A stochastic evolutionary model generating a mixture of exponential distributions

Trevor Fenner, Mark Levenea, and George Loizou

Department of Computer Science and Information Systems, Birkbeck, University of London, London WC1E 7HX, UK

Received 30 November 2015 / Received in final form 14 January 2016
Published online 17 February 2016 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2016

Abstract. Recent interest in human dynamics has stimulated the investigation of the stochastic processes that explain human behaviour in various contexts, such as mobile phone networks and social media. In this paper, we extend the stochastic urn-based model proposed in [T. Fenner, M. Levene, G. Loizou, J. Stat. Mech. 2015, P08015 (2015)] so that it can generate mixture models, in particular, a mixture of exponential distributions. The model is designed to capture the dynamics of survival analysis, traditionally employed in clinical trials, reliability analysis in engineering, and more recently in the analysis of large data sets recording human dynamics. The mixture modelling approach, which is relatively simple and well understood, is very effective in capturing heterogeneity in data. We provide empirical evidence for the validity of the model, using a data set of popular search engine queries collected over a period of 114 months. We show that the survival function of these queries is closely matched by the exponential mixture solution for our model.

1 Introduction

Recent interest in complex systems, such as social networks, the world-wide-web, email networks and mobile phone networks [1], has led researchers to investigate the processes that could explain the dynamics of human behaviour within these networks. Barabási [2] has suggested that the bursty nature of human behaviour, for example, when measuring the inter-event response time of email communication, is a result of a decision-based queuing process. In particular, humans tend to prioritise actions, for example, when deciding which email to respond to, and therefore a priority queue model was proposed in reference [2], leading to a heavy-tailed power-law distribution of inter-event times. The availability of large data sets, such as mobile phone records, has widened the applicability of human dynamics investigation, for example, in an attempt by Schneider et al. [3] to uncover the characteristics of daily mobility patterns. Human dynamics is not limited to the study of behaviour within communication networks, as can be seen, for example, by a recent proposal of Mitnitski et al. [4], who apply a simple stochastic queueing model to the complex phenomenon of ageing, in order to illustrate how health deficits accumulate with age. Moreover, sociophysics [5–7] has similar aims and, in particular, uses concepts and methods from statistical physics to investigate social phenomena and political behaviour. This methodology often involves the empirical investigation of collective choice dynamics, resulting in the formation of statistical laws governing, for example, how the popularity of movies emerges from social behaviour.

In this context, we have recently modelled the survival of search engine queries in the top-10 list of the most popular queries that users have submitted, as an application of human dynamics [8]; see [9] for attempts to predict the future popularity of online content.

Survival analysis [10,11] provides statistical methods to estimate the time until an event occurs, known as the survival or failure time. Typically, an event in a survival model is referred to as a failure, since it often has negative connotations, such as mortality or the contraction of a disease, although it could, in principle, also be positive, such as the time to return to work or to recover from a disease. In the context of email communication mentioned above, an event might be a reply to an email; in the context of popular queries, an event might be dropping out of the top-10 most popular queries. Traditional applications of survival analysis are in clinical trials [12], and in understanding the mechanisms in biological ageing [13]. The methods used in survival analysis overlap with those used in engineering for reliability life data analysis [14,15]. Reliability analysis in engineering has many applications, for example, in manufacturing processes, and in software design and testing. However, one can envisage that survival analysis would find application in newer human dynamics scenarios in complex systems, such as those arising in social and communication networks [2,16,17].

We have been particularly interested in formulating generative models in the form of stochastic processes by which complex systems evolve and give rise to power laws or other distributions [8,18,19]. This type of research...
builds on the early work of Simon [20], and the more recent work of Barabási’s group [21] and other researchers [22]. In the context of human dynamics, the priority queue model mentioned above [2] is a generative model characterised by a heavy-tailed distribution.

In the bigger picture, one can view the goal of such research as being similar to that of social mechanisms [23], which looks into the processes, or mechanisms, that can explain observed social phenomena. Using an example given in [24], the growth in the sales of a book can be explained by the well-known logistic growth model [25], and more recently we have shown that the process of conference registration with an early bird deadline can be modelled by bi-logistic growth [26]. Such research can also be grounded within the field of sociophysics, mentioned above, where critical phenomena are important in demonstrating how the transition to global behaviour can emerge from the interactions of many individuals. The individuals may be neurons, as in reference [27], where criticality emerges as neuron cooperation, or people’s decisions, as in reference [28], where the popularity of movies emerges as collective choice behaviour.

In this work we build on the generative model proposed in reference [8], which captures the essential dynamics of survival analysis applications. In the urn-based model presented in reference [8], the actors are called balls, and a ball being present in urn $i$, the $i$th urn, indicates that the actor represented by the ball has so far survived for $i$ time steps. An actor could, for example, be a subject in a clinical trial, an email that has not yet been replied to, or an ongoing phone call. As a simplification, we assume that time is discrete and that, at any given time, a single ball may join the system with a fixed birth probability. In addition, existing balls in the system may be discarded, i.e. one or more mortality (or death) events may occur, according to some specified mortality function, which may take a general form. It is evident that, at any given time $t$, we may have at most one ball in urn $i$, for all $i \leq t$.

The probability that, after $t$ steps of the stochastic process outlined above, there is a surviving ball in urn $i$, is known as the survival function [10]. In the context of a medical study, this could represent the probability that a patient survives for longer than a specified time. In reference [8] we have shown how the survival function depends on the mortality function. Two important cases are: (1) when the mortality function is constant then the survival function follows an exponential distribution, and (2) when the mortality function follows a power-law distribution then the survival function follows a Weibull distribution [29]; the latter is often applied to the modelling of life data (also known as survival data, time to event data, or time to failure data) [10,15].

It is interesting to observe that the distribution of the survival function has two parameters, $\theta$ and $t$ (cf. Ref. [19]), whereas most previously studied generative stochastic models [21], including those in our previous work, for example [18], result in steady state distributions that are asymptotic in $t$ to a distribution with a single parameter $\theta$.

Exponential mixture models [30] have been exploited in survival analysis in a variety of applications, as mentioned in Section 2 below. Moreover, it has been shown that the survival function of an exponential mixture model has a decreasing failure rate [31,32], which was first observed in reference [33] in the context of air-conditioning systems for a fleet of airplanes. To generate exponential mixtures, in Section 3, we extend the urn-based stochastic process of [8] by adding a first step of tossing a biased coin to decide which of two processes becomes active; although the extension is defined for a mixture of two exponentials, it extends naturally to a mixture of several exponentials. We provide empirical evidence for the validity of the model by analysing the longitudinal data set of popular search engine queries over 114 months that we used in reference [8]. We show that the survival function of these queries closely follows the distribution generated by our model.

The rest of the paper is organised as follows. In Section 2, we review mixture models in the context of survival data analysis. In Section 3, we present our stochastic urn-based model that provides us with a mechanism to model the mixture of exponential distributions. In Section 4, we apply the model to the survival of popular search engine queries posted on Google Trends (www.google.com/trends/hottrends). Finally, in Section 5, we give our concluding remarks.

## 2 Mixture models for failure time analysis

Lifetime analysis, also referred to as survival time analysis or failure time analysis [11], has a long history of exploiting mixture models [30], in particular exponential mixture models, in a variety of applications. Examples of areas where exponential mixtures have been employed for lifetime analysis modelling are: the reliability of system components [33], network performance [34], actuarial losses [35], reliability engineering [32], and the likelihood of volcanic eruption [36].

The application we consider in Section 4, in the context of survival analysis, is that of predicting the duration of the popularity of search engine queries. We note that this is related to the problem of predicting the popularity of movies [28] or online content [9]. Mixture models are a natural tool for modelling such applications as they are very effective in capturing heterogeneity. In particular, when considering the popularity of queries, a mixture model can differentiate between two types of queries: ones that are very popular for a short period of time and others whose popularity lasts much longer; such an observation has already been made in the context of the popularity of movies [28].

The survival function $S(\theta)$ represents the probability that an actor survives for longer than a specified time $\theta$. Here we restrict ourselves to a mixture of two exponential survival functions, noting that this can readily be extended to a mixture of more than two such functions. The mixture rates of the distribution, $p_1$ and $p_2$, sum to one, and the mortality rates, $\lambda_1$ and $\lambda_2$, of the mixture model
(also known as the failure rates) are constants. These define a discrete hazard function [37], i.e. the conditional probability that a failure occurs at a specified time instant given that no failure occurred previously. We therefore define the mixture survival function \( S(\theta) \) as
\[
S(\theta) = p_1 \exp(-\lambda_1 \theta) + p_2 \exp(-\lambda_2 \theta),
\]
where \( p_1 + p_2 = 1 \).

The mortality rate of the mixture model [32] is now given by:
\[
\lambda(\theta) = \pi_1(\theta)\lambda_1 + \pi_2(\theta)\lambda_2,
\]
where the time-dependent probabilities \( \pi_k \) are defined as:
\[
\pi_k(\theta) = \frac{p_k \exp(-\lambda_k \theta)}{p_1 \exp(-\lambda_1 \theta) + p_2 \exp(-\lambda_2 \theta)}.
\]
In references [31,32] it was shown that the mortality rate \( \lambda(\theta) \) of the mixture model is decreasing in \( \theta \). Moreover, if we assume that \( \lambda_1 \leq \lambda_2 \), then \( \lambda(\theta)/\lambda_1 \) converges to 1 as \( \theta \) tends to infinity; this is referred to in the literature as a decreasing failure rate.

For comparison purposes, we will also make use of the Weibull survival function considered in reference [8], with shape parameter \( \rho \) and scale parameter \( \gamma \), which is given by:
\[
\exp\left(-\gamma \theta^{\rho+1}\right).
\]
It is important to note that, when the shape parameter satisfies \(-1 < \rho < 0\), then it can be expressed as a mixture of exponentials, since it is a completely monotone function [34,38]; see [39] for the underlying theory.

3 An evolutionary urn transfer model

We next formalise a stochastic urn model based on the survival model for human dynamics introduced in reference [8], in order to generate the mixture of two exponential distributions. This is done by first generating a mixture distribution by tossing a coin with mixture probabilities \( p_1 \) and \( p_2 \), where \( p_1 + p_2 = 1 \), to decide which process is active, and then employing a constant mortality rate parameter \( \lambda_k \) in order to generate an exponential distribution, as in reference [8]. We note that this mechanism extends naturally to a mixture of more than two exponentials by deploying a multi-sided “coin”.

We assume a countable number of urns, \( \text{urn}_0, \text{urn}_1, \ldots \), where a ball (or actor) in \( \text{urn}_i \) is said to be of age \( i \). Initially, all the urns are empty. At the start of the stochastic process, we toss the coin, as described above, to decide which of two processes will be active. From then on, at any time \( t, t \geq 0 \), a new ball may be born, with probability \( b \), and any such new ball is inserted into \( \text{urn}_0 \); in addition, an existing ball of age \( i \) can die by being discarded from \( \text{urn}_i \), for all \( i > 0 \).

For a given age \( i \) and time \( t \), we let the mortality rate \( \mu_k(i,t) \) be the probability that a ball in \( \text{urn}_i \) dies at time \( t \), where \( k = 1 \) or 2 depending on which process is active. We always require that \( \mu_k(0,t) = 0 \) for all \( t \), and that \( \mu_k(i,t) = \lambda_k \) for \( i > 0 \), where \( \lambda_k \) is a constant, satisfying \( 0 < \lambda_k < 1 \).

At time \( t \), the stochastic process then proceeds as follows in order to obtain the configuration at time \( t+1 \), where \( t \geq 0 \).

(i) For each \( i, 1 \leq i \leq t \), if \( \text{urn}_i \) is non-empty then, with probability \( \mu_k(i,t) = \lambda_k \), a mortality event occurs, i.e. the ball in \( \text{urn}_i \) is discarded.

(ii) Next, the ages of all balls remaining in the system are incremented by 1, i.e. a ball in \( \text{urn}_i \) is moved to \( \text{urn}_{i+1} \), for each \( i \).

(iii) Finally, with probability \( b \), where \( 0 < b < 1 \), a birth event occurs, i.e. a ball is inserted into \( \text{urn}_0 \).

We observe that, at any time \( t \), \( \text{urn}_i \) is empty for \( i > t \). Note that we are assuming that the mortality rate is constant, i.e. \( \mu_k(i,t) = \lambda_k \), which leads to the survival function of an exponential distribution, as was shown in reference [8]. However, in principle, the mortality rate could be an arbitrary function, for example, power-law mortality leads to the survival function of a Weibull distribution. Note also that we are using a single birth probability \( b \), rather than one for each process, since, as can be seen from (7) below, \( b \) is merely a scaling factor. It follows that having different values for \( b \) for the two processes is equivalent to adjusting the mixture probabilities.

We now let \( F_k(i,t) \geq 0 \), with \( k = 1 \) or 2, be a discrete function denoting the probability that there is a ball in \( \text{urn}_i \), at time \( t \), given that process \( k \) is active. Initially, we set \( F_k(0,0) = b \) and \( F_k(i,0) = 0 \) for all \( i > 0 \).

The dynamics of the model are captured by the following two equations, where we assume that \( k = 1 \) or 2 according to the outcome of the coin toss:
\[
F_k(0,t) = b \quad \text{for} \quad t \geq 0,
\]
and
\[
F_k(i+1,t+1) = F_k(i,t) - \mu_k(i,t)F_k(i,t) \quad \text{for} \quad 0 \leq i \leq t.
\]
We can expand (6) to obtain
\[
F_k(i+1,t+1) = b \prod_{j=1}^{i} (1 - \mu_k(j,i+j)) = b(1 - \lambda_k)^i.
\]
Using the approximation \( \ln(1-x) \approx -x \), which holds for small \( x \), we see that the distribution is approximately exponential:
\[
F_k(i+1,t+1) \approx b \exp(-\lambda_k i)\] (8)
We note that, in equation (8), we could use the more accurate second order approximation \( \ln(1-x) \approx -x - x^2/2 \).

We now define \( F(i,t) \) accordingly to \( F_k(i,t) \) to take into account which of the two processes is active, i.e.
\[
F(i,t) = p_1 F_1(i,t) + p_2 F_2(i,t).
\]
This yields the initial conditions \( F(0,0) = b, F(i,0) = 0 \) for all \( i > 0 \), and also \( F(0,t) = b \) for \( t \geq 0 \). Similarly,
from (7) and (8) for \(0 \leq i \leq t\), we obtain the exponential mixture:

\[
F(i + 1, t + 1) = p_1 b (1 - \lambda_1)^i + p_2 b (1 - \lambda_2)^i \\
\approx p_1 b \exp(-\lambda_1 i) + p_2 b \exp(-\lambda_2 i).
\]  

We note that, as mentioned in the beginning of this section, the model can be extended to more than two components; in this case the sums in equation (9) would have the appropriate number of terms.

### 4 Application of the mixture model

In the context of human dynamics, survival analysis has recently been applied to large data sets. These include the analysis of phone call durations [40], the investigation of how long Wikipedia editors remain active [41], and for predicting the likelihood of online content to become popular [42].

The survival function \(S(\theta)\) introduced in Section 2 is usually estimated via a step function by computing the probability that an actor survives until time \(\theta\), for \(\theta = 1, 2, \ldots, t\); this step function is known as the Kaplan-Meier estimator [10, 43]. By comparing (7) with the Kaplan-Meier estimators for the survival functions (43, Eq. (2b)) of the components of the mixture model, these are seen to be analogous to \(F_k(i, t)\) for \(k = 1, 2\), and an actor that was born at time \(t - i\). We thus obtain

\[
S(i) \approx \frac{p_1 F_1(i, t) + p_2 F_2(i, t)}{b} = \frac{F(i, t)}{b}.
\]

In theory, the survival function \(S(\theta)\) does not depend on the length \(t\) of the trial. However, in practice, the Kaplan-Meier estimate will be more accurate for longer trials. Nevertheless, this estimate is more accurate when most of the actors are still present in the study, since the estimate may be inaccurate when there are only a few actors remaining [44].

As a proof of concept for the exponential mixture model, we analysed the survival of queries in the top-10 Google Trends “hot searches” (http://www.google.com/trends/hottrends), that we used in reference [8]. Data was collected monthly for the top-10 “hot searches” over 114 months, from January 2004 until June 2013, for six categories (together with their subcategories in each case): Business and Politics (or simply Business), Entertainment, Nature and Science (or simply Science), Shopping & Fashion (or simply Shopping), Sports, and Travel and Leisure (or simply Travel). The number of distinct queries per category over the period is shown in Table 1. It is apparent from this statistic that the top-10 queries from Shopping change the least, while those from Entertainment change the most.

In this data set, the balls are top-10 queries, a mortality event occurs when a query leaves the top-10 in a given month, and a birth event occurs when a new query joins the top-10 in a given month (note that time is discrete and is measured in months). As we are considering an exponential mixture model with constant mortality for each component, queries leave the top-10 according to equation (2).

We next outline the methodology we have used to validate and evaluate the stochastic urn model presented in Section 3. We then give further details, before discussing and analysing the results.

(i) First, to obtain estimates of \(p_1\), \(\lambda_1\) and \(\lambda_2\), we performed nonlinear regression to the right-hand side of equation (9) using the Kaplan-Meier estimates computed from the raw data, for \(i = 1, 2, \ldots, 114\).

(ii) We used the estimates of \(p_1\), \(\lambda_1\) and \(\lambda_2\) from (i) to compute, for each \(i\), the right-hand side of equation (9); we call this the mixture data. We then repeated the nonlinear regression using the mixture data as a quick “sanity check” that the new values of \(p_1\), \(\lambda_1\) and \(\lambda_2\) obtained were consistent with those from (i).

(iii) Next we ran simulations using the parameters \(p_1\), \(\lambda_1\) and \(\lambda_2\) from (i), and \(b = 0.9\). Using the averaged values of \(F(i, t)\) from the simulations, we again repeated the nonlinear regression to obtain new values for \(p_1\), \(\lambda_1\) and \(\lambda_2\); these were compared with those from (i) for consistency.

(iv) We computed the \(D\) values for Kolmogorov-Smirnov tests to ascertain whether the Kaplan-Meier estimates, the mixture data and the averaged simulation data are likely to have all come from the same distribution.

We first obtained the Kaplan-Meier estimates from the raw data sets for the six individual categories and for their aggregation (all categories). Recalling that the survival function is approximated as in equation (10), following (i), we used Matlab to obtain estimates for \(p_1\), \(\lambda_1\) and \(\lambda_2\) by nonlinear least-squares regression of \(S(i)\) on \(i\) for fitting the right-hand side of equation (9), for \(i = 1, 2, \ldots, 114\). The estimated parameters \(p_1\), \(\lambda_1\) and \(\lambda_2\) are shown in the rows of Table 2, together with the coefficient of determination \(R^2\) [45]; the \(R^2\) values show a very good fit for all of the categories. Nonlinear regression using the mixture data values computed from the right-hand side of equation (9), as in (ii), yielded almost perfect fits, as expected.

To test the validity of the model, as in (iii), we then carried out simulations in Matlab of the stochastic urn transfer model using the values of \(p_1\), \(\lambda_1\) and \(\lambda_2\) from Table 2. We chose the value \(b = 0.9\) for all simulations, after running some sample simulations with other values of \(b\). The value of \(b\) is not critical since, as can be seen from

| Data set     | Number of queries |
|--------------|-------------------|
| Business     | 318               |
| Entertainment| 1672              |
| Science      | 150               |
| Shopping     | 107               |
| Sports       | 774               |
| Travel       | 342               |
| All categories | 3363             |

Table 1. Number of top-10 queries collected from Google Trends.
Table 2. Nonlinear least-squares regression to a mixture of exponentials using the Kaplan-Meier estimates.

| Data set        | $p_1$   | $\lambda_1$ | $\lambda_2$ | $R^2$  |
|-----------------|---------|--------------|--------------|--------|
| Business        | 0.4061  | 0.0136       | 0.4095       | 0.9893 |
| Entertainment   | 0.2013  | 0.0405       | 0.4793       | 0.9965 |
| Science         | 0.6693  | 0.0071       | 0.4391       | 0.9884 |
| Shopping        | 0.7403  | 0.0091       | 0.4880       | 0.9915 |
| Sports          | 0.2275  | 0.0215       | 0.4599       | 0.9829 |
| Travel          | 0.4483  | 0.0090       | 0.4364       | 0.9823 |
| All categories  | 0.2600  | 0.0154       | 0.4337       | 0.9894 |

Table 3. Nonlinear least-squares regression to a mixture of exponentials using the simulated data with $p_1$, $\lambda_1$ and $\lambda_2$ taken from Table 2.

| Data set        | $p_1$   | $\lambda_1$ | $\lambda_2$ | $R^2$  |
|-----------------|---------|--------------|--------------|--------|
| Business        | 0.4052  | 0.0135       | 0.4037       | 0.9999 |
| Entertainment   | 0.1991  | 0.0402       | 0.4683       | 0.9996 |
| Science         | 0.6686  | 0.0071       | 0.4289       | 0.9994 |
| Shopping        | 0.7394  | 0.0091       | 0.4760       | 0.9996 |
| Sports          | 0.2269  | 0.0214       | 0.4504       | 0.9997 |
| Travel          | 0.4472  | 0.0089       | 0.4286       | 0.9997 |
| All categories  | 0.2598  | 0.0153       | 0.4260       | 0.9998 |

equation (9), $b$ is merely a scaling factor. The simulations were run for 114 steps, one for each month, for each of the categories, and this was repeated $10^6$ times. For each category, we then calculated the average value of $F(i,t)$ for $i = 1, 2, \ldots, t$, over the $10^6$ runs. Repeating the nonlinear regression using these average values gave the results shown in Table 3. Comparing Table 3 with Table 2 shows that all the values of $p_1$, $\lambda_1$ and $\lambda_2$ are very close. Moreover, the very high values for $R^2$ confirm the validity of equation (9) as the solution to our model.

To compare the Kaplan-Meier estimates, the mixture data and the averaged simulation data, as in (IV), we performed three Kolmogorov-Smirnov 2-sample 2-tailed tests, as described in Section 6.6.4 of reference [46]. Taking the null hypothesis to be that the Kaplan-Meier estimates, the mixture data and the averaged simulation data all come from the same population distribution, the critical value at significance level $\alpha = 0.05$ is 0.1801 for a sample of 114 (number of months). The $D$ values for the three pairwise tests are shown in Table 4. It can be seen that, in all cases, the values of the test statistic $D$ are well below the critical value. Hence, we cannot reject the null hypothesis at significance level $\alpha = 0.05$. The values in the Sim-Mix column show that the distributions of the mixture data and averaged simulation data are extremely close, which is unsurprising since these are both derived directly from our model. We note that, even at significance level $\alpha = 0.10$, where the critical value is 0.1616, the null hypothesis cannot be rejected.

In reference [8], we used nonlinear least-squares regression to fit Weibull survival functions to the Kaplan-Meier estimates. The estimates obtained for the parameters $\gamma$ and $\rho$ are shown in Table 5, taken from reference [8].

We note that the shape parameters $\rho$ satisfy $-1 < \rho < 0$, which, as mentioned at the end of Section 2, implies that the Weibull survival functions can be expressed as mixtures of exponentials. In order to compare our approximation by exponential mixtures with the Weibull survival functions obtained in reference [8], we used least-squares curve fitting to fit Weibulls (4) to the exponential mixtures obtained from equations (9) and (10), using the parameters from Table 2, in a similar manner to [47]. The estimated parameters $\gamma$ and $\rho$ are shown in Table 6, together with $R^2$. It can be seen that these estimates are very close to those in Table 5.

Finally, we note from Table 2 that $\lambda_1$ is significantly smaller than $\lambda_2$. It therefore follows, from the comment made after (3) regarding the decreasing failure rate, that a substantial number of queries are popular for a short duration, because $\lambda_2$ is large, while the popularity of others lasts for much longer, because $\lambda_1$ is small. We note that a similar observation, in the context of the popularity of movies, was made in reference [28]. In particular, for the Google Trends data set, about 40% of the queries survive for only a single month.
5 Concluding remarks

In this paper, we propose a stochastic evolutionary urn-based model that can be applied to survival analysis scenarios in the context of human dynamics. In our model, actors survive in the system until they die and leave the system according to a constant mortality rate. A solution to the equations describing the model was obtained in (9), and it was seen that this solution is approximately an exponential mixture. We then successfully applied our model to data on the survival of popular search engine queries. It is also potentially applicable to other data sets relating to human behaviour.

One important problem to be addressed is how to choose the number of components in the mixture model. As discussed in reference [30], Chapter 6, several methods have been suggested to obtain a parsimonious model. Among these, a number of prominent methods are based on maximising a penalised log-likelihood function.

Generative models, like the one presented here, have the potential to explain observed social phenomena and, more specifically, social mechanisms and the emergence of collective behaviour, as discussed in the introduction. They allow us to gain insight into the underlying processes; in addition, they may be useful for more accurate prediction of, for example, online content, as discussed in reference [9]. With regards to our theme of human dynamics [7], our stochastic mixture urn model contributes to a better understanding of collective phenomena, such as popularity, and how such global behaviour emerges from the decisions of individuals from whom the statistics have been collected. Another possible avenue for investigation is to consider other distribution mixtures, such as finite Weibull mixture models (see Ref. [48], where it is argued that the reliability of a system can be suitably modelled by a Weibull mixture model).

We would like to thank Sunee Kingrani, who collected the Google Trends data and computed the Kaplan-Meier estimates for the data set.

References

1. A.-L. Barabási, IEEE Control Syst. Mag. 27, 33 (2007)
2. A.-L. Barabási, Nature 435, 207 (2005)
3. C.M. Schneider, V. Belik, T. Couronné, M.C. González, J. R. Soc. Interface 10, 20130246 (2013)
4. A. Mitnitski, X. Song, K. Rockwood, Biogerontology 14, 709 (2013)
5. S. Galam, J. Mod. Phys. C 19, 409 (2008)
6. P. Sen, B.K. Chakrabarti, Sociophysics: An Introduction (Oxford University Press, Oxford, 2014)
7. C. Castellano, S. Fortunato, V. Loreto, Rev. Mod. Phys. 81, 591 (2009)
8. T. Fenner, M. Levene, G. Loizou, J. Stat. Mech. 2015, P08015 (2015)
9. A. Tatar, M.D. de Amorim, S. Fdida, P. Antoniadis, J. Internet Services Appl. 5, 20 (2014)
10. D.G. Kleinbaum, M. Klein, Survival Analysis, A Self-Learning Text, 3rd edn. Springer Science+Business Media (Springer-Verlag, New York, 2012)
11. J.F. Lawless, Statistical Models and Methods for Lifetime Data, 2nd edn. Wiley Series in Probability and Statistics (John Wiley & Sons, 2003)
12. T.R. Fleming, D.Y. Lin, Biometrics 56, 971 (2000)
13. L.A. Gavrilov, N.S. Gavrilova, J. Theor. Biol. 213, 527 (2001)
14. M. Finkelstein, Failure Rate Modelling for Reliability and Risk, Springer Series in Reliability Engineering (Springer-Verlag, London, 2008)
15. P.D.T. O’Connor, A. Kleyner, Practical Reliability Engineering, Wiley Series in Telecommunications, 5th edn. (John Wiley & Sons, Chichester, 2012)
16. J. Mathiesen, L. Angheluta, P.T.H. Ahlgren, M.H. Jensen, in Proceedings of the National Academy of Sciences of the United States of America, 2013, Vol. 110, pp. 17259–17262
17. J. Candia, M.C. González, P. Wang, T. Schoenhar, G. Madey, A.-L. Barabási, J. Phys. A 41, 224015 (2008)
18. T. Fenner, M. Levene, G. Loizou, Social Networks 29, 70 (2007)
19. T. Fenner, M. Levene, G. Loizou, IEEE Trans. Comput. Intell. AI Games 4, 84 (2012)
20. H.A. Simon, Biometrika 42, 425 (1955)
21. R. Albert, A.-L. Barabási, Rev. Mod. Phys. 74, 47 (2002)
22. S. Börner, S. Sanyal, A. Vespignani, Ann. Rev. Inform. Sci. Technol. 41, 537 (2007)
23. P. Hedström, R. Swedberg, Social mechanisms: An introductory essay, in Social Mechanisms: An Analytical Approach to Social Theory, edited by P. Hedström, R. Swedberg (Cambridge University Press, Cambridge, 1998), pp. 1–31
24. T.C. Schelling, Social mechanisms and social dynamics, in Social Mechanisms: An Analytical Approach to Social Theory, edited by P. Hedström and R. Swedberg (Cambridge University Press, Cambridge, 1998), pp. 32–44
25. A. Tsoularis, J. Wallace, Math. Biosci. 179, 21 (2002)
26. T. Fenner, M. Levene, G. Loizou, Central Eur. J. Phys. 11, 904 (2013)
27. E. Lovecchio, P. Allegrini, E. Geneston, B.J. West, Phys. Rev. Lett. 106, 170602 (2011)
28. R.K. Pan, S. Sinha, New J. Phys. 12, 115004 (2010)
29. H. Rinne, The Weibull Distribution: A Handbook (CRC Press, Boca Raton, 2009)
30. G. McLachlan, D. Peel, Finite Mixture Models, 2nd edn. (John Wiley & Sons, New York, 2000)
31. H. Block, H. Joe, Lifetime Data Analysis 3, 269 (1997)
32. M. Finkelstein, Appl. Stoch. Models Bus. Ind. 25, 643 (2009)
33. F. Proschan, Technometrics 5, 375 (1963)
34. A. Feldman, W. Whitt, Performance Evaluation 31, 245 (1998)
35. C. Keatinge, in Proceedings of the Casualty Actuarial Society, 1999, Vol. LXXXVI, pp. 654–698
36. A.T. Mendoza-Rosas, S. de la Cruz-Reyna, Nat. Hazards Earth Syst. Sci. 9, 425 (2009)
37. C. Bracquemond, O. Gaudoin, Int. J. Reliab. Qual. Safety Eng. 10, 69 (2003)
38. N.P. Jewell, Ann. Statist. 10, 479 (1982)
39. W. Feller, in *An Introduction to Probability Theory and its Applications*, 2nd edn. (John Wiley & Sons, New York, 1971), Vol. II

40. P.O.S. Vaz De Melo, L. Akoglu, C. Faloutsos, A.A.P. Loureiro, in *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECML PKDD), Barcelona, 2010, pp. 354–369.

41. D. Zhang, K. Prior, M. Levene, in *Proceedings of the 8th International Symposium on Wikis and Open Collaboration (WikiSym)*, Linz, Austria, 2012

42. J.G. Lee, S. Moon, K. Salamatian, *Neurocomputing* **76**, 134 (2012)

43. E.L. Kaplan, P. Meier, *J. Am. Stat. Assoc.* **53**, 457 (1958)

44. J.T. Rich, J.G. Neely, R.C. Paniello, C.C.J. Voeller, B. Nussenbaum, E.W. Wang, *Otolaryngology-Head and Neck Surgery* **143**, 331 (2010)

45. H. Motulsky, *Intuitive Biostatistics* (Oxford University Press, Oxford, 1995)

46. S. Siegel, N.J. Castellan Jr., *Nonparametric Statistics for the Behavioral Sciences*, 2nd edn. (McGraw-Hill, New York, 1988)

47. T. Jin, L.K. Gonigunta, *J. Stat. Comput. Simul.* **80**, 273 (2010)

48. T. Bucar, M. Nagode, M. Fajdiga, *Reliab. Eng. Syst. Safety* **84**, 241 (2004)