Manuscript submitted to
AIMS' Journals
Volume **X**, Number **0X**, XX **200X**

doi:10.3934/xx.xx.xx.xx

pp. **X–XX**

MODELLING DYNAMIC NETWORK EVOLUTION AS A PITMAN-YOR PROCESS

Francesco Sanna Passino and Nicholas A. Heard

Department of Mathematics, Imperial College London
180 Queen's Gate, London – SW7 2AZ (United Kingdom)

(Communicated by the associate editor name)

Abstract. Dynamic interaction networks frequently arise in biology, communications technology and the social sciences, representing, for example, neuronal connectivity in the brain, internet connections between computers and human interactions within social networks. The evolution and strengthening of the links in such networks can be observed through sequences of connection events occurring between network nodes over time. In some of these applications, the identity and size of the network may be unknown *a priori* and may change over time. In this article, a model for the evolution of dynamic networks based on the Pitman-Yor process is proposed. This model explicitly admits power-laws in the number of connections on each edge, often present in real world networks, and, for careful choices of the parameters, power-laws for the degree distribution of the nodes. A novel empirical method for the estimation of the hyperparameters of the Pitman-Yor process is proposed, and some necessary corrections for uniform discrete base distributions are carefully addressed. The methodology is tested on synthetic data and in an anomaly detection study on the enterprise computer network of the Los Alamos National Laboratory, and successfully detects connections from a red-team penetration test.

1. Introduction.

A network can be represented as a directed graph $\mathbb{G} = (V, E)$, where V is a set of nodes and $E \subseteq V \times V$ is a set of links, or *edges*, indicating the pairs of nodes which have interacted. Statistical models for networks are well studied and understood in the literature [8]. Less attention is devoted to dynamic networks. In dynamic networks, a stochastic process is observed on $V \times V$. Graphs having this structure are commonly observed in the social network literature; for example, emails sent between employees within an enterprise network, or messages between users on Facebook. The motivating application for this article is in cyber-security, where the nodes are computers or Internet Protocol (IP) addresses and the objective is to detect anomalous, potentially compromised nodes. Network-wide modelling of large computer networks typically requires a trade-off between mathematical complexity and computational feasibility, which is addressed in this article by proposing a simple model for network evolution. The main advantage of the proposed model over other approaches to dynamic network modelling suggested in the literature [21, 17] is that the model is simpler, and allows to directly obtain a predictive probability for a link.

2010 *Mathematics Subject Classification.* Primary: 90B15; Secondary: 91D30.

Key words and phrases. anomaly detection, dynamic networks, statistical cyber-security, Pitman-Yor process, p-value combination.

Computer network graphs have directed edges, since every connection is initiated by a *source* node making a request to another *destination* node, so $(x,y)\in E \not\Rightarrow (y,x)\in E$. Let $V_S\subseteq V$ be the subset of nodes which are ever observed to initiate a connection, and $V_D\subseteq V$ the set of nodes arising as a destination of one or more connections. This article examines a joint network-wide generative model for the sequence of links $(x_1,y_1),(x_2,y_2),\ldots\in V_S\times V_D$, based on the Bayesian nonparametric Pitman-Yor process [23, 12]. The Pitman-Yor process, also known as the two-parameter Poisson-Dirichlet process, provides a natural extension of the Dirichlet process [6] where the probability of observing infrequently observed events can be further discounted. The Pitman-Yor process has been successfully used in natural language processing and computational linguistics applications [9, 29] to appropriately model the frequency of words in languages, which often exhibit power-law behaviour [2]. The power-law is explicitly modelled by the interplay between a strength parameter α and a discount parameter $d\in[0,1)$.

Statistical modelling of categorical sequences is a well established branch of statistics and natural language processing, with relevant applications in speech recognition. Common methods mainly include n-gram models, based on n-order Markov assumptions (see the survey of [13]), neural models, in particular recurrent neural networks [3, 18], maximum entropy or exponential models [25] and positional models [16]. The model presented in this article is much simpler, but well-suited for an efficient network-wide implementation on networks with large numbers of nodes and high frequency activity.

In the present article, the Pitman-Yor is shown to have interesting properties for suitably modelling events within real-world networks. The article also proposes an empirical procedure for estimation of the parameters within a “big data” framework. Furthermore, approximations and corrections for parameter estimation in Pitman-Yor processes with uniform discrete base distributions are also carefully addressed.

2. Modelling sequences using the Pitman-Yor process.

Consider an infinitely exchangeable sequence of nodes $x_1,x_2,\ldots\in V$. After observing n events $\boldsymbol{x}_n=(x_1,\ldots,x_n)$, let $(x_1^\star,\ldots,x_{K_n}^\star)$ be the K_n unique observations in $\boldsymbol{x}_n$. Let F_0 be a *base* probability distribution on V. Then the distribution of observed nodes is a Pitman-Yor process $\mathrm{PY}(\alpha,d,F_0)$, if the predictive distribution for the next observed node is

$$p(x_{n+1}|\boldsymbol{x}_n)=\frac{\alpha+dK_n}{\alpha+n}F_0(x_{n+1})+\sum_{j=1}^{K_n}\frac{N_{jn}-d}{\alpha+n}\delta_{x_j^\star}(x_{n+1}), \tag{1}$$

where $d\in[0,1)$ is a *discount* parameter, $\alpha>-d$ is a *strength* parameter, $\delta_v(u)=1$ if $u=v$, 0 otherwise, and $N_{jn}=\sum_{i=1}^n\delta_{x_i}(x_j^\star)$ is the number of occurrences of the value $x_j^\star$ in $\boldsymbol{x}_n$. Note that $d=0$ corresponds to the more familiar Dirichlet process, $\mathrm{PY}(\alpha,0,F_0)\equiv\mathrm{DP}(\alpha,F_0)$.

The dynamics of the process can be most easily understood using the Chinese Restaurant metaphor [1], which will be extensively used in this article. Starting from an empty restaurant with a potentially infinite number of tables, the first customer sits at the first table and then for $n=1,2,\ldots$ the $(n+1)$-th customer either sits at a new table or chooses an already occupied table with the following probabilities:

$$\mathbb{P}(\text{new table})=\frac{\alpha+dK_n}{\alpha+n},\qquad \mathbb{P}(j\text{-th table})=\frac{N_{jn}-d}{\alpha+n},$$

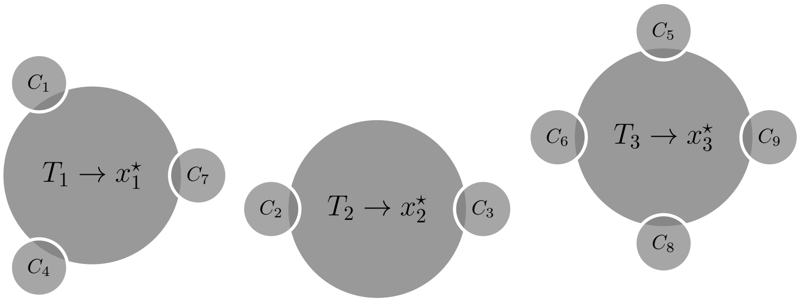

Figure 1. Cartoon example of the Chinese Restaurant metaphor for the Pitman-Yor process, where C_j represents the jth customer and $x_j^\star$ is the unique dish served at table T_j. The vector $\boldsymbol{x}_n = (x_1, \dots, x_n)$ denotes the observed sequence of dishes eaten by each customer. Customer C_i being seated at table T_j is denoted $C_i \to T_j$. Then, for example, the tenth customer will sit at T_1 with probability $(3-d)/(\alpha+9)$.

for $j = 1, \dots, K_n$. In this metaphor, N_{jn} is the number of customers sitting at the j-th table when n customers are in the restaurant. If each table is associated with a random dish drawn from a non-atomic base distribution F_0, the sequence $x_1, x_2, \dots$ corresponding to the dishes eaten by successive customers is an exchangeable stochastic process with De Finetti measure $\mathrm{PY}(\alpha, d, F_0)$. Figure 1 presents a cartoon representation of this metaphor for the Pitman-Yor process.

In this article, for an observed sequence of links $(x_1, y_1), \dots, (x_N, y_N) \in V_S \times V_D$, it is assumed that the joint distribution has the following hierarchical structure:

$$\begin{aligned} x_i|y_i &\overset{d}{\sim} F_{x|y_i},\ i = 1, 2 \dots, N, \\ y_i &\overset{iid}{\sim} G,\ i = 1, 2 \dots, N, \\ F_{x|y} &\overset{d}{\sim} \mathrm{PY}(\alpha_y, d_y, F_0),\ y \in V_D, \\ G &\overset{d}{\sim} \mathrm{PY}(\alpha_0, d_0, G_0), \end{aligned} \tag{2}$$

where $\{F_{x|y}\}$ and G are unknown probability mass functions on the node set V. The probability of obtaining a link (x, y) is decomposed in two parts: $p(x, y) = p(x|y)p(y)$. The sequence of destination nodes $y_1, \dots, y_N \in V_D$ is assumed to be exchangeable with a Pitman-Yor hierarchical distribution. Similarly, conditional on the destination node $y \in V_D$, it is assumed that the sequence of source nodes $x_1, x_2, \dots \in V_S$ connecting to y is again exchangeable with a Pitman-Yor hierarchical distribution, with parameters depending on the specific value of y. For the application to computer networks, the decomposition $p(x, y) = p(x|y)p(y)$ is preferred over $p(x, y) = p(y|x)p(x)$: for identifying red-team behaviour, anomalous client computers must be selected, hence each event is first given a score measuring how surprising the server found the connection from the corresponding client, and then all such measures are combined for a given client.

Note that care is needed when the base distribution of the process is atomic. In this case, using the Chinese Restaurant metaphor, it is possible that two or more tables serve the same dish, which means that the same draw from the base distribution is associated with multiple tables. Hence, the predictive probability (1) cannot be identified, since the underlying K_n, and consequently N_{jn}, are not determinable from $\boldsymbol{x}_n$. In contrast, with continuous base distributions, the draws are distinct with probability 1. Attention in the literature is mostly devoted to non-atomic base distributions. For the case of atomic F_0, [4] suggest introducing latent *multiplicities* and *table indicators*, corresponding to the number of tables contributing to the total number of times a value $x_j^\star$ is observed. An efficient sampler for this representation is derived in [5]. In this article, the node set will be assumed to be countable, and simple adaptation techniques will be used in the estimation of the parameters to take this problem into account.

3. **Empirical estimation of the hyperparameters.** Consider a Pitman-Yor process with non-atomic base distribution and $0 \leq d < 1$, and let K_n be the number of occupied tables after n customers have entered the restaurant. [22] shows that $\mathbb{E}(K_n) = d^{-1}(\alpha+d)_n/(\alpha+1)_{n-1} - \alpha/d$ for $d \neq 0$, where the subscripts represent the Pochhammer symbol. For large n, using Stirling's approximation, $n^d \approx \Gamma(n+\alpha+d)/\Gamma(n+\alpha)$, this expectation can be approximated as

$$\mathbb{E}(K_n) \approx \begin{cases} \alpha\log(n) & \text{if } d=0, \\ \dfrac{\Gamma(1+\alpha)n^d}{d\Gamma(d+\alpha)} - \dfrac{\alpha}{d} & \text{if } d>0. \end{cases} \tag{3}$$

Similarly, the expectation of the number of tables of size m after observing n customers enter, H_{mn}, can be approximated for large n [22, 31] as:

$$\mathbb{E}(H_{mn}) \approx \frac{\Gamma(1+\alpha)n^d}{\Gamma(d+\alpha)m!}\prod_{j=1}^{m-1}(j-d). \tag{4}$$

Therefore, after N observations, simple method of moments or empirical Bayes estimates $(\hat{\alpha}, \hat{d})$ of the hyperparameters α and d are obtained by solving the following non-linear system of equations:

$$H_{1N} = \frac{\Gamma(1+\hat{\alpha})N^{\hat{d}}}{\Gamma(\hat{d}+\hat{\alpha})}, \qquad K_N = \frac{H_{1N}-\hat{\alpha}}{\hat{d}}. \tag{5}$$

A further estimate could be based on an alternative approximation of $\mathbb{E}(K_n)$, often used in the literature [22, 31], obtained by noting that $\lim_{n\to\infty} n^d = \infty$ when $d > 0$. Therefore, when N is large, the value of $\Gamma(1+\alpha)n^d/d\Gamma(d+\alpha)$ will dominate α/d in (3), and hence

$$\hat{d} \approx \frac{H_{1N}}{K_N}. \tag{6}$$

The estimate (6) is particularly illuminating: asymptotically, d is the long term ratio between the tables with just one customer and the total number of tables. Under this approximation, if $\hat{d} = 0$ then the process can be approximated by a Dirichlet process, and from (3),

$$\hat{\alpha} = \frac{K_N}{\log(N)};$$

otherwise, from (5) $\hat{\alpha}$ can be obtained numerically as a solution to the equation

$$K_N = \frac{\Gamma(1+\hat{\alpha})N^{\hat{d}}}{\hat{d}\Gamma(\hat{d}+\hat{\alpha})}. \tag{7}$$

4. A correction for discrete uniform base distributions.

Under the Pitman-Yor generative process, only the sequence of ordered dishes $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ is observed. For discrete base distriubtions, the number of unique dishes, $\tilde{K}_n$, observed in $\boldsymbol{x}_n$ provides only a lower bound for the number of occupied tables, K_n, since the same dish may be drawn from the base distribution multiple times. Similarly, the number of of dishes eaten by only one customer, $\tilde{H}_{1n}$, provides only a lower bound for the number of single-customer tables, H_{1n}.

Assuming a uniform discrete base distribution on a sufficiently large set of nodes V, the approximations in (3) and (4) might be acceptable, but in this particular scenario it is also possible to use a simple correction. Conditioning on the true but unobserved number of draws from the base distribution (equivalently, the number of tables) K_n, a uniform base distribution implies that the expectation of the number of unique draws $\tilde{K}_n$ can be obtained as a generalisation of the *birthday problem* with $|V|$ days in a year:

$$\mathbb{E}(\tilde{K}_n|K_n) = |V| \left\{ 1 - \left(\frac{|V|-1}{|V|} \right)^{K_n} \right\}.$$

Substituting $\mathbb{E}(\tilde{K}_n|K_n)$ with $\tilde{K}_n$ and solving for K_n yields the approximation

$$\hat{K}_n = \log\left(1 - \frac{\tilde{K}_n}{|V|}\right) \log^{-1}\left(\frac{|V|-1}{|V|}\right). \tag{8}$$

Similarly,

$$\mathbb{E}(\tilde{H}_{1n}|H_{1n}, K_n) = H_{1n} \left(\frac{|V|-1}{|V|} \right)^{K_n - 1},$$

and hence the approximation

$$\hat{H}_{1n} = \tilde{H}_{1n} \left(\frac{|V|}{|V|-1} \right)^{\hat{K}_n - 1}. \tag{9}$$

The performance of (8) and (9) as rival estimates to $\tilde{K}_n$ and $\tilde{H}_{1n}$ will be assessed via simulations.

5. Power-laws and Pitman-Yor dynamic graphs.

In this section, some properties of the graph generated from distinct Pitman-Yor processes on each destination node are analysed. After an observation period $[0, T)$, it is possible to construct an adjacency matrix $\mathbf{A} \in \{0,1\}^{|V_S| \times |V_D|}$, where $A_{ij} = 1$ if the source node x_i connected to the destination node y_j at least once, and 0 otherwise. The row and column sums of $\mathbf{A}$ correspond to the out-degree and in-degree sequences:

$$\kappa_i^{\text{out}} = \sum_{j=1}^{|V_D|} A_{ij}, \qquad \kappa_j^{\text{in}} = \sum_{i=1}^{|V_S|} A_{ij}.$$

Real world graphs are commonly characterised by power-law degree distributions: $\kappa \overset{d}{\sim} \text{PL}(\gamma)$ if $p(\kappa) \propto \kappa^{-\gamma}$, $\kappa \geq 1$ for some $\gamma > 1$, typically $2 \leq \gamma \leq 3$ [19]. The in-degree and out-degree distributions in a graph generated from a Pitman-Yor process for each destination node can be controlled by the discount and strength parameters and the base distribution.

First, note that $\kappa_j^{\text{in}} = \tilde{K}_{N_j}$, where N_j is the total number of connections to the destination node y_j. Conditional on the parameters α and d of the process, and for a sufficiently large number of atoms, $\tilde{K}_{N_j} \approx K_{N_j}$, hence $\kappa_j^{\text{in}} \propto N_j^d$ from (3). Therefore, if $N_j \overset{d}{\sim} \text{PL}(\gamma)$, which holds for a suitable choice of α_0 and d_0 in (2), then approximately $\kappa_j^{\text{in}} \overset{d}{\sim} \text{PL}\{(\gamma-1+d)/d\}$.

The out-degree distribution is slightly more difficult to control: it is highly dependent on the choice of the base distribution F_0. For a discrete uniform base distribution over V_S:

$$\mathbb{E}(\kappa_i^{\text{out}}|K_{N_j}) = \sum_{j=1}^{|V_D|}\left\{1-\left(\frac{|V_S|-1}{|V_S|}\right)^{K_{N_j}}\right\},$$

which is not a power law. In order to generate a power law distribution, one can set a non-uniform discrete base distribution $F_0(x) = \pi_x \mathbb{1}_{V_S}(x)$, $\sum_{x\in V_S}\pi_x = 1$, where $\{\pi_x\} \overset{d}{\sim} \text{Dirichlet}(\{\theta_x\})$, and $\theta_x \overset{iid}{\sim} \text{PL}(\gamma)$. In this case:

$$\mathbb{E}(\kappa_i^{\text{out}}|K_{N_j},\{\theta_i\}) = \sum_{j=1}^{|V_D|}\left\{1-\frac{\Gamma(\sum_{i'}\theta_{i'})\Gamma(K_{N_j}+\sum_{i'\neq i}\theta_{i'})}{\Gamma(\sum_{i'\neq i}\theta_{i'})\Gamma(K_{N_j}+\sum_{i'}\theta_{i'})}\right\}$$

and a similar formula can be derived for $\mathbb{E}(\kappa_j^{\text{in}}|K_{N_j},\{\theta_i\})$. Hence, the distribution of κ_i^{out} strongly depends on the interplay between N_j, which contributes to K_{N_j}, and θ_i, which means that it is controlled by α_0, d_0 and G_0 in (2).

6. **Calculating *p*-values for anomaly detection.** For any given destination node $y \in V_D$, a p-value can be computed for each observed source node x_{n+1}, given the history $\boldsymbol{x}_n$. From the posterior predictive (1), the p-value p_{n+1} associated with the $(n+1)$-th connection to the destination node y is

$$p_{n+1} = \sum_{x\in V_S} p(x|\boldsymbol{x}_n)\mathbb{1}_{[0,p(x_{n+1}|\boldsymbol{x}_n)]}\{p(x|\boldsymbol{x}_n)\}. \tag{10}$$

The p-values (10) are discrete and stochastically larger than standard uniform random variables. For this reason, it can be preferable to consider *mid p-values* [15]. Defining the stochastically smaller quantity

$$p_{n+1}^* = \sum_{x\in V_S} p(x|\boldsymbol{x}_n)\mathbb{1}_{[0,p(x_{n+1}|\boldsymbol{x}_n))}\{p(x|\boldsymbol{x}_n)\},$$

where the indicator function instead acts on the half-open interval $[0,p(x_{n+1}|\boldsymbol{x}_n))$ and thus sums all possibilities strictly less probable than the event observed, the mid p-value is given by

$$q_{n+1} = \frac{p_{n+1}+p_{n+1}^*}{2}.$$

In many problems, mid-p-values have been shown to outperform standard p-values, in particular in anomaly detection procedures in computer networks [26].

Similarly, the p-value for the joint event (x_{n+1},y_{n+1}), conditional on $(\boldsymbol{x}_n,\boldsymbol{y}_n)$, is

$$\sum_{x\in V_S,y\in V_D} p(x,y|\boldsymbol{x}_n,\boldsymbol{y}_n)\mathbb{1}_{[0,p(x_{n+1},y_{n+1}|\boldsymbol{x}_n,\boldsymbol{y}_n)]}\{p(x,y|\boldsymbol{x}_n,\boldsymbol{y}_n)\}, \tag{11}$$

where $p(x_{n+1}, y_{n+1}|\boldsymbol{x}_n, \boldsymbol{y}_n) = p(x_{n+1}|y_{n+1}, \boldsymbol{x}_n)p(y_{n+1}|\boldsymbol{y}_n)$ by conditional independence assumptions. The calculation of (11) is intractable, but the p-value could be suitably approximated using a p-value combiner. Given a sequence of p-values $p_1, p_2, \dots, p_\ell$, common choices for p-value combiners are [11]:

- Fisher's method [7]: $S_F = -2\sum_{i=1}^{\ell} \log(p_i) \overset{d}{\sim} \chi^2_{2\ell}$,
- Pearson's method [20]: $S_P = -2\sum_{i=1}^{\ell} \log(1-p_i) \overset{d}{\sim} \chi^2_{2\ell}$,
- Tippett's method [30]: $S_T = \min\{p_i\} \overset{d}{\sim} \mathrm{Beta}(1, \ell)$,
- Stouffer's method [28]: $S_S = \sum_{i=1}^{\ell} \Phi^{-1}(p_i) \overset{d}{\sim} \mathbb{N}(0, \ell)$, where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal distribution.

For approximation of the p-value in (11), $\ell = 2$, with p_1 and p_2 the corresponding p-values for y_{n+1} and $x_{n+1}|y_{n+1}$.

7. **Applications and results.**

The proposed model and estimation procedure have been validated and tested on synthetic networks. Furthermore, the goodness-of-fit of the Pitman-Yor process has been assessed on a real-world network released by the Los Alamos National Laboratory [14].

7.1. **Estimation of the Pitman-Yor hyperparameters.**

The different techniques proposed for the estimation of the Pitman-Yor hyperparameters have been extensively compared using a simulation study, constructed as follows: $S = 10{,}000$ sequences of table allocations, each of length $N = 100{,}000$, have been simulated using the Chinese Restaurant metaphor of the Pitman-Yor process, setting $\alpha = 7$ and $d = 0.25$. For each of the S sequences, the tables have been assigned a label drawn at random, with replacement, from a discrete uniform distribution with $|V| = 1{,}000$ and $|V| = 16{,}230$ atoms (corresponding to the number of source machines in the Los Alamos National Laboratory enterprise network). Matching the sequence of table allocations with the table labels gives a sample sequence from a Pitman-Yor process with uniform discrete base distribution F_0 with $|V|$ atoms. From the resulting sequence, the values of $\tilde{K}_N$ and $\tilde{H}_{1N}$ are calculated, and their corrected counterparts in (8) and (9), $\hat{K}_N$ and $\hat{H}_{1N}$. The values are compared with the true K_N and H_{1N}, available from the simulated table allocation. The parameter estimates $\hat{\alpha}$ and $\hat{d}$ obtained using the pair (6) and (7) are referred to as *Method 1* and (5) as *Method 2*; both are computed using either the uncorrected estimates $(\tilde{H}_{1N}, \tilde{K}_N)$ or the corrected pair $(\hat{H}_{1N}, \hat{K}_N)$, producing four different estimates of the parameters. Kernel density estimates across the simulations for each parameter and each estimation method are plotted in Figures 2 and 3. The kernel bandwidth choice is based on Silverman's rule of thumb [27].

In Figure 2 and 3, the plots *(a)* and *(b)* show that parameter estimation based on the most popular approximation for $\mathbb{E}(K_n)$ in the literature, Method 1, gives biased results. On the other hand, when Method 2 is used to estimate the pair (α, d), and the corrections (8) and (9) are used, the results are excellent, and the proposed estimation procedure on average correctly recovers the exact values of the parameters used to simulate the data. When the correction is not applied, the performance of Method 2 is still fairly reliable for α, but performs poorly for estimating d. Also, the estimates obtained from the uncorrected values K_N and H_{1N}, in plots *(c)* and *(d)* can deviate significantly from the true values.

The plots *(c)* and *(d)* in Figure 2 and 3 show again the importance of the correction for a small number of atoms in the base distribution. For $|V| = 16{,}230$, the

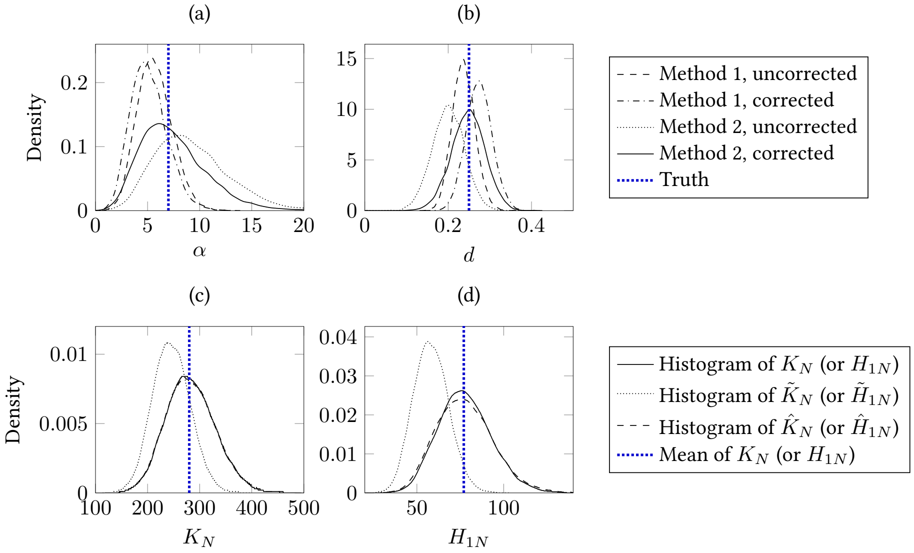

FIGURE 2. Kernel density estimates of the parameter estimates from 10,000 simulations, $|V| = 1{,}000$.

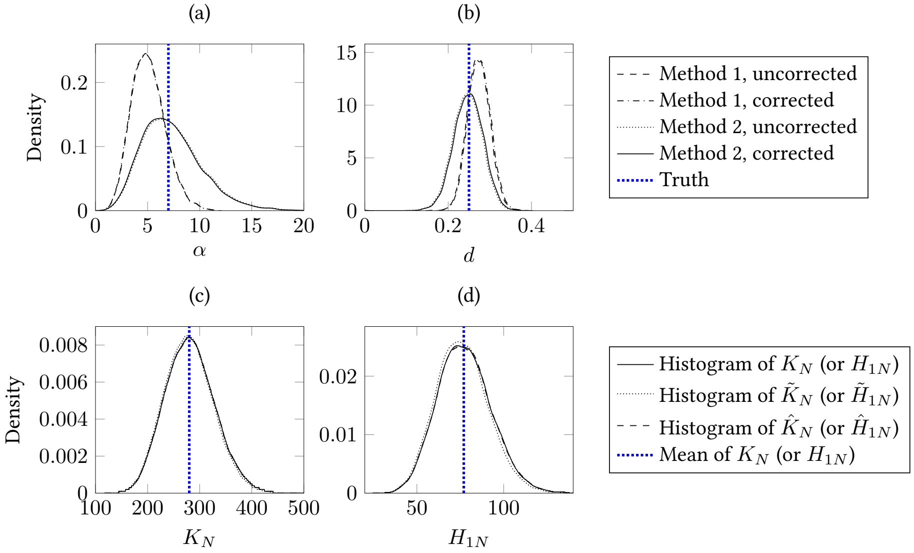

FIGURE 3. Kernel density estimates of the parameter estimates from 10,000 simulations, $|V| = 16{,}230$.

observed value $\tilde{K}_N$ is on average close to the true K_N. This does not happen for $|V| = 1{,}000$, and the approximation becomes crucial.

Overall, the simulation confirms the reliability of the parameter estimation procedure based on (5) for a Pitman-Yor process. The performance of the proposed procedure is far superior to the method based on the estimates (6) and (7), which are obtained from the most common approximation of $\mathbb{E}(K_n)$ used in the literature.

Furthermore, for discrete uniform base distributions, the corrections proposed in equations (8) and (9) are shown to be highly beneficial for parameter estimation, especially when the number of atoms in the base distribution $|V|$ is not large.

7.2. Description of the LANL authentication data.

The proposed model and estimation procedure have been applied to the user-authentication data [14] released by the Los Alamos National Laboratory[1]. An example entry is:

```
1,C567$@DOM1,C567$@DOM1,C574,C988,Kerberos,...
```

The entries of interest in the data-line above are the source computer `C574` and destination computer `C988`, and the arrival time `1` of the event. In total, the data contain 1,051,430,459 events, involving 16,230 source computers and 15,895 destination computers, for a total of 17,684 unique machines and 419,744 unique observed edges. Interestingly, the data also contain 48,079 records labelled as red-team events, resulting from a simulated intrusion. The compromised source nodes are `C17693`, `C18025`, `C19932` and `C22409`. A reliable model must be able to identify unusual activity associated with those source nodes and therefore associate a high anomaly score.

7.3. Empirical assessment of the goodness-of-fit.

In this section, the empirical estimation procedure of the hyperparameters and goodness-of-fit of the Pitman-Yor process is evaluated on selected nodes in the Los Alamos National Laboratory network. In order to assess the model fit, it is possible to examine the Q-Q plot of the p-values, which are approximately uniformly distributed in $(0,1)$ under a correct model specification. Examples of the Q-Q plots observed for six destination nodes are plotted in Figure 4, obtained using the estimate (5) and the corrections (8) and (9). The base distribution F_0 of the Pitman-Yor process was chosen to be uniform on the set of source computers V_S: $F_0(x) = 1/16{,}230 \times \mathbb{1}_{V_S}(x)$. The estimated values of the parameters of the Pitman-Yor process, and additional summary statistics, are reported in Table 1. The Q-Q plots show, for some of the nodes, an excellent fit on real data under the strong assumptions made in this modelling framework (see plots for `C5716`, `U7`, `C2525`, and `C1877`). On the other hand, in some other cases (see plots for `C395` and `C423`) the distribution of the p-values is not uniform, but follows the theoretical distribution only on the left tail.

It is also possible to compare the performance of the method for different choices of α and d. A destination node `C1438` is used as an example. The computer `C1438` appears as destination in $N = 136{,}743$ connections, with $K_N = 394$ unique edges

[1] The data set is available online at `https://csr.lanl.gov/data/cyber1/`.

Table 1. Estimated Pitman-Yor parameters for 6 destination nodes.

Destination	N	$\tilde{K}_n$	$\hat{K}_n$	$\tilde{H}_{1n}$	$\hat{H}_{1n}$	$\hat{\alpha}$	$\hat{d}$
`C5716`	113,987	3,401	3,816.418	144	182.164	651.519	0.047
`U7`	138,286	2,700	2,952.989	350	419.819	302.093	0.142
`C2525`	204,532	1,555	1,634.571	97	107.272	183.319	0.066
`C1877`	342,766	5,095	6,114.758	226	329.390	866.520	0.054
`C395`	518,058	5,957	7,422.437	442	698.259	841.357	0.094
`C423`	2,426,512	2,705	2,958.988	166	199.188	230.040	0.067

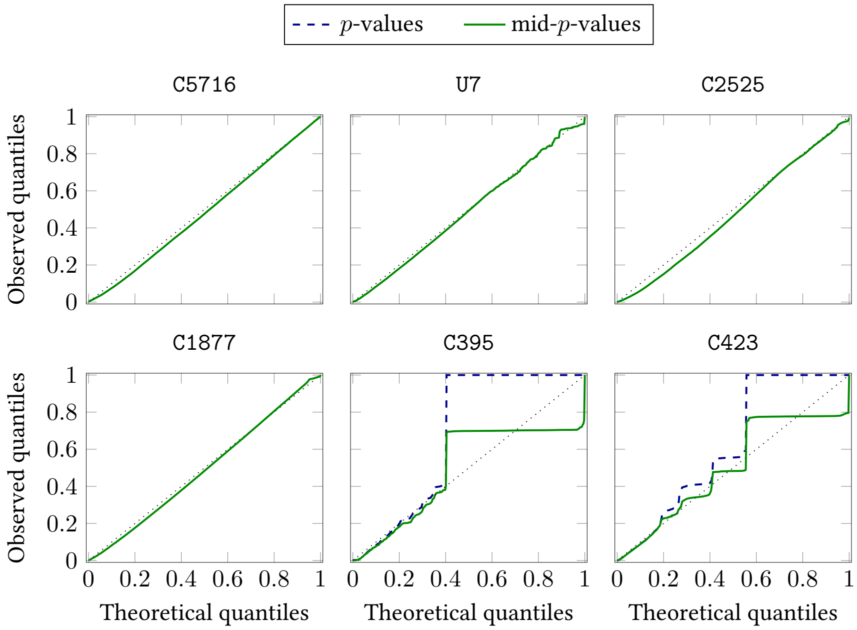

Figure 4. Uniform Q-Q plot for the Pitman-Yor process fitted to six destination nodes.

and $H_{1N} = 72$ source computers that connected only once, resulting in $\hat{\alpha} = 27.434$ and $\hat{d} = 0.113$ after solving (5) with corrections (8) and (9). The three plots in Figure 5 present the sequential values of $\tilde{K}_n$ and $\tilde{H}_{1n}$ and their ratio $\tilde{H}_{1n}/\tilde{K}_n$, corresponding corrected estimates, and average sample paths obtained from Pitman-Yor processes with a number of different values of the parameters.

It is immediately clear from the plots that the only sample path which correctly captures the behaviour of $\tilde{K}_n$, $\tilde{H}_{1n}$ and $\tilde{H}_{1n}/\tilde{K}_n$ simultaneously is the Pitman-Yor process with estimates $\hat{\alpha}$ and $\hat{d}$ obtained using Method 2, with or without correction for discrete F_0. The fit is excellent, and the observed trajectories almost coincides with the average estimates obtained from multiple simulations of the process. The Dirichlet process fails to track $\tilde{K}_n$ and $\tilde{H}_{1n}$ individually, and gives a slightly better performance when modelling the ratio $\tilde{H}_{1n}/\tilde{K}_n$. The Pitman-Yor process with estimates obtained from Method 1 only marginally tracks the data, reaching approximately the observed values $\tilde{K}_N$ and $\tilde{H}_{1N}$ only at the end of the process ($N = 136{,}743$). On the other hand, the ratio is not modelled in a satisfactory way.

Overall, the Pitman-Yor fit is far superior to the Dirichlet process, even without an optimal choice of the parameters, showing that in this case adding the discount parameter d is beneficial. The simulation empirically confirms that the Pitman-Yor process seems to be a suitable choice, but the parameters should be estimated carefully.

7.4. **Network-wide anomaly detection.** Let us suppose that $p_1, \dots, p_N$ are the p-values computed for the N events observed on a given edge $(x, y) \in E$, and $q_1, \dots, q_N$ are the corresponding mid-p-values. Those p-values can be obtained from the sequence $y_{n+1}|\boldsymbol{y}_n$, or from $x_{n+1}|y_{n+1}, \boldsymbol{x}_n$, or from a combination of the

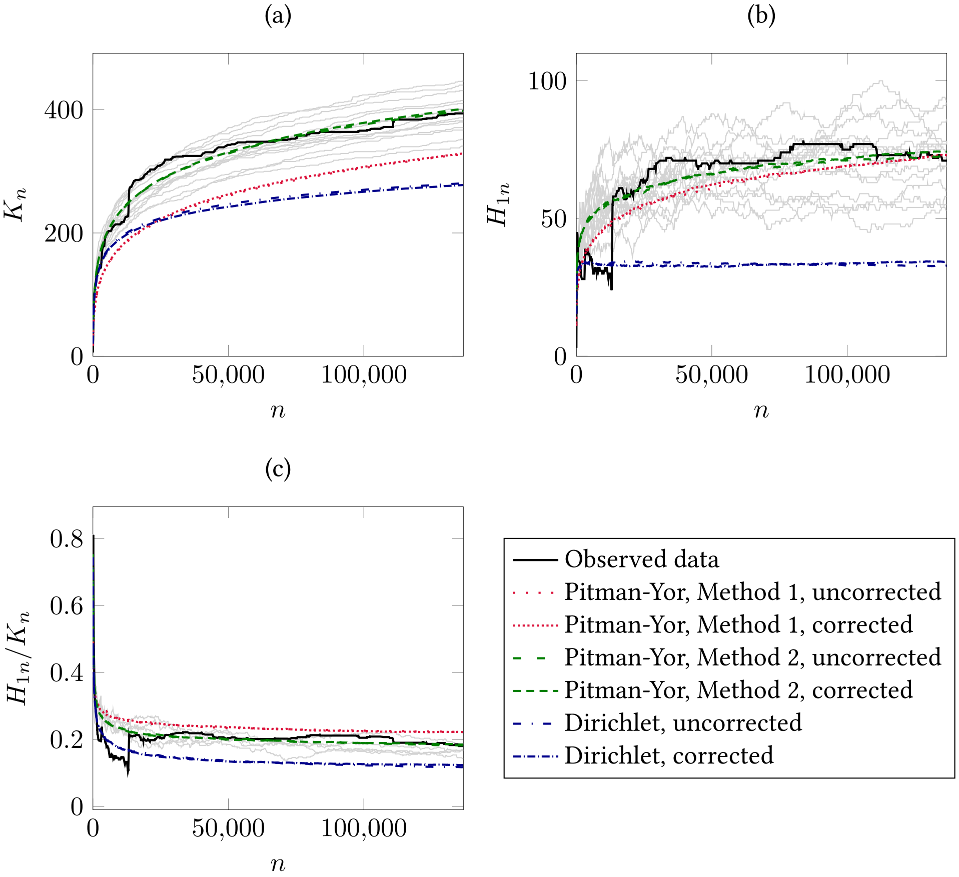

FIGURE 5. Plot of the corrected K_n (a), H_{1n} (b) and their ratio H_{1n}/K_n (c) as a function of n for the connections to the destination computer C1438, and averaged sample paths from 100 samples from a Pitman-Yor process, obtained using different estimates of the parameters. The grey lines correspond to 10 realised trajectories of simulated Pitman-Yor processes, obtained using the corrected estimate (5).

two p-values at the event level, as described in earlier sections. For this analysis, uniform base distributions F_0 and G_0 were used: $F_0(x) = 1/16{,}230 \times \mathbb{1}_{V_S}(x)$, and $G_0(x) = 1/17{,}684 \times \mathbb{1}_V(x)$.

Given the sequence of p-values, it is possible to use combine the scores into a single distribution for each edge. [10] suggests to use the minimum p-value method, or Tippett's method. Note that in this framework the distributional result is only approximate, because of the discreteness of the p-values. Then, the lower tail area of the Beta$(1, N)$ distribution evaluated at $\min\{p_1, \dots, p_N\}$ gives the p-value p_{xy} associated with the edge (x, y).

Following [10], we define as E_x set of edges in the network graph with source node x on which connections have been observed: $E_x = \{(x, y) : y \in V_D \cap (x, y) \in E\}$. The p-values p_{xy} on each edge can be combined into a single score s_x for each

TABLE 2. Anomaly rankings for the four red-team source computers.

		Events $x_{n+1}\|y_{n+1}$ only. Standard p-values p_{n+1}			Events $x_{n+1}\|y_{n+1}$ only. Mid-p-values q_{n+1}		
Edge level combiner: Tippett		Node level combiner:			Node level combiner:		
		Fisher	Pearson	Stouffer	Fisher	Pearson	Stouffer
Source computer	C17693	5	2	4	2	1	5
	C18025	138	75	78	151	74	105
	C19932	3831	8870	8877	3571	2754	3151
	C22409	3767	15773	15764	3450	6984	3756
		Events y_{n+1} only. Mid p-values q_{n+1}			Event level combiner: Tippett. Mid-p-values q_{n+1}		
Edge level combiner: Tippett		Node level combiner:			Node level combiner:		
		Fisher	Pearson	Stouffer	Fisher	Pearson	Stouffer
Source computer	C17693	6	5	5	6	5	5
	C18025	2806	1536	1674	142	96	107
	C19932	5407	8882	8914	3813	2264	3232
	C22409	12126	15808	15878	3803	6516	4196
		Event level combiner: Fisher. Standard p-values p_{n+1}			Event level combiner: Fisher. Mid-p-values q_{n+1}		
Edge level combiner: Tippett		Node level combiner:			Node level combiner:		
		Fisher	Pearson	Stouffer	Fisher	Pearson	Stouffer
Source computer	C17693	3	5	5	5	2	5
	C18025	151	88	101	155	90	106
	C19932	6339	3818	4879	4937	3017	3996
	C22409	6120	14799	5379	4451	6695	5236

source node using one of the combiners previously presented. For example, using the Pearson's combiner, under a normal behaviour of the network, the theoretical distribution of s_x is $s_x \overset{d}{\sim} \chi^2_{2|E_x|}$. Therefore, a p-value p_x for the source computer x is given by the left tail probability of the $\chi^2_{2|E_x|}$ distribution, given the observed s_x. From the list $\{p_x,\ x \in V_S\}$, where V_S is the set of source nodes, it is then possible to identify the most anomalous source computers by ranking the p-values. A similar procedure can be carried out sequentially using the mid-p-values $q_1, \dots, q_N$, or different combiners at the event, edge or node level. The procedure is fully parallelisable and particularly suitable for implementation on standard platforms for Big Data analysis like Hadoop MapReduce, and has been applied to the Los Alamos National Laboratory data.

For each destination computer $y \in V_D$, α_y and d_y are estimated using (5) and the corrections (8) and (9), and similarly for α_0 and d_0. Table 2 reports the results obtained using the minimum p-value method at the edge level and a number of different combiners at the event and node level.

Two of the compromised computers are consistently ranked among the top-1% most anomalous nodes. In particular, C17693 is sometimes ranked as most anomalous machine overall. For the two remaining nodes, associated with more subtle activity within the network, impressive improvements are achieved when using mid-p-values, but the malicious activity on these nodes is still difficult to detect.

Overall, the results confirm the conclusions in [26] about the improved detection performance when using mid-p-values. This is particularly evident in the case of the two least anomalous compromised machines: using mid-p-values, the ranking improves significantly. The activity associated with `C17693` is anomalous even when only the p-values associated with the sequence of destination nodes $y_1, y_2, \ldots$ is considered. For the other three machines, the rankings significantly improve when the edge activity from the p-values for $x_{n+1}|y_{n+1}$ is used for computing the anomaly scores. Using the combined scores at the event level does not significantly improve the results, showing that most of the malicious activity might be due to edge-related anomalies.

8. Conclusion.

In this article, a model for events in dynamic networks is presented. The sequence of edges is modelled by using the Pitman-Yor process, a two-parameter extension of the Dirichlet process. Empirical methods for choosing the hyperparameters, based on large sample results, have been proposed and tested on the Los Alamos National Laboratory authentication data, showing excellent fit on the activity of real destination computers. Furthermore, corrections for discrete base distributions are carefully addressed. The Pitman-Yor process allows for more flexible and accurate modelling of the tails of the predictive distribution, which also implies more accurate modelling of the new links in the network. The model has been tested on synthetic data and applied to the Los Alamos National Laboratory authentication dataset, showing good performance in a network-wide anomaly detection study.

The Pitman-Yor model might be used as a building block for more complex models. For example, in computer networks, the arrival time of each event is also available, and suitable models for arrival times on each edge have been proposed in the literature [24]. The overall performance of the method is impressive, since only two pieces of information are used in this article: the source and destination node. The methodology can be potentially extended to consider any additional information which is usually available for each event: for example, in computer network data, authentication type and logon type.

Acknowledgements. The authors gratefully acknowledge funding from the EPSRC and the Heilbronn Institute for Mathematical Research.

REFERENCES

[1] D. Aldous, Exchangeability and related topics, in *École d'Été de Probabilités de Saint-Flour XIII-1983*, 1985, 1–198.

[2] A. L. Barabási, The origin of bursts and heavy tails in human dynamics, *Nature*, **435** (2005), 207–211.

[3] Y. Bengio, R. Ducharme, P. Vincent and C. Janvin, A neural probabilistic language model, *Journal of Machine Learning Research*, **3** (2003), 1137–1155.

[4] W. Buntine and M. Hutter, A Bayesian view of the Poisson-Dirichlet process, preprint, arXiv:1007.0296.

[5] C. Chen, L. Du and W. Buntine, Sampling table configurations for the hierarchical Poisson-Dirichlet process, in *Machine Learning and Knowledge Discovery in Databases* (eds. D. Gunopulos, T. Hofmann, D. Malerba and M. Vazirgiannis), Springer Berlin Heidelberg, 2011, 296–311.

[6] T. S. Ferguson, A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, **1** (1973), 209–230.

[7] R. A. Fisher, *Statistical Methods for Research Workers*, vol. 4th ed., Edinburgh: Oliver & Boyd, 1934.

[8] A. Goldenberg, A. X. Zheng, S. E. Fienberg and E. M. Airoldi, A survey of statistical network models, *Foundations and Trends in Machine Learning*, **2** (2009), 129–233.

[9] S. Goldwater, T. L. Griffiths and M. Johnson, Interpolating between types and tokens by estimating power-law generators, in *Proceedings of the 18th International Conference on Neural Information Processing Systems*, MIT Press, 2005, 459–466.

[10] N. A. Heard and P. Rubin-Delanchy, Network-wide anomaly detection via the Dirichlet process, in *Proceedings of the IEEE workshop on Big Data Analytics for Cyber-security Computing*, 2016.

[11] N. A. Heard and P. Rubin-Delanchy, Choosing between methods of combining p-values, *Biometrika*, **105** (2018), 239–246.

[12] H. Ishwaran and L. F. James, Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association*, **96** (2001), 161–173.

[13] D. Jurafsky, J. H. Martin, P. Norvig and S. Russell, *Speech and Language Processing*, Pearson Education, 2014.

[14] A. D. Kent, Cybersecurity data sources for dynamic network research, in *Dynamic Networks and Cyber-Security*, World Scientific, 2016.

[15] H. Lancaster, Statistical control of counting experiments, *Biometrika*, **39** (1952), 419–422.

[16] Y. Lv and C. X. Zhai, Positional language models for information retrieval, in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2009, 299–306.

[17] C. Matias and V. Miele, Statistical clustering of temporal networks through a dynamic stochastic block model, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79** (2017), 1119–1141.

[18] T. Mikolov, M. Karafiát, L. Burget, J. Černocký and S. Khudanpur, Recurrent neural network based language model, in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, International Speech Communication Association, 2010, 1045–1048.

[19] M. E. J. Newman, Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*, **46** (2005), 323–351.

[20] K. Pearson, On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random, *Biometrika*, **25** (1933), 379–410.

[21] P. O. Perry and P. J. Wolfe, Point process modelling for directed interaction networks, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75** (2013), 821–849.

[22] J. Pitman, *Combinatorial stochastic processes*, Technical Report 621, Department of Statistics University of California at Berkeley, 2002.

[23] J. Pitman and M. Yor, The two-parameter Poisson-Dirichlet distribution derived from a stable sub-ordinator, *Annals of Probability*, **25** (1997), 855–900.

[24] M. Price-Williams and N. A. Heard, Nonparametric self-exciting models for computer network traffic, *Statistics and Computing*, 2019.

[25] R. Rosenfeld, A maximum entropy approach to adaptive statistical language modelling, *Computer Speech & Language*, **10** (1996), 187 – 228.

[26] P. Rubin-Delanchy, N. A. Heard and D. J. Lawson, Meta analysis of mid-p-values: some new results based on the convex order, *Journal of the American Statistical Association*, 2018.

[27] B. W. Silverman, *Density Estimation*, London: Chapman and Hall, 1986.

[28] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star and R. M. Williams, *The American Soldier. Adjustment During Army Life*, Princeton, New Jersey: Princeton University Press, 1949.

[29] W. Y. Teh, 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 of Computational Linguistics*, 2006, 985–992.

[30] L. H. C. Tippett, *The Methods of Statistics*, London: Williams and Norgate, 1931.

[31] S. M. Wallach, S. T. Jensen, L. Dicker and K. A. Heller, An alternative prior process for nonparametric Bayesian clustering, in *Proceedings of the Thireenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, 2010, 892–899.

Received June 2019; revised August 2019.

E-mail address: francesco.sanna-passino16@imperial.ac.uk
E-mail address: n.heard@imperial.ac.uk