UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS
Y DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

ON NETWORK MODELS BASED ON RANDOM MEASURES

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS

PRESENTA:
Bernardo Flores López

DIRECTOR
Dr. Ramsés Humberto Mena Chávez
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

CIUDAD DE MÉXICO, FEBRERO 2021.

A mi abuela †

CONTENTS

INTRODUCTION

The rapid advances in computing have made the study of relational data an important tool to understand complex phenomena. Reconstructing a phylogenetic tree, modelling atomic fission, the spread of a disease, all can be fitted into a solid mathematical framework by representing them as networks and trees. This entails the challenge of finding suitable families of probability measures that can replicate the behaviors observed empirically.

Networks, understood as a collection of relations between entities, can be encompased within the rich theory of random graphs, from where the challenge of building models that can mimic the power-law behaviour most real world networks possess arises (Crane, 2018).

A common assumption among probabilistic models for networks is exchangeability over vertices (Orbanz and Roy, 2015), (Crane, 2018), appealing to Aldous-Hoover theorem (Aldous, 1985) to construct prior distributions over classes of random graphs. While theoretically appealing, this framework produces dense graphs with probability one (Orbanz and Roy, 2015), which limits their applicability.

In response to this limitation, over the past few years several families of models have been developed that move away from vertex exchangeability, instead relying on an exchangeable point process representation of adjacency matrices (Caron and Fox, 2017), (Veitch and Roy, 2015), (Borgs et al., 2016), or on switching the exchangeability assumption from vertices to edges (Crane and Dempsey, 2018), (Cai, Campbell, and Broderick, 2016a), for example. These models, under certain conditions, can produce sparse networks with power-law degree distributions.

In this thesis we will review some of the most used models for random graphs, focusing on how to generate sparsity and in the drawbacks the basic models have. We will also discuss some of the newer models that have appeared on the literature, finishing with an extension of the edge-exchangeable model to accommodate for temporal dynamics, based on (Palla, Caron, and Teh, 2016).

This work will be divided in three chapters. The first chapter will serve as an introduction, providing the necessary background on random measures, bayesian statistics and (random) graphs and partitions.

The second chapter will be dedicated to random graph models. The first half will cover some of the classic models, starting with the initial work by Erdős and Rényi and then following with generalizations of it and some other generating algorithms that sample more flexible graph structures. On the second half we will study the random graphs arising from the theory of graph limits and two of its extensions using random measures. We will discuss the general theory and provide its sparsity and power law properties.

Finally, on chapter 3 we will propose a model for dynamic networks in continuous time based on the edge-exchangeable framework. We will state some of its basic properties and perform a simulation study in order to analyze the sparsity of the graphs it generates.

1

PRELIMINARIES

1.1 RANDOM MEASURES

Random measures and point processes play a key role in several areas of probability and statistics, such as bayesian nonparametrics (Ghosal and Vaart, 2017), branching processes (Kallenberg, 2017) and stochastic geometry (Chiu et al., 2013). This section will develop the basic theory behind them and lay the ground for many of the topics covered in this work.

Let $\mathbb{X}$ be a Polish space, $\mathcal{X}$ its Borel σ-field and $\mathbb{M}_{\mathbb{X}}^{\mathfrak{b}}$ the set of locally bounded measures over $(\mathbb{X}, \mathcal{X})$, that is, the set of measures $\{\nu\}$ such that $\nu(A) < \infty$ for every bounded Borel set A. We will endow $\mathbb{M}_{\mathbb{X}}^{\mathfrak{b}}$ with the σ-field generated by the mappings $\pi_A : \mu \mapsto \mu(A)$ for every $A \in \mathcal{X}$, $\mathcal{M}$.

Definition 1.1. *A **random measure** is a measurable mapping from $(\mathbb{X}, \mathcal{X})$ to $(\mathcal{M}_{\mathbb{X}}^{\mathfrak{b}}, \mathcal{M})$. Equivalently, a random measure is a locally bounded kernel $\xi : \mathbb{X} \times \mathcal{X} \to [0, \infty]$, i.e., a function ξ such that $\xi(\cdot, A)$ is measurable for every $A \in \mathcal{X}$ and $\xi(\omega, \cdot)$ is a locally bounded measure for fixed $\omega \in \mathbb{X}$.*

Now we will define several functionals that will prove to be useful througout the rest of this section. The ω argument will be omitted. Let ξ be a random measure.

- The *intensity* or *mean measure*, $\mathbb{E}[\xi]$, is given by
$$\mathbb{E}[\xi(A)].$$
- For every bounded, positive and measurable function, f, the *Laplace functional* of ξ is defined as
$$\mathcal{L}_\xi(f) = \mathbb{E}\left[\exp\left\{-\int f \, d\xi\right\}\right].$$
- The *characteristic functional* of ξ is defined for every bounded, measurable function, f, as
$$\Phi_\xi(f) = \mathbb{E}\left[\exp\left\{i\int f \, d\xi\right\}\right].$$

Note that the mean measure is a measure and both functionals uniquely determine ξ (Kallenberg, 2017).

Definition 1.2. (Kingman, 1967). *A **completely random measure** (CRM) is a random measure, ξ, such that if $A_1, A_2, \ldots, A_n$ are disjoint members of $\mathcal{X}$, the random variables*
$$\xi(A_1), \xi(A_2), \ldots, \xi(A_n)$$
are independent.

Completely random measures provide a tractable way of constructing distributions over the space of probability measures (James, Lijoi, and Prünster, 2009), which makes them a central piece of many applications, particularly bayesian statistics. We will first focus on the main concept required to understand the theory: the Poisson point process.

Definition 1.3. *A **Poisson point process** (PPP) ξ, with intensity measure μ is a completely random measure such that for every μ-finite measurable set*[1] *A, $\xi(A) \sim Po(\mu(A))$.*

Poisson point processes belong to a more general type of random measures called *point processes*: integer-valued random measures. Every point process ξ has the following representation (Daley and Vere-Jones, 2007)

$$\xi = \sum_i \delta_{X_i},$$

where X_i are random elements of $\mathbb{X}$. If all X_i are distinct a.s., the process is called *simple*. This implies that $\xi(A) \leqslant |A|$ a.s. for every simple point process ξ and every measurable set A. For the purpose of this dissertation we will only work with simple Poisson point processes and for the sake of simplicity we will omit the adjective.

Not all measures can be the intensity measure of a simple Poisson point process. Indeed, suppose μ has an atom at x. Then

$$\mathbb{P}(\xi(\{x\}) > 1) = 1 - e^{-\mu(\{x\})} - \mu(\{x\})\, e^{\mu(\{x\})} > 0,$$

which contradicts the previous statement. Consequently, a mean measure must be non atomic.

Proposition 1.1. *The Laplace functional of a Poisson point process ξ, with mean measure ν is given by*

$$\mathbb{E}\left[e^{-f\xi}\right] = \exp\left\{\int_{\mathbb{X}} \left(1 - e^{-f}\right) d\nu\right\},$$

for all measurable $f : \mathbb{X} \to \mathbb{R}_+$.

Now consider the set functions

$$\lambda_\theta(A) = -\log \mathbb{E}\left[e^{-\theta\xi(A)}\right], \quad \theta > 0.$$

It can be proven (Kingman, 1967) that these functions are indeed measures and by uniqueness of the Laplace transforms, they determine the measures $\xi(A)$ for all measurable A.

If the λ_θ from a completely random measure ξ are σ-finite for all θ, we say ξ is a Σ-finite random measure. This property allows us to decompose the law of ξ in three components.

Theorem 1.1. *(Kingman, 1967). A Σ-finite completely random measure, ξ, can be decomposed as the sum of three independent components:*

$$\xi = \beta + \Psi + \Phi,$$

where

1 We consider the zero-measure case to be degenerate.

1. β*: a non atomic deterministic* σ*-finite Borel measure.*

2. $\Phi = \sum_{a_j \in \mathcal{A}} \Phi(\{a_j\})\delta_{a_j}$*, where* $\mathcal{A}$ *is countable, making* Φ *a random measure with fixed atoms.*

3. Ψ*: a completely random measure with law given by*

$$\mathbb{E}\left[e^{-t\Psi(A)}\right] = \exp\left\{-\int_{A\times\bar{\mathbb{R}}^+} \left(1 - e^{-ts}\right)\, \nu(dx, ds)\right\},$$

where ν *is a measure on* $\mathbb{X} \times (0, \infty]$ *such that* $\nu(\{x\} \times [0, \infty)) = 0$ *and* $\int_{\mathbb{X}} \int (s \wedge 1)\, \nu(dx, ds) < \infty$.

The ψ component is a purely atomic random measure whose atoms correspond to a Poisson process on $\mathbb{X} \times (0, \infty]$, and from now on we will only consider CRM with both $\beta, \Phi = 0$. Then, for every nonnegative measurable function $f : \mathbb{X} \to \mathbb{R}$, the Laplace functional of this type of CRM has the form

$$\exp\left\{-\int \left(1 - e^{-sf(x)}\right)\, \nu(dx, ds)\right\}$$

so that its law is fully characterized by the so called *intensity measure*, ν, and can be represented as the functional of a Poisson point process $\tilde{N}$ with intensity measure ν

$$\xi(B) = \int_{B\times\mathbb{R}^+} s\, \tilde{N}(\, dx,\, ds), \qquad \text{for all } B \in \mathcal{X}.$$

Intensity measures for this type of CRM can be classified in two cases (James, Lijoi, and Prünster, 2009). Let μ be a non-atomic and σ-finite measure on $\mathbb{X}$. We say ν is:

1. *Homogeneous* if for some measure ρ on $\mathbb{R}^+$,

$$\nu(dx, ds) = \rho(ds)\mu(dx).$$

2. *Non-homogeneous* if

$$\nu(dx, ds) = \rho(ds|x)\mu(dx),$$

where ρ is a conditional measure obtained via disintegration (Kallenberg, 2002).

A complete random measure ξ over the positive real line with the usual topology generates a stochastic process $\phi : \mathbb{R} \to \mathbb{R}$

$$\phi(t) = \begin{cases} \xi((0, t]) & \text{if } t \geqslant 0 \\ \xi([t, 0)) & \text{if } t < 0 \end{cases}$$

This construction ensures that ϕ is right-continuous and increasing, defining a random cumulative distribution function, which determines ξ uniquely. The independence of ξ on disjoint sets implies that for numbers $t_1 < t_2 < \cdots < t_n$, the increments

$$\phi(t_i) - \phi(t_{i-1}) = \xi((t_{i-1}, t_i])$$

are independent. This type of process, increasing and with independent increments, is called a *subordinator*.

Subordinators are a special case of *Lévy processes*, a stochastic process that starts at zero almost surely and that has stationary, independent increments. Lévy processes can be characterized in the following way:

Theorem 1.2. *(Lévy-Itô). Let X be a Lévy process in $\mathbb{R}$. Then for all $t \geqslant 0$ and $\eta \in \mathbb{R}$ there exists $a \in \mathbb{R}$, $\sigma \in \mathbb{R}$ and a σ-finite measure ν called the **Lévy measure** concentrated on $\mathbb{R} \setminus \{0\}$ satisfying $\int_{\mathbb{R}} (1 \wedge x^2)\, \nu(dx) < \infty$ such that the characteristic function of X has the form $\mathbb{E}\left[e^{i\eta X_t}\right] = e^{t\psi_\eta}$, where*

$$\psi_\eta = ai\eta + \frac{1}{2}\sigma^2\eta^2 + \int_{\mathbb{R}\setminus\{0\}} \left(e^{i\eta x} - 1 - i\eta x \mathbb{I}_{|x|<1}\right)\, \nu(dx). \qquad (1)$$

When X is a subordinator, ψ_η takes the form

$$\psi_\eta = ai\eta + \int_0^\infty \left(e^{i\eta x} - 1 - i\eta x \mathbb{I}_{0<x<1}\right)\, \nu(dx)$$

Note that (1) implies that the law of the process (X_t) is fully determined by the law of X_1, which in turn is characterized by the *characteristic triplet* (a, σ^2, ν).

When working with subordinators, instead of the characteristic function, it is more convenient to use the Laplace transform, which is given by

$$\mathbb{E}\left[e^{-\eta X_1}\right] = \exp\left\{-\left(a\eta + \int_0^\infty \left(1 - e^{-\eta x}\right)\, \nu(dx)\right)\right\}, \quad \eta \geqslant 0.$$

In terms of the increments if $s < t$,

$$\mathbb{E}\left[e^{-\eta(X_t - X_s)}\right] = \exp\left\{-(t-s)\left(a\eta + \int_0^\infty \left(1 - e^{-\eta x}\right)\, \nu(dx)\right)\right\}$$

Kingman's decomposition together with this last expression implies that the increments $X_t - X_s$ of a subordinator form a completely random measure with constant $\beta = a$ and intensity measure ν, which gives a bijection between subordinators and CRMs.

Example 1.1. (Gamma process). Let $(X_t)_{t\geqslant 0}$ be a Lévy process with characteristic triplet $(0, 0, \nu)$, where $\nu(ds) = \frac{e^{-s}}{s}\, ds$. Then the Laplace transform of the increments is a Frullani integral which takes the form

$$\begin{aligned}
\mathbb{E}\left[e^{-\eta(X_t - X_s)}\right] &= \exp\left\{-(t-s)\int_0^\infty \left(1 - e^{-\eta x}\right)\, \nu(dx)\right\} \\
&= \exp\left\{-(t-s)\int_0^\infty \left(1 - e^{-\eta x}\right)\frac{e^{-x}}{x}\, dx\right\} \\
&= \frac{1}{(1+\eta)^{t-s}}.
\end{aligned}$$

Hence $X_t - X_s \sim \text{Gamma}(t-s, 1)$. By an extension procedure this generates a homogeneus completely random measure ξ over $\mathbb{R} \times \mathbb{R}^+$ by $\xi(A) = \xi(A) \sim \text{Gamma}(\lambda(A), 1)$, where $\lambda(\cdot)$ is the Lebesgue measure over $\mathbb{R}$. This can be extended (James, Lijoi, and Prünster, 2009) to every Polish space $\mathbb{X}$ by replacing the Lebesgue measure by a non-atomic, σ-finite measure.

Example 1.2. (σ-stable process). Let $\sigma \in (0,1)$ and α a non-atomic measure over a Polish space $\mathbb{X}$, and consider a CRM ξ_σ with Lévy measure

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{\sigma}{\Gamma(1-\sigma)s^{1+\sigma}}\,\mathrm{d}s\,\alpha(\mathrm{d}x).$$

Then ξ_σ is called a σ-stable process with parameter measure α. Moreover, for any measurable function $f : \mathbb{X} \to \mathbb{R}^+$ such that $\int |f|^\sigma \,\mathrm{d}\alpha < \infty$, the Laplace functional is $\mathbb{E}\left[e^{-\int f \,\mathrm{d}\xi_\sigma}\right] = e^{-\int f^\alpha \,\mathrm{d}\alpha}$, which implies that the Laplace transform of $\xi_\alpha(A)$ has the form $\mathbb{E}\left[e^{-\lambda \xi_\sigma(A)}\right] = e^{-\lambda^\sigma \alpha(A)}$, corresponding to a positive stable distribution.

Random probability measures can be obtained by normalizing almost surely finite completely random measures. A CRM ξ with Lévy measure ν has finite and positive total mass $\xi(\mathbb{X})$ a.s. if, respectively, $\nu(\mathbb{X} \times \mathbb{R}^+) = \infty$ and

$$\int_{\mathbb{X}} \int_{\mathbb{R}^+} \left(1 - e^{-\lambda s}\right)\, \nu(\mathrm{d}x, \mathrm{d}s) < \infty$$

for all positive λ (James, Lijoi, and Prünster, 2009). In that case, we define a normalized random measure with independent increments (NRMI) $\tilde{p}$ by

$$\tilde{p}(\mathrm{d}x) = \frac{\xi(\mathrm{d}x)}{\xi(\mathbb{X})}.$$

It is worth noting that NRMIs select only discrete distributions almost surely (James, 2003), which implies that $\tilde{p}$ admits the following representation

$$\tilde{p} = \sum_{i \geqslant 1} \tilde{p}_i \delta_{X_i},$$

where X_i are random elements of $\mathbb{X}$ sampled from the atomless factor of its intensity measure and the $\tilde{p}_i$ sum up to one a.s.

Example 1.3. Dirichlet process. (Ferguson, 1973). Let α be a non-atomic, finite measure on a Polish space $\mathbb{X}$. A random measure P is called a Dirichlet process with base measure $P_0 = \alpha/\alpha(\mathbb{X})$ and precision parameter $M = \alpha(\mathbb{X})$, $P \sim \text{DP}(M, P_0)$, if for every finite measurable partition $B_1, \ldots, B_n$ the joint distribution of $(P(B_1), \ldots, P(B_n))$ is a k-dimensional Dirichlet distribution with parameters $\alpha(B_1), \ldots, \alpha(B_n)$.

Ever since its introduction by Ferguson, the Dirichlet process has been one of the fundamental models of Bayesian nonparametrics. There are several ways to prove its existence (Ferguson, 1973), but we will focus on normalizing a gamma process with intensity measure α.

It is well known (Ferguson, 1973) that if $Z_1, \ldots, Z_n$ are independent gamma random variables with $Z_j \sim \text{Gamma}(\alpha_j, 1)$, $j = 1, \ldots, n$, then

$$\left(\frac{Z_1}{\sum_1^n Z_j}, \ldots, \frac{Z_n}{\sum_1^n Z_j}\right) \sim \text{Dirichlet}\left(\alpha_1, \ldots, \alpha_n\right).$$

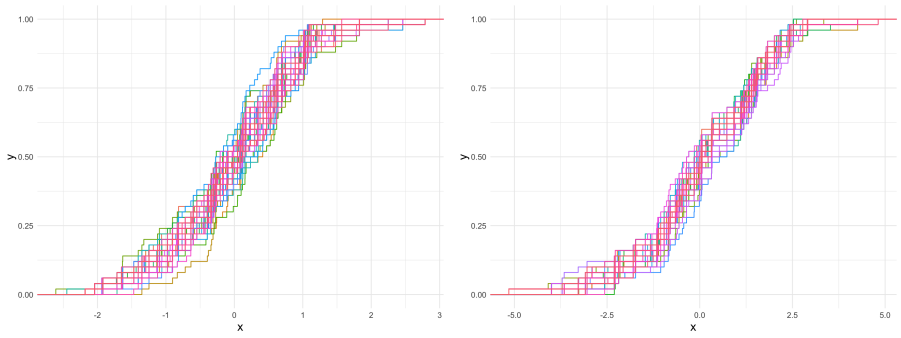

Figure 1: Two samples of size 100 of a Dirichlet process with base measure $N(0,1)$. Left, precision parameter 1; right, 0.5.

Now let us consider a measurable partition $B_1, \dots, B_n$ and a Φ gamma CRM with Lévy measure α. Then, by independence over disjoint sets,

$$\left(\frac{\Phi(B_1)}{\sum_1^n \Phi(B_j)}, \dots, \frac{\Phi(B_n)}{\sum_1^n \Phi(B_j)} \right) \sim \text{Dirichlet}\left(\alpha(B_1), \dots, \alpha(B_n)\right),$$

which corresponds to the definition of a Dirichlet process. Since the partition was arbitrary, we can conclude that a normalized gamma random measure coincides in law with a Dirichlet process.

The Dirichlet process possesses another representation particularly useful for simulation, the *stick-breaking* (Sethuraman, 1994).

Let $v_1, v_2 \dots \overset{\text{iid}}{\sim} \text{Beta}(1, M)$ and $\theta_1, \theta_2, \dots \overset{\text{iid}}{\sim} P_0$. Set

$$w_1 = v_1 \quad \text{and} \quad w_i = v_i \prod_{j=1}^{i-1} (1 - v_j), \quad i \geqslant 2.$$

Then

$$\sum_{i=1}^{\infty} w_i \delta_{\theta_i} \sim \text{DP}(\alpha, M).$$

This type of representation can be used to construct other random measures, such as the *two-parameter Poisson-Dirichlet process* or *Pitman-Yor process*.

Definition 1.4. *Let $\sigma \in (0,1)$ and $\theta > -\sigma$. Let $v_1, v_2, \dots$ be a sequence of independent random variables with $v_k \sim \text{Beta}(\theta + k\sigma, 1 - \sigma)$. Define the weights*

$$w_1 = v_1 \quad \textit{and} \quad w_i = v_i \prod_{j=1}^{i-1} (1 - v_j), \quad i \geqslant 2.$$

Then the random measure defined as

$$\sum_{j \geqslant 1} w_j \delta_{\theta_j}$$

is called a Pitman-Yor process with parameters (σ, θ).

1.2 BAYESIAN STATISTICS

The uncertainty present in complex systems can be splitted into two different kinds: epistemic and aleatory (Goldstein, n.d.). Aleatory uncertainty is that which relates to the intrinsic to the system and cannot be resolved except by direct observation, whereas epistemic uncertainty captures our lack of knowledge that can be reduced by gathering more information.

Statistical analysis represents both uncertainties using a probabilistic model, which can be regarded as a family of probability measures over the sample space[2], $\{\mathcal{P}_\theta\}$, indexed by a parameter $\theta \in \vartheta$, such that the data $X \sim \mathcal{P}_\theta$ for a given θ. Within this context, the distributions represent the aleatory uncertainty, whereas the parameter represents the epistemic one.

Bayesian statistics solves the epistemic uncertainty by placing a distribution over θ called the *prior distribution*, and conditioning it with respect to the data, obtaining the *posterior distribution*.

Formally, if the parameter space is equiped with a σ-field, $\mathcal{A}$, the prior distribution is the push-forward measure of the mesaurable function $\Theta : (\mathbb{X}, \mathcal{X}) \to (\vartheta, \mathcal{A})$, and the distribution $\mathcal{P}_\theta \mid \theta$, called the *likelihood*, is a regular conditional probability over the sample space, $(\mathbb{X}, \mathcal{X})$, then (X, θ) has a well defined joint distribution over the product space $(\mathbb{X} \times \Theta, \mathcal{X} \otimes \mathcal{A})$. This, in turn, lets us define the posterior distribution $\mathbb{P}(\theta \in A \mid X)$ for all $A \in \mathcal{A}$ as a regular conditional probability on the parameter space. Note that all the previous conditional measures are guaranteed to exist if both $\mathbb{X}$ and Θ are Polish.

When the $\mathcal{P}_\theta$ are all absolutely continuous with respect to a σ-finite measure ν on $(\mathbb{X}, \mathcal{X})$, the marginal distribution of X, μ_X, can be calculated using Fubini's theorem as

$$\begin{aligned}\mu_X(A) &= \int_\vartheta \int_A f_{X|\theta}(x \mid \theta)\, \nu(\mathrm{d}x)\, \mu_\theta(\mathrm{d}\theta) \\ &= \int_A \int_\vartheta f_{X|\theta}(x \mid \theta)\, \mu_\theta(\mathrm{d}\theta)\, \nu(\mathrm{d}x),\end{aligned}$$

where $f_{X|\theta}$ denotes the Radon-Nikodym derivative $\frac{\mathrm{d}\mathcal{P}_\theta}{\mathrm{d}\nu}$. It follows that μ_X is absolutely continuous with respect to ν with density

$$f_X(x) = \int_\vartheta f_{X|\theta}(x \mid \theta)\, \mu_\theta(\mathrm{d}\theta).$$

This measure is usually called the *prior predictive distribution of* X.

The same calculation can be repeated for the posterior distribution, obtaining the *posterior predictive distribution*, that is, the distribution of new data, X^*, given the previous observations, X,

$$f_{X^*|X}(x^* \mid x) = \int_\vartheta f_{X|\Theta}(x \mid \theta)\, \mu_{\Theta|X}(\mathrm{d}\theta).$$

2 For example, all normal distributions with unit variance.

In the case where the data points $x_1, \dots, x_n$ are conditionally iid given θ, both distributions take the form

$$f_X(x) = \int_\vartheta \prod_{i=1}^{n} f_{X|\Theta}(x_i \mid \theta)\, \mu_\theta(\mathrm{d}\theta),$$

$$f_{X^*|X}(x^* \mid x) = \int_\vartheta \prod_{i=1}^{m} f_{X|\Theta}(x_i^* \mid \theta)\, \mu_{\Theta|X}(\mathrm{d}\theta),$$

where $x_1^*, \dots, x_m^*$ are the new observations.

In most standard cases, the posterior distribution can be calculated using Bayes' theorem.

Theorem 1.3. *(Bayes' theorem) (Schervish, 1995). Suppose that the conditions above hold. Suppose that for all θ, $\mathcal{P}_\theta \ll \nu$ for some σ-finite measure on $(\mathbb{X}, \mathcal{X})$, and let $f_{X|\Theta}(x \mid \theta)$ be the conditional density with respect to ν of X given $\Theta = \theta$. Let μ_Θ be the prior distribution. Then $\mu_{\Theta|X} \ll \mu_\Theta$ a.s. with respect to the marginal distribution of X and the Radon-Nikodym derivative is*

$$\frac{\mathrm{d}\mu_{\Theta|X}}{\mathrm{d}\mu_\Theta}(\theta \mid x) = \frac{f_{X|\theta}(x \mid \theta)}{\int_\vartheta f_{X|\theta}(x \mid t)\ \mathrm{d}\mu_\Theta(t)}$$

A probabilistic model that fulfills the condition that for all θ, $\mathcal{P}_\theta \ll \nu$ for some σ-finite measure ν is said to be *dominated*. This is not a trivial assumption. In fact, most cases where ϑ is infinite-dimensional[3] are not dominated (Ghosal and Vaart, 2017). Particularly, NRMIs are not dominated.

Proposition 1.2. *Consider the random discrete distribution*

$$\Xi = \sum w_k \delta_{\theta_k},$$

where the weights w_k sum up to one and $\theta_1, \theta_2, \dots \overset{iid}{\sim} G$, with G an atomless probability measure on a Polish space $(\mathbb{X}, \mathcal{X})$. Then the familiy $\{\Xi_\theta \mid \theta\}$ is not dominated.

Proof. Suppose that we observe $\theta_1 = x_1$. Then Ξ would take values on $A = \{\text{Probability measures on } (\mathbb{X}, \mathcal{X}) \text{ with an atom at } x_1\}$., which entails that

$$\mathbb{P}(\Xi \in A \mid \theta_1 = x_1) = 1,$$

whereas, since G is atomless,

$$\mathbb{P}(\Xi \in A) = 0.$$

Hence the posterior is not dominated by the prior, implying that the prior cannot be dominated. □

Since NRMIs locations are sampled from an atomless distribution, this last proposition entails that they are not dominated either, so Bayes' theorem cannot be used to calculate their posteriors. However, as shown in (James, Lijoi, and Prünster, 2009), a closed-form formula can be found.

3 In this case we say the model is *nonparametric*, as opposed to a *parametric* model whose parameter space has a finite dimension.

Theorem 1.4. *(James, Lijoi, and Prünster, 2009). Let $\tilde{p}$ be a NRMI with intensity $\nu(dx, ds) = \rho(ds \mid x)H(dx)$ obtained by normalizing $\tilde{\mu}$. Then*

$$\tilde{\mu} \mid (\mathbf{X}, U_n) \sim \tilde{\mu}_{U_n} + \sum_{i=1}^{k} J_i^{(U_n)} \delta_{X_i^*},$$

where $\tilde{\mu}_{U_n}$ is a CRM with Lévy intensity $\nu^{(U_n)}(dx, ds) = e^{-U_n s}\rho(ds \mid x)\,\alpha(dx)$, the nonnegative jumps $J_n^{(U_n)}$ are mutually independent and independent from $\tilde{\mu}$ with (Lebesgue) density function $f_i(s) \propto s^{n_i} e^{-U_n s}\rho(ds \mid X_i^*)$. Moreover

$$\tilde{p} \mid (X, U_n) \sim w\frac{\tilde{\mu}_{U_n}}{\tilde{\mu}_{U_n}(\mathbb{X})} + (1-w)\frac{\sum_{i=1}^{k} J_i^{(U_n)}\delta_{X_i^*}}{\sum_{r=1}^{k} J_r^{(U_n)}},$$

where $w = \tilde{\mu}_{U_n}(\mathbb{X}) / \left[\tilde{\mu}_{U_n}(\mathbb{X}) + \sum_{i=1}^{k} J_i^{(U_n)}\right]$.

This means that the posterior distribution of a NRMI has fixed atoms at the observation points. In the case of a Dirichlet process $\tilde{P} \sim DP(M, P_0)$, this last formula reduces to

$$\tilde{P} \mid x_1, \ldots, x_n \sim DP\left(\frac{M}{M+n}P_0 + \frac{n}{M+n}\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, M+n\right).$$

Note that by taking M close to zero we can make the prior less informative. Conditioning and margianlizing $\tilde{P}$, the posterior predictive is

$$X_{n+1} \mid X_1 = x_1, \ldots, X_n = x_n \sim \frac{M}{M+n}P_0 + \frac{1}{M+n}\sum_{i=1}^{n}\delta_{x_i}.$$

Even though the Bayesian approach to statistical learning provides us with a solid mathematical framework to do inference, we have not yet justified why is considering the parameters random a plausible assumption. To do this we will start by defining a weaker dependence assumption than independence: exchangeability.

Definition 1.5. *A vector of random variables $(X_1, \ldots, X_n)$ is said to be* exchangeable *if*

$$(X_1, \ldots, X_n) \overset{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)})$$

for every permutation π of the indices.

A sequence $X_1, X_2 \ldots$ of random variables is exchangeable if its finite-dimensional distributions are exchangeable.

Exchangeability possesses a remarkable property discovered by Bruno de Finetti that links it with the usual iid assumption from frequentist statistics.

Theorem 1.5. *(de Finetti)*[4]*. Let $\mathbb{X}$ be a Polish space equipped with its Borel σ-field $\mathcal{X}$. Let $\mathcal{P}_{\mathbb{X}}$ be the space of all probability measures over $(\mathbb{X}, \mathcal{X})$. An*

4 The general version here presented was first proved by Hewitt and Savage in 1955(Orbanz and Roy, 2015).

infinite sequence of random elements of $\mathbb{X}$, $\{X_i\}_{i=1}^{\infty}$, *is exchangeable if and only if there exists a random probability measure* Q *over* $\mathcal{P}_{\mathbb{X}}$ *such that for all* n *and all measurable* $A_1, \ldots, A_n$,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} P(X_i \in A_i)\, Q(\mathrm{d}P). \tag{2}$$

Furthermore,

$$Q = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i} \text{ a.s.}$$

Proof. We will use a proof by Aldous (Aldous, 1985) to prove the equivalent statement: the sequence is conditionally independent given Q.

$$X_1, X_2, \ldots | Q = P \overset{\text{iid}}{\sim} P.$$

For the sake of simplicity we will restrict ourselves to the case $\mathbb{X} = \mathbb{R}$. Note that the reverse implication is immediate, i. e., a conditionally iid sequence is always exchangeable.

First assume $X_1, X_2, \ldots$ are exchangeable random variables and define the decreasing familiy of σ-fields

$$\mathscr{F}_n = \sigma(X_{n+1}, X_{n+2}, \ldots).$$

Note that $\mathscr{T} := \bigcap_{n=1}^{\infty} \mathscr{F}_n$ is the tail σ-algebra.

We will prove that the sequence $\{X_n\}_n$ is conditionally independent given $\mathscr{T}$. Fix $m \geqslant 1$. By exchangeability,

$$(X_m, X_{m+1}, X_{m+2}, \ldots) \overset{\mathrm{d}}{=} (X_m, X_{n+1}, X_{n+2}, \ldots),$$

and since $\mathscr{F}_n \subseteq \sigma(X_m, X_{n+1}, X_{n+2}, \ldots)$, by the Disintegration Theorem,

$$\mathbb{E}[\phi(X_m) \mid \mathscr{F}_{m+1}] \overset{\mathrm{d}}{=} \mathbb{E}[\phi(X_m) \mid \mathscr{F}_n]$$

for each bounded measurable $\phi : \mathbb{R} \to \mathbb{R}$. Also, as $\mathbb{E}[\phi(X_m) \mid \mathscr{F}_n] \to \mathbb{E}[\phi(X_m) \mid \mathscr{T}]$ a.s. by the Backwards Martingale Convergence Theorem, $\mathbb{E}[\phi(X_m) \mid \mathscr{F}_{m+1}] \overset{\mathrm{d}}{=} \mathbb{E}[\phi(X_m) \mid \mathscr{T}]$. The equality in law and the set inclusion $\mathscr{T} \subseteq \mathscr{F}_m$ imply that for all $A \in \mathscr{F}_{m+1}$

$$\int_A \mathbb{E}[\phi(X_m) \mid \mathscr{F}_{m+1}] \ \mathrm{d}\mathbb{P} = \int_A \mathbb{E}[\phi(X_m) \mid \mathscr{T}] \ \mathrm{d}\mathbb{P},$$

thus $\mathbb{E}[\phi(X_m) \mid \mathscr{F}_{m+1}] = \mathbb{E}[\phi(X_m) \mid \mathscr{T}] = \mathbb{E}[\phi(X_m) \mid \mathscr{F}_{m+1}, \mathscr{T}]$ a.s., which in turn means that X_m and $\mathscr{F}_{m+1}$ are conditionally independent given $\mathscr{T}$. Since m is arbitrary, $X_1, X_2, X_3, \ldots$ are conditionally independent given $\mathscr{T}$.

Furthermore, exchangeability also implies that

$$(X_1, X_{n+1}, X_{n+2}, \ldots) \overset{\mathrm{d}}{=} (X_n, X_{n+1}, X_{n+2}, \ldots)$$

for all n, hence by the same argument as before

$$\mathbb{E}[X_n \mid \mathscr{F}_n] = \mathbb{E}[X_1 \mid \mathscr{F}_n] \text{ a.s.}$$

Thus conditioning on $\mathscr{T}$ and by the tower property, $\mathbb{E}[X_n \mid \mathscr{T}] = \mathbb{E}[X_1 \mid \mathscr{F}_n \mid \mathscr{T}]$ a.s. and therefore the sequence $X_1, X_2, \dots$ is iid given $\mathscr{T}$.

Now take the empirical distribution

$$\Xi_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

and denote $\Xi_\infty := \lim_{n\to\infty} \Xi_n$. $\Xi_\infty(\cdot, A)$ is $\mathscr{T}$-measurable for all measurable A, since for all k,

$$\Xi_\infty = \underbrace{\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{k} \delta_{X_i}}_{0} + \lim_{n\to\infty} \frac{1}{n}\sum_{i=k+1}^{n} \delta_{X_i} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=k+1}^{n} \delta_{X_i},$$

so Ξ_∞ is $\mathscr{F}_k$-measurable for all k, and hence for the tail σ-field as well. From this we can conclude that the sequence $X_1, X_2, \dots$ is iid conditioned to the random measure Ξ_∞.

The convergence statement then follows from the Strong Law of Large Numbers, since conditionally to Ξ_∞, the Ξ_n converge almost surely to the (conditional) distribution of $X_1, X_2, \dots$, which by unicity is a.s. equal to Ξ_∞. In consequence, by the Disintegration and Bounded Convergence theorems, for all bounded and measurable f

$$\begin{aligned}
\mathbb{E}[f(\Xi_\infty, \Xi_n)] &= \mathbb{E}\left[\int f(\Xi_\infty, s)\mathbb{P}(\Xi_n \in ds \mid \Xi_\infty)\right] \\
&\overset{n\to\infty}{\longrightarrow} \mathbb{E}\left[\int f(\Xi_\infty, s)\mathbb{P}(\Xi_\infty \in ds \mid \Xi_\infty)\right] \\
&= \mathbb{E}[f(\Xi_\infty, \Xi_\infty)] \quad \Xi_\infty - \text{a.s.}
\end{aligned}$$

Hence

$$\frac{1}{n}(\delta_{X_1} + \delta_{X_2} + \cdots + \delta_{X_n}) \overset{n\to\infty}{\Longrightarrow} \Xi_\infty \quad \Xi_\infty - \text{ a.s.}$$

□

De Finetti's representation then guaranties the existence of an almost sure unique prior distribution under which the data becomes independent, and gives a way to approximate it. Moreover, it justifies using random parameters instead of fixed ones.

1.3 RANDOM GRAPHS AND NETWORKS

Random graphs are among the most used models for studying relational data (Crane, 2018). Yet, no general framework of how to model real-world networks has ben found. This section will lay out the main properties and definitions of network theory, giving a particular emphasis on how can random graphs be used as prior distributions over structured data.

Let us first recall the basic definitions of graph theory. The first half of this chapter will be based on (Hofstad, 2016). A graph G is a pair

(E, V), where E is a countable set of edges and V is the set of vertices. If G is *directed*, the edges are indicated as ordered pairs (u, v), whereas if it is *undirected*, as sets $\{u, v\}$. From this point forward, unless stated otherwise, we will only consider finite graphs and the number of edges will be written as $|G|$.

A graph can be represented using its *adjacency matrix*, a matrix A whose A_{ij} entry equals 1 if there is an edge that goes from vertex i to vertex j, and equals zero otherwise. Note that a graph is undirected if and only if its adjacency matrix is symmetric.

Now let us consider connectivity. The *degree* of a node v, $\deg(v)$ is the number of edges containing v. If the graph is directed, the *out-degree* and the *in-degree* are defined as the number of outgoing and ingoing links, respectively. The *degree distribution* is the discrete probability mass function $k \mapsto \mathbb{P}(D_n = k)$, where D_n is de degree of a randomnly chosen vertex, and represents the empirical distribution of the degrees in the graph.

The degree distribution encodes information about the *density* of the graph, e. g., how close the number of edges is to the maximum possible number of connections, which in case of an undirected graph of size n it is equal to $\binom{n}{2}$. Then, a *dense graph* (V, E) has number of edges $\Omega(|V|^2)$, in contrast of a *sparse graph*, with number of edges $O(|V|)$ (Orbanz and Roy, 2015). This can be summarized with the *density number* $|E|/|V|^2$, a graph being sparse if it has density number of order $1/|V|$.

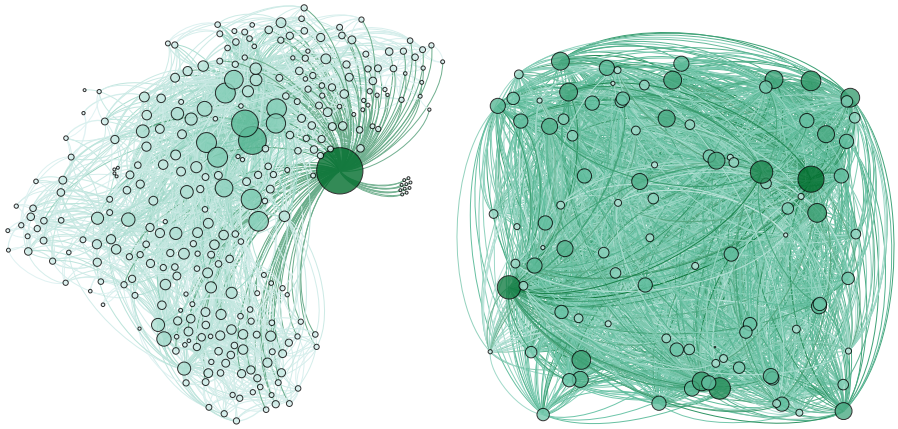

Figure 2: Left: the neural network of a *Caenorhabditis elegans* nematode with density number 4.9%. Right: a randomly generated network with density number 57.3%. Color darkness and size of nodes are directly proportional to their degree. Data obtained from https://snap.stanford.edu/data/C-elegans-frontal.html.

Many real-world networks are observed sequentially, growing in size as we observe them. This makes it reasonable to consider *graph sequences* $\{G_n\}_n$, where n denotes the number of vertices, the size of the graph.

Denote the rate of nodes with degree k in G_n by $P_k^{(n)}$,

$$P_k^{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\left\{d_i^{(n)} = k\right\}},$$

where d_i^n denotes the degree of the ith vertex in G_n.

Using $P_k^{(n)}$, the graph G_n being sparse can be equivalently formulated as it having

$$\lim_{n\to\infty} P_k^{(n)} = p_k, \quad k \geqslant 0,$$

for some deterministic probability distribution $\{p_k\}_k$. This alternative definition allows us to define a behaviour often encountered in applications: power-law degree distributions. When considering *random graphs*, that is, graphs whose adjacency matrix is random, the convergence must be restated as convergence in probability (or equivalently, in distribution).

Definition 1.6. *We call a graph sequence* $\{G_n\}_{n\geqslant 1}$ scale-free *with exponent* τ *when it is sparse and*

$$\lim_{k\to\infty} \frac{\log[1-F(k)]}{\log(1/k)} = \tau - 1,$$

where $F(k) = \sum_{l\leqslant k} p_l$ *denotes the cumulative distribution function corresponding to* $\{p_k\}$ *defined as in the paragraphs above.*

This last expression is equivalent to

$$p_k \propto k^{-\tau},$$

so that the limiting law follows a *power law*, a heavy-tailed probability distribution, which implies the existence of highly connected nodes.

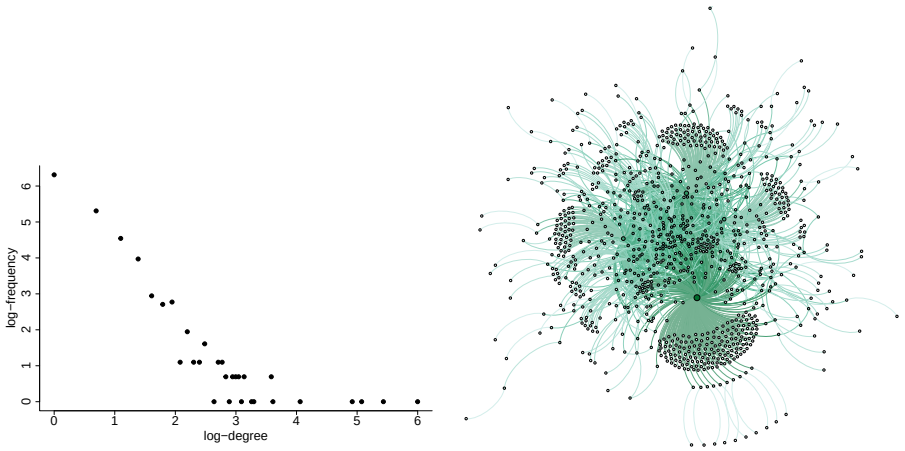

Figure 3: The image on the left depicts the degree distribution on a log-log scale of the graph on the right generated according to a power law. As expected, the degree distribution in this scale shows an almost linear behaviour.

Often, graphs exhibit a clustering behaviour in which a large fraction of the vertices lies in a single connected component. For a graph $G_n = (V, E)$ let $\mathscr{G}(v) = \{u \in [n] : \mathrm{dist}_G(u, v) < \infty\}$ for every node v, where $\mathrm{dist}_G(u, v)$ is the minimal number of edges in a path linking u and v, equalling ∞ when there is no such path. Let $\mathscr{C}_{\max}$ be the largest connected component, satisfying $|\mathscr{C}_{\max}| = \max_{v\in V} |\mathscr{C}(v)|$. Note that $\mathscr{C}_{\max}$ is not necessarily unique. In case it is not, we define $\mathscr{C}_{\max}$ as any of the maximums with equal probability.

Definition 1.7. *A graph sequence* $\{G_n\}_{n\geqslant 1}$ *is called* ***highly connected*** *when*

$$\liminf_{n\to\infty} |\mathscr{C}_{max}|/n > 0,$$

in which case $\mathscr{C}_{max}$ *is called the* ***giant component****. Furthermore, for a highly connected graph sequence, its giant component is called unique when*

$$\liminf_{n\to\infty} \left|\mathscr{C}_{(2)}\right|/n = 0,$$

where $\mathscr{C}_{(2)}$ *is the second largest cluster. This intuitively means that*

The giant component of a graph can be thought of as beign a large, densely connected blob of vertices surrounded by a cloud of mostly unconnected nodes. The uniqueness condition above described intuitively means that after the largest cluster, all others sharply decrease in size. These connected components are often refered to as *communities*.

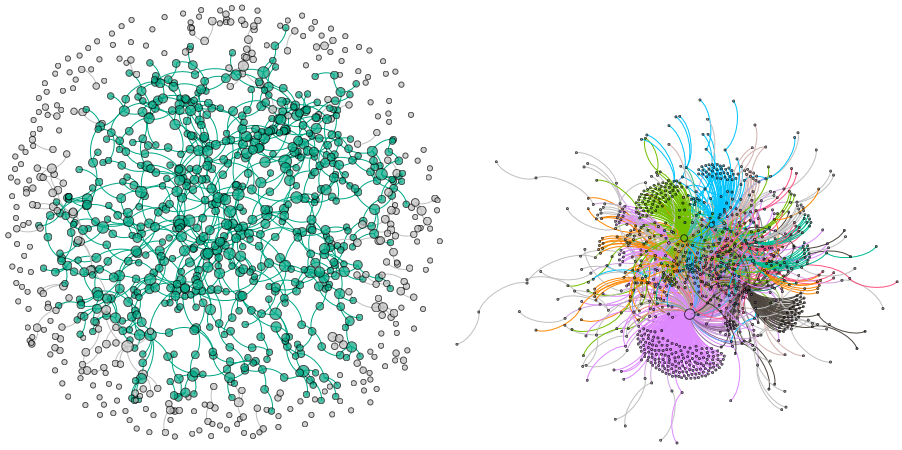

Figure 4: On the left, a random graph with a giant component colored in green. On the right, a sparse graph whose nodes are colored according to the community they belong to.

Another connectivity behaviour often found empirically is when most nodes are not neighbours of one another and yet, most are connected by a small number of steps. This can be formalized using the *typical distance*, defined as $H_n = \mathrm{dist}_G(U_1, U_2)$, where U_1 and U_2 are two vertices sampled uniformly from a graph G of size n. A graph with this characteristic is called *small-world*.

Definition 1.8. *We say that a graph sequence* $\{G_n\}_{n\geqslant 1}$ *is a* ***small world*** *when there exists a constant* $K < \infty$ *such that*

$$\lim_{n\to\infty} \mathbb{P}\left(H_n \leqslant K \log n\right) = 1.$$

Further, we say that $\{G_n\}n \geqslant 1$ *is an* ***ultra-small world*** *when, for every* $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(H_n \leqslant \varepsilon \log n\right) = 1.$$

This definition can be informally stated as the typical distance begin proportional to $\log n$.

Now let us focus on a modelling perspective, based on (Orbanz and Roy, 2015). When representing relational data as edges in a graph, it is

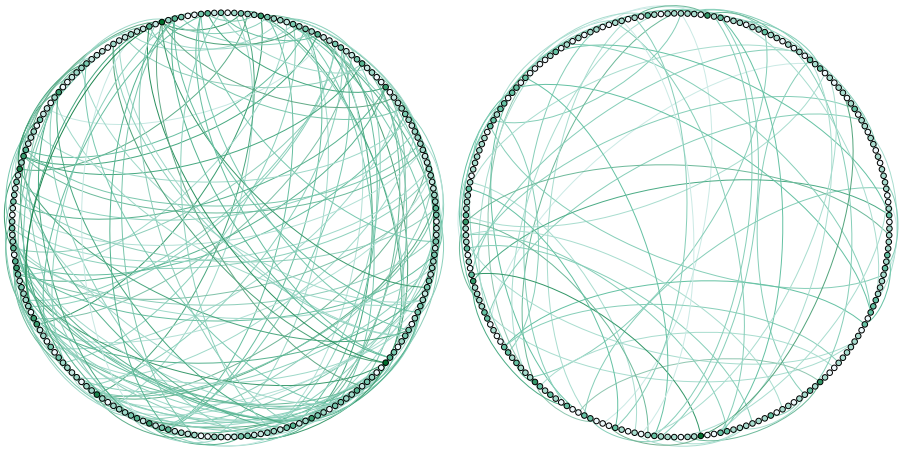

Figure 5: Comparison between a non-small-world (left) and a small-world graph (right). Both have roughly the same number of nodes. Note the low typical distance between nodes on the right plot.

usually assumed that as more data is observed, the graph grows. This violates the usual frequentist statistics assumptions, where multiple independent graphs are expected. A way to circumvent this is by assuming some form of symmetry, dependence that allows statistical learning.

The Bayesian approach for sequences of random variables using de Finetti's theorem poses a similar situation. Herein it is assumed that data points are sampled from an infinite, exchangeable random vector, and then made iid by conditioning on a random measure constructed as the limit of the empirical distribution. This idea can be generalized to random graphs by considering *exchangeable random graphs*, i. e., random graphs law-invariant to permutation of its vertices.

Definition 1.9. *A $\mathbb{X}$-valued* d-array *is a collection $x_{k_1}, x_{k_2}, \ldots, x_{k_d}$ of elements of $\mathbb{X}$ indexed by $k_1, \ldots, k_d \in \mathbb{N}$.*

A d-array can be thought of as a d-dimensional sequence. Let us consider 2-arrays, i. e., arrays of the form $\{x_{ij}\}_{1 \leqslant i,j < \infty}$.

Definition 1.10. *Let $\{X_{ij}\}$, $1 \leqslant i, j < \infty$, be random 2-arrays. They are* ***separately exchangeable*** *if*

$$(X_{ij})_{1 \leqslant i,j \leqslant n} \overset{d}{=} (X_{\sigma(i)\tau(j)})_{1 \leqslant i,j \leqslant n}, \tag{3}$$

for all n and all permutations σ, τ of $\mathbb{N}$, and where $\overset{d}{=}$ denotes equality in distribution. They are ***jointly exchangeable*** *if (3) holds in the special case $\tau = \sigma$.*

Exchangeable arrays have a de Finetti-type representation reformulated as random functions instead of random measures.

Theorem 1.6. (Aldous-Hoover). *An $\mathbb{X}$-valued random 2-array (X_{ij}) is jointly exchangeable if and only if there is a random function $F : [0,1]^3 \to \mathbb{X}$ such that*

$$(X_{ij}) \overset{d}{=} \left(F\left(U_i, U_j, U_{\{i,j\}}\right)\right), \tag{4}$$

where $\{U_i\}_{i \in \mathbb{N}}$, $\{U_{\{i,j\}}\}$ are iid Uniform$[0,1]$ random variables. Note that the $\{i,j\}$ subindex is ordered.

This theorem can be specialized to the case of separate exchangeability.

Corollary 1.1. (Aldous). *An $\mathbb{X}$-valued random 2-array (X_{ij}) is separately exchangeable if and only if there is a random function $F : [0,1]^3 \to \mathbb{X}$ such that*

$$(X_{ij}) \stackrel{d}{=} \left(F\left(U_i, U_j, U_{ij}\right)\right),$$

where $\{U_i\}_{i\in\mathbb{N}}$, $\{U_j\}_{j\in\mathbb{N}}$ and $\{U_{i,j}\}$ are iid Uniform$[0,1]$ random variables. In this case, the U_{ij} are symmetric in the sense that $U_{ij} = U_{ji}$.

At a first glance, this does not seems related to de Finetti's theorem. However, (2) can be restated as follows(Orbanz and Roy, 2015).

Corollary 1.2. *A sequence $X_1, X_2, \ldots$ of random elements of a Polish space $\mathbb{X}$ is exchangeable if and only if there exists a random function $F : [0,1] \to \mathbb{X}$ such that if $U_1, U_2, \ldots$ is an iid sequence of Uniform$[0,1]$ random variables,*

$$(X_1, X_2, \ldots) \stackrel{d}{=} (F(U_1), F(U_2), \ldots).$$

In the case where $\mathbb{X} = [a, b]$, this reformulation is easily justified by the fact that any random variable X with cumulative distribution function F has the same law as $F^-(U)$, where $U \sim \text{Uniform}[0,1]$ and F^- is a right-continuous pseudo-inverse of F. The de Finetti measure can then be replaced by its random CDF, confirming the corollary above.

Clearly, a random graph is exchangeable if and only if its adjacency matrix is jointly exchangeable. Hence (4) can be used to build prior measures over graphs. First we must refine some technical details.

If a random 2-array X is binary, is undirected and has a zero diagonal[5] is jointly exchangeable, the random function in (4) takes the form

$$F\left(U_i, U_j, U_{\{i,j\}}\right) = \begin{cases} 1 & \text{if } U_{\{i,j\}} < W(U_i, U_j) \\ 0 & \text{otherwise} \end{cases},$$

for a random function $W : [0,1]^2 \to [0,1]$. If we consider a 2-array as an infinite dimensional adjacency matrix, each $U_{\{i,j\}}$ is then associated to and edge, and each U_i to a vertex. This last representation is equivalent to

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

$$X_{\{i,j\}} \sim \text{Bernoulli}\left(W\left(U_i, U_j\right)\right)$$

Thus Aldous-Hoover theorem implies that an infinite exchangeable simple random graph can be obtained by sampling from a Bernoulli distribution parametrized by a random function W. This random function is called a *graphon*, and is unique up to measure-preserving bijections. Conversely, all exchangeable graphs with n nodes define an *empirical graphon* obtained by dividing $[0,1]^2$ into a $n \times n$ grid and assigning each square i, j the value zero if there is no edge between node i and node j, and 1 otherwise, obtaining a checkerboard-like function.

5 A graph with this property is called *simple*.

With this scheme, inference for a single observed graph instead of many can be solved in the same way as in the one-dimensional case by assumming that the collected data represents a finite-dimensional restriction of a graph with infinite vertices. Nevertheless, a careful analysis must be done to work around the lack of unicity of the graphon by considering equivalent graphons such that the generating arrays are equal up to reordering of rows and columns (Kallenberg, 2002).

Graphons can be further studied as *graph limits*. Roughly speaking, a sequence of simple graphs $\{G_n\}$ converges if for every fixed graph F, the proportion of copies of F in G_n converges. The formal definition requires a lengthy exposition outside the bounds of this dissertation.

If a sequence of graphs converges, the object it converges to is a graphon. Furthermore, $G_n \to W$ if and only if $W_n \Rightarrow W$ weakly, where W_n are the empirical graphons.

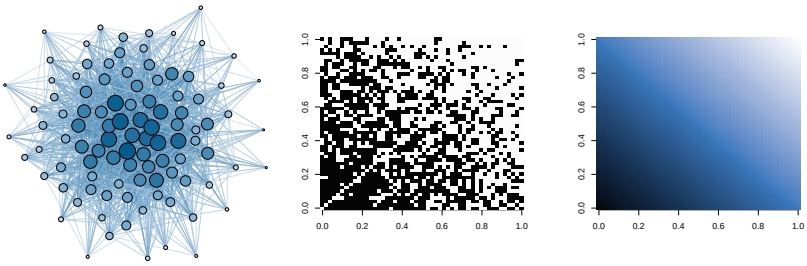

Figure 6: The two rightmost images depict the empirical graphon and the graphon the random graph on the left was sampled from. For the graphon, darker shades represent larger values, and for the empirical graphon, black represents 1 and white 0.

Graph convergence can be metrizised using the **cut distance**.

Definition 1.11. *Let w be a graphon. The **cut norm** is defined as*

$$\|w\|_\square = \sup_{S,T\subseteq[0,1]} \int_{S\times T} w(x,y)\,dx\,dy.$$

*with its induced **cut metric** denoted by $d_\square$.*

Now let us define the equivalence relation $\sim$ such that for two graphons w_1, w_2, $w_1 \sim w_2$ if they both generate the same random graphs. Then the cut metric induces a pseudometric by

$$\delta_\square(w_1,w_2) := \inf_{w'\in[w_2]} d_\square(w_1,w'),$$

where $[w_2]$ is the equivalence class of w_2. We can use this *cut pseudometric* to redefine graph convergence: a graph sequence $\{G_n\}$ converges if $\delta_\square(w_{G_n},w) \to 0$ for some measurable function $w : [0,1]^2 \to [0,1]$, where w_{G_n} denotes the empirical graphons.

If W is the space of all graphons, we can also use $\sim$ to construct the quotient space $W\backslash\sim$, making $(W\backslash\sim, \delta_\square)$ a compact metric space, from which an integral version of Aldous-Hoover theorem can be derived: G is an exchangeable 2-array if and only if

$$\mathbb{P}(G \in A) = \int_{W\backslash\sim} P_{\hat{w}}(A)\,\nu(d\hat{w}),$$

with $P_{\hat{w}}$ is the distribution over exchangeable arrays induced by a class representative of $\hat{w}$ sampled from a unique probability measure ν over $W\backslash\sim$. Note that using the quotient space allows us to make the representation unique.

Exchangeable random graphs have an unfortunate restriction: they can only be dense.

Proposition 1.3. *An exchangeable random graph is either dense or empty with probability one.*

Proof. Let W be a random graphon and let G_n be an undirected random graph with n vertices sampled from G_n using Aldous-Hoover. Then the expected proportion of edges in G_n is given by

$$\frac{1}{2}\int_{[0,1]^2} W(x,y)\,dx\,du,$$

where the $1/2$ was introduced because the adjacency matrix is symmetric. If this last integral equals zero, then since W is non-negative G_n will be empty almost surely. At the same time, if the integral is a positive constant p, then the expected number of edges is $p\binom{n}{2} = \Theta(n^2)$, so by the Strong Law of Large Numbers, the number of edges is $\Theta(n^2)$ with probability one. □

This last proposition entails a major disfunction when working with exchangeable random graphs, since most real-world graphs are sparse, not dense (Crane, 2018).

Another possible source of model misspecification is sampling. Consider an algorithm that extracts a sample graph from a data base. Can this graph be representative of the whole data? The answer to this question relies on the algorithm used. When coupled with a probabilistic model, that is, an infinite random graph from where we assume the observations were sampled from, this question translates as whether the sampling method is *projective* and *consistent* (Crane, 2018), (Orbanz, 2017).

Take an injection $\psi : [n] \to [N]$, $1 \leqslant n \leqslant N$ and a binary matrix y^N of dimensions $N \times N$. A *sampling operator* is a map

$$S^{\psi}_{N\to n}\left(y^N\right) = \left(y^N_{\psi(i)\psi(j)}\right)_{1\leqslant i,j\leqslant n}$$

i. e., a restriction to n vertices of the original structure.

A *sampling scheme* is defined (Orbanz, 2017) as a random sampling operator, making ψ aleatory.

Definition 1.12. *A sampling scheme is **projective** if for all* n

$$S^{\psi}_{N\to n}(y^N) \preccurlyeq S^{\psi}_{N\to n+1}(y^N),$$

where $x \preccurlyeq y$ *if* x *is a subgraph of* y.

Definition 1.13. *A sampling scheme is **consistent** if given an infinite graph* y *there exists a random variable* $S_\infty(y)$,

$$S^{\psi}_{n\to k}(y) \overset{n\to\infty}{\Longrightarrow} S_\infty(y)\big|_k,$$

where $\Rightarrow$ *denotes the weak convergence and the restriction operator* $\cdot\big|_k$ *extracts a subgraph composed of the first* k *vertices of its input (Orbanz, 2017).*

These two properties do not state anything regarding the appropiateness of the sampling scheme to preserve the characteristics of the graph it samples from. Akin to regular experimental design, the choice of the sampling algorithm depends on the particular characteristics of the graph one wishes to maintain.

Now we will introduce some common sampling schemes for undirected graphs as presented in (Kolaczyk, 2009), noting that modifying them to accomodate directed graphs is relatively straightforward. Let $G = (V, E)$ be an undirected graph with adjacency matrix (y_{ij}).

- Induced and Incident Subgraph Sampling. In this method we select a random sample of vertices and edges, respectively, and then recover the corresponding edge or vertex pair.

  Induced sampling arises naturally in social network settings, where people are first chosen and then asked about their relationships with each other. At the same time, an example where incindent sampling can be used is in the reconstruction of telephone call graphs or other types of telecommunication, where the calls are first observed and then paired to their respective senders and receivers.

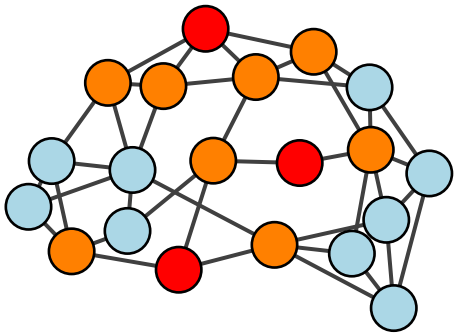

Figure 7: Example of induced sampling. Starting nodes are depicted in red, with their adjacing nodes in orange.

- Snowball sampling. Start with an initial vertex v^* and a fixed radius $r = 1, 2, \ldots$. Then we observe the neighbours of order r of v^* by doing

$$N(v^*, 1) = \{v \in V : y_{v^*v} = 1\}$$

$$N(v^*, 2) = \bigcup_{\{v \in V : N(v^*,1)\}} \{v' \in V : y_{vv'} = 1\} \setminus N(v^*, 1)$$

$$\vdots$$

$$N(v^*, k) = \bigcup_{\{v \in V : N(v^*, k-1)\}} \{v' \in V : y_{vv'} = 1\} \setminus \bigcup_{j=1}^{k-1} N(v^*, j)$$

$$\vdots$$

This sampling method can be exemplified by the behaviour some spiders used by web search engines exhibit, where a website is visited, the hyperlinks it contains are visited, then all the hyperlinks these websites contain are visited too, and the process is repeated up to a fixed number of steps.

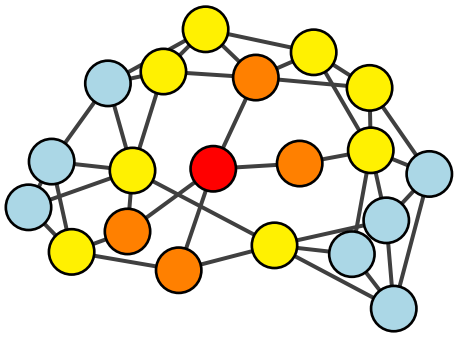

Figure 8: Example of snowball sampling of degree 2. The starting node is colored in red, the first order neighbours in orange and the second order neighbours in yellow.

- Path tracing. Here, we start with a set of vertices $\{v_i\}_i$ and then we find paths between all possible pairs, that is, a connected subgraph of G containing $\{v_i\}_i$. This type of sampling is used during contact tracing in epidemiology, wherein the possible contagion routes are traced between sick individuals.

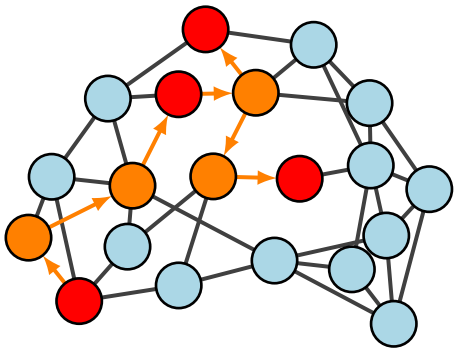

Figure 9: Example of path tracing. Starting nodes are depicted in red and their corresponding paths in yellow

1.4 RANDOM PARTITIONS

This section will be based on (Bertoin, 2006). Let us begin with a few definitions.

Definition 1.14. *A **partition** of $B \subseteq \mathbb{N}$ is a countable collection $\pi = \{\pi_i\}_{i\in\mathbb{N}}$ of pairwise disjoint subsets of B, called blocks, such that $\bigcup_{i\in\mathbb{N}} \pi_i = B$, which are always enumerated in increasing order of their least element, that is*

$$\min \pi_i \leqslant \min \pi_j,$$

for all $i \leqslant j$ and with the convention that $\min \emptyset = \infty$.

1. *We write $\mathcal{P}_B$ for the set of partitions of B. In the special case when $B = [k] = \{1, \dots, k\}$ for some $k \in \mathbb{N} \cup \{\infty\}$, we simply write $\mathcal{P}_k := \mathcal{P}_{[k]}$; in particular $\mathcal{P}_\infty := \mathcal{P}_{\mathbb{N}}$.*

2. *We denote by*

$$\#\pi := \#\{i \in \mathbb{N} : \pi_i \neq \emptyset\} = \max\{i \in \mathbb{N} : \pi_i \neq \emptyset\}$$

the cardinal of the set of non-empty blocks of a partition π.

If $B' \subset B$ is a subset of a block B and $\pi \in \mathcal{P}_B$ a partition of B, we write $\pi|_{B'}$ for the restriction of π to B', that is the partition of B' $\pi|_{B'} = \{\pi_i \cap B'\}_{i\in\mathbb{N}}$. This restriction yields the notion of compatibility.

Definition 1.15. *A sequence $\pi^{(1)}$, $\pi^{(2)}$,... of partitions of [1], [2], ... is called **compatible** if for all integers $k' \leqslant k$, $\pi^{(k')} = \pi^{(k)}|_{[k']}$.*

Clearly, if $\pi \in \mathcal{P}_\infty$ is a partition of $\mathbb{N}$, then the sequence of restrictions $\left\{\pi|_{[n]}\right\}_{n\in\mathbb{N}}$ is compatible. The converse is also true, i. e.given a compatible sequence of partitions $\left\{\pi^{(n)} : \pi^{(n)} \in \mathcal{P}_n\right\}_{n\in\mathbb{N}}$ there exists $\pi \in \mathcal{P}_\infty$ such that $\pi|_{[n]} = \pi^{(n)}$ for all n. Indeed, if we take

$$\pi_i = \bigcup_{n\in\mathbb{N}} \pi_i(n), \quad i \in \mathbb{N},$$

then $\{\pi_i\}_{i\in\mathbb{N}}$ is a partition π of $\mathbb{N}$, and $\pi|_{[n]} = \pi^{(n)}$ for all n.

A topology over $\mathcal{P}_\infty$ can be induced from endowing it with the following metric.

Proposition 1.4. *The space $\mathcal{P}_\infty$ is endowed with the ultrametric*[6]

$$d(\pi, \pi') = \max\left\{k \in \mathbb{N} : \pi|_{[k]} = \pi'|_{[k]}\right\}^{-1}, \quad \pi, \pi' \in \mathcal{P}_\infty,$$

with the convention that $1/\max\mathbb{N} = 0$. Then $(\mathcal{P}_\infty, d)$ is compact.

Now we need to introduce the concept of mass partition.

Definition 1.16. *A **mass partition** is an infinite numerical sequence $s = (s_1, s_2, \dots)$ such that $s_1 \geqslant s_2 \geqslant \cdots \geqslant 0$ and $\sum_{i=1}^{\infty} s_i \leqslant 1$. The terms s_i are called the **fragments** of s, and the space of mass partitions is denoted by $\mathcal{P}_m$.*

6 An ultrametric is a metric such that the triangle inequality is strengthened to $d(x, y) \leqslant \max\{d(x, z), d(z, y)\}$.

We can also define $s_0 := 1 - \sum_{i=1}^{\infty} s_i$, the mass of the amount of $[0,1]$ not contained in s. A mass partition is called *proper* if $s_0 = 0$ and improper otherwise.

Akin to $\mathcal{P}_\infty$, $\mathcal{P}_m$ can also be made to have "nice" topological properties by endowing it with the proper metric.

Proposition 1.5. *The space $\mathcal{P}_m$ endowed with the uniform metric*

$$d(s, s') = \max_{i \in \mathbb{N}} \{|s_i - s'_i|\}, \quad s, s' \in \mathcal{P}_m$$

is compact, and the induced topology coincides with that of pointwise convergence.

A map between $\mathcal{P}_\infty$ and $\mathcal{P}_m$ can be given by considering the notion of asymptotic frequency of blocks.

Definition 1.17. 1. *We say that a block B possesses an **asymptotic frequency** if and only if the limit*

$$|B| := \lim_{n\to\infty} \frac{1}{n} \#(B \cap [n]) = \lim_{n\to\infty} \frac{1}{n} \#\{k \in B : k \leqslant n\}$$

exists.

2. *If each block of some partition π has an assymptotic frequency, then we say π possesses asymptotic frequencies. We then write $|\pi| = (|\pi_1|, \dots)$ and then $|\pi|^{\downarrow} = \left(|\pi|_1^{\downarrow}, \dots\right)$ for the mass partition given by the decreasing rearrangement of the sequence $|\pi|$.*

3. *We say that a partition π has **proper** asymptotic frequencies if π possesses asymptotic frequencies with*

$$\sum_{i=1}^{\infty} |\pi_i| = 1.$$

Note that the last sum always exists by the fact that it is always bounded by 1, fact given by Fatou's lemma.

When some block of a partition does not have an asymptotic frequency we can add the extra point $|\pi| = |\pi|^{\downarrow} = \partial$ to $\mathcal{P}_m$, which allows us to define the measurable (but not continuous) map $\pi \mapsto |\pi|^{\downarrow}$ from $\mathcal{P}_\infty \to \mathcal{P}_m \cup \{\partial\}$. The lack of continuity can be easily seen from the fact that given $\pi \in \mathcal{P}_\infty$ and $s \in \mathcal{P}_m$, by compactness one can always find two sequences $\{\pi^{(n)}\}_n$ and $\{s^{(n)}\}_n$ of partitions and mass partitions, respectively, such that $\pi^{(n)} \to \pi$ and $s^{(n)} \to s$, and then glue them together by taking

$$\tilde{\pi}^{(n)} = \pi^{(n)}|_{[n]} \cup \pi^{s^{(n)}}|_{\mathbb{N}\setminus[n]},$$

where $\pi^{s^{(n)}}$ is any partition whose asymptotic frequencies equal $s^{(n)}$[7]. By this construction, $\left\{\tilde{\pi}^{(n)}\right\}_n$ converges to π and $\left\{\left|\tilde{\pi}^{(n)}\right\}_n\right|^{\downarrow}$ converges to s.

7 For example, a partition sampled from a paint-box, defined further ahead.

Partitions can also be identified through an equivalence relation $\sim_\pi$ such that $i \sim_\pi j$ if and only if i and j belong to the same block. This characterization allows us to define the group action of $\mathrm{Sym}([n])$ on $\mathcal{P}_n$ by taking a permutation $\sigma \in \mathrm{Sym}([n])$, a partition $\pi \in \mathcal{P}_n$ and defining the equivalence relation $\sim_{\sigma(\pi)}$

$$i \sim_\pi j \Leftrightarrow \sigma(i) \sim_{\sigma(\pi)} \sigma(j),$$

which constitutes a partition itself. This leads to the concept of *exchangeable random partitions*.

Definition 1.18. *Let $n \in \mathbb{N} \cup \{\infty\}$. A random partition π of $[n]$ is called exchangeable if for every permutation σ of $[n]$, $\sigma(\pi)$ and π have the same law.*

Recalling that the asymptotic frequency of a block B is defined as

$$|B| = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{i \in B\},$$

then for an exchangeable random partition Π, $\{\mathbb{I}\{i \in B\}\}$ is an exchangeable sequence and hence by de Finetti's theorem, $|B|$ exists almost surely. As consequence, the law of π is fully characterized by its asymptotic frequencies, that is,

$$\mathbb{P}(\Pi = \pi) = \mathbb{P}\left(|\Pi|^{\downarrow} = |\pi|^{\downarrow}\right).$$

Mass partitions and partitions can be connected by a *paint-box distribution*. Take a mass partition $s = (s_1, s_2, \dots)$ and consider an interval representation, that is a collection of disjoint subintevals of $(0,1)$ $\vartheta = \{I\}_j$ whose Lebesgue measures correspond to the elements of s, for instance, $\{[s_{i-1}, s_{i-1} + s_i]\}_i$. Let $U_1, U_2 \overset{\text{iid}}{\sim} \mathcal{U}[0,1]$, and consider the random partition π of $\mathbb{N}$ generated by assigning $n \in \mathbb{N}$ to block b_j if $U_n \in I_j$, and asigning it to its own singleton block if $U_j \in [0,1] \setminus \vartheta$. This is equivalent to the following equivalence relation:

$$i \sim_\pi j \Leftrightarrow (i = j) \text{ or } (U_i \text{ and } U_j \text{ belong to the same block of } \vartheta)$$

The name paint-box comes from thinking the mass partition divides the unit interval into different colors, randomly coloring with them every natural number.

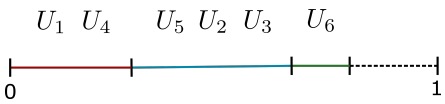

Figure 10: Representation of a paint-box sampling procedure. In this case, the assigned blocks are $\{\{1,4\},\{2,3,5\},\{6\}\}$.

Now take a permutation σ and a paint-box partition $\sim_\pi$. Then

$$i \sim_{\sigma(\pi)} j \Leftrightarrow (i = j) \text{ or } (U_{\sigma(i)} \text{ and } U_j \text{ belong to the same block of } \vartheta).$$

Since the U_i are iid, $\sigma(\pi)$ and π are equal in distribution, which implies that paint-box partitions are exchangeable.

Another important property of paint-boxes is the fact that they are invariant to the exact interval representation of the underlying mass partition used. As the elements of every interval representation always have the same Lebesgue measure by definition, the probability that U_i and U_j belong to the same block does not change when the representation does, thus the law of the paint-box remains unchanged.

The partitions sampled from a paint-box share the following elementary properties.

Proposition 1.6. *Let π be a paint-box based on a mass-partition $s \in \mathcal{P}_\infty$. Then the following assertions hold:*

1. *The paint-box posseses asymptotic frequencies equaling the mass-partition, i.e., $|\pi|^\downarrow = s$.*
2. *If $|\pi_i| = 0$, then π_i is either a singleton or is empty almost surely, and singletons occur if and only if s is improper, the set of singletons π_0 having an assymptotic frequency given by $1 - \sum_{i=1}^\infty s_i$.*

Proof. Each random block of a paint-box has the form $B = \{n \in \mathbb{N} : U_n \in A\}$, where A is a measurable interval member of an interval representation of the underlying mass-partition. Since the U_i are iid, by the Strong Law of Large Numbers

$$\frac{B \cap [n]}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(U_i \in A) \overset{n\to\infty}{\longrightarrow} |A| \text{ a.s.}$$

Also, $|\pi_0| = 1 - \sum_{i=1}^\infty s_i$.

The remaining claims follow from the last argument and by construction of the paint-box. □

Exchangeable random partitions can be represented in a similar way as graphs and sequences.

Theorem 1.7. *(Kingman) Let π be an exchangeable random partition of $\mathbb{N}$. Then the law of π can be expressed as a mixture of paint-boxes*

$$\mathbb{P}(\pi \in \cdot) = \int_{\mathcal{P}_\infty} \mathbb{P}\left(|\pi|^\downarrow \in \mathrm{d}s\right) \rho_s(\cdot),$$

where ρ_s is the law of the paint-box based on s.

Proof. This proof is due to Aldous (Aldous, 1985). We will prove the equivalent statement that there exists a probability measure μ on $\mathcal{P}_\infty$ such that conditioning on it, $\pi \sim \mu$.

Let $b : \mathbb{N} \to \mathbb{N}$ be a function that maps all elements of a block of π onto the same member of said block, for instance, the minimum. Also let $U_1, U_2, \ldots \overset{\text{iid}}{\sim} U[0,1]$ independent of π and b.

Take $\xi_i = U_{b(i)}$ and σ a permutation of $\mathbb{N}$. Then $i \sim_\pi j$ if and only if $\xi_i = \xi_j$, and $\xi_{\sigma(i)} = U_{b(\sigma(i))}$.

Since the U_i are iid and independent of π and b, $(\{U_i\}_i, \pi) \overset{\mathrm{d}}{=} \left(\{U_{\sigma(i)}\}_i, \sigma(\pi)\right)$, which makes $\{\xi_i\}_i$ an exchangeable sequence, and so by de Finetti's theorem there exists a random probability measure

μ such that conditioning on it, the $\xi_1, \dots$ are iid with law μ. Define the quantile function

$$q(s) = \inf\{x \in (0,1) : \mu([0,x]) \geqslant s\}$$

and introduce the set of flat points of q

$$\vartheta = \{x \in (0,1) : \exists \varepsilon > 0 \text{ such that } q(x) = q(y) \text{ whenever } |x-y| < \varepsilon\},$$

so we can define an interval representation of ϑ by making the lengths of the intervals components coincide with the masses of the atoms of ϑ.

Now let $V_1, \dots \overset{\text{iid}}{\sim} U[0,1]$ independent of π, U_i and b, Then conditionally on μ,

$$(q(V_1), q(V_2), \dots) \overset{d}{=} (\xi_1, \xi_2, \dots),$$

thus $i \sim_\pi j$ if and only if $q(V_i) = q(V_j)$, which only happens if V_i and V_j fall in the same interval component of ϑ, which is a paint-box construction. □

Kingman's representation allows us to circumvent the lack of continuity of the map $\pi \to |\pi|^\downarrow$ by considering convergence in distribution, making this map a bijective correspondence continuous in law called *Kingman's correspondence*.

Theorem 1.8. *Let $\left\{\pi^{(n)}\right\}_n$ and $\pi^{(\infty)}$ be random exchangeable partitions. The following conditions are equivalent:*

- *When $n \to \infty$, $\left|\pi^{(n)}\right|^\downarrow$ converges in distribution on $\mathcal{P}_m$ to $\left|\pi^{(\infty)}\right|^\downarrow$.*
- *When $n \to \infty$, $\pi^{(n)}$ converges in distribution on $\mathcal{P}_\infty$ to $\pi^{(\infty)}$.*

The law of an exchangeable partition π can be specified by working with its restrictions. Consider a partition $\varphi = (B_1, \dots, B_k, \emptyset, \dots)$ of $[n]$ with $B_k \neq \emptyset$. The exchangeability of π implies that

$$\mathbb{P}\left(\pi|_{[n]} = \phi\right) = p\left(\#B_1, \dots, \#B_k\right),$$

where p is a symmetric function over a finite (but not fixed) amount of positive integers called the *exchangeable partition probability function* (EPPF) of π. When $\pi \in \mathcal{P}_\infty$, the compatibility condition translates to the *addition rule*:

$$p(n_1, \dots, n_k) = p(n_1, \dots, n_k, 1) + \sum_{j=1}^{k} p(n_1, \dots, n_{j-1}, n_j + 1, n_{j+1}, \dots, n_k).$$

Proposition 1.7. *Let $s \in \mathcal{P}_m$ be a proper mass partition such that all the terms of s are distinct, and ρ_s the law of s. Let also $x_1, \dots, x_k$ be k terms sampled without replacement from s, and $(B_1, \dots, B_k, \emptyset, \dots)$ be some partition of $[n]$ with $B_k \neq \emptyset$. Then we have*

$$\rho_s\left(\pi|_{[n]} = (B_1, \dots, B_k, \emptyset, \dots), |\pi_1| = k_1, \dots, |\pi_k| = x_k\right) = x_1^{\#B_1} \cdots x_k^{\#B_k},$$

and as consequence,

$$\rho_s\left(\pi|_{[n]} = (B_1, \dots, B_k, \emptyset, \dots) \,|\, |\pi_1| = k_1, \dots, |\pi_k| = x_k\right) = \left(\prod_{i=1}^{k} x_k^{\#B_k - 1}\right)\left(\prod_{j=1}^{k-1}\left(1 - (x_1 + \cdots + x_j)\right)\right)$$

A particularly important family of random partitions are the Poisson-Dirichlet partitions. The most general one is the two-parameter Poisson Dirichlet distribution $\mathrm{PD}(\sigma, \theta)$ defined at the end of the first section of this chapter. When $\sigma \in (0, 1)$ and $\theta = 0$, one obtains the partition generated by the atoms of a normalized stable subordinator, and when $\theta > 0$ and $\sigma = 0$, the $\mathrm{PD}(0, \sigma)$ coincides with the partition generated by the atoms of a Dirichlet process.

Now consider $p_{\sigma,\theta}$ the EPPF of a $\mathrm{PD}(\sigma, \theta)$ partition and pick integers $n_1, \dots, n_k$ with $k \leqslant n$ and $n_1 + \cdots + n_k = n$. Then $p_{\sigma,\theta}$ has the following explicit formulas:

- For $\sigma = 0$ we have the *Ewens sampling formula*:

$$p_{\sigma,\theta}(n_1, \dots, n_k) = \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{i=1}^{k} (n_i - 1)!$$

- For $\sigma > 0$ we have the *Pitman sampling formula*:

$$p_{\sigma,\theta}(n_1, \dots, n_k) = \frac{(\theta/\sigma)_{k\uparrow}}{(\theta)_{n\uparrow}} \prod_{i=1}^{k} -(-\sigma)_{n_i\uparrow},$$

  where for any integer $l \geqslant 1$ and real number a,

$$(a)_{0\uparrow} = 1 \text{ and } (a)_{l\uparrow} = a(a+1)\cdots(a+l-1)$$

Two-parameter Poisson Dirichlet partitions can be recursively generated using a procedure called a *Chinese restaurant process*, described as follows:

Algorithm 1 Chinese restaurant process (σ, θ)

1: Assign $1 \to \{1\}$.
2: **for** $i = 2, 3, \dots$ **do**
3: Make i join an existing block B_k with probability $\frac{\#B_k - \sigma}{\theta + i - 1}$.
4: Assign $i \to \{i\}$ with probability $\frac{\theta + K\sigma}{\theta + i - 1}$, with K the number of currently existing blocks.
5: **end for**

The name "Chinese restaurant" comes from imagining a seating costumers at tables with infinite capacity in a Chinese restaurant, so that the first client arrives and sits at the first table, the second one then comes and either sits at the same table as costumer one or on its own table, and in general, every client must choose between sitting at an occupied table with probability directly proportional to the number of people seated at it or at an unoccupied table.

1.5 DISTRIBUTIONAL SYMMETRIES

This last section is based on (Orbanz and Roy, 2015).

Representation theorems like de Finetti's, Aldous-Hoover and Kingman's allow us to build Bayesian models over different types of structured data by parametrizing the law of exchangeable structures, giving a way to specify a prior over the parameter space and, sometimes, providing convergence results that add some form of consistency to the model.

The general scheme for these three theorems can be written as follows:

- Consider $\mathbb{X}_\infty$ a space of infinite structures.
- Identify a parameter space $\mathbb{O}$ for exchangeable random structures X_∞ of $\mathbb{X}_\infty$.
- Find a family of distributions on $\mathbb{X}_\infty$ called the *ergodic measures* such that each element $\theta \in \mathbb{O}$ determines an ergodic distribution p_θ.
- If ν is a probability measure over $\mathbb{O}$, the law of any exchangeable random element $X_\infty \in \mathbb{X}_\infty$ can be represented as a mixture of the ergodic distributions

$$\mathbb{P}(X_\infty \in \cdot) = \int_{\mathbb{O}} p_\theta(\cdot)\, \nu(\mathrm{d}\theta) \tag{5}$$

This last integral representation can be rewritten as a sampling scheme that defines a Bayesian model over $\mathbb{X}_\infty$:

$$\Theta \sim \nu$$

$$X_\infty \mid \Theta \sim p_\Theta,$$

and when coupled with a law of large numbers, the last scheme allows us to do statistical inference by approximating the infinite structure through a finite sample.

This framework can be further generalized to any probabilistic symmetry, not just exchangeability. Consider a group $\mathbb{G}$ of transformations of $\mathbb{X}_\infty$. We call a random structure X $\mathbb{G}$-*invariant* if $g(X) \overset{\mathrm{d}}{=} X$ for all $g \in \mathbb{G}$. For example, in the case of Aldous-Hoover theorem, the transformations correspond to the subgroup generated by row and column permutations.

Under some technical conditions, particularly, that $\mathbb{G}$ is locally compact and second-countable, the law of any $\mathbb{G}$-invariant random structure can be decomposed as in (5). This, however does not mean that a statistical model can be built with it, we first need a convergence result that justifies inference from a finite sample.

Definition 1.19. *A countable group $\mathbb{G}$ is **amenable** if there is a sequence $A_1, A_2, \ldots$ of finite subsets of $\mathbb{G}$ so that for some $c > 0$ and all $g \in \mathbb{G}$,*

$$\frac{|gA_k \cap A_k|}{|A_k|} \xrightarrow{k\to\infty} 1$$

If this last condition holds, we get the following result (Orbanz, 2017):

Theorem 1.9. *Let X be a random element of a Polish space $\mathbb{X}$ and $f, \{f_i\}_{i\in\mathbb{N}}$ function from $\mathbb{X} \to \mathbb{R}$ in $L_1(\mathbb{X})$, such that $f_i \to f$ X-almost-surely. Let T be a measurable action of a countable, amenable group $\mathbb{G}$, and F be the empirical measure defined by the sequence $\{A_k\}$ obtained by amenability*

$$(x, k) \mapsto F_k^x(\cdot) = \frac{1}{|A_k|} \sum_{g\in A_k} \delta_{T_g(x)}(\cdot),$$

Assume that (5) exists for $\mathbb{G}$, and let ξ be the random ergodic measure for X as in (5), then

$$F_k^X(f_k) \overset{k\to\infty}{\longrightarrow} \nu(f) \text{ a.s.}$$

Moreover, if the f_k are all dominated by an integrable function g, then the convergence is also in $L_1(\mathbb{X})$.

When taking $\mathbb{G}$ the finitary symmetric group[8], this theorem specializes for general exchangeable random structures, particularly exchangeable sequences and partitions.

8 The subgroup FSymm of the symmetric group such that every $g \in$ FSymm, $\text{Supp}(g) = \{\omega : \omega g \neq \omega\}$ is finite.

2

GENERATIVE MODELS FOR RANDOM GRAPHS

This chapter will be divided in two sections. First, based on (Hofstad, 2016), we will present some of the basic random graph models, covering Erdős-Rényi, preferential attachment, stochastic block models and the configuration model. On the second part we will review various of the more sophisticated models that attempt to overcome the drawbacks of the ones presented in the first section.

2.1 CLASSICAL MODELS

2.1.1 *Erdős-Rényi*

Generate a random graph of n vertices by sampling independently each possible edge with probability p. This is called the Erdős-Rényi graph, $\mathrm{ER}_p(n)$, and it is considered the keystone of the field (Hofstad, 2016).

When p is constant, the probability of a graph with m edges equals $p^m(1-p)^{n-m}$, so we can get an uniform distribution over the set of graphs with n vertices by taking an Erdős-Rényi graph with *edge probability* $p = 0.5$. In the following we will be concerned about the case $p = \lambda/n$, $\lambda > 0$.

Theorem 2.1. *Let $\mathbb{P}_\lambda$ denote the law of an $ER_{\lambda/n}(n)$ random graph. Then*

- *(*Subcritical regime*). If $\lambda < 1$, then*

$$\frac{|\mathscr{C}_{\max}|}{\log n} \xrightarrow{\mathbb{P}_\lambda} \frac{1}{\lambda - 1 - \log \lambda},$$

  where $\xrightarrow{\mathbb{P}_\lambda}$ denotes convergence in probability and $|\mathscr{C}_{\max}|$ is the size of the largest connected component.

  In other words,

$$|\mathscr{C}_{\max}| = O(\log n) \text{ as } n \to \infty \text{ with high probability (WHP).}$$

- *(*Supercritical regime*). If $\lambda > 1$, then the graph will have a giant component of size $O(n)$ WHP, and the distance between any two vertices in the giant component will be $O(\log n)$ asymptotically WHP.*

- *(*Critical regime*). If $\lambda = 1$, the largest component will have size $O(n^{2/3})$ WHP.*

This last theorem shows that in the subcritical regime the largest component size is logarithmically small; whereas in the supercritical regime its size is larger than $\log n$.

The critical case is far more complex than the other two. An unicylic graph is a graph with k vertices and k edges containing exactly one

cycle. When $\lambda \neq 1$, if $A < \infty$ is a constant, the expected number of unicyclic components with less than $A \log n$ vertices is of order $\lambda(A \log n)^2$, hence they generally are rare. Conversely, in the critical case when $\lambda = 1$, the expected number without size restrictions is approximately $\frac{1}{6} \log n$.

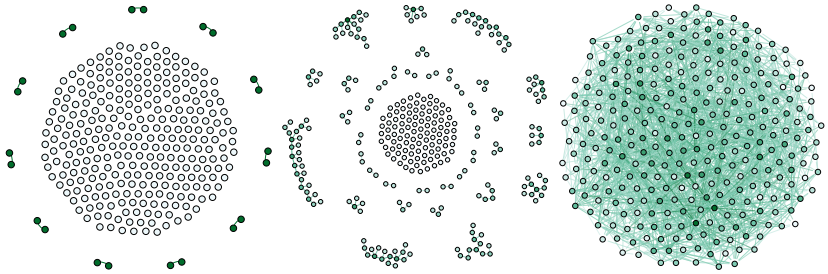

Figure 11: Three $\mathrm{ER}_{\lambda/300}(300)$ random graphs with $\lambda = 0.1, 1$ and 10, respectively. Note the different behaviors according to their criticalities: the subcritical case shows very few connected vertices, the critical one exhibits many unicyclic components and the supercitical graph has a single giant component with many edges.

Erdős-Rényi random graphs have some characteristics that make them unable to replicate real-world behaviors, the most prominent ones being independence and their degree distribution.

Since every edge is Bernoulli-distributed, the degree distribution for each vertex v equals

$$\mathbb{P}\left(\deg(v) = k\right) = \binom{n-1}{k} p^k (1-p)^{n-1-k},$$

so if $p = \lambda/n$, $\lambda > 0$, the degree distribution converges in law to a Poisson(λ) random variable, making highly connected nodes very unlikely. When paired with independence between edges, this translates to low clustering, a property often observed in real-world networks. Furthermore, since for large k the Poisson probability mass function is much smaller than $k^{-\tau}$ for any $\tau > 0$, the Erdős-Rényi random graph is not scale-free.

This last convergence result can be strengthened to the proportion of edges with a given degree k converging in probability to a Poisson probability mass function evaluated at k, impliying that $\mathrm{ER}_{\lambda/n}(n)$ is a sparse random graph.

Another related definition for an Erdős-Rényi is when the number of edges is fixed. Here, we begin with n vertices and sample uniformly without replacement m edges between them. If we write $\mathbb{P}_\lambda$ for the law of $\mathrm{ER}_{\lambda/n}(n)$ and $\mathbb{P}_m$ for the one of this last model, then the following relation allows us to translate results for $\mathrm{ER}_{\lambda/n}(n)$ to this alternative model:

$$\mathbb{P}_\lambda(\cdot) = \sum_{M=1}^{n(n-1)/2} \mathbb{P}_m(\cdot)\mathbb{P}\left(\mathrm{Binom}(n(n-1)/2, \lambda/n) = M\right).$$

2.1.2 *Stochastic block models*

The Erdős-Rényi random graph can be generalized to produce graphs containing communities by first labelling each node according to the community it belongs to and then connecting it to the rest of the nodes with different probabilities depending on their respective memberships. This can be formalized as follows.

Definition 2.1. *Let* $n \in \mathbb{N}$ *be the number of vertices,* k *the number of communities,* p *the* k*-dimensional vector of probabilities of belonging to each community and* W *a symmetric* $\kappa \times k$ *matrix whose entries are the connectivity probabilities. The pair* (X, G) *is a* ***stochastic block model*** *with parameters* n*,* p *and* W $SBM(n, p, W)$ *if* X *is an* n*-dimensional random vector with iid components distributed according to* p *and* G *is a simple graph with* n *vertices where vertices* i, j *are connected with probability* W_{X_i, X_j} *independently of other pairs of vertices. We also define the community sets by* $\Omega_i = \{v \in [n] : X_v = i\}$ *for all* $i \in [k]$*.*

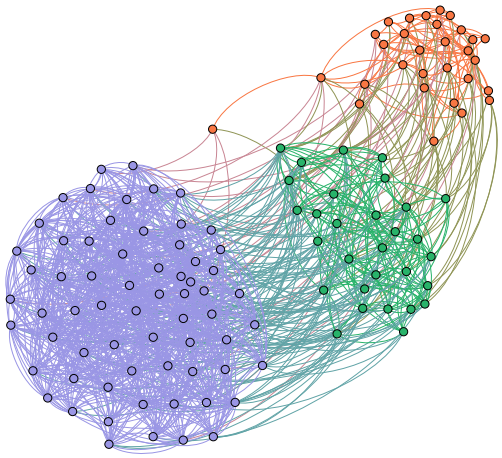

Figure 12: A graph sampled from a stochastic block model with three communities, each represented by a different color.

Note that when all the entries of W are the same the Erdős-Rényi is recovered.

When p is uniform and W takes the same values on the diagonal and outside the diagonal, i. e.,

$$W = \begin{pmatrix} a & b & b & \cdots & b \\ b & a & b & \cdots & b \\ b & b & a & \cdots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b & b & b & \cdots & a \end{pmatrix},$$

then one obtains a *symmetric stochastic block model* with parameters n, k, a and b, denoted $\mathrm{SBM}(n, k, a, b)$.

An important task of this type of model is how precisely we can recover the labels X by observing G.

Definition 2.2. *The **agreement** between two community vectors $x, y \in [k]^n$ is obtained by maximizing the common components between x and any rellabeling of y, i. e.*

$$A(x, y) = \max_{\pi \in Sym([k])} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i = \pi(y_i)),$$

where $Sym([k])$ is the group of permutations of $[k]$.

Definition 2.3. *Let $(X, G) \sim SBM(n, p, W)$. If there exists an algorithm that inputs G and outputs $\hat{X}$, then one can have*

- *Exact recovery:* $\mathbb{P}\left(A(X, \hat{X}) = 1\right) = 1 - o(1)$.
- *Almost exact recovery:* $\mathbb{P}\left(A(X, \hat{X}) = 1 - o(1)\right) = 1 - o(1)$.
- *Partial recovery:* $\mathbb{P}\left(A(X, \hat{X} \geqslant \alpha)\right) = 1 - o(1)$, $\alpha \in (0, 1)$.

Intuitively, exact recovery means that the entire partition is correctly recovered, almost exact allows for an increasingly smaller proportion of misclassified vertices and partial recovery allows for a constant fraction of misclassified nodes.

The connectivity of a symmetric stochastic block model exhibits a similar behavior to the Erdős-Rényi model.

Theorem 2.2. *Let n, k be positive integers, then the two following statements hold:*

- *For $a, b > 0$, $SBM(n, k, a \log n/n, b \log n/n)$ is connected with high probability if and only if $d = [a + (k-1)b]/k > 1$.*
- *$SBM(n, k, a/n, b/n)$ has a giant component if and only if $d > 1$.*

These two connectivity properties can be used to state recovery conditions for both the logarithmic and linear regime SBMs, since it can be proven that if $d < 1$ exact recovery in $SBM(n, k, a \log n/n, b \log n/n)$ and almost exact recovery in $SBM(n, k, a/n, b/n)$ are not solvable.

2.1.3 *Configuration model*

The configuration model is a model that generates random graphs from a given degree sequence. This can be accomplished by constructing a multigraph, i. e.a graph that allows multiple connection between nodes, with the given degree sequence in such a way that the number of multiple edges and self-loops converges to zero as the number of nodes increases.

A multigraph with n vertices and degree sequence $d = \{d_j\}_j$ can be constructed by taking k half-edges for each node j, choosing two half-edges uniformly at random and connecting them to form an edge. This procedure is repeated until all the half-edges are connected, and the resulting multigraph is called the *configuration model with degree sequence* d, abbreviated $CM_n(d)$.

If $G = (x_{ij})_{i,j\in[n]}$ is a multigraph on the vertices $[n]$ such that $d_i = x_{ii} + \sum_{j\in[n]} x_{ij}$, then

$$\mathbb{P}\left(CM_n(d) = G\right) = \frac{1}{(l_n-1)!!}\frac{\prod_{i\in[n]} d_i!}{\prod_{i\in[n]} 2^{x_{ii}}\prod_{1\leqslant i\leqslant j\leqslant n} x_{ij}!},$$

where $(l_n-1)!! = \prod_{k=1}^{l_n/2}(2k-1)$. Hence conditioned on the event $\{CM_n(d) \text{ is a simple graph}\}$, the configuration model has a uniform law over all simple graphs with degree sequence d.

While this sampling scheme produces random multigraphs with any given degree sequence[1], the model is not capable to produce simple graphs for every degree sequence; first they must satisfy the following regularity conditions:

1. **Weak convergence of vertex weight**. There exists a distribution function F such that
$$D_n \Longrightarrow D,$$
where F_n and F are the distribution functions of D_n and D, respectively, and D_n is the degree of a uniformly chosen vertex U, thus making $F_n = n^{-1}\sum_{j\in[n]} \mathbb{I}\{d_j \leqslant x\}$. We further assume that $D \geqslant 1$ a.s.

2. **Convergence of average vertex degrees**.
$$\lim_{n\to\infty} \mathbb{E}\left[D_n\right] = \mathbb{E}\left[D\right].$$

3. **Convergence of second moment vertex degrees**.
$$\lim_{n\to\infty} \mathbb{E}\left[D_n^2\right] = \mathbb{E}\left[D^2\right].$$

Theorem 2.3. *Assume $d = \{d_k\}_{k\in[n]}$ satisfy the above conditions. Then the probability that $CM_n(d)$ is a simple graph is asymptotically equal to $e^{-\nu/2-\nu^2/4}$, where*
$$\nu = \frac{\mathbb{E}\left[D(D-1)\right]}{\mathbb{E}\left[D\right]}.$$

There are two typical ways to obtain degree sequences that satisfy the regularity conditions.

- Fixed degrees moderated by a distribution function F.

  Fix a distribution function F take the number of vertices with degree k equal to
  $$n_k = \lceil nF(k)\rceil - \lceil nF(k-1)\rceil,$$
  and take the corresponding degree sequence $d = \{d_k\}$ as the unique ordered degree sequence compatible with $\{n_k\}_{k\geqslant 0}$. This implies that
  $$F_n(k) = \frac{1}{n}\lceil nF(k)\rceil,$$
  and all the conditions are fulfilled providing that D has finite second moment.

1 assuming its total degree l_n is an even number.

- Iid degrees.

  This situation comes with the problem that the sum of the sequence might not be odd, which impossibilitates the generation of a simple undirected graph. One way to overcome this is by adding a half-edge to the n vertex, increasing d_n by 1. Then the conditions above hold by the Strong Law of Large Numbers.

It is worth noting that if the second moment of D is infinite, then the probability of sampling a simple graph from the configuration model converges to zero.

2.1.4 *Preferential attachment*

Preferential attachment models describe a growing sequence of graphs $\left\{PA_t^{(m,\delta)}\right\}_{t\geqslant 0}$ such that $PA_t^{(m,\delta)}$ is a graph with t vertices and mt edges for some $m = 1, 2, \ldots$ and $\delta \geqslant -m$ and the probability of adding an edge to a vertex is an increasing affine function of its degree, promoting the emergence of hubs.

We will first define the model for $m = 1$. Henceforth, the vertices of $PA_t^{(m,\delta)}$ will be denoted by $v_1^{(1)}, v_2^{(1)}, \ldots, v_t^{(1)}$ and their degrees by $D_i(t)$. A self-loop increases the degree by 2, by convention.

The growth mechanism is defined inductively as follows:

Algorithm 2 Preferential attachment $PA_t^{(1,\delta)}$

1: Start with $PA_1^{(1,\delta)}$ having a single vertex with a single self-loop.
2: **for** $t = 2, 3, \ldots$ **do**
3: Add a vertex $v_{t+1}^{(1)}$ by connecting it with probability

$$\mathbb{P}\left(v_{t+1}^{(1)} \to v_i^{(1)} \,\middle|\, PA_t^{(1,\delta)}\right) = \begin{cases} \frac{1+\delta}{t(2+\delta)+(1+\delta)} & \text{for } i = t+1 \\ \frac{D_i(t)+\delta}{t(2+\delta)+(1+\delta)} & \text{for } i \in [t] \end{cases}$$

4: **end for**

Note how new edges are more likely to connect to high degree vertices, making their degrees even larger.

The mechanism for $m > 1$ can be defined accordingly as below:

Algorithm 3 Preferential attachment $PA_t^{(m,\delta)}$

1: Generate $PA_{mt}^{(1,\delta/m)}$.
2: **for** $j = 1, 2, \ldots$ **do**
3: Collapse the vertices $v_{(j-1)m+1}^{(1)}, \ldots, v_{jm}^{(1)}$ to form the vertex $v_j^{(m)}$ of $PA_t^{(m,\delta)}$.
4: **end for**

This construction implies that an edge in $PA_t^{(m,\delta)}$ is attached to vertex $v_j^{(m)}$ with probability proportional to the total weight of the m

vertices $v^{(1)}_{(j-1)m+1}, \dots, v^{(1)}_{jm}$, and since the total sum of their degrees is equal to the degree of $v^{(m)}_j$, the probability of adding an edge to it is proportional to its degree plus δ.

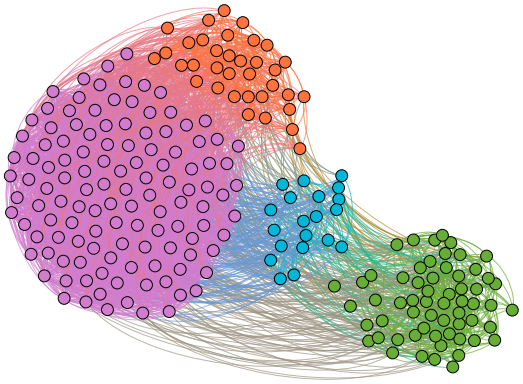

Figure 13: A graph sampled from a preferential attachment model using the Barabási-Albert algorithm.

There is an alternative definition for $m \geqslant 2$. We make the convention $D_{t+1}(0,t) = 0$.

Algorithm 4 Preferential attachment $PM_t^{(m,\delta)}$

1: Start with $PA_1^{(m,\delta)}$ having a single vertex with m self-loops.
2: **for** $t = 2, 3, \dots$ **do**
3: Add a vertex $v^{(1)}_{t+1}$ to $PA_t^{(m,\delta)}$ with m edges attached to it.
4: Connect each eth edge to $v^{(m)}_i$, $i \in [t]$ with probability

$$\propto \begin{cases} D_{t+1}(0,t) + 1 + e\delta/m & \text{for } i = t+1 \\ D_i(e-1,t) + \delta & \text{for } i \in [t] \end{cases}$$

5: **end for**

The above model for $\delta = 0$ is called the *Barabási-Albert model*, and several different properties arise when varying the parameters m and δ. For example, when $m = 1$ and $\delta \geqslant -1$, self-loops do not occur.

Preferential attachment models have the advantage of producing scale-free graphs. If we take the empirical degree distribution

$$P_k(t) = \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}\{D_i(t) = k\},$$

and for $m \geqslant 1$ and $\delta \geqslant -m$ we define $\{p_k\}_{k \geqslant 0}$ by $p_k = 0$ for $k = 0, \dots, m-1$ and for $k \geqslant m$,

$$p_k = \left(2 + \frac{\delta}{m}\right) \frac{\Gamma(k+\delta)\Gamma\left(m + 2 + \delta + \frac{\delta}{m}\right)}{\Gamma(m+\delta)\Gamma\left(k + 3 + \delta + \frac{\delta}{m}\right)}.$$

By rearranging $\sum p_k$ to a telescopic sum we can prove that $\{p_k\}$ is a probability mass function and

$$\mathbb{P}\left(\max_k |P_k(t) - p_k| \geqslant C\sqrt{\frac{\log t}{t}}\right) = o(1),$$

thus $P_k(t) \xrightarrow{\mathbb{P}} p_k$ in probability for any fixed k. p_k can be represented as the expectation of the functional of a negative binomial random variable X with random probability of success $U^{1/(2+\delta/m)}$, $U \sim \mathcal{U}[0,1]$,

$$p_k = \mathbb{E}\left[\mathbb{P}\left(X = k\right)\right],$$

from where we can deduce that, asymptotically

$$p_k = c_{m,\delta} k^{-\tau} \left(1 + O(1/k)\right)$$

with

$$c_{m,\delta} = \left(2 + \frac{\delta}{m}\right) \frac{\Gamma\left(m + 2 + \delta + \frac{\delta}{m}\right)}{\Gamma(m+\delta)} \text{ and } \tau = 3 + \frac{\delta}{m} > 2.$$

Therefore the asymptotic degree distribution of a stochastic block model is close to a power law with exponent τ, and any $\tau > 2$ can be obtained by properly tuning m and δ.

2.2 ALDOUS-HOOVER AND BEYOND

The sampling scheme induced by the Aldous-Hoover theorem can be utilized to build nonparametric models for (vertex) exchangeable random graphs by reconstructing graphons from an observed, finite network. This can be done by placing a suitable prior on the space of measurable functions from $[0,1]^3 \to [0,1]$.

The lack of uniqueness of the graphon representation makes the regression problem ill-posed, requiring a more structured approach. Two different directions have been explored (Cai, Ackerman, and Freer, 2015): estimating the graphon function W and estimating the values of the graphon function, i. e., inferring the graphon by directly recovering a measurable function up to equivalence from the data and estimating the latent probabilities $W(U_i, U_j)$ without considering the value of the graphon in any other points not contained in the original sample.

In the latter case the problem can be written as

$$A_{ij} = W(U_i, U_j) + \varepsilon_{ij} = P_{ij} + \varepsilon_{ij},$$

where A is the observed adjacency matrix, W the graphon and $U_k \overset{\text{iid}}{\sim} \mathcal{U}[0,1]$. This is reminiscent of the problem stated in factor analysis, where the loadings of some latent factors are estimated. In (Zhang, Levina, and Zhu, 2017) a smoothing estimator is defined by averaging over an apropiate set of neighbours of i, $\mathcal{N}_i$ such that $P_{i'j} \approx P_{ij}$ for all $i' \in \mathcal{N}_i$,

$$\hat{P}_{ij} = \frac{1}{|\mathcal{N}_i|} \sum_{i' \in \mathcal{N}_i} A_{i'j}.$$

When estimating the entire graphon an approximation in the cut distance can be made using piece-wise constant functions (Cai, Ackerman, and Freer, 2015) where a measurable partition of $[0,1]$ $S_1, \ldots, S_k$

called the *steps* of W is induced such that W is constant in every product set $S_i \times S_j$. The partition can be estimated, for instance, by using a stochastic block model approach or a Chinese restaurant process (Orbanz and Roy, 2015).

Another approach is to estimate a continuous graphon by placing a gaussian process with a suitable covariance function prior over the function W. Since continuous functions are dense in the space of measurable functions, we can approximate any graphon by this method. In (Lloyd et al., 2012) the graphon is decomposed into two functions H and Θ

$$F(U_i, U_j, U_{ij}) = H(U_{ij}, \Theta(U_i, U_j)).$$

A Gaussian process prior is then placed over Θ and H is assumed to be equal in law to

$$H(U_{ij}, \Theta(U_i, U_j)) \overset{d}{=} \Theta(U_i, U_j) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma).$$

Although vertex exchangeability gives an elegant way to model networks, it has some inherently problematic properties that makes it unfeasible for most applications. The first one is that it only produces dense graphs. The second, and probably the most dire one, is the fact that graphons cannot replicate collective behaviors.

Consider a large or infinite graph y. Fitting a graphon model to it is equivalent to the following sampling algorithm (Orbanz, 2017):

Algorithm 5 Sampling a subgraph of size k from a graph y of size n

1: Draw $\Phi_n \sim U(S_n)$.
2: Permute the graph $X_n := \Phi_n(y)$.
3: Return the subgraph $S_{n\to k}(y) = X_n\big|_k$.

In other words, this means that every random graph sampled from this algorithm is vertex exchangeable and for every graphon w there is a graph y such that its induced random graph is equal in law to $S_\infty(y)$, where S_∞ fulfills

$$S_\infty(y)\big|_k = \lim_{n\to\infty} S_{n\to k}(y).$$

Now consider that a sample of y $X_k = S_\infty(y)\big|_k$ contains a subgraph of size[2] $j \leqslant k$, x_j. By exchangeability, y must contain an infinite amount of copies of x_j. If x_j appears m times in X_k, then since there are $\binom{k}{j}$ possible subgraphs of size j in X_k, then the graphon model would assume that a frequency $m/\binom{k}{j}$ of all subgraphs of size j in y matches x_j, thus as j gets smaller, the probability that a reconstructed graph from a graphon contains x_j gets larger. In other words,

Small observed patterns have a higher probability than larger ones.

As a corollary of this last statement, since isolated subgraphs of X_k have a size k in order to explicitly include the missing edges, we have that graphon models do not reproduce, in general, partially isolated subgraphs (Orbanz, 2017).

2 In this contex the size is defined as the number of vertices.

2.2.1 *Caron-Fox model*

Caron and Fox (Caron and Fox, 2017) proposed a probabilistic model capable of producing sparse graphs that represents a random adjacency matrix as a simple point process on $\mathbb{R}_+$. We will describe said model next following (Caron and Fox, 2017) and (Veitch and Roy, 2015).

Consider an infinite symmetric adjacency matrix Z. It can be represented as

$$Z = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} z_{ij}\, \delta_{(\theta_i, \theta_j)}, \tag{6}$$

where $z_{ij} = z_{ji} = 1$ if θ_i and θ_j are connected and zero otherwise, and $\theta_k \in \mathbb{R}_+$ represents the time at which a potential node enters the network.

Each node θ_i is endowed with a *sociability parameter* w_i that define the the link probabilities

$$\mathbb{P}\left(z_{ij} = 1 \mid \{w_i\}\right) = \begin{cases} 1 - \exp\left\{-2 w_i w_j\right\} & i \neq j \\ 1 - \exp\left\{-w_i^2\right\} & i = j \end{cases}$$

Finally, the sociability parameters and node locations are sampled from a homogeneous completely random measure with no deterministic component with Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta) = \rho(\mathrm{d}w)\lambda(\mathrm{d}\theta)$, where ρ is a Lévy measure on $(0, \infty)$ and λ is the Lebesgue measure.

A more general model can be defined by considering directed multigraphs by changing the z_{ij} for n_{ij} the number of connections from i to j. In this case, given a CRM W a directed multigraph D can be generated from a Cox process with intensity given by the product measure $W \otimes W$ on $\mathbb{R}_+^2$. An undirected graph with self-edges can then be produced by connecting two nodes if there is an interaction between them, i. e., setting $z_{ij} = 1$ if $n_{ij} + n_{ji} > 0$ and o otherwise. This yields the following sampling scheme:

$$\begin{aligned} W &= \sum_{i=1}^{\infty} w_i \delta_{\theta_i} \sim \mathrm{CRM}(\rho, \lambda) \\ \mathrm{D} \mid W &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} n_{ij} \delta_{(\theta_i, \theta_j)} \sim \mathrm{PPP}(W \otimes W) \\ Z &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \min\{n_{ij} + n_{ji}, 1\} \delta_{\theta_i, \theta_j} \end{aligned} \tag{7}$$

This construction is equivalent to the simpler model (6). Indeed, if we identify the weights of the CRM as the sociability parameters, for $i \neq j$,

$$\mathbb{P}\left(z_{ij} = 1 \mid w\right) = \mathbb{P}\left(n_{ij} + n_{ji} > 0 \mid w\right) = 1 - \mathbb{P}\left(n_{ij} + n_{ji} = 0 \mid w\right),$$

where $w = \{w_i\}$. By the hierarchical construction above the n_{ij} are independent Poisson$(w_i w_j)$ random variables, thus $n_{ij} + n_{ji} \sim$

Poisson$(2w_iw_j)$ and the desired result follows. The case $i = j$ can be derived in an analogous way.

This model can be simulated from drawing upon exchangeability for point processes. A point process ξ is said to be *jointly exchangeable* if for any measure-preserving transformation $f : \mathbb{R}_+ \to \mathbb{R}_+$

$$\xi \stackrel{d}{=} \xi \circ (f \otimes f)^{-1},$$

where $\otimes$ is the tensor product. Note that this implies that for any intervals $A_i = [h(i-1), hi]$ with $i \in \mathbb{N}$ and $h > 0$,

$$\xi\left(A_i \times A_j\right) \stackrel{d}{=} \xi\left(A_{\pi(i)} \times A_{\pi(j)}\right)$$

for $(i, j) \in \mathbb{N}^2$ and any permutation π of $\mathbb{N}$.

Exchangeable random measures have a representation theorem due to Kallenberg (Veitch and Roy, 2015).

Theorem 2.4. *A random measure ξ on $\mathbb{R}_+^2$ is jointly exchangeable if and only if almost surely*

$$\begin{aligned}\xi = &\sum_{ij} f(\alpha, \vartheta_i, \vartheta_j, \zeta_{\{i,j\}})\delta_{\theta_i,\theta_j} + \\ &\sum_{jk}\left[g(\alpha, \vartheta_j, \chi_{jk})\delta_{\theta_j,\sigma_{jk}} + g'(\alpha, \vartheta_j, \chi_{jk})\delta_{\sigma_{jk},\theta_j}\right] + \\ &\sum_{k}\left[l(\alpha, \eta_k)\delta_{\rho_k,\rho'_k} + l'(\alpha, \eta_k)\delta_{\rho'_k,\rho_k}\right] + \\ &\sum_{j}\left[h(\alpha, \vartheta_j)(\delta_{\theta_k} \otimes \lambda) + h'(\alpha, \vartheta_j)(\lambda \otimes \delta_{\theta_k})\right] + \beta\lambda_D + \gamma\lambda^2\end{aligned}$$

for some measurable functions $f \geqslant 0$ on $\mathbb{R}_+^4$, $g, g' \geqslant 0$ on $\mathbb{R}_+^3$ and $h, h', l, l' \geqslant 0$ on $\mathbb{R}_+^2$, some collection of iid uniform random variables $\left(\zeta_{\{i,j\}}\right)$ on $[0, 1]$, some independent unit rate Poisson processes $\{(\theta_j, \vartheta_j)\}$ and $\{(\sigma_{ij}, \xi_{ij})\}_j$ for $i \in \mathbb{N}$ on $\mathbb{R}_+^2$ and $\{(\rho_j, \rho'_j, \eta_j)\}$ on $\mathbb{R}_+^3$ and some independent set of random variables $\alpha, \beta, \gamma \geqslant 0$. Here λ and λ_D are the Lebesgue measure on $\mathbb{R}_+$ and on the diagonal of $\mathbb{R}_+^2$ $\{(s, s) \in \mathbb{R}_+^2\}$

It can be proven (Caron and Fox, 2017) that the random measure (7) is jointly exchangeable and thus the above representation exists. In particular, if (θ_i, ϑ_i) is a Poisson process on $\mathbb{R}_+^2$ and we define the *tail Lévy intensity*

$$\tilde{\rho} := \int_x^\infty \rho(\,dw),$$

then the CRM $W = \sum_i w_i\delta_{\theta_i}$ with Lévy measure $\rho(\,dw)\lambda(\,d\theta)$ can be constructed from the bidimensional point process by taking $w_i = \tilde{\rho}^{-1}(\vartheta_i)$, and the Kallenberg representation for the model for undirected graphs is formulated by taking $\alpha = \beta = \gamma = 0$, $h = h' = l = l' = 0$ and

$$f(\alpha, \vartheta_i, \vartheta_j, \zeta_{\{i,j\}}) = \begin{cases} 1 & \zeta_{\{i,j\}} \leqslant M(\vartheta_i, \vartheta_j) \\ 0 & \text{otherwise} \end{cases},$$

where $M : \mathbb{R}_+^2 \to [0,1]$ is defined by

$$M(\vartheta_i, \vartheta_j) = \begin{cases} 1 - \exp\left\{-2\tilde{\rho}^{-1}(\vartheta_i)\tilde{\rho}^{-1}(\vartheta_j)\right\} & \vartheta_i \neq \vartheta_j \\ 1 - \exp\left\{-\tilde{\rho}^{-1}(\vartheta_i)^2\right\} & \vartheta_i = \vartheta_j \end{cases}$$

The function f is equivalent to the Aldous-Hoover representation for exchangeable arrays, and W in this case is akin to a graphon.

In order to study the sparsity properties of model (7) we must first introduce its restriction to the bounded squares $[0,\alpha]^2$, $\alpha > 0$. If the CRM has finite activity, there will be a finite number of (potential) nodes in $[0,\alpha]$, whereas if it has infinite activity, an infinite number is expected. Now we define the number of nodes with at least degree one N_α and the total number of edges $N_\alpha^{(e)}$,

$$N_\alpha = \#\{\theta_i \in [0,\alpha] : Z(\{\theta_i\} \times [0,\alpha]) > 0\}$$

$$N_\alpha^{(e)} = Z\left(\{(x,y) \in \mathbb{R}_+^2 : 0 \leqslant x \leqslant y \leqslant \alpha\}\right),$$

and we shall only consider Lévy measures ρ such that $\int_0^\infty \rho(\,dw) > 0$, since the case $\int_0^\infty \rho(\,dw) = 0$ corresponds to $N_\alpha = N_\alpha^{(e)} = 0$ almost surely. We also write for two almost surely divergent positive stochastic processes X_α and Y_α,

$$X_\alpha = o(Y_n) \text{ a.s.} \Leftrightarrow \lim_{n\to\infty} X_\alpha/Y_\alpha = 0 \text{ a.s.}$$

$$X_\alpha = \Theta(Y_n) \text{ a.s.} \Leftrightarrow \limsup_{n\to\infty} X_\alpha/Y_\alpha < \infty \text{ and } \limsup_{n\to\infty} Y_\alpha/X_\alpha < \infty \text{ a.s.}$$

Theorem 2.5. *Consider a point process Z representing an undirected graph as in (7), and assume that the directing Lévy measure ρ satisfies $\int_0^\infty w\,\rho(\,dw) < \infty$.*

- *If ρ has finite activity, then the number of edges scales quadratically with the number of observed nodes, $N_\alpha^{(e)} = \Theta(N_\alpha^2)$ almost surely when $\alpha \to \infty$, and thus the graph is dense.*
- *If ρ has infinite activity, the number of edges scales subquadratically with the number of observed nodes, $N_\alpha^{(e)} = o\left(N_\alpha^2\right)$ almost surely when $\alpha \to \infty$, implying that the graph is sparse.*

If we assume to have observed a set of undirected connections $\{z_{ij}\}_{1\leqslant i,j\leqslant N_\alpha}$ or undirected connections $\{n_{ij}\}_{1\leqslant i,j\leqslant N_\alpha}$, we order the location of the observed nodes $0 < \theta_1 < \cdots < \theta_{N_\alpha} < \alpha$ and we write $w_i = W(\{\theta_i\})$ the set of sociability parameters and ϕ the set of hyperparameters of the Lévy measure, then the posterior distribution for the node locations, sociability parameters of the observed nodes and the sum of the weights of the sociability parameters of the unobserved nodes w^* can be written as follows.

Theorem 2.6. *For $N_\alpha \geqslant 1$ let $\theta_1 < \ldots < \theta_{N_\alpha}$ be the set of support points of the measure $D_\alpha = \sum_{1 \leqslant i,j \leqslant N_\alpha} n_{ij} \delta_{(\theta_i,\theta_j)}$. Let $w_i = W_\alpha(\{\theta_i\})$ and $w_* = W^*_\alpha - \sum_{i=1}^{N_\alpha} w_i$. We have*

$$\mathbb{P}\left(\{w_i \in dw_i\}_{1 \leqslant i \leqslant N_\alpha}, w_* \in dw_* \mid \{n_{ij}\}_{1 \leqslant i,j \leqslant N_\alpha}, \phi\right) \propto$$

$$\exp\left\{-\left(\sum_{i=1}^{N_\alpha} w_i + w_*\right)^2\right\} \left\{\prod_{i=1}^{N_\alpha} w_i^{m_i} \rho(dw_i)\right\} G^*_\alpha(dw_*),$$

where $m_i = \sum_{j=1}^{N_\alpha} (n_{ij} + n_{ji}) > 0$ for $1 \leqslant i \leqslant N_\alpha$ are the node degrees of the multigraph and G^*_α is the Stiltjes measure of the random variable W^*_α with Laplace transform

$$\mathbb{E}\left[\exp\{-tW^*_\alpha\}\right] = \exp\{-\alpha\psi(t)\}.$$

Conditionally on observing an empty graph we have

$$\mathbb{P}\left(w_* \in dw_* \mid N_\alpha = 0, \phi\right) \propto \exp\left\{-w_*^2\right\} G^*_\alpha(dw_*).$$

The model we last described can be inscribed in a bigger family of random graphs models driven by *graphices* called *Kallenberg exchangeable graphs*, which will be described next in accordance with (Veitch and Roy, 2015). Let us first give some formal definitions. Let G be an undirected graph whose vertices are labeled with values in $\mathbb{R}_+$ and let $e(G)$ and $v(E)$ the set of edges and vertices of G, respectively.

Definition 2.4. *An **adjacency measure** is a locally finite symmetric simple measure in $\mathbb{R}_+$. The restriction of an adjacency measure ξ to $[0,\alpha]^2$, $\xi(\cdot \cap [0,\alpha]^2)$ is called an α-truncation. The adjacency measure of G is the adjacency measure $\sum_{(x,y)\in e(G)} \delta_{(x,y)}$.*

Given an adjacency measure $\sum_{i<\kappa} \delta_{e_i}$, where $e_1, e_2, \cdots \in \mathbb{R}^2_+$, we can construct its associated graph by taking the edge set equal to $\{e_i\}_{i<\kappa}$ and vertex set $\{x : e_i = (x,y)$ for some $i < \kappa, y \in \mathbb{R}_+\}$.

The correspondence between graphs and adjacency measures allows us to characterize random graphs invariant to the relabeling of their vertices using Kallenberg's scheme for exchangeable random measures, considering exchangeable adjacency maesures instead of exchangeable arrays. In this case, Kallenberg's representation assumes the following form:

Theorem 2.7. *Let ξ be a random adjacency measure. ξ is jointly exchangeable if and only if almost surely*

$$\xi = \sum_{ij} \mathbb{I}(\zeta_{\{i,j\}} \leqslant W(\alpha, \vartheta_i, \vartheta_j)))\delta_{\theta_i,\theta_j} +$$

$$\sum_{jk} \mathbb{I}(\xi_{jk} \leqslant S(\alpha, \vartheta_j)) \left(\delta_{\theta_j,\sigma_{jk}} + \delta_{\sigma_{jk},\theta_j}\right) +$$

$$\sum_{k} \mathbb{I}(\eta_k \leqslant I(\alpha)) \left(\delta_{\rho_k,\rho'_k} + \delta_{\rho'_k,\rho_k}\right)$$

for some measurable functions $S : \mathbb{R}_+^2 \to \mathbb{R}_+$, $I : \mathbb{R}_+ \to \mathbb{R}_+$ *and* $W : \mathbb{R}_+^3 \to [0,1]$ *with* $W(a,\cdot,\cdot)$ *symmetric for all positive* a, *some collection of iid uniform random variables on* $[0,1]$ $\{\zeta_{\{i,j\}}\}$ *some independent unit rate Poisson processes* $\{(\theta_j,\vartheta_j)\}$ *and* $\{(\sigma_{ij},\xi_{ij})\}_j$ *for* $i \in \mathbb{N}$ *on* $\mathbb{R}_+^2$ *and* (ρ_j,ρ'_j,η_j) *on* $\mathbb{R}_+^3$ *and an independent random variable* $\alpha \geqslant 0$ *that can be chosen deterministic if and only if* ξ *is an ergodic element.*

The second term corresponds to stars centered at the points $\{\theta_j\}$ and the third term to isolated edges. Since α is not random, the ergodic measures are completely determined by the triple (I,S,W). These functions cannot be arbitrarly chosen, we need to impose restrictions in order to guarantee that the restrictions to bounded rectangles are finite. These conditions are stated in the following definition:

Definition 2.5. *A* ***graphex*** *is a triple* (I,S,W), *where* $I \geqslant 0$ *is a non-negative real,* $S : \mathbb{R}_+ \to \mathbb{R}_+$ *is integrable and* $W : \mathbb{R}_+^2 \to [0,1]$ *is symmetric and satisfies the following conditions:*

1. $\Lambda\{\mu_W = \infty\} = 0$ *and* $\Lambda\{\mu_W > 1\} < \infty$,.
2. $\displaystyle\int_{\mathbb{R}_+^2} W(x,y)\mathbb{I}(\mu_W(x) \leqslant 1)\mathbb{I}(\mu_W(y) \leqslant 1)\,dx\,dy < \infty,$
3. $\displaystyle\int_{\mathbb{R}_+} W(x,y)\,dx\,dy < \infty,$

where Λ *denotes the Lebesgue measure on* $\mathbb{R}_+$ *and* $\mu_W(x) = \int_{\mathbb{R}_+} W(x,y)\,dy$ *is called the* ***graphex marginal****.*

Theorem (2.7) gives a simple algorithm to sample graphs from a graphex.

Algorithm 6 Sampling from a graphex

1: Draw nodes (θ_i,ϑ_i) from a unit rate Poisson process.
2: Connect them with probability $W(\vartheta_i,\vartheta_j)$.
3: **for** $i = 1,2,\ldots$ **do**
4: Sample nodes from a Poisson process with rate $S(\vartheta_i)$.
5: Connect them only to θ_i.
6: **end for**
7: Sample pairs of vertices connected only to each other from a Poisson process with rate I.

This lets us define the class of graphs sampled from a graphex: *Kallenberg exchangeable graphs.*

Definition 2.6. *A* ***Kallenberg exchangeable graph*** *(KEG) associated with a graphex* (I,S,W) *is the random graph* G *whose adjacency measure* χ *has the form* (2.7). G *induces a Kallenberg exchangeable graph model, a family of* α*-truncations* $G_\alpha = \xi(\cdot \cap [0,\alpha]^2)$, $\alpha \in \mathbb{R}$.

Now we shall consider only KEGs with graphex $(0,0,W)$, since the structural complexity of the adjacency matrix is only given by W. In this case, we have the following density result:

Theorem 2.8. *Let* G *be a Kallenberg exchangeable graph with graphex* $(0, 0, W)$*. If* W *is compactly supported, then* G *is dense almost surely. Conversely, if* W *is integrable and not compactly supported, then* G *is sparse almost surely.*

Consider W with compact support. Then it corresponds to the rescaling of a graphon. Now consider a graphon $\tilde{W}$. In correspondence, we can define a graphex with compact support$W$ by

$$W(x, y) = \begin{cases} \tilde{W}(x/c, y/c) & x \leqslant c, y \leqslant c \\ 0 & \text{otherwise} \end{cases}.$$

In this case points (θ, ϑ) of the latent Poisson point process will not connect to any edge if $\vartheta > c$, and thus for the finite size α-restriction G_α the underlying point process is a Poisson point process on $[0, \alpha] \times [0, c]$, rendering the following sampling scheme:

$$N_\alpha \sim \text{Poisson}(c\alpha)$$

$$\{\theta_i\} \mid N_\alpha \overset{\text{iid}}{\sim} U[0, \nu]$$

$$\{\vartheta_i\} \mid N_\alpha \overset{\text{iid}}{\sim} U[0, 1]$$

$$(\theta_i, \theta_j) \mid \vartheta_i, \vartheta_j \overset{\text{iid}}{\sim} \text{Bernoulli}\left(\tilde{W}(\vartheta_i, \vartheta_j)\right),$$

which corresponds to the usual graphon model with iid uniform labels instead of integer ones, proving that the only compactly supported graphices are graphons.

Another specific case of a graphex model corresponds to the Caron-Fox model. Indeed, we can choose

$$W(x, y) = \begin{cases} 1 - \exp\left\{-2\tilde{\rho}^{-1}(x)\tilde{\rho}^{-1}(y)\right\} & x \neq y \\ 1 - \exp\left\{-\tilde{\rho}^{-1}(x)^2\right\} & x = y \end{cases},$$

where $\tilde{\rho}^{-1}$ is the inverse tail Lévy intensity of an underlying completely random measure, then we recover the Caron-Fox model.

2.2.2 *Edge-exchangeable graphs*

Imagine uniformly sampling a list of emails. This can be represented as a set of tuples $\{(x_i, x_j)\}_{i,j}$, where x_i represents the sender and x_i the receiver. If we uniformly sample this list, then the law of the resulting sample would be invariant to *edge* relabeling, that is, the probability of observing a set of emails is the same regardless of which specific emails are included in the sample. This type of invariance is called *edge exchangeability*, and can be defined as the law of a graph being invariant to the order of arrival of its edges.

The theory behind edge-exchangeable graphs has been parallely developed in (Cai, Campbell, and Broderick, 2016a) and (Crane and Dempsey, 2018), and while they are somewhat equivalent, we will

restrict ourselves to the construction developed in (Cai, Ackerman, and Freer, 2015) and (Cai, Campbell, and Broderick, 2016b).

Consider a projective sequence of graphs, i. e., a sequence $G_1 = (V_1, E_1), G_2 = (V_2, E_2), \dots$ such that $E_n \subseteq E_{n+1}$ for all n. Let us only consider graphs without isolated vertices, so the set of vertices can be defined as the union of the edges, making each G_n completely determined by E_n. Using this perspective, each graph G_n can be represented as a set of vertices, each vertex being itself a set containing the indices of the edges it is attached to, for instance, $\{\{1,2,3\},\{2\},\{1,3\}\}$.

Define for each graph G_n the *step-augmented edge set* E'_n as the collection of edges decorated with the step in which it was added to the sequence of graphs, $E'_n = \{(e_1, s_1), \dots, (e_m, s_m)\}$, for some m. Then the *step-augmented graph sequence* can be defined as a sequence of step-augmented edge sets $\{E'_n\}_n = (E'_1, E'_2, \dots)$.

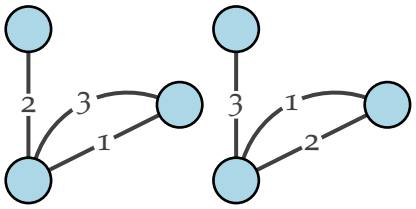

Figure 14: Left: a planar representation of the graph encoded by $\{\{1,2,3\},\{2\},\{1,3\}\}$. Right: the same graph with its edge labels reordered. Under edge exchangeability both graphs have the same probability.

Definition 2.7. *Let $\{G_n\}_n$ be a random graph sequence with step-augmented edge sets E'_n. We call $\{G_n\}_n$ **edge-echangeable** if for every n and every permutation π of $[n]$ $G_n \overset{d}{=} \tilde{G}_n$, where $\tilde{G}_n$ has step-augmented edge set $\pi E'_n := \{(e_1, \pi(s_1)), \dots, (e_m, \pi(s_m))\}$.*

Intuitively, edge exchangeability of a random graph sequence means that its law is invariant to the order of arrival of its edges.

A de Finetti-like representation for this symmetry can be stated by defining a *graph paint-box*. Heuristically, a graph paint-box corresponds to first sampling a sequence of random open subsets of $(0,1)$ representing the vertices, $\{C_k\}_k$, C'_1 and C'_2, and sampling iid sequence of uniform random variables on $(0,1)$, $\{V_n\}_n$, each V_n corresponding to an edge. The first sequence of vertices $\{C_k\}$ represents *regular vertices* and the other two the *dust vertices*, i. e., vertices that occur in only one edge (regular vertices being vertices that are not dust vertices). We then have the next three cases:

- if $V_n \in C_k \cap C_j$ for $k \neq j$, then the nth edge connects vetices k and j, and we restrict V_n to only be contained in two of the C_i at a time.
- If $V_n \in C_k$ and $V_n \in C'_1$, the nth edge connects a regular vertex with a dust vertex that only ever connects with the nth edge.
- If $V_n \in C'_1$ and $V_n \in C'_2$, the nth edge links two different dust vertices, producing an isolated component.

Definition 2.8. *An edge-exchangeable graph sequence* $\{G_n\}_n$ *possesses a graph paint-box if there exists random open Lebesgue measurable subsets* $\{C_k\}_k$ *and* C_1' *and* C_2' *of* $(0,1)$ *such that*

1. *the sets* C_2' *and* $\bigcup_k C_k \cup C_1'$ *are disjoint,*
2. *any* $V \in (0,1)$ *is an element of either* C_2' *or exactly two sets* $(C_k)_k$ *and* C_1'.

The graph sequence can then be generated by doing

1. *Sample* $V_i \overset{iid}{\sim} U(0,1)$.
2. *Set* $I_k = \{n : V_n \in C_k\}$ *for all* k.
3. *Set* $I_1' = \{\{n\} : V_k \in C_i'\}$ *and* $I_2' = \{\{n\}, \{n\} : V_n \in C_2'\}$.
4. *For each* n, *let* G_n *be the collection* $\{I_k\}_k, I_1'$ *and* I_2' *truncated up to the first* n *indices.*

Clearly every graph sampled from a graph paint-box is edge-exchangeable since the V_i are independent. The converse also holds, that is, an edge-exchangeable random graph always has a graph paint-box, yielding the desired integral representation (Cai, Campbell, and Broderick, 2016b). Thus, the law of every edge-exchangeable graph is fully characterized by a random measure on $\mathbb{N}^2$.

The distribution of an edge-exchangeable graph has a decomposition akin to the exchangeable partition probability functions studied in the last chapter. Let $\bar{G}_n$ be the multiset containing the degree of each vertex, e. g., for the graph in figure 14, $\bar{G}_n = \{3,1,2\}$. Also let $\kappa(G_n)$ be the number of unique orderings of the sets in G_n. For instance, in the last example, $\kappa(G_n) = 3! = 6$. We say G_n has an *exchangeable vertex probability function (EVPF)* if the distribution of G_n depends only on $\kappa(G_n)$, n and $\bar{G}_n$, the multiset of degrees of the vertices of G_n .

Definition 2.9. *A random graph* G_n *has an exchangeable vertex probability function if its law can be decomposed as*

$$\mathbb{P}(G_n) = \kappa(G_n)\, p\left(n, \bar{G}_n\right),$$

for some function $p : \mathbb{N} \times \mathcal{G} \to \mathbb{R}_+$, *where* $\mathcal{G}$ *is the set of all finite graphs.*

Because the distribution of a random graph having an EVPF only depends on the degrees of the vertices, it is edge exchangeable; however, not all edge exchangeable random graphs have an EVPF. In fact, edge exchageable graphs having one are called *frequency models* (Cai, Campbell, and Broderick, 2016a).

Take an infinite amount of latent vertices indexed by the natural numbers $\mathbb{N}$, an infinity of edge labels $\{\theta_{\{i,j\}}\}$ in a set Θ and positive edge frequencies $\{w_{\{i,j\}}\}$ in $\mathbb{R}_+$. Either $\{\theta_{\{i,j\}}\}$ or $\{w_{\{i,j\}}\}$ can be random, and we define the random measure on Θ

$$W := \sum_{\{i,j\}:i,j\in\mathbb{N}} w_{\{i,j\}} \delta_{\theta_{\{i,j\}}}. \tag{8}$$

There are two posiible ways to construct an edge-exchangeable graph from this last measure. If the weights are normalized, i. e., $\sum_{i,j} w_{\{i,j\}} = 1$, then $\{w_{i,j}\}_{i,j\in\mathbb{N}}$ is a probability distribution over all possible vertex pairs, and then a multigraph can be sampled by the following algorithm:

Algorithm 7 Normalized case

1: Draw $\{w_{\{i,j\}}\}$ and $\{\theta_{\{i,j\}}\}$.
2: Set $E_0 = \emptyset$.
3: **for** $k = 1, 2, \ldots$ **do**
4: Sample an edge e from the distribution $\{w_{\{i,j\}}\}$.
5: Set $E_k = E_{k-1} \cup \{e\}$.
6: **end for**

If the weights do not sum up to one, then the measure is not a probability measure. Consequently we introduce $f(m \mid w)$ a distribution over the non-negative integers given some rate $w \in \mathbb{R}_+$. The multigraph then is sampled as follows.

Algorithm 8 Non-normalized case

1: Draw $\{w_{\{i,j\}}\}$ and $\{\theta_{\{i,j\}}\}$.
2: Set $E_0 = \emptyset$ and $F = \emptyset$.
3: **for** $k = 1, 2, \ldots$ **do**
4: For every edge $e = \{i, j\}$ draw its multiplicity $m_e \sim f(\cdot|w_e)$.
5: Add m_e copies of edge e to F.
6: Set $E_k = E_{k-1} \cup F$.
7: **end for**

In both cases, since conditionally on $\{w_{\{i,j\}}\}$ the edges are iid, the resulting multigraphs are edge-exchangeable, and these algorithms are called *frequency models*.

Theorem 2.9. *A regular graph sequence has a frequency model if and only if it has an EVPPF.*

This last theorem implies that by restricting to the frequency models we can obtain a tractable familiy of models suitable for estimation, the particular properties of the graphs sampled from each model depending on the choice of the random measure W.

At a first glance the representation (8) is quite similar to Kallenberg exchangeable graphs; albeit it has some differences. In the construction developed by Caron and Fox (Caron and Fox, 2017) the weight measure $W = \sum_i w_i \delta_{\theta_i}$ is first sampled and then the graph measure

$$G = \sum_{i,j} g_{ij} \delta_{\{\theta_i, \theta_j\}}$$

is constructed by sampling the g_{ij} once given w_{ij} for each pair i, j, whereas graph frequency models repeatedly sample the number of edges g_{ij} and add the result. Furthermore, the graphex-based model

generates a projective family of graphs by restricting the graph measure to the rectangle $[0, y] \times [0, y]$ for some $y \in \mathbb{R}_+$, and then progressively increasing the value of y; in contrast, frequency models fix y beforehand and grow the network by the resampling of the g_{ij}.

Example 2.1. Let $W \sim \mathrm{PPP}(\nu)$ for some non-atomic, $\sigma-$finite measure ν satisfying $\nu([0,1]) = \infty$ and $\int_0^1 w \, d\nu < \infty$, so that W is a countable infinite set of rates and $\sum_{w \in W} w < \infty$ almost surely. Take $w_{\{i,j\}} = w_i w_j$ if $i \neq j$ and $w_{\{i,i\}} = 0$ for all i.

We use algorithm 8 to sample a multigraph (V, E) by taking $f(\cdot \mid w)$ be Bernoulli(w), which makes the multiplicity of each edge after n steps $\mathrm{Binom}(n, w_i w_j)$. A graph $(\bar{V}, \bar{E})$ can be then generated by setting $\bar{V} = V$ and $\bar{E}$ equal to the set of edges with multiplicity at least one.

The density properties of the generated graphs can be summarized by the following theorem.

Theorem 2.10. *Suppose ν has a regularly varying tail, so there exist $\alpha \in (0,1)$ and $l : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $c > 0$, $\lim_{x\to\infty} \frac{l(cx)}{l(x)} = 1$ and*

$$\int_x^1 \nu(\, dw) \sim x^\alpha l(x^{-1}) \text{ when } x \to 0.$$

Then as $n \to \infty$,

$$\#V_n \overset{a.s.}{=} \Theta\left(n^\alpha l(n)\right),$$

$$\#E_n \overset{a.s.}{=} \Theta(n) \text{ and}$$

$$\#\bar{E}_n \overset{a.s.}{=} O\left(l(n^{1/2}) \min\left\{n^{(1+\alpha)/2}, l(n) n^{3\alpha/2}\right\}\right).$$

This last theorem implies that the multigraph is sparse when $\alpha \in (1/2, 1)$ and thus its restriction to a graph is sparse when $\alpha \in (0, 1)$.

Now let us consider a variation of the graph frequency model. Take $w_{ij} = w_i w_j$ and $W \sim \mathrm{PD}(\alpha, \theta)$. This model is called the *Hollywood model* (Crane, 2018), and a computationally tractable way to simulate from this model can be derived by harnessing the stick-breaking representation of the two-parameter Poisson-Dirichlet distribution, constructing the edge frequencies and vertex labels at the same time.

Assume at time t the network has $t-1$ edges and a random number of vertices N_t with $N_1 = 0$. Label these vertices as $1, \ldots, N_t$ and take $D_t(i)$ as the degree of node i before the tth edge is added. When the $t-$th edge arrives, its two attached vertices $v_1(t)$ and $v_2(t)$ are randomly chosen between the existing vertices $1, \ldots, N_t$ and a new vertex N_{t+1} as follows.

$$\mathbb{P}\left(v_1(t) = i\right) \propto \begin{cases} D_t(i) - \alpha, & i = 1, \ldots, N_t \\ \theta + \alpha N_t & i = N_t + 1 \end{cases} \tag{9}$$

After $v_1(t)$ is chosen, $v_2(t)$ is sampled using (9) replacing $N_t = N_t + 1$ if $v_1(t) = N_t + 1$ and leaving it as N_t if no new node has been added.

This generates a growing sequence of networks $\{G_n\}_{n\in\mathbb{N}}$ with n edges and a random number N_n of vertices, whose expected value satisfies as $n \to \infty$ (Crane, 2018)

$$\mathbb{E}[N_n] \sim \frac{\Gamma(\theta)+1}{\alpha\Gamma(\theta+\alpha)}(2n)^{\alpha}.$$

Moreover, if we write $N_k(g)$ as the number of vertices of degree k in a graph g with vertex set $V(g)$, this construction gives a closed-form formula for the distribution of G_n,

$$\mathbb{P}(G_n = g) = \alpha^{\#V(g)}\frac{(\theta/\alpha)^{\uparrow v(g)}}{\theta^{\uparrow(2n)}}\prod_{k=2}^{\infty}\exp\left\{N_k(g)\log(1-\alpha)^{\uparrow(k-1)}\right\},$$

where $x^{\uparrow j} = x(x+1)\cdots(x+j-1)$ is the rising factorial.

We then get a result regarding the sparsity and power-law properties of the Hollywood process.

Theorem 2.11. *Let $G = \{G_n\}_{n\in\mathbb{N}}$ be a realization of a Hollywood process with parameters α and θ. Then*

- *$\{G_n\}_{n\in\mathbb{N}}$ is sparse almost surely if $1/2 < \alpha < 1$, and*
- *$\{G_n\}_{n\in\mathbb{N}}$ has almost surely a power law degree distribution with exponent $\alpha+1$ if $0 < \alpha < 1$.*

3

EDGE-EXCHANGEABLE DYNAMIC NETWORKS

Consider a sequence of time-evolving interactions that can appear and disappear over time, *e.g.*, the interactions between elements in a complex chemical reaction. This type of phenomena can be modelled through dynamic random graphs by considering time-dependent connection probabilities. In this section we will adapt the methods developed in (Palla, Caron, and Teh, 2016) to the context of edge-exchangeable frequency models.

Start with a frequency model

$$W = \sum_{i,j} w_i \delta_{\theta_{\{i,j\}}}$$

$$G \mid W = \sum_{i,j} e_{\{i,j\}} \delta_{(i,j)}$$

$$e_{ij} \overset{\text{iid}}{\sim} \text{Ber}(w_i w_j)$$

Here we fix the atoms $\theta_{i,j}$ belonging to a compact space Θ and following (Palla, Caron, and Teh, 2016) we incorporate continuous-time dynamics by sampling the latent sociabilities from a time-dependent random measure on Θ

$$W_t = \sum_{i,j} w_i(t) w_j(t) \delta_{\theta_{\{i,j\}}}, \quad t \geqslant 0,$$

conceived by obtaining the weights from a continuous-time Feller process known as an SF-Harris process (Anzarut and Mena, 2019) with transition semigroup

$$K_t f(x) = \left(1 - e^{-\lambda t}\right) Wf + e^{-\lambda t} f(x), \quad \lambda > 0,$$

where $W = (w_i)_i$ is a Poisson point process on $[0, 1]$ with a non-atomic, σ-finite intensity measure ν satisfying $\nu([0, 1]) = \infty$ and $\int_0^1 w \, \nu(\mathrm{d}w) < \infty$, guaranteeing that W_t is countably infinite and that $\sum_{w \in W_t} w < \infty$ almost surely.

The SF-Harris process is time-reversible and has W as invariant measure, which makes it strictly stationary. Further, it is uniformly ergodic and has mean $\mathbb{E}[W_t] = \mathbb{E}[W] = \nu$. Having an arbitrary stationary measure allows us to incorporate power-law behaviours, making the model more flexible.

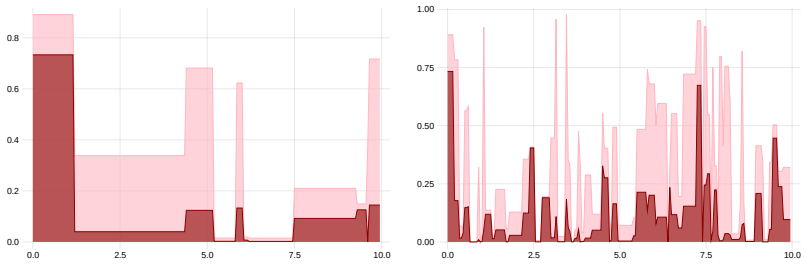

Figure 15: Sample trajectories of the two largest weights of an SF-Harris process with a standard beta process with parameters $\alpha = 1$ and $\gamma = 5$ as its invariant measure and two different values of λ: 1 on the left and 10 on the right. Note how higher values of λ correspond to a higher amount of jumps.

We also model the presence of a connection between each pair of nodes as a continuous-time two-state chain such that infinitesimally, given W_t,

$$\mathbb{P}\left(e_{ij}(t+\mathrm{d}t) = 1 \mid e_{ij}(t) = 0\right) = w_i(t)w_j(t)\,\mathrm{d}t + o(\mathrm{d}t) \tag{10}$$

$$\mathbb{P}\left(e_{ij}(t+\mathrm{d}t) = 0 \mid e_{ij}(t) = 1\right) = \rho\,\mathrm{d}t + o(\mathrm{d}t)$$

$$\mathbb{P}\left(e_{ij}(t+\mathrm{d}t) = 1 \mid e_{ij}(t) = 1\right) = (1-\rho)\,\mathrm{d}t + o(\mathrm{d}t)$$

$$\mathbb{P}\left(e_{ij}(t+\mathrm{d}t) = 0 \mid e_{ij}(t) = 0\right) = \left(1 - w_i(t)w_j(t)\right)\,\mathrm{d}t + o(\mathrm{d}t)$$

Thus the infinitesimal generator $\mathcal{A}^{(ij)}(t)$ for this process has the form

$$\mathcal{A}^{(ij)}(t) = \begin{bmatrix} -w_i(t)w_j(t) & w_i(t)w_j(t) \\ \rho & -\rho \end{bmatrix}$$

This implies that for each pair of nodes θ_i, θ_j new edges arrive according to a Poisson process with intensity $w_i(t)w_j(t)$ and are deleted after an exponentially-distributed random time with intensity ρ.

Note how since the state-space is finite then (11) is a Feller process (Kallenberg, 2002). Moreover, we consider the $e_{ij}(t)$ independent given W_t, hence the process $(e_{ij}(t))_{ij}$ is still Feller since by proposition 19.3 of (Kallenberg, 2002), weak convergence of the random sequence is equivalent to the weak convergence of each individual component, thus fulfilling the necessary conditions.

This scheme yields the following model:

$$W_t \sim \text{SF-Harris}(\alpha, W) \tag{11}$$

$$e_{ij}(t) \mid W_t = \sum w_i(t)\,\delta_{\theta_i} \sim \text{2-State-Chain}(\rho, w_i w_j) \tag{12}$$

$$G_t \mid e_{ij}(t) = \sum_i^\infty e_{ij}(t)\,\delta_{\theta_{(i,j)}}$$

An additional assumption required to better model real-world phenomena is that at each time t only a finite number of interactions

are present, which can be achieved by imposing restrictions on the intensity measure of W.

This sampling procedure can be formalized by considering the pair

$$\begin{pmatrix} G_t \mid W_t \\ W_t \end{pmatrix} \tag{13}$$

taking values on $\mathcal{M}_\theta^{\text{sim}} \times \mathcal{M}_\theta^{[0,1]}$, the product of spaces of simple point processes and atomic measures on $[0,1]$ with a fixed set of atoms θ.

The time-homogeneous transition semigroup of this process takes the form

$$T_t f = \int\int f(w,e)\, P_t^w(e,\mathrm{d}e)\, S_t(w,\mathrm{d}w), \quad f \in C_b\left(\mathcal{M}_\theta^{\text{sim}} \times \mathcal{M}_\theta^{[0,1]}\right),$$

where S_t is the semigroup of the process W_t and P_t^w the one of $G_t|W_t$.

This last semigroup can be obtained by considering the transformation $\phi_\theta : \{0,1\}^{\mathbb{N}} \to \mathcal{M}_\theta^{\text{sim}}$,

$$(e_{ij})_{i,j} \xrightarrow{\varphi} \sum_{i,j} e_{ij}\, \delta_{\theta_{ij}},$$

where $\{0,1\}^{\mathbb{N}}$ has the product topology induced by endowing each $\{0,1\}$ with the discrete metric, which coincides with the ℓ_1 distance. This makes $\{0,1\}^{\mathbb{N}}$ compact and Hausdorff (as a subset of $[0,1]$).

Proposition 3.1. *The transformation φ_θ is a homeomorphism.*

Proof. First let us show that φ_θ is continuous. To do this notice that a sequence $(x_i)_{i\in\mathbb{N}}$ in $\{0,1\}^{\mathbb{N}}$ convergences to $x = (x^1, x^2, \dots) \in \{0,1\}^{\mathbb{N}}$ if and only if each of its projections converges in $\{0,1\}$, and thus for each $\varepsilon > 0$ and each n we can take $2^{-n-1}\varepsilon$ and find an $N_n > 0$ such that $|x_i^n - x_i| < 2^{-n-1}\varepsilon$ for all $n \geqslant N_n$, hence

$$\sum_{n=1}^{\infty} |x_i^n - x_i| < \sum_{n=1}^{\infty} 2^{-n-1}\varepsilon = 2\frac{\varepsilon}{2} = \varepsilon,$$

which entails that convergence in the product topology implies convergence in ℓ_1. Hence for all bounded measurable functions f (each bounded by K_f) and all $\mu, \nu \in \mathcal{M}_\theta^{\text{sim}}(\Theta)$

$$\int f\, \mathrm{d}\mu - \int f\, \mathrm{d}\nu = \sum_i e_{ij}^\mu f(\theta_{ij}) - \sum_{ij} e_{ij}^\nu f(\theta_{ij}) \leqslant K \sum |e_{ij}^\mu - e_{ij}^\nu|,$$

and thus φ_θ is continuous on its coefficients e_{ij}; moreover, it also has a continuous inverse

$$\sum_{i,j} e_{ij}\, \delta_{\theta_{ij}} \xrightarrow{\varphi^{-1}} (e_{ij})_{i,j},$$

since if a sequence $x_n = \{(x_n^1, x_n^2, \dots)\} \subset \{0,1\}^{\mathbb{N}}$ converges in ℓ_1 to $x \in \{0,1\}^{\mathbb{N}}$ then for all i

$$|x_n^i - x^i| \leqslant \sum_{i=1}^{\infty} |x_n^i - x^i|,$$

implying that each component of x_n converges in ℓ_1 and therefore in the discrete topology, so x_n converges in the product space. The continuity of φ^{-1} follows by taking the constant function equal to 1 in $\mathcal{M}^{\text{sim}}(\Theta)$. □

Then by taking the transition semigroup of (e_{ij}) given W_t, i. e.

$$P_t^* f = \mathbb{E}\left[f\left((e_{ij})\right) \mid \mathscr{F}_t^e, W_t\right],$$

where $\mathscr{F}_t^e$ is the natural filtration generated by the process (e_{ij}), we can define for all $f \in C_b\left(\mathcal{M}^{\text{sim}}(\Theta)\right)$

$$P_t^w f = P_t^*(f \circ \varphi_\theta) \circ \varphi_\theta^{-1},$$

which by the continuity of both φ_θ and φ_θ^{-1} is a well defined transition semigroup on $\mathcal{M}^{\text{sim}}(\Theta)$.

Now we shall give conditions for the finiteness of the number of edges at each time. Note that by Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left(\#E_t = \infty\right) &= \mathbb{P}\left(\bigcap_{k=1}^{\infty} \#E_t \geqslant k\right) \\
&= \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{k=1}^{n} \#E_t \geqslant k\right) \\
&= \lim_{n\to\infty} \mathbb{P}\left(\#E_t \geqslant n\right) \leqslant \lim_{n\to\infty} \frac{\mathbb{E}\left[\#E_t\right]}{n},
\end{aligned}$$

which implies that if the expected number of connections is finite then $\#E_t$ is almost surely finite. A sufficient condition for integrability can be stated in terms of the Lévy measure of the stationary distribution of W_t.

Theorem 3.1. *The expected number of edges* $\#E_t$ *and vertices* $\#V_t$ *is given by*

$$\mathbb{E}\left[\#E_t\right] = \int\int \frac{wv}{\rho + wv}\, \nu(\,dw)\, \nu(\,dv)$$

and if $\rho > 1$

$$\mathbb{E}\left[\#V_t\right] = \int 1 - \exp\left\{-\int 1 - \frac{\rho}{\rho + wv}\, \nu(\,dv)\right\} \nu(\,dw).$$

Proof. First we note that the form of the transition semigroup of the SF-Harris process implies that if W is the stationary measure then by Campbell's theorem for all ν-integrable functions f,

$$\mathbb{E}\left[f(W_t)\right] = \mathbb{E}\left[f(W)\right] = \int f(w)\; d\nu(\,dw).$$

Further, if for each $e_{ij}(t)$ we take the initial measure

$$\mu(\,de \mid w_i, w_j) = \frac{w_i w_j}{w_i w_j + \rho}\delta_1 + \frac{w_i w_j}{w_i w_j + \rho}\delta_0,$$

then the law of $e_{ij}(t)$ equals

$$\mathcal{L}\left(e_{ij}(t)\right) = \frac{w_i(t)w_j(t)}{w_i(t)w_j(t)+\rho}\delta_1 + \frac{\rho}{w_i(t)w_j(t)+\rho}\delta_0.$$

This can be seen by noting that the two-state Markov chain is time-homogeneus and has that exact invariant measure, making it strictly stationary, and hence by starting the randomized chain on μ the relation holds.

This, along with Fubini's theorem gives

$$\begin{aligned}
\mathbb{E}\left[\#E_t\right] &= \mathbb{E}\left[\sum_{i,j=1}^{\infty} e_{ij}(t)\right] \\
&= \sum_{i,j=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[e_{ij}(t) \mid W_t, e_{ij}(s) = 0\right]\right] \\
&= \sum_{i,j=1}^{\infty} \mathbb{E}\left[\frac{w_i w_j}{\rho + w_i w_j}\right].
\end{aligned}$$

Rearranging the sums and using Slivnyak-Mecke's[1] and Fubini's theorems,

$$\begin{aligned}
\mathbb{E}\left[\#E_t\right] &= \sum_{w\in W} \mathbb{E}\left[\sum_{v\in W\setminus\{w\}} \frac{wv}{\rho+wv}\right] \\
&= \sum_{w\in W} \int \mathbb{E}\left[\frac{wv}{\rho+wv}\right] \nu(\,\mathrm{d}v) \\
&= \int \mathbb{E}\left[\sum_{w\in W} \frac{wv}{\rho+wv}\right] \nu(\,\mathrm{d}v)
\end{aligned}$$

Then, by Campbell's theorem,

$$\mathbb{E}\left[\#E_t\right] = \int \mathbb{E}\left[\sum_{w\in W} \frac{wv}{\rho+wv}\right] \nu(\,\mathrm{d}v) = \int\int \frac{wv}{\rho+wv}\, \nu(\,\mathrm{d}w)\, \nu(\,\mathrm{d}v).$$

Now for the vertices, by the conditional independence of the e_{ij} and Fubini's theorem,

$$\begin{aligned}
\mathbb{E}\left[\#V_t\right] &= \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{I}(\mathrm{Deg}_i(t) > 0)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{P}\left(\mathrm{Deg}_i(t) > 0 \mid W_t, e_{ij}(s) = 0\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\infty} \left(1 - \prod_{j=1}^{\infty} \mathbb{P}\left(e_{ij}(t) = 0 \mid e_{ij}(s) = 0, W_t\right)\right)\right] \\
&= \mathbb{E}\left[\sum_{w\in W_t} \prod_{v\in W_t\setminus\{w\}} \left(1 - \frac{\rho}{\rho+wv}\right)\right].
\end{aligned}$$

1 See appendix A.

Using Slivnyak-Mecke's theorem

$$\begin{aligned}\mathbb{E}\left[\#V_t\right] &= \mathbb{E}\left[\sum_{w\in W_t}\left[1-\prod_{v\in W_t\setminus\{w\}}\frac{\rho}{\rho+wv}\right]\right]\\ &= \int \mathbb{E}\left[1-\prod_{v\in W_t}\frac{\rho}{\rho+wv}\right]\nu(\,\mathrm{d}w)\\ &= \int 1-\mathbb{E}\left[\exp\left\{\sum_{v\in W_t}\log\left(\frac{\rho}{\rho+wv}\right)\right\}\right]\nu(\,\mathrm{d}w)\end{aligned}$$

Finally, since $w \in (0,1)$ and noting that $\left|\log\frac{\rho}{\rho+wv}\right| = -\log\frac{\rho}{\rho+wv}$, exponentiating we get

$$\frac{\rho+wv}{\rho} = 1+\frac{w}{\rho}v \leqslant 1+v \leqslant e^v \quad \text{if } \rho > 1,$$

and $\int_0^1 v\,\nu(\,\mathrm{d}v) < \infty$, $\int_0^1 \log\left(\frac{\rho}{\rho+wv}\right)\nu(\,\mathrm{d}v) < \infty$, so by Campbell's theorem,

$$\begin{aligned}\mathbb{E}\left[\#V_t\right] &= \int 1-\mathbb{E}\left[\exp\left\{\sum_{v\in W_t}\log\left(\frac{\rho}{\rho+wv}\right)\right\}\right]\nu(\,\mathrm{d}w)\\ &= \int 1-\exp\left\{-\int 1-\frac{\rho}{\rho+wv}\,\nu(\,\mathrm{d}v)\right\}\nu(\,\mathrm{d}w).\end{aligned}$$

□

In the context of (13) this last result is true for any initial measure $\mu^{\otimes\infty}(\,\mathrm{d}e \mid w)\,\nu_0(\,\mathrm{d}w)$, where ν_0 is whichever starting measure for W and $\mu_{\phi_x}^{\otimes\infty}$ is the pushforward measure of ϕ_x with respect to the infinite-product measure $\mu^{\otimes\infty}$ obtained from the Daniell-Kolmogorov extension theorem by the conditional independence of the $e_{ij}(t)$.

Furthermore, by the geometric ergodicity of the two-state Markov chain it is not unreasonable to conjecture the asymptotic veracity of the statement for any initial measure of the process (13).

3.1 SIMULATION

Sampling for the aforementioned stochastic process can be done in a two-stage procedure:

1. Fix a sequence of times $t_1,\ldots,t_n$ and simulate a trajectory of the sociability parameters,

$$(W_{t_1},\ldots W_{t_n}),$$

2. then sample the edges' paths,

$$\left(e_{ij}(t_1),\ldots,e_{ij}(t_n)\right)_{ij}.$$

While the second stage of this procedure is straight-forward to do, the feasibility of the first one depends on the exact random measure chosen as the stationary distribution of the SF-Harris process, since one must simulate from some finite-dimensional approximation of it.

Some random measures, like the Dirichlet or the beta process[2], can be managed by truncating their stick-breaking representation (Sethuraman, 1994), (Teh, Görür, and Ghahramani, 2007); but this approach is not always feasible nor computationally tractable (Favaro, Lijoi, and Prünster, 2012).

In the general setting, one wishes to find a series representation for the random measure W, that is, a way to expand W as

$$W = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}.$$

Doing so would yield an iterative sampling method that can be stopped at any time to obtain a finite-dimensional representation of W. Now we will present three different ways this can be achieved, based on the survey provided in (Campbell et al., 2019): the inverse Lévy, Bondesson and thinning representations. In the following, let ν be the jump part of the Lévy measure of a homogeneous completely random measure W and let $\Gamma_k = \sum_{l=1}^{k} E_l$, with $E_l \overset{\text{iid}}{\sim} \text{Exp}(1)$ be the ordered jumps of a homogeneous unit-rate Poisson process on $\mathbb{R}_+$.

Inverse Lévy representation

Let $\bar{\nu}(x) = \nu([x, \infty))$ be the tail Lévy measure and define its generalized inverse

$$\bar{\nu}^{-1}(u) = \inf_x \{x : \nu([x, \infty)) \leqslant u\}.$$

Then, in distribution,

$$W \overset{d}{=} \sum_{k=1}^{\infty} \nu^{-1}(\Gamma_k)\ \delta_{\theta_k}.$$

The weights of this representation are non-increasing, guaranteeing the sampling of atoms with the higher mass first; however, the inverse tail Lévy measure is not analitically tractable in most of the cases (Campbell et al., 2019).

Bondesson representation

W has a Bondesson representation if for $c > 0$ and a density g on $\mathbb{R}$,

$$W \overset{d}{=} \sum_{k=1}^{\infty} V_k e^{-\Gamma_k/c} \delta_{\theta_k}, \quad V_k \overset{\text{iid}}{\sim} g.$$

Not all completely random measures have said representation; indeed, if $\nu(\mathrm{d}\theta) = \nu(\theta)\,\mathrm{d}\theta$, i. e., ν has a (Lebesgue) density such that $\theta\nu(\theta)$

2 A beta process is a CRM on $[0, 1]$ with Lévy measure $\nu(\mathrm{d}w) = \gamma\alpha w^{-1}(1-w)^{1-\alpha}\,\mathrm{d}w$.

is nonincreasing, $\lim_{\theta\to\infty}\theta\nu(\theta) = 0$, $c_\nu := \lim_{\theta\to 0}\theta\nu(\theta) < \infty$ and $c_\nu > 0$ then we can take

$$g = -c_\nu^{-1}\frac{d}{dv}v\nu(v)$$

and ν has a Bondesson representation with density g.

Thinning representation

We say that W has a thinning representation if there exists a probability measure g on $\mathbb{R}_+$ such that ν is absolutely continuous with respect to g and

$$W \overset{d}{=} \sum_{k=1}^{\infty} V_k \mathbb{I}\left(\frac{d\nu}{dg}(V_k) \geqslant \Gamma_k\right)\delta_{\theta_k}, \quad V_k \overset{iid}{\sim} g.$$

This has the disadvantage that as $\Gamma_k \to \infty$ a.s. when $k \to \infty$, the probability that $\frac{d\nu}{dg}(V_k) \geqslant \Gamma_k$ decreases in k and thus the amount of atoms sampled with zero mass becomes increasingly frequent, making this representation inefficient for a large amount of atoms.

Example 3.1. The 3-parameter beta process on $[0,1]$ is defined as having Lévy measure

$$\nu(\,dw) = \gamma\frac{\Gamma(1+\alpha)}{\Gamma(1-\beta)\Gamma(\alpha+\beta)}w^{-1-\beta}(1-w)^{\alpha+\beta-1}\,dw,$$

where $\gamma > 0$ and $\beta \in (0,1)$. This CRM has finite total mass if $\alpha > -\beta$, and has been known to exhibit power-law behaviours when normalized (Campbell et al., 2019).

If $\beta = 0$ then it corresponds to the standard beta process and if we further let $\alpha = 1$, it has inverse Lévy representation

$$\sum_{k=1}^{\infty} W_i\,\delta_{\theta_i}, \quad W_k = \prod_{i=1}^{k}\beta_i,\ \beta_i \overset{iid}{\sim} \text{Beta}(\gamma,1). \tag{14}$$

Moreover, if $\alpha > 1$ and $\beta = 0$, then $\theta\nu(w) = \gamma\alpha(1-w)^{1-\alpha}$, which is nonincreasing in $[0,1]$. Further,

$$c_\nu = \gamma\alpha \quad \text{and} \quad g(v) = (\alpha-1)(1-v)^{\alpha-2}.$$

This corresponds to the density of a $\text{Beta}(1,\alpha-1)$ random variable, from which the weights of the Bondesson representation are then sampled.

If we take $g = \text{Beta}(1-\beta,\alpha+\beta)$ the thinning representation for the beta process takes the form

$$\sum_{k=1}^{\infty} V_k\,\mathbb{I}\left(V_k\Gamma_k \leqslant \gamma\right)\,\delta_{\theta_k}, \quad V_k \overset{iid}{\sim} \text{Beta}\left(1-\beta,\alpha+\beta\right).$$

Now we will illustrate the empirical behaviour of the graph sequences sampled from this model by employing a standard beta process as the stationary measure of the sociability parameters.

We truncate the outer sum to 10,000 and fix the parameters of the beta process to $\alpha,\rho = 1$ and let $\gamma = 5,100$. To simulate we use the

inverse Lévy representation (14) at times $t = 0, 1, 2$ and the transition of the two-state chain:

$$\begin{cases} p_{00}(t) &= \dfrac{\rho}{w_i w_j + \rho} + \dfrac{w_i w_j}{w_i w_j + \rho} e^{-(w_i w_j + \rho)t} \\ p_{01}(t) &= \dfrac{w_i w_j}{w_i w_j + \rho} - \dfrac{w_i w_j}{w_i w_j + \rho} e^{-(w_i w_j + \rho)t} \\ p_{10}(t) &= \dfrac{\rho}{w_i w_j + \rho} - \dfrac{\rho}{w_i w_j + \rho} e^{-(w_i w_j + \rho)t} \\ p_{11}(t) &= \dfrac{w_i w_j}{w_i w_j + \rho} + \dfrac{\rho}{w_i w_j + \rho} e^{-(w_i w_j + \rho)t} \end{cases},$$

which yields the following algorithm:

Algorithm 9 Graph sampling with a beta process base measure

Sample $v_i \overset{\text{iid}}{\sim} \text{Beta}(\gamma, 1)$ for $i = 1, \dots n$.
Take $w_i(0) = \prod_{j \leqslant i} v_i$ for all i.
Sample $u \sim U(0, 1)$.
Sample G_0 by adding an edge between nodes i and j if

$$u < \frac{w_i w_j}{w_i w_j + \rho}.$$

for $t = t_1, t_2, \dots$ **do**
    Sample $u \sim U(0, 1)$.
    if $u \leqslant e^{-\lambda t}$ **then**
        Set $W_{t_k} = W_{t_{k-1}}$
    else
        Simulate $v_i \overset{\text{iid}}{\sim} \text{Beta}(\gamma, 1)$ for $i = 1, \dots n$.
        Take $w_i(t_k) = \prod_{j \leqslant i} v_i$ for all i.
    end if
    Sample $u \sim U(0, 1)$.
    if $e_{ij}(t_{k-1}) == 0$ **then**
        if $u < \frac{w_i(t_k) w_j(t_k)}{w_i(t_k) w_j(t_k) + \rho} - e^{-(w_i(t_k) w_j(t_k) + \rho)t} \frac{w_i(t_k) w_j(t_k)}{w_i(t_k) w_j(t_k) + \rho}$ **then**
            Set $e_{ij}(t_k) = 1$
        else
            Set $e_{ij}(t_k) = 0$
        end if
    else
        if $u < \frac{\rho}{w_i(t_k) w_j(t_k) + \rho} - e^{-(w_i(t_k) w_j(t_k) + \rho)t} \frac{\rho}{w_i(t_k) w_j(t_k) + \rho}$ **then**
            Set $e_{ij}(t_k) = 0$
        else
            Set $e_{ij}(t_k) = 1$
        end if
    end if
end for

In order to illustrate that the model produces sparse graphs we sample a sequence of dynamic networks by varying the truncation

number and calculating the number of edges present at each time. The results are presented in figure 16.

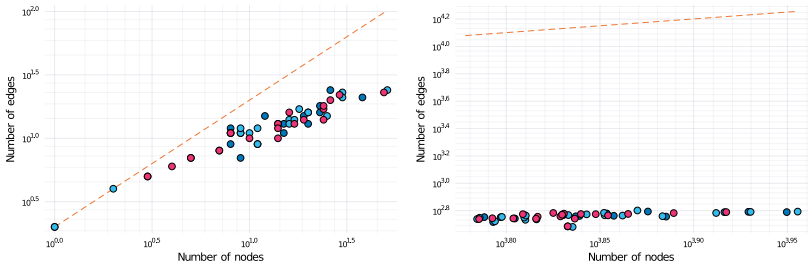

Figure 16: Simulated data from a model with weights generated by a beta process. The plot on the left was generated using $\gamma = 3$, while the one on the right with $\gamma = 50$. Each color represents a different sampling time. The dashed line is a line with slope 2. Both axes are on a log scale.

To check whether each graph was sparse we plotted the number of vertices against the number of edges (on a log scale) and compared it to a line of slope 2. In all cases the point cloud appears to follow a straight line with a slope smaller than two, which indicates that the graphs are sparse since, empirically, $\#E_n \sim o\left(\#V_n^2\right)$. Also we note that higher levels of γ give sparser networks.

As a complement we simulated again a graph sequence with parameters $\rho, \alpha = 1$ and $\gamma = 15$. The degree distribution of each graph is plotted on figure 17. Note how the shape does not seem to correspond to a power law, which is expected since the beta process fails to reproduce that kind of behaviour (Broderick, Jordan, and Pitman, 2012).

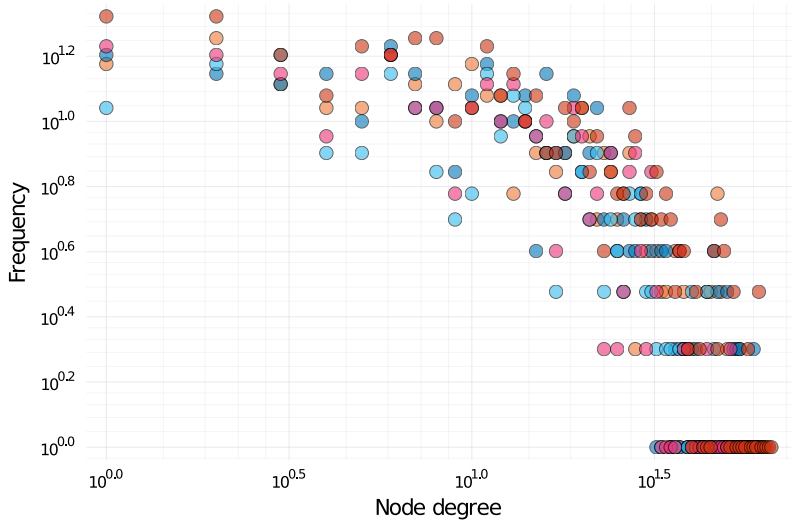

Figure 17: Degree distribution of a graph sequence across times 1, 2, 3, 4 and 5. The number of atoms was truncated to 200. Each color represents a different sampling time. Both axes are in a log scale.

4

CONCLUSIONS

Building probabilistic models for relational data that can mimic real-world behaviour has proven to be a daunting challenge. Most of the classic models fail to reproduce sparcity and power laws, or are too restrictive to be useful for applications.

In recent years a new trend of models that exploit the rich theory of Bayesian nonparametrics has arisen. They move past the framework provided by dense graph limits by encoding the connection probabilities as a random measure whose atoms correspond to the nodes in the graph it produces.

Throughout this dissertation we have reviewed some of these models, namely the ones based on graphices (Caron and Fox, 2017), (Veitch and Roy, 2015) and on edge-exchangeability (Crane and Dempsey, 2018), (Cai, Campbell, and Broderick, 2016a). These are able to produce sparse graphs and power laws, provided their base completely random measures have said property.

On the last chapter we built a continuous-time graph-valued Markov process based on the edge-exchangeable framework and explored its sparsity properties through simulation. While we have found that the graps it samples are indeed sparse, the specific role of its parameters, particularly the death rate of its edges, needs to be further studied and verified theoretically. Moreover, a closer examination of its properties as a Markov process can be useful, specifically stability can continuity results, e. g., ergodicity and whether the process has the Feller property.

In the same fashion as in (Palla, Caron, and Teh, 2016), a direct extension of this model is to consider a linear-birth-death process instead of a two-state chain to generalize it to multigraphs. This however would complicate the topology of its state-space, making it more difficult to handle.

Finally, an estimation procedure still needs to be researched. Markov chain Montecarlo is the usual choice for Bayesian models, since it can separate each individual parameter and perform an iterative algorithm. This carries the problem of efficiency, since usually many iterations need to be performed in order to guarantee convergence. Another possibility, not necessarily incompatible with MCMC, is the usage of the many algorithms deviced in the machine learning community for estimating hidden Markov processes in order to propagate the information of the graph into the latent connection probabilities.

APPENDIX A. FUNCTIONALS OF POISSON POINT PROCESSES

In this appendix we will state two different theorems useful for computing expectations of functionals of Poisson point processes. The first one is Campbell's theorem.

Theorem A.1. *(Campbell) (Kingman, 1967). Let Π be a Poisson point process on S with rate measure ν and let $f : S \to \mathbb{R}$ be measurable. Then*

$$\mathbb{E}\left[\sum_{x\in\Pi} f(x)\right] = \int_S f(x)\,\nu(\,dx).$$

Moreover, if

$$\int_S 1 \wedge |f(x)|\,\nu(\,dx) < \infty,$$

then for all $\theta \in \mathbb{C}$

$$\mathbb{E}\left[e^{\theta\left(\sum_{x\in\Pi} f(x)\right)}\right] = \exp\left\{\int_S \left(e^{\theta f(x)} - 1\right)\,\nu(\,dx)\right\}.$$

The second part of the theorem is particularly useful to calculate expectations of products by transforming them to a sum using the logarithm and then composing it with the exponential function; provided the integrability condition holds.

The second theorem is a corollary of the Slivnyak-Mecke's theorem for reduced Palm distributions of point processes, which intuitively correspond to conditioning the process to having a point on some location.

Theorem A.2. *(Slivnyak-Mecke) (Cai, Campbell, and Broderick, 2016a). Let Π be a Poisson point process on S with rate measure ν and let $f : S \times \Omega \to \mathbb{R}_+$ be measurable. Then*

$$\mathbb{E}\left[\sum_{x\in\Pi} f\left(x, \Pi \setminus \{x\}\right)\right] = \int_S \mathbb{E}\left[f(x,\Pi)\right]\,\nu(\,dx).$$

B

APPENDIX B. ASYMPTOTICS NOTATION

Here we will establish the notation used throughout this work, directly transcribing appendix A.1 in (Caron and Fox, 2017). Let $(X_t)_{t\geqslant 0}$ and $(Y_t)_{t\geqslant 0}$ two almost surely divergent stochastic processes defined on the same probability space. Then

$$X_t = O(Y_t) \text{ a.s. if } \limsup_{t\to\infty} \frac{X_t}{Y_t} < \infty \text{ a.s.}$$

$$X_t = o(Y_t) \text{ a.s. if } \limsup_{t\to\infty} \frac{X_t}{Y_t} = 0 \text{ a.s.}$$

$$X_t = \Omega(Y_t) \text{ a.s. if } Y_t = O(X_t) \text{ a.s.}$$

$$X_t = \omega(Y_t) \text{ a.s. if } Y_t = o(X_t) \text{ a.s.}$$

$$X_t = \Theta(Y_t) \text{ a.s. if } X_t = O(Y_t) \text{ and } X_t = \Omega(Y_t) \text{ a.s.}$$

These can be interpreted as

$X_t = O(Y_t)$ if X_t does not grow at a faster rate than Y_t

$X_t = o(Y_t)$ if X_t grows at a strictly slower rate than Y_t

$X_t = \Omega(Y_t)$ if X_t does not grow at a slower rate than Y_t

$X_t = \omega(Y_t)$ if X_t strictly grows at a faster rate than Y_t

$X_t = \Theta(Y_t)$ if X_t and Y_t grow at the same rate.

C

APPENDIX C. GLOSSARY

This small glossary is designed to help readers understand the key concepts behind this dissertation, providing short and less technical explanations than the ones given in the main text.

- **Dense graph**. We say a graph is dense if the number of edges is roughly of the same order as the maximum number of possible edges, which in the case when the graph has n edges corresponds to it being of order n^2.

  This kind of graphs are usually not suitable for real-world applications since most graphs observed empirically are not dense (Orbanz and Roy, 2015).

- **Kallenberg exchangeability**. Kallenberg exchangeability of a graph consists of its adjacency matrix represented as a point process being exchangeable, that is, being invariant (in law) to area-preserving transformations. In terms of network theory, this is translated to the moment of arrival of each new node to the network being exchangeable. As an example consider Facebook. If its profiles and the friendships between them are Kallenberg-exchangeable then its distribution is invariant to the order in which new users sign up, not taking into account whether they have any connections whatsoever.[1]

- **Power laws**. A power law describes the behaviour of the tail of a probability distribution, making it decay slowly. In the case of random graphs, it models situations where nodes with a distinguishably larger amount of edges than the average are observed, for instance, an epidemic in which certain individuals can spread the disease to an unexpectedly large number of people.

  In contrast to exchangeable graphs, a network with a power law exhibits unhomogeneous behaviours better suitable for many real-world applications.

- **Sparse graph**. A graph is sparse if, in average, has the same amount of edges as vertices. Intuitively this implies that the number of connections is much less than the maximum amount possible, in contrast with a dense graph.

  This definition of sparsity is somewhat arbitrary, tailored to work on graph sequences with an increasing number of nodes rather than on a single, isolated finite graph (Crane, 2018). The general definition is rather vague, a graph being sparse simply if the number of edges is negligible compared to the number of nodes.

1 These interpretation and example were taken from section 7.3.3 of (Crane, 2018)

- **Vertex exchangeability**. A random network is vertex-exchangeable if its distribution remains invariant to the reordering of its vertices. This is a homogeneity condition. Indeed, since the distribution of the graph must look the same from the perspective of every vertex, the graph is not likely to have any discernable patterns. According to (Crane, 2018), this also corresponds to any finite sample of the network being representative of it.

BIBLIOGRAPHY

Aldous, David J. (1985). "Exchangeability and Related Topics". In: *École d'Été de Probabilités de Saint-Flour XIII — 1983*. Ed. by David J. Aldous, Illdar A. Ibragimov, Jean Jacod, and P. L. Hennequin. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer, pp. 1–198.

Anzarut, Michelle and Ramsés H. Mena (Apr. 1, 2019). "A Harris Process to Model Stochastic Volatility". In: *Econometrics and Statistics* 10, pp. 151–169.

Bertoin, Jean (2006). *Random Fragmentation and Coagulation Processes*. Cambridge Studies in Advanced Mathematics. Cambridge: Cambridge University Press. ISBN: 978-0-521-86728-3. DOI: 10.1017/CBO9780511617768.

Borgs, Christian, Jennifer Chayes, Henry Cohn, and Nina Holden (Jan. 26, 2016). "Sparse Exchangeable Graphs and Their Limits via Graphon Processes". In: *Journal of Machine Learning Research* 18.

Broderick, Tamara, Michael I. Jordan, and Jim Pitman (June 2012). "Beta Processes, Stick-Breaking and Power Laws". In: *Bayesian Analysis* 7.2, pp. 439–476.

Cai, Diana, Nathanael Ackerman, and Cameron Freer (May 11, 2015). *An Iterative Step-Function Estimator for Graphons*. arXiv: 1412.2129 [math, stat].

Cai, Diana, Trevor Campbell, and Tamara Broderick (2016a). "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. 4249–4257.

Cai, Diana, Trevor Campbell, and Tamara Broderick (2016b). "Paintboxes and Probability Functions for Edge-Exchangeable Graphs". In.

Campbell, Trevor, Jonathan H. Huggins, Jonathan P. How, and Tamara Broderick (May 2019). "Truncated Random Measures". In: *Bernoulli* 25.2, pp. 1256–1288. ISSN: 1350-7265. DOI: 10.3150/18-BEJ1020.

Caron, François and Emily B. Fox (2017). "Sparse Graphs Using Exchangeable Random Measures". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.5, pp. 1295–1366.

Chiu, Sung Nok, Dietrich Stoyan, Wilfrid S. Kendall, and Joseph Mecke (Sept. 10, 2013). *Stochastic Geometry and Its Applications*. Edición: 3. Chichester, West Sussex, United Kingdom: Wiley. 582 pp. ISBN: 978-0-470-66481-0.

Crane, Harry (Apr. 17, 2018). *Probabilistic Foundations of Statistical Network Analysis*. 1 edition. Chapman and Hall/CRC. 234 pp.

Crane, Harry and Walter Dempsey (July 3, 2018). "Edge Exchangeable Models for Interaction Networks". In: *Journal of the American Statistical Association* 113.523, pp. 1311–1326.

Daley, D. J. and David Vere-Jones (Sept. 1, 2007). *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure: 2*. Edición: 2nd ed. New York: Springer. ISBN: 978-0-387-21337-8.

Favaro, S., A. Lijoi, and I. Prünster (2012). "On the Stick-Breaking Representation of Normalized Inverse Gaussian Priors". In: *Biometrika* 99.3, pp. 663–674. ISSN: 0006-3444.

Ferguson, Thomas S. (Mar. 1973). "A Bayesian Analysis of Some Nonparametric Problems". In: *Annals of Statistics* 1.2, pp. 209–230.

Ghosal, Subhashis and Aad van der Vaart (June 26, 2017). *Fundamentals of Nonparametric Bayesian Inference*. Edición: 1. Cambridge ; New York: Cambridge University Press. 670 pp. ISBN: 978-0-521-87826-5.

Goldstein, Micheal (n.d.). *Observables and Models: Exchangeability and the Inductive Argument*. Oxford University Press. ISBN: 978-0-19-174416-7. URL: https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199695607.001.0001/acprof-9780199695607-chapter-1 (visited on 02/26/2021).

Hofstad, Remco van der (2016). *Random Graphs and Complex Networks*. Vol. 1. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. ISBN: 978-1-107-17287-6. DOI: 10.1017/9781316779422.

James, Lancelot F (Dec. 15, 2003). "A Simple Proof of the Almost Sure Discreteness of a Class of Random Measures". In: *Statistics & Probability Letters* 65.4, pp. 363–368.

James, Lancelot F., Antonio Lijoi, and Igor Prünster (2009). "Posterior Analysis for Normalized Random Measures with Independent Increments". In: *Scandinavian Journal of Statistics* 36.1, pp. 76–97.

Kallenberg, Olav (2002). *Foundations of Modern Probability*. 2nd ed. Probability and Its Applications. New York: Springer-Verlag.

Kallenberg, Olav (Apr. 25, 2017). *Random Measures, Theory and Applications*. Edición: 2017 ed. New York, NY: Springer. ISBN: 978-3-319-41596-3.

Kingman, J. F. C. (1967). "Completely Random Measures." In: *Pacific Journal of Mathematics* 21.1, pp. 59–78. ISSN: 0030-8730.

Kolaczyk, Eric D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. New York: Springer-Verlag. ISBN: 978-0-387-88145-4. DOI: 10.1007/978-0-387-88146-1.

Lloyd, James, Peter Orbanz, Zoubin Ghahramani, and Daniel M Roy (2012). "Random Function Priors for Exchangeable Arrays with Applications to Graphs and Relational Data". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., pp. 998–1006.

Orbanz, Peter (Oct. 11, 2017). *Subsampling Large Graphs and Invariance in Networks*. arXiv: 1710.04217 [math, stat].

Orbanz, Peter and Daniel M. Roy (Feb. 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.

Palla, Konstantina, François Caron, and Yee Whye Teh (July 6, 2016). *Bayesian Nonparametrics for Sparse Dynamic Networks*. arXiv: 1607.01624 [stat].

Schervish, Mark J. (1995). *Theory of Statistics*. Springer Series in Statistics. New York: Springer-Verlag. ISBN: 978-0-387-94546-0. DOI: 10.1007/978-1-4612-4250-5.

Sethuraman, Jayaram (Jan. 1, 1994). "A Constructive Definition of the Dirichlet Prior". In: *Statistica Sinica* 4, pp. 639–650.

Teh, Yee, Dilan Görür, and Zoubin Ghahramani (Jan. 1, 2007). "Stick-Breaking Construction for the Indian Buffet Process." In: *Journal of Machine Learning Research - Proceedings Track* 2, pp. 556–563.

Veitch, Victor and Daniel M. Roy (2015). "The Class of Random Graphs Arising from Exchangeable Random Measures". In.

Zhang, Yuan, Elizaveta Levina, and Ji Zhu (Dec. 1, 2017). "Estimating Network Edge Probabilities by Neighbourhood Smoothing". In: *Biometrika* 104.4, pp. 771–783.