Characterizing Cascade Dynamics in A Microblogging System

Shengkai Shi*, Zhi Wang†, Chuan Wu*, and Xiaojun Lin‡
*The University of Hong Kong, †Tsinghua University, ‡Purdue University
Email: {skshi, cwu}@cs.hku.hk, wangzhi04@mails.tsinghua.edu.cn, linx@purdue.edu

Abstract—**Online microblogging sites have become increasingly important platforms for information diffusion in today's world, where users post short messages and follow various messages posted by people that they are interested in. It is intriguing to qualitatively study the temporal dynamics of an information cascade in a microblogging system, in terms of the number of users influenced at any given time, which may provide valuable input to facilitate emerging applications such as online advertising and content distribution. In this paper, we model information diffusion in a microblogging network as an *age-dependent branching process*, based on practical observations from Tencent Weibo, a popular microblogging site in China. This model enables careful characterization of the diffusion topology, the different delays for users to respond to new information, and the evolution of the size of the information cascade over time. We derive the expected cascade size at any time. We validate our model based on Tencent Weibo traces, and demonstrate its effectiveness in capturing information diffusion dynamics in the real world.**

I. Introduction

With their rapid proliferation in today's Internet, online social networks have remarkably revolutionized how individuals communicate and connect with each other. As a major type of online social networking services, online microblogging (*e.g.*, Twitter, Weibo) allows users to post short messages, including texts, images, and links to videos. Such a short message is generally referred to as a microblog. Followers of a microblog user read the microblog and may further repost it, resulting in cascading-style information diffusion [1]. By March 21 2013, the leading microblogging service, Twitter, had achieved in total 200 million active users who were creating more than 400 million tweets on a daily basis [2]. Online microblogging is gaining an increasingly important role in information dissemination in today's society [3].

A thorough understanding of the dynamics of information cascade in a typical microblogging system can provide valuable guidelines for operating emerging applications such as online advertising and content distribution. It is especially useful and intriguing to characterize the temporal evolution of cascade sizes and the influential factors underneath. A number of studies have been devoted to modeling information diffusion in online social networks [4]. The epidemic model has been a popular choice among these work. Leskovec *et al.* [1] investigate the blog link propagation using an *SIS* (*Susceptible-Infectious-Susceptible*) epidemic model. With an *SIS* model, the epidemic is assumed to persist in the system and ultimately infect everyone, which is inconsistent with the fact that most information cascades in an online social network tend to be extremely small [5]. Cheng *et al.* [5] propose an enhanced *SIIRP* (*Susceptible-Immune-Infectious-Recovered-Permanent*) model to accommodate diverse user behaviors in online social video sharing. However, a constant transition probability from one stage of user behavior to another stage is assumed, which may be insufficient for capturing the temporal dynamics of the cascade, and only descriptive analysis of the information diffusion process is provided. Another representative category of work investigates the influence maximization problem, *i.e.*, finding K nodes which will influence the most number of other nodes in the network, using influence models. Kempe *et al.* [6] base their study on two basic influence models, *Independent Cascade* (*IC*) Model and *Linear Threshold* (*LT*) Model. Yang *et al.* [7] propose a linear influence model, which uses an influence function at each node to quantify how many subsequent infections can be attributed to that node, as learned by regression methods. Most studies in this category focus on the final coverage of the influence cascade, but not the temporal dynamics of the diffusion process.

We seek a simple yet effective model based on the branching process, to analytically study the temporal dynamics of information cascades. One representative branching process model is the Galton-Watson process model. Li *et al.* [8] build a modified Galton-Watson model to capture the branching factor and share rate derived from measurement results on information diffusion in a social website. Wang *et al.* [9] extend the classic Galton-Watson branching model with a killing process to describe the process of information spreading in a microblogging network, where a detailed description of temporal dynamics is not provided. In addition, the standard Galton-Watson process is mainly determined by one random variable, the distribution of the number of offsprings, which is insufficient for capturing the temporal dynamics of information diffusion.

In this paper, we apply an *age-dependent branching process* [10] to characterize cascade dynamics in a microblogging system. We first carefully study a large volume of traces from Tencent Weibo [11], one of the largest microblogging websites in China. The empirical observations from our measurement study provide practical and useful guidance in modeling real-world microblog diffusion using the highly abstracted age-dependent branching process. To the best of our knowledge,

This work is supported in part by Hong Kong RGC under the contract HKU 717812E.

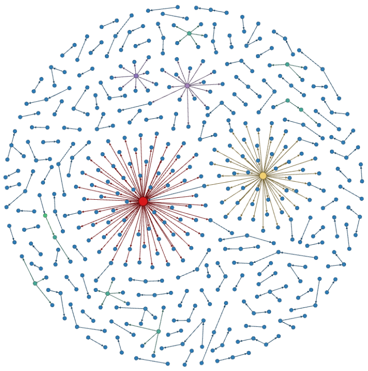


Fig. 1. Example diffusion cascades in Tencent Weibo.

we are among the first to apply such an age-dependent branching process model to carefully investigate the temporal dynamics of microblog propagation. Two key parameters, the distribution of the followers' repost delays and the distribution of the degrees in the diffusion tree, are chosen to characterize the diffusion process. We derive the expected cascade size at any time. We validate our model based on Tencent Weibo traces, and demonstrate its effectiveness in capturing diffusion dynamics in the real world.

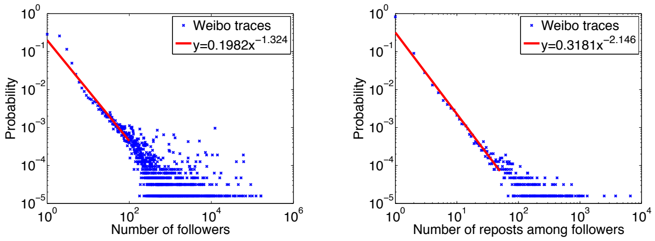
The rest of this paper is organized as follows. We present our measurement observations in Sec. II, model the diffusion process in Sec. III, evaluate the model using trace-driven experiments in Sec. IV, and conclude the paper in Sec. V.

II. Measurement of Information Diffusion in A Microblogging System

We collected large datasets from the technical team of Tencent Weibo, one of the largest microblogging websites in China [11]. Tencent Weibo was launched in April 2010, and had reportedly achieved 540 million registered users and more than 100 million active users on a daily basis by the end of 2012. On its platform, a user can *post* messages, images, and links to videos as microblogs. The followers of the user, *i.e.*, users who are socially connected to the user in the microblogging system, can *repost* a microblog, leading to an information diffusion cascade.

A. Dataset Description

Our datasets contain 20-day runtime traces of the system during October 9 to 29, 2011. Each entry in the traces corresponds to one microblog, including (i) the ID, name and IP address of the user who posted the microblog, time stamp when the microblog was posted, (ii) the ID of the parent user from which the microblog was received and the ID of root user who initiated the microblog, if it is a repost, as well as (iii) contents of the microblog. Our trace collection focused on microblogs containing links to videos shared in external video sharing websites, *e.g.*, a link to a video on YouKu [12]. In particular, we collected about 2 million microblogs containing links to over 350 thousand videos in the 20-day span. In addition, we also collected the profiles of users who posted these microblogs, which include the lists of their followers.



(a)

(b)

Fig. 2. Distribution of the number of followers of users, and the number of reposts to their microblogs.

B. Key Observations

1) Tree-like Microblog Propagation with No Repeated Reposting: Fig. 1 presents example diffusion cascades of different microblogs in Tencent Weibo. Each dot represents a user, and an arrow represents the diffusion from one user to a follower. We made the following observations: (1) Most of the diffusion cascades are trees (where a root node is the user who initially posts the microblog), and there is rarely a circle in the diffusion graph, indicating that it is very unlikely for a user to repost the same microblog twice; (2) A dominating fraction of the diffusion trees are very small, *e.g.*, there is a large fraction of 2-node trees.

2) Power-law Distributions of the Number of Followers and the Number of Reposts: Fig. 2(a) illustrates the distribution of the number of followers of $63,546$ users, randomly selected in our traces. Each sample represents the percentage of users (indicated by the y value) with the same number of followers (denoted by the x value). We see that this distribution is highly skewed. The number of users with very large numbers of followers is small. The distribution can be fitted by a power-law distribution $y = 0.1982x^{-1.324}$.

Fig. 2(b) plots the distribution of the number of reposts by the followers of the $63,546$ users, summarized from reposts of all the microblogs posted by those users. Each sample represents the percentage of microblogs (y value) with the same number of reposts (x value). Similarly, we observe that this distribution can be fitted by a power-law distribution $y = 0.3181x^{-2.146}$.

3) Evolution of Cascade Size: We plot the evolution of three microblogs, carrying links to three representative types of videos, in Fig. 3. Each curve represents the cumulative number of reposts of one microblog over time. The numbers of reposts to all the three microblogs stop increasing at certain time points, indicating that a cascade in a microblogging system typically has a limited duration. We observe that the three curves share similar growth patterns, that after a fast increase in the first several hours, the cascade size remains at a stable level. This observation indicates that in representative microblog diffusion, information spread is most effective at the early stage.

4) Gamma Distribution of the Response Delays: A follower of a user may not repost a microblog immediately after the user has posted/reposted the microblog, as the follower may not be online at the time and will only find out the post

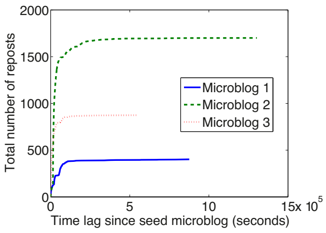


Fig. 3. The total number of reposts versus the time lag since when the seed microblogs are posted.

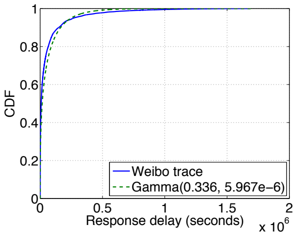


Fig. 4. CDF of response delays of all reposts in our traces.

later. We define the period from when a user posts/reposts a microblog to the time when a follower of the user reposts the microblog, as the *response delay* of the follower for this specific microblog. Such a response delay is a key factor to decide the microblog diffusion process. Fig. 4 illustrates the cumulative distribution function (CDF) of the response delays for all microblogs in our traces. We observe that this distribution can be well fitted using the CDF of a Gamma distribution $Gamma(k, \theta)$, where $k = 0.336$ is the shape parameter and $\theta = 5.967 \times 10^{-6}$ is the scale parameter.

III. An Age-dependent Branching Process Model

Fig. 1 illustrates the tree-like cascades of microblog diffusion, which enables us to apply a branching process to describe each cascade. We aim to characterize the detailed temporal evolution of the cascade size of a microblog, posted by its source user at time 0. In particular, we seek to answer the following question: How many users in total are expected to have reposted the microblog after a certain time t?

A. Mapping a Microblog Diffusion Cascade to an Age-Dependent Branching Process

A branching process models a population in which each individual gives birth to a random number of offsprings independently according to a certain probability distribution. An age-dependent branching process is a more general type of branching process, where the lifetimes of individuals are considered based on a lifetime distribution [10]. In an age-dependent branching process, the *seed node* born at time 0 remains active for a random lifetime according to a probability distribution; at the end of its life, the seed node produces a random number of offspring nodes following a probability distribution, and turns inactive. Similarly, each offspring node keeps active for a certain period of time, and then generates more offsprings and turns inactive. The process results in a branching tree of nodes, as illustrated in Fig. 5(2), where node 1 produces 3 offsprings at time d_1, and each of the latter produces one offspring after time d_2, d_3 and d_4, respectively. The blue and red nodes in the branching tree represent inactive and active nodes, respectively. In such a standard age-dependent branching process, all the offsprings of a node are born at the same time. In the case of microblog diffusion, followers of a user may well repost a microblog at different times after the parent user has posted/reposted the microblog. An example microblog diffusion process is given in Fig. 5(1): seed user 0 posts a microblog at time 0, user 1 reposts it at time d_1, and then three followers of user 1 repost the microblog further after d_2, d_3 and d_4 time, respectively. Next, we map the microblog diffusion process to the standard age-dependent branching process, in order to study the temporal dynamics of the microblog diffusion cascade by analyzing the evolution of the corresponding branching tree.

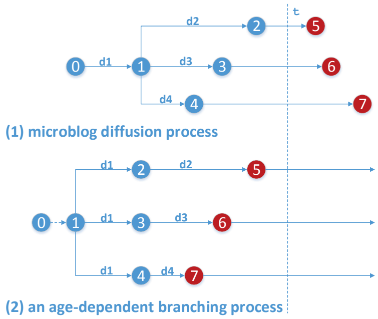

t
0
d1
1
d2
2
5
d3
3
6
d4
4
7
(1) microblog diffusion process
0
1
d1
2
d2
5
d1
3
d3
6
d1
4
d4
7
(2) an age-dependent branching process

Fig. 5. Mapping between the microblog diffusion cascade and an age-dependent branching tree.

The example microblog diffusion cascade in Fig. 5(1) is mapped to the age-dependent branching process in Fig. 5(2) as follows: each repost is mapped to a node in the branching tree which is born at the time when the parent post/repost in the microblog cascade occurs, *e.g.*, repost 1 is mapped to tree node 1 which is born at time 0, reposts 2, 3, 4 correspond to tree nodes 2, 3, 4 which occur at the same time d_1, and the offspring repost 5 of repost 2 is mapped to the offspring node 5 of tree node 2 born further after time d_2, and so on. The seed user 0 in the microblog cascade is mapped to the seed node 0 in the diffusion tree. Consider time t indicated by the slash line in Fig. 5. Based on our mapping, the total number of reposts in the diffusion cascade by time t equals the number of inactive nodes in the branching tree at t (excluding node 0). This mapping enables us to study the size of the microblog diffusion cascade at any time based on the total number of inactive nodes in the corresponding branching tree.

In addition, Fig. 5 shows the case that there is only one repost (“1”) directly from the seed post (“0”). In the case that multiple followers of the seed user repost the microblog, the microblog cascade can be mapped to multiple stochastically identical and independent branching trees, each initiated by the seed node 0 in the branching process. The next subsection will first focus on studying the expected size of one of the branching trees (*e.g.*, the one rooted at node ‘1’ in Fig. 5), and then extend the result to the case of multiple trees.

B. The Temporal Dynamics of Microblog Diffusion Cascade

Let $X(t)$, $Y(t)$ and $Z(t)$ denote the number of inactive nodes, the total number of nodes and the number of active nodes in a branching tree at time t, respectively. We use lower-case notation $x(t)$, $y(t)$ and $z(t)$ to represent the expectations of random variables $X(t)$, $Y(t)$ and $Z(t)$, respectively. Here $x(t)$ corresponds to the expected size at time t of a diffusion cascade starting from one of the direct reposts of the seed post, based on our mapping in Sec. III-A.

TABLE I
Important Notation

$X(t)$	the number of inactive nodes in a branching tree at t
$Y(t)$	the total number of nodes in a branching tree at t
$Z(t)$	the number of active nodes in a branching tree at t
$\tilde{x}(t)$	the expected total size of a microblog cascade at t
$x(t)$	the expected number of inactive nodes in a branching tree at t / the expected size of a diffusion sub-cascade at t
$y(t)$	the expected total number of nodes in a branching tree at t
$z(t)$	the expected number of active nodes in a branching tree at t
p_k	distribution of the number of offspring nodes at each node
μ	the reproductive number of nodes in a branching tree
$G(\tau)$	the cumulative distribution function of the lifetimes of nodes in a branching tree / the cumulative distribution function of response delays in a microblog cascade
$F(s,t)$	probability generating function of $Z(t)$
$U_\mu(t)$	the renewal function
$E(\omega)$	the Laplace transformation of $G(\tau)$
$L(\omega)$	the Laplace transformation of $x(t)$

Let $P(R=k)=p_k$ denote the probability density function of the number of offsprings of a node in the branching tree, *i.e.*, the degree distribution, where R is the random variable of the number of offsprings of a node. It corresponds to the probability distribution of the number of reposts following a previous repost in the microblog diffusion cascade. Let $\mu = \sum_{k=0}^{\infty} p_k k$ denote the expected degree of each node, referred to as a *reproductive number* of a node in the branching process. Let $G(\tau)$ be the cumulative distribution function (CDF) of the lifetimes of nodes in a branching process, corresponding to the CDF of response delays in the microblog diffusion process (*e.g.*, that plotted in Fig. 4). The two distribution functions are key parameters deciding the temporal dynamics of the diffusion process. Our following analysis focuses on the derivation of $x(t)$, based on $x(t)=y(t)-z(t)$. Table I summarizes the important notation in this paper.

We first seek to derive the expected number of active nodes in the branching tree, $z(t)$, based on the two distribution functions. We construct the probability generating function (PGF) of $Z(t)$, as $F(s,t)=\sum_{k=0}^{\infty} P(Z(t)=k)s^k$ [10]. We will make use of an important property of PGF of a random variable, that the expectation of random variable $Z(t)$, $z(t)$, is a bounded limit of $\frac{\partial F(s,t)}{\partial s}$ as $s \to 1$ [10]. Especially, we will derive the expression for $F(s,t)$, in order to derive $z(t)$ based on this property.

We start by deriving the probability density function $P(Z(t)=k)$. We consider two cases according to the lifetime of the seed node. (1) Case 1: the seed node is alive at time t, so no offspring has been born by t, which happens with probability $1-G(t)$ (since $G(t)$ is the probability that a node becomes inactive before t). Given that there is only one node at time t, we have $P(Z(t)=k)=[1-G(t)]\delta_{1k}$, where δ_{1k} is 1 if $k=1$ and 0 otherwise. (2) Case 2: the seed node becomes inactive at some time $\tau < t$, with probability $dG(\tau)$, and it produces j successors at time τ with probability p_j. In the remaining time $t-\tau$, these j successors give birth to a total of k offsprings. As the probability for one node to produce k offsprings in time $t-\tau$ is $P(Z(t-\tau)=k)$, the probability for j successors to do so is $P^{*j}(Z(t-\tau)=k)$, where P^{*j} is the j-fold convolution of probability density function $P(Z(t-\tau)=k)$. We can thus derive

$$\begin{aligned} P(Z(t)=k) =& [1-G(t)]\delta_{1k} \\ &+ \int_0^t dG(\tau)\sum_{j=0}^{\infty} p_j P^{*j}(Z(t-\tau)=k). \end{aligned}$$

Hence the PGF of $Z(t)$ can be derived as:

$$\begin{aligned} F(s,t) =& \sum_{k=0}^{\infty} P(Z(t)=k)s^k \\ =& [1-G(t)]\sum_{k=0}^{\infty} s^k \delta_{1k} \\ &+ \int_0^t dG(\tau)\sum_{j=0}^{\infty} p_j \sum_{k=0}^{\infty} P^{*j}(Z(t-\tau)=k)s^k. \end{aligned}$$

We note that $\sum_{k=0}^{\infty} P^{*j}(Z(t-\tau)=k)s^k = F^j(s,t-\tau)$, where $F^j(s,t-\tau)$ stands for the jth power of the PGF of $Z(t-\tau)$, and $\sum_{k=0}^{\infty} s^k\delta_{1k} = s$. We can thus derive

$$F(s,t) = s[1-G(t)] + \int_0^t h[F(s,t-\tau)]dG(\tau), \quad (1)$$

where $h[F(s,t-\tau)] = \sum_{j=0}^{\infty} p_j F^j(s,t-\tau)$. Since $F(s,t)$ is a convergent power series for any $0 < s < 1$, we can differentiate both sides of (1) over s and derive

$$\frac{\partial F(s,t)}{\partial s} = [1-G(t)] + \int_0^t h^{'}[F(s,t-\tau)]\frac{\partial F(s,t-\tau)}{\partial s}dG(\tau). \quad (2)$$

Because $0 < s < 1$, $F(s,t-\tau) < 1$. We have $h^{'}[F(s,t-\tau)] < h^{'}[1] = \mu$. Since $z(t)$ is a bounded limit of $\frac{\partial F(s,t)}{\partial s}$ as $s \to 1$, by taking limit $s \to 1$ in both sides of (2), we have

$$z(t) = [1-G(t)] + \mu\int_0^t z(t-\tau)dG(\tau). \quad (3)$$

We can derive the expected total number of nodes in the branching tree, $y(t)$, by deriving the probability generating function of $Y(t)$, using very similar steps as how we have derived $z(t)$. We omit the steps due to space constraint, but directly give the result:

$$y(t) = 1 + \mu\int_0^t y(t-\tau)dG(\tau). \quad (4)$$

Define a renewal function $U_\mu(t) = \sum_{n=0}^{\infty} \mu^n G^{*n}(t)$, where $G^{*n}(t)$ is the n-fold convolution of $G(t)$. Eqn.s (3) and (4) have the form of a renewal equation, for which the renewal theory provides solutions [10]. For example, if $\gamma H(0+) < 1$, an equation of the form

$$S(t) = \xi(t) + \gamma\int_0^t S(t-\tau)dH(\tau) \quad (5)$$

has a unique solution $S(t) = \xi(t) * U_\gamma(t)$ which is bounded on any finite interval of t, where $U_\gamma(t) = \sum_{n=0}^{\infty} \gamma^n H^{*n}(t)$ [10]. We can thus derive the expressions of $z(t)$ and $y(t)$ as follows:

$$z(t) = [1-G(t)] * U_\mu(t), \quad (6)$$

$$y(t) = U_\mu(t). \tag{7}$$

Based on (6) and (7), we finally derive

$$x(t) = y(t) - z(t) = G(t) * U_\mu(t). \tag{8}$$

We can see that $x(t)$ is decided by μ, the reproductive number in the branching process (*i.e.*, the expected degree of nodes in the microblog diffusion cascade), and $G(\tau)$, the distribution of lifetimes of nodes (*i.e.*, the distribution of response delays in the diffusion cascade).

Since $U_\mu(t)$ contains an infinite series of functions, it is difficult to calculate $x(t)$ based on the definition of convolution directly. We further seek to compute an explicit expression of $x(t)$ using Laplace transformation. Let $E(\omega)$ denote the Laplace transformation of distribution $G(\tau)$. Define

$$V_\mu(t) = U_\mu(t) - 1 = \sum_{n=1}^{\infty} \mu^n G^{*n}(t). \tag{9}$$

Based on the basic properties of Laplace transformation [13], we can derive that the Laplace transformation of $V_\mu(t)$ is $\frac{\mu E(\omega)}{1-\mu\omega E(\omega)}$. Since $G(t)$ is a CDF, we have $x(t) = G(t) + G(t) * V_\mu(t) = G(t) + V_\mu(t) * G(t) = G(t) + \int_0^t V_\mu(t-\psi)dG(\psi) = G(t) + \int_0^t V_\mu(t-\psi)G^{'}(\psi)d\psi$. $G^{'}(t)$ is the corresponding probability density function of $G(t)$. We can obtain that the Laplace transformation of $G^{'}(t)$ is $\omega E(\omega)$. Hence, the Laplace transformation of $x(t)$ is:

$$L(\omega) = E(\omega)/(1 - \mu\omega E(\omega)). \tag{10}$$

When the concrete forms of the distributions p_k and $G(\tau)$ are given, we can compute μ (the expectation of distribution p_k) and $E(\omega)$ (the Laplace transformation of distribution $G(\tau)$). Then we can derive $L(\omega)$ (the Laplace transformation of $x(t)$) according to (10). Finally, we are able to compute the explicit form of $x(t)$, through inverse Laplace transform.

Recall that $x(t)$ is the expected size of one of the branching trees originated from seed node '0', *i.e.*, the expected size of a cascade starting from one of the direct reposts following the seed post in the microblog diffusion. Since the expected number of direct reposts from the seed post is μ, we can derive that the overall size of a microblog cascade is $\tilde{x}(t) = \sum_{k=0}^{\infty} p_k k x(t) + 1 = \mu x(t) + 1$, where 1 corresponds to the seed post.

C. The Case of Time-varying Degree Distribution

In a microblogging system, the popularity of a microblog may change over time, *i.e.*, more users tend to repost the microblog at the early stage after it is initially posted, and less reposts may happen when it has been around for a while. Our model in Sec. III-B can capture such a scenario with a time-varying degree distribution p_k, *i.e.*, the expected repost degree at a node, μ, can be larger when a microblog is newly posted, and its value decreases over time.

Consider the staged variation of the degree distribution p_k: there exists a time sequence $0, T_1, T_2, \ldots$; the degree distribution remains fixed within each interval (*e.g.* $[0, T_1]$, $(T_1, T_2]$) but can vary from one interval to the next. We calculate the cascade size as follows. The cascade size by time T_1, *i.e.*, $\tilde{x}(T_1)$, can be derived using the method in Sec. III-B. There are $z(T_1)$ active nodes in the branching tree at T_1; treating each active node as one new root node, we can further estimate the size of the cascade tree rooted at each active node beyond T_1, *i.e.*, $\tilde{x}(t - T_1)$, using the method in Sec. III-B, and then the size of the entire cascade can be computed by $\tilde{x}(T_1) + z(T_1)\tilde{x}(t - T_1)$, for $t \in (T_1, T_2]$. The similar method can be applied to derive the cascade size when $t > T_2$.

When the expected degree μ of the degree distribution decreases to a very small value (≈ 0), the number of active nodes approaches zero and the size of the cascade essentially stops growing. The final cascade size can be derived using the method described above.

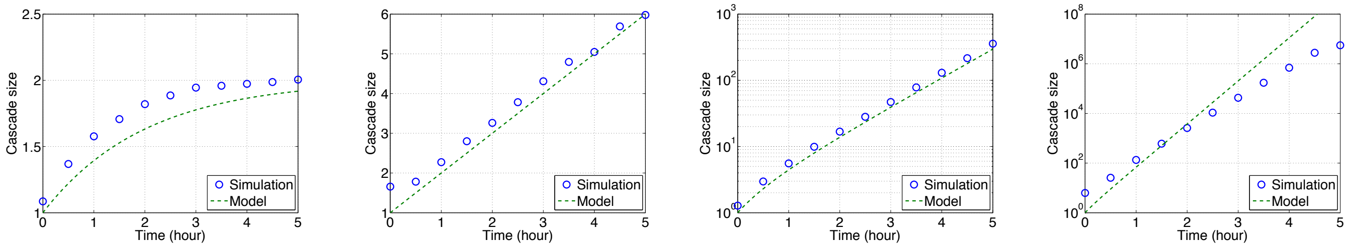
IV. Trace-driven Performance Evaluation

A. Experiment Setup

We simulate a microblogging network, where the number of followers of a user in the network follows the power-law distribution in Fig. 2(a). We simulate a microblog diffusion process as follows. At time slot 0, we randomly pick one node in the microblogging network as the seed user who posts a microblog. A number of followers of this seed user repost the microblog after a random response delay where the number of followers is picked according to the offspring distribution p_k and the response delays follow the distribution $G(\tau)$. The process repeats from the repost users: after a user has reposted the microblog, a particular number of its followers (who have not reposted the microblog) are randomly selected to repost the microblog after certain response delays. Both an exponential distribution and a Gamma distribution are used in our experiments for the response delays. We compare the simulation results with the calculated expected cascade size $\tilde{x}(t)$ from Sec. III-B and Sec. III-C, as well as with statistics from the traces, wherever applicable. For every set of parameters, the simulation is run for 10^4 times in order to obtain converged statistics.

B. Experiment Results

1) Evolution of Cascade Size: We first study the effectiveness of our model in capturing the evolution of cascade size over time, by comparing the cascade sizes computed using our model with those from the simulations. In this set of experiments, we use exponentially distributed response delays following $G(\tau) = 1 - e^{-a\tau}$ where $a = 1$ such that the average response delay is 1 hour. We will examine the case of the Gamma distribution from the traces in the next subsection. We use the power-law distribution in Fig. 2(b) as the degree distribution p_k. Fig. 6 compares the cascade sizes under different values of the expected degree μ. When $\mu > 1$, the diffusion cascade grows exponentially; when $\mu = 1$, the increase of the cascade size is approximately linear; when $\mu < 1$, the increase of the cascade size is slow and the total size becomes stable soon (*i.e.*, the cascade stops growing soon). The simulation results that we show are the average of multiple



(a) $\mu = 0.5$. (b) $\mu = 1$. (c) $\mu = 2$. (d) $\mu = 5$.

Fig. 6. Comparison of the evolution of cascade sizes generated by simulations and our model.

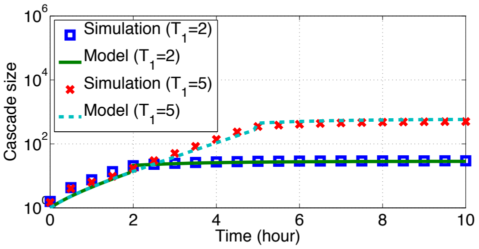
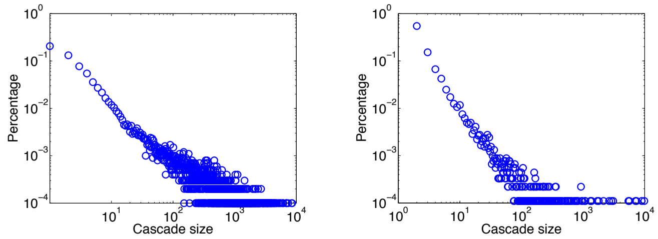


Fig. 7. Comparison of the evolution of cascade sizes generated by simulations and our model: two-stage μ.



(a) Simulation (b) Tencent Weibo

Fig. 8. Distribution of final cascade sizes.

runs, which explains the fraction numbers in Fig. 6(a). In all cases, the computed sizes fit well with the simulation results.

We plot in Fig. 7 the cascade size over time with μ changing in two stages, *i.e.*, a large $\mu = 2$ in the early stage (*i.e.*, $t \leq T_1$) and a small $\mu = 0.5$ in the later stage (*i.e.*, $t > T_1$). We observe that the computed results based on our model in Sec. III-C also fit well with the simulation results.

2) Final Cascade Size: Using the Gamma distribution in Fig. 4 for $G(\tau)$ and a time-decaying μ according to Sec. III-C, we further evaluate the final cascade size generated by simulations according to our model. We plot in Fig. 8(a) the distribution of final cascade sizes from our simulations where the value of μ decreases exponentially from one hour to the next following $\mu(t) = \mu_0 e^{-0.2t}$ with an initial $\mu_0 = 2$ (its value remains fixed within each hour). The rationale is that in our measurement study in Sec. II-B3, we have observed exponential decreases of repost rates over time for representative microblogs. Each sample in Fig. 8 (a) is the percentage of cascades with the same final cascade size, generated in our simulations. We observe that the final cascade size follows a Zipf-like distribution. Using Weibo traces, we also plot in Fig. 8(b) the distribution of final cascade sizes of microblogs in Tencent Weibo. We observe that in the traces, the final cascade sizes also closely follow a similar Zipf-like distribution. This further validates that using a repost degree distribution that varies in stages, our model is able to closely capture the evolution of cascade sizes in real-world traces.

V. Concluding Remarks

Effective information diffusion modeling is critical for a large variety of social applications in today's world. This paper presents our first step towards a qualitative understanding of the information diffusion process in a microblogging system. In particular, we reveal several facts on microblog propagation based on a large-scale measurement study, which motivates our adoption of an age-dependent branching process to investigate the temporal dynamics of cascade sizes in a microblogging network. We give detailed mathematical derivation of the expected cascade size at any time during a microblog diffusion process. We evaluate our model using trace-based simulation experiments and demonstrate its effectiveness. In our model, we have used time-invariant probability distribution of the number of reposts p_k and the CDF of response delays in the microblog diffusion process $G(\tau)$. We leave the study of time-varying p_k and $G(\tau)$ for future work.

References

[1] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading Behavior in Large Blog Graphs," *arXiv preprint arXiv:0704.2803*, 2007.

[2] "https://blog.twitter.com/company."

[3] "http://www.guardian.co.uk/news/datablog/2013/jul/04/brazilian-protesters-twitter-microsoft."

[4] J. Leskovec, "Tutorial: Analytics and Predictive Models for Social Media," in *Proc. of International Conference on World Wide Web (WWW)*, 2011.

[5] X. Cheng, H. Li, and J. Liu, "Video Sharing Propagation in Social Networks: Measurement, Modeling, and Analysis," in *Proc. of IEEE INFOCOM*, 2013.

[6] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the Spread of Influence Through a Social Network," in *Proc. of ACM SIGKDD*, 2003.

[7] J. Yang and J. Leskovec, "Modeling Information Diffusion in Implicit Networks," in *Proc. of IEEE ICDM*, 2010.

[8] H. Li, J. Liu, K. Xu, and S. Wen, "Understanding Video Propagation in Online Social Networks," in *Proc. of IEEE International Workshop on Quality of Service (IWQoS)*, 2012.

[9] W. Dong, P. Hosung, X. Gaogang, M. Sue, M.-A. Kaafar, and K. Salamatian, "A Genealogy of Information Spreading on Microblogs: A Galton-Watson-based Explicative Model," in *Proc. of IEEE INFOCOM*, 2013.

[10] K. B. Athreya and P. E. Ney, *Branching Processes*. Dover Publications, 2004.

[11] "http://t.qq.com/."

[12] "http://www.youku.com."

[13] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and Systems*. Prentice-Hall International, 1997.