Beyond the Chinese Restaurant and Pitman-Yor processes: Statistical Models with double power-law behavior

Fadhel Ayed [*1] **Juho Lee** [*1 2] **François Caron** [1]

Abstract

Bayesian nonparametric approaches, in particular the Pitman-Yor process and the associated two-parameter Chinese Restaurant process, have been successfully used in applications where the data exhibit a power-law behavior. Examples include natural language processing, natural images or networks. There is also growing empirical evidence suggesting that some datasets exhibit a two-regime power-law behavior: one regime for small frequencies, and a second regime, with a different exponent, for high frequencies. In this paper, we introduce a class of completely random measures which are doubly regularly-varying. Contrary to the Pitman-Yor process, we show that when completely random measures in this class are normalized to obtain random probability measures and associated random partitions, such partitions exhibit a double power-law behavior. We present two general constructions and discuss in particular two models within this class: the beta prime process (Broderick et al. (2015, 2018) and a novel process called generalized BFRY process. We derive efficient Markov chain Monte Carlo algorithms to estimate the parameters of these models. Finally, we show that the proposed models provide a better fit than the Pitman-Yor process on various datasets.

1. Introduction

Power-law distributions appear to arise in a wide range of contexts, including natural languages, natural images or networks. For example, the empirical distribution of the word frequencies in natural languages is well approximated by a power-law distribution, an observation attributed to Zipf (1935). That is, the frequency $f_{(k)}$ of the kth most frequent word in a corpus satisfies, within some range

$$f_{(k)} \simeq Ck^{-\xi}$$

where C is some constant and $\xi > 0$ is the power-law exponent which is typically close to 1 for natural languages. These empirical findings have motivated the development of numerous generative models that can reproduce this power-law behavior; see the reviews of (Mitzenmacher, 2004) and (Newman, 2005).

Amongst these generative models, Bayesian nonparametric hierarchical models based on infinite-dimensional random measures have been successfully used to capture the power-law behavior of various datasets. Applications include natural language processing (Goldwater et al., 2006; Teh, 2006; Wood et al., 2009; Mochihashi et al., 2009; Sato & Nakagawa, 2010), natural image segmentation (Sudderth & Jordan, 2009) or network analysis (Caron, 2012; Caron & Fox, 2017; Crane & Dempsey, 2018; Cai et al., 2016). A very popular model is the Pitman-Yor (PY) process (Pitman, 1995; Pitman & Yor, 1997; Pitman, 2006), an infinite-dimensional random probability measure whose properties induce a power-law behavior. It admits two parameters ($0 \leq \alpha < 1, \theta > -\alpha$). The PY random probability measure is almost surely discrete, with weights $\pi_{(1)} \geq \pi_{(2)} \geq, \ldots$ following the so-called two-parameter Poisson-Dirichlet distribution $\mathrm{PD}(\alpha, \theta)$ (Pitman & Yor, 1997). For $\alpha > 0$, the random weights satisfy

$$\pi_{(k)} \sim k^{-1/\alpha} S \quad \text{almost surely as } k \to \infty$$

where S is a random variable. That is, small weights asymptotically follow a power-law distribution whose exponent is controlled by the parameter α. The PY process also enjoys tractable alternative constructions via the two-parameter Chinese restaurant process or the stick-breaking construction which explains its great popularity amongst models with similar properties. Other popular infinite-dimensional random measures that have been used for their similar power-law properties include the stable Indian buffet process (Teh & Gorur, 2009) or the generalized gamma process (Hougaard, 1986; Brix, 1999).

[*]Equal contribution [1]Department of Statistics, University of Oxford, Oxford, United Kingdom [2]AITRICS, Seoul, Republic of Korea. Correspondence to: Fadhel Ayed <fadhel.ayed@stats.ox.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

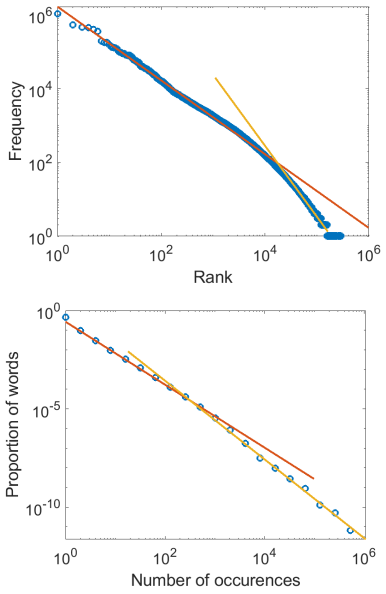


Figure 1. (Top) Ranked word frequencies from the American National Corpus (circles) and power-law fit (straight lines). (Bottom) Proportion of words with a given number of occurences for the same dataset (circles) and power-law fit (straight lines).

Double power-law in empirical data. There is a growing empirical evidence that some datasets may exhibit a double power-law regime when the sample size is large enough. Examples include word frequencies in natural languages (Ferrer i Cancho & Solé, 2001; Montemurro, 2001; Gerlach & Altmann, 2013; Font-Clos et al., 2013), Twitter rates and retweet distributions (Bild et al., 2015), or degree distributions in social (Csányi & Szendrői, 2004), communication (Seshadri et al., 2008) or transportation networks (Paleari et al., 2010). In the case of word frequencies for example, it is conjectured that high frequency words approximately follow a power-law with Zipfian exponent approximately equal to 1, while the low frequency words follow a power-law with a higher exponent. An illustration is given in Figure 1, which shows the word frequencies of about 300,000 words from the American National Corpus[1].

In this paper, we introduce a class of completely random measures (CRMs), named doubly regularly-varying CRMs. We show that, when a random measure in this class is normalized to obtain a random probability measure P, and one repeatedly samples from P, the resulting frequencies exhibit a double power-law behavior. Informally, the ranked frequencies satisfy

$$f_{(k)} \simeq \begin{cases} C_1 k^{-1/\tau} & \text{for small rank } k \\ C_2 k^{-1/\alpha} & \text{for large rank } k \end{cases} \tag{1}$$

[1]http://www.anc.org/data/anc-second-release/frequency-data/

where $\tau > 0$, $\alpha \in (0,1)$ and $C_1, C_2 > 0$. The above statement is made mathematically accurate later in the article. We describe two general constructions to obtain doubly regularly varying CRMs, and consider two specific models within this class: the beta prime process of Broderick et al. (2015; 2018) and a novel process named generalized BFRY process. We show how these two CRMs can be obtained from transformations of the generalized gamma and stable beta processes. We derive Markov chain Monte Carlo inference algorithms for these models, and show that such models provide a good fit compared to a Pitman-Yor process on text and network datasets.

2. Background on (normalized) completely random measures

CRMs, introduced by Kingman (1967), are important building blocks of Bayesian nonparametric models (Lijoi & Prünster, 2010). A homogeneous CRM on a Polish space Θ, without deterministic component nor fixed atoms, is almost surely (a.s.) discrete and takes the form

$$W = \sum_{k \geq 1} w_k \delta_{\theta_k} \tag{2}$$

where $(w_k, \theta_k)_{k \geq 1}$ are the points of a Poisson point process with mean measure $\rho(dw)H(d\theta)$. H is some probability distribution on Θ, and ρ is a Lévy measure on $(0, \infty)$. We write $W \sim \mathrm{CRM}(\rho, H)$. A popular CRM is the generalized gamma process (GGP) (Hougaard, 1986; Brix, 1999) with Lévy measure

$$\rho_{\mathrm{GGP}}(dw; \sigma, \zeta) = \frac{1}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-\zeta w} dw \tag{3}$$

where $\sigma \in (0,1)$ and $\zeta \geq 0$ or $\sigma \leq 0$ and $\zeta > 0$. The GGP admits as special case the gamma process ($\sigma = 0, \zeta > 0$) and the stable process ($\sigma \in (0,1), \zeta = 0$). If

$$\int_0^\infty \rho(dw) = \infty \tag{4}$$

then the CRM is said to be infinite-activity: it has an infinite number of atoms and the weights satisfy $0 < W(\Theta) = \sum_{k=1}^\infty w_k < \infty$ a.s. We can therefore construct a random probability measure P by normalizing the CRM (Regazzini et al., 2003; Lijoi et al., 2007)

$$P = \frac{W}{W(\Theta)}. \tag{5}$$

We call P a normalized CRM (NCRM) and write $P \sim \mathrm{NCRM}(\rho, H)$. The Pitman-Yor process with parameters $\theta \geq 0$ and $\alpha \in [0,1)$ and distribution H, written $P \sim \mathrm{PY}(\alpha, \theta, H)$ admits a representation as a (mixture of)

CRMs (Pitman & Yor, 1997, Proposition 21). If $\theta, \alpha > 0$ it is a mixture of normalized generalized gamma processes

$$\eta \sim \text{Gamma}\left(\frac{\theta}{\alpha}, \frac{1}{\alpha}\right) \tag{6}$$

$$P \mid \eta \sim \text{NCRM}(\eta\, \rho_{\text{GGP}}(\cdot; \alpha, 1), H) \tag{7}$$

and for $\theta = 0$, it is a normalized stable process

$$P \sim \text{NCRM}(\rho_{\text{GGP}}(\cdot; \alpha, 0), H). \tag{8}$$

Although this representation is more complicated than the usual stick-breaking or urn constructions of the PY, it will be useful later on when we will discuss its asymptotic properties. The above construction essentially tells us that the PY has the same asymptotic properties as the normalized GGP for $\theta > 0$ and the stable process for $\theta = 0$.

3. Doubly regularly varying CRMs

3.1. General definition

We first introduce a few definitions on regularly varying functions (Bingham et al., 1989).

Definition 3.1 (Slowly varying function) *A positive function ℓ on $(0, \infty)$ is slowly varying at infinity if for all $c > 0$ $\ell(ct)/\ell(t) \to 1$ as $t \to +\infty$ Examples of slowly varying functions are constant functions, functions converging to a strictly positive constant, $(\log t)^a$ for any real a, etc.*

Definition 3.2 (Regularly varying function) *A positive function f on $(0, \infty)$ is said to be regularly varying at infinity with exponent $\xi \in \mathbb{R}$ if $f(x) = x^{\xi}\ell(x)$ where ℓ is a slowly varying function. Similarly, a function f is said to be regularly varying at 0 if $f(1/x)$ is regularly varying at infinity, that is $f(x) = x^{-\xi}\ell(1/x)$ for some $\xi \in \mathbb{R}$ and some slowly varying function ℓ.*

Informally, regularly varying functions with exponent $\xi \neq 0$ behave asymptotically similarly to a "pure" power-law function $g(x) = x^{\xi}$.

A homogeneous CRM W on Θ with mean measure $\rho(dw)H(d\theta)$ is said to be doubly regularly varying if its tail Lévy intensity

$$\overline{\rho}(x) = \int_x^{\infty} \rho(dw) \tag{9}$$

is regularly varying at 0 and ∞, that is

$$\overline{\rho}(x) \sim \begin{cases} x^{-\alpha}\ell_1(1/x) & \text{as } x \to 0 \\ x^{-\tau}\ell_2(x) & \text{as } x \to \infty \end{cases} \tag{10}$$

where $\alpha \in [0, 1]$, $\tau \geq 0$ and ℓ_1 and ℓ_2 are slowly varying functions. The CRM is said to be doubly power-law if it is doubly regularly varying with exponents $\alpha > 0$ and $\tau > 0$. Note that in this case, the CRM necessarily satisfies condition (4) and is therefore infinite activity.

3.2. Properties

In the following, let $w_{(1)} \geq w_{(2)} \geq \ldots$ denote the ordered weights of the CRM. The first proposition states that, if the CRM is regularly varying at 0 with exponent $\alpha > 0$, the small weights asymptotically scale as a power-law (up to a slowly varying function). Its proof is given in the Supplementary material.

Proposition 1 *A CRM, regularly varying at 0 with exponent $\alpha > 0$, satisfies*

$$w_{(k)} \sim k^{-1/\alpha}\ell_1^*(k) \quad \text{as } k \to \infty \tag{11}$$

where ℓ_1^* is a slowly varying function whose expression, which depends on ℓ_1 and α, is given in the supplementary material.

The next proposition states that, if the CRM is regularly varying at infinity with $\tau > 0$ and the scaling factor of the Lévy measure is large, the CRM has a power-law behavior for large weights.

Proposition 2 ***[Kevei & Mason (2014, Theorem 1.2)]*** *Consider a CRM with mean measure $\eta\rho(dw)H(d\theta)$, regularly varying at ∞ with $\tau > 0$. Then, for any $k_1, k_2 \geq 1$*

$$\frac{w_{(k_1+k_2)}^{\tau}}{w_{(k_1)}^{\tau}} \xrightarrow{d} \text{Beta}(k_1, k_2) \quad \text{as } \eta \to \infty. \tag{12}$$

Note that Equation (12) indicates a power-law behavior with exponent $1/\tau$, as for large η and $k \gg 1$, $w_{(k)} \simeq w_{(1)}k^{-1/\tau}$.

GGP and stable process. The GGP with parameter $\zeta > 0$ is regularly varying at 0 with exponent $\alpha = \max(0, \sigma)$. Hence, it satisfies Proposition (1). However, the exponential decay of the tails of the Lévy measure implies that it is not regularly varying at ∞. Large weights therefore decay exponentially fast. The stable process, which is a GGP with parameter $\zeta = 0$ and $\sigma \in (0, 1)$, is doubly regularly-varying with the same power-law exponent σ at 0 and ∞. Hence, it satisfies Proposition 1. Additionally, Pitman & Yor (1997, Proposition 8) showed that the result of Proposition 2 holds non-asymptotically for the stable process. In particular, for all $k \geq 1$, $w_{(k+1)}/w_{(k)} \sim \text{Beta}(k\sigma, 1)$.

In Section 3.3, we describe two general constructions for obtaining doubly regularly varying CRMs. Then we describe two specific processes with doubly regularly varying tail Lévy measure where one can flexibly tune both exponents. In the rest of the paper, we assume that the Lévy measure ρ is absolutely continuous with respect to the Lebesgue measure, and use the same notation for its density $\rho(dw) = \rho(w)dw$.

3.3. Construction of doubly regularly varying CRMs

Scaled-CRM. A first way of constructing a doubly regularly varying CRMs is to consider a CRM, regularly varying

at 0, and to divide its weights by independent and identically distributed (iid) random variables, whose cumulative density function (cdf) is also regularly varying at 0. More precisely, let

$$W = \sum_{k\geq 1} \frac{w_{0k}}{z_k}\delta_{\theta_k} \tag{13}$$

where $(z_1, z_2, \ldots)$ are strictly positive, continuous and iid random variables with cumulative density function $F_Z(z)$ and locally bounded probability density function $f_z(z)$, and

$$W_0 = \sum_{k\geq 1} w_{0k}\delta_{\theta_k} \sim \mathrm{CRM}(\rho_0, H)$$

where $\overline{\rho}_0(x)$ and $F_Z(z)$ are both regularly varying functions at 0, that is, for some $\alpha \in (0,1)$ and $\tau > \alpha$,

$$\overline{\rho}_0(x) \sim x^{-\alpha}\ell_1(1/x) \tag{14}$$
$$F_Z(z) \sim z^{\tau}\ell_2(1/z). \tag{15}$$

The random measure W is a CRM $W \sim \mathrm{CRM}(\rho, H)$ where

$$\rho(w) = \int_0^\infty z f_Z(z)\rho_0(wz)dz.$$

The next proposition shows that W is doubly regularly varying.

Proposition 3 *Assume that $\overline{\rho}_0$ and F_Z verify Equations* (14) *and* (15)*. Additionally, suppose $x\rho(x)$ and f_z are ultimately bounded and that there exists $\beta > \tau$, such that $\mu_\beta = \int_0^\infty w^\beta \rho_0(w)dw < \infty$. Then the CRM W defined by Equation* (13) *is doubly regularly varying, with*

$$\overline{\rho}(x) \sim \begin{cases} \mathbb{E}(Z^{-\alpha})x^{-\alpha}\ell_1(1/x) & \text{as } x \to 0 \\ \mu_\tau x^{-\tau}\ell_2(x) & \text{as } x \to \infty \end{cases}.$$

where Z is a random variable with cdf F_Z.

In Sections 3.4 and 3.5 we present two specific models constructed via a scaled GGP.

Discrete Mixture. An alternative to the scaled-CRM construction is to consider that the CRM is the sum of two CRMs, one regularly varying at 0 (hence infinite activity), the second one regularly varying at infinity. More precisely, consider the Lévy density

$$\rho(w) = \rho_0(w) + \beta f(w) \tag{16}$$

where ρ_0 is a Lévy measure, regularly varying at 0, and f is the probability density function of a random variable with power-law tails. That is ρ_0 satisfies (14) and

$$\int_x^\infty f(t)dt \sim x^{-\tau}\ell_2(x) \quad \text{as } x \to \infty.$$

If we additionally assume that $\overline{\rho}_0(x)$ has light tails at infinity (e.g. exponentially decaying tails), then the resulting CRM ρ is then doubly regularly varying and satisfies Equation (10). For example, one can take for ρ_0 the Lévy density (3) of a GGP, and for f the pdf of a Pareto, generalized Pareto or inverse gamma distribution.

3.4. Generalized BFRY process

Consider the Lévy density

$$\rho(w) = \frac{1}{\Gamma(1-\sigma)}w^{-1-\tau}\gamma(\tau-\sigma, cw) \tag{17}$$

where $\gamma(\kappa, x) = \int_0^x u^{\kappa-1}e^{-u}du$ is the lower incomplete gamma function and the parameters satisfy $\sigma \in (-\infty, 1)$, $\tau > \max(0, \sigma)$ and $c > 0$. We have

$$\overline{\rho}(x) \sim \frac{\Gamma(\tau-\sigma)}{\tau\Gamma(1-\sigma)}x^{-\tau} \tag{18}$$

as x tends to infinity and, for $\sigma > 0$,

$$\overline{\rho}(x) \sim \frac{c^{\tau-\sigma}}{\sigma(\tau-\sigma)\Gamma(1-\sigma)}x^{-\sigma} \tag{19}$$

as x tends to 0. When $\sigma \leq 0$, $\overline{\rho}(x)$ is a slowly varying function, with $\lim_{x\to 0}\overline{\rho}(x) = \infty$ if $\sigma = 0$ and $\lim_{x\to 0}\overline{\rho}(x) < \infty$ if $\sigma < 0$. $\overline{\rho}(x)$ therefore satisfies Equation (10) with $\alpha = \max(\sigma, 0)$. When $\sigma > 0$, it is doubly power-law with exponent $\sigma \in (0,1)$ and $\tau > 0$.

The Lévy density (17) admits the following latent construction as a scaled-GGP. Note that

$$\rho(w) = \frac{c^{\tau-\sigma}}{\tau}\int_0^1 z\,\rho_{\mathrm{GGP}}(wz; \sigma, c)f_Z(z)dz$$

where $f_Z(z) = \tau z^{\tau-1}$ is the probability density function of a $\mathrm{Beta}(\tau, 1)$ random variable. We therefore have the hierarchical construction. For $k \geq 1$,

$$w_k = \frac{w_{0k}}{\beta_k}, \beta_k \sim \mathrm{Beta}(\tau, 1).$$

where $(w_{0k})_{k\geq 1}$ are the points of a Poisson process with mean measure $c^{\tau-\sigma}/\tau\,\rho_{\mathrm{GGP}}(w_0; \sigma, c)dw_0$.

The process is somewhat related to, and can be seen as a natural generalization of the BFRY distribution (Pitman & Yor, 1997; Winkel, 2005; Bertoin et al., 2006). The name was coined by Devroye & James (2014) after the work of Bertoin, Fujita, Roynette and Yor. This distribution has recently found various applications in machine learning (Lee et al., 2016; 2017). Taking $c = 1$, $\tau \in (0,1)$ and $\sigma = \tau - 1 < 0$, we have

$$\rho(w) \propto w^{-\tau-1}(1-e^{-w})$$

which corresponds to the unnormalized pdf of a BFRY random variable. The BFRY random variable admits a representation as the ratio of a gamma and beta random variable, and the stochastic process introduced in this section, which admits a similar construction, can be seen as a natural generalization of the BFRY distribution, and we call this process a generalized BFRY (GBFRY) process. In Section 4 of the supplementary material, we provide more details on the BFRY distribution and its generalization.

3.5. Beta prime process

Consider the Lévy density

$$\rho(w) = \frac{\Gamma(\tau-\sigma)}{\Gamma(1-\sigma)} w^{-1-\sigma}(c+w)^{\sigma-\tau} \tag{20}$$

where $\sigma \in (-\infty, 1)$, $\tau > 0$ and $c > 0$. This density is an extension of the beta prime (BP) process, with an additional tuning parameter. This process was introduced by Broderick et al. (2015) and generalized by Broderick et al. (2018), as a conjugate prior for odds Bernoulli process. We have

$$\overline{\rho}(x) \sim \frac{\Gamma(\tau-\sigma)}{\tau\Gamma(1-\sigma)} x^{-\tau} \tag{21}$$

as x tends to infinity and, for $\sigma > 0$,

$$\overline{\rho}(x) \sim \frac{c^{\sigma-\tau}\Gamma(\tau-\sigma)}{\sigma\Gamma(1-\sigma)} x^{-\sigma} \tag{22}$$

as x tends to 0. When $\sigma \leq 0$, $\overline{\rho}(x)$ is a slowly varying function, with $\lim_{x\to 0}\overline{\rho}(x) = \infty$ if $\sigma = 0$ and $\lim_{x\to 0}\overline{\rho}(x) < \infty$ if $\sigma < 0$. $\overline{\rho}(x)$ therefore satisfies Equation (10) with $\alpha = \max(\sigma, 0)$. When $\sigma > 0$, it is doubly power-law with exponent $\sigma \in (0,1)$ and $\tau > 0$.

The BP process is related to the stable beta process (Teh & Gorur, 2009) with Lévy density

$$\frac{\alpha\Gamma(\tau-\sigma)}{c^{\tau}\Gamma(1-\sigma)} u^{-1-\sigma}(1-u)^{\tau-1}\mathbb{1}_{u\in(0,1)},$$

via the transformation $w = \frac{cu}{1-u}$. Similarly to the generalized BFRY model, the beta prime process can also be obtained via a scaled GGP. Note that

$$\rho(w) = \Gamma(\tau)c^{-\tau}\int_0^{\infty} y\,\rho_{\mathrm{GGP}}(wy;\sigma,c) f_Y(y)dy$$

where $f_Y(y) = \frac{c^{\tau}y^{\tau-1}e^{-cy}}{\Gamma(\tau)}$ is the density of a $\mathrm{Gamma}(\tau, c)$ random variable. We therefore have the following hierarchical construction, for $k \geq 1$

$$w_k = \frac{w_{0k}}{\gamma_k}, \quad \gamma_k \sim \mathrm{Gamma}(\tau, c)$$

where $(w_{0k})_{k\geq 1}$ are the points of a Poisson process with mean measure $c^{-\tau}\Gamma(\tau)\,\rho_{\mathrm{GGP}}(w_0;\sigma,1)dw_0$.

4. Normalized CRMs with double power-law

For some probability distribution H, Lévy measure ρ satisfying Equation (4) and $\eta > 0$, let

$$P = \frac{W}{W(\Theta)} \text{ where } W \sim \mathrm{CRM}(\eta\rho, H)$$

and for $i = 1, \ldots, n$, $X_i \mid P \overset{i.i.d.}{\sim} P$. As P is a.s. discrete, there will be repeated values within the sequence $(X_i)_{i\geq 1}$. Let $K_n \leq n$ be the number of unique values in $(X_1, \ldots, X_n)$, and $m_{n,(1)} \geq m_{n,(2)} \geq \ldots \geq m_{n,(K_n)}$ their ranked multiplicities. For $k = 1, \ldots, K_n$, denote $f_{n,(k)} = \frac{m_{n,(k)}}{n}$ the ranked frequencies.

4.1. Double power-law properties

The following theorem provides a precise formulation of Equation (1) and shows that the ranked frequencies have a double power-law regime when the CRM is doubly regularly varying with stricly positive exponents.

Theorem 1 *The ranked frequencies satisfy*

$$(f_{n,(1)}, f_{n,(2)}, \ldots) \to \left(\frac{w_{(1)}}{W(\Theta)}, \frac{w_{(2)}}{W(\Theta)}, \ldots\right) \tag{23}$$

almost surely as n tends to infinity. If the CRM is regularly varying at 0 with exponent $\alpha > 0$ we have

$$\frac{w_{(k)}}{W(\Theta)} \sim W(\Theta)^{-1}k^{-1/\alpha}\ell_1^*(k) \text{ as } k \to \infty. \tag{24}$$

If the CRM is regularly varying at ∞ with exponent $\tau > 0$ we have, for any $k_1, k_2 \geq 1$

$$\frac{w_{(k_1+k_2)}^{\tau}}{w_{(k_1)}^{\tau}} \xrightarrow{d} \mathrm{Beta}(k_1, k_2) \text{ as } \eta \to \infty. \tag{25}$$

Equation (23) in Theorem 1 follows from (Gnedin et al., 2007, Proposition 26). Equations (24) and (25) follow from Propositions 1 and 2. Instead of expressing the power-law properties in terms of the ranked frequencies, we can alternatively look at the asymptotic behavior of the number $K_{n,j}$ of elements with multiplicity $j \geq 1$, defined by

$$K_{n,j} = \sum_{k=1}^{K_n} \mathbb{1}_{m_{n,(k)}=j}. \tag{26}$$

Let $p_{n,j} = \frac{K_{n,j}}{K_n}$. Note that $\sum_{j\geq 1} p_{n,j} = 1$. The following is a corollary of Equation (24). It follows from Proposition 23 and Corollary 21 in (Gnedin et al., 2007).

Corollary 2 *If the CRM is regularly varying at 0 with exponent α, we have*

$$p_{n,j} \to p_j \text{ a.s. as } n \to \infty \tag{27}$$

where

$$p_j = \frac{\sigma\Gamma(j-\alpha)}{j!\Gamma(1-\alpha)} \sim \frac{\alpha}{\Gamma(1-\alpha)}\frac{1}{j^{1+\alpha}} \quad \textit{for large } j.$$

Figure 2 shows some illustration of these empirical results for the GBFRY model.

Remark 1 *The GGP with parameter $\sigma > 0$ is regularly varying at 0, but not at infinity. Hence, the normalized GGP with $\zeta > 0$ and the related Pitman-Yor process with $\theta > 0$ satisfy Equation* (24) *and* (27) *but not* (25)*, due to the exponentially decaying tails of the Lévy measure of the GGP. The normalized GGP with $\zeta = 0$, which is the same as the Pitman-Yor with $\theta = 0$, satisfies both equations, but with the same exponent $\sigma \in (0, 1)$, lacking the flexibility of the three models presented in Section 3.*

4.2. Posterior Inference

In this subsection, we briefly discuss the inference procedure for estimating the parameters of the normalized CRMs we introduced in Section 3. Additional details are provided in the supplementary material. Assume that the Lévy measure ρ is parameterised by some parameters ϕ we want to estimate, in particular the two power-law exponents. We write $\rho(w;\phi)$ to emphasize this, and let $p(\phi)$ be the prior density. The objective is to approximate the posterior density of the parameters given the ranked counts $p(\phi \mid (m_{n,(k)})_{k=1,\dots,K_n})$.

Parametrisation. Since we are working with normalized CRMs, multiplying W by any positive constant $\xi > 0$ gives the same random probability measure P. In particular, the normalized CRMs with Lévy densities $\rho(w)$ and $\widetilde{\rho}(w) = \xi\rho(\xi w)$ have the same distribution. To avoid overparameterisation we set the parameter $c = 1$ in the GBFRY and BP processes, and estimate the parameter $\phi = (\sigma, \tau, \eta)$.

We introduce a latent variable $U \mid W \sim \mathrm{Gamma}(n, W(\Theta))$. Using Proposition 3 of James et al. (2009) (see also Pitman (2003)) and Equation (2.2) of Pitman (2006)), the joint density is written as

$$p\left((m_{n,(k)})_{k=1,\dots,K_n}, u, \phi\right) \propto p(\phi)u^{n-1}e^{-\psi(u;\phi)}\prod_{k=1}^{K_n}\kappa(m_{n,(k)}, u;\phi) \tag{28}$$

where the normalizing constant only depends on n and the ranked counts, and

$$\psi(t;\phi) = \eta\int_0^\infty (1-e^{-tw})\rho(w;\phi)dw, \tag{29}$$

$$\kappa(m,t;\phi) = \eta\int_0^\infty w^m e^{-tw}\rho(w;\phi)dw. \tag{30}$$

If ψ and κ have analytic forms, one can derive a MCMC sampler to approximate the posterior by successively updating U and ϕ . Unfortunately, this is not the case for our models. For instance, in the generalized BFRY process case, we have

$$\psi(t;\phi) = \frac{\eta}{\sigma}\int_0^c ((y+t)^\sigma - y^\sigma)y^{\tau-\sigma-1}dy \tag{31}$$

$$\kappa(m,t;\phi) = \frac{\eta\Gamma(m-\sigma)}{\Gamma(1-\sigma)}\int_0^c \frac{y^{\tau-\sigma-1}}{(y+t)^{m-\sigma}}dy. \tag{32}$$

We may resort to a numerical integration algorithm to approximate ψ as only one evaluation of this function is needed at each iteration. We could do the same for κ. However, this would require K_n numerical integrations at each step of the MCMC sampler, which is computationally prohibitive for large K_n. Instead, building on the construction of the generalized BFRY as a scaled generalized gamma process described in Section 4, we introduce a set of latent variables $Y = (Y_k)_{j=1,\dots,K_n}$ whose conditional density is written as

$$p(y_k|u, (m_{n,(k)})_{k=1,\dots,K_n}) \propto \frac{y_k^{\tau-\sigma-1}}{(y_k+u)^{m_{n,(k)}-\sigma}}\mathbb{1}_{0<y_k<c},$$

and this gives the joint density

$$p\left((m_{n,(k)})_{k=1,\dots,K_n}, u, y, \phi\right) \propto p(\phi)u^{n-1}e^{-\psi(u;\phi)}\prod_{k=1}^{K_n}\frac{\eta\Gamma(m_{n,(k)}-\sigma)y_k^{\tau-\sigma-1}}{\Gamma(1-\sigma)(y_k+u)^{m_{n,(k)}-\sigma}}.$$

where the normalizing constant only depends on n and the ranked counts. Then we can alternate between updating ϕ and U via Metropolis-Hastings and updating Y via Hamiltonian Monte-Carlo (HMC) (Duane et al., 1987; Neal et al., 2011) to estimate the posterior. See the supplementary material for more details. A similar strategy can be used for the beta prime process.

5. Experiments

We run the algorithms described in Section 4.2 for the GBFRY and BP models. We fix $c = 1$ to avoid overparameterisation, as explained in Section 4.2. We use standard normal prior on $\log\eta$, $\log\tau$ and $\mathrm{logit}\,\sigma$. The proposed models are compared to the normalized GGP with the same priors on η and σ and fixed $\zeta = 1$, and the PY process with standard normal prior on $\log\theta$ and $\mathrm{logit}\,\alpha$. We also considered the discrete mixture construction described in Section 3.3 with ρ_0 taken to be a GGP, and f a Pareto or generalized Pareto distribution. While we were able to recover the parameters on simulated data, this model was underperforming on real data, and results are not reported. The codes to replicate our experiments can be found in `https://github.com/OxCSML-BayesNP/doublepowerlaw`.

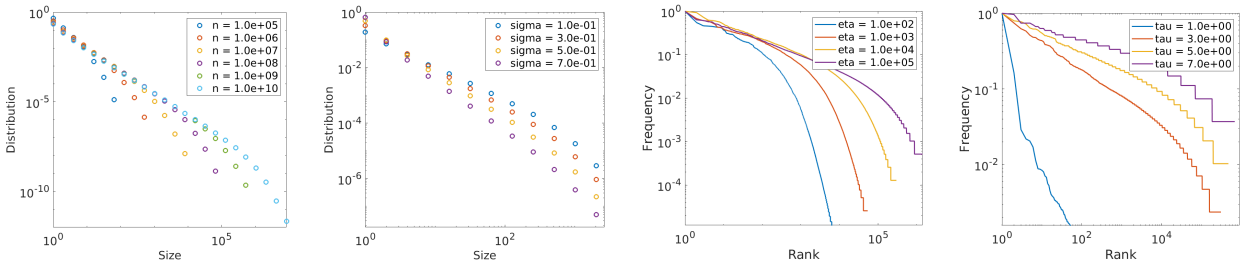


Figure 2. Simulated data from the normalized GBFRY model. Proportion of clusters of a given size for (First) $\eta = 4000, \tau = 3, \sigma = 0.2$ with varying n and (Second) $\eta = 4000, \tau = 3, n = 10^7$ with varying of σ. Ordered frequencies, normalized by the largest one, for (Third) $n = 10^6, \sigma = 0.2, \tau = 3$ with varying η and (Fourth) $n = 10^6, \sigma = 0.2, \eta = 50000$ with varying τ .

We stress that the objective is to show that the proposed models provide a better fit than alternative models, not to test the double power-law assumption.

5.1. Synthetic data

We sample simulated datasets from the normalized GBFRY and the BP with parameters $\sigma = 0.1$, $\tau = 2$, $c = 1$ and $\eta = 4000$. We run the MCMC algorithm described in Section 4.2 with 100 000 iterations. The 95% credible intervals are $\sigma \in (0.09, 0.12)$, $\tau \in (1.6, 2.2)$ for the BFRY and $\sigma \in (0.08, 0.11)$, $\tau \in (1.8, 2.3)$ for the BP, indicating that the MCMC recovers true parameters. Trace plots are reported in the supplementary material.

5.2. Real data

We then consider five real datasets, four of which are word frequencies in natural languages, and the last is the out-degree distribution of a Twitter network. We first provide a description of the different datasets.

Word frequencies. Each dataset is composed of n words $X_1, .., X_n$, with $K_n \leq n$ unique words. The counts $m_{n,(k)}$ represent the number of occurences of the kth most frequent word in the dataset. The first dataset is the written dataset of the American National Corpus[2] (ANC), composed of about 18 million word occurences and 300 000 unique words. The second and third datasets are the words of a collection of most popular English books and French books, downloaded from the Project Gutenberg[3]. The English books dataset is composed of about 3 million words and 71 000 unique words, the French books of about 7 million words and around 135 000 unique words. The fourth dataset represents the words of a thousand papers from the NIPS conference. It contains about 2 million word occurences and 68 000 unique words.

Twitter network. We consider a rank-1 edge-exchangeable model for directed multigraphs (Crane & Dempsey, 2018; Cai et al., 2016). In this case, the atoms of W represent the nodes of the graph, and each directed edge (X_i, Y_i) from node X_i to node Y_i is sampled independently from $P \times P$. Note that when P is a Pitman-Yor process, the associated model corresponds to the urn-based Hollywood model of Crane & Dempsey (2018). Here, we only consider the out-degree distribution. Therefore, n represents the number of directed edges and $X_1, .., X_n$ the source nodes of the directed edges sampled from the normalized CRM P. $m_{n,(k)}$ corresponds to the kth largest out-degree in the network. We consider a subset of 25 millions tweets of August 2009 from Twitter (Yang & Leskovec, 2011). We construct a directed multigraph by adding an edge (X_i, Y_i) whenever user X_i mentions user Y_i (with @) in tweet i. The resulting graph contains about 4 millions edges and 300 000 source nodes.

[2]http://www.anc.org/data/anc-second-release/frequency-data/
[3]http://www.gutenberg.org/

Table 1. Average Kolmogorov-Smirnov divergence between the data and the posterior predictive. Lower is better.

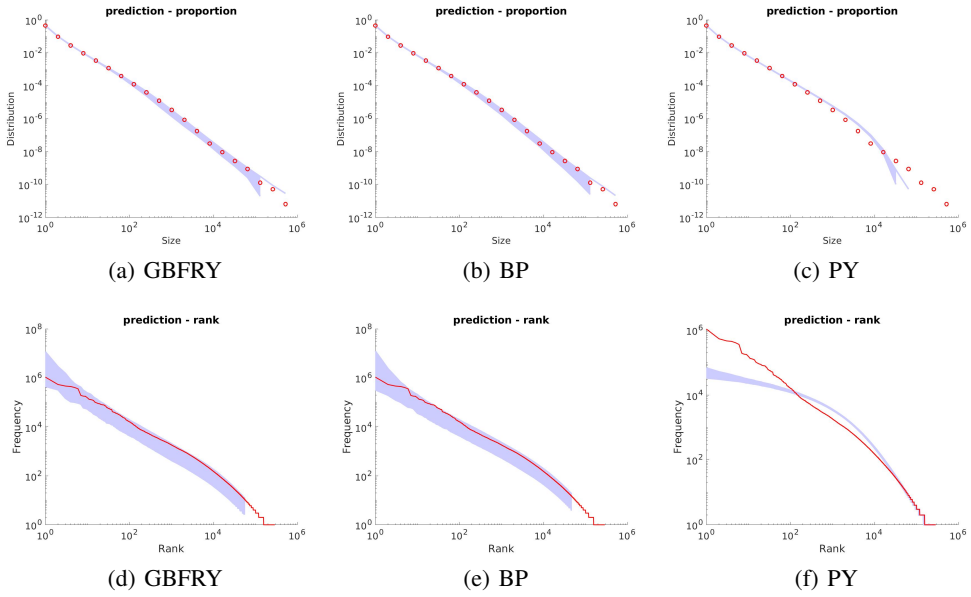
Dataset	GBFRY	Beta Prime	GGP	PY
Englishbooks	0.072	**0.041**	0.12	0.12
Frenchbooks	0.064	**0.032**	0.11	0.11
NIPS1000	**0.041**	0.081	0.08	0.059
ANC	**0.033**	0.034	0.082	0.081
Twitter	0.10	**0.047**	0.25	0.26

Results

For each of the four models and each dataset, we approximate the posterior distribution of the parameters ϕ of the Lévy measure, and sample new datasets from the posterior predictive. The 95% credible intervals of the posterior predictive for the proportion of occurences and ranked frequencies are reported in Figure 3 for the ANC dataset (plots for the other datasets are given in the supplementary material). As the results for the normalized GGP and PY are almost identical, we only show the plot for the PY model. As can clearly be seen from the posterior predictive plots, all models provide a good fit for low frequencies. However, the

Table 2. 95% posterior credible intervals of the power-law exponents.

	GBFRY		Beta Prime		GGP	PY
Dataset	σ	τ	σ	τ	σ	σ
Englishbooks	(0.351, 0.362)	(0.912, 0.980)	(0.345, 0.358)	(0.974, 1.078)	(0.416, 0.423)	(0.416, 0.423)
Frenchbooks	(0.368, 0.375)	(0.967, 1.039)	(0.363, 0.371)	(1.04, 1.175)	(0.407, 0.412)	(0.407, 0.412)
NIPS1000	(0.538, 0.545)	(1.338, 1.906)	(0.538, 0.545)	(1.541, 2.286)	(0.542, 0.548)	(0.542, 0.549)
ANC	(0.433, 0.438)	(0.998, 1.055)	(0.431, 0.436)	(1.09, 1.17)	(0.461, 0.465)	(0.461, 0.465)
Twitter	(0.282, 0.287)	(1.590, 1.600)	(0.099, 0.116)	(1.336, 1.411)	(0.272, 0.277)	(0.272, 0.277)



(a) GBFRY (b) BP (c) PY

(d) GBFRY (e) BP (f) PY

Figure 3. Results on the ANC dataset: 95% credible interval of the posterior predictive in blue, data in red. (Top) Proportion of occurences of a given size. (Bottom) Ranked occurences.

PY model (and similar the normalized GGP) fail to capture the power-law behavior for large frequencies. This behavior is better captured by the GBFRY and BP models. To illustrate quantitatively the comparison, we compute the average reweighted Kolmogorov-Smirnov divergence (Clauset et al., 2009) between the true data and the posterior predictive for each model, and report the results in Table 1. Finally, we report in Table 2 the 95% credible intervals of the parameters for each model and dataset. We can remark that to the exception of the NIPS dataset, we recover the Zipfian exponent $\tau = 1$ for large frequencies in text datasets.

6. Conclusion

In this paper we presented a novel class of random measures with double power-law behavior. We focused on the case of iid sampling from a normalized completely random measure. More generally, one could build on this class of models for other CRM-based constructions. In particular, it would be interesting to explore the asymptotic degree distribution when such models are used for random graph models based on exchangeable point processes (Caron & Fox, 2017). Building hierarchical versions of such models as for the hierarchical Pitman-Yor process (Teh, 2006) would also be of interest. Finally, it would be useful to explore the connections between the models presented here and the two-stage urn process suggested by Gerlach & Altmann (2013) and investigate if other urn schemes could be derived that provably exhibit a double power-law behavior.

Acknowledgments The authors thank Valerio Perrone for providing the NIPS dataset. JL and FC's research leading to these results has received funding from European Research Council under the European Unions Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071 and from EPSRC under grant EP/P026753/1. FC acknowledges support from the Alan Turing Institute under EPSRC grant EP/N510129/1. JL acknowledges support from IITP grant funded by the Korea government(MSIT) (No.2017-0-01779, XAI) and Samsung Research Funding & Incubation Center under project number SRFC-IT1702-15.

References

Bertoin, J., Fujita, T., Roynette, B., and Yor, M. On a particular class of self-decomposable random variables: the durations of Bessel excursions straddling independent exponential times. 2006.

Bild, D. R., Liu, Y., Dick, R. P., Mao, Z. M., and Wallach, D. S. Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology (TOIT)*, 15(1):4, 2015.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. *Regular variation*, volume 27. Cambridge university press, 1989.

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

Broderick, T., Mackey, L., Paisley, J., and Jordan, M. I. Combinatorial clustering and the beta negative binomial process. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):290–306, 2015.

Broderick, T., Wilson, A. C., and Jordan, M. I. Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli*, 24(4B):3181–3221, 2018.

Cai, D., Campbell, T., and Broderick, T. Edge-exchangeable graphs and sparsity. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4249–4257. Curran Associates, Inc., 2016.

Caron, F. Bayesian nonparametric models for bipartite graphs. In *Advances in neural information processing systems*, 2012.

Caron, F. and Fox, E. B. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1295–1366, 2017.

Clauset, A., Shalizi, C. R., and Newman, M. E. Power-law distributions in empirical data. *SIAM review*, 51(4): 661–703, 2009.

Crane, H. and Dempsey, W. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, 113(523):1311–1326, 2018.

Csányi, G. and Szendrői, B. Structure of a large social network. *Physical Review E*, 69(3):036131, 2004.

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

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Ferrer i Cancho, R. and Solé, R. V. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8(3):165–173, 2001.

Font-Clos, F., Boleda, G., and Corral, A. A scaling law beyond Zipf's law and its relation to Heaps' law. *New Journal of Physics*, 15(9):093033, 2013.

Gerlach, M. and Altmann, E. G. Stochastic model for the vocabulary growth in natural languages. *Physical Review X*, 3(2):021006, 2013.

Gnedin, A., Hansen, B., and Pitman, J. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability surveys*, 4:146–171, 2007.

Goldwater, S., Johnson, M., and Griffiths, T. L. Interpolating between types and tokens by estimating power-law generators. In *Advances in neural information processing systems*, pp. 459–466, 2006.

Hougaard, P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396, 1986.

James, L. F., Lijoi, A., and Prünster, I. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.

Kevei, P. and Mason, D. M. The limit distribution of ratios of jumps and sums of jumps of subordinators. *ALEA*, 11 (2):631–642, 2014.

Kingman, J. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

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

Lee, J., Heaukulani, C., Ghahramani, Z., James, L. F., and Choi, S. Bayesian inference on random simple graphs with power law degree distributions. In *International Conference on Machine Learning*, pp. 2004–2013, 2017.

Lijoi, A. and Prünster, I. Models beyond the Dirichlet process. *Bayesian nonparametrics*, 28(80):3, 2010.

Lijoi, A., Mena, R. H., and Prünster, I. Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007.

Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.

Mochihashi, D., Yamada, T., and Ueda, N. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pp. 100–108. Association for Computational Linguistics, 2009.

Montemurro, M. A. Beyond the Zipf Mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3-4):567–578, 2001.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

Paleari, S., Redondi, R., and Malighetti, P. A comparative study of airport connectivity in China, Europe and US: which network provides the best service to passengers? *Transportation Research Part E: Logistics and Transportation Review*, 46(2):198–210, 2010.

Pitman, J. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2): 145–158, 1995.

Pitman, J. Poisson-Kingman partitions. *Lecture Notes-Monograph Series*, pp. 1–34, 2003.

Pitman, J. *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002*. Springer, 2006.

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

Regazzini, E., Lijoi, A., and Prünster, I. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2): 560–585, 2003.

Sato, I. and Nakagawa, H. Topic models with power-law using pitman-yor process. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 673–682. ACM, 2010.

Seshadri, M., Machiraju, S., Sridharan, A., Bolot, J., Faloutsos, C., and Leskovec, J. Mobile call graphs: beyond power-law and lognormal distributions. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 596–604. ACM, 2008.

Sudderth, E. B. and Jordan, M. I. Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in neural information processing systems*, pp. 1585–1592, 2009.

Teh, Y. W. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 985–992. Association for Computational Linguistics, 2006.

Teh, Y. W. and Gorur, D. Indian buffet processes with power-law behavior. In *Advances in neural information processing systems*, pp. 1838–1846, 2009.

Winkel, M. Electronic foreign-exchange markets and passage events of independent subordinators. *Journal of applied probability*, 42(1):138–152, 2005.

Wood, F., Archambeau, C., Gasthaus, J., James, L., and Teh, Y. W. A stochastic memoizer for sequence data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1129–1136. ACM, 2009.

Yang, J. and Leskovec, J. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 177–186. ACM, 2011.

Zipf, G. The psycho-biology of language: an introduction to dynamic philology. 1935.