Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A stochastic evolutionary model for survival dynamics

Trevor Fenner, Mark Levene *, George Loizou

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

HIGHLIGHTS

- A generative model that captures the essential dynamics of survival analysis.
- Solution of the mean-field equations for the model as a type of power-law.
- Validation of the model using two data sets from the survival analysis literature.

ARTICLE INFO

Article history:
Received 23 January 2014
Received in revised form 12 March 2014
Available online xxxx

ABSTRACT

The recent interest in human dynamics has led researchers to investigate the stochastic processes that explain human behaviour in different contexts. Here we propose a generative model to capture the essential dynamics of survival analysis, traditionally employed in clinical trials and reliability analysis in engineering. In our model, the only implicit assumption made is that the longer an actor has been in the system, the more likely it is to have failed. We derive a power-law distribution for the process and provide preliminary empirical evidence for the validity of the model from two well-known survival analysis data sets.

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 may explain the dynamics of human behaviour within these networks. For example, Barabási [2] 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 Ref. [2], leading to a heavy-tailed power-law distribution of inter-event times.

The tail of a power-law distribution decays polynomially in contrast to the exponential decay characteristic of the Gaussian distribution, which is why it is also referred to as a heavy-tailed distribution. A *power-law* distribution takes the general form

$$g(i) = \frac{C}{i^{\rho}},$$

representing the proportion of observations having the value i, where C and ρ are positive constants; we call ρ the *exponent* of the distribution [3].

Survival analysis [4] provides statistical methods to estimate the time until an event will occur, known as the *survival time*. Typically, an event in a survival model is referred to as a *failure*, as it often has negative connotations, such as death or

* Corresponding author. Tel.: +44 2076316711.
E-mail addresses: trevor@dcs.bbk.ac.uk (T. Fenner), mlevene@dcs.bbk.ac.uk, mark@dcs.bbk.ac.uk (M. Levene), george@dcs.bbk.ac.uk (G. Loizou).

http://dx.doi.org/10.1016/j.physa.2014.05.038
0378-4371/© 2014 Published by Elsevier B.V.

the contraction of a disease, although it could 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. Traditional applications of survival analysis are in clinical trials [5], and reliability engineering [6], the analogue of survival analysis for mechanical systems. 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 [2,7,8].

Of particular interest to us has been the formulation of a *generative model* in the form of a stochastic process by which a complex system evolves and gives rise to a power law or other distribution [9–11]. This type of research builds on the early work of Simon [12], and the more recent work of Barabási's group [13] and other researchers [14]. In the context of human dynamics, the priority queue model [2] mentioned above 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* [15], which looks into the processes, or mechanisms, that can explain observed social phenomena. Using an example given in Ref. [16], the growth in the sales of a book can be explained by the well-known logistic growth model [17].

The motivation of this paper is in the formulation of a simple generative model that will capture the essential dynamics of survival analysis applications. For this purpose, we make use of an urn-based stochastic model, where the actors are called *balls*, and a ball being present in urn_i, the ith 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 or an email that has not yet been replied to. As a simplification, we assume that time is discrete and that, at any given time, one ball may join the system with a fixed probability. As a result, at any given time, say t, we may have at most one ball in urn_i, for all $i \le t$. Alternatively, with a fixed probability, an existing ball in the system may be chosen uniformly at random and discarded. We note that at any time t, if $i < j$, the probability that urn_i is empty is less than the probability that urn_j is empty, as a ball in urn_j could have been discarded at any of the previous j time steps, whereas a ball in urn_i could have only been discarded at any of the previous i time steps, and a ball to be discarded is chosen uniformly.

This mechanism can be contrasted with the preferential attachment rule in evolving networks [13] (also known as the "rich get richer" phenomenon [1]), where the probability of adding a new link to a node (or actor) is proportional to the number of existing links that the node already has. In our case, actors are chosen uniformly rather than preferentially; however, the model keeps a record of the time for which an actor has survived so for.

Our main result is to derive a power-law distribution for the probability that, after t steps of the stochastic process outlined above, there is a surviving ball in urn_i, where $i \le t$. Thus, in our model, the *survivor function* [4], which gives the probability that a patient (in our model a ball) survives for more than a given time, can be approximated by a power-law distribution. It is interesting to observe that the resulting distribution has two parameters, i and t, as in Ref. [11], whereas most previously studied generative stochastic models [13,3], including those in our previous work [9,10], result in steady state distributions that are asymptotic in t to a distribution with a single parameter i. As a proof of concept, we demonstrate the validity of our model by analysing two well-known data sets from the survival analysis literature [4].

The rest of the paper is organised as follows. In Section 2, we present our stochastic urn-based model that provides us with a mechanism to model the essential dynamics of survival models, and we derive the resulting power-law distribution. In Section 3, we apply our generative model to two well-known data sets from survival analysis, and finally, in Section 4, we give our concluding remarks.

2. An evolutionary urn transfer model

In this section we formalise our stochastic urn model for modelling the dynamic aspects of a time-varying system and present an approximate solution to the mean field equations describing the model.

We assume a countable number of urns, $urn_1, urn_2, \ldots$. Initially all the urns are empty except urn_1, which has one ball in it. Let $F_i(t)$ be 1 or 0, respectively, according to whether or not there is a ball in urn_i at time t of the stochastic process. Initially we set $F_1(1) = 1$, and for all other urns $F_i(1) = 0$. The *age* of a ball in urn_i is defined to be i. Then, at time $t + 1$ of the stochastic process, where $t \ge 1$, one of two things may occur:

(i) with probability p, where $0 < p < 1$, a new ball is put into urn_1 (i.e. its initial age is 1), or
(ii) with probability $q(1 - p)$, where $0 < q < 1$, an urn is selected, with urn_i being selected with probability proportional to $F_i(t)$, i.e. one of the non-empty urns is selected uniformly at random, and the ball in the selected urn is discarded.

Next, the age of all balls remaining in the system, apart from a new ball that may have just been put into urn_1 during this time step, is incremented by 1, i.e. any ball in urn_i is moved to urn_{i+1} for each i.

We observe that, in this model, $F_i(t)$ is equal to either 1 or 0, since at most one new ball is generated at time t and the age of all other balls increases by one at time t. Moreover, it can be seen that, at any given time $t > 1$, the probability of there being a ball in urn_1 is p.

We constrain the system so that on average more balls are added to the system than are discarded, i.e.

$$p > q(1 - p),$$

otherwise the system would almost surely degenerate into a state of emptiness, i.e. having no balls in the system. This constraint is an instance of the gambler's ruin problem [18], from which it follows that the probability that the urn transfer process will *not* terminate with all the urns being empty is strictly positive [9].