A detailed analysis of Twitter-based information cascades is performed, and it is demonstrated that branching process hypotheses are approximately satisfied. Using a branching process framework, models of agent-to-agent transmission are compared to conclude that a limited attention model better reproduces the relevant characteristics of the data than the more common independent cascade model. Existing and new analytical results for branching processes are shown to match well to the important statistical characteristics of the empirical information cascades, thus demonstrating the power of branching process descriptions for understanding social information spreading.

Keywords: branching processes; networks; cascades.

1. Introduction

The transmission of information via online social networks is increasingly ubiquitous. The volume of freely available data offers unprecedented opportunities for data-driven mathematical modelling of human behaviour. Twitter, for example, is a directed social network wherein users ‘follow’ other users in order to receive their broadcast transmissions, called ‘tweets’. All tweets are public, making analysis of Twitter data particularly popular among data scientists. Twitter users may retweet messages they receive from the users they follow, and in this way cascades of information may stem from a single tweet event that we call the ‘seed’. In the search for mathematical models to describe such structures, branching processes [1, 2] are an appealing option. As stochastic processes, they can potentially capture the wide variability
observed in tweeting patterns and human behaviour, while offering a wealth of theoretical results that can be tested against data from online social networks.

Branching processes have already been applied in several studies of Twitter and other online fora. The recent review by Aragón et al. [3] surveys models of discussion threads, including Twitter reply cascades. Several of the generative models cited in Ref. [3] are based on branching processes, but most find it necessary to modify classical branching processes with some novel features in order to match to data. For example, Nishi et al. [4] studied reply cascades in Twitter (as distinct from the retweet cascades that we examine here) that were seeded by celebrities and found that these could not be fitted by classical Galton–Watson processes, so they introduced a modified version of the branching process. On the other hand, Galton–Watson processes (albeit with special seed offspring distributions) were successfully applied to discussion trees from Reddit by Medvedev et al. [5], and time-dependent continuous-time branching processes were fitted to a viral marketing campaign in Refs [6, 7]. Golub and Jackson [8] reanalysed data from Ref. [9] to show that although standard branching processes did not appear to reproduce the features of the email cascades studied in Ref. [9], when selection bias was added—to model the fact that large, viral and chains are more likely to be observed than small chains—then the biased Galton–Watson process fitted quite well.

Although branching processes have been fitted to data to form the basis for simulation and prediction in several studies, the application of analytical results from branching processes theory has been mostly limited to a small selection of features. The most common [3] is the cascade size, that is, the accumulated number of tweets or replies to a single seeding post, for which the well-known Galton–Watson result for the expected total number of progeny has been used for prediction, for example [10]. However, determining the entire distribution of cascade sizes, not just its mean, is quite feasible [11, 12], as are analytical (or semi-analytical) methods for the calculating the length and depth of cascade trees, as well as other measures. One such measure that we examine in Section 4.3 is the ‘structural virality’ of a cascade, as introduced by Goel et al. [13]. Using this and other measures of cascade trees, Goel et al. performed large-scale numerical simulations of a simple transmission model on networks to fit to data from Twitter. The transmission model of [13] is a discrete-time version of the susceptible-infected-recovered disease-spread model, also known as the ‘independent cascade model’ (ICM) [14]. Other network-based simulation models [15, 16] use variations of such dynamics (such as susceptible-infected-susceptible disease-spread models [17]) to understand the effects of network structure upon spreading.

In this article, we focus on three aspects of branching process models for Twitter retweet cascades, using a reanalysis of two previously studied data sets [18, 19]. First, in Section 2, we extract the empirical offspring distributions from the tree structures and show that these remain approximately stable across a range of generations. This is a necessary condition for classical branching processes to provide accurate models for detailed features of cascade trees, and the simplicity of this result contrasts with models where explicit time-decay of novelty [20–22] or generation-dependent branching numbers [23] are required.

Secondly, we consider in Section 3, how the structure of the underlying social network and the modelling of user-to-user transmission mechanisms can affect the offspring distribution for cascade trees. By comparing with the empirical results of Section 2, we examine whether the offspring distribution is better modelled by the independent cascade model, or by an alternative model that accounts for limited attention of users of social media [24].

Finally, in Section 4, we use the branching process framework to derive predictions for features of cascades, focussing on both the distribution of metrics of interest across the entire data set and on analytical
results for expected values. For completeness, we first derive (in Sections 4.1 and 4.2) some well-known [1, 2, 25] results for cascade sizes and durations. We then build on this approach in Section 4.3 to derive new results for the expected value of tree depths and the structural virality measure of Goel et al. [13], and apparent novelty decay factors [20, 21]. These results include integral expressions for expected structural virality and tree depth that we have not been able to locate in existing literature, and which are amenable to asymptotic analysis. We conclude the article with a discussion of the results, limitations and potential extensions in Section 5.

2. Data

2.1 Data sources

As a motivation and test for branching process hypotheses, we reanalyse two independent Twitter data sets. Both data sets have been previously analysed and described in Refs [18] and [19, 26], but here, we use the identified cascade structures to focus on the accuracy of branching process models.

The first data set, which we call ‘Marref’, is comprised of tweets related to the 2015 Irish same-sex marriage referendum, collected between 8 May and 23 May 2015. As described in Ref. [18], all tweets containing either of the hashtags #marref or #marriageref were gathered and analysed. Our focus in this article is on the tree structures of retweeting behaviour, and the extent to which this can be accurately modelled by branching processes. A ‘particle’ or ‘node’ in a tree (see Fig. 1 e.g.) represents a retweet event, which may cause further retweets, that is, a next generation of particles in the tree, called the ‘children’ particles of the ‘parent’ node. The children of a given node event are identified using the tree-reconstruction methodology described in Goel et al. [13], as implemented in Ref. [27]. This algorithm aims to identify a single parent node (tweet) for each retweet event, by using the text of the tweet and (in cases of multiple possible parents) the timestamps of the tweeting events. The output of the Goel et al. algorithm is a tree structure corresponding to each cascade of retweets, as described in Section 2.2 below. The data collection procedure is restricted to cascades of size greater than one, so that seeds that generate no children are not recorded.

The second data set, called ‘URL’, is comprised of tweets containing Uniform Resource Locators (URLs) for internet addresses that were posted on Twitter during October 2010 [28]. The collection and processing of the data are described in detail in Refs [26] and [19]. The URLs chosen for tracking were discovered by sampling tweets gathered using Twitter’s Gardenhose API [19] and then searching for all appearances of these URLs. As for the Marref data set, trees were reconstructed using the algorithm of Goel et al. [13]. We note that there is some selection bias in this data as URLs with larger cascade sizes were more likely to be selected by the initial sampling than relatively unpopular URLs. Also, as in the Marref data, the URLs data set only contains those cascades whose size (total number of tweets) is greater than one.

In order to ensure that each of the studied trees has had sufficient time to develop all its generations within the finite-length time window of data collection, we select only those trees where the ‘birth’ of the tree (i.e. the timestamp of the seed node) occurs during the first half of the time window for data collection. In the case of the Marref data, this avoids trees that are born during the increased activity around the referendum day and its aftermath. Although the original tweets occur in continuous time, we confine our attention in this article to the discrete generations of the tree structures. This choice simplifies the analysis to the discrete-time case of Galton–Watson processes, but it means that we do not investigate important aspects, such as the time between tweet arrival and retweet, which would require an extension.
of our approach to continuous-time branching processes [6, 7]. The extracted tree structures for both data sets are made available in anonymized form [29].

2.2 Tree structures

In this section, we analyse the characteristics of the trees extracted from the data as described in Section 2.1. For each data set, we consider an ensemble of $M$ trees, with each tree made up of particles (or nodes) in multiple generations, see Fig. 1. We define $Z_{m,n}$ to be the number of particles in generation $n$ of tree $m$ (where $m = 1, 2, \ldots, M$ and $n = 0, 1, \ldots$ ). The individual trees have very heterogeneous characteristics (size, number of generations, etc.), so we first consider the ensemble as a whole.

Defining $z_n$ as the total number of generation-$n$ particles observed across all trees, that is,

$$z_n = \sum_{m=1}^{M} Z_{m,n},$$

we plot in the top panels of Fig. 2 the dependence of $z_n$ on the generation $n$ using log-linear scales. Figure 2(a) is from the Marref data set ($M = 7,736$) and Fig. 2(b) is from the URL data set ($M = 39,547$). An approximately exponential dependence of $z_n$ upon $n$ is shown by the nearly linear shape of the function on log-linear axes; such a dependence is consistent with a subcritical branching process. Note that small-number fluctuations occur when $z_n$ is relatively small: we choose $10^3$ as a threshold level (shown by the black dashed line in Fig. 2(a and b)) and focus on $z_n$ values which are above this threshold.

In Fig. 2(c and d) (for Marref and URL, respectively), we show the effective branching number [23]

$$\xi_n = \frac{z_{n+1}}{z_n},$$

Fig. 1. Schematic of the ensemble of trees, indicating the $Z_{m,n}$ values for the first two trees in the ensemble.
which gives the average number of children particles for a particle of the $n$th generation. Observe that $\xi_n$ is approximately constant for the range of generations in which $z_n$ is sufficiently large (i.e. above the threshold marked in Fig. 2(a and b)). Figure 2(d) shows that the early generations of the URL data set exhibit a lot of fluctuations in the $\xi_n$ values, consistent with the possible biasing of the data towards larger trees (see Section 2.1). In both cases, the branching number $\xi_0$ of the seed generation appears to be anomalously high; this is partly due to the biasing introduced by the fact that no trees of size less than two are recorded (but see also the discussion leading to Eq. (10) below). The dashed green lines in Fig. 2 highlight the range of generations over which the branching number $\xi_n$ appears to be approximately constant. For each data set, we calculate an average value $\bar{\xi}$ of the branching number over the range shown by the dashed green line. The URL data set, with a value $\bar{\xi} = 0.90$, has a high virality (recall the critical branching number of 1 separates the regime of subcritical cascades from that of supercritical cascades), while the Marref data set has lower virality ($\bar{\xi} = 0.46$).

Next, we make a stronger test of the branching process hypothesis, by examining the empirical offspring distribution at each generation. For each particle, $i$ in generation $n$ we record the number $Z_{n+1}^{(i)}$ of its offspring particles, that is, the number of users in generation $n + 1$ that are identified as children of particle $i$. Gathering the ensemble of $Z_{n+1}^{(i)}$ values across all trees, we calculate the empirical offspring
distribution of generation \( n \) as

\[
\tilde{q}_{\ell,n} = \text{Prob} \left( Z_{n+1}^{(i)} = \ell \mid \text{particle } i \text{ in generation } n \right),
\]

that is, \( \tilde{q}_{\ell,n} \) is the probability that a particle in generation \( n \) spawns \( \ell \) children particles in generation \( n + 1 \) and we have used the fact that the maximum-likelihood estimate of the probability of having \( \ell \) children is given by the fraction of nodes in the data with \( \ell \) children [8].

In Fig. 3(a and b), we plot the empirical offspring distributions for several generations. Because of the data collection restriction to cascades of size exceeding one (and also because of the network structure, see Section 3 below), the seed generation offspring distribution \( \tilde{q}_{\ell,0} \) differs substantially from the other generations. However, for the Marref data set (Fig. 3a), observe that the \( \tilde{q}_{\ell,n} \) distributions for \( n = 1 \) through \( n = 4 \) (which is the range of generations giving \( z_n \) values above threshold in Fig. 2a) are very
similar to each other: the curves in Fig. 3a are almost indistinguishable. This collapse of the empirical offspring distributions is consistent with a branching process model in which the offspring distributions are identical for all generations with \( n \geq 1 \), see Section 3.

In the URL data set, the low-generation distributions \( \bar{q}_{\ell,n} \) do not show as clean a collapse as seen in the Marref case, see the inset of Fig. 3b. However, this may be due to the selection bias in the data collection, which means that small trees (those with fewer generations) are likely to have been omitted from the collected set of trees. Larger trees are more likely to be properly represented in the data set, and these trees are also likely to consist of a large number of generations. Accordingly, we plot also the \( \bar{q}_{\ell,n} \) curves for \( n = 10, 15, 20, 25, 30, 35, 40 \) in Fig. 3b (note the range of generations chosen matches the green dashed line in Fig. 2(b and d)), and we observe a good collapse of these distributions, which is again consistent with a branching process model.

From the evidence of Figs 2 and 3(a and b), we conclude that a branching process model may give a good approximation to the heterogeneous cascades represented by the trees extracted from the data sets. In the next section, we will derive a mathematical model that explicitly links the network structure and various hypotheses on the information-spreading mechanism to predict offspring distributions, which we then compare with the empirical results of Fig. 3(a and b).

3. Modelling information spread as a branching process

We consider a directed network whose structure is minimally described (in the configuration-model sense [30]) by the joint distribution \( p_{jk} \) of nodes’ in-degree \( j \) and out-degree \( k \): in other words, \( p_{jk} \) is the probability that a randomly chosen node has \( j \) friends and \( k \) followers. We model the dynamics of information spreading at the level of \( (j, k) \) classes also, defining the vulnerability \( v_{jk} \) as the probability that a \( (j, k) \)-class node will retweet a message that it has received from one of its \( j \) friends [31].

Consider a message that is tweeted by a node to its followers. Under the configuration-model assumption, the probability that a follower is in the \( (j, k) \) class is given by \( j p_{jk} / \langle j \rangle \), where \( \langle j \rangle = \sum_{j,k} j p_{jk} \) is the mean in-degree (mean number of friends) over the network. This follower will retweet the message if he is vulnerable, which occurs with probability \( v_{jk} \), and in doing so, he will expose all \( k \) of his followers to the message. Thus, the probability that a randomly chosen follower will retweet a message he receives is given by

\[
\rho = \sum_{j,k} \frac{j}{\langle j \rangle} p_{jk} v_{jk}. \tag{4}
\]

If we know that a follower has retweeted the message (i.e. if we condition on retweeting) then the probability that he is in the \( (j, k) \) class is

\[
\frac{1}{\rho} \frac{j}{\langle j \rangle} p_{jk} v_{jk}. \tag{5}
\]

1 Following [19, 26], we call the nodes followed by node \( i \) (the in-neighbours of \( i \)) its friends. The followers of \( i \) are the out-neighbours of \( i \), where we consider the direction of the edges to be the direction of information flow, that is, edges point from a node to its followers.
In particular, the probability that a retweeter has $k$ followers is given by summing over all possible $j$ values:

$$\sum_j \frac{1}{\rho \langle j \rangle} p_{jk} v_{jk}. \quad (6)$$

Assuming each of the $k$ followers to be independently vulnerable with probability $\rho$, the number $\ell$ of followers who themselves retweet has the binomial distribution

$$\binom{k}{\ell} \rho^\ell (1 - \rho)^{k-\ell}. \quad (7)$$

Combining these probabilities, we have derived the offspring distribution $q_\ell$ which gives the probability that a retweeting by a node will lead to $\ell$ further retweets by followers of that node as

$$q_\ell = \sum_k \sum_j \frac{1}{\rho \langle j \rangle} p_{jk} v_{jk} \binom{k}{\ell} (1 - \rho)^{k-\ell}. \quad (8)$$

The corresponding probability generating function (pgf) for the offspring distribution, $f(x) = \sum_\ell q_\ell x^\ell$, is

$$f(x) = \sum_{k,j} \frac{1}{\rho \langle j \rangle} p_{jk} v_{jk} (1 - \rho + \rho x)^k. \quad (9)$$

In the derivation of Eq. (9), we began by considering a node that receives the message from one of its friends. However, the initial source (or seed) of the cascade has a different dynamic, meaning that the seed generation of the branching process has an offspring distribution different from Eq. (8). We assume that the seed node for a cascade is chosen uniformly at random from all the nodes. This means that the seed node is in the $(j,k)$ class with probability $p_{jk}$. As above, the number $\ell$ of its $k$ followers who will retweet the message is given by Eq. (7), and so the seed-generation offspring distribution is

$$\tilde{q}_\ell = \sum_{k,j} p_{jk} \binom{k}{\ell} \rho^\ell (1 - \rho)^{k-\ell}, \quad (10)$$

with corresponding pgf

$$\tilde{f}(x) = \sum_{\ell=0}^\infty \tilde{q}_\ell x^\ell = \sum_{k,j} p_{jk} (1 - \rho + \rho x)^k. \quad (11)$$

Having defined how the offspring distribution of the branching process is determined by the network structure ($p_{jk}$) and the dynamics (via the vulnerability $v_{jk}$), we next examine two possible models for contagion dynamics.

---

2 We use tildes to differentiate the seed-generation offspring distribution and pgf from those defined in Eqs (8) and (9).
3.1 The independent cascade model

In the independent cascade model (ICM) [14], each ‘infected’ node (i.e. node who tweets or retweets the message of interest) gets one attempt to infect each of its out-neighbours; the infection attempt is successful (meaning that the follower also retweets the message) with probability $C$, where $C$ is the single parameter of the model. In our modelling framework, this implies that the ICM vulnerability of every node is equal to $C$, regardless of the node’s $(j,k)$ class:

$$v_{jk}^{ICM} = C \quad \forall j,k. \quad (12)$$

Note that in this case, the retweet probability $\rho$ is determined from Eq. (4) to be $\rho = C$. Moreover, in the special case of uncorrelated in- and out-degrees (i.e. if the number of friends $j$ and the number of followers $k$ of a node are uncorrelated), the joint distribution $p_{jk}$ factorises into the product $p_{jk}^{in}p_{jk}^{out}$ and the offspring $q_{j}$ and $\tilde{q}_{j}$ are identical. However, in the more realistic case where the in- and out-degrees of nodes (the numbers of friends and followers of users) are correlated (see, e.g. Fig. 2 of [32]), the offspring distribution of the seed generation differs from the offspring distributions of subsequent generations.

The branching number for (non-seed) nodes in cascades generated by this model is determined from Eqs (9) and (12) as

$$\xi = f'(1) = \frac{<jk>}{<j>} C, \quad (13)$$

which will be useful when fitting to empirical data in Sec. 3.3.

3.2 The limited attention model

A number of researchers have pointed out that the limitations of human cognition impose an effective limit on how much information can be absorbed and shared by an individual. For a user on Twitter, having a larger number of friends $j$ leads to a faster influx of information into the user’s stream, with a consequent dividing of attention among the many tweets. Empirical analyses [19, 24] and models of information-sharing dynamics [11, 12, 33] both indicate that the probability that a user retweets a particular piece of information she has received can be modelled as being approximately inversely proportional to the number $j$ of her friends. In our notation, the vulnerability $v_{jk}$ of a $(j,k)$-class user in the limited attention model (LAM) is inversely proportional to $j$:

$$v_{jk}^{LAM} = \frac{B}{j}, \quad (14)$$

where $B$ is a parameter of the model, and we assume no nodes have $j = 0$.

In the LAM, the probability $\rho$ of a random follower retweeting is given in terms of $B$ by Eq. (4):

$$\rho = \frac{B}{<j>} \quad (15)$$

---

3 This is easily seen from the corresponding pgfs, where the sum over $j$ can be performed in Eqs (9) and (11) to give $f^{ICM}(x) = \tilde{f}^{ICM}(x) = \sum_k p_{jk}^{out} (1 - Cx)^k$ if $p_{jk} = p_{jk}^{in}p_{jk}^{out}$.
Interestingly, under the assumption that the network has no nodes with \( j = 0 \) then the LAM offspring distributions for the seed generation and for later generation are identical, even if the in- and out-degrees of nodes are correlated (unlike the ICM model):

\[
f^{\text{LAM}}(x) = \tilde{f}^{\text{LAM}}(x) = \sum_{k,j} p_{jk} (1 - \rho + \rho x)^k \quad \text{if } p_{00}^\text{in} = 0. \tag{16}
\]

The branching number for this model is related to the parameter \( B \) by

\[
\xi = f'(1) = \frac{\langle k \rangle}{\langle j \rangle} B, \tag{17}
\]

which again is useful for fitting.

### 3.3 Comparing ICM and LAM with empirical offspring distributions

Using the empirical network structure for the Marref and URL data sets, specifically the in-degree \( j_i \) and out-degree \( k_i \) of each node \( i \) in the network, we construct the offspring distribution predicted by the independent cascade model and by the limited attention model, using Eqs (12) and (14), respectively, in Eqs (4), (8) and (10). In each case, we fit the parameters \( C \) and \( B \) by matching the branching number to the average value \( \bar{\xi} \) calculated in Section 2.2, using Eqs (13) and (17). The sums over \( j \) and \( k \) are replaced by sums over the \( N \) nodes: Equation (4), for example, becomes

\[
\rho = \sum_{i=1}^{N} \bar{j}_i \frac{1}{\bar{j}} \sum_{i=1}^{N} j_i, \tag{18}
\]

where \( \bar{j} \) is the sample mean of the in-degrees: \( \bar{j} = \frac{1}{N} \sum_{i=1}^{N} j_i \). (In effect, we replace \( p_{jk} \) by \( 1/N \) and replace sums over \( j \) and \( k \) by a sum over all nodes.)

The black (for ICM) and magenta (for LAM) curves in Fig. 3 show how these predictions compare with the empirical offspring distribution. Evidently, the LAM predictions are closer to the empirical offspring distributions than the ICM predictions, at least for the relatively low values of \( \ell \) in Fig. 3(a and b). To examine the empirical offspring distributions at higher values of \( \ell \) we reduce the low-number fluctuations by averaging the distributions over the generations marked with the green line in Fig. 2(a and b), that is, those generations for which the effective branching number is approximately constant. This averaged offspring distribution is shown by the blue symbols in Fig. 3(c and d): note we plot \( \ell + 1 \) on the horizontal axis in order to make the \( \ell = 0 \) case visible on the logarithmic scale.

Noting the near-linear decay of the offspring distribution on the log–log plot, we fit the empirical averaged offspring distribution with a truncated power law:

\[
q_\ell \propto (\ell + 1)^{-\beta} e^{-\ell \theta}. \tag{19}
\]

This distribution is chosen for its good fit and analytical convenience\(^4\); calculations with this distribution can be more easily reproduced than by using the full ICM or LAM distributions, which require knowledge

---

\(^4\) Its pgf is \(Li_\beta \left(e^{-\frac{1}{\theta}} x\right) / \left(x Li_\beta \left(e^{-\frac{1}{\theta}}\right)\right)\), where \(Li_\beta\) is the polylogarithm function of order \(\beta\).
of the full set of node degrees \((j, k)\). To fit the parameters \(\beta\) and \(\theta\) in Eq. (19), we match the first and second moments of the distribution with the corresponding moments of the averaged empirical distribution. The fitted parameters are given in Table 1, and the red curves in Fig. 3(c and d) show that the fitted offspring distribution is reasonably close to the empirical distribution. A similar procedure is used to fit a seed generation offspring distribution \(\tilde{q}_\ell\), using the form of Eq. (19) with parameters \(\beta\) and \(\theta\) replaced by \(\beta_0\) and \(\theta_0\), and with the domain restricted to \(\ell > 1\) because every seed node in the empirical data set has at least one child.\(^5\)

To summarize this section: we have derived a general formulation for the offspring distribution that results from cascades on a network with a given distribution \(p_{jk}\) of in- and out-degrees. We used the vulnerability \(v_{jk}\) to describe different models of information transmission, focussing on comparing the ICM with the LAM. In Fig. 3, we see that there are observable differences between the offspring distributions predicted by the two models, with the LAM case generally closer to the empirical observations. Finally, we fitted a standard distribution (Eq. (19)) to the empirical distribution to make our results in the next section more tractable and readily reproducible. Note, however, that in principle the data on the structure of the network (e.g. the \(p_{jk}\) distribution) and the assumed vulnerability \(v_{jk}\) suffice to determine the offspring distribution, and this opens the possibility of examining further hypotheses on the dependence of information spreading on the nodes’ in- and out-degrees [34].

It is also worth noting that the network structure, through the correlations in the \(p_{jk}\) distribution, has an impact on the offspring distribution (and hence, as we show in the next section, the predictions of the cascade structure); this point has recently been recognized by Ma et al. [32]. We point out that the \(p_{jk}\) distribution of the network should therefore be included, when possible, in analysis of information spreading. This is not current practice in Ref. [12, 13], for example, large-scale simulations are performed on synthetic networks with specified out-degree distributions but without considering the correlation structure between the in- and out-degrees of nodes.

### 4. Application of branching process model

In this section, we focus on analytical predictions of branching process theory that can be compared to statistical features of the two data sets. We begin with a discrete-time branching process, where—as in Section 3—the number of offspring of the seed particle is distributed according to \(\text{pgf } \tilde{f}(x) = \sum_{\ell=0}^{\infty} \tilde{q}_\ell x^\ell\) while all later generations of the tree have offspring numbers generated by \(f(x) = \sum_{\ell=0}^{\infty} q_\ell x^\ell\). We consider

---

\(^5\) The pgf for the seed generation is \(\left( x - e^x \text{Li}_\beta \left( e^{-\frac{1}{2}} x \right) \right) / \left( x - xe^x \text{Li}_\beta \left( e^{-\frac{1}{2}} x \right) \right)\).
the seed of the tree to be generation 0, and we are interested in various properties of the trees as observed a number of generations later. In Sections 4.1 and 4.2, we use a slightly unusual approach to derive known results on the distribution of cascade durations and sizes. We then extend this methodology to the calculation of other metrics in Sections 4.3 and 4.4.

4.1 Number of particles in generation n; distribution of cascade lifetimes

As a first example, we define the random (non-negative integer) variable $\tilde{Z}_n$ to represent the number of particles in generation $n$ of the tree (the small nodes in Fig. 4). As schematically represented in Fig. 4, these particles are the descendants of the generation-0 seed node, observed $n$ generations after the seed. They can also be considered as the sum of the particles contained in all the subtrees that are seeded at generation 1 and which are observed $n - 1$ generations later. Conditioning on the number $k$ of particles in generation 1, we define $Z_{n-1}^{(i)}$ to be the number of particles in the subtree that is seeded by the $i$th particle in generation 1, as observed $n - 1$ generations after the subtree is born (i.e. at generation $n$ of the parent tree). Since all the subtrees are i.i.d., each of the $k$ random variables $Z_{n-1}^{(i)}$ has the same distribution.

We define the pgf $\tilde{F}_n$ for the random variable $\tilde{Z}_n$ as

$$\tilde{F}_n(s) = E \left( s^{\tilde{Z}_n} \right) = \sum_{j=0}^{\infty} \text{Prob} (\tilde{Z}_n = j) s^j,$$

(20)

where $E$ denotes expectation over the ensemble of trees and $s$ is a dummy variable. If we condition on the number $k$ of particles in generation 1, we can write $\tilde{Z}_n$ as the sum of the $k$ subtree variables $Z_{n-1}^{(i)}$ (the superscript $i$ denotes the $i$th i.i.d. copy):

$$\tilde{Z}_n = \sum_{i=1}^{k} Z_{n-1}^{(i)},$$

(21)
and so
\[
E \left( s^{Z_n} \mid k \text{ particles in generation 1} \right) = E \left( s^{k \sum_{i=1}^{k} Z_{n-1}^{(i)}} \right) \\
= E \left( s^{Z_{n-1}^{(1)}} \right) E \left( s^{Z_{n-1}^{(2)}} \right) \ldots E \left( s^{Z_{n-1}^{(k)}} \right) \\
= \left[ E \left( s^{Z_{n-1}} \right) \right]^k,
\]
(22)
where we have used the independence of the subtrees and the i.i.d. nature of the \(Z_{n-1}^{(i)}\) variables.

Writing \(F_{n-1}(s)\) for the pgf \(E \left( s^{Z_{n-1}} \right)\) and summing over all possible values of \(k\) (recall that \(\tilde{q}_k\) is the probability that there are \(k\) children of the seed particle, i.e. \(k\) particles in generation 1) yields
\[
\tilde{F}_n(s) = E \left( s^{Z_n} \right) = \sum_{k=0}^{\infty} \tilde{q}_k E \left( s^{Z_n} \mid k \text{ particles in generation 1} \right) \\
= \sum_{k=0}^{\infty} \tilde{q}_k \left[ E \left( s^{Z_{n-1}} \right) \right]^k \\
= \tilde{f} \left( F_{n-1}(s) \right). \quad (23)
\]

This equation relates the pgf for \(\tilde{Z}_n\) to the pgf for the subtree quantities \(Z_{n-1}\). The next step is to derive an equation that recursively links \(Z_n\) (the number of particles in a subtree \(n\) generations after its birth) to \(Z_{n-1}\).

Figure 5 is a schematic view of this relationship. The main subtree in Fig. 5 is born with the first particle shown (left of the figure) and we condition on the number \(k\) of children particles of this first particle; recall that \(k\) is a random variable with pgf \(f(x)\). The number \(Z_m\) of particles in the main subtree after \(m\) generations is equal to the sum of the \(k\) i.i.d. variables \(Z_{m-1}^{(i)}\):
\[
Z_m = \sum_{i=1}^{k} Z_{m-1}^{(i)}, \quad (24)
\]
and so

\[ E \left( s^m \mid k \text{ particles in generation 1} \right) = E \left( s^{\sum_{i=1}^{k} Z_{m-1}^{(i)}} \right) = \left[ E \left( s^{Z_{m-1}} \right) \right]^k , \]  

(25)

as in Eq. (23). Summing over the possible values of \( k \) then yields

\[ F_m(s) = E \left( s^m \right) = \sum_{k=0}^{\infty} q_k E \left( s^m \mid k \text{ particles in generation 1} \right) = \sum_{k=0}^{\infty} q_k \left[ E \left( s^{Z_{m-1}} \right) \right]^k = f \left( F_{m-1}(s) \right) . \]  

(26)

Equation (26) gives a recursion relation for the pgf \( F_m(s) \), starting from the initial condition \( F_0(s) = s \), corresponding to the tree being seeded from a single particle. Using the result of the recursion Eq. (26) in Eq. (23) then gives the pgf for the number of nodes in generation \( n \) of the tree. This characterization of the branching process is called the backward approach in Ref. [25], in analogy with the backward Chapman–Kolmogorov equation of Markov processes. An alternative forward approach—wherein the states of particles in generation \( n+1 \) is predicted from the state of the process after \( n \) generations—is often used to derive Eq. (26), but we will find the backward approach easily generalizable to other quantities of interest.

The probability that the tree is terminated at or before generation \( n \) is equal to the probability of the tree having zero nodes in generation \( n \), which is \( \tilde{F}_n(0) \). The probability that the tree terminates precisely at generation \( n \) (i.e. that there is a nonzero number of particles in generation \( n \) but each of these has zero offspring) is therefore

\[ \Omega_n = \tilde{F}_n(0) - \tilde{F}_{n-1}(0) = f \left( F_{n-1}(0) \right) - f \left( F_{n-2}(0) \right) , \]  

(27)

where \( F_n(0) \) is calculated by iteration from Eq. (23) and the initial condition \( F_0(0) = 0 \). We call \( \Omega_n \) the lifetime distribution of trees, as it gives the probability that the observed lifetime of a tree is \( n \) generations. See Fig. 6 for a comparison of the empirical lifetime distribution with the predictions of Eq. (27), using the offspring distribution fitted in Eq. (19).

### 4.2 Cascade size

A similar approach can be applied to calculate the distribution of tree (cascade) sizes, that is, the total number of particles that are in all generations of the tree, from the seed at generation 0 up to the last generation of the tree (this quantity is sometimes called the total progeny of the tree). We define the random variable \( \tilde{X}_n \) to be the size of the tree observed \( n \) generations after its seed particle is born. As before, \( \tilde{X}_n \) can be decomposed into the sum of contributions from each of the subtrees born in generation
Fig. 6. Lifetime distribution of cascades in Marref (left) and URL (right) data sets. Blue symbols are empirical values; red line shows the theoretical distribution from Eq. (27), using the offspring distribution of Eq. (19).

1. Conditioning on the seed node having \( k \) children particles, we write

\[
\tilde{X}_n = 1 + \sum_{i=1}^{k} X_{n-1}^{(i)},
\]

where \( X_{n-1}^{(i)} \) represents the \( i \)th i.i.d. subtree size as observed after \( n - 1 \) generations, and the first term counts the seed node of the tree (see Fig. 4). Using identical arguments to those leading to Equations (22) and (23), we obtain

\[
E \left( x^{\tilde{X}_n} \right) = x \left[ E \left( x^{X_{n-1}} \right) \right]^k
\]

and the pgf \( \tilde{G}_n(x) = E \left( x^{\tilde{X}_n} \right) = \sum_{j=0}^{\infty} \text{Prob} \left( \tilde{X}_n = j \right) x^j \) is then given by

\[
\tilde{G}_n(x) = \sum_{k=0}^{\infty} \tilde{q}_k E \left( x^{\tilde{X}_n} \middle| k \text{ particles in generation 1} \right)
= \sum_{k=0}^{\infty} \tilde{q}_k x \left[ E \left( x^{X_{n-1}} \right) \right]^k
= x \tilde{f} \left( G_{n-1}(x) \right),
\]

where \( G_{n-1}(x) = E \left( x^{X_{n-1}} \right) \) is the pgf for the size of a subtree after \( n - 1 \) generations. Referring now to Fig. 5, the recursion relation for the subtree sizes is derived by first assuming \( k \) children in the first generation of the subtree:

\[
X_m = 1 + \sum_{i=1}^{k} X_{m-1}^{(i)},
\]
and then proceeding as in Equations (26) and (30) to obtain the recursion relation

\[ G_m(x) = xf(G_{m-1}(x)), \]  

with initial condition \( G_0(x) = x \).

By iterating Eq. (32) for \( m = 1, 2, \ldots, n-1 \) and then substituting \( G_{n-1}(x) \) into Eq. (30), we obtain the desired pgf \( \tilde{G}_n(x) \) describing the distribution of cascade sizes after \( n \) generations. In order to invert the pgf to obtain the distribution of cascade sizes, we iterate Eqs (32) and (30) for a set of \( x \) values that are uniformly spaced around the unit circle in the complex \( x \)-plane, and use a fast Fourier transform to approximate the Cauchy integral

\[ \text{Prob}(\tilde{X}_n = j) = \frac{1}{j!} \left. \frac{d^j \tilde{G}_n}{dx^j} \right|_{x=0} = \frac{1}{2\pi i} \oint_{C} \tilde{G}_n(x)x^{-(j+1)}dx, \]

as in Section S2 of [11].

Figure 7 shows the large-\( n \) limit of the cascade size distribution, and compares it with the empirical distribution. The good agreement between this theoretical prediction and the empirical results gives further support to the usage of branching process descriptions for such data.

4.3 Average tree depth and structural virality

In this subsection, we build on the approach used in Section 4.1 to derive results for measures of the shape of cascade trees, which are of considerable interest in analyses of Twitter [13, 18]. We focus on the distribution (and expected value) of two quantities [13]: the average depth of a tree, and the structural virality of a tree.

4.3.1 Average tree depth

To calculate the average depth of a sample tree, we first sum the depths (generation numbers) of all particles in the tree to obtain the cumulative depth of the tree, and then divide this by the size of the tree (the total number of particles in the tree), see Fig. 8. In this subsection, we generalize the methods used in Section 4.2 to calculate the joint distribution of tree size and cumulative depth, and hence to find a formula for the expected average tree depth (EATD). In the ensemble of trees generated by the branching process, each tree has its own average depth, and the EATD is the mean of the average depths over all trees in the ensemble. We believe the formula we derive (Eq. (46)) is novel.

We extend the approach of Section 4.2 to consider the joint distribution of \( \tilde{X}_n \) (the tree size after \( n \) generations) and of \( \tilde{Y}_n \), which is the random variable giving the cumulative depth of the tree after \( n \) generations. We define the two-variable pgf \( \tilde{H}_n(x,y) \) as

\[ \tilde{H}_n(x,y) = E \left( x^{\tilde{X}_n} y^{\tilde{Y}_n} \right) = \sum_{j,\ell=0}^{\infty} \text{Prob}(\tilde{X}_n = j \text{ and } \tilde{Y}_n = \ell) x^j y^\ell. \]  

(34)

As in earlier sections, we relate the variables \( \tilde{X}_n \) and \( \tilde{Y}_n \) to subtree quantities, and begin by assuming that the seed node (in Fig. 4 e.g.) has \( k \) children. Each of the \( k \) children generates a subtree with (after \( n-1 \) generations of the subtree) respective random-variable pairs \( \left( \tilde{X}_{n-1}^{(i)}, \tilde{Y}_{n-1}^{(i)} \right) \) for \( i = 1, 2, \ldots, k \).

The relationship between \( \tilde{X}_n \) and \( \tilde{X}_{n-1}^{(i)} \) is given by Eq. (28) but we must now also find an analogous expression for \( \tilde{Y}_n \). We define \( \tilde{Y}_{n-1}^{(i)} \) to be the cumulative depth of the \( i \)th i.i.d. subtree. Notice (see Fig. 8)
Fig. 7. Cascade size distributions: pdfs (top panels) and ccdfs (bottom panels) for Marref (left) and URL (right). Blue symbols are empirical values; red line shows the theoretical distribution from Section 4.2, using the offspring distribution of Eq. (19).

Fig. 8. This is a tree of size 5. Each of the 5 particles is labelled by its depth (its generation). The cumulative depth of the tree is $0 + 1 + 1 + 2 + 2 = 6$, and so the average depth of this tree is $6/5$. Note that the cumulative depth of the top subtree is $0 + 1 + 1 = 2$, also that each node in the subtree has a depth that is one larger than its value when considered as part of the main tree: See Eq. (35).
that when we add the $Y_{n-1}^{(i)}$ values for all the subtrees, each node of the subtree has a depth that is one less that its depth in the main tree. Therefore, the $i$th subtree contributes to $\tilde{Y}_n$ a total of $Y_{n-1}^{(i)} + X_{n-1}^{(i)}$, where the second term adds one for each node in the subtree. From this relationship, we obtain

$$\tilde{Y}_n = \sum_{i=1}^{k} \left( Y_{n-1}^{(i)} + X_{n-1}^{(i)} \right), \tag{35}$$

and with Eq. (28) we find the pgf relations

$$\tilde{H}_n(x, y) = \sum_{k=0}^{\infty} \tilde{q}_k E \left( x^{\tilde{X}_n} y^{\tilde{Y}_n} \bigg| k \text{ particles in generation 1} \right)$$

$$= \sum_{k=0}^{\infty} \tilde{q}_k E \left( x^{1+\sum_{i=1}^{k} X_{n-1}^{(i)}} y^{\sum_{i=1}^{k} (Y_{n-1}^{(i)} + X_{n-1}^{(i)})} \right)$$

$$= x \sum_{k=0}^{\infty} \tilde{q}_k E \left( (xy)^{\sum_{i=1}^{k} X_{n-1}^{(i)}} y^{\sum_{i=1}^{k} Y_{n-1}^{(i)}} \right)$$

$$= x \sum_{k=0}^{\infty} \tilde{q}_k [E \left( (xy)^{\tilde{X}_{n-1}} y^{\tilde{Y}_{n-1}} \right)]^k$$

$$= x \tilde{H}_{n-1}(x, y), \tag{36}$$

where $H_{n-1}(x, y) = E \left( x^{\tilde{X}_{n-1}} y^{\tilde{Y}_{n-1}} \right)$.

Addressing the recursion relation for the subtrees in a similar fashion leads (as in Section 4.2) to

$$H_m(x, y) = x f \left( H_{m-1}(x, y) \right), \tag{37}$$

with initial condition $H_0(x, y) = x$ (since a single particle is a tree of size 1, with zero depth). Iterating Eq. (37) for $m = 1, 2, \ldots, n - 1$ and substituting into Eq. (36) yields the pdf $\tilde{H}_n(x, y)$ for the joint distribution of trees size and cumulative depth after $n$ generations.

We can use this joint distribution to calculate the EATD for trees of $n$ generations as

$$d_n = \sum_{j=0}^{\infty} \sum_{\ell=0}^{\infty} \text{Prob} \left( \tilde{X}_n = j \text{ and } \tilde{Y}_n = \ell \right) \frac{\ell}{j}$$

$$= \int_0^1 \frac{1}{x} \left. \frac{\partial \tilde{H}_n}{\partial y} \right|_{y=1} dx, \tag{38}$$

as can be verified by term-by-term differentiation and integration of the series in Eq. (34). Taking the $n \to \infty$ limit in order to include all trees, Eq. (37) give a self-consistent equation for $H_\infty(x, y)$ and it can be differentiated with respect to $x$ to yield

$$\left. \frac{\partial H}{\partial x} \right|_{y=1} = \frac{f \left( H(x, 1) \right)}{1 - xf' \left( H(x, 1) \right)}, \tag{39}$$
where, for simplicity, we drop the subscript from $H_{\infty}$ for the remainder of this section. Similarly, differentiation of Eq. (37) with respect to $y$ gives

$$\left. \frac{\partial H}{\partial y} \right|_{y=1} = x f'(H(x, 1)) \left[ \left. \frac{x \partial H}{\partial x} \right|_{y=1} + \left. \frac{\partial H}{\partial y} \right|_{y=1} \right],$$

(40)

which can be solved for $\left. \frac{\partial H}{\partial y} \right|_{y=1}$, after substituting for $\left. \frac{x \partial H}{\partial x} \right|_{y=1}$ from Eq. (39):

$$\frac{\partial H}{\partial y} \left|_{y=1} = \frac{x f'(H(x, 1)) f(H(x, 1))}{\left[ 1 - x f'(H(x, 1)) \right]^2}. \right.$$

(41)

Differentiating the $n \to \infty$ limit of Eq. (36) with respect to $y$ yields

$$\left. \frac{\partial \tilde{H}}{\partial y} \right|_{y=1} = x \tilde{f}'(H(x, 1)) \left[ \left. \frac{x \partial H}{\partial x} \right|_{y=1} + \left. \frac{\partial H}{\partial y} \right|_{y=1} \right],$$

(42)

and substituting from Eqs (39) and (41) gives

$$\left. \frac{\partial \tilde{H}}{\partial y} \right|_{y=1} = \frac{x \tilde{f}'(H(x, 1)) f(H(x, 1))}{\left[ 1 - x \tilde{f}'(H(x, 1)) \right]^2}. \right.$$

(43)

Thus, the expected average tree depth over all trees is given by Eq. (38) as

$$d = \int_0^1 x \tilde{f}'(H(x, 1)) f(H(x, 1)) \frac{dx}{\left[ 1 - x \tilde{f}'(H(x, 1)) \right]^2}.$$

(44)

Noting that Eq. (37) relates $H(x, 1)$ to $x$ through the implicit relation

$$H(x, 1) = x f'(H(x, 1)), \right.$$

(45)

we make the change of integration variable $x \mapsto h$ defined implicitly by $x = h/f(h)$ (with $dx = (f(h) - hf'(h))/f^2(h) dh$), to yield a simple integral formula for the EATD:

$$d = \int_0^1 \frac{h \tilde{f}'(h)}{f(h) - hf'(h)} dh.$$

(46)

This remarkably simple formula is easily evaluated once the offspring distributions $f$ and $\tilde{f}$ of the branching process are given. In Table 2, we show that it agrees with Monte Carlo simulations and also gives quite a reasonably accurate estimate of the values found from the empirical data.
Fig. 9. Ccdfs of average tree depth (top panels) and structural virality (bottom panels) for Marref (left) and URL (right). Blue symbols are empirical distributions; red symbols are from Monte Carlo simulations of branching processes with offspring distribution $q_\ell$ given by Eq. (19), and by the corresponding distribution $q_\ell'$ for the seed’s offspring.

Table 2  Expected average tree depth (EATD) from the integral formula of Eq. (46), compared with Monte Carlo simulations ($10^5$ realizations) and the data values. Bootstrap intervals given for the latter two cases show quantile 0.025 to quantile 0.975 (i.e. 95% of cases) for the expected average tree depth, using $10^3$ bootstrap samples

<table>
<thead>
<tr>
<th></th>
<th>Marref</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral</td>
<td>0.862</td>
<td>1.22</td>
</tr>
<tr>
<td>Monte Carlo</td>
<td>0.862 (0.859, 0.866)</td>
<td>1.22 (1.21, 1.23)</td>
</tr>
<tr>
<td>Data</td>
<td>0.899 (0.887, 0.912)</td>
<td>1.34 (1.32, 1.36)</td>
</tr>
</tbody>
</table>

4.3.2 Structural virality  The structural virality of a tree with size $n > 1$ was introduced in Goel et al. [13] as

$$
\frac{1}{n(n - 1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij},
$$

(47)
where $d_{ij}$ is the graph distance from node $i$ to node $j$. The distribution of this metric across an ensemble of trees was used to fit models to data in [13].

As noted in Ref. [13], the structural virality of a tree is closely related to its Wiener index, defined by $\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}$. If we consider the expected value of the Wiener index across the ensemble of trees generated by the branching process, we can usefully adapt the approach of Entringer et al. [35], with the aim of calculating the expected structural virality for the ensemble. Entringer et al. define a generating function $W(x)$ so that the coefficient of $x^n$ in the power series is the contribution of trees of size $n$ to the ensemble-averaged Wiener index (note that $W(x)$ is not a pgf). Their Eq. (3.5) is

$$W(x) = D(x) + x f'(G) W(x) + x f''(G) [D(x) + x G'] x G',$$

where $D(x)$ is our $\frac{\partial H}{\partial y} \bigg|_{y=1}$ from Eq. (41) and $G(x) = H(x, 1)$ is the $m \to \infty$ cascade size pgf from Eq. (32). The first term in Eq. (48) comes from considering pairs of vertices $u$ and $v$ where one of $u$ or $v$ is the root of the tree. The second term arises from the case where $u$ and $v$ belong to the same subtree, and the third term stems from $u$ and $v$ belonging to different subtrees, see Section 3 of Ref. [35] for details.

We extend the approach of Eq. (48) to the case where the seed node of the tree has offspring distribution with pgf $\tilde{f}$, to get an analogous equation for $\tilde{W}(x)$:

$$\tilde{W}(x) = \tilde{D}(x) + x \tilde{f}'(G) W(x) + x \tilde{f}''(G) [D(x) + x G'] x G',$$

where $\tilde{D}(x)$ is $\tilde{D}(x)$ as given by Eq. (43).

Solving Eq. (48) for $W(x)$ and substituting into Eq. (49) enables us to determine $\tilde{W}(x)$. The expected structural virality for the ensemble of trees is then given by

$$s = \sum_{n=2}^{\infty} \frac{\tilde{W}_n}{n(n-1)},$$

where $\tilde{W}_n$ is the coefficient of $x^n$ in the power series of $\tilde{W}(x)$. The value of $s$ can be calculated from the generating function $\tilde{W}(x)$ by a double integration:

$$s = \int_{0}^{1} \int_{0}^{y} \frac{\tilde{W}(x)}{x^2} \, dx \, dy$$

$$= \int_{0}^{1} \frac{\tilde{W}(x)}{x^2} (1-x) \, dx,$$

where the second equation follows from changing the order of integration, that is, using the identity

$$\int_{0}^{1} \int_{0}^{y} (\cdot) \, dx \, dy = \int_{0}^{1} \int_{y}^{1} (\cdot) \, dy \, dx.$$  

More precisely, $W(x)$ is defined in Ref. [35] as $W(x) = \sum_{n=1}^{\infty} w_n x^n$ where $w_n$ is defined as $\sum c(T_n) Q(T_n)$ with the sum being over all $n$-vertex trees $T_n$, the weights $c(T_n)$ are the likelihoods of tree $T_n$ in the ensemble, and $Q(T_n)$ is the Weiner index of tree $T_n$.  

Table 3: Expected structural virality from the integral formula of Eq. (53), compared with Monte Carlo simulations ($10^5$ realizations) and the data values. Bootstrap intervals given for the latter two cases show quantile 0.025 to quantile 0.975 (i.e. 95% of cases) for the expected structural virality, using $10^3$ bootstrap samples.

<table>
<thead>
<tr>
<th></th>
<th>Marref</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral</td>
<td>1.44</td>
<td>1.81</td>
</tr>
<tr>
<td>Monte Carlo</td>
<td>1.440 (1.436, 1.443)</td>
<td>1.82 (1.81, 1.83)</td>
</tr>
<tr>
<td>Data</td>
<td>1.47 (1.46, 1.49)</td>
<td>1.77 (1.75, 1.78)</td>
</tr>
</tbody>
</table>

Combining these results and then making the same change of variable as for Eq. (46) yields an integral formula for the expected structural virality:

$$s = \int_0^1 f (f - h) \left( \tilde{f} \tilde{f}'' + hf \tilde{f}'' - hf \tilde{f}' + h^2 \left( f'' \tilde{f}' - \tilde{f}'' f' \right) \right) \frac{dh}{(f - hf')^3}.$$  (53)

Table 3 shows that this formula agrees with Monte Carlo simulations of the branching process, and also matches reasonably well to the average structural virality of the ensemble of empirical trees in both data sets.

4.3.3 Asymptotic analysis The integral formulas derived for the expected average tree depth (Eq. (46)) and the expected structural virality (Eq. (53)) enable us to analytically study the impact of the spreading process upon these measures. Such understanding can assist in the fitting of information-spreading models to empirical data. In Fig. 2 of Ref. [13], for example, large-scale numerical simulations are used to calculate the dependence of the expected structural virality on the branching number, and this information is then used to guide model parameter fitting.

We are therefore motivated to examine how the integrals in Eqs (46) and (53) depend upon the form of the offspring distribution (through its pgf $f(x)$) and in particular on the branching number $\xi = f'(1)$. For simplicity we will restrict ourselves in this section to the case where $\tilde{f}(x) = f(x)$, that is, assuming that the seed node’s offspring distribution is the same as that of the later generations.

First, we note that both integrals may be performed exactly in the special case of a binary fission process [1], where each parent has either zero or two children:

$$f_{bf}(x) = \left(1 - \frac{\xi}{2}\right) + \frac{\xi}{2}x^2.$$  (54)

The exact integrals for EATD and expected structural virality in this case are

$$d_{bf} = 2 \sqrt{\frac{2 - \xi}{\xi}} \text{ArcTanh} \sqrt{\frac{\xi}{2 - \xi}} - 2,$$

$$s_{bf} = 1 - \frac{\xi}{2} + (2 - \xi) \log \left( \frac{2 - \xi}{2 - 2\xi} \right) - \sqrt{\frac{2 - \xi}{\xi}} \text{ArcTanh} \sqrt{\frac{\xi}{2 - \xi}}.$$  (55)
and each shows a logarithmic divergence as the branching number $\xi$ approaches the critical value of 1 from below (see dashed curves in Fig. 10).

In fact, this logarithmic divergence as $\xi \to 1$ is not unique to the exactly-solvable binary fission example. Indeed, asymptotic analysis of the integrals shows that a similar divergence occurs for any offspring distribution that has a finite value of $f''(1)$, meaning that the second moment of the offspring distribution is finite. The integrands in Eqs (46) and (53) are singular at $h = 1$, and the form of the singularity can be understood using the expansion of $f(h)$ about $h = 1$:

$$f(h) \sim f(1) + f'(1)(h - 1) + \frac{1}{2}f''(1)(h - 1)^2 + \ldots$$

$$= 1 - \xi(1 - h) + \frac{1}{2}f''(1)(1 - h)^2 + \ldots \text{ as } h \to 1^-.$$ (56)

The integrand of Eq. (46), for example, has leading-order expansion

$$\sim \frac{\xi}{1 - \xi + f''(1)(1 - h)} \text{ as } h \to 1^- $$ (57)

and so the integral diverges logarithmically as $\xi \to 1$; the same asymptotic behaviour is found for the integrand in Eq. (53). Hence the behaviour of the dashed curves in Fig. 10 is quite generic for offspring distributions with finite second moments.

Offspring distributions with infinite second moments are also of interest, as they relate to heavy-tailed follower distributions in the Twitter network [12, 36]. An important example is the case of a power-law tail, that is,

$$q_\ell \sim D \ell^{-\gamma} \text{ as } \ell \to \infty,$$ (58)
for constant $D$ and for values of the exponent between 2 and 3. The asymptotic series for $f(h)$ as $h \to 1^-$ is given in this case by [11, 37]

$$f(h) \sim 1 - \xi (1 - h) + D\Gamma(1 - \gamma)(1 - h)^{\gamma - 1} \text{ as } h \to 1^-, \quad (59)$$

where $\Gamma(\cdot)$ is the Gamma function. Using this asymptotic series, the integrands in both Eqs (46) and (53) have the leading order behaviour $\sim (1 - h)^{2 - \gamma}$ as $h \to 1^-$ at the critical value of $\xi = 1$. Since this singularity is integrable, the resulting values of $d$ and $s$ are both finite at $\xi = 1$, in contrast to the divergence seen in the case where $f''(1)$ is finite. The example of the solid curves in Fig. 10 is for the offspring distribution where $q_\ell \propto \ell^{-\gamma}$ for all $\ell \geq 1$ (and $q_0 = 1 - \sum_{\ell>0} q_\ell$), with power-law exponent $\gamma = 2.5$. The finite limits of $d$ and $s$ as the branching number approaches 1 are evident.

4.4 Novelty decay in a multiplicative process model

Multiplicative stochastic processes have been used in a number of papers to model popularity growth [20, 21]. In our notation, the assumption of the multiplicative model is that the total number of tweets by generation $n$ (i.e. the tree size $\tilde{X}_n$) can be considered as proportional to the number of tweets that occurred in all previous generations ($\tilde{X}_{n-1}$), multiplied by a random factor $W_n$ that is modulated by a novelty decay factor $r_n$:

$$\tilde{X}_n = (1 + r_n W_n) \tilde{X}_{n-1}. \quad (60)$$

Here, the random variables $W_1, W_2, \ldots$ are assumed to be positive, independent, and identically distributed for each tree, while $r_n$ is a deterministic novelty decay factor that is common to all trees.

The novelty decay factor for this model can be obtained from Eq. (60) by rewriting it as

$$r_n W_n = \frac{\tilde{X}_n - \tilde{X}_{n-1}}{\tilde{X}_{n-1}} \quad (61)$$

and taking expectations (i.e. averaging over all trees). The deterministic novelty decay factor $r_n$ is then proportional to the expectation of the right hand side of Eq. (61). Using the fact that the number of particles $\tilde{Z}_n$ in generation $n$ can be related to the tree size by

$$\tilde{Z}_n = \tilde{X}_n - \tilde{X}_{n-1}, \quad (62)$$

we, therefore, consider the calculation of the quantity

$$\tilde{r}_n = E\left(\frac{\tilde{Z}_n}{\tilde{X}_n - \tilde{Z}_n}\right) \quad (63)$$

which, up to a multiplicative constant, is the novelty decay factor in such models. (The multiplicative constant is often set, as in Ref. [20] e.g., by normalizing the value of $r_1$).

Similar to Eq. (34), we consider here the joint distribution, at generation $n$, of tree size $\tilde{X}_n$ and number of particles $\tilde{Z}_n$, defining the two-variable pgf $\tilde{K}_n(x, z)$ as

$$\tilde{K}_n(x, z) = E\left(x^{\tilde{X}_n} z^{\tilde{Z}_n}\right) = \sum_{j, \ell=0}^{\infty} \text{Prob}(\tilde{X}_n = j \text{ and } \tilde{Z}_n = \ell) x^j z^\ell. \quad (64)$$
The iteration equation for $\tilde{K}_n$ is, similar to Eqs (30) and (23),

$$\tilde{K}_n(x, z) = x\tilde{f} (K_{n-1}(x, z)), \quad (65)$$

where $K_n(x, z)$ satisfies

$$K_n(x, z) = x\tilde{f} (K_{n-1}(x, z)), \quad (66)$$

and $K_0(x, z) = xz$.

We observe that if we modify the second argument of $\tilde{K}$ as follows

$$\tilde{K}_n(x, z, x) = \sum_{j, \ell=0}^{\infty} \text{Prob} (\tilde{X}_n = j \text{ and } \tilde{Z}_n = \ell) x^j \left(\frac{z}{x}\right)^\ell, \quad (67)$$

then we can write, analogous to Eq. (38),

$$\tilde{\tau}_n = \sum_{j, \ell=0}^{\infty} \text{Prob} (\tilde{X}_n = j \text{ and } \tilde{Z}_n = \ell) \frac{\ell}{j - \ell}, \quad (69)$$

$$= \int_0^1 \frac{1}{x} \left. \frac{\partial \tilde{J}_n}{\partial z} \right|_{z=1} dx, \quad (70)$$

where $\tilde{J}_n(x, z) = \tilde{K}_n(x, z)$. The iteration equation for $\tilde{J}_n(x, z)$ is obtained from Eq. (65) as

$$\tilde{J}_n(x, z) = x\tilde{f} (J_{n-1}(x, z)), \quad (71)$$

where

$$J_n(x, z) = K_n(x, z) = x\tilde{f} (J_{n-1}(x, z)), \quad (72)$$

and $J_0(x, z) = z$.

To evaluate the integral in Eq. (70), it is convenient to define the single-argument function

$$\tilde{I}_n(x) = \frac{1}{x} \left. \frac{\partial \tilde{J}_n}{\partial z} \right|_{z=1} = \frac{1}{x} \left. \frac{\partial \tilde{K}_n(x, z)}{\partial z} \right|_{z=1}. \quad (73)$$

Then we obtain from Eq. (71) that $\tilde{I}_n$ can be expressed as

$$\tilde{I}_n(x) = x\tilde{f} (J_{n-1}(x, 1)) I_{n-1}(x), \quad (74)$$
where \( L_n(x) \), defined as
\[
L_n(x) = \frac{1}{x} \left. \frac{\partial J_n}{\partial z} \right|_{z=1},
\]
obeys the iteration equation
\[
L_n(x) = x f'(J_{n-1}(x,1)) L_{n-1}(x),
\]
with \( L_0(x) = 1/x \). Iterating Eqs (74), (76) and (72) for values of \( x \) that partition the interval \([0, 1]\) enables us to calculate the integral
\[
\tilde{r}_n = \int_0^1 \tilde{L}_n(x) dx
\]
using the trapezoidal rule.

Thus, we have shown how a subcritical branching process model can give rise to an apparent novelty decay factor, even though the offspring distribution does not change from generation to generation. The ′apparent′ nature of the decay in the novelty factor does not reflect any change in the likelihood of retweeting by a user who receives the tweet; rather it is the mechanism needed in the multiplicative process model of Eq. (60) to deal with the finite lifetimes of cascades. At each generation of the branching process fewer trees remain alive, and so the growth rate of the total number of tweets must decline with \( n \), and in the multiplicative process this is mediated by the decay of the novelty factor. Figure 11 shows the decay of the novelty factor in the empirical data sets; this compares well with the theoretical result of Eq. (77).

5. Discussion

In Section 2, we demonstrated that two data sets from Twitter can be approximately described by branching processes, at least when we examine the discrete generation-by-generation structure. An examination of
the details of a continuous-time branching process that could produce these structures is left for further work. In Section 3, we argued that the observed offspring distributions were better fitted by a model based on the assumption that Twitter users have limited attention—so those who follow many others are less likely to notice and retweet any single message they receive—than by the more usual independent cascade model, with its assumption of equal transmission probability for each infection attempt.

Taking the fitted offspring distributions as inputs, in Section 4, we derived analytical and semi-analytical results using branching process theory. We began with well-established results (Eq. (26) and Eq. (32)) on the distribution of cascade lifetimes and of cascade sizes, and then extended the arguments used to derive results for other measures. Our novel results include the derivation of integral formulas for the expected average tree depth (equation (46)) and for the expected structural virality (equation (53)); we also showed that these provide a good match to the data. The integral formulas are also amenable to asymptotic analysis to understand the behaviour of the metrics as the branching number approaches the critical value. These results should assist in the fitting of transmission models to large-scale data sets, as was done (albeit using billions of numerical simulations rather than analytical methods) in Goel et al. [13]. Finally, we derived a novel formula that enables the calculation of the apparent novelty factor, as would be used in a multiplicative stochastic model for the cascades under study. In the branching process model, information does not decrease in its transmission likelihood over generations, but the fact that the processes are subcritical means that the number of users who receive a cascading tweet decreases over time (Fig. 3). In a multiplicative model, the stochastic lifetimes of cascade trees must be imposed through the assumption of novelty decay, and our results in Section 4.4 show how the two modelling approaches can be directly compared. We believe that the insights of the branching process approach will help inform applications of the multiplicative model, while the formula linking the offspring distribution to the apparent novelty decay (equation (77)) will allow the application of branching process theory to data sets that previously were studied only via the multiplicative model.

Our study has, of course, several limitations. The nature of cascades on Twitter is that they are rather short-lived, so our observation of a stable offspring distribution might not generalize to cascades on other social media where the attention given to topics is longer-lived, and hence where novelty decay might be more likely. We have implicitly assumed that all cascade topics are equally attractive to the Twitter users and so the identification of cascade-specific ‘fitnesses’ [38] has not been addressed here. As noted above, a study based on continuous-time branching processes could potentially extend our results to include age-dependent effects [7], but we expect that the results presented here would remain valid in the long-time limit where all cascades have reached their final state. In conclusion, we hope that the results and the methodology presented here will prove useful to researchers investigating those aspects of human behaviour that are mediated by online social networks.

Acknowledgements

We acknowledge the work of the authors of Refs[18, 19, 26] in gathering the initial data sets and making them available for study.

Funding

This work was supported by Science Foundation Ireland [grant numbers 16/IA/4470, 16/RC/3918, 12/RC/2289 P2, 18/CRT/6049] with co-funding from the European Regional Development Fund (J.G.), by the James S. McDonnell Foundation (P.F.) and by JSPS KAKENHI [grant number JP19K14618 (T.O.)].
REFERENCES