Powered Chinese Restaurant Process - Controlling the "Rich-Get-Richer" Assumption in Bayesian Clustering

Anonymous Author(s)

ABSTRACT

One of the most used priors in Bayesian clustering is the Dirichlet prior. It can be expressed as a Chinese Restaurant Process. This process allows nonparametric estimation of the number of clusters when partitioning datasets. Its key feature is the "rich-get-richer" property, which assumes a cluster has an *a priori* probability to get chosen linearly dependent on population. In this paper, we show that such prior is not always the best choice to model data. We derive the Powered Chinese Restaurant process from a modified version of the Dirichlet-Multinomial distribution to answer this problem. We then develop some of its fundamental properties (expected number of clusters, convergence). Unlike state-of-the-art efforts in this direction, this new formulation allows for direct control of the importance of the "rich-get-richer" prior.

CCS CONCEPTS

• **Mathematics of computing** → *Discrete mathematics*; *Stochastic processes*; *Bayesian computation*; *Cluster analysis.*

KEYWORDS

Chinese restaurant process, Rich-get-richer, Dirichlet process, Clustering, Bayesian prior

ACM Reference Format:

Anonymous Author(s). 2018. Powered Chinese Restaurant Process - Controlling the "Rich-Get-Richer" Assumption in Bayesian Clustering. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The notion of clustering has been initially introduced by anthropologists Driver and Kroeber in 1932 [15] in the classification of human psychological traits. It has been later used successfully in a broad range of applications, ranging from scientific research to data compression, marketing, and medicine. Over the past decades, it also became a central problem in machine learning[1].

The Bayesian clustering approach received broad attention in the last years. A non-exhaustive list of application includes medicine, [12], natural language processing [4, 33], genetics [18, 22, 25], recommender systems [1, 9, 24], sociology [5, 11], etc. The key idea is to simulate a corpus of independent observations by drawing them from a set of latent variables (clusters). Those clusters are each associated with a probability distribution on the observations, whose parameters are drawn from a prior distribution, as we will formulate mathematically later. Now, an often desirable property of Bayesian models is to make them nonparametric. In our case, it means that both the number of clusters and their associated distributions are inferred. A very popular prior on clusters distributions that allows this is the Dirichlet process. It incorporates a chance for a new cluster to be created in the prior probability of a distribution (often when an observation is not likely to be explained by existing clusters). Otherwise, the observation is associated *a priori* with an existing cluster with a probability proportional to that cluster's population. Note that the model the prior is associated with might use of this prior information, but might as well ignore it by design.

However, the Dirichlet process (and the related Pitman-Yor process) prior comes with a hypothesis on the way observations are allocated to various clusters: the *rich-get-richer* property [6]. As stated before, a new observation *a priori* belongs to a cluster with a probability proportional to the number of observations already present in the cluster; large clusters have a greater chance to get associated with new observations. While this can be a relevant property in some cases, it implies a strong assumption on the way data is generated. It has already been pointed out [30] that there is a need for more flexible priors.

This is particularly relevant in the case of imbalanced data and scale-dependent clustering. A cluster made of fewer entities might go unnoticed due to a rich-get-richer prior. As an example of the imbalance problem, consider a case where data is treated sequentially –which is often the case when it comes to Dirichlet process prior. The first observation from a new cluster would then have a much larger *a priori* probability to belong to a populated but irrelevant cluster, than to open a new one (this probability decreases as $\frac{1}{N_{obs}}$). This typically happens when sampling topics from news streams [29, 32]. In the case of scale-dependent clustering, a similar problem arises. Consider clustering people pinpointed on a map. Tiny clusters (at the scale of cities, for instance) might go unnoticed at larger scales (countries, for instance). To spot city clusters on a world map, the "rich-get-richer" assumption becomes irrelevant and a "rich-get-no-richer" prior would be preferred (such as [29]); the optimal solution might as well be in-between these two priors, as in Fig.4. We aim to design a method to bridge the variety of possible priors between the Dirichlet process and the Uniform process in a continuous fashion.

Little effort has been put into exploring alternative forms of priors for nonparameteric Bayesian modeling. In the present work, we offer to address this problem by deriving a more general form of the Dirichlet process that explicitly controls the importance of

[1] As an illustration, scraping Google Scholar shows that the yearly number of publications containing the keyword "Clustering" averages to 250.000.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Woodstock '18, June 03–05, 2018, Woodstock, NY

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

the "rich-get-richer" assumption. Explicitly, we derive the Powered Chinese Restaurant Process (PCRP) that allows control of the "rich-get-richer" property while generalizing state-of-the-art works. We show that controlling the "rich-get-richer" prior of simple models yields better results on synthetic and real-world datasets.

2 BACKGROUND

2.1 Motivation

This work is motivated by the need to control the "rich-get-richer" assumption's importance in Dirichlet process (DP) priors. The "rich-get-richer" property of the DP may not always be the most suitable prior for modeling a given dataset. The usual motivation for using a DP prior is that a new observation has a probability of being assigned to any cluster proportional to its population in the absence of external information (such as inter-points distance in case of spatial clustering, for instance). However, this assumption might be flawed (see Introduction).

Most state-of-the-art works rely on tuning a parameter α (see Eq.1) to get the "right" number of clusters (this parameter shifts the distribution of the number of clusters as $\mathbb{E}(K|N) \propto \alpha \log N$ with K the number of clusters and N the number of observations). However, we argue this is a bad practice in some cases. Imagine sampling topics in a news stream: there is no specific reason for topics to appear at a rate $\alpha \log N$ as in the regular DP. The logarithmic dependence on N cannot be tuned using α. Such prior is then unfit to describe the data correctly as the number of observations grows; worst, it can lead the model it is coupled with on the wrong track. Moreover, when considering observations streams [29, 32], there is usually no specific *a priori* reason for a new observation to belong to a cluster with a probability depending linearly on its size, as in the regular Dirichlet and Pitman-Yor processes. To alleviate those assumptions, we develop a more general form of the DP process allowing a natural control of the "rich-get-richer" property.

2.2 Previous works

2.2.1 Dirichlet process. A well-known metaphor for the Dirichlet process is referred to as "Chinese restaurant". The corresponding process is named "Chinese Restaurant Process" (CRP). It can be illustrated as follows: if a n^{th} client arrives in a Chinese restaurant, she will sit at one of the K already occupied table with a probability proportional to the number of persons already sat at this table. She can also go to a new table in the restaurant and be the first client to sit there with a probability inversely proportional to the total number of clients already sat at other tables. It can be written formally as:

$$CRP(C_i = c|\alpha, C_1, C_2, ..., C_{i-1}) = \begin{cases} \frac{N_c}{\alpha+N} & \text{if c = 1, 2, ..., K} \\ \frac{\alpha}{\alpha+N} & \text{if c = K+1} \end{cases} \quad (1)$$

Where c is the cluster chosen by the i^{th} customer, N_k is the population of cluster k, K is the number of already occupied tables and α the concentration parameter. When the number of clients goes to infinity, this process is equivalent to a draw from a Dirichlet distribution over an infinite number of clusters with an identical initial probability to get chosen proportional to α. The form of Eq.1 is helpful to understand the underlying dynamics of the process and the contribution of seminal works we will detail now. It can be shown that the expected number of clusters after N observations evolves as $\log N$ [2].

The two best-known variations of the regular Dirichlet process that address the "rich-get-richer" property control are the seminal Pitman-Yor process and the Uniform process. Each of them can be expressed in a similar form as Eq.1.

2.2.2 Pitman-Yor process. Following the Chinese Restaurant process metaphor, the Pitman-Yor process [16, 23] proposed to incorporate a *discount* when a client opens a new table. Mathematically, the process can be formulated as:

$$PY(C_i = c|\alpha, \beta, C_1, C_2, ..., C_{i-1}) = \begin{cases} \frac{N_c-\beta}{\alpha+N} & \text{if c = 1, 2, ..., K} \\ \frac{\alpha+\beta K}{\alpha+N} & \text{if c = K+1} \end{cases} \quad (2)$$

The introduction of the parameter $\beta > 0$ increases the probability of creating new clusters. A table with a low number of customers has significantly less chances to gain new ones, while the probability of opening a new table increases significantly. It can be shown that the number of tables evolves with the number of clients N as N^{β} [10, 28]. However, this process does not control the arguable "rich-get-richer" hypothesis [30], since the relation to the population of a table remains linear; it only shifts this dependence of a value β. It makes so by creating clusters based on the number of existing clusters and the total number of observations, but not according to the population of already existing clusters. Those play the same role in the Pitman-Yor process as in the DP. The Pitman-Yor process thus comes with two limitations. First, since $\beta > 0$, it cannot modify the process to generate fewer clusters. Second, the discount parameter does not modify the linear dependence on previous observations for cluster allocations — rich still get richer; the prior is as peaky on large clusters as before. The present work offers to address those two limitations.

Another work based on the Generalized Gamma Process proposed a similar discount idea to increase the probability of opening new clusters in [20]. The proposed prior ([20]-Eq.4) modifies a cluster's probability to get chosen by subtracting a constant term to each cluster's population. Thus, the rich-get-richer property is not alleviated in their approach either, since the dependence on cluster's population is still linear. As for the PY process, this formulation only allows to increase the number of clusters and does not alleviate the "rich-get-richer" hypothesis.

2.2.3 Uniform process. Another process that aims at breaking the "rich-get-richer" property is the Uniform process. It has been used in some occasions [18, 25] without proper definition. More recently, it has been formalized and studied in comparison with the regular Dirichlet and Pitman-Yor processes [29]. It can be written as follows:

$$UP(C_i = c|\alpha, C_1, C_2, ..., C_{i-1}) = \begin{cases} \frac{1}{\alpha+K} & \text{if c = 1, 2, ..., K} \\ \frac{\alpha}{\alpha+K} & \text{if c = K+1} \end{cases} \quad (3)$$

This formulation completely gets rid of the "rich-get-richer" property. The probability of a new client joining an occupied table is a uniform distribution over the number of occupied tables; it does not depend on the tables' population. In [29], it has been shown that the expected number of tables evolves with N as $\sqrt{N}$. Removing the "rich-get-richer" property leads to a flat prior. As we show later, our

formulation allows to retrieve such flat priors and thus generalizes the Uniform Process.

2.3 Contributions

In the present work, we derive the Powered Chinese Restaurant Process (PCRP) that allows controlling the "rich-get-richer" property while generalizing state-of-the-art works — rich-get-no-richer (Uniform process), rich-get-less-richer, "rich-get-richer" (DP), and rich-get-more-richer. Doing so, we define the Powered Dirichlet-Multinomial distribution. We detail some key-properties of the Powered Dirichlet Process (convergence, expected number of clusters). Finally, we show that controlling the "rich-get-richer" prior of simple models yields better results on synthetic and real-world datasets.

3 THE MODEL

3.1 The Dirichlet-Multinomial distribution

We recall:

$$Dir(\vec{p}|\vec{\alpha}) = \frac{\prod_k p_k^{\alpha_k - 1}}{B(\vec{\alpha})} \qquad Mult(\vec{N}|N,\vec{p}) = \frac{\Gamma(\sum_k N_k + 1)}{\prod_k \Gamma(N_k+1)} \prod_k p_k^{N_k} \tag{4}$$

With $\vec{N} = (N_1, N_2, ..., N_K)$ where N_k is the integer number of draws assigned to cluster k, $N = \sum_k N_k$ the total number of draws, $\Gamma(x) = (x-1)!$ and $B(\vec{x}) = \prod_k \Gamma(x_k)/\Gamma(\sum_k x_k)$.

The regular Dirichlet process can be derived from the Dirichlet-Multinomial distribution. The Dirichlet-Multinomial distribution is defined as follows:

$$\begin{aligned} p(\vec{N}|\vec{\alpha}, n) &= \int_{\vec{p}} p(\vec{N}|\vec{p}, n) p(\vec{p}|\vec{\alpha}) d\vec{p} \\ &= \frac{(n!)\Gamma(\sum_k \alpha_k)}{\Gamma(n + \sum_k \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha_k)}{(N_k!)\Gamma(\alpha_k)} \\ \text{where } \vec{p} \sim Dir(\vec{p}|\vec{\alpha}) &; \vec{N} \sim Mult(\vec{N}|n,\vec{p}) \end{aligned} \tag{5}$$

In Eq.5, we sample n values over a space of K distinct clusters each with probability $\vec{p} = (p_1, p_2, ..., p_K)$, using a Dirichlet prior with parameter $\vec{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)$. As we will show in the next section, we can derive the Dirichlet process equation by iterating the Dirichlet-Multinomial distribution. More precisely, one has to compute a new observation's conditional distribution to belong to any cluster given the allocation of all the previous random variables when $K \to \infty$.

3.2 Powered conditional Dirichlet prior

In the derivation of the standard Dirichlet-Multinomial posterior predictive, one considers a categorical distribution coupled with a Dirichlet prior on its parameter $\vec{p}$. Usually, this prior is linearly dependent on previous draws from the distribution. We propose to modify this assumption by using a Dirichlet prior that depends on the history of draws as:

$$Dir_r(\vec{p}|\vec{\alpha},\vec{N}) = \frac{1}{B(\vec{\alpha} + \vec{N}^r)} \prod_k p_k^{\alpha_k + N_k^r - 1} \tag{6}$$

In Eq.6, the vector $\vec{N}^r$ shifts the parameter $\vec{\alpha}$ according to the count of draws allocated to each cluster k up to the nth draw. The parameter $r \in \mathbb{R}^+$ controls the intensity of this shift for each entry of $\vec{X}$.

We demonstrate that the Powered Dirichlet distribution is a conjugate prior of the Multinomial distribution, by writing Eq.6 as:

$$\begin{aligned} Dir_r(\vec{p}|\vec{\alpha},\vec{N}) &= \frac{1}{B(\vec{\alpha} + \vec{N}^r)} \prod_k p_k^{\alpha_k - 1} \prod_k p_k^{N_k^r} \\ &\overset{\text{Eqs.4}}{=} \frac{B(\vec{\alpha}) \prod_k N_k^r!}{B(\vec{\alpha} + \vec{N}^r)(\sum_k N_k^r)!} Dir(\vec{p}|\vec{\alpha}) Mult(\vec{N}^r | \sum_k N_k^r, \vec{p}) \end{aligned} \tag{7}$$

where the prior on vector $\vec{N}$ is a regular Multinomial distribution of parameter $N = \sum_k N_k^r$. Note that for certain values of r, the vector $\vec{N}^r$ might not be made of integer values; the resulting Multinomial prior on $\vec{N}^r$ must then be expressed in terms of Γ functions (see Eq.4) to be valid for $\vec{N}^r \in \mathbb{R}^{|\vec{N}|}$. Distributions of non-integer counts are not new in the literature [7, 19, 21] and are essentially allowed by the generalized definition of the factorial function in terms of the gamma function. When $r = 1$, we recover the standard Dirichlet-Multinomial prior on $\vec{p}$ for the n^{th} draw; the history of draws $\vec{N}$ can be expressed as the result of N independent draws of equal probability $\vec{p}$. When $r \neq 1$, the prior on $\vec{N}$ is sampled from a Multinomial distribution in which the number of samples drawn depends on r as $\sum_k N_k^r$. For instance, let $\vec{N} = (1, 2)$ and $r = 2$: the resulting powered conditional Dirichlet prior would then be sampled from a Multinomial distribution $Mult(\vec{N} = (1, 4)|N = 5, \vec{p} = (p, 1-p))$.

3.3 Posterior predictive

We now derive the posterior distribution for the n^{th} draw to belong to a cluster c given all previous draws. We assume that $\vec{C_-}$ represents all previous realizations up to $n-1$, that is, the cluster to which each previous draw has been associated. For simplicity of notation, we define the population of a cluster k at time $n-1$ as $N_k = |\{C_i | i = k\}_{i=1,2,...,n-1}|$. We are now looking at the probability distribution of its n^{th} draw to belong to c. It is expressed as the probability of a draw from the categorical distribution given all previous observations (because there is only one new draw, it is the same as a Multinomial distribution with parameter $N = 1$) combined with the powered Dirichlet prior defined Eq.6. Then:

$$\begin{aligned} DirCat_r(C_n = c|\vec{\alpha},\vec{C_-}) &= \int_{\vec{p}} Cat(C_n = c|\vec{p}) \underbrace{Dir_r(\vec{p}|\vec{\alpha},\vec{N})}_{\textbf{Eq.6}} \\ &= \int_{\vec{p}} \frac{1}{B(\vec{\alpha} + \vec{N}^r)} \prod_k p_k^{c_k + \alpha_k + N_k^r - 1} \\ &= \frac{B(\vec{c} + \vec{\alpha} + \vec{N}^r)}{B(\vec{\alpha} + \vec{N}^r)} \end{aligned} \tag{8}$$

where $\vec{c}$ is a vector of the same length as $\vec{\alpha}$ and $\vec{C}$ whose cth entry equals 1, and 0 anywhere else. Alternative demonstrations of this result are possible [17, 31].

3.4 Powered Chinese Restaurant process

We finally derive an expression for the Powered Chinese Restaurant process from Eq.8. We recall that $N_k = |\{C_{-i}|i = k\}_{i=1,2,...,n-1}|$. Taking back the conditional probability for the n^{th} observation to belong to cluster c (Eq.8), we have:

$$\begin{aligned} p(C_n =& c|\vec{C_{-}}, \vec{\alpha}) = DirCat_r(C_n = c|\vec{C_{-}}, \vec{\alpha}) \\ =& B(\vec{c} + \vec{N^r} + \vec{\alpha})/B(\vec{N^r} + \vec{\alpha}) \\ =& \Gamma(N_c^r + \alpha_c + 1)\frac{\prod_{k\neq c}\Gamma(N_k^r + \alpha_k)}{\Gamma(1 + \sum_k N_k^r + \alpha_k)}\frac{\Gamma(\sum_k N_k^r + \alpha_k)}{\prod_{k\neq c}\Gamma(N_k^r + \alpha_k)} \\ =& \frac{(N_c^r + \alpha_c)}{\sum_k N_k^r + \alpha_k}\frac{\prod_k \Gamma(N_k^r + \alpha_k)}{\Gamma(\sum_k N_k^r + \alpha_k)}\frac{\Gamma(\sum_k N_k^r + \alpha_k)}{\prod_k \Gamma(N_k^r + \alpha_k)} \\ =& \frac{N_c^r + \alpha_c}{\sum_k N_k^r + \alpha_k} \end{aligned} \tag{9}$$

Every cluster with $N_c = 0$ (empty clusters) has an identical probability of getting chosen. Besides, the result is identical if either of them gets chosen. Therefore, we can express the probability of choosing any empty clusters as a function of $\alpha = \sum_k \alpha_k$. Finally, taking the limit $K \to \infty$ and defining the limit value $\lim_{K\to\infty} \sum_k^K \alpha_k = \alpha$, we find the Powered Chinese Restaurant Process (*PCRP*):

$$PCRP(C_i = c|\alpha, r, C_1, C_2, ..., C_{i-1}) = \begin{cases} \frac{N_c^r}{\alpha+\sum_k^K N_k^r} & \text{if c = 1, 2, ..., K} \\ \frac{\alpha}{\alpha+\sum_k^K N_k^r} & \text{if c = K+1} \end{cases} \tag{10}$$

The formal derivation of the Powered Chinese Restaurant process in Eq.10 and the demonstration of its link to the conditional Dirichlet prior on $\vec{p}$ are the first main contribution of this work. Besides, this demonstration uncovers the link between the prior in Eq.6 and an exotic formulation of the Multinomial distribution, which has never been considered before. As stated in the introduction, special cases of the process have already been used in some occasions [18, 25, 29] but never demonstrated. Furthermore, this formulation generalizes the Uniform process when $r \to 0$ [29], the Dirichlet process when $r \to 1$ and the Pitman-Yor process when $r_k(N_k) = (log(1 - \beta/N_k) + log(N_k))/log(N_k)$ and $\alpha(K) = \alpha + \beta K$ (see Eq.2, we recall that $e^{\log x} = x$). The present expression explicitly allows for controlling the importance of the "rich-get-richer" property as well as recovering state-of-the-art processes.

We illustrate the change on prior probability for an existing cluster to get chosen induced by the Powered Chinese Restaurant process in Fig.1 – we do not plot the prior probability for a new cluster to be created. This figure plots the population of clusters (grey bars) and their associated prior probability of getting chosen. When $r > 1$, the most populated clusters are associated with a more significant prior probability than in the standard CRP, whereas the less populated ones have even less chances to get chosen; rich-get-more-richer, the prior on population is more peaky on large clusters. On the other hand, when $r < 1$, most populated clusters have less chances to get chosen than in CRP, whereas less populated ones have an increased chance of getting chosen; rich-get-less-richer, the prior on population is flatter across clusters of different sizes. In the limit case $r = 0$, the clusters' population does not play any role anymore; rich get-no-richer, the prior is flat over all clusters. Note

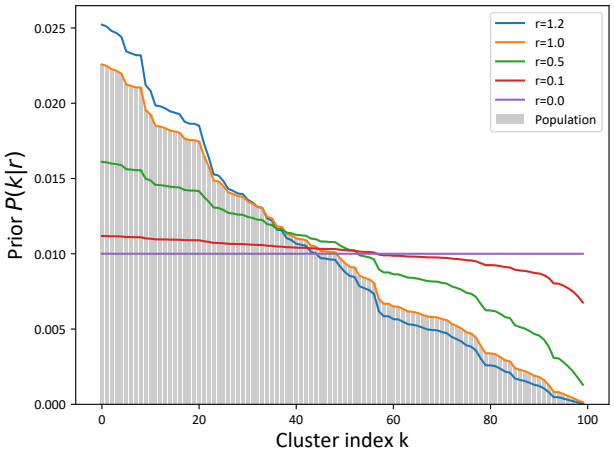

Figure 1: Effect of r on the Powered Chinese Restaurant process prior probability.

that if we wanted to represent the Pitman-Yor process prior in this figure, it would correspond to the plot for $r = 1$ vertically shifted of $-\beta$ (such as defined Eq.2) leading to an increased probability of creating a new cluster of βK (not represented in the plot) [23]. Varying the parameter α of $\Delta\alpha$ plays a similar role as β in this situation. It would uniformly shift the prior probability for each existing cluster to get chosen by $\frac{\Delta\alpha}{K}$ and increase the probability of creating a new one by $\Delta\alpha$. For Both Pitman-Yor and Dirichlet processes, the linear dependence of each cluster's population does not change.

In Fig1, we understand that the Powered Chinese Restaurant process allows for defining priors from clusters population that are not possible when tuning the Chinese Restaurant or Pitman-Yor processes. Introducing non-linearity in the dependence on previous observations allows giving any importance to the "rich-get-richer" property.

4 PROPERTIES OF THE POWERED CHINESE RESTAURANT PROCESS

We will now investigate some key properties of the Powered Chinese Restaurant process. We recall that N_k is the population of the cluster k, and $N = \sum_k N_k$.

4.1 Convergence

Proposition 1. *For $N \to \infty$, the Powered Chinese Restaurant process converges towards a stationary distribution. When $r < 1$, it converges towards a uniform distribution over all the possible clusters, and when $r > 1$, it converges towards a Dirac distribution on a single cluster.*

Proof. We consider a simple situation where only 2 clusters are involved. The generalization to the case where K clusters are involved is straightforward. When clusters' population is large

enough, we make the following Taylor approximation:

$$(N_i+1)^r = N_i^r(1+\frac{1}{N_i^r}) = N_i^r + rN_i^{r-1} + O(N^{r-2}) \quad (11)$$

Since the population of a cluster N_i is a non-decreasing function of N, we assume that first order Taylor approximation holds when $N \to \infty$. Given clusters population at the N^{th} observation, we perform a stability analysis of the gap between probabilities $\Delta p(N) = p_1(N) - p_2(N)$. We recall that the probability for cluster i to get chosen is $p_i(N) = N_i^r/(\sum_k N_k^r)$ and that either of the clusters is chosen with this probability at the next step (at step $N+1$, $\Delta p(N+1) = p_1(N+1) - p_2(N)$ with probability $p_1(N)$ and $\Delta p(N+1) = p_1(N) - p_2(N+1)$ with probability $p_2(N)$). Explicitly the variation of the gap between probabilities when N grows is written as:

$$\begin{aligned}
&\frac{p_1(N)(p_1(N+1)-p_2(N)) + p_2(N)(p_1(N)-p_2(N+1)) - \Delta p(N)}{\Delta p(N)} \\
&\overset{\text{Eq.11}}{\approx} \frac{1}{p_1(N)-p_2(N)} \\
&\quad \times \left(p_1(N)\frac{N_1^r - N_2^r + rN_1^{r-1}}{N_1^r + N_2^r + rN_1^{r-1}} + p_2(N)\frac{N_1^r - N_2^r - rN_2^{r-1}}{N_1^r + N_2^r + rN_2^{r-1}} \right) \\
&= \frac{2rN_1^rN_2^r}{(N_1^r + N_2^r + rN_1^{r-1})(N_1^r + N_2^r + rN_2^{r-1})} \left(\frac{N_1^{r-1} - N_2^{r-1}}{N_1^r - N_2^r} \right)
\end{aligned} \quad (12)$$

We see in Eq.12 that the sign of the variation of the gap between probabilities depend only on the term $\frac{N_1^{r-1}-N_2^{r-1}}{N_1^r-N_2^r}$. We can therefore perform a stability analysis of the Powered Chinese Restaurant process using only this expression.

When $0 < r < 1$, the following relation holds: $N_1^{r-1} - N_2^{r-1} < 0 \Leftrightarrow N_1^r - N_2^r > 0 \ \forall N_1, N_2$; that makes right hand side of Eq.12 negative. Therefore adding a new observation statistically reduces the gap between the probabilities of the two clusters. We could forecast this prediction from Eq.11 by seeing that adding a new observation to a large cluster increases its probability to get chosen lesser than for a small cluster – rich-get-less-richer. Moreover, we see from Eq.11 that a crowded cluster (such as $N_1^r \gg N_2^r$) see its probability evolve as N^{r-1}. Asymptotically, the only fixed point of Eq.12 when $N \to \infty$ is $N_1 \to N_2$, which implies a uniform distribution.

On the contrary, when $r > 1$ we have the following relation: $N_1^{r-1} - N_2^{r-1} > 0 \Leftrightarrow N_1^r - N_2^r > 0 \ \forall N_1, N_2$; ; that makes right hand side of Eq.12 positive. Adding a new observation statistically increases the gap between probabilities. From Eq.11, we see that adding an observation to a large cluster increases its probability with its population – rich-get-more-richer. In this case, Eq.12 has $K+1$ fixed points, with K the number of clusters. The uniform distribution is an unstable fixed point, while K Dirac distributions (each on one cluster) are stable fixed points of the system. It means the gap converges to 1, that is a probability of 1 for one cluster and a probability of 0 for the others.

When $r = 1$, the right hand side of Eq.12 is null. It means the gap remains statistically constant $\forall N_i$, which is a classical result for the regular Dirichlet process. This convergence has already been studied on many occasions [2, 6].

We note that as $r \to 0$, Eq.12 is not defined anymore. That is because the probability for a cluster to be chosen does not depend on its population anymore. In this case, $p_1(N) - p_2(N) \propto N_1^0 - N_2^0 = 0$: the probability for any cluster to be chosen is equal, hence the Uniform process – "rich-get-no-richer".

□

4.2 Expected number of tables

Proposition 2. *When N is large, $\sum_k N_k^r$ varies with N as $N^{\frac{r^2+1}{2}}$ when $r < 1$, and with N^r when $r \geq 1$.*

Proof. Taking back Eq.10, we are interested in the variation of $p_i = \frac{N_i^r}{\sum_k N_k^r}$ according to N when N_i^r is large:

$$p_i(N+1) - p_i(N) \approx \begin{cases} \frac{rN_i^{r-1} + O(N^{r-2})}{\sum_k N_k^r} & \text{if } N_i \text{ grows} \\ 0 & \text{else} \end{cases} \quad (13)$$

We see in Eq.13 that for $r < 1$, the larger N_i the slower the variation of p_i. It means that for large N_i^r, we can write $N_i \propto Np_i$, with p_i a constant of N. Since N_i is either way a non-decreasing function of N, we reformulate the constraint N_i^r large in N^r large.

For $r > 1$, the probability p_i varies greatly with N and quickly converges to 1 for large N (see Proposition 1), and so $N_i \approx N$ for cluster i and $N_{j\neq i} \ll N_i \ \forall j$.

Since the sum $\sum_k N_k^r$ essentially varies according to large N_k, we can approximate $\sum_k N_k^r \approx N^r \sum_k p_k^r$ for large N^r.

Besides, we showed in Proposition 1 that for large N the process converges towards a uniform distribution for $r < 1$ and towards a Dirac distribution when $r > 1$. Therefore, we can express $\sum_k^K p_k^r$ as:

$$\sum_k^K p_k^r \overset{N\gg 1}{\approx} \begin{cases} K^{1-r} & \text{for } r < 1 \\ 1 & \text{for } r \geq 1 \end{cases} \quad (14)$$

Based on the demonstration of Eq.4 in [29], we suppose that K evolves with N as $N^{\frac{1-r}{2}}$ when $r < 1$. We verify that this assumption holds in the Experiment section.

Therefore, we can write:

$$\sum_k N_k^r \approx N^r \sum_k^K p_k^r \approx \begin{cases} N^r \left(N^{\frac{1-r}{2}}\right)^{1-r} = N^{\frac{1+r^2}{2}} & \text{for } r < 1 \\ N^r & \text{for } r \geq 1 \end{cases} \quad (15)$$

□

Proposition 3. *The expected number of tables of the Powered Chinese Restaurant process evolves with $N \gg 1$ as $H_{\frac{r^2+1}{2}}(N)$ for $r < 1$ and as $H_r(N)$ when $r \geq 1$, where $H_m(n)$ is the generalized harmonic number.*

Proof. In general, the expected number of clusters at the N^{th} step can be written as:

$$\mathbb{E}(K|N,r) = \sum_1^N \frac{\alpha}{\sum_k N_k^r + \alpha} \overset{N^r\gg 1}{\propto} \sum_1^N \frac{1}{\sum_k N_k^r} \quad (16)$$

We showed in Proposition 2 that we can rewrite $\sum_k N_k^r \propto N^{\frac{r^2+1}{2}}$ when $r < 1$ and $\sum_k N_k^r \propto N^r$ when $r \geq 1$. Injecting this result in

Eq.16 for r, we get:

$$\mathbb{E}(K|N,r) \overset{N^r \gg 1}{\propto} \begin{cases} \sum_1^N \frac{1}{N^{\frac{r^2+1}{2}}} = H_{\frac{r^2+1}{2}}(N) \\ \sum_1^N \frac{1}{N^r} = H_r(N) \end{cases} \tag{17}$$

□

For $r = 1$, $\mathbb{E}(K|N, r = 1) \propto H_1(N) \approx \gamma + \log(N)$ where γ is the Euler–Mascheroni constant, which is a classical result for the regular Dirichlet process.

When $r > 1$ and $N \to \infty$, the term $H_{\frac{r^2+1}{2}}(N)$ converges towards a finite value and the sum $\sum_k p_k^r$ goes to 1 (see Proposition 1). By definition $\mathbb{E}(K|N, r > 1) \overset{N\to\infty}{\propto} \zeta(\frac{r^2+1}{2})$, where ζ is the Riemann Zeta function.

When $r < 1$, we can approximate the harmonic number in a continuous setting. We rewrite Eq.17 as:

$$\begin{aligned} \mathbb{E}(K|N,r) &\overset{N^r \gg 1}{\propto} \sum_{n=1}^{N} \frac{1}{n^{\frac{r^2+1}{2}}} \overset{N^r \gg 1}{\approx} \int_1^N n^{-\frac{r^2+1}{2}} dn \\ &= \frac{2}{1-r^2}(N^{\frac{1-r^2}{2}} - 1) \end{aligned} \tag{18}$$

One can show that $\frac{N^{1-x}-1}{1-x} = H_x(N) + O(\frac{1}{N^x})$. Therefore, the Powered Chinese Restaurant process exhibits a power-law behaviour similar to the Pitman-Yor process Eq.2 for $r = \sqrt{1-2\beta}$ for $0 < r < 1$. For values of $r > 1 \Leftrightarrow \beta < 0$, the equivalent Pitman-Yor process is not defined unlike the Powered Chinese Restaurant process. Note that there is *a priori* no reason for r to be constrained in the domain of real number. Complex analysis of the process might be an interesting lead for future works.

5 EXPERIMENTS

5.1 Numerical validation of propositions

First of all, we present numerical confirmations of propositions stated above (Propositions 1, 2, 3) by simulating 100 independent Powered Chinese Restaurant processes with parameter $\alpha = 1$ for various values of r. We present the results of numerical simulations in Fig.2.

In the **top** part, we plot the evolution of the probability for each cluster to be chosen as N grows for $r = 0.2$ for one run. We see that the probabilities do not remain constant but instead diminish as the number of clusters grows. The figure suggests they all converge to a common value (a uniform probability) as shown in Proposition 1. The black line shows the probability of a uniform distribution. We chose not to show the results for $r > 1$; in this case, one probability goes to 1 as the other fades to 0 as N grows, as expected.

In the **middle** part of the figure, we plot the expression for $\sum_k N_k^r$ derived in Proposition 2 (solid lines) versus the value of the sum from experimental results (transparent lines), averaged over 100 runs. Note that plots are in a log-log scale and that curves have been shifted vertically for visualization purposes. As assumed in Proposition 2, the approximation holds for all values of r.

Finally in the **bottom** picture, we plot the evolution of the number of clusters K versus N according to Proposition 3 (solid lines) and experiments (transparent lines). The error bars correspond to the standard deviation over the 100 runs. We see that the expression derived in Proposition 3 accounts well for the evolution of the number of clusters. Note that plots are in a log-log scale and that curves have been shifted vertically for visualization purposes. We must point out that there is a constant shift from experiments to the theory that does not appear on the plot (because of the rescaling). This shift comes from the approximation of large N^r which is not valid at the beginning of the process. However, it does not play any role in the evolution of K as N grows large enough.

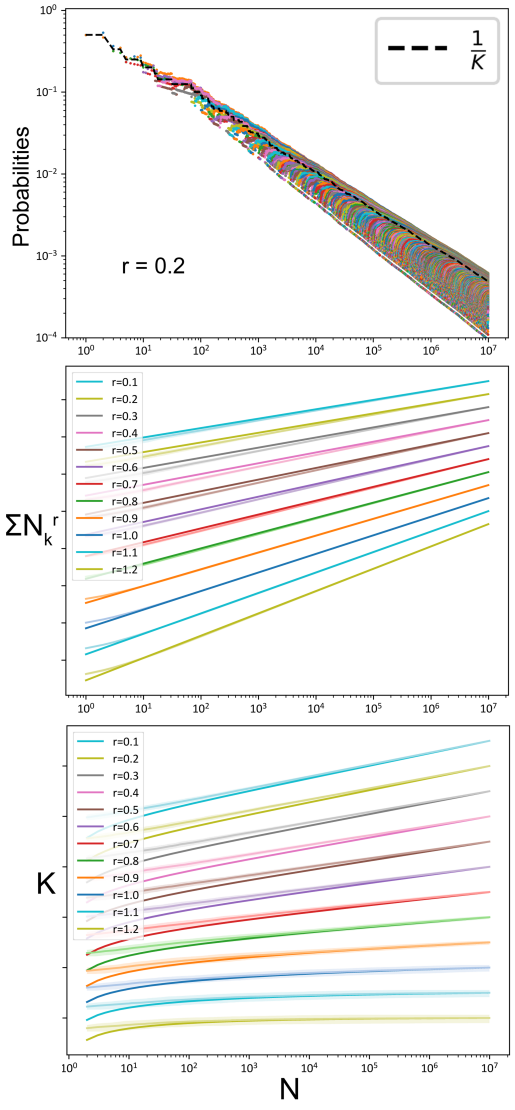

Figure 2: Numerical validation of Propositions 1 (top), 2 (middle), 3 (bottom). In the first plot, K is the number of non-empty clusters. In the second and third plots, the theoretical results are the solid lines and the associated numerical results are the transparent lines of same color. Except for small N, the difference between theory and experiments is almost indistinguishable.

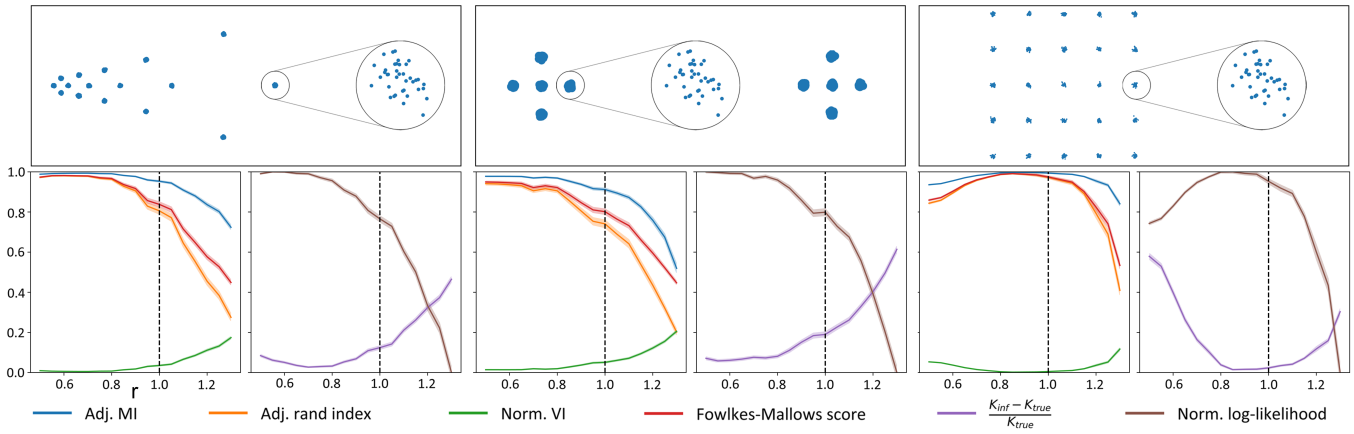

Figure 3: Application on synthetic data. (Top) Original datasets used for the experiments (Density, Diamond and Grid). (Bottom) Results for various values of r; the x and y axes all the same. The dashed line indicates the regular DP prior as $r = 1$. The error correspond to the standard error of the mean over all runs.

5.2 Use case: infinite Gaussian mixture model

We now illustrate the usefulness of a prior that alleviates the "rich-get-richer" property on several synthetic datasets and on a real-world application. We choose to consider as an illustration its use as a prior in the infinite Gaussian mixture model[2]. We choose this application to ease visual understanding of the implications of the PCRP, but the argument holds for other models using DP priors as well (text modeling, gene expression clustering, etc.).

We consider a classical infinite Gaussian mixture model coupled with a Powered Dirichlet process prior. We fit the data using a standard collapsed Gibbs sampling algorithm for IGMM [26, 29, 33], with a Normal Inverse Wishart prior on the Gaussians' parameters. The input data is shuffled at each iteration to reduce the ordering bias from the dataset. Note that we cannot completely get rid of the bias because the Powered Dirichlet Process is not exchangeable for all r. The problem has been addressed on numerous occasions (Uniform process [29], distance-dependent CRP [3, 8], spectral CRP [27]) and shown to induce negligible variations of results in the case of Gibbs sampling. We stop the sampler once the likelihood of the model reaches stability ; we repeat this procedure 100 times for each value of r. Finally, the parameter α is set to 1 in all experiments (see Section 2.1).

5.2.1 Synthetic data. We present the results on synthetic data in Fig.3 and in Table1. We consider standard metrics in clustering evaluation with a non-fixed number of clusters: mutual information score and rand index both adjusted for chance (**Adj.MI** and **Adj.RI**), normalized variation of information (**Norm.VI**, lower is better), Fowlkes-Mallow score, marginal likelihood (normalized for visualization) and absolute relative variation of the inferred number of clusters according to the number used in the generation process ($\frac{K_{inf}-K_{true}}{K_{true}}$, lower is better). Note that we purposely chose stereotypical cases to illustrate the argument better. The Density dataset on the **left** of Fig.3 is informative about the change induced by r. Here, clusters are distributed at various scales in the dataset; we see that the lower the value of r, the better the results. Indeed, when r is small, the model can distinguish clusters in the dense area better, whereas when r is closer to 1, the clusters in the dense area are put together in a larger cluster. The same happens with the Diamond dataset in the **middle** of Fig.3, where clusters are distributed according to two different scales. Finally, on the Grid

[2]All codes and datasets can be found at https://anonymous.4open.science/r/PDP-91ea587e-fba6-4ba0-887e-79d87abf0b31/

Table 1: Numerical results of the Powered Dirichlet process, Uniform process and Dirichlet process priors coupled to a standard Infinite Gaussian Mixture Model, for the 3 synthetic datasets plotted Fig. 3 and 2 real-world datasets (100 runs each). We see that using PDP as prior makes the model outperform the baselines consistently on every metric. The standard error on the last digit(s) is given in shorthand form (for instance $0.123(12) \Leftrightarrow 0.123 \pm 0.012$).

		Adj.MI	Adj.RI	Norm.VI	$\frac{K_{inf}-K_{true}}{K_{true}}$
Density	PDP (r=0.60)	**0.992(1)**	**0.980(2)**	**0.006(1)**	**0.045(5)**
	DP	0.951(4)	0.797(17)	0.037(3)	0.128(10)
	UP	0.939(2)	0.854(4)	0.050(1)	0.548(1)
Diamond	PDP (r=0.50)	**0.982(2)**	**0.956(5)**	**0.011(1)**	**0.063(7)**
	DP	0.909(7)	0.731(19)	0.053(4)	0.202(12)
	UP	0.927(2)	0.844(6)	0.051(2)	0.544(2)
Grid	PDP (r=0.85)	**0.997(1)**	**0.990(2)**	**0.003(1)**	**0.014(2)**
	DP	0.995(1)	0.977(4)	**0.004(1)**	**0.018(3)**
	UP	0.811(1)	0.517(3)	0.154(1)	2.12(1)
Iris	PDP (r=0.90)	**0.868(4)**	**0.866(7)**	**0.057(2)**	**0.000(0)**
	DP	0.843(6)	0.820(12)	0.065(2)	0.030(10)
	UP	0.544(2)	0.295(3)	0.303(2)	2.777(32)
Wines	PDP (r=0.10)	**0.712(15)**	**0.637(20)**	**0.102(5)**	**0.157(17)**
	DP	0.589(19)	0.461(16)	0.128(4)	0.327(13)
	UP	**0.713(17)**	**0.657(21)**	**0.103(5)**	**0.147(17)**
20-NG	PDP (r=0.80)	**0.421(4)**	**0.119(3)**	**0.477(3)**	-
	DP	0.404(4)	0.105(4)	0.491(3)	-
	UP	0.000(4)	0.000(0)	0.830(3)	-

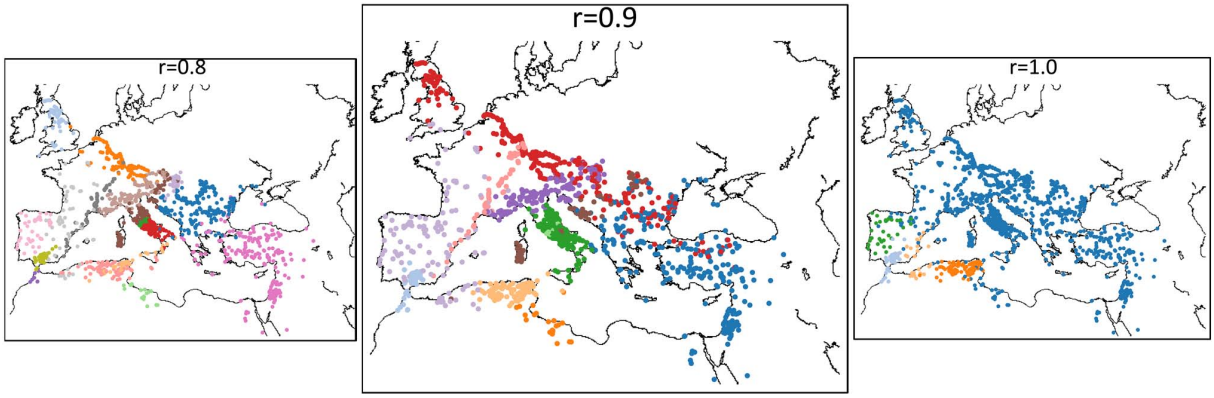

Figure 4: Application to spatial clustering on geolocated data for $r = 0.8$ (left), $r = 1$ (right) and $r = 0.9$ (middle). We see that the model using a Powered Dirichlet process prior for $r = 0.9$ and $r = 0.8$ describes the data better than the same model using a Dirichlet process prior ($r = 1$).

dataset on the **right** part of Fig.3, we see an optimum r exists to distinguish the clusters distributed on a grid; it makes sense since only one scale in clusters distribution is involved in this dataset. In Table1, we explicitly report the values of the PDP optimal r and compare them to the values yield by the same model using either a Dirichlet Process (**DP**) prior or a Uniform Process (**UP**) prior.

5.2.2 Real-world data. In Table 1, we report the results for two well known real-world datasets (Iris and Wines). We also run an additional experiment on the 20Newsgroup (20-NG) dataset, which is a collection of 18 000 users posts published on Usenet, organized in 20 Newsgroup (which are our target thematic clusters). As a model, we consider a modified version of LDA that uses a PDP prior instead of DP. Note that because the number of clusters has to be given to LDA, we do not compute $\frac{K_{inf}-K_{true}}{K_{true}}$. We see PDP yields improved performances on every real-world dataset.

We now illustrate the interest of using an alternate form of prior for the Infinite Gaussian Mixture model on real-world data. We consider a dataset of 4.300 roman sepulchral inscriptions comprising the substring "Antoni" that have been dated between 150AC and 200AC and assigned with map coordinates. The dates correspond to the reign of Antoninus Pius over the Roman empire. The dataset is available on Clauss-Slaby repository[3]. It was common to give children or slaves the name of the emperor; the dataset gives a global idea of the main areas of the roman empire at that time [14]. The task is to discover spatial clusters of individuals named after the emperor. We expect to find geographical clusters around: Italy, Egypt, Gauls, Judea, and all along the *limes* (borders of the roman empire, which concentrate lots of sepulchral inscriptions for war-related reasons) [13]. We present the results in Fig.4.

We see that when $r = 1$, the classical DP prior is not fit for describing this dataset, as it misses most of the clusters. On the other hand, when $r = 0.9$, the infinite Gaussian mixture model retrieves the expected clusters. It also makes some clusters that were not expected, such as the north Italian cluster or the long cluster going through Spain and France that corresponds to roman roads layout (via Augusta and via Agrippa; it was common to bury the dead on roads edges). Finally, when $r = 0.8$, we get even more detail: some of the main clusters are broken into smaller ones (Italy breaks into Rome, North Italy, and South Italy; Britain becomes an independent cluster, etc.). In this case, changing r controls the level of details of the clustering. Note that we do not compute metrics for this experiment in the absence of "ground-truth" clustering; there is no such thing as the right clustering in this case. Applied to spatial data, the PDP prior allows to control clustering granularity. We see how different results can be according to the extent the model relies on the "rich-get-richer" prior and how its control is needed to make modeling relevant to a given situation.

[3]http://www.manfredclauss.de/fr/index.html

6 CONCLUSION

In this article, we discuss the necessity of controlling the "rich-get-richer" property that arises from the common Chinese Restaurant Process usual formulation. We discuss cases where this modeling hypothesis must be alleviated or strengthened. To allow it, we derive the Powered Chinese Restaurant Process. This formulation allows reducing the expected number of clusters, which is not possible in the standard Pitman-Yor processes, while generalizing the standard Dirichlet process and the Uniform process. The principal feature of this formulation is that it allows for direct control of the "rich-get-richer" priors' importance. We derive elementary results on convergence and the expected number of clusters of the new process. Finally, we show that it yields better results on synthetic data when coupled with a standard Gaussian mixture model and illustrates a possible use case with real-world data. For future works, it might be interesting to investigate cases where r takes non-positive values (which might lead to a "poor-get-richer" kind of process), and to develop a procedure to infer it automatically for specific problems (by minimizing a dispersion criterion for instance).

The regular Chinese Restaurant process has been used for decades as a powerful prior in many real-world applications. However, alternate forms for this prior have been little explored. It would be very interesting to study the changes brought to state-of-the-art models based on Dirichlet priors by varying the importance of the "rich-get-richer" assumption as proposed in this paper.

REFERENCES

[1] Edoardo Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing. 2008. Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research* 9 (2008), 1991–1992.

[2] Richard Arratia, A. D. Barbour, and Simon Tavaré. 1992. Poisson Process Approximations for the Ewens Sampling Formula. *The Annals of Applied Probability* 2, 3 (1992), 519–535.

[3] D.M. Blei and Peter Frazier. 2011. Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research* 12 (08 2011), 2461–2488.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022.

[5] Duch J. Cobo-López S., Godoy-Lorite A. 2018. Optimal prediction of decisions and model selection in social dilemmas using block models. *EPJ Data Sci* 7(48) (2018).

[6] Thomas S. Ferguson. 1973. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics* 1, 2 (1973), 209 – 230.

[7] Yair Ghitza and Andrew Gelman. 2013. Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups. *American Journal of Political Science* 57 (07 2013). https://doi.org/10.1111/ajps.12004

[8] Soumya Ghosh, Michalis Raptis, Leonid Sigal, and Erik B. Sudderth. 2014. Nonparametric Clustering with Distance Dependent Hierarchies *(UAI'14)*. 260–269.

[9] Antonia Godoy-Lorite, Roger Guimerà, Cristopher Moore, and Marta Sales-Pardo. 2016. Accurate and scalable social recommendation using mixed-membership stochastic block models. *PNAS* 113, 50 (2016), 14207–14212.

[10] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2011. Producing Power-Law Distributions and Damping Word Frequencies with Two-Stage Language Models. *JMLR* 12, 68 (2011).

[11] Roger Guimera, Alejandro Llorente, and Marta Sales-Pardo. 2012. Predicting Human Preferences Using the Block Structure of Complex Social Networks. *PLOS One* 7, 9 (2012).

[12] Roger Guimerà and Marta Sales-Pardo. 2013. A Network Inference Method for Large-Scale Unsupervised Identification of Novel Drug-Drug Interactions. *PLoS Comput Biol* (2013).

[13] John William Hanson. 2016. *An urban geography of the Roman world, 100 BC to AD 300*. Vol. 18. Archaeopress Oxford.

[14] J. W. Hanson, S. G. Ortman, and J. Lobo. 2017. Urbanism and the division of labour in the Roman Empire. *Journal of The Royal Society Interface* 14, 136 (2017), 20170367.

[15] Driver H.E. and Kroeber A.L. 1932. In *Quantitative expression of cultural relationships*. University of California Press.

[16] Hemant Ishwaran and Lancelot James. 2003. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica* 13 (10 2003), 1211–1235.

[17] Sethuraman J. 1994. A constructive definition of Dirichlet priors. *Statistica sinica* 4, 4 (1994), 639–650.

[18] S. Jensen and J. Liu. 2008. Bayesian clustering of transcription factor binding motifs. In *Journal of the American Statistical Association*, Vol. 103. 188–200.

[19] Anwer Khurshid, Mohammad Ageel, and Rizwan Lodhi. 2005. On Confidence Intervals for the Negative Binomial Distribution. *Revista Investigacion Operacional* 26 (01 2005), 59–70.

[20] Antonio Lijoi, Ramsés H. Mena, and Igor Prünster. 2007. Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69, 4 (2007), 715–740. https://doi.org/10.1111/j.1467-9868.2007.00609.x arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2007.00609.x

[21] Davis J. McCarthy, Y. Chen, and G. Smyth. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* 40 (2012), 4288 – 4297.

[22] Ian C McDowell, Dinesh Manandhar, Christopher M Vockley, Amy K Schmid, Timothy E Reddy, and Barbara E Engelhardt. 2018. Clustering gene expression time series data using an infinite Gaussian process mixture model. *PLoS computational biology* 14, 1 (2018), e1005896.

[23] Jim Pitman and Marc Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* 25, 2 (1997), 855 – 900.

[24] Gaël Poux-Médard, Julien Velcin, and Sabine Loudcher. 2020. Interactions in information spread: quantification and interpretation using stochastic block models. *arXiv* (2020). arXiv:2004.04552 [cs.LG]

[25] Z. S. Qin, L. A. McCue, W. Thompson, L. Mayerhofer, C. E. Lawrence, and J. S. Liu. 2003. Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. In *Nature Biotechnology*, Vol. 21. 435–439.

[26] Carl Edward Rasmussen. 1999. The Infinite Gaussian Mixture Model *(NIPS'99)*. MIT Press, 554–560.

[27] Richard Socher, Andrew Maas, and Christopher Manning. 2011. Spectral Chinese Restaurant Processes: Nonparametric Clustering Based on Similarities. *JMLR - Proceedings* 15 (2011), 698–706.

[28] Erik Sudderth and Michael Jordan. 2009. Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes. In *NIPS*, Vol. 21.

[29] Hanna Wallach, Shane Jensen, Lee Dicker, and Katherine Heller. 2010. An alternative prior process for nonparametric Bayesian clustering. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR, 892–899.

[30] Max Welling. 2006. Flexible priors for infinite mixture models. In *Workshop on learning with non-parametric Bayesian methods*.

[31] S.S. Wilks. 1992. Mathematical statistics. (1992), section 7.

[32] W. Xu, Y. Li, and J. Qiang. 2021. Dynamic clustering for short text stream based on Dirichlet process. *Appl Intell* (2021). https://doi.org/10.1007/s10489-021-02263-z

[33] Jianhua Yin and Jianyong Wang. 2014. A Dirichlet Multinomial Mixture Model-Based Approach for Short Text Clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, New York, USA) *(KDD '14)*. Association for Computing Machinery, New York, NY, USA, 233–242.