prod_{m=1}^{j-1} \frac{\Gamma(S_m - \alpha m + p/\rho)\Gamma(S_m - \alpha(m-1))}{\Gamma(S_m - \alpha m)\Gamma(S_m - \alpha(m-1) + p/\rho)} \\
&= \mathbb{E}[\psi_j^p M_j^p] \prod_{m=1}^{j-1} \mathbb{E}[V_m^{p/\rho}] \ ,
\end{aligned}$$

which verifies (5.56). (5.57) is obtained in a similar way as its $\mathbf{EGP}(\alpha, \theta)$ counterpart, (5.46): Each application of $B_j^{1/\rho}$ "shifts" $\psi_j M_j$ to $\psi_{j+1} M_{j+1}$. Multiplication by $V_j^{1/\rho}$ results in the correct moments.

When $\alpha = 0$, $V_m = 1$ for all m, yielding simplified moments for ξ_j that depend only on S_j. This property is unique to the $\mathbf{YS}(\beta)$ family among all $\mathbf{nPP}(\alpha, \Lambda_\tau^{\text{i.i.d.}})$ partitions because $\lambda_\beta^{(k)}(t+1) = \beta$ is memoryless. Among all discrete probability distributions, only the Geometric distribution is memoryless. □

5.4 Random graphs from random partitions

There are multiple ways to create a graph G from a partition Π. The simplest is to treat each observation as a vertex, and to add an edge between vertex v and vertex v' if they occupy the same block of Π. More precisely, let $b(t)$ be the block of Π that contains t. Then

$$v \bullet\!\!-\!\!\bullet\, v' \iff b(v) = b(v') \ .$$

However, this rule creates graphs with $k(t)$ disjoint components, each of which is the complete graph on $|A_j|$ vertices. See Aldous (1997) for a different construction that connects the component sizes in a Erdős-Rényi random graph to the multiplicative coalescent.

In this section we use a different procedure. Given a partition Π_{2t} of $[2t]$ with k blocks, form a multigraph with t edges, $G_t = \varphi_t(\Pi_{2t})$ as follows:

- For each block $j = 1, \ldots, k$, create vertex v_j.
- For each odd $s < 2t$, create the edge $(b(s), b(s+1))$ with label $\lceil s/2 \rceil$.

The maps $\varphi_t : \widetilde{\mathcal{P}}_{2t} \to \mathcal{G}_t$ are bijective because the edges are labeled in order of appearance. If the edge labels are discarded, G_t corresponds to more than one partition, and an equivalent way to generate G_t is to sample pairs of elements of Π_{2t} uniformly without replacement,

and connect each pair with an edge. In either case, the induced mapping between blocks and vertices is bijective.

A number of existing models for random graphs can be constructed this way, each with a different distribution for Π_{2t}. These include:

Preferential attachment graphs (Barabási and Albert, 1999; Berger, Borgs, Chayes, and Saberi, 2014; Peköz, Ross, and Röllin, 2014). Given parameters α and $\ell \in \mathbb{N}_+$ and a seed graph G_{n_0} on n_0 vertices with e_0 edges, construct a random multigraph with the following sequential process: For $n > n_0$, add a new vertex v_{n+1}. For $m = 1, \ldots, \ell$, add a new edge with one end connected to v_{n+1}, and the other end connected to vertex $v' \in \mathbf{V}(G_n) \cup \{v_{n+1}\}$ with probability

$$\frac{d_{v'}(n+m-1) - \alpha}{e_0 + \ell(n-n_0) - \alpha(n+1) + m - 1} \,. \tag{5.58}$$

The PA model and its variations have been subject to extensive study. The local weak limit was described in Berger, Borgs, Chayes, and Saberi (2014) using a nested paintbox similar to Algorithm 5.3. The joint degree sequence for general seed graphs was characterized in Peköz, Ross, and Röllin (2014), and studied from a different perspective by James (2015); similar distributional results can be obtained from Theorems 5.3 and 5.4. The $\mathbf{ACL}(\beta)$ model discussed in Chapter 3 is a generalization of the PA model that is closely related to $\mathbf{YS}(\beta)$ partitions.

Factorizable edge-exchangeable models (Crane and Dempsey, 2015a, 2016; Cai,

Campbell, and Broderick, 2016). As introduced in Section 2.3.1, let $E_1, E_2, \cdots = (V, V')_1, (V, V')_2, \ldots$ be an exchangeable sequence of edges, with entries taking values in the measurable product space $\mathcal{V}_* \times \mathcal{V}_*$ (with σ-algebra $\mathcal{A}_{\mathcal{V}_*}$). A graph formed from the edges and the induced vertex set is called an *edge-exchangeable* graph, as studied in Crane and Dempsey (2016) and Cai, Campbell, and Broderick (2016). de Finetti's theorem, adapted to edge-exchangeable graphs in Crane and Dempsey (2016), says that every edge-exchangeable graph can be represented as a mixture of i.i.d. draws from a probability distribution $\mu : \mathcal{V}_* \times \mathcal{V}_* \to [0, 1]$. Consider a mixing distribution that places mass only on *factorizable* μ, i.e. there is some $\tilde{mu}$ such that

$$\mu(A \times B) = \tilde{\mu}(A)\tilde{\mu}(B), \quad \text{for all} \quad \mathrm{A}, B \in \mathcal{A}_{\mathcal{V}_*} .$$

Such a model generates an exchangeable sequence of ends of edges, or edge *stubs* (Newman, 2009; Riordan, 2012); it might be said to be "stub-exchangeable". The specific model studied in Crane and Dempsey (2016), called the Hollywood model because it was applied to a dataset from the Internet Movie Database (IMDb), is precisely the Ewens–Pitman $\mathbf{EGP}(\alpha, \theta)$ process adapted as a stub-exchangeable model for hypergraphs.[4] The Hollywood model inherits a well-known property of $\mathbf{EGP}(\alpha, \theta)$ processes: The law of the limiting block size proportions sorted into non-increasing order, denoted as $(\xi_j^{\downarrow})_{j \geq 1}$, is the so-called **Poisson-Dirichlet** distribution (Pitman and Yor, 1997; Pitman, 2006) $\mathbf{PD}(\alpha, \theta)$. Therefore, $\tilde{\mu}$ in the Hollywood model is $\mathbf{PD}(\alpha, \theta)$.

[4] Crane and Dempsey (2016) generalized to hypergraphs, where an edge is defined to be a set of k vertices, with a possibly random k for each edge. For fixed $k = 2$, multigraphs from the Hollywood model are the multigraph sequence formed from an $\mathbf{EGP}(\alpha, \theta)$ process. In the more general case, the same basic intuition applies.

5.5 Sparsity and degree distributions

The previous sections examined the limiting behavior of the ordered block count sequences; the bijectivity of the mapping between blocks and vertices allows those properties to be transferred without modification to the degree sequences of the resulting random graphs. Of interest in many settings is the degree distribution. Intuitively, the degree distribution is the probability that a uniformly sampled vertex will have degree d. More precisely, let $N_e(t)$ and $N_v(t)$ be the number of edges and vertices, respectively, in G_t. Let $m_d(t)$ be the number of vertices in G_t with degree d. The *degree distribution* is the empirical histogram vector $N_v(t)^{-1}(m_d(t))_{d\geq 1} := ((p_d(t))_{d\geq 1})_{t\geq 1}$, as in Section 3.2.2. Two of the most widely cited empirical properties of real networks, repeated from Chapter 2, are:

- **Sparsity.** Roughly, the average degree grows more slowly than the number of edges needed for a complete graph on the same number of vertices. Let $N_e(t)$ and $N_v(t)$ denote the number of edges and vertices, respectively, in G_t. For $1 \leq \varepsilon < 2$, we call a graph sequence $(G_t)_{t>0}$ ε-sparse if

$$\limsup_{t\to\infty} \frac{N_e(t)}{N_v(t)^{\varepsilon}} = c_\varepsilon > 0 \, . \tag{5.59}$$

  If $\varepsilon \geq 2$, the network is called *dense*. Note that $\varepsilon > 2$ is only possible for multigraphs.

- **Power law degree distribution.** A graph sequence $(G_t)_{t>0}$ exhibits a power law

degree distribution with exponent $\eta > 1$ if

$$p_d(t) := \frac{m_d(t)}{N_v(t)} \overset{t\uparrow\infty}{\sim} L(d)d^{-\eta} \quad \text{for all large } d \text{ as} \quad t \to \infty\,, \tag{5.60}$$

for some slowly varying function $L(d)$, that is, $\lim_{x\to\infty} L(rx)/L(x) = 1$ for all $r > 0$ (e.g. Feller, 1971; Bingham, Goldie, and Teugels, 1989), and where $a(t) \overset{t\uparrow\infty}{\sim} b(t)$ indicates $\lim_{t\to\infty} a(t)/b(t) \to 1$.

For models based on a nested Pólya urn process, the two properties are intimately related. It is easy to see that sparsity is entirely controlled by the distribution of the arrival times. Perhaps harder to see is that the tail behavior of the degree distribution is also controlled by Λ. Intuitively, if the number of vertices grows quickly, there will be many vertices of small degree, placing high probability mass on small values of d. If vertex growth is too high, the preferential attachment mechanism will have minimal effect and the tails will fall off quickly; the degenerate cases of $\mathbf{YS}(1, \alpha)$ or $\mathbf{EGP}(1, \theta)$, for example, result in all vertices of degree $d = 1$. As the new vertex rate of growth decreases, the effect of the preferential attachment mechanism grows stronger and more vertices of higher degree will occur, spreading probability mass away from small d. If the rate of growth is too low, the degree distribution concentrates entirely on large d.

The tail behavior of the limiting degree distributions for **EGP** models and **YS** models makes this clear. In the **EGP** case, the limiting degree distribution is (Pitman, 2006, Lemma 3.11)

$$p_\alpha^{\mathbf{EGP}}(d) = \alpha \frac{\Gamma(d-\alpha)}{\Gamma(d+1)\Gamma(1-\alpha)} \tag{5.61}$$

$$= \frac{\alpha}{\Gamma(1-\alpha)} d^{-(1+\alpha)}(1 + O(d^{-1})) \,. \tag{5.62}$$

In the **YS** case, the limiting degree distribution is a generalization of the well-known Yule–Simon distribution introduced in Chapter 3 (see Appendix D.1)

$$\begin{aligned} p^{\mathbf{YS}}_{\beta,\alpha}(d) &= \rho(\beta,\alpha) \frac{\Gamma(d-\alpha)\Gamma(1-\alpha+\rho(\beta,\alpha))}{\Gamma(1-\alpha)\Gamma(d+1-\alpha+\rho(\beta,\alpha))} \quad \text{where} \quad \rho(\beta,\alpha) := \frac{1-\beta\alpha}{1-\beta} \\ &= \rho(\beta,\alpha) \frac{\Gamma(1-\alpha+\rho(\beta,\alpha))}{\Gamma(1-\alpha)} d^{-(1+\rho(\beta,\alpha))}(1 + O(d^{-1})) \,. \end{aligned} \tag{5.63}$$

In the **EGP** process, exchangeability requires that the rate of vertex arrivals decrease asymptotically to zero; as a result, the degree exponent is $\eta \in (1, 2)$. On the other hand, the rate of vertex arrivals is constant (in expectation) in the non-exchangeable **YS** process; the degree exponent is $\eta \in (2, \infty)$.

We make this intuition precise for random graph models built from exchangeable random partitions in the following:

PROPOSITION 5.9. *Let $(G_t)_{t\geq 1} = (\varphi_t(\Pi_{2t}))_{t\geq 1}$, where the law of $(\Pi_{2t})_{t\geq 1}$ is exchangeable. Then the law of $(G_t)_{t\geq 1}$ is invariant under the symmetric group acting on stubs, and the graph sequence is either dense ($\varepsilon \geq 2$) or has sparsity $\varepsilon > 1$. If, as $t \to \infty$, the degree distribution obeys a power law, its exponent is in the interval $(1, 2)$. Moreover, if the graph is both ε-sparse and has power law degree distribution, the power law exponent is $\eta = 1 + \varepsilon^{-1} \in (3/2, 2)$.*

PROOF. For $\alpha < 0$, the $N_v(t)$ is finite for all t greater than some finite T (Pitman, 2006), resulting in dense graph sequences that do not exhibit power law degree distributions.

For $\alpha = 0$, $N_v(t) \approx \log(2t)$, also resulting in dense sequences without power law degree distributions.

For $0 < \alpha < 1$, all claims follow from Pitman (2006), Lemma 3.11: $N_v(t)/(2t)^\alpha \xrightarrow{\text{a.s.}} Z_\alpha$ as $t \to \infty$, where Z_α is a strictly positive random variable, yielding $\epsilon = 1/\alpha$. The definition of sparsity requires $\alpha > 1/2$. Furthermore, the asymptotic degree distribution is given in (5.61) so $\eta = 1 + \alpha \in (1, 2)$. When the graph sequence is sparse, $3/2 < \eta < 2$. □

Remark. Crane and Dempsey (2016) define sparsity for hypergraphs; adapting ε-sparsity to their definition, a hypergraph sequence is sparse if $\varepsilon < \mu_v$, where μ_v is the average number of vertices participating in each edge. Analogous claims can be shown to hold; in that case, a sequence exhibiting both sparsity (i.e. $\alpha > 1/\mu_v$) and power law degree distribution has $\eta \in (1 + 1/\mu_v, 2)$. ◁

For random graphs constructed from $\mathbf{YS}(\beta, \alpha)$ partition processes, there is a counterpart:

PROPOSITION 5.10. *Let $(G_t)_{t\geq 1} = (\varphi_t(\Pi_{2t}))_{t\geq 1}$, where $(\Pi_{2t})_{t\geq 1}$ has law $\mathbf{YS}(\beta, \alpha)$. Then the graph sequence is almost surely ε-sparse with $\varepsilon = 1$, and the degree distribution obeys a power law with exponent $\eta = 1 + \frac{1-\beta\alpha}{1-\beta} \in (2, \infty)$, as $t \to \infty$.*

PROOF. The sparsity claim is a straightforward result of the law of large numbers for a sum of i.i.d. Bernoulli(β) random variables: The number of vertices scales as $1 + \beta(2t - 1)/2$. The asymptotic degree distribution is as in (5.63), which proves the claim. □

Remark. We note that $1/\rho(\beta, \alpha)$ is equal to γ, the scaling of the block count sequence for $\mathbf{YS}(\beta, \alpha)$ models in Proposition 5.8. The relationship suggests the following conjecture. ◁

CONJECTURE 5.11. *Let* $(G_t)_{t\geq 1} = (\varphi_t(\Pi_{2t}))_{t\geq 1}$, *where* $(\Pi_{2t})_{t\geq 1}$ *has law* $\mathbf{nPP}(\alpha, \Lambda_\tau^{\text{i.i.d.}})$, *and* $0 < \mu_\tau < \infty$ *is the expectation of a* τ*-distributed random variable. Then the degree distribution obeys a power law with exponent*

$$\eta = 1 + \frac{\mu_\tau - \alpha}{\mu_\tau - 1} \in (2, \infty) \, .$$

Finally, we characterize the degree distribution of all $\mathbf{EGP}(\alpha, v)$ and $\mathbf{YS}(\beta, \alpha)$ random graph models as compound beta-geometric distributions:

PROPOSITION 5.12. *Let* $(G_t)_{t\geq 1} = (\varphi_t(\Pi_{2t}))_{t\geq 1}$, *where* $(\Pi_{2t})_{t\geq 1}$ *either has law* $\mathbf{EGP}(\alpha, v)$ *or* $\mathbf{YS}(\beta, \alpha)$, *for* $0 < \alpha < 1$. *Then the asymptotic degree distribution is a compound beta-geometric distribution with the following representation:*

$$p(d) = \mathbb{E}[X(1-X)^{d-1}] \quad \textit{where} \quad X \sim \begin{cases} \text{Beta}(\alpha, 1-\alpha) & \textit{if } (\Pi_{2t})_{t\geq 1} \sim \mathbf{EGP}(\alpha, v) \\ \text{Beta}(\rho(\beta,\alpha), 1-\alpha) & \textit{if } (\Pi_{2t})_{t\geq 1} \sim \mathbf{YS}(\beta, \alpha) \end{cases} . \tag{5.64}$$

PROOF. If $X \sim \text{Beta}(a, b)$, then for any $p > -a$, $q > -b$,

$$\mathbb{E}[X^p(1-X)^q] = \frac{\Gamma(a+p)\Gamma(b+q)\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+p+q)} \, .$$

Then

$$\begin{aligned} p_\alpha^{\mathbf{EGP}}(d) &= \frac{\Gamma(1+\alpha)\Gamma(d-\alpha)}{\Gamma(\alpha)\Gamma(1-\alpha)\Gamma(d+1)} \\ &= \alpha \frac{\Gamma(d-\alpha)}{\Gamma(1-\alpha)\Gamma(d+1)} \, , \end{aligned}$$

and

$$p^{\mathbf{YS}}_{\beta,\alpha}(d) = \frac{\Gamma(1+\rho(\beta,\alpha))\Gamma(d-\alpha)\Gamma(1-\alpha+\rho(\beta,\alpha))}{\Gamma(\rho(\beta,\alpha))\Gamma(1-\alpha)\Gamma(d+1-\alpha+\rho(\beta,\alpha))}$$
$$= \rho(\beta,\alpha)\frac{\Gamma(d-\alpha)\Gamma(1-\alpha+\rho(\beta,\alpha))}{\Gamma(1-\alpha)\Gamma(d+1-\alpha+\rho(\beta,\alpha))} ,$$

which proves the claim. □

Remark. The construction of the Yule–Simon distribution (5.63) with $\alpha = 0$ is typically given as a compound exponential-geometric distribution,

$$p(d) = \mathbb{E}[e^{-X}(1-e^{-X})^{d-1}] \quad \text{where} \quad X \sim \text{Exponential}(\rho(\beta,0)) .$$

The construction follows directly from properties of Yule's birth process (Yule, 1925). For $B \sim \text{Beta}(a,1)$ and $X \sim \text{Exponential}(a)$, there is the equality in distribution $B \stackrel{d}{=} e^{-X}$. Therefore, it is straightforward to show that this construction is equivalent to the $\mathbf{YS}(\beta,0)$ case in (5.64). ◁

5.6 Discussion

The results in this chapter shed some light on the behavior of the tails of the degree distribution for preferential attachment-type models. At a high level, a power law requires that new blocks be added to a partition at a sufficiently high rate. As Proposition 5.10 shows, one way to produce power law behavior is to have i.i.d. inter-arrival times with Geometric(β) distribution. Even when $\alpha = 0$, a power law is obtained. The exchangeable case, however, requires $\alpha > 0$ to generate a power law distribution. Propositions 5.9 and 5.10 indicate

that PA-type models for random graphs formed from edge sequences have two complementary regimes: exchangeable and non-exchangeable. A similar (non-rigorous) argument was made by Crane and Dempsey (2015b), based on the rate of growth of the number of vertices; the growth rates corresponding to the two regimes are precisely the ε-sparsity levels of Propositions 5.9 and 5.10

The technique of Poissonization has been used to analyze asymptotic properties of exchangeable urn models (Gnedin, Hansen, and Pitman, 2007). Such an approach seems natural in the non-exchangeable case. In particular, Yule's original model was mapped to a time-changed Poisson point process by Kendall (1966), and the connections here between EGPs and Yule–Simon processes seem promising.

Finally, there is a literature on the limits of permutations and partitions with various restrictions placed on them (e.g. Gnedin, 2006; Chen and Winkel, 2013; Gnedin and Gorin, 2015). To our knowledge, none of the existing work applies directly to the models studied here; deeper connections may exist.

Higher-order models. Given that the number of vertices and the degree sequence are predictive sufficient statistics at each t, it is perhaps not surprising that the class of models studied in this chapter only allows modeling control of sparsity and degree properties. It is in this sense that we call the edge density *zeroeth-order* and the degree sequence *first-order* statistics. While it is possible to calculate the properties of higher-order statistics for models based on zeroeth- and first-order statistics, such models do not offer sufficient flexibility as statistical models to capture structure of higher orders.

It is natural to ask what sorts of higher-order statistics could be used as the basis for

more flexible statistical models. While a number of possibilities may exist, one is provided by realizing that the normalized degree sequence of a graph is the stationary distribution of a simple random walk on the graph (and therefore also the dominant left eigenvector of the transition matrix $\mathbf{W}$). In a sequential model, inserting an edge by sampling twice from the degree-biased distribution is equivalent in distribution to performing two infinite-length random walks. Considering finite-length walks provides motivation for the random walk models in Chapter 3.

Bibliography

Aiello, William, Fan Chung, and Linyuan Lu (2001). "A random graph model for power law graphs". In: *Experiment. Math.* 10.1, pp. 53–66.

— (2002). "Random Evolution in Massive Graphs". In: *Handbook of Massive Data Sets*. Ed. by James Abello, Panos M. Pardalos, and Mauricio G. C. Resende. Vol. II, pp. 97–122.

Airoldi, Edoardo M., David M. Blei, Stephen E. Fienberg, and Eric P. Xing (2008). "Mixed Membership Stochastic Blockmodels". In: *J. Mach. Learn. Res.* 9, pp. 1981–2014.

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

— (1983). "Random walks on finite groups and rapidly mixing Markov chains". In: *Séminaire de Probabilités XVII 1981/82: Proceedings*. Ed. by Jacques Azéma and Marc Yor. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 243–297.

— (1997). "Brownian excursions, critical random graphs and the multiplicative coalescent". In: *Ann. Probab.* 25.2, pp. 812–854.

Andrieu, Christophe, Arnaud Doucet, and Roman Holenstein (2010). "Particle Markov chain Monte Carlo methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.3, pp. 269–342.

Andrieu, Christophe and Gareth O. Roberts (2009). "The pseudo-marginal approach for efficient Monte Carlo computations". In: *Ann. Statist.* 37.2, pp. 697–725.

Arthur, W. Brian, Yu. M. Ermoliev, and Yu. M. Kaniovski (1987). "Path-dependent processes and the emergence of macro-structure". In: *European Journal of Operational Research* 30.3, pp. 294–303.

Athreya, Krishna B., Arka P. Ghosh, and Sunder Sethuraman (2008). "Growth of preferential attachment random graphs via continuous-time branching processes". English. In: *Proceedings Mathematical Sciences* 118.3, pp. 473–494.

Bacallado, Sergio, Stefano Favaro, and Lorenzo Trippa (2013). "Bayesian nonparametric analysis of reversible Markov chains". In: *Ann. Statist.* 41.2, pp. 870–896.

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

Batagelj, Vladimir and Adrej Mrvar (2006). "Pajek datasets". Available at: `http://vlado.fmf.uni-lj.si/pub/networks/data/`.

Beaumont, Mark A. (2003). "Estimation of Population Growth or Decline in Genetically Monitored Populations". In: *Genetics* 164.3, pp. 1139–1160.

Belkin, Mikhail, Irina Matveeva, and Partha Niyogi (2004). "Regularization and Semi-supervised Learning on Large Graphs". In: *Learning Theory: 17th Annual Conference on Learning Theory, COLT 2004, Banff, Canada, July 1-4, 2004. Proceedings*. Ed. by John Shawe-Taylor and Yoram Singer. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 624–638.

Bender, Edward A. and E. Rodney Canfield (1978). "The asymptotic number of labeled graphs with given degree sequences". In: *Journal of Combinatorial Theory, Series A* 24.3, pp. 296–307.

Benjamini, Itai and Oded Schramm (2001). "Recurrence of Distributional Limits of Finite Planar Graphs". In: *Electron. J. Probab.* 6, pp. 1–13.

Berger, Noam, Christian Borgs, Jennifer T. Chayes, and Amin Saberi (2005). "On the Spread of Viruses on the Internet". In: *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '05. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, pp. 301–310.

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

Bertoin, Jean (2006). *Random Fragmentation and Coagulation Processes*. Cambridge Studies in Advanced Mathematics 102. Cambridge University Press.

Bianconi, G. and A.-L. Barabási (2001). "Competition and multiscaling in evolving networks". In: *Europhysics Letters* 54.4, p. 436.

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

Blackwell, David and David G. Kendall (1964). "The Martin Boundary for Pólya's Urn Scheme, and an Application to Stochastic Population Growth". In: *Journal of Applied Probability* 1.2, pp. 284–296.

Blackwell, David and James B. MacQueen (1973). "Ferguson Distributions Via Pólya Urn Schemes". In: *Ann. Statist.* 1.2, pp. 353–355.

Blei, David M. (2012). "Probabilistic Topic Models". In: *Commun. ACM* 55.4, pp. 77–84.

Bollobás, Béla (1980). "A Probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs". In: *European Journal of Combinatorics* 1.4, pp. 311–316.

Bollobás, Béla, Svante Janson, and Oliver Riordan (2007). "The phase transition in inhomogeneous random graphs". In: *Random Structures & Algorithms* 31.1, pp. 3–122.

Bollobás, Béla, Oliver Riordan, Joel Spencer, and Gábor Tusnády (2001). "The degree sequence of a scale-free random graph process". In: *Random Structures & Algorithms* 18.3, pp. 279–290.

Borgs, Christian, Jennifer T. Chayes, Henry Cohn, and Nina Holden (2016). "Sparse exchangeable graphs and their limits via graphon processes". In: arXiv: `1601.07134 [math.PR]`. URL: `http://arxiv.org/abs/1601.07134`.

Borgs, Christian, Jennifer T. Chayes, Henry Cohn, and Yufei Zhao (2014). "An L^p theory of sparse graph convergence I: limits, sparse random graph models and power law distributions". In: arXiv: `1401.2906 [math.CO]`. URL: `http://arxiv.org/abs/1401.2906`.

Borgs, Christian, Jennifer T. Chayes, Constantinos Daskalakis, and Sebastien Roch (2007). "First to market is not everything: an analysis of preferential attachment with fitness". In: *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, pp. 135–144.

Box, George E. P. (1980). "Sampling and Bayes' Inference in Scientific Modelling and Robustness". English. In: *Journal of the Royal Statistical Society. Series A (General)* 143.4, pp. 383–430.

Boyd, Stephen, Persi Diaconis, Pablo Parrilo, and Lin Xiao (2009). "Fastest Mixing Markov Chain on Graphs with Symmetries". In: *SIAM Journal on Optimization* 20.2, pp. 792–819.

Butland, Gareth, Jose Manuel Peregrin-Alvarez, Joyce Li, Wehong Yang, Xiaochun Yang, Veronica Canadien, Andrei Starostine, Dawn Richards, Bryan Beattie, Nevan Krogan, Michael Davey, John Parkinson, Jack Greenblatt, and Andrew Emili (2005). "Interaction network containing conserved and essential protein complexes in Escherichia coli". In: *Nature* 433.7025, pp. 531–537.

Butler, Steve (2008). "Eigenvalues and structures of graphs". PhD thesis. U.C. San Diego.

— (2015). "Using twins and scaling to construct cospectral graphs for the normalized Laplacian". In: *Electronic Journal of Linear Algebra* 28.1.

— (2016). "Algebraic aspects of the normalized Laplacian". In: *Recent Trends in Combinatorics*. Ed. by Andrew Beveridge, Jerrold R. Griggs, Leslie Hogben, Gregg Musiker, and Prasad Tetali. Springer International Publishing, pp. 295–315.

Cai, Diana, Trevor Campbell, and Tamara Broderick (2016). "Edge-exchangeable graphs and sparsity". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., pp. 4242–4250.

Caron, Francois and Emily B. Fox (2015). "Sparse graphs using exchangeable random measures". In: arXiv: `1401.1137 [stat.ME]`. URL: `http://arxiv.org/abs/1401.1137`.

Chatterjee, Sourav, Persi Diaconis, and Allan Sly (2011). "Random graphs with a given degree sequence". In: *Ann. Appl. Probab.* 21.4, pp. 1400–1435.

Chen, Bo and Matthias Winkel (2013). "Restricted exchangeable partitions and embedding of associated hierarchies in continuum random trees". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 49. 3. Institut Henri Poincaré, pp. 839–872.

Chopin, Nicolas and Sumeetpal S. Singh (2015). "On particle Gibbs sampling". In: *Bernoulli* 21.3, pp. 1855–1883.

Chung, Fan (1997). *Spectral Graph Theory*. Vol. 92. CBMS Regional Conference Series in Mathematics. American Mathematical Society.

Chung, Fan and Linyuan Lu (2002). "Connected Components in Random Graphs with Given Expected Degree Sequences". In: *Annals of Combinatorics* 6.2, pp. 125–145.

— (2003). "The Average Distance in a Random Graph with Given Expected Degrees". In: *Internet Math.* 1.1, pp. 91–113.

— (2006). *Complex Graphs and Networks*. Vol. 107. CBMS Regional Conference Series in Mathematics. American Mathematical Society.

Chung, Fan, Linyuan Lu, T. Gregory Dewey, and David J. Galas (2003). "Duplication Models for Biological Networks". In: *Journal of Computational Biology* 10.5, pp. 677–687.

Çinlar, Erhan (2011). *Probability and Stochastics*. Graduate Texts in Mathematics. Springer-Verlag New York.

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

Cooper, Colin and Alan Frieze (2003). "A general model of web graphs". In: *Random Structures & Algorithms* 22.3, pp. 311–335.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. London: Chapman & Hall, p. 511.

Crane, Harry (2016). "The Ubiquitous Ewens Sampling Formula (with discussion and a rejoinder by the author)". In: *Statist. Sci.* 31.1, pp. 1–19.

Crane, Harry and Walter Dempsey (2015a). "A framework for statistical network modeling". In: arXiv: 1509.08185 [math.ST]. URL: https://arxiv.org/abs/1509.08185.

— (2015b). "Atypical scaling behavior persists in real world interaction networks". In: arXiv: 1509.08184 [cs.SI]. URL: https://arxiv.org/abs/1509.08184.

— (2016). "Edge exchangeable models for network data". In: arXiv: 1603.04571 [math.ST]. URL: https://arxiv.org/abs/1603.04571.

De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero (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.

de Finetti, Bruno (1930). "Fuzione caratteristica di un fenomeno aleatorio". In: *Mem. R. Acc. Lincei* 6.4, pp. 86–133.

— (1937). "La prévision: ses lois logiques, ses sources subjectives". In: *Ann. Inst. H. Poincaré* 7, pp. 1–68.

de Solla Price, Derek J. (1965). "Networks of Scientific Papers". In: *Science* 149.3683, pp. 510–515.

Del Moral, Pierre (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Probability and Its Applications. Springer New York.

Del Moral, Pierre and Lawrence M. Murray (2015). "Sequential Monte Carlo with highly informative observations". In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1, pp. 969–997.

Deng, Kun and Dayu Huang (2015). "Optimal Kullback–Leibler approximation of Markov chains via nuclear norm regularisation". In: *International Journal of Systems Science* 46.11, pp. 2029–2047.

Diaconis, Persi and Svante Janson (2007). "Graph limits and exchangeable random graphs". In: *Rendiconti di Matematica, Serie VII* 28, pp. 33–61.

Doucet, Arnaud and Adam M. Johansen (2011). "A tutorial on particle filtering and smoothing: fifteen years later". In: *Oxford Handbook of Nonlinear Filtering*, pp. 656–704.

Dufresne, Daniel (2010). "G distributions and the beta-gamma algebra". In: *Electron. J. Probab.* 15, pp. 2163–2199.

Durrett, Rick (2006). *Random Graph Dynamics*. Cambridge Series in Statistical and Probabilistic Mathematics. New York, NY, USA: Cambridge University Press.

Dyk, David A van and Taeyoung Park (2008). "Partially Collapsed Gibbs Samplers". In: *Journal of the American Statistical Association* 103.482, pp. 790–796.

Eggenberger, F. and G. Pólya (1923). "Über die Statistik verketteter Vorgänge". In: *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik* 3.4, pp. 279–289.

Ewens, W. J. (1972). "The sampling theory of selectively neutral alleles". In: *Theoretical Population Biology* 3.1, pp. 87–112.

Faloutsos, Michalis, Petros Faloutsos, and Christos Faloutsos (1999). "On power-law relationships of the internet topology". In: *ACM SIGCOMM computer communication review*. Vol. 29. 4. ACM, pp. 251–262.

Feller, William (1971). *An Introduction to Probability Theory and Its Applications*. 2nd. Vol. 2. Wiley.

Fortini, Sandra, Lucia Ladelli, and Eugenio Regazzini (2000). "Exchangeability, predictive distributions and parametric models". In: *Sankhyā Ser. A* 62.1, pp. 86–109.

Fortini, Sandra and Sonia Petrone (2012). "Predictive construction of priors in Bayesian nonparametrics". In: *Braz. J. Probab. Stat.* 26.4, pp. 423–449.

Fouss, F., Luh Yen, A. Pirotte, and M. Saerens (2006). "An Experimental Investigation of Graph Kernels on a Collaborative Recommendation Task". In: *Data Mining, 2006. ICDM '06. Sixth International Conference on*, pp. 863–868.

Freedman, David A. (1965). "Bernard Friedman's Urn". In: *Ann. Math. Statist.* 36.3, pp. 956–970.

Gelman, Andrew, Xiao-li Meng, and Hal Stern (1996). "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies". In: *Statistica Sinica* 6.4, pp. 733–807.

Globerson, A., G. Chechik, F. Pereira, and N. Tishby (2007). "Euclidean Embedding of Co-occurrence Data". In: *The Journal of Machine Learning Research* 8, pp. 2265–2295.

Gnedin, Alexander (2006). "Constrained exchangeable partitions". In: *Fourth Colloquium on Mathematics and Computer Science Algorithms, Trees, Combinatorics and Probabilities*. Discrete Mathematics and Theoretical Computer Science, pp. 391–398.

Gnedin, Alexander and Vadim Gorin (2015). "Record-dependent measures on the symmetric groups". In: *Random Structures & Algorithms* 46.4, pp. 688–706.

Gnedin, Alexander, Ben Hansen, and Jim Pitman (2007). "Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws". In: *Probab. Surveys* 4, pp. 146–171.

Gnedin, Alexander and Jim Pitman (2006). "Exchangeable Gibbs partitions and Stirling triangles". In: *Journal of Mathematical Sciences* 138.3, pp. 5674–5685.

Goldenberg, Anna, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi (2010). "A Survey of Statistical Network Models". In: *Foundations and Trends in Machine Learning* 2.2, pp. 129–233.

Gouet, Raúl (1989). "A martingale approach to strong convergence in a generalized Pólya-Eggenberger urn model". In: *Statistics & Probability Letters* 8.3, pp. 225–228.

— (1993). "Martingale Functional Central Limit Theorems for a Generalized Polya Urn". In: *Ann. Probab.* 21.3, pp. 1624–1639.

— (1997). "Strong convergence of proportions in a multicolor Pólya urn". In: *Journal of Applied Probability* 34.2, pp. 426–435.

Gradshteyn, I.S. and I.M. Ryzhik (2015). *Table of Integrals, Series, and Products*. Ed. by Daniel Zwillinger and Victor Moll. Eighth Edition. Boston: Academic Press.

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

Griffiths, Robert C. and Simon Tavaré (1994a). "Ancestral Inference in Population Genetics". In: *Statist. Sci.* 9.3, pp. 307–319.

— (1994b). "Simulating Probability Distributions in the Coalescent". In: *Theoretical Population Biology* 46.2, pp. 131–159.

Helmstaedter, Moritz (2015). "The Mutual Inspirations of Machine Learning and Neuroscience". In: *Neuron* 86.1, pp. 25–28.

Herlau, Tue, Mikkel N Schmidt, and Morten Mørup (2016). "Completely random measures for modelling block-structured sparse networks". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., pp. 4260–4268.

Hewitt, Edwin and Leonard J. Savage (1955). "Symmetric measures on Cartesian products". In: *Transactions of the American Mathematical Society* 80.2, pp. 470–501.

Hjort, Nils Lid, Chris Holmes, Peter Müller, and Stephen G Walker (2010). *Bayesian nonparametrics*. Vol. 28. Cambridge University Press.

Ho, Ngoc-Diep and Paul Van Dooren (2008). "Non-negative matrix factorization with fixed row and column sums". In: *Linear Algebra and its Applications* 429.5–6, pp. 1020–1025.

Hoff, Peter D. (2008). "Modeling homophily and stochastic equivalence in symmetric relational data". In: *Advances in Neural Information Processing Systems 20*. Ed. by J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis. Curran Associates, Inc., pp. 657–664.

Hofstad, Remco van der (2016). *Random Graphs and Complex Networks*. Vol. 1. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Holland, Paul W. and Samuel Leinhardt (1971). "Transitivity in structural models of small groups". In: *Comparative Group Studies* 2.2, pp. 107–124.

— (1976). "Local structure in social networks". In: *Sociological methodology* 7, pp. 1–45.

— (1977). "A dynamic model for social networks". In: *The Journal of Mathematical Sociology* 5.1, pp. 5–20.

— (1981). "An Exponential Family of Probability Distributions for Directed Graphs". In: *Journal of the American Statistical Association* 76.373, pp. 33–50.

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

Hoppe, Fred M. (1984). "Pólya-like urns and the Ewens' sampling formula". In: *Journal of Mathematical Biology* 20.1, pp. 91–94.

Hripcsak, George and David J Albers (2012). "Next-generation phenotyping of electronic health records". In: *Journal of the American Medical Informatics Association* 20.1, p. 117.

Huggins, Jonathan H. and Daniel M. Roy (2015). "Convergence of Sequential Monte Carlo-based Sampling Methods". In: arXiv: `1503.00966 [math.ST]`.

Hunter, David R., Steven M. Goodreau, and Mark S. Handcock (2008). "Goodness of Fit of Social Network Models". In: *Journal of the American Statistical Association* 103.481, pp. 248–258.

Hunter, David R., Pavel N. Krivitsky, and Michael Schweinberger (2012). "Computational Statistical Methods for Social Network Models". In: *Journal of Computational and Graphical Statistics* 21.4, pp. 856–882.

James, Lancelot F. (2015). "Generalized Mittag Leffler distributions arising as limits in preferential attachment models". In: arXiv: `1509.07150 [math.PR]`. URL: `http://arxiv.org/abs/1509.07150`.

Janson, Svante (2006). "Limit theorems for triangular urn schemes". In: *Probability Theory and Related Fields* 134.3, pp. 417–452.

— (2010). "Moments of Gamma type and the Brownian supremum process area". In: *Probab. Surveys* 7, pp. 1–52.

— (2016). "Graphons and cut metric on sigma-finite measure spaces". In: arXiv: `1608.01833 [math.CO]`. URL: `https://arxiv.org/abs/1608.01833`.

Jensen, Peter B., Lars J. Jensen, and Søren Brunak (2012). "Mining electronic health records: towards better research applications and clinical care". In: *Nature Reviews Genetics* 13.6, pp. 395–405.

Johnson, N.L. and S. Kotz (1977). *Urn models and their application: an approach to modern discrete probability theory*. Wiley.

Kallenberg, Olav (2005). *Probabilistic Symmetries and Invariance Principles*. Springer.

Kemp, Charles, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda (2006). "Learning Systems of Concepts with an Infinite Relational Model". In: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*. AAAI'06. Boston, Massachusetts: AAAI Press, pp. 381–388.

Kendall, David G. (1966). "Branching Processes Since 1873". In: *Journal of the London Mathematical Society* s1-41.1, pp. 385–406. ISSN: 1469-7750.

Kerov, S. V. (2006). "Coherent random allocations, and the Ewens-Pitman formula". In: *Journal of Mathematical Sciences* 138.3, pp. 5699–5710.

Kingman, J. F. C. (1978a). "Random Partitions in Population Genetics". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 361.1704, pp. 1–20.

— (1978b). "The Representation of Partition Structures". In: *Journal of the London Mathematical Society* s2-18.2, pp. 374–380.

Kirichenko, Alisa and Harry van Zanten (2015). "Estimating a smooth function on a large graph by Bayesian Laplacian regularisation". In: arXiv: `1511.02515 [math.ST]`. URL: `https://arxiv.org/abs/1511.02515`.

Kivelä, Mikko, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter (2014). "Multilayer networks". In: *Journal of Complex Networks* 2.3, p. 203.

Kohane, Isaac S. (2011). "Using electronic health records to drive discovery in disease genomics". In: *Nature Reviews Genetics* 12.6, pp. 417–428.

Kolaczyk, Eric D. (2009). *Statistical Analysis of Network Data*. Springer-Verlag New York.

Kondor, Risi and John D. Lafferty (2002). "Diffusion Kernels on Graphs and Other Discrete Input Spaces". In: *Proceedings of the Nineteenth International Conference on Machine Learning*. ICML '02, pp. 315–322.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, pp. 1106–1114.

Lafferty, John D. and Guy Lebanon (2005). "Diffusion Kernels on Statistical Manifolds". In: *J. Mach. Learn. Res.* 6, pp. 129–163.

Leskovec, Jure, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani (2010). "Kronecker Graphs: An Approach to Modeling Networks". In: *J. Mach. Learn. Res.* 11, pp. 985–1042.

Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos (2007). "Graph evolution: Densification and shrinking diameters". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1, p. 2.

Levin, David Asher, Yuval Peres, and Elizabeth Lee Wilmer (2009). *Markov chains and mixing times*. American Mathematical Society.

Libbrecht, Maxwell W. and William S. Noble (2015). "Machine learning applications in genetics and genomics". In: *Nature Reviews Genetics* 16.6, pp. 321–332.

Lin, L., K. F. Liu, and J. Sloan (2000). "A noisy Monte Carlo algorithm". In: *Phys. Rev. D* 61 (7), p. 074505.

Lindsten, Fredrik, Michael I. Jordan, and Thomas B. Schön (2014). "Particle Gibbs with Ancestor Sampling". In: *J. Mach. Learn. Res.* 15, pp. 2145–2184.

Lindsten, Fredrik and Thomas B. Schön (2012). "On the use of backward simulation in the particle Gibbs sampler". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3845–3848.

— (2013). "Backward Simulation Methods for Monte Carlo Statistical Inference". In: *Foundations and Trends in Machine Learning* 6.1, pp. 1–143.

Lloyd, James R., Peter Orbanz, Zoubin Ghahramani, and Daniel M. Roy (2012). "Random function priors for exchangeable arrays". In: *Adv. in Neural Inform. Processing Syst. 25*, pp. 1007–1015.

Loomis, Charles P., Julio O. Morales, Roy A. Clifford, and Olen E. Leonard (1953). *Turrialba: Social Systems and the Introduction of Change*. Glencoe, Ill.: The Free Press.

Lovász, L. (2013). *Large Networks and Graph Limits*. American Mathematical Society.

Mahmoud, Hosam (2008). *Pólya Urn Models*. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC.

Miller, Kurt, Michael I. Jordan, and Thomas L. Griffiths (2009). "Nonparametric Latent Feature Models for Link Prediction". In: *Advances in Neural Information Processing Systems 22*. Ed. by Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta. Curran Associates, Inc., pp. 1276–1284.

Mitzenmacher, Michael (2003). "A Brief History of Generative Models for Power Law and Lognormal Distributions". In: *Internet Math.* 1.2, pp. 226–251.

Móri, Tamás F. (2005). "The Maximum Degree of the Barabási-Albért Random Tree". In: *Comb. Probab. Comput.* 14.3, pp. 339–348.

Nacu, Şerban (2006). "Increments of Random Partitions". In: *Combinatorics, Probability & Computing* 15.4, pp. 589–595.

Naesseth, Christian Andersson, Fredrik Lindsten, and Thomas B. Schön (2014). "Sequential Monte Carlo for Graphical Models". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, pp. 1862–1870.

Newman, M. E. J. (2009). *Networks. An Introduction*. Oxford University Press.

Nowicki, Krzysztof and Tom A. B Snijders (2001). "Estimation and Prediction for Stochastic Blockstructures". In: *Journal of the American Statistical Association* 96.455, pp. 1077–1087.

Orbanz, Peter and Daniel 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, Erol A., Nathan Ross, and Adrian Röllin (2014). "Joint degree distributions of preferential attachment random graphs". In: arXiv: `1402.4686 [math.PR]`. URL: `http://arxiv.org/abs/1402.4686`.

Pemantle, Robin (2007). "A survey of random processes with reinforcement". In: *Probab. Surveys* 4, pp. 1–79.

Pitman, Jim (1995). "Exchangeable and partially exchangeable random partitions". In: *Probability Theory and Related Fields* 102.2, pp. 145–158.

— (1996). "Some developments of the Blackwell-MacQueen urn scheme". In: *Statistics, probability and game theory*. Ed. by T. S. Ferguson, L. S. Shapley, and J. B. MacQueen. Vol. Volume 30. Lecture Notes–Monograph Series. Hayward, CA: Institute of Mathematical Statistics, pp. 245–267.

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

Pitman, Jim and Marc Yor (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator". In: *Ann. Probab.* 25.2, pp. 855–900.

Pitt, Michael, Ralph Silva, Paolo Giordani, and Robert Kohn (2010). "Auxiliary Particle filtering within adaptive Metropolis-Hastings Sampling". In: arXiv: `1006.1914 [stat.ME]`. URL: `https://arxiv.org/abs/1006.1914`.

Riordan, Oliver (2012). "The Phase Transition in the Configuration Model". In: *Combinatorics, Probability and Computing* 21.1-2, pp. 265–299.

Robert, Christian P. and George Casella (2004). *Monte Carlo Statistical Methods*. Springer New York.

Roy, Daniel M. and Yee Whye Teh (2009). "The Mondrian Process". In: *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Curran Associates, Inc., pp. 1377–1384.

Salakhutdinov, Ruslan and Andriy Mnih (2008). "Probabilistic Matrix Factorization". In: *Advances in Neural Information Processing Systems 20*. Vol. 20.

Shalizi, Cosma Rohilla and Alessandro Rinaldo (2013). "Consistency under sampling of exponential random graph models". In: *Ann. Statist.* 41.2, pp. 508–535.

Simkin, M. V. and V. P. Roychowdhury (2011). "Re-inventing Willis". In: *Physics Reports* 502.1, pp. 1–35.

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

Smola, A. J. and Risi Kondor (2003). "Kernels and Regularization on Graphs". In: *Proceedings of the Annual Conference on Computational Learning Theory*. Ed. by B. Schölkopf and M. Warmuth. Lecture Notes in Computer Science. Springer.

Srivastava, Ashok N. and Mehran Sahami, eds. (2009). *Text Mining: Classification, Clustering, and Applications*. Data Mining and Knowledge Discovery. Chapman & Hall/CRC.

Thorne, Thomas and Michael P. H. Stumpf (2012). "Graph spectral analysis of protein interaction network evolution". In: *Journal of The Royal Society Interface*.

Travers, Jeffrey and Stanley Milgram (1969). "An Experimental Study of the Small World Problem". In: *Sociometry* 32.4, pp. 425–443.

Tricomi, F. G. and A. Erdélyi (1951). "The asymptotic expansion of a ratio of gamma functions." In: *Pacific J. Math.* 1.1, pp. 133–142.

Veitch, Victor and Daniel M. Roy (2015). "The Class of Random Graphs Arising from Exchangeable Random Measures". In: arXiv: `1512.03099 [math.ST]`. URL: `http://arxiv.org/abs/1512.03099`.

— (2016). "Sampling and Estimation for (Sparse) Exchangeable Graphs". In: arXiv: `1611.00843 [math.ST]`. URL: `https://arxiv.org/abs/1611.00843`.

von Luxburg, Ulrike (2007). "A tutorial on spectral clustering". English. In: *Statistics and Computing* 17.4, pp. 395–416.

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

Wang, Junshan, Ajay Jasra, and Maria De Iorio (2014). "Computational Methods for a Class of Network Models". In: *Journal of Computational Biology* 21.2, pp. 141–161.

Watts, Duncan J. and Steven H. Strogatz (1998). "Collective dynamics of 'small-world' networks". In: *Nature* 393.6684, pp. 440–442.

Whiteley, Nick (2010). "Discussion of "Particle Markov Chain Monte Carlo Methods" by Andrieu et. al." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.3, pp. 306–307.

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

Wiuf, Carsten, Markus Brameier, Oskar Hagberg, and Michael P. H. Stumpf (2006). "A likelihood approach to analysis of network data". In: *Proceedings of the National Academy of Sciences* 103.20, pp. 7566–7570.

Xu, Zhao, Volker Tresp, Kai Yu, and Hans Peter Kriegel (2006). "Infinite Hidden Relational Models". In: *Proceedings of the 22nd International Conference on Uncertainity in Artificial Intelligence (UAI 2006)*. Cambridge, MA, USA: AUAI Press.

Yule, G. Udny (1925). "A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S." In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 213.402-410, pp. 21–87.

Zhou, Mingyuan (2015). "Infinite edge partition models for overlapping community detection and link prediction". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Guy Lebanon and S. V. N. Vishwanathan. Vol. 38, pp. 1135–1143.

Appendix A

Proofs for Chapter 3

A.1 Proof of Proposition 3.1

If K has law P with probability generating function $G(z)$, then

$$\begin{aligned}
\mathbb{E}_P[(D^{-1}A)^K] &= \sum_k (D^{-1}A)^K P(K=k) = D^{-1/2}\left[\sum_k (\mathbb{I}_n - \mathbf{L})^k P(K=k)\right] D^{1/2} \\
&= D^{-1/2}\left[\sum_k \sum_{i=1}^n (1-\sigma_i)^k \boldsymbol{\psi}_i \boldsymbol{\psi}_i' P(K=k)\right] D^{1/2} \\
&= D^{-1/2}\left[\sum_{i=1}^n \sum_k (1-\sigma_i)^k \boldsymbol{\psi}_i \boldsymbol{\psi}_i' P(K=k)\right] D^{1/2} \\
&= D^{-1/2}\left[\sum_{i=1}^n G_P(1-\sigma_i) \boldsymbol{\psi}_i \boldsymbol{\psi}_i'\right] D^{1/2} \\
&= D^{-1/2} G_P(\mathbb{I}_n - \mathbf{L}) D^{1/2} \,. \qquad \text{(A.1)}
\end{aligned}$$

$\mathbf{L}$ is symmetric and positive semi-definite, so the eigenvalue decomposition always exists. Moreover, it is well known that the eigenvalues of $\mathbf{L}$ are $\sigma_i \in [0, 2]$ (Chung, 1997), so the proof holds if $G(z)$ exists for $|z| \leq 1$.

A.2 Proof of Proposition 3.2

The result can be read from Proposition 3.1, but we include the following calculations for completeness. If K has law $\text{Poisson}_+(\lambda)$, (3.1) becomes

$$\begin{aligned} Q^\lambda &= \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!}(D^{-1}A)^{k+1} = D^{-1/2}\left[(\mathbb{I}_n - \mathbf{L})\, e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k\,(\mathbb{I}_n-\mathbf{L})^k}{k!}\right] D^{1/2} \\ &= D^{-1/2}\left[(\mathbb{I}_n - \mathbf{L})\, e^{-\lambda\mathbf{L}}\right] D^{1/2} = D^{-1/2}\left[(\mathbb{I}_n - \mathbf{L})\,\mathbf{K}^\lambda\right] D^{1/2}. \end{aligned}$$

Since the spectral norm of $\mathbb{I} - \mathbf{L}$ is 1, series is absolutely convergent. For $K \sim \text{NB}_+(r,p)$,

$$\sum_{k=0}^{\infty}\frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)}p^k(1-p)^r(D^{-1}A)^{k+1} = D^{-1/2}\left[(\mathbb{I}_n-\mathbf{L})\left(\mathbb{I}_n + \frac{p}{1-p}\mathbf{L}\right)^{-r}\right] D^{1/2}$$

similarly yields $Q^{r,p} = D^{-1/2}\left[(\mathbb{I}_n - \mathbf{L})\,\mathbf{K}^{r,p}\right] D^{1/2}$.

A.3 Proof of Theorem 3.3

We begin with a lemma used in the proofs of Theorem 3.3 and Theorem 3.4. Let $\delta_j^{(1)}(t)$ be an indicator variable that is equal to 1 if the first end of the t-th edge is attached to vertex v_j, and 0 otherwise; likewise define $\delta_j^{(2)}(t)$ for the second end of the t-th edge.

LEMMA A.1. *Suppose a sequence of graphs is generated with law* $\mathbf{RW}_{\text{SB}}(\beta, P)$*, such that* P *is a probability distribution on* $\mathbb{N}$*. Then the size-biased distribution* $\mathbb{S}_t$ *is left-invariant under the mixed random walk probability* Q_t *induced by* P*, that is*

$$\mathbb{S}_t' Q_t = \mathbb{S}_t' \qquad \textit{where } [Q_t]_{uv} := \mathbb{P}\{V_{end} = v \mid V_0 = u, G_t\} \tag{A.2}$$

holds for every $t \in \mathbb{N}_+$. Furthermore, for each $v_j \in \mathbf{V}(G_t)$, the following holds:

$$\mathbb{E}[\delta_j^{(1)}(t+1) \mid G_t] = \frac{1}{1-\beta}\mathbb{E}[\delta_j^{(2)}(t+1) \mid G_t] = \frac{\deg_t(v_j)}{2t} \,. \tag{A.3}$$

Proof. Substituting (3.1) into $\mathbb{S}_t' Q_t$ yields

$$\mathbb{S}_t' Q_t = \sum_k \mathbb{S}_t'(D^{-1}A)^k P\{K=k\} = \sum_k \mathbb{S}_t' P\{K=k\} = \mathbb{S}_t' \,. \tag{A.4}$$

The $\mathbf{RW}_{\text{SB}}$ model samples the first end of each edge from the size-biased distribution:

$$\mathbb{E}[\delta_j^{(1)}(t+1) \mid G_t] = \mathbb{P}[V_{t+1} = v_j \mid G_t] = \frac{\deg_t(v_j)}{2t} \,. \tag{A.5}$$

For the second end, denote by $\mathbb{1}_{j,t}$ the indicator vector for vertex v_j. Then

$$\mathbb{E}[\delta_j^{(2)}(t+1) \mid G_t] = \sum_u \mathbb{P}[V_{t+1} = u \mid G_t]\big[Q_t\big]_{uv} = (1-\beta)\mathbb{S}_t' Q_t \mathbb{1}_{j,t} = (1-\beta)\frac{\deg_t(v_j)}{2t} \,.$$

□

In the setup of Theorem 3.3, asymptotic expected degree counts are as follows:

Lemma A.2. *Let a sequence of multigraphs $(G_1, G_2, \dots)$ have law $\mathbf{RW}_{\text{SB}}(\beta, P)$, for a probability distribution P on $\mathbb{N}$ satisfying Equation* (3.11)*. Then*

$$\frac{\mathbb{E}[m_{d,t}]}{\beta t} \xrightarrow{a.s.} p_d \qquad \text{for each } d \in \mathbb{N} \text{ as } t \to \infty \,, \tag{A.6}$$

where p_d is given in (3.9)*.*

PROOF. Equations (A.3) and (3.11) yield the recurrence

$$\mathbb{E}(m_{d,t+1}) = \beta \cdot \delta_1(d) + \mathbb{E}(m_{d,t})\Big(1 - \frac{(2-\beta)d}{2t}\Big) + \mathbb{E}(m_{d-1,t})\frac{(2-\beta)(d-1)}{2t} + o(1) \ ,$$

with $m_{0,t} = 0$ for all t. This can be written more generally as

$$M(d, t+1) = (1 - b(t)/t)M(d,t) + g(t) \ . \tag{A.7}$$

If $b(t) \to b$ and $g(t) \to g$, then $M(d,t)/t \to g/(b+1)$ (e.g. Durrett, 2006, Lemmas 4.1.1, 4.1.2). For $d = 1$, we have $b(t) = b = (2-\beta)/2$ and $g(t) = g = \beta$, so $\mathbb{E}(m_{1,t})/t \to \frac{2\beta}{4-\beta} = \beta p_1$. Proceeding by induction, $b = (2-\beta)d/2$ and $g = \beta p_{d-1}(2-\beta)(d-1)/2$ yield

$$\frac{\mathbb{E}[m_{d,t}]}{\beta t} \quad \to \quad p_{d-1}\tfrac{(2-\beta)(d-1)}{2+(2-\beta)d} = \frac{2}{4-\beta}\prod_{j=1}^{d-1}\frac{j}{\frac{2}{2-\beta}+j+1} = \frac{2}{2-\beta}\frac{\Gamma(d)\Gamma(2+\frac{\beta}{2-\beta})}{\Gamma(d+2+\frac{\beta}{2-\beta})} \ ,$$

for $d > 1$, which is just p_d as defined in (3.9). □

The following result, the proof of which can be found in Chung and Lu (2006, Section 3.6), shows that the random variable $m_{d,t}$ concentrates about its mean:

LEMMA A.3. *Let a sequence of multigraphs* $(G_1, G_2, \dots)$ *have law* $\mathbf{RW}_{\text{SB}}(\beta, P)$*, for a probability distribution* P *on* $\mathbb{N}$ *satisfying Equation* (3.11)*. Then*

$$\mathbb{P}\Big[\ \Big|\frac{m_{d,t}}{\beta t} - p_d\Big| \le 2\sqrt{\frac{d^3 \log t}{\beta^2 t}}\ \Big] \ge 1 - 2(t+1)^{d-1}t^{-d} = 1 - o(1) \ . \tag{A.8}$$

PROOF OF THEOREM 3.3. Lemma A.3 shows that $\frac{m_{d,t}}{\beta t}$ concentrates around its mean, and with Lemma A.2 showing that the mean is p_d, the proof of the theorem is complete. □

A.4 Proof of Theorem 3.4

By definition of **RW** models (see Algorithm 3.1), at each step t a new vertex appears with probability β, which is represented as a collection of Bernoulli(β) random variables $(B_t)_{t\geq 2}$. Alternatively, there is the collection of steps $S_1, S_2, \ldots, S_j, \ldots$ at which new vertices appear, such that S_j is the step at which the j-th vertex appears. By the fundamental relationship between Bernoulli and Geometric random variables, the sequence $S_1, S_2, \ldots, S_j, \ldots$ can be sampled independently of the graph sequence as $S_1 = S_2 = 1$ and

$$S_j = S_{j-1} + \Delta_j, \text{ where } \Delta_j \overset{\text{iid}}{\sim} \text{Geometric}(\beta), \text{ for } j > 2. \tag{A.9}$$

In what follows, we condition on the sequence $(S_j)_{j\geq 1}$ unless explicitly stated otherwise. We begin by calculating the expected degree of the j-th vertex.

Lemma A.4. *Let a sequence of multigraphs $(G_1, G_2, \ldots)$ have law $\mathbf{RW}_{\text{SB}}(\beta, P)$ such that Lemma A.1 applies. Let $d_j(t) := \deg_t(v_j)$ be the degree of v_j in graph G_t, where v_j is the j-th vertex to appear in the graph sequence, and let $\rho = 1 + \frac{\beta}{2-\beta}$. Then conditional on S_j, $d_j(t)t^{-1/\rho}$ converges almost surely to a random variable ξ_j as $t \to \infty$, and*

$$\mathbb{E}[d_j(t) \mid S_j] = \frac{\Gamma(S_j)\Gamma(t + \frac{1}{\rho})}{\Gamma(S_j + \frac{1}{\rho})\Gamma(t)} \tag{A.10}$$

Proof. Let $\mathcal{A}_t$ denote the σ-algebra generated after t steps. Then for $t \geq S_j$,

$$\mathbb{E}[d_j(t+1) \mid \mathcal{A}_t] = d_j(t) + \mathbb{E}[\delta_j^{(1)}(t+1) \mid \mathcal{A}_t] + \mathbb{E}[\delta_j^{(2)}(t+1) \mid \mathcal{A}_t] = d_j(t)(1 + \tfrac{1}{\rho t}) \ ,$$

where the final identity follows from Lemma A.1, and $d_j(S_j) = 1$. The sequence

$$M_j(t) := d_j(t)\frac{\Gamma(S_j + \frac{1}{\rho})}{\Gamma(S_j)} \cdot \frac{\Gamma(t)}{\Gamma(t + \frac{1}{\rho})} \qquad \text{for } t \geq S_j$$

is a non-negative martingale with mean 1, by (A.4). Therefore, $M_j(t)$ converges almost surely to a random variable M_j as $t \to \infty$. Taking expectations on both sides and rearranging (A.4) yields (A.10). With Stirling's formula,

$$\frac{d_j(t)}{t^{1/\rho}} \xrightarrow{\text{a.s.}} M_j \frac{\Gamma(S_j)}{\Gamma(S_j + \frac{1}{\rho})} := \xi_j \ . \tag{A.11}$$

□

We now consider the joint distribution of the limiting random variables $(\xi_j)_{j\geq 1}$. The approach, which is adapted from Móri (2005) (Durrett, 2006; Hofstad, 2016, see also), is to analyze a martingale that yields the moments of $(\xi_j)_{j\geq 1}$. First, define for $t \geq S_j$,

$$R_{j,k}(t) := \frac{\Gamma(d_j(t) + k)}{\Gamma(d_j(t))\Gamma(k + 1)} \ , \tag{A.12}$$

At a high level, $R_{j,k}(t) \approx d_j(t)^k/k!$ for large t, so (A.11) shows that properly scaled $R_{j,k}(t)$ should converge to $\xi_j^k/k!$ for each j. We make this precise in what follows.

Let $\mathbf{j} := (j_i)$ be an ordered collection of vertices such that $1 \leq j_1 < j_2 < \cdots < j_r$, and let $\mathbf{k} := (k_i)$ be a corresponding vector of moments. To define a suitable martingale, we

abbreviate

$$\mu_{\mathbf{k}} := \sum_{i=1}^{r} k_i \quad \text{and} \quad \Sigma_{\mathbf{j},\mathbf{k}}(t) := \sum_{i,l=1}^{r} \left(\frac{1}{2}\right)^{\mathbb{1}_{i=l}} \frac{k_i(k_l - \mathbb{1}_{i=l})}{d_{j_l}(t) + \mathbb{1}_{i=l}} [Q_t]_{j_i j_l} ,$$

and note $[Q_t]_{jj'} = 0$ for all $t < \max\{S_j, S_{j'}\}$. Define

$$c_{\mathbf{j},\mathbf{k}}(t) = \Gamma(t)\Big[\prod_{s=0}^{t-1}\big(s + \tfrac{\mu_{\mathbf{k}}}{\rho} + \tfrac{1-\beta}{2}\Sigma_{\mathbf{j},\mathbf{k}}(s)\big)\Big]^{-1} , \tag{A.13}$$

which is a random variable since Q_t is random, and is $\mathcal{A}_t$-measurable for each t. Since $[Q_t]_{j_i j_l} < 1$, and since $d_j(t) \xrightarrow{\text{a.s.}} \infty$ as $t \to \infty$ by by Lemma A.4, $\Sigma_{\mathbf{j},\mathbf{k}}(t) \xrightarrow{\text{a.s.}} 0$ as $t \to \infty$, which yields

$$c_{\mathbf{j},\mathbf{k}}(t) = \frac{\Gamma(t)}{\Gamma(t + \frac{1}{\rho}\mu_{\mathbf{k}})}(1 + o(1)) = t^{-\mu_{\mathbf{k}}/\rho}(1 + o(1)) \quad \text{as } t \to \infty . \tag{A.14}$$

Note that

$$\lim_{t\to\infty} c_{\mathbf{j},\mathbf{k}}(t) \prod_{i=1}^{r} R_{j_i,k_i} = \prod_{i=1}^{r} \frac{\xi_{j_i}^{k_i}}{\Gamma(k_i + 1)} , \tag{A.15}$$

if the limit exists. Existence is based on the following result:

LEMMA A.5. *Let $\mathcal{A}_t$ denote the σ-algebra generated by a $\mathbf{RW}_{\text{SB}}(\beta, P)$ multigraph sequence up to step t. Let $r > 0$ and $1 \le j_1 < j_2 < \cdots < j_r$ be integers, and real-valued $k_1, k_2, \ldots, k_r > -1$. Then with $R_{j,k}(t)$ defined in* (A.12) *and $c_{\mathbf{j},\mathbf{k}}(t)$ defined in* (A.13),

$$Z_{\mathbf{j},\mathbf{k}}(t) := c_{\mathbf{j},\mathbf{k}}(t) \prod_{i=1}^{r} R_{j_i,k_i}(t) \tag{A.16}$$

is a nonnegative martingale for $t \geq \max\{S_{j_r}, 1\}$. *If* $k_1, k_2, \ldots, k_r > -\frac{1}{2}$, *then* $Z_{\mathbf{j},\mathbf{k}}(t)$ *converges in* L_2.

PROOF. It can be shown that

$$\mathbb{E}\Big[\prod_{i=1}^{r} R_{j_i,k_i}(t+1) \;\Big|\; \mathcal{A}_t\Big] = \Big(\prod_{i=1}^{r} R_{j_i,k_i}(t)\Big)\Big(1 + \frac{\mu_{\mathbf{k}}}{\rho t} + \frac{1-\beta}{2t}\Sigma_{\mathbf{j},\mathbf{k}}(t)\Big)\,, \tag{A.17}$$

from which it follows that

$$\mathbb{E}[Z_{\mathbf{j},\mathbf{k}}(t+1) \mid \mathcal{A}_t] = c_{\mathbf{j},\mathbf{k}}(t+1)R_{\mathbf{j},\mathbf{k}}(t)\Big(1 + \frac{\mu_{\mathbf{k}}}{\rho t} + \frac{1-\beta}{2t}\Sigma_{\mathbf{j},\mathbf{k}}(t)\Big) = c_{\mathbf{j},\mathbf{k}}(t)R_{\mathbf{j},\mathbf{k}}(t) = Z_{\mathbf{j},\mathbf{k}}(t)\,. \tag{A.18}$$

Furthermore, $Z_{\mathbf{j},\mathbf{k}}(\max\{S_{j_r}, 1\}) > 0$. By (A.14) and properties of the gamma function,

$$Z_{\mathbf{j},\mathbf{k}}(t)^2 \leq Z_{\mathbf{j},2\mathbf{k}}(t)\prod_{i=1}^{r}\binom{2k_i}{k_i}\,. \tag{A.19}$$

By (A.18), $Z_{\mathbf{j},2\mathbf{k}}(t)$ is a martingale with finite expectation for $2k_1, \ldots, 2k_r > -1$. Therefore, $Z_{\mathbf{j},\mathbf{k}}(t)$ is an L_2-bounded martingale and hence converges in L_2 (and also in L_1) for $k_1, \ldots, k_r > -\frac{1}{2}$. □

Combining the auxiliary results above, we can now give proof of the result.

PROOF OF THEOREM 3.4. The limit of $Z_{\mathbf{j},\mathbf{k}}(t)$ is (A.15), which enables calculation of the moments of ξ_j. In particular, for any vertex v_j, $j \in \mathbb{N}_+$, $k \in \mathbb{R}$ such that $k > -\frac{1}{2}$,

$$\mathbb{E}\Big[\tfrac{\xi_j^k}{\Gamma(k+1)} \;\Big|\; S_j\Big] = \lim_{t\to\infty}\mathbb{E}[Z_{j,k}(t) \mid S_j] = \mathbb{E}[Z_{j,k}(S_j) \mid S_j] = \tfrac{\Gamma(S_j)}{\Gamma(S_j+\frac{k}{\rho})}\,. \tag{A.20}$$

Although the joint moments involve $\Sigma_{\mathbf{j},\mathbf{k}}$, the moments (A.20) characterize the marginal distribution of $\xi_j \mid S_j$ via the Laplace transform. Since

$$\mathbb{E}[\xi_j^k \mid S_j] = \frac{\Gamma(S_j)\Gamma(k+1)}{\Gamma\big(S_j + \frac{k}{\rho}\big)} = \frac{\Gamma\left(\rho(S_j-1)+1+k\right)\Gamma(S_j)}{\Gamma\left(\rho(S_j-1)+1\right)\Gamma\big(S_j+\frac{k}{\rho}\big)} \frac{\Gamma\left(\rho(S_j-1)+1\right)\Gamma(k+1)}{\Gamma\left(\rho(S_j-1)+1+k\right)} ,$$

$\mathbb{E}[\xi_j^k \mid S_j] = \mathbb{E}[M_j^k B_j^k \mid S_j]$, where M_j is a generalized Mittag-Leffler$(\rho^{-1}, S_j - 1)$ variable (James, 2015), and B_j is $B_j \sim \text{Beta}(1, \rho(S_j - 1))$. It follows that $M_j \perp\!\!\!\perp_{S_j} B_j$.

It remains to show (3.15). We begin by noting that the left-hand side of (A.20) is an expectation that conditions on S_j. Define $\tilde{S}_j := S_j - 1 - (j-2)$, which is marginally distributed as $\text{NB}(j-2, 1-\beta)$. Then

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}[\xi_j^k \mid S_j]\big] &= \mathbb{E}\Big[\frac{\Gamma(\tilde{S}_j + 1 + (j-2))\Gamma(k+1)}{\Gamma(\tilde{S}_j + 1 + (j-2) + \frac{k}{\rho})}\Big] \\
&= \sum_{t=0}^{\infty} \frac{\Gamma(t+1+(j-2))\Gamma(k+1)}{\Gamma(t+1+(j-2)+\frac{k}{\rho})} \frac{\Gamma(t+j-2)}{\Gamma(j-2)\Gamma(t+1)} (1-\beta)^t \beta^{j-2} \qquad \text{(A.21)} \\
&= \frac{\Gamma(k+1)\Gamma(j-1)}{\Gamma(j-1+\frac{k}{\rho})} \beta^{\frac{k}{\rho}} \, {}_2F_1(1+\frac{k}{\rho}, \frac{k}{\rho}; j-1+\frac{k}{\rho}; 1-\beta) ,
\end{aligned}$$

where ${}_2F_1(a,b;c;z)$ is the ordinary hypergeometric function. For $j \to \infty$,

$$\lim_{j\to\infty} \mathbb{E}\big[\mathbb{E}[\xi_j^k \mid S_j]\big] = \lim_{j\to\infty} \tfrac{\Gamma(k+1)\Gamma(j-1)}{\Gamma(j-1+\frac{k}{\rho})} \beta^{\frac{k}{\rho}} (1 + O(j^{-1})) = \Gamma(k+1)\beta^{\frac{k}{\rho}} j^{-\frac{k}{\rho}} (1 + O(j^{-1})) .$$

follows using the series expansion of ${}_2F_1(a,b;c;z)$. □

A.5 Proof of Proposition 3.5

Proof of Proposition 3.5. Existence of the limit as $\lambda \to \infty$ follows from the existence of the limiting (stationary) distribution for each $t \in \mathbb{N}_+$. For equivalence, it suffices to show that given any connected graph G, the conditional distribution over graphs G' is the same for $\mathbf{RW}_{\text{SB}}(\beta, \infty)$ and $\mathbf{ACL}(\beta)$. The probability of attaching a new vertex to an existing vertex v in both models is $\beta \text{deg}(v)/\text{vol}(G)$. With probability $1-\beta$, a new edge is inserted between existing vertices u and v, and so it remains to show that in this case the distribution over pairs of vertices, $\nu(u,v)$, is the same. We have

$$\nu(u,v) = 2\frac{\deg(u)}{\text{vol}(G)}\frac{\deg(v)}{\text{vol}(G)} \quad \text{and} \quad \nu(u,v) = \frac{\deg(u)}{\text{vol}(G)}[Q^\lambda]_{uv} + \frac{\deg(v)}{\text{vol}(G)}[Q^\lambda]_{vu}$$

for the ACL and $\mathbf{RW}_{\text{SB}}$ model respectively, with Q^λ. First, consider $(\mathbb{I}_n - \mathbf{L})\mathbf{K}^\lambda$ in terms of the spectrum of $\mathbf{L}$. Let $0 = \sigma_1 \leq \cdots \leq \sigma_n \leq 2$ be the eigenvalues for a graph on n vertices, with eigenvectors $\boldsymbol{\psi}_i$. Then

$$(\mathbb{I}_n - \mathbf{L})\mathbf{K}^\lambda = (\mathbb{I}_n - \mathbf{L})e^{-\lambda \mathbf{L}} = \sum_{i=1}^{n}(1-\sigma_i)e^{-\lambda\sigma_i}\boldsymbol{\psi}_i\boldsymbol{\psi}_i' \,. \tag{A.22}$$

When $\lambda \to \infty$, only the eigenvector $\boldsymbol{\psi}_1$ corresponding to the eigenvalue $\sigma_1 = 0$ contributes to the random walk probabilities, i.e. $\lim_{\lambda\to\infty}(\mathbb{I}_n - \mathbf{L})e^{-\lambda\mathbf{L}} = \boldsymbol{\psi}_1\boldsymbol{\psi}_1'$. The limit satisfies $\boldsymbol{\psi}_1 \propto D^{1/2}\mathbb{1}$, where $\mathbb{1}$ is the vector of all ones (Chung, 1997, Ch. 2). Therefore,

$$\lim_{\lambda\to\infty}[Q^\lambda]_{uv} = \big[\mathbb{1}\mathbb{1}'D\tfrac{1}{\text{vol}(G)}\big]_{uv} = \frac{\deg(v)}{\text{vol}(G)} \,,$$

and the result follows. □

Appendix B

Proofs for Chapter 4

B.1 Proof of Proposition 4.1

Proof of Proposition 4.1. Proposition 2.2 of Del Moral and Murray (2015) shows that

$$\mathbb{E}[f(G_{1:T}) \mid G_1, G_T] = L_\theta^1(G_1)^{-1}\mathbb{E}\big[f(G_{1:T})L_\theta^{T-1}(G_{T-1})\tfrac{h_t(G_t)}{h_t(G_{T-1})}\textstyle\prod_{s=2}^{T-1}\tfrac{h_s(G_s)}{h_{s-1}(G_{s-1})} \mid G_1\big] ,$$

for any approximation function h_t that is positive outside a null set. For our choice of h_t, the proposal density (4.12) only places probability mass on graphs with non-zero bridge likelihood, making $\mathbb{1}\{h_t = 0\}$ a probability zero event. For an SMC approximation to be consistent, the target density must be absolutely continuous with respect to the proposal density at each step t (see, e.g. Doucet and Johansen, 2011; Robert and Casella, 2004). In other words, the proposal density must be a valid importance sampling distribution at each step. The target, $\mathcal{L}_\theta(G_{1:t} \mid G_T, G_1)$ is absolutely continuous with respect to $\prod_{s=2}^{t} r_\theta(G_s \mid G_{s-1})$ from (4.12) by construction if properties (P1) and (P2) hold, and the result follows.

□

B.2 Proof of Proposition 4.2

Proposition 4.2 is a special case of the following:

PROPOSITION B.1. *Let q_θ^t, for $t = 1, \ldots, T$, be the Markov kernels defining a sequential network model that satisfies conditions (P1) and (P2), and let r_θ^t be the corresponding proposal kernels. Furthermore, let $h_t(G_T \mid G_t)$ for $t = 1, \ldots, T-2$ be a sequence of fixed functions that are strictly positive if $G_t \subseteq G_T$. Given an observation G_T and a fixed G_1, define the weights*

$$\tilde{w}_t^i := \begin{cases} 1, & \text{for} \quad t = 1 \\ \dfrac{1}{h_t(G_T \mid G_{t-1}^i)} \dfrac{q_\theta^t(G_t^i \mid G_{t-1}^i)}{r_\theta^t(G_t^i \mid G_{t-1}^i)}, & \text{for} \quad 2 \leq t < T-1 \\ \dfrac{q_\theta^t(G_T \mid G_t)}{h_t(G_T \mid G_{t-1}^i)} \dfrac{q_\theta^t(G_t^i \mid G_{t-1}^i)}{r_\theta^t(G_t^i \mid G_{t-1}^i)}, & \text{for} \quad t = T-1 \end{cases}, \tag{B.1}$$

and w_t^i the corresponding weights normalized across the N particles. Define the estimator

$$\begin{aligned} \hat{L}_\theta^1 &:= \prod_{t=2}^{T-1} \left[\left(\sum_{i=1}^N \frac{\tilde{w}_t^i}{N} \right) \left(\frac{\sum_{i=1}^N h_{t-1}(G_T \mid G_{t-1}^i) \tilde{w}_{t-1}^i}{\sum_{i=1}^N \tilde{w}_{t-1}^i} \right) \right] \\ &:= \prod_{t=2}^{T-1} \hat{L}_\theta(G_t \mid G_{t-1}) \, . \end{aligned} \tag{B.2}$$

Then $\hat{L}_\theta^1$ is unbiased, that is $\mathbb{E}[\hat{L}_\theta^1] = L_\theta^1(G_1) = p_\theta(G_T \mid G_1)$, for any $N \geq 1$.

PROOF. Unbiasedness will be established by iterating expecations, following the approach in Pitt, Silva, Giordani, and Kohn (2010). Let $\mathcal{S}_t$ be the set of particles and weights $\{G_t^i; \tilde{w}_t^i\}$

at step t. Then we have that

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=1}^{N}\frac{\tilde{w}_{T-1}^{i}}{N}\right)\,\middle|\,\mathcal{S}_{T-2}\right]\\
&=\sum_{i=1}^{N}\int\frac{q_{\theta}^{t}(G_T\mid G_{T-1})}{h_{T-2}(G_T\mid G_{T-2}^{i})}\frac{q_{\theta}^{t}(G_{T-1}\mid G_{T-2}^{i})}{r_{\theta}^{t}(G_{T-1}\mid G_{T-2}^{i})}\frac{r_{\theta}^{t}(G_{T-1}\mid G_{T-2}^{i})h_{T-2}(G_T\mid G_{T-2}^{i})\tilde{w}_{T-2}^{i}}{\sum_{j=1}^{N}h_{T-2}(G_T\mid G_{T-2}^{j})\tilde{w}_{T-2}^{j}}dG_{T-1}\\
&=\sum_{i=1}^{N}\frac{p_{\theta}^{T,T-2}(G_T\mid G_{T-2}^{i})\tilde{w}_{T-2}^{i}}{\sum_{j=1}^{N}h_{T-2}(G_T\mid G_{T-2}^{j})\tilde{w}_{T-2}^{j}}\,.
\end{aligned}$$

Therefore,

$$\mathbb{E}[\hat{L}_{\theta}(G_{T-1}\mid G_{T-2})\mid\mathcal{S}_{T-2}]=\sum_{i=1}^{N}p_{\theta}^{T,T-2}(G_T\mid G_{T-2}^{i})w_{T-2}^{i}\,.$$

Likewise, it is straightforward to show that

$$\begin{aligned}
&\mathbb{E}[\hat{L}_{\theta}(G_{T-1}\mid G_{T-2})\hat{L}_{\theta}(G_{T-2}\mid G_{T-3})\mid\mathcal{S}_{T-3}]\\
&\quad=\mathbb{E}[\,\mathbb{E}[\hat{L}_{\theta}(G_{T-1}\mid G_{T-2})\mid\mathcal{S}_{T-2}]\,\hat{L}_{\theta}(G_{T-2}\mid G_{T-3})\mid\mathcal{S}_{T-3}]\\
&\quad=\sum_{i=1}^{N}p_{\theta}^{T,T-3}(G_T\mid G_{T-3}^{i})w_{T-3}^{i}\,.
\end{aligned}$$

Iterating for $t=T-4,\ldots,1$, we have

$$\begin{aligned}
&\mathbb{E}\left[\prod_{t=2}^{T-1}\left\{\left(\sum_{i=1}^{N}\frac{\tilde{w}_{t}^{i}}{N}\right)\left(\frac{\sum_{i=1}^{N}h_{t-1}(G_T\mid G_{t-1}^{i})\tilde{w}_{t-1}^{i}}{\sum_{i=1}^{N}\tilde{w}_{t-1}^{i}}\right)\right\}\,\middle|\,\mathcal{S}_{1}\right]\\
&=\sum_{i=1}^{N}p_{\theta}^{T,1}(G_T\mid G_{1}^{i})w_{1}^{i}=p_{\theta}(G_T\mid G_1)\,,
\end{aligned}$$

which proves the claim. □

B.3 Proof of Proposition 4.3 and Proposition 4.4

Proof of Proposition 4.3. As in the proof of the analogous Theorem 4 in Andrieu, Doucet, and Holenstein (2010), if (a) the resampling scheme in the SMC algorithm is unbiased, i.e. each particle is resampled with probability proportional to its weight; and (b) the estimate $\hat{L}_\theta^1(G_1)$ is positive and unbiased, then Algorithm 4.2 yields a MH update on the extended space that includes the state of the SMC variables. If, additionally, (c) $\mathcal{L}_\theta(G_{1:t} \mid G_T, G_1)$ is absolutely continuous with respect to $\prod_{s=2}^t r_\theta(G_s \mid G_{s-1})$ for any θ; and (d) the MH sampler targeting $P_{[\Theta]}(\Theta)$ is irreducible and aperiodic, then by Theorem 1 of Andrieu and Roberts (2009), the marginal law converges to the desired density. Condition (a) is satisfied by the multinomial, residual, and stratified resampling schemes. As discussed in the proof of Proposition 4.1, the condition (b) holds by construction if properties (P1) and (P2) hold. Finally (c) holds by Proposition 4.2 and (d) holds by assumption. □

Proof of Proposition 4.4. Algorithm 4.3 is a Gibbs sampler on the extended space that includes the state of the SMC variables. That it has the correct marginal law follows from Theorem 5 in Andrieu, Doucet, and Holenstein (2010), which requires absolute continuity as in (c) in the previous proof, and that the Gibbs sampler defined by the updates $(\beta, \lambda) \mid G_{1:T}$ and $G_{1:T} \mid (\beta, \lambda)$ is irreducible and aperiodic. Absolute continuity holds by construction, as before; the irreducibility and aperiodicity are satisfied by inspection. □

Appendix C

Particle Gibbs updates

(Section 4.3)

The particle Gibbs sampler updates are performed in three blocks: (β, λ), $(B_{2:T}, K_{2:T})$, and $G_{2:(T-1)}$. As we describe in the following subsections, we make use of conditional conjugacy and tools from spectral graph theory to increase sampling efficiency. We discuss each block of updates in turn.

Updating the parameters β and λ. The model is specified in terms of distributions for the latent variables, and placing conjugate priors on their parameters we have,

$$B_t \mid \beta \overset{\text{iid}}{\sim} \text{Bernoulli}(\beta), \quad \beta \sim \text{Beta}(a_\beta, b_\beta) \tag{C.1}$$

$$K_t \mid \lambda \overset{\text{iid}}{\sim} \text{Poisson}_+(\lambda), \quad \lambda \sim \text{Gamma}(a_\lambda, b_\lambda). \tag{C.2}$$

The Gibbs updates are hence $\beta \mid B_{2:T} \sim \text{Beta}\,(\chi, \omega)$ and $\lambda \mid K_{2:T} \sim \text{Gamma}\,(\kappa, \tau)$, with

$$\chi := a_\beta + \sum_{t=2}^{T} B_t, \quad \omega := b_\beta + (T-1) - \sum_{t=2}^{T} B_t \tag{C.3}$$

$$\kappa := a_\lambda + \sum_{t=2}^{T} K_t, \quad \tau := b_\lambda + (T-1)\,. \tag{C.4}$$

Updating the latent variables $B_{2:T}$ and $K_{2:T}$. The sequential nature of the model allows each of the latent variables B_t and K_t to be updated individually. We use conditional independence to marginalize out sampling steps where possible, since such marginalization increases the sampling efficiency (Dyk and Park, 2008). For both $B_{2:T}$ and $K_{2:T}$, we marginalize twice: The first transforms the conditional distributions $P(B_t \mid \beta)$ and $P(K_t \mid \lambda)$ into the predictive distributions $P(B_t \mid B_{-t})$, and $P(K_t \mid K_{-t})$, which yields

$$B_t \mid B_{-t} \sim \text{Bernoulli}\Big(\frac{\chi_{-t}}{\chi_{-t} + \omega_{-t}}\Big) \quad \text{and} \quad K_t \mid K_{-t} \sim \text{NB}_+\Big(\kappa_{-t}, \frac{1}{1+\tau_{-t}}\Big)\,.$$

The second improvement is made by marginalizing the conditional dependence of B_t on K_t, and of K_t on B_t. That requires some notation: $\delta_t(V_t, U_t)$ indicates that a new edge is added to the graph between vertices V_t and U_t. When a new vertex is attached to V_t, we write $\delta_t(V_t, u^*)$. We abbreviate $\bar{\tau} := \frac{1}{1+\tau}$, and write $\widetilde{\mathcal{N}}(v)$ for the $\{0,1\}$-ball of vertex v, i.e. v and its neighbors. The updates for $B_{2:T}$ are

$$P(B_t = 1 \mid G_{(t-1):t}, K_{-t}, B_{-t}) \propto \frac{\delta_t(V_t, u^*)\mu_{t-1}(V_t)\chi_{-t}}{\chi_{-t} + \omega_{-t}}$$

$$P(B_t = 0 \mid G_{(t-1):t}, K_{-t}, B_{-t}) \propto \delta(V_t, U_t) + \frac{\delta(V_t, u^*)\mu_{t-1}(V_t)\omega_{-t}}{\chi_{-t} + \omega_{-t}} \sum_{u \in \widetilde{\mathcal{N}}(V_t)} \big[Q_{t-1}^{\kappa_{-t}, \bar{\tau}_{-t}}\big]_{V_t, u}$$

with $Q_{t-1}^{\kappa_{-t},\bar{\tau}_{-t}}$ as in (A.2) with $r = \kappa_{-t}$, $p = \bar{\tau}_{-t}$. The updates for $K_{2:T}$ are

$$
\begin{aligned}
P(K_t = k \mid G_{(t-1):t}, K_{-t}, B_{-t}) &\propto \frac{\Gamma(\kappa_{-t}+k)}{\Gamma(k+1)\Gamma(\kappa_{-t})}(1-\bar{\tau}_{-t})^{\kappa_{-t}}\bar{\tau}_{-t}^{k} \ \dots \\
\times \ \Big[\ \delta_t(V_t, u^*)\mu_{t-1}(V_t)\Big(\frac{\chi_{-t}}{\chi_{-t}+\omega_{-t}} &+ \frac{\omega_{-t}}{\chi_{-t}+\omega_{-t}} \sum_{u \in \widetilde{\mathcal{N}}(V_t)} \big[Q_{t-1}^{k+1}\big]_{V_t,u}\Big) \dots \\
+ \ \delta_t(V_t, U_t)\frac{\omega_{-t}}{\chi_{-t}+\omega_{-t}}&\Big(\mu_{t-1}(V_t)\big[Q_{t-1}^{k+1}\big]_{V_t,U_t} + \mu_{t-1}(U_t)\big[Q_{t-1}^{k+1}\big]_{U_t,V_t}\Big)\Big]
\end{aligned}
$$

with $Q_{t-1}^{k+1} = D_{t-1}^{-1/2}(\mathbb{I}_{t-1} - \mathbf{L}_{t-1})^{k+1}D_{t-1}^{1/2}$, the probability of a random walk of length $k+1$ from u to v. For implementation, the distribution for K_t must be truncated at some finite k, which can safely be done at three or four times the diameter of G_T: The total remaining probability mass can be calculated analytically, and the mass is placed on larger k is negligible.

Sampling $G_{2:(T-1)}$. Implementation of Algorithm 4.1 is straight-forward, with one exception: at each SMC step $G_{t-1} \to G_t$, we collapse the dependence on the particular values of B_t, K_t in that sampling iteration so that edges are proposed from the collapsed transition kernel $q_\varphi^t(G_t \mid G_{t-1}, B_{-t}, K_{-t})$. Thus, G_t is composed of G_{t-1} plus a random edge e_t sampled as

$$
\begin{aligned}
&P(e_t = (v,u)) \propto \\
&\quad \mathbb{1}\{(v,u) = (V_t, u^*)\}\mu_{t-1}(v)\Big(\tfrac{\chi_{-t}}{\chi_{-t}+\omega_{-t}} + \tfrac{\omega_{-t}}{\chi_{-t}+\omega_{-t}}\Big(\sum_{u \in \widetilde{\mathcal{N}}(v)} \Big[Q_{t-1}^{\kappa_{-t},\bar{\tau}_{-t}}\Big]_{v,u}\Big)\Big) + \dots \\
&\quad \mathbb{1}\{(v,u) = (V_t, U_t)\}\frac{\omega_{-t}}{\chi_{-t}+\omega_{-t}}\left(\mu_{t-1}(v)\left[Q_{t-1}^{\kappa_{-t},\bar{\tau}_{-t}}\right]_{v,u} + \mu_{t-1}(u)\left[Q_{t-1}^{\kappa_{-t},\bar{\tau}_{-t}}\right]_{u,v}\right) .
\end{aligned}
$$

Appendix D

Proofs for Chapter 5

D.1 Asymptotic degree distribution for $\mathbf{YS}(\beta, \alpha)$ models.

We derive the asymptotic distribution of degrees (equivalently, block sizes) used in Proposition 5.10 for the $\mathbf{YS}(\beta, \alpha)$ model. The setup is similar to that for the proof of Theorem 3.3 in Appendix A.3: We derive the asymptotic expected number of vertices of each degree, $m_{d,t}$, and then show that $m_{d,t}$ concentrates around its mean.

It is straightforward to show that the expectation satisfies the recurrence (with k_t the number of vertices at step t)

$$\mathbb{E}(m_{d,t+1}) = \beta \cdot \delta_1(d) + \mathbb{E}(m_{d,t})\Big(1 - \frac{(1-\beta)(d-\alpha)}{t-\alpha k_t}\Big) + \mathbb{E}(m_{d-1,t})\frac{(1-\beta)(d-1-\alpha)}{t-\alpha k_t} \ .$$

Now, $\frac{d-\alpha}{t-\alpha k_t} \to \frac{d-\alpha}{t(1-\alpha\beta)}$, so

$$\frac{\mathbb{E}[m_{1,t}]}{\beta t} \to p_1 := \frac{\frac{1-\beta\alpha}{1-\beta}}{1-\alpha+\frac{1-\beta\alpha}{1-\beta}}$$

and

$$\frac{\mathbb{E}[m_{d,t}]}{\beta t} \quad \rightarrow \quad p_d = p_{d-1}\frac{d-\beta-1}{d-\alpha+\frac{1-\beta\alpha}{1-\beta}} = p_1 \prod_{j=2}^{d} \frac{j-1-\alpha}{d-\alpha+\frac{1-\beta\alpha}{1-\beta}}$$
$$= \frac{1-\beta\alpha}{1-\beta}\frac{\Gamma(d-\alpha)\Gamma(1-\alpha+\frac{1-\beta\alpha}{1-\beta})}{\Gamma(1-\alpha)\Gamma(d+1-\alpha+\frac{1-\beta\alpha}{1-\beta})} ,$$

which is (5.63). Lemma A.3 can be applied here, as well, which establishes the concentration of $m_{d,t}/(\beta t)$ about its mean.