Long-tailed distributions of inter-event times as mixtures of exponential distributions

Makoto Okada¹, Kenji Yamanishi¹ and Naoki Masuda²,³,⁴

¹Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo Bunkyo, Tokyo 113-8656, Japan
²Department of Engineering Mathematics, University of Bristol, Woodland Road, Clifton, Bristol BS8 1UB, UK
³Department of Mathematics, University at Buffalo, State University of New York, Buffalo, NY 14260-2900, USA
⁴Computational and Data-Enabled Science and Engineering Program, University at Buffalo, State University of New York, Buffalo, NY 14260-5030, USA

Inter-event times of various human behaviour are apparently non-Poissonian and obey long-tailed distributions as opposed to exponential distributions, which correspond to Poisson processes. It has been suggested that human individuals may switch between different states, in each of which they are regarded to generate events obeying a Poisson process. If this is the case, inter-event times should approximately obey a mixture of exponential distributions with different parameter values. In the present study, we introduce the minimum description length principle to compare mixtures of exponential distributions with different numbers of components (i.e. constituent exponential distributions). Because these distributions violate the identifiability property, one is mathematically not allowed to apply the Akaike or Bayes information criteria to their maximum-likelihood estimator to carry out model selection. We overcome this theoretical barrier by applying a minimum description principle to joint likelihoods of the data and latent variables. We show that mixtures of exponential distributions with a few components are selected, as opposed to more complex mixtures in various datasets, and that the fitting accuracy is comparable to that of state-of-the-art algorithms to fit power-law distributions to data. Our results lend support to Poissonian explanations of apparently non-Poissonian human behaviour.

1. Introduction

Many social and economic processes are a consequence of human behaviour. Technological advances are increasingly enabling us to
record various human behaviours in quantitative manners. Such quantitative understanding often challenges traditional assumptions underlying mathematical modelling of human behaviour, a major instance of which is a lack of Poissonian properties in a range of human behavioural data. In other words, when a sequence of time-stamped events, such as email correspondence, is observed from a human individual, the event sequence often deviates from Poisson processes, in which inter-event times independently obey an exponential distribution. Rather, empirical inter-event times from human behaviour and other data often obey long-tailed distributions [1, 2]. Such non-Poissonian processes on nodes and edges are building blocks of temporal networks on which dynamical processes such as epidemic processes often behave differently from the same processes on static networks [3, 4].

Two classes of models that generate long-tailed distributions of inter-event times are priority queue models and modulated Markov processes [3, 4]. In priority queue models, one assumes that a human individual is a queue that receives tasks across time and prioritizes some particular tasks for execution [1]. By contrast, in modulated Markov processes, one assumes that a human individual generates events according to a Poisson process (i.e. one determines whether to have an event right now at a constant rate without memory) but modulates the event rate. The event rate may be assumed to take one of a few values [5–10] or continuously many values [11, 12]. When we assume a few states, where each state corresponds to a Poisson process, an underlying assumption about human behaviour is that an individual transits among a few distinguishable states, such as an active state and an inactive state. If this is the case, we should be able to fit a mixture of exponential distributions rather than power-law distributions to empirical distributions of inter-event times with a reasonable accuracy. This is because a subset of inter-event times, i.e. those produced in each state, is expected to obey an exponential distribution.

In fact, a mixture of a small number of exponential distributions and a power-law distribution with an exponential cut-off can look similar. In the present study, we fit mixtures of exponential distributions to several datasets of inter-event times. The idea of approximating empirical distributions resembling power-law distributions by mixtures of exponential distributions is not a new idea [13]. We perform model selection to determine the number of component exponential distributions fitted to data and also compare the mixture of exponential distributions with two types of conventionally used power-law distributions. A technical challenge is that the maximum-likelihood estimator of the mixture of exponential distributions does not satisfy the asymptotic normality such that it is not allowed to use the Akaike information criterion (AIC) or Bayes information criterion (BIC) (e.g. [14, p. 203]). Therefore, based on the minimum description length (MDL) criterion [15, 16], we perform the procedure so-called latent variable completion [17–20] to derive variants of MDL justified for mixtures of exponential distributions. We find that mixtures of up to three exponential distributions are the best performer in many cases. Python codes for estimating and selecting the mixture of exponential distributions are available at Github (https://github.com/naokimas/exp_mixture_model).

2. Methods

2.1. Mixture of exponential distributions and its maximum-likelihood estimation

We fit mixtures of exponential distributions, which we refer to as the exponential mixture models, abbreviated as EMMs, to distributions of inter-event times, denoted by \( r \in [0, \infty) \), and compare the fit with the case of power-law distributions. We denote by \( k \) the number of the exponential distributions to be mixed, which we refer to as the number of component distributions. The mixing weight, i.e. the probability that the \( j \)th exponential distribution is used, is denoted by \( \pi_j \) (1 \( \leq j \leq k \)). The probability density function for an EMM is given by

\[
p(r; \pi, \mu) = \sum_{j=1}^{k} \pi_j \exp\left(-\frac{r}{\mu_j}\right),
\]

where \( \pi = (\pi_1, \ldots, \pi_k) \), \( \mu = (\mu_1, \ldots, \mu_k) \) and \( \mu_j \) is the mean of the \( j \)th exponential distribution.

Given a series of inter-event times, \( \tau = [\tau_1, \ldots, \tau_n] \), and the number of components, \( k \), we estimate the values of the model parameters, \( \theta = (\pi, \mu) \), as follows.

First, we run the expectation-maximization (EM) algorithm [21]. As an initial condition, we set \( \pi_j^{(0)} = 1/k \) (1 \( \leq j \leq k \)), where the superscript indicates the iteration number in the EM procedure. We also draw \( \mu_j^{(0)} \) (1 \( \leq j \leq k \)) such that \( \log_{10} \mu_j^{(0)} \) independently obeys the uniform density on \([\log_{10}\tau_{\text{min}}, \log_{10}\tau_{\text{max}}]\).
\[ \log(10 \tau_{\text{max}}), \text{ where } \tau_{\text{min}} \text{ and } \tau_{\text{max}} \text{ are the minimum and maximum of the inter-event time in the given dataset, respectively. In this manner, we intend to generate a sufficiently broad initial distribution of } \mu_j^{(0)} \text{ while a majority of the } \mu_j^{(0)} \text{ values is relatively small. Intuitively, a large value of } \mu_j^{(0)} \text{, which sometimes occurs, reflects the long tail of the distribution of inter-event times. We iterate the following EM steps } t_{\text{max}} \text{ times by alternating the expectation (E) and maximization (M) steps. The E step is given by} \]

\[ \gamma_j^{(t)} = \frac{\pi_j^{(t)}}{\sum_{j=1}^{k} \pi_j^{(t)}}, \]

\[ \text{where } t = 0, 1, 2, \ldots, t_{\text{max}} - 1. \text{ One can interpret } \gamma_j^{(t)} \text{ as the probability that } \tau_i \text{ is generated from the } j \text{th exponential distribution. The M step is given by} \]

\[ \pi_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_j^{(t)}, \]

\[ \mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_j^{(t)} \tau_i}{\sum_{j=1}^{k} \gamma_j^{(t)}}. \]

After the iteration of the E and M steps \( t_{\text{max}} \) times, one obtains \( \pi_1^{(\text{max})}, \ldots, \pi_k^{(\text{max})} \) and \( \mu_1^{(\text{max})}, \ldots, \mu_k^{(\text{max})} \), which is an approximate maximum-likelihood estimator. This estimator corresponds to the case in which one does not estimate the latent variables that explicitly indicate which exponential distribution out of the \( k \) exponential distributions has generated each \( \tau_i \).

When the values of the latent variables are required, we compute them as follows. We denote the latent variable corresponding to each \( \tau_i (1 \leq i \leq n) \) by \( z_i (1 \leq z_i \leq k) \). In other words, \( \tau_i \) is estimated to be generated from the \( z_i \)th exponential distribution. We first estimate the latent variables by

\[ \hat{z}_i = \arg \max_{j=1, \ldots, k} \gamma_j^{(\text{max})} \]

\[ = \arg \max_{j=1, \ldots, k} [\pi_j^{(\text{max})} p(\tau_i; \mu_j^{(\text{max})})]. \]

We then estimate the other parameters by

\[ \tilde{\pi}_j(\hat{z}) = \frac{n_j}{n}, \]

\[ \text{where } \hat{z} = [\hat{z}_1, \ldots, \hat{z}_n], n_j \text{ is the number of inter-event times whose estimated latent variable } \hat{z}_i \text{ is equal to } j, \text{ and} \]

\[ \tilde{\mu}_j(\hat{z}) = \frac{1}{n_{\hat{z}_j}} \sum_{i=1, z_i = j}^{n} \tau_i. \]

The \( \tilde{\pi}_j \) and \( \tilde{\mu}_j \) values given by equations (2.5) and (2.6) are a maximum-likelihood estimator when the joint distribution of the inter-event times (i.e. \( \tau_1, \ldots, \tau_n \)) and the latent variables (i.e. \( z_1, \ldots, z_n \)) is estimated. They are different from the values estimated by the EM algorithm (i.e. \( \pi_1^{(\text{max})}, \ldots, \pi_k^{(\text{max})} \) and \( \mu_1^{(\text{max})}, \ldots, \mu_k^{(\text{max})} \)).

To cope with the problem of local maxima, we estimate the values of \( \tilde{\pi}_j, \tilde{\mu}_j, \) and \( \hat{z}_i (1 \leq j \leq k, 1 \leq i \leq n) \) 10 times starting from different initial conditions. Then, among the 10 maximum-likelihood estimators, we employ the one that has yielded the largest likelihood of the joint distribution of \( \tau \) and \( \hat{z} \).

### 2.2. Model selection criteria for mixtures of exponential distributions

We consider the following six model selection criteria.

#### 2.2.1. AIC and BIC

For an EMM with \( k \) components without an explicit consideration of the latent variables, the AIC [22] and the BIC [23] are given by

\[ \text{AIC}(\tau) = -\log p(\tau; \hat{\theta}_{\text{EM}}(\tau)) + 2k - 1 \]
In equations (2.9), (2.10) and (2.11), \( \hat{\theta}_{\text{EM}}(\tau) \) is the EM estimator of the parameters of the EMM, i.e. \( \pi_1^{(n_{1,1})}, \ldots, \pi_k^{(n_{1,1})}, \mu_1^{(n_{1,1})}, \ldots, \mu_k^{(n_{1,1})} \).

### 2.2.2. AIC and BIC with latent variable completion

An EMM is non-identifiable, which by definition dictates that a distribution does not correspond to a parameter set in a one-to-one manner [14]. For example, if \( k=2 \) and \( \mu_1 = \mu_2 \), the distribution of \( \tau \) is the same exponential distribution regardless of the value of \( \pi_1 (1 - \pi_2) \). When a distribution is non-identifiable, the maximum-likelihood estimator is not asymptotically normal, and the conventional AIC and BIC, which are derived on the basis of asymptotic normality, lose justification [14].

A method for overcoming this theoretical barrier is to apply model selection criteria to a complete variable model, in which one completes the values of the latent variables, rather than to use a marginalized model [17–20]. We call this method the latent variable completion [20] throughout this paper. Under latent variable completion, one explicitly incorporates the likelihood of the latent variables, \( z_i (1 \leq i \leq n) \), without marginalizing them, into a model selection criterion. Then, the joint distribution of \( \tau_i \) and \( z_i (1 \leq i \leq n) \) is an identifiable distribution such that the use of the AIC and BIC is justified. Also in practice, latent variable completion is better at describing some empirical data (for example, at estimating an appropriate number of components, \( k \)) than model selection criteria that do not use latent variable completion [17,18].

We refer to the corresponding AIC or BIC as the AIC or BIC with latent variable completion [17,19] and denote them by \( \text{AIC}_{\text{LVC}} \) and \( \text{BIC}_{\text{LVC}} \), respectively. We obtain

\[
\text{AIC}_{\text{LVC}}(\tau, z) = - \log p(\tau, z; \hat{\theta}(\tau, z)) + 2k^* - 1 \tag{2.9}
\]

and

\[
\text{BIC}_{\text{LVC}}(\tau, z) = - \log p(\tau, z; \hat{\theta}(\tau, z)) + \frac{k^* - 1}{2} \log n + \frac{1}{2} \sum_{j=1}^{k^*} \log n_j. \tag{2.10}
\]

In equations (2.9) and (2.10), \( \hat{\theta}(\tau, z) = \{ \hat{\pi}_1(z), \ldots, \hat{\pi}_{k^*}(z), \hat{\mu}_1(z), \ldots, \hat{\mu}_{k^*}(z) \} \), derived in equations (2.5) and (2.6), is the maximum-likelihood estimator for the joint distribution of inter-event times \( \tau \) and latent variables \( z \). The likelihood in equations (2.9) and (2.10) is given by

\[
p(\tau, z; \hat{\theta}(\tau, z)) = \prod_{j=1}^{k^*} (\hat{\pi}_j)^{n_j} p((\pi)_j^{n_{1,1}}) = \prod_{j=1}^{k^*} \left( \frac{\pi_j}{\hat{\mu}_j} \right)^{n_j}. \tag{2.11}
\]

In equations (2.9), (2.10) and (2.11), \( k^* \) is the number of components that are used at least once under the estimated latent variable values. The latent variable specifies which of the \( k \) exponential distributions has produced \( \tau_i \) for each \( i \). Therefore, as a result of estimating the latent variables, a \( j \)th exponential distribution may not have any inter-event times \( \tau_i \) belonging to it (i.e. \( n_j = 0 \)). We exclude such exponential distributions, i.e. those with \( n_j = 0 \), and only consider the remaining \( k^* \) exponential distributions. To calculate \( \text{AIC}_{\text{LVC}} \) and \( \text{BIC}_{\text{LVC}} \), we use \( k^* \) instead of \( k \) because \( k^* \) is the actual number of components given the estimated latent variable values. For the same reason, we will use \( k^* \) instead of \( k \) in the two model selection criteria introduced in the following sections, which also explicitly use the estimated latent variables and latent variable completion.

### 2.2.3. Normalized maximum-likelihood codelength

Here, we introduce model selection criteria on the basis of the MDL principle [16]. This principle asserts that the best model should be the one that attains the shortest codelength required for encoding the data as well as the model itself. The codelength is the number of bits required for encoding the data into a binary sequence under the information-theoretic requirement that each codeword can be uniquely decodable even without commas. In fact, MDL approaches to mixture modelling have a long history, as represented by the Snob program [24,25].

The normalized maximum-likelihood (NML) codelength is a type of MDL [16,26]. The NML codelength minimizes the worst-case (i.e. in terms of data \( x \)) regret value, which is equal to the actual
codelength minus the ideal codelength, i.e. \( \min_{\theta \in \Theta} (-\log p(x; \theta)) \) [27]. The other two model selection criteria, which we will explain in the next two sections, are based on the NML. Therefore, we explain the NML codelength in this section. Although we assume that inter-event times \( \tau \) are continuous-valued, we first explain the NML codelength for a sequence of discrete-valued variables \( x = \{x_1, \ldots, x_n\} \) for expository purposes.

To shorten the codelength for the given data \( x \), one should in principle minimize the Shannon information given by \(-\log p(x; \theta)\), where we assume throughout the present paper that \( \log \) is in base \( e \) unless we specify the base. The minimizer is obviously given by the maximum-likelihood estimator, \( \hat{\theta}(x) \). However, the maximum-likelihood estimator generally yields

\[
\sum_x p(x; \hat{\theta}(x)) > 1,
\]  
(2.12)

because \( \hat{\theta} \) depends on \( x \). In equation (2.12), the summation is taken over all possible values of \( x \). Therefore, we use the normalized probability distribution given by

\[
p_{\text{NML}}(x) = \frac{p(x; \hat{\theta}(x))}{\sum_x p(x'; \hat{\theta}(x'))}
\]  
(2.13)

to encode the data [28]. In other words, we set

\[
L_{\text{NML}}(x) = -\log p_{\text{NML}}(x) = -\log p(x; \hat{\theta}(x)) + \log C(n),
\]  
(2.14)

where

\[
C(n) = \sum_x p(x'; \hat{\theta}(x'))
\]  
(2.15)

is called the parametric complexity. In equation (2.14), the first term on the right-hand side is small when the model fits the data well, and the second term represents the complexity of the model. In general, \( C(n) \) tends to increase as \( k \) increases because an increase in \( k \) leads to the expansion of the parameter space in which one searches \( \theta \), which leads to an increase in \( C(n) \) [29].

One can similarly derive the NML codelength for a sequence of continuous-valued variables by replacing the summation by the integral. For example, the equivalent of equation (2.15) in the case of continuous-valued variables is given by

\[
C(n) = \int p(x'; \hat{\theta}(x')) \, dx'_1 \cdots dx'_n.
\]  
(2.16)

In the remainder of this section, we explain the NML codelength for an exponential distribution [19,30], not for an EMM, for two reasons. First, analytically calculating the NML codelength for an EMM seems formidable. Second, we will use the NML codelength for single exponential distributions in deriving the two types of the NML-based codelengths for EMMs in the following sections.

Consider an exponential distribution given by

\[
p(\tau; \mu) = \frac{1}{\mu} \exp \left(-\frac{\tau}{\mu}\right).
\]  
(2.17)

For this distribution, we obtain

\[
L_{\text{NML}}(\tau) = -\log p(\tau; \hat{\mu}(\tau)) + \log C_{\exp}(n)
\]  
(2.18a)

and

\[
C_{\exp}(n) = \int_{T[\mu_{\min}, \mu_{\max}]} p(\tau'; \hat{\mu}(\tau')) \, d\tau'_1 \cdots d\tau'_n,
\]  
(2.18b)

where \( T[\mu_{\min}, \mu_{\max}] \) is a region of \( \tau = \{\tau_1, \ldots, \tau_n\} \in \mathbf{R}^n \) such that the maximum-likelihood estimator given by

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \tau'_i
\]  
(2.19)

is contained in \([\mu_{\min}, \mu_{\max}]\). We have to confine the domain of integration to \( T \) in this manner because the
integral in equation (2.18b) diverges if \( T \) is replaced by \( \mathbb{R}^n \). In practice, if we take \( \mu_{\text{min}} (> 0) \) small enough and \( \mu_{\text{max}} \) large enough, \( \hat{\mu} \in [\mu_{\text{min}}, \mu_{\text{max}}] \) would be satisfied for empirical sequences of inter-event times \( \tau \). The standard method to evaluate the integral in equation (2.18b) is to transform the integral to that in the parameter space \([19,31]\). This method yields an integral of the form similar to \( \int_{\mu_{\text{min}}}^{\mu_{\text{max}}}(1/\hat{\mu})d\hat{\mu} \), which is manageable \([19]\). Region \( T \) is a useful heuristic for avoiding such a divergence in NML-related calculations \([31]\).

To incorporate the codelength necessary for encoding the value of \( \mu_{\text{min}} \) and \( \mu_{\text{max}} \) into the NML codelength, we rewrite \( \mu_{\text{min}} = \exp(m_{\text{min}}) \) and \( \mu_{\text{max}} = \exp(m_{\text{max}}) \), where \( m_{\text{min}} \) and \( m_{\text{max}} \) are integers. Then, we encode \( m_{\text{min}} \) and \( m_{\text{max}} \) instead of \( \mu_{\text{min}} \) and \( \mu_{\text{max}} \).

The first term on the right-hand side of equation (2.18a) is given by

\[
\log p(\tau, \hat{\mu}(\tau)) = \log \prod_{i=1}^{n} \frac{1}{\hat{\mu}} \exp \left( -\frac{\tau_i}{\hat{\mu}} \right) = -n \log \hat{\mu} - \frac{1}{\hat{\mu}} \sum_{i=1}^{n} \tau_i = -n \log \hat{\mu} - n \tag{2.20}
\]

where \( \hat{\mu} \) is given by equation (2.19). To obtain the second term on the right-hand side of equation (2.18a), we substitute equation (2.17) into equation (2.18b), which leads to

\[
\log C_{\text{exp}}(n) = n \log n - n - \log \Gamma(n) + \log \frac{\mu_{\text{max}}}{\mu_{\text{min}}} = n \log n - n - \log \Gamma(n) + \log (m_{\text{max}} - m_{\text{min}}), \tag{2.21}
\]

where \( \Gamma(n) \) is the gamma function.

By substituting equations (2.20) and (2.21) into equation (2.18a), one obtains

\[
L_{\text{NML}}(\tau) = n \log \hat{\mu} + n \log n - \log \Gamma(n) + \log (m_{\text{max}} - m_{\text{min}}). \tag{2.22}
\]

To carry out model selection, we add the codelength for \( m_{\text{min}} \) and \( m_{\text{max}} \), denoted by \( \ell(m_{\text{min}}) \) and \( \ell(m_{\text{max}}) \), to \( L_{\text{NML}}(\tau) \) to obtain

\[
\tilde{L}_{\text{NML}}(\tau) = L_{\text{NML}}(\tau) + \ell(m_{\text{min}}) + \ell(m_{\text{max}}). \tag{2.23}
\]

The derivation of \( \ell(m_{\text{min}}) \) and \( \ell(m_{\text{max}}) \) is given in the electronic supplementary material.

### 2.2.4. Normalized maximum-likelihood codelength with latent variable completion

As a first NML type of the model selection criterion for EMMs, we consider a latent variable completion of the NML codelength. We refer to the resulting criterion by NML\(_{\text{LVC}}\) and the codelength by \( L_{\text{LVC}} \). Although an analytical expression for the NML codelength for an EMM without latent variable completion is unavailable, we can analytically calculate the NML codelength for the joint distribution of inter-event times and the latent variables.

Similar to the derivation in the case of a mixture of Gaussian distributions \([19]\), we derive the codelength for EMMs as follows:

\[
L_{\text{LVC}}(\tau, \hat{\theta}; \hat{\tau}) = - \log p(\tau, \hat{\theta}; \hat{\tau}) + \log C_{\text{EMM}}(n, k^*), \tag{2.24}
\]

where

\[
C_{\text{EMM}}(n, k^*) = \sum_{\hat{\tau}_1}^{k_1} \cdots \sum_{\hat{\tau}_k}^{k_k} \int_{T(\mu_{\text{min}}, \mu_{\text{max}})} p(\hat{\tau}_1, \cdots, \hat{\tau}_k; \hat{\theta}(\hat{\tau}_1, \cdots, \hat{\tau}_k)) d\hat{\tau}_1 \cdots d\hat{\tau}_k. \tag{2.25}
\]

We remind that \( p(\hat{\tau}_1, \cdots, \hat{\tau}_k; \hat{\theta}(\hat{\tau}_1, \cdots, \hat{\tau}_k)) \) is given by equation (2.11). Region \( T'(\mu_{\text{min}}, \mu_{\text{max}}) \) of \( \tau \in \mathbb{R}^n \) is defined such that each of \( \hat{\mu}_1, \cdots, \hat{\mu}_k \) in the maximum-likelihood estimator \( \hat{\theta}(\hat{\tau}_1, \cdots, \hat{\tau}_k) \) when the latent variables are explicitly estimated (i.e. equation (2.6)) is contained in \([\mu_{\text{min}}, \mu_{\text{max}}]\). Region \( T'(\mu_{\text{min}}, \mu_{\text{max}}) \) coincides with region \( T(\mu_{\text{min}}, \mu_{\text{max}}) \) when \( k^* = 1 \). Similarly to the case of equation (2.18b), we confine the domain of integration to \( T' \) to avoid the divergence of the integral in equation (2.25). We set

\[
\mu_{\text{min}} = \exp(m_{\text{min}}), \tag{2.26}
\]
where

\[ m'_{\text{min}} = \left\lfloor \log \left( \min_{j=1, \ldots, k'} \hat{\mu}_j \right) \right\rfloor, \tag{2.27} \]

and

\[ \mu'_{\text{max}} = \exp \left( m'_{\text{max}} \right), \tag{2.28} \]

where

\[ m'_{\text{max}} = \left\lfloor \log \left( \max_{j=1, \ldots, k'} \hat{\mu}_j \right) \right\rfloor, \tag{2.29} \]

for two reasons. First, \( \log C_{\text{EMM}}(n, k') \) is small when \( \mu'_{\text{min}} \) is large or \( \mu'_{\text{max}} \) is small. Second, encoding of \( \mu'_{\text{min}} \) and \( \mu'_{\text{max}} \) is facilitated by the introduction of integer variables, as we did for the NML codelength (§2.2.3).

By substituting

\[ -\log p(\tau, \hat{z}; \hat{\theta}(\tau, \hat{z})) = -\log \prod_{j=1}^{k'} \left\{ \left( \frac{n_j}{\bar{n}} \right)^{n_j} p(\tau_{j=1,\ldots,\hat{z}_j}^{\hat{\mu}_j}) \right\} = -n \sum_{j=1}^{k'} n_j \log \frac{n_j}{\bar{n}} - \sum_{j=1}^{k'} \log p(\tau_{j=1,\ldots,\hat{z}_j}^{\hat{\mu}_j}) \]

into equation (2.24), where \( H(q_i, \ldots, q_k) = \sum_{j=1}^{k'} -q_j \log q_j \) represents the entropy, one obtains

\[ L_{\text{LVC}}(\tau, \hat{z}) = nH\left( \frac{n_1}{n}, \ldots, \frac{n_k}{n} \right) + \sum_{j=1}^{k'} n_j \log \hat{\mu}_j + n + \log C_{\text{EMM}}(n, k'). \tag{2.31} \]

By substituting equation (2.11) in equation (2.25), one obtains the parametric complexity as follows:

\[ C_{\text{EMM}}(n, k') = \sum_{\bar{n}_1, \ldots, \bar{n}_k = n} \frac{n!}{\bar{n}_1! \cdots \bar{n}_k!} \prod_{j=1}^{k'} \left\{ \left( \frac{n_j}{n} \right)^{\bar{n}_j} C_{\text{EMM}}(\bar{n}_j, 1) \right\}, \tag{2.32} \]

where

\[ C_{\text{EMM}}(n, 1) = \int_{\mathcal{T}(\mu_{\text{max}}, \mu_{\text{min}})} p(\tau; \bar{\mu}(\tau)) d\tau_1 \cdots d\tau_n \]

\[ = \left( \frac{n}{e} \right)^{m'_{\text{max}} - m'_{\text{min}}} T(n). \tag{2.33} \]

Note that we have used equation (2.21) to derive equation (2.33) and that \( C_{\text{EMM}}(n, 1) \) coincides with \( C_{\exp}(n) \). Because we cannot calculate \( C_{\text{EMM}}(n, k') \) using equation (2.32) due to combinatorial explosion, we calculate \( C_{\text{EMM}}(n, k') \) using the recursive relationship [19] given by

\[ C_{\text{EMM}}(n, k+1) = \sum_{r_1, r_2 = 2} \left( \frac{n}{r_1} \right)^{r_1} \left( \frac{n}{r_2} \right)^{r_2} C_{\text{EMM}}(r_1, k) C_{\text{EMM}}(r_2, 1). \tag{2.34} \]

The calculation of \( C_{\text{EMM}}(n, k') \) using this recursive relationship, which dominates the computation time for calculating \( L_{\text{LVC}}(\tau, \hat{z}) \), requires \( O(n^2 k') \) time.

Finally, we add the codelength to encode integers \( m'_{\text{min}} \) and \( m'_{\text{max}} \) to obtain

\[ \tilde{L}_{\text{LVC}}(\tau, \hat{z}) = L_{\text{LVC}}(\tau, \hat{z}) + \ell(m'_{\text{min}}) + \ell(m'_{\text{max}}). \tag{2.35} \]

### 2.2.5. Decomposed NML codelength

The second type of the NML codelength with latent variable completion, which we call the decomposed NML (DNML) codelength [20,32], also completes the latent variable and then calculates the NML codelength. The DNML does so after decomposing the joint distribution into the distribution of the latent variables, \( \hat{z} \), and that of the observables, \( \tau \), conditioned on the value of \( z \) [20]. The DNML codelength was originally formulated for the case in which both observables and latent variables were...
discrete [20]. Here, we regard the inter-event time, \( \tau_i \), as continuous-valued and the latent variable, \( z_i \), as discrete-valued.

We start by considering

\[
L_{DNML}(\tau, z) = L_{NML}(\tau|z) + L_{NML}(z).
\]

Equation (2.22) yields

\[
L_{NML}(\tau|z) = \sum_{j=1}^{k-1} L_{NML}(|\tau|_{i=1,z_{i-j}}) = \sum_{j=1}^{k'} |n_j \log \mu_j + n_j \log n_j - \log \Gamma(n_j)| + k' \log (\max_m - \min_m),
\]

where \( \min_m \) and \( \max_m \) are given by equations (2.27) and (2.29), respectively. It should be noted that we could prepare integers \( \min_m \) and \( \max_m \) for each \( j \) (1 \( \leq j \leq k' \)) to code \( |\tau|_{i=1,z_{i-j}} \) through \( \mu_j \) similarly to equation (2.23), to derive an alternative of equation (2.37). However, that coding method requires \( 2k' \) integers and therefore is less efficient than using only two integers \( \min_m \) and \( \max_m \).

The NML codelength \( L_{NML}(z) \) on the right-hand side of equation (2.36) is for a multinomial distribution of \( z_i \) and is given by

\[
L_{NML}(z) = nH(\frac{n_1}{n}, \ldots, \frac{n_{k'}}{n}) + \log \mult(c, k').
\]

In equation (2.38), \( \mult(c, k') \) is the parametric complexity for the multinomial distribution having \( k' \) elements. Using equation (2.5), one obtains

\[
\mult(c, k') = \sum_{\bar{z}_i=1}^{k} \cdots \sum_{\bar{z}_{i-1}=1}^{k} \prod_{j=1}^{k'} (\frac{\bar{z}_j}{\bar{z}_i})^{n_j} = \sum_{\bar{z}_i=1}^{k} \cdots \sum_{\bar{z}_{i-1}=1}^{k} \prod_{j=1}^{k'} (\frac{\bar{z}_j}{n})^{n_j},
\]

where \( n_j \) is the number of inter-event times whose latent variable \( \bar{z}_i \) is equal to \( j \). One can recursively calculate \( \mult(c, k') \) [33] by

\[
\begin{align*}
\mult(c, 1) & = 1, \\
\mult(c, 2) & = \sum_{i=0}^{c} \frac{n!}{i!(n-i)!} \left( \frac{i}{n} \right)^{t} \left( \frac{n-i}{n} \right)^{n-t}, \\
\mult(c, k) & = \mult(c, k-1) + \frac{n}{k-2} \mult(c, k-2) \quad (k \geq 3).
\end{align*}
\]

The computational time for solving the set of recursive equations is given by \( O(n + k') \) and dominates the computational time for \( L_{DNML}(\tau, z) \).

Finally, similarly to the case of NML-LVC, we add the codelength to encode integers \( \min_m \) and \( \max_m \) to obtain

\[
L_{DNML}(\tau, z) = L_{DNML}(\tau, z) + \ell(\min_m) + \ell(\max_m).
\]

2.3. Power-law distributions

We also fitted two types of power-law distributions to the empirical distributions of inter-event times. The first type is the Pareto distribution whose probability density function is given by

\[
p(\tau, a, b) = \frac{a-1}{b} (\frac{\tau}{b})^{-a},
\]

which is defined for \( \tau \geq b \). We use the maximum-likelihood estimator given by [34,35]

\[
\hat{b}(\tau) = \min_{1 \leq i \leq n} \tau_i
\]

and

\[
\hat{a}(\tau) = 1 + \frac{1}{(1/n) \sum_{i=1}^{n} \log(\tau_i) - \log(b)}.
\]

The second estimator was the one proposed by Clauset et al. [36]. In this method, which we refer to as the PLFit algorithm, one selects the \( \hat{b} \) value such that the distribution of the data points satisfying \( \tau_i \geq \hat{b} \) is as close as possible to a power law. The power-law exponent given the \( \hat{b} \) value is
estimated by equation (2.44). It should be noted that the PLFit does not intend to fit a power-law or different distribution to the data points whose values are less than $b$. We used a publicly available Python implementation of the PLFit [37].

3. Results

3.1. Data

We used the following seven datasets of time-stamped event sequences obtained from human behaviour. Basic properties of each dataset are shown in table 1.

First, we used the data that the SocioPatterns project collected from individuals in an office building that hosted scientific departments [38]. We refer to this dataset as the Office dataset. The individuals wore a sensor on their chest, with which physical proximity between pairs of individuals was detected. The recording lasted for two weeks. A time-stamped event was defined as a contact that lasted at least 20 s, which was the time resolution of the data. For each individual, we ignored the partner of the contact and used the beginning time of each contact as the time of the event. In this manner, we examined a sequence of time-stamped events for each individual in this and the following datasets. Note that, because a contact is a symmetric relationship between two individuals, any event between nodes $i$ and $j$ is considered for both $i$ and $j$.

The inter-event time was defined as the difference between the starting time of the two consecutive events. There was no event during the night-time due to the circadian rhythm. The effect of the circadian rhythm on analysis of inter-event times may be considerable but beyond the scope of the present study. Therefore, we excluded the inter-event times when the two events belonged to different days unless otherwise stated. When there were multiple events in consecutive 20 s time windows between the same pair of individuals, we regarded that they constitute a single event lasting over 20 s rather than a sequence of events with inter-event times equal to 20 s. In practice, in this case, we discarded all but the first event in each sequence of the multiple events in consecutive 20 s time windows and then calculated the inter-event times. We also aggregated events for a focal node that occurred at the same time with different partners into one event. Therefore, the minimum inter-event time is equal to 20 s. For the individuals that have at least 100 inter-event times, the mean and standard deviation of the inter-event times are shown in table 1.

Second, we used the Hypertext 2009 dynamic contact network (Hypertext for short) [39]. Under the SocioPatterns project, the data were collected from approximately 75% of the participants in the HT09 conference in Torino over three days. As was the case for the Office dataset, we ignored inter-event times that span multiple days.

The third dataset originates from the Reality Mining Project (Reality for short), which provides time-stamped physical proximity relationships between the participants of an experiment, who are students or
faculty members of MIT. The data were obtained from Bluetooth logs of mobile phones [40]. We did not skip inter-event times overnight because the data were collected over months and the mean inter-event time was much longer than that for the Office and Hypertext datasets (table 1). We used the data from September 2004 to May 2005, which were the months in each of which there were at least 103 events in total.

Fourth, Bitcoin OTC is an over-the-counter marketplace where users trade with bitcoin [41]. In Bitcoin OTC, users rate other users regarding trustworthiness. We neglected the rate score and who a focal user i rated. For each user i, we examined a series of time-stamped events, where an event was defined by rating behaviour by user i toward anybody. We refer to this dataset as Bitcoin.

The Email-Eu-core (Email for short) dataset was collected from a large European research institution between October 2003 and May 2005 [42]. A node was an email address. An edge was defined between two nodes if and only if they sent email to each other at least once. A time-stamped event for node i in the network was defined by an email that i sent to anybody, disregarding the identity of the recipient of the email.

The CollegeMsg (College for short) data were collected from students belonging to an online community at the University of California, Irvine, between April and December 2004 [43]. The event was defined as the message sent from a user to another user. We downloaded the Bitcoin, Email and College data from the Stanford Network Analysis Project (SNAP) website (http://snap.stanford.edu/).

The sexual contact (Sexual for short) dataset provides times of online sexual encounters between escorts and sex buyers [44]. We used event sequences for each sex buyer by regarding a commercial sexual activity with any escort as an event.

3.2. Fitting of the EMM to the individual with the largest number of events and comparison with power-law distributions

We fitted EMMs with different numbers of components, k, to the sequence of inter-event times obtained from each individual in each dataset. We examined k = 1, 2, …, 9, 10, 20, 50 and 100 under each criterion. For the Sexual dataset, k = 1, 2 or 3 were ties when AICLVC, BICLVC, NMLLVC or DNML was used. However, these four k values effectively produced the same exponential distribution with k’ = 1. Therefore, in the table, we wrote k = 1 for these four criteria. Note that, for any criterion, the same k’ value induced by different initial k values may yield different criterion values because the estimated EMM may be different between the different k values even if the k’ value is the same. On the other hand, if different criteria are maximized at the same k value, they are maximized by the same EMM in addition that they share the k’ value.

| data     | n   | AIC | BIC | AICLVC | BICLVC | NMLLVC | DNML |
|----------|-----|-----|-----|--------|--------|--------|------|
| Office   | 403 | (2, 2) | (2, 2) | (2, 2) | (2, 2) | (2, 2) | (2, 2) |
| Hypertext| 659 | (2, 2) | (2, 2) | (2, 2) | (2, 2) | (2, 2) | (2, 2) |
| Reality  | 1198 | (3, 3) | (3, 3) | (2, 2) | (2, 2) | (2, 2) | (2, 2) |
| Bitcoin  | 604 | (5, 4) | (3, 3) | (4, 3) | (4, 3) | (4, 3) | (4, 3) |
| Email    | 3948 | (8, 6) | (8, 6) | (4, 4) | (4, 4) | (4, 4) | (4, 4) |
| College  | 1090 | (8, 6) | (4, 4) | (3, 3) | (3, 3) | (3, 3) | (3, 3) |
| Sexual   | 118 | (1, 1) | (1, 1) | (1, 1) | (1, 1) | (1, 1) | (1, 1) |
Next, we compared the performance between the estimated EMMs and power-law estimators in approximating the empirical distributions. We have fitted power-law distributions because they have widely been applied to empirical distributions of inter-event times [1–4]. For the same individuals as those used in table 2, the empirical and estimated distributions of inter-event times are compared in figure 1. When the different model selection criteria yielded different EMMs (i.e. EMMs with different \(k\) and \(k^*\) values), we showed all the selected EMMs overlaid in the figure. We compared the different distributions in terms of the survival probability (i.e. the probability with which the inter-event time is larger than \(\tau\)), i.e. \(\int_{\tau}^{\infty} p(\tau')d\tau'\) and the odds ratio defined by \(\text{OR}(\tau) = [1 - (n'/n) \int_{\tau}^{\infty} p(\tau')d\tau'] / [(n'/n) \int_{\tau}^{\infty} p(\tau')d\tau']\). Note that the odds ratio is particularly good at capturing differences between distributions at small values of \(\tau\) [45].

Figure 1 elicits the following observations. First, in five datasets (i.e. except Bitcoin and Email) the right tails of the distribution seem to be more accurately approximated by an EMM than the power-law distributions. This may be because the two power-law estimators employed here assume a strictly power-law tail, whereas empirical data usually have exponential cut-offs. For the Bitcoin and Email datasets, the power-law estimate obtained by the PLFit algorithm seems to be roughly as accurate as EMMs in approximating right tails of the distribution. Second, the odds ratio plots show that, at small \(\tau\) values, the approximation looks the most accurate with the Pareto distribution in all but the College dataset.

To examine whether these casual observations extend to the entire set of inter-event times from the same individuals, we investigated the likelihood of the empirical data in each model. The likelihood for
the three models, i.e. the selected EMM, maximum-likelihood estimator of the Pareto distribution, and PLFit, is compared in table 3. Because the PLFit discards small inter-event times, we start by comparing the EMM and Pareto distribution when all inter-event times are used. The table indicates that the EMM realizes a larger likelihood value than the Pareto distribution in five datasets and vice versa in the other two datasets (i.e. Hypertext and Reality). The EMM has more parameters than the Pareto distribution, which one may suspect as a reason why the EMM fits the data better than the Pareto distribution in more datasets than vice versa. However, the results are qualitatively the same when the two models are compared in terms of the AIC and BIC (electronic supplementary material, table S1). We remark that the maximum-likelihood estimator for the Pareto distribution is not asymptotically normal such that the application of the AIC and BIC to the Pareto distribution as well as to EMMs is not justified. It should be noted that the other four criteria are not relevant to the Pareto distribution because it is not a latent variable model.

For the Hypertext and Reality datasets, for which the likelihood is larger for the Pareto distribution than the optimal EMM, there are relatively many data points with $\tau = \tau_{\text{min}}$ (table 3). The Pareto distribution beats the EMM probably because it avoids devoting the probability mass to values less than $\tau_{\text{min}}$ by definition, and it is accurate at many data points that have $\tau = \tau_{\text{min}}$. To clarify this point, we compared the likelihood when the data points with $\tau = \tau_{\text{min}}$ were excluded. The results are shown in table 3. To simplify the discussion, for the EMM we employed the value of $k$ that minimized the AIC. For the Reality data, the EMM fits the data better than the Pareto distribution when we only consider the inter-event times larger than $\tau_{\text{min}}$. For the Hypertext dataset, the Pareto distribution is still better than the EMM, but the difference in the likelihood is much smaller than when all the inter-event times are used. Note that the Reality dataset has a larger fraction of inter-event times with $\tau = \tau_{\text{min}}$ than the Hypertext dataset (table 3).

The PLFit only fits a power-law distribution to relatively large inter-event times. Therefore, we compared the likelihood of the EMM, Pareto distribution, and PLFit for the set of inter-event time values satisfying $\tau \geq \hat{b}$, where $\hat{b}$ is the threshold estimated by the PLFit. The PLFit is the best
performer among the three fitting schemes (table 3). This is probably because the threshold is optimized for the PLFit in this comparison and the PLFit is allowed to discard small inter-event times. By contrast, the EMM intends to fit the entire set of inter-event times. Furthermore, in this comparison, the EMM is disadvantageous by the amount of probability mass that it has to devote to inter-event times smaller than \( \hat{b} \). It should be noted that the PLFit discards a considerable fraction of data points, ranging between 27 and 94\%, except for the College dataset (table 3). It should also be noted that the EMM yields a larger likelihood value than the Pareto distribution for all the datasets with this \( \hat{b} \) value.

We do not conclude whether or not the EMM outperforms the Pareto distribution or PLFit. Both the Pareto distribution and PLFit benefit from carefully setting the lower bound on the inter-event times, \( \hat{b} \). When all inter-event time data have to be modelled and we do not know the smallest possible value of the inter-event time, EMMs may be preferred to the Pareto distribution and the PLFit.

So far, we have excluded the inter-event times across consecutive days in the Office and Hypertext datasets. We also ran the same analysis for these two datasets without excluding such long inter-event times. The results were qualitatively the same except that the EMM estimated the number of components larger by one (electronic supplementary material). The new component (i.e. a single exponential distribution in the EMM) corresponded to the across-day inter-event times, with a considerably longer mean inter-event time than that for the other estimated components. Furthermore, the mean inter-event time of the constituent exponential distributions that exist both when across-day inter-event times are included and when excluded are almost the same in the two cases (electronic supplementary material). We conclude that inter-event times on the timescale of a day or longer contribute one exponential distribution with a large mean to the entire EMM without interfering with the constituent exponential distributions on smaller time scales.

### 3.3. Population results

In the previous section, we showed that mixtures of a small number of exponential distributions tended to be selected according to the different model selection criteria. To assess the generality of this result, we carried out the model selection for all individuals having at least 100 inter-event times. Because the Sexual dataset has only two individuals with at least 100 inter-event times, we instead used the 56 individuals that had at least 50 inter-event times for this dataset. The histogram of the optimal \( k \) value for each combination of the dataset and criterion is shown in figure 2. Consistently with the results for the individuals with the largest number of inter-event times in each dataset, the effective number of components, \( k' \), was estimated to be at most three in four out of the seven datasets (i.e. Office, Hypertext, Reality and Sexual). For the Bitcoin dataset, the distributions of \( k' \) had the mode at 3, and we obtained \( 2 \leq k' \leq 4 \) in most cases. For the other two datasets (i.e. Email and College), the AIC and BIC yielded distributions of \( k' \) whose mode was four and the largest value was seven. Under the other four criteria, which are the ones justified for EMMs, the mode was equal to three, and the largest \( k' \) was equal to five.

To exclude the possibility that longer data (i.e. larger \( n \)) tend to yield a larger \( k' \) value, we calculated the Pearson correlation coefficient between \( n \) and the optimal \( k' \) for each combination of the dataset and criterion. To calculate the Pearson correlation coefficient, we only used the individuals with at least 100 inter-event times (and at least 50 inter-event times for the Sexual dataset), similar to figure 2. The Pearson correlation coefficient values, denoted by \( r \), and its \( p \)-value are shown in table 4. The correlation coefficient was small and insignificant except for the Email dataset combined with all criteria and the College dataset combined with the AIC or BIC. When we confined the analysis to the individuals with larger numbers of inter-event times, i.e. at least 500 and 200 inter-event times for the Email and College datasets, respectively, the Pearson correlation coefficient was smaller. In particular, for the AIC\textsubscript{LVC}, BIC\textsubscript{LVC}, NML\textsubscript{LVC} and DNML, the correlation was close to zero and insignificant when we only considered the individuals with many inter-event times. Note that the largest number of inter-event times for these two datasets were sufficiently larger than the new threshold values (Email: \( n = 3948 \), College: \( n = 1090 \)) and that there remained sufficiently many individuals with the new threshold values (Email: 117 individuals, College: 61 individuals). The results suggest that our main result that the optimal number of components, \( k' \), tends to be small is expected to remain true for long sequences of inter-event times.

### 4. Discussion

We showed that EMMs (i.e. mixtures of exponential distributions) with a small number of components, up to three in many cases and four in some cases, were a reasonably good fit to various empirical
distributions of inter-event times. In general, there are various mechanisms behind power-law distributions of an observable, and many of them are believed to be relevant as generative mechanisms of long-tailed degree distributions of empirical networks [46]. By contrast, the present results suggest that empirical long-tailed distributions can be also approximated by EMMs. Furthermore, EMMs can be obtained as an outcome of stochastic point processes in which the system’s state switches between a small number of discrete states in each of which the process generates events as a Poisson process. In practice, a human or animal individual may maintain a

Figure 2. Distributions of the effective number of components in the EMM, $k^*$. (a) Office, (b) Hypertext, (c) Reality, (d) Bitcoin, (e) Email, (f) College and (g) Sexual. For the AIC and BIC, the model selection is carried out in terms of $k$, and the $k^*$ values shown in this figure are those corresponding to the selected $k$ values. We calculated the distributions on the basis of the individuals with at least 100 inter-event times with the exception of the Sexual dataset, for which we used the individuals with at least 50 inter-event times (see table 1 for the number of such individuals).
small number of relatively discrete states, such as active and rest. The event rate may depend on the discrete state. It should be noted that, within each state, the mechanism to generate event sequences is maximally simple: Poisson process with a constant rate. To the best of our knowledge, Poissonian views of long-tailed distributions of inter-event times were first introduced in [5]. The contribution of the present work is to have introduced a formal model selection framework to directly compare EMMs with different numbers of components and having carried out some comparisons with power-law distributions.

There is a recent debate regarding the abundance of scale-free (i.e. power-law) degree distributions in empirical networks. When a pure power-law tail is assumed for the degree distribution, there are empirically few scale-free networks [47]. By contrast, scale-free degree distributions are abundant when impurity in the power-law distributions via the regularly varying distributions is considered [48]. In the present study, we fitted the pure power-law distributions via the PLFit algorithm [36]. The PLFit was more accurate than EMMs while the PLFit discarded small inter-event times, which occupied a non-negligible fraction of inter-event times in each dataset. We did not examine the method presented in [48]. This is because the purpose of the present study was not to find an accurate estimator or to test if a power-law distribution of inter-event times was a good fit. Rather, our purpose was to test the hypothesis that the system generating time-stamped events, such as humans, transit between a small number of discrete states each of which is a Poisson process generator.

A reasonably accurate fit of the EMMs to the empirical data revealed in the present study mainly comes from a high accuracy at small \( r \) values, because there are many data points with small \( r \). Such short inter-event times yield a component exponential distribution with a small mean inter-event time and may be produced as a result of the visit of the individual to an active state. Given these considerations, looking at the entire distribution of the data rather than focusing on the distribution’s tail may help us to understand mechanisms generating the data.

The authors of [10] reached a similar conclusion to ours, where they split the inter-event times into two groups by thresholding. Then, they fitted an exponential distribution to the inter-event times exceeding the threshold and another exponential distribution to the inter-event times less than the
threshold. Their models are EMMs with two components. In contrast to their study, we systematically fitted EMMs and inferred the number of components using model selection criteria valid for EMMs. Testing our method on various other distributions of inter-event times to clarify the reason why EMMs with a small number of components apply accurately to some datasets and not others warrants future work.

Data accessibility. The data are open resources. Python codes for estimating and selecting the mixture of exponential distributions are available at Github (https://github.com/nakimas/esp_mixure_model).

Authors’ contributions. K.Y. and N.M. conceived the study. N.M. designed the study. M.O. carried out the theoretical analysis and performed the data analysis. M.O., K.Y. and N.M. discussed the results and drafted the manuscript. All authors read and approved the manuscript.

Competing interests. The authors declare that they have no competing interests.

Funding. K.Y. and N.M. acknowledge the support provided through JST CREST grant no. JPMJCR1304, Japan.

Acknowledgements. We thank the SocioPatterns collaboration (http://www.sociopatterns.org) for providing the two datasets used in the present paper.

References

1. Barabási AL. 2005 The origin of bursts and heavy tails in human dynamics. Nature 435, 207–211. (doi:10.1038/nature03499)
2. Vázquez A, Oliveira JG, D dessz Z, Goh KI, Kondor I, Barabási AL. 2006 Modeling bursts and heavy tails in human dynamics. Phys. Rev. E 73, 066127. (doi:10.1103/PhysRevE.73.066127)
3. Holme P, Saramäki J. 2012 Temporal networks. Phys. Rep. 519, 97–125. (doi:10.1016/j.physrep.2012.03.001)
4. Masuda N, Lambiotte R. 2016 A guide to temporal networks. Singapore: World Scientific.
5. Malmgren RD, Stouffer DB, Motter AE, Amaral LAN. 2008 A Poissonian explanation for heavy tails in e-mail communication. Proc. Natl Acad. Sci. USA 105, 18.153–18.158. (doi:10.1073/pnas.0808332105)
6. Malmgren RD, Stouffer DB, Campanhão ASU, Amaral LAN. 2009 On universality in human correspondence activity. Science 325, 1696–1700. (doi:10.1126/science.1174562)
7. Karsai M, Kaski K, Barabási AL, Kertész J. 2012 Universal features of correlated bursty behaviour. Sci. Rep. 2, 2012. (doi:10.1038/srep00397)
8. Vajna S, Toth B, Kertész K. 2013 Modelling bursty time series. New J. Phys. 15, 103023. (doi:10.1088/1367-2630/15/10/103023)
9. Raghavan V, Ver Steeg G, Galstyan A, Tartakovsky AG. 2014 Modeling temporal activity patterns in dynamic social networks. IEEE Trans. Comput. Soc. Netw. 1, 89–107. (doi:10.1109/TCSN.2013.2307453)
10. Ji ang ZQ, Xie WJ, Li MX, Zhou WX, Sornette D. 2016 Two-state Markov-chain Poisson nature of individual cellphone call statistics. J. Stat. Mech. 2016, 073210. (doi:10.1088/1742-5468/2016/ 07/073210)
11. Masuda N, Takaguchi T, Sato N, Yano K. 2013 Self-exciting process modeling of conversation event sequences. In Temporal networks (eds P Holme, J Saramäki), pp. 245–264. Berlin, Germany: Springer.
12. Masuda N, Rocha LEC. 2018 A Gillespie algorithm for non-Markovian stochastic processes. SIAM Rev. 60, 95–115. (doi:10.1137/ 16M1055876)
13. Fiedmann A, Whitt W. 2002 Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. Perform. Eval. 31, 245–279. (doi:10.1016/S0166-5316(97)00030-5)
14. McLachlan G, Peel D. 2000 Finite mixture models. Hoboken, NJ: John Wiley & Sons.
15. Rissanen J. 1978 Modeling by shortest data description. Automatica 14, 465–471. (doi:10.1016/0005-1098(78)90005-5)
16. Rissanen J. 2012 Optimal estimation of parameters. Cambridge, UK: Cambridge University Press.
17. Kontkanen P, Myllymäki P, Buntine W, Rissanen J, Timi H. 2005 An MDL framework for data clustering. In Advances in minimum description length: theory and applications (eds PD Grunwald, U Myung, MA Pitt), pp. 323–353. Cambridge, MA: MIT Press.
18. Geuliev G, Forbes F, Robert CP, Titterington DM. 2006 Deviance information criteria for missing data models. Bayesian Anal. 1, 651–673. (doi:10.1214/06-BA122)
19. Hirai S, Yamashiki K. 2013 Efficient computation of normalized maximum likelihood codes for Gaussian mixture models with its applications to clustering. IEEE Trans. Inf. Theory 59, 7718–7727. (doi:10.1109/TIT.2013.2276036)
20. Wu T, Sugawara S, Yamashiki K. 2017 Decomposed normalized maximum likelihood codelength criterion for selecting hierarchical latent variable models. Data Mining Knowl. Disc. 33, 1017–1058. (doi:10.1007/s10618-019-00624-4)
21. Kontkanen P, Myllymäki P. 2007 A linear-time algorithm for computing the multinomial distribution. J. R. Stat. Soc. Ser. B 69, 123–157. (doi:10.1111/j.1467-9868.2007.00553.x)
22. Alkaik H. 1974 A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723. (doi:10.1109/TAC.1974.1100705)
23. Schwarz G. 1978 Estimating the dimension of a model. Ann. Stat. 6, 461–464. (doi:10.1214/ aos/1176344136)
24. Wallace CS, Boulton DM. 1968 An information measure for classification. Comput. J. 11, 185–194. (doi:10.1093/comjnl/11.2.185)
25. Wallace CS, Dove DL. 2000 MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. Stat. Comput. 10, 73–83. (doi:10.1023/A:1008992619036)
26. Rissanen J. 1996 Fisher information and stochastic complexity. IEEE Trans. Inf. Theory 42, 40–47. (doi:10.1109/18.481776)
27. Starkev YM. 1987 Universal sequential coding of single messages. Probl. Peredachi Inf. 23, 3–17.
28. Rissanen J. 1986 Stochastic complexity and modeling. Ann. Stat. 14, 1080–1100. (doi:10. 1214/aos/1176350551)
29. Barron A, Rissanen J, Yu B. 1998 The minimum description length principle in coding and modeling. IEEE Trans. Inf. Theory 44, 2734–2760. (doi:10.1109/18.720554)
30. Schmidt DF, Makalic E. 2009 Universal models for the exponential distribution. IEEE Trans. Inf. Theory 55, 3087–3090. (doi:10.1109/TIT.2009.2018331)
31. Rissanen J. 2000 MDL denoising. IEEE Trans. Inf. Theory 46, 2537–2541. (doi:10.1109/18.887661)
32. Yamashiki K, Wu T, Sugawara S, Okada M. 2019 The decomposed normalized maximum likelihood code-length criterion for selecting hierarchical latent variable models. Data Mining Knowl. Disc. 33, 1107–1107. (doi:10.1007/s10618-019-00624-4)
33. Kontkanen P, Myllymäki P. 2007 A linear-time algorithm for computing the multimodal stochastic complexity. Inf. Process. Lett. 103, 227–233. (doi:10.1016/j.ipl.2007.04.003)
34. Malik HJ. 1970 Estimation of the parameters of the Pareto distribution. Metrika 15, 126–132. (doi:10.1007/BF01803565)
35. Galati S, Shapiro S. 2008 Goodness-of-fit tests for Pareto distribution. In Statistical models and methods for biomedical and technical systems (eds F Vonta, M Nikulin, N Limnios, C HuberCarol), pp. 259–274. Boston, MA: Birkhäuser Boston.
36. Clarizia A, Shalizi CR, Newman MEJ. 2009 Power-law distributions in empirical data. SIAM Rev. 51, 661–703. (doi:10.1137/ 070710111)
37. Alstott J, Bullmore E, Plenz D. 2014 powerlaw: a Python package for analysis of heavy-tailed distributions. PLoS ONE 9, e85777. (doi:10.1371/journal.pone.0085777)

38. Genois M, Vestergaard CL, Fouquet J, Panisson A, Bornain I, Barrat A. 2015 Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. Netw. Sci. 3, 326–347. (doi:10.1017/nws.2015.10)

39. Isella L, Stehle J, Barrat A, Cattuto C, Pinton JF, Van den Broeck W. 2011 What’s in a crowd? Analysis of face-to-face behavioral networks. J. Theor. Biol. 271, 166–180. (doi:10.1016/j.jtbi.2010.11.033)

40. Eagle N, Pentland A. 2006 Reality mining: sensing complex social systems. Pers. Ubiquit. Comput. 10, 255–268. (doi:10.1007/s00779-005-0046-3)

41. Kumar S, Spezzano F, Subrahmanian V$, Faloutsos C. 2016 Edge weight prediction in weighted signed networks. In 2016 IEEE 16th Intl. Conf. Data Mining (ICDM), Barcelona, Spain, 12–15 December, pp. 221–230. IEEE.

42. Leskovec J, Kleinberg J, Faloutsos C. 2007 Graph evolution: densification and shrinking diameters. ACM Trans. Knowl. Disc. Data 1, 2. (doi:10.1145/1217299.1217301)

43. Panzarasa P, Opsahl T, Carley KM. 2009 Patterns and dynamics of users’ behavior and interaction: network analysis of an online community. J. Amer. Soc. Inf. Sci. Technol. 60, 911–932. (doi:10.1002/asi.21015)

44. Rocha LEC, Liljeros F, Holme P. 2010 Information dynamics shape the sexual networks of internet-mediated prostitution. Proc. Natl Acad. Sci. USA 107, 5706–5711. (doi:10.1073/pnas.0914080107)

45. Vaz De Melo POS, Faloutsos C, Assunção R, Alves R, Loureiro AAF. 2015 Universal and distinct properties of communication dynamics: how to generate realistic inter-event times. ACM Trans. Knowl. Disc. Data 9, 24. (doi:10.1145/2700399)

46. Newman MEJ. 2005 Power laws, Pareto distributions and Zipf’s law. Contem. Phys. 46, 323–351. (doi:10.1080/00107510500052444)

47. Broido AD, Clauset A. 2019 Scale-free networks are rare. Nat. Commun. 10, 1017. (doi:10.1038/s41467-019-08746-5)

48. Voitalov I, van der Hoorn P, van der Hofstad R, Krioukov D. 2019 Scale-free networks well done. Phys. Rev. Res. 1, 033034. (doi:10.1103/PhysRevResearch.1.033034).

49. Voitalov I, van der Hoorn P, van der Hofstad R, Krioukov D. 2019 Scale-free networks well done. Phys. Rev. Res. 1, 033034. (doi:10.1103/PhysRevResearch.1.033034).