A Genealogy of Information Spreading on Microblogs: a Galton-Watson-based Explicative Model

Dong Wang* † §, Hosung Park‡, Gaogang Xie*, Sue Moon‡, Mohamed-Ali Kaafar§, Kave Salamatian¶
*Institute of Computing Technology, Chinese Academy of Sciences, China
†Graduate School of Chinese Academy of Sciences,China
‡Department of Computer Science, KAIST, Korea
§INRIA, France
¶LISTIC Lab, Universite de Savoie, France
Email: wangdong01@ict.ac.cn

***Abstract*—In this paper, we study the process of information diffusion in a microblog service developing Galton-Watson with Killing (GWK) model. Microblog services offer a unique approach to online information sharing allowing microblog users to forward messages to others. We describe an information propagation as a discrete GWK process based on Galton-Watson model which models the evolution of family names. Our model explains the interaction between the topology of the social graph and the intrinsic interest of the message.**

We validate our model on dataset collected from Sina Weibo and Twitter microblog. Sina Weibo is a Chinese microblog web service which reached over 100 million users as for January 2011. Our Sina Weibo dataset contains over 261 thousand tweets which have retweets and 2 million retweets from 500 thousand users. Twitter dataset contains over 1.1 million tweets which have retweets and 3.3 million retweets from 4.3 million users. The results of the validation show that our proposed GWK model fits the information diffusion of microblog service very well in terms of the number of message receivers. We show that our model can be used in generating tweets load and also analyze the relationships between parameters of our model and popularity of the diffused information. To the best of our knowledge, this paper is the first to give a systemic and comprehensive analysis for the information diffusion on microblog services, to be used in tweets-like load generators while still guaranteeing popularity distribution characteristics.

I. Introduction

Microblog services offer a unique approach to online information sharing that differs from online newsgroups, bulletin boards, or social networking services. The size of a posting cannot grow indefinitely and is limited (to 140 characters in case of Sina Weibo and Twitter). The relationship between users typically requires no reciprocal approval and is often considered as a form of subscription. Although postings can be configured to be private, most users keep theirs public, as well as their follower information. Since its launch in 2006, Twitter has now grown to a global service with hundreds of millions of users where news crowdsourcing is making historical changes as we have witnessed in recent MENA (Middle East North Africa) events. As the favorable convenience and interaction, the recording of information sharing and spreading has reached an unprecedented level with today's microblog services.

Word-of-mouth spreading is an integral mechanism of human information sharing and has long been studied in many disciplines of sciences. Information diffusion on online social media bears much relevance and similarity to offline word-of-mouth spreading, but no past mechanism can beat it in terms of speed and efficiency. As the tendency to share information online will only accelerate, the study of online information diffusion has direct implications on opinion mining, viral marketing, and political campaigns, just to name a few.

Analysis of online network topologies and information spreading patterns has laid foundation for explicative models of information diffusion. In this work we build an explicative model that takes the network topology of actual information diffusion and characteristics of contents into consideration and describes the process of diffusion comprehensively. We take an analogy between the family name evolution and diffusion by retweeting where a family name is carried on only by male descents with offspring and information on mircoblogs spreads only by those who choose to retweet it. The Galton-Watson process is a branching stochastic process that has been used for the evolution and extinction of family names, therefore we employ the Galton-Watson (GW) process to our modeling of information diffusion on microblog services. However, since information diffusion stops rather quickly online because the novelty of online news wears out with time, while family names die out much slowly, we include a killing process in our GW model to take into account such peculiar feature of the online information diffusion. We collect data from Sina Weibo and Twitter in order to use them in our analysis and evaluation. The results of our experiments demonstrate that the Galton-Watson process with Killing can describe the pattern of information diffusion on microblogs very well and can be efficiently used to generate synthetic loads of microblog online information while still guaranteeing the statistical characteristics in terms of tweets popularity. What's more, our GWK model and its parameters reveal the key features of popular tweets.

978-1-4673-5946-7/13/$31.00 ©2013 IEEE

The structure of this paper is as follows. Section II covers a survey of the related work. Section III describes our datasets. Sections IV and V introduce the the Galton-Watson model with the killing process and evaluate with our datasets. Finally, in Section VII we conclude.

II. Related Work

In this section we review prior work on the online social networks (OSNs) and social media; online information diffusion and its analytical models. Most previous work to model information diffusion have considered Independent Information Cascades and Linear Threshold models as building blocks and estimated the properties of the obtained cascades. Different from all the previous work, our Galton-Watson model with a killing process introduced in this paper, takes both the topology of microblog social graph and the intrinsic interest of the message into consideration and therefore can describe the online information diffusion more comprehensively and in an accurate way. This is supported by our validation and comparison between empirical tweets distributions and the synthetic model-based information patterns. Specially, as opposed to previous work, in addition to modeling the diffusion and popularity of online information, we also present an asymptotic analysis of the proposed process, which in turn allows us to not only validate the model to fit the actual tweets propagation, but also to use it for tweet load synthesis. Besides, this paper is to the best of our knowledge among the first to take an analogy between the family names evolution and the diffusion in microblog services, and as such to adapt the Galton-Watson model to information diffusion.

A. Online social networks and social media

From citation networks to call graphs and group dynamics in newsgroups, human dynamics in a great many forms of interaction has long been studied. The following two have analyzed the topological characteristics of online social networks and online social media which are of particular relevance to this work. Mislove *et al.* analyze four popular online social networks including Flickr, Livejournal, Orkut, and Youtube and find some basic features about OSNs such as a small world phenomenon and high clustering coefficients [20]. Kwak *et al.* report on news-media-like characteristics of Twitter [16].

B. Online information diffusion

These online social services offer a massively amount of data on human interaction and have spurred research on information sharing. Generally speaking, there are two directions for the online information diffusion researches, characteristics descriptions and analytical models.

Cha *et al.* provide an in-depth study of YouTube, including an analysis of popularity evolution [4]. Guo *et al.* analyze the popularity of various user-generated contents (UGC) and find that the observed rank-ordered popularity distribution is not power-law as expected but is a stretched exponential distribution [12]. Lee *et al.* use a Cox proportional hazard regression model to predict the popularity of online contents [17]. Zaman *et al.* give a probabilistic collaborative filtering model for predicting the popularity of information in Twitter and find that the most important features for information propagation in Twitter are the identity of the source of tweet and retweeter [23]. Lerman and Gosh conduct an empirical description of news spread process on Digg and Twitter [18]. Ye *et al.* show how breaking news spread through Twitter and provide metrics for social influence of users [29].Goetz *et al.* use "zero-crossing" approach to research the temporal dynamics of the blogosphere [8]. Gomez-Rodriguez *et al.* develop an efficient approximation algorithm to infer the information diffusion network [9]. Work listed above can be construed as of descriptive nature and do not answer causality of the phenomena.

Other work focused on building analytical models of popularity and diffusion. Two models are widely used for online information diffusion researches, Independent Cascades(IC) and Linear Threshold (LT) models. Kempe *et al.* introduce in [14] the LT model to find the influential users and then maximize the information spread on online social networks. Galuba *et al.* analyze the diffusion of URLs in Twitter and propose to use the LT model to predict which users will propagate which URLs [27]. Yang *et al.* develop in [28] a Linear Influence Model (LIM) based on LT models which can predict interactions between nodes in the information dissemination process without requiring the knowledge of the social network graph. On the other hand, IC models are firstly used to analyze the information spread on blogosphere [19], [7]. And epidemic model, which is a variation of an IC model is also proposed to make the microscopic characterization of information diffusion process [1], [10]. Cha *et al.* introduce the cascade model into the research of information dissemination on online social network such as Flickr [5], [6]. Myers *et al.* improve the traditional cascade model in which information can reach a node not only via the links of the social network but also through the influence of external sources [21]. Guille *et al.* develop a concrete model which relies on the IC model and is based on machine learning techniques to predict the temporal dynamics of diffusion in social networks [11]. Similarly, [24] proposes a K-tree based model, established to correct for missing data in information cascades, which makes such a model suitable for all types of cascades.

III. Dataset Description

In this section, we give a brief overview of Sina Weibo and Twitter datasets including the collection methodology and their basic properties.

A. Sina Weibo and Twitter

Twitter is a well-known microblog service, where a user can unidirectionally follow other users and subscribe to their tweets. Sina Weibo (weibo in Chinese means microblog) is a Chinese microblog service with Twitter-like unidirectional follow relationships. Followers on both services can retweet some of the messages they receive from their followings and these retweets are seen by their own followers. The distance

TABLE I
SINA AND TWITTER DATASET SUMMARY

Dataset	Time	Tweets forwarded	Retweets	Users
S_{user}	8 $\sim$ 12/2011	261,833	1,996,170	500,000
T_{user}	8/2011	1,133,568	3,316,609	4,332,445

between a retweeter and the tweet's original publisher is measured in *hops* where the publisher is considered at the 0^{th} hop. Retweeting is an easy and popular mechanism to share a tweet with followers.

Sina Weibo makes public two statistics per tweet: the number of retweets and the number of comments for the tweet, both of which represent the popularity of the tweet. However, Twitter does not offer comments to a tweet and thus we only use the number of retweets in this work. In latter sections, we measure the popularity of a tweet by the number of retweets.

B. Data collection methodology

For our study, we need the complete set of retweets per tweet. Both Twitter and Sina Weibo provide a search API, to which we input a tweet's identification number (ID) and are returned all retweets of the tweet where the retweeting pathes from the original publisher to retweeters are provided in detail. While for Sina Weibo we use this API, we use the methodology from Kwak *et al.* [15] in Twitter to collect followers, followings, tweets, and retweets of *all* Korean Twitter users. We refer to this Korean Twitter dataset as T_{user}.

There are over 100 million users on Sina Weibo as of January 2011, of which size is too big for us to manage without Sina Weibo company's cooperation. Instead we use sampling to reduce the size of the dataset for our work. In Sina Weibo, each user has a 10-digit user ID, whose first digit is 1 or 2. We generate uniformly random numbers from 1×10^9 to $(3 \times 10^9 - 1)$ and use them as user IDs to sample nodes from Sina Weibo. With this uniform sampling method we obtain an unbiased sample of $500,000$ users [13].

If a user with the uniformly sampled ID actually exists in Sina Weibo and hence is successfully located, we collect all his tweets which have retweets and are published from Aug. 1st, 2011 to Dec. 1st, 2011. For each tweet we have sampled, we then collect all its retweets. We refer to this user unbiased dataset from Sina Weibo as S_{user}.

C. Data description

Table I summarizes our datasets S_{user} and T_{user} characteristics.

Fig. 1 plots the CCDF (Complimentary Cumulative Distribution Function) of the users' followers from S_{user} and T_{user}. Neither has a simple power-law distribution. Today's online social networking services often have hundreds of millions of users and are used not only for personal communication but in numerous types of communication, including political campaigns and advertisements. The degree distributions from Cyworld, Twitter, and Orkut have been reported to deviate from a strict power-law distribution [2], [16], [20].

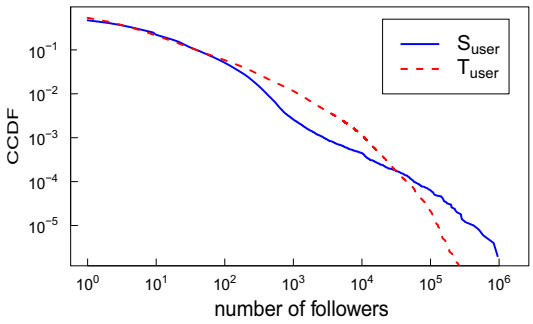

Fig. 1. CCDF of the number of followers

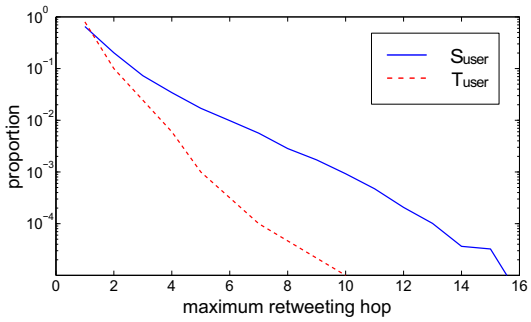

Fig. 2. Maximum retweet hops distribution in two datasets

For each original tweet (message) M, we build the corresponding retweeting tree $T(M)$ as follows. When node S publishes the original tweet M, S is considered as the root of $T(M)$ and all of S's followers which received the tweet are the children nodes of S in $T(M)$. For all children nodes, if node A retweets the tweet M from his parent, then A generates children nodes composed of all his followers, otherwise node A is considered as a leaf node.

Fig. 2 plots the distribution of the maximum retweeting hops in our two datasets, S_{user} and T_{user}. This distribution shows the depth of the diffusion and will be used in the forthcoming as a validation metric.

IV. A GALTON-WATSON MODEL

The Galton-Watson model has been used with success to model the evolution of family names [22]. The information diffusion process via retweeting online bears a striking similarity to the family name evolution. Family names are transferred patrilineally, while information spreads only by those who retweet. Family generations are analogous to retweeting hops, which indicates the distance between the source of the original tweet and the particular retweet.

One important factor in the above model is the decision to retweet or not. Such a decision depends mainly on the content of the tweet. In family dynamics, it would correspond to fertility. On the other hand, the distribution of number of followers (or descendants) is a topological property of the microblog social graph (or the family tree). Thus we take two

important aspects of information diffusion in the retweeting process: the intrinsic interest of the tweet message and the topology of the social graph.

Let us now formalize our model. A GW is a branching stochastic process $\{X_n\}$, where X_n represents the number of users on the microblog service that receive a particular tweet through a path of n retweeting hops. The process $\{X_n\}$ is then evolving according to the recurrence formula: $X_0 = 1$ and $X_{n+1} = \sum_{j=1}^{X_n} \xi_j$, where for each generation n, ξ_j is a sequence of Independent and Identically Distributed (IID) discrete random variables following a distribution $f(k)$ representing the number of offspring of a node. For the purpose of our tweet propagation modeling we assume that:

$$f(k) = (1-\alpha)\mathbb{1}_{\{k=0\}} + \alpha\mathcal{D}(k)\mathbb{1}_{\{k>0\}} \tag{1}$$

where α is the probability that a user receiving a tweet message can retweet it, and $\mathcal{D}(k)$ is the degree distribution of the microblog social graph, *i.e.* the distribution of number of followers for the microblog.

It is however important to notice that our tweets propagation modeling and the original GW family name process may differ from the process termination's perspective. Typically, a tweet has a shorter lifetime than a family name (in terms of hops) and will inevitably die faster, *i.e.* no more retweeting activity is observed. While a genealogical tree, depending on the distribution of the number of male offspring is more likely to last longer. This can be explained by the fact that online content might be more prone to a platitude effect due to lack of content novelty. In order to account for this peculiar property, we model an extinction process that will govern the original GW trees.

A. Analytic approach

First, we describe the GW trees without considering the extinction process. The mean evolution of a GW tree can be easily analyzed through the Wald equality that gives $\mathbb{E}\{X_{n+1}\} = \mu\mathbb{E}\{X_n\}$, where $\mu = \mathbb{E}\{\xi\}$ is the mean number of offspring of members, *i.e.* the mean number of people receiving a micromessage retweeted from one user directly in one retweeting process. In other terms, we have:

$$\mathbb{E}\{X_{n+1}\} = \mu^n \tag{2}$$

where $n \geq 0$.

Using the previous assumptions expressed in Eq. 1 we can rewrite μ as $\mu = \alpha\delta$ where α is the probability of retweeting a message and δ is the mean number of followers of microblog users that have a least one followers, resulting finally in:

$$\mathbb{E}\{X_{n+1}\} = (\alpha\delta)^n \tag{3}$$

where $n \geq 0$.

It should be noticed that for the retweeting process, X_1 doesn't meet Eq. 3. When the information source publishes a tweet, all followers of the publisher can receive the message at the first hop which means X_1 is equal to δ but not to $\alpha\delta$. In this case, Eq. 3 should be amended as follows to describe the retweeting process:

$$\mathbb{E}\{X_{n+1}\} = \delta(\alpha\delta)^{n-1} = (\alpha)^{n-1}(\delta)^n \tag{4}$$

where $n > 0$ and $X_0 = 1$.

Eq. 4 has the interesting property of separating two effects on the information spreading on microblog services: the intrinsic interest of the message represented by α and the properties of the social graph represented by δ. The mean total number of users receiving a tweet can be derived as $\bar{M} = \sum_{i=1}^{\infty} X_i$ and the mean total number of retweets is derived as $\bar{T} = \sum_{i=1}^{\infty} \alpha X_i$.

A more refined analysis of the evolution of GW trees can be done through the Probability Generating Function (PGF) that is defined for a discrete random variable X with pdf $p(k)$ as :

$$\phi(s) = \mathbb{E}\left\{s^X\right\} = \sum_{k=0}^{\infty} p(k)s^k \tag{5}$$

We define $\phi_n(s) = \mathbb{E}\left\{s_n^X\right\}$ as the PGF of X_n. In the context of the GW tree it is easy to prove that $\phi_{n+1}(s) = \phi_n(\phi(s))$. Deriving the precise value of the PGF in general is hard and closed form for it is available in a very limited number of cases. However the PGF relationship is useful for deriving asymptotic properties of the GW process. It is easy to see that $\phi_n(0) = p(0)$, which is the probability that the n^{th} generation is the last generation and that $\phi_n'(1) = \mathbb{E}\{X_n\}$, which is the mean number of users receiving the message at n^{th} generation.

There are two cases of interest here. First case is subcritical and happens when, $\delta\alpha < 1$, and the second case is supercritical when $\delta\alpha > 1$. One important parameter is the probability of extinction, *i.e.* the probability that $X_k = 0$ for a k. This probability can be derived as the smallest positive solution q of the equation $s = \phi(s)$, where $\phi(s)$ is the PGF of the number of offspring of a node. Let's here derive the parameters for the subcritical and supercritical cases that are observed in practice.

1) Subcritical case: $\mu < 1$: when $m = \phi'(1) < 1$, it can be proved that $q = 1$ is the smallest positive solution. This means that the diffusion tree will surely die. The number of generations (τ) of a subcritical GW diffusion tree can also be bounded as [26]:

$$\begin{cases} \frac{\log\delta}{|\log\mu|}\left(1 - \frac{\log\log\delta - |\log\mu|}{\log\delta}\right)\left(1 - \frac{1}{\delta}\right) \leq \mathbb{E}\{\tau\} \\ \mathbb{E}\{\tau\} \leq \frac{\log\delta}{|\log\mu|} + \frac{2-m}{1-m} \end{cases} \tag{6}$$

The mean total number of users receiving a tweet (M) and the mean total number of retweets (T) are derived as :

$$\bar{M} = \frac{1}{1-\alpha\delta},\ \bar{T} = \frac{\alpha\delta}{1-\alpha\delta} \tag{7}$$

2) Supercritical case: $\mu > 1$: When $m = \phi'(1) > 1$, we have $q < 1$ and the tree will continue to grow with a probability $1 - q$. More precisely, we can observe the asymptotic behavior, when $n \to \infty$, and in this supercritical case X_n converges either to ∞ with a probability $1 - q$ or to 0 with a probability q. In other words, when $\mu > 1$, one

can consider that asymptotically the process will die with probability q at each stage, or express differently which means that $\mathbb{P}\text{rob}\{X_n > 0\} = 1 - q$ for n sufficiently large. In the supercritical case, the number of generations can be potentially infinite and results in an infinite number of members of the GW tree. However, if we assume that the tree will be extinct at some point we can derive the expected numbers of members as:

$$\bar{M} = \frac{1}{1 - \phi'(q)},\ \bar{T} = \frac{\alpha}{1 - \phi'(q)} \tag{8}$$

where ϕ is the PGF of X_n as defined above.

B. Killing process

So far, we can already derive from the GW process the probability of extinction and the mean number of generations for a tweet spreading. However as we will observe in the next section on our empirical data , the dynamic of the GW process might capture an overestimated model for the propagation of tweets. In essence, we can observe that in real life the hop depth of the propagation trees and the mean number of users receiving a tweet are lower than what are predicted by the GW process. As discussed previously, this difference might be explained intuitively by the difference in nature between the genealogy of offspring which is modeled by the GW process and the actual information spreading that might be influenced by the content novelty.

We therefore need to introduce a killing probability π, which represents the probability that the GW process is killed prematurely at the n^{th} generation, resulting in a Galton-Watson process with Killing (GWK). The probability that the process is killed after k generations becomes $\pi(1-\pi)^{k-1}$.

One might at the first glance think that a GWK process with retweeting probability α is equivalent to a classical GW process with retweeting probability $\alpha_K = (1-\pi)\alpha$. However, this is not the case because when killing happens in the GWK process, in the last hop where all nodes are stopped together, we cannot assume anymore that nodes have offsprings independently from each others, which is a different assumption compared with the classical GW model. Nonetheless, the GW process with retweeting probability α_K can be considered as a lower bound of the GWK process, with α, the probability of having offspring at any hop (except the last one) in the GW process, being strictly lower than the corresponding probability in the GWK process. For this reason, one can then expect the number of receivers in a GW process with a retweeting probability α_K and $\mu_K = \alpha_K\delta$ to be a strict lower bound of the number of receivers in a GWK process.

In general, in a GWK process, a GW tree falls in one of three situations: either it is finished because of natural extinction of the GW process, or it is killed because of the killing process or it can also grow infinitely. If $\mu < 1$ then the probability of the third situation happening is 0. This typically means that the probability of generating a finite GW tree is larger for the GWK process than for GW. However, a major issue is that one can't disambiguate the reasons why a tree would stop growing because of a natural extinction or a killing process. We will see in the next section that this might create problems during the estimation of the killing probability π.

C. Asymptotic analysis of the GWK process

As described earlier, the PGF of X_n in a GW process is obtained recursively from the PGF of the number of offspring $\phi(s)$ through $\phi_{n+1}(s) = \phi_n(\phi(s))$. When a killing probability is added to the GW process the PGF of the overall number of members of the GW tree is given by :

$$\phi_M(s) = \sum_{n=1}^{\infty} \phi_n(s)\pi(1-\pi)^{n-1} \tag{9}$$

and it verifies the following equation $\phi_M(s) = \pi\phi(s) + (1-\pi)\phi_M(\phi(s))$. The mean number of members of the GW tree, or equivalently the mean number of users receiving a tweet, can be derived as :

$$\bar{M} = \phi'_M(1) = \sum_{n=1}^{\infty} \phi'_n(1)\pi(1-\pi)^{n-1} \tag{10}$$

where $\phi'_n(1) = \mu^n$ for GW process. We have therefore this relation for a GW process with a killing probability π:

$$\bar{M} = \begin{cases} \frac{\mu\pi}{1-\mu+\mu\pi}, & \text{when } \mu_K < 1 \\ \infty & \text{when } \mu_K \geq 1 \end{cases} \tag{11}$$

In [22] it is showed that if the offspring distribution has finite mean $\mu = \phi'(1)$, the dominant tail will be $P(M = m) \approx R(m)m^{-1-\kappa}$ where

$$\kappa = \frac{\log(1-\pi)^{-1}}{\log \mu} \tag{12}$$

This result means that we expect the distribution of the number of users receiving a tweet to have a power law behavior with an exponent $1 + \kappa$.

V. Validation

In this section we will validate our usage of the GW and GWK model for describing information diffusion over microblogs. We first describe how to estimate the model parameters and thereafter show that the GWK can be applied to the retweeting trees. However, because of lack of space, we will just show the results of the validation over the Sina Weibo dataset, S_{user}. The similar results are obtained over the Twitter dataset T_{user}.

A. Parameter estimation

The model described in the previous sections depends on three main parameters:α_i the retweeting probability of the tweet i, δ the mean number of followers of microblog users that have at least one followers, and π, the probability that a tweet diffusion tree is killed at each hop. We first need to propose way of extracting these parameters over a dataset. Out of these three parameters the first one has to be derived for each tweet and the last two have to be derived over the whole dataset and can be considered as properties of the microblog site. A major issue that we have to deal with is relative to the ambiguity of the cause of death of a diffusion tree. A diffusion

TABLE II
GALTON-WATSON PARAMETERS ON MICROBLOG

Dataset	δ	π	$\%(\mu_i < 1)$
S_{user}	93.4	0.49	54.2%
T_{user}	103.92	0.53	77.5%

tree can be finished because of natural extinction caused by the GW process or of being killed by the killing process. As α is relative to the GW process and π to the killing process, being able to separate these two effects is very important.

1) Inference of δ: For the forthcoming analysis we need to obtain the distribution of the number of followers and to derive from it the mean and its PGF. As explained earlier, the initial user unbiased Sina Weibo dataset contains 500,000 microblog users. As the number of followers of a user is the open and available information, deriving the follower statistics is straightforward. We observe on this dataset a mean of 93.4 followers, and a very large deviation equal to 4520.2, showing the very large variability of the number of followers in the dataset. Twitter dataset has a mean of 103.92 followers, and a deviation equal to 1436.8. We use the resulting distribution as $\mathcal{D}(k)$ in Eq. 1.

2) Estimation of α_i: The second important parameter to estimate is the retweeting probability α_i. However this value should be estimated when the diffusion tree is not in the "killed" state. To ensure this, we estimate α_i in the retweeting tree of i as the proportion of users that retweeted the message among all users who received the message excluding these in the last diffusion hop.

3) Estimation of π: The last parameter to infer is π, the probability of killing a tweet at each hop. We therefore need to ensure that a tweet is not naturally extinct , before using it for estimating π. In order to achieve this, we calculate for each diffusion tree the probability that the tree is naturally extinct at the last hop. If we assume that a tweet i has a tree with N receiving users in its last hop, and that the retweeting probability of this tweet is α_i, then the probability of extinction is given by $P_e(i) = (1-\alpha_i)^N$. We derive for all tweets this value and decide to put aside all tweets that have a probability of extinction larger than 5%, or in other terms, we focus on all tweets which the probability of being killed at the last hop is larger than 95%, resulting in 37,866 tweets. We thereafter derive the value π by fitting the formula $\pi(1-\pi)^l$ to the distribution of maximum retweeting hop number l measured over these tweets. We show in Fig. 3 the fit over the distribution of maximum hop number of S_{user}. We can see that the distribution of generation number follows the expected exponential decrease. We estimate $\pi = 0.49$ for Sina Weibo and $\pi = 0.53$ for Twitter.

B. GW model validation

As explained in section IV-A, there is a fundamental difference among the two cases: $\mu < 1$ and $\mu > 1$. The GWK process parameters calibrated over S_{user} and T_{user} are shown in Table II. Here we can estimate for each tree a μ_i using

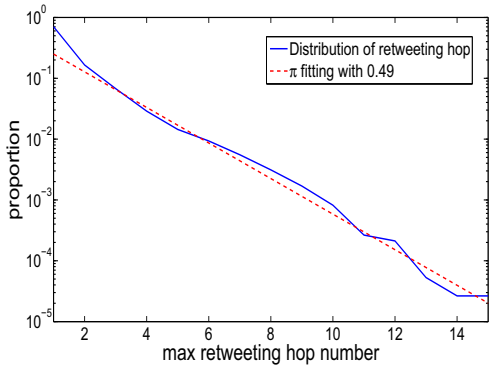

Fig. 3. π fitting with distribution of maximum retweeting hop number of S_{user}

the following estimator that is known to be the maximum likelihood estimator of μ:

$$\hat{\mu} = \frac{\sum_{i=2}^{L} X_i}{\sum_{i=1}^{L-1} X_i}, \tag{13}$$

where L is the maximum hop length of the retweeting tree. Note that we do not consider the number of offspring for X_0(*i.e.* X_1) in numerator, because as stated in section IV-A, Eq. 3 is not suitable for X_1.

Table II also gives the proportion of tweets in $\mu < 1$ case for each dataset. We have first to assess if the killing process is needed or not. For this purpose we can do two tests: first to check if the number of observed hops is compatible with the formula given in Eq. 6, and to check if the mean number of receiving nodes is compatible with Eqs. 7 and 8. The first method is only applicable to subcritical tweets and the second method is applicable to all tweets. Using the lower bound in Eq. 6, we observed that 89% of tweets have a maximum retweeting hop number less than the lower bound given for the GW model. Moreover, we show in Fig. 4, a graph showing the number of receivers as predicted by the Galton-Watson model and what is observed. We observed that for 83% of tweets, the observed receivers number is less than the value predicted by the Galton-Watson model (below the line $y = x$ line). Indeed, one can expect that tweets which are naturally extinct will have a mean receivers number following the GW equations. However we can observe that there is considerable proportion of tweets that have a number of receivers much less than the mean predicted by the GW model. The two above results validate that we need to add a killing process that will account for the reduction of number of receivers and the smaller hop lengths.

We explained in section IV-B, that the number of receivers in a GWK process can be lower bounded by the number of receivers in a modified GW process with retweeting probability α_K. We show in Fig. 5 the comparison of the number of receivers predicted over the modified GW process with what is observed. The figure confirms that the modified GW process acts as a strict lower bound to the GWK process. However as expected this bound is not very tight. The above analysis

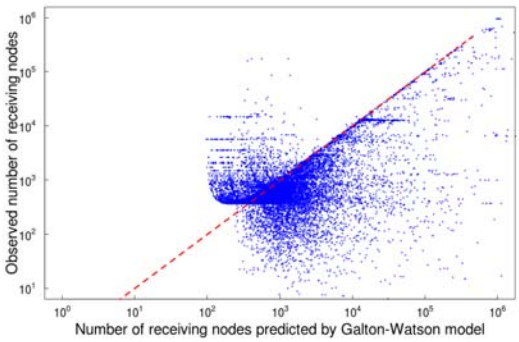

Fig. 4. Comparison of number of receivers in a tweet tree as predicted by the GW model with what observed.

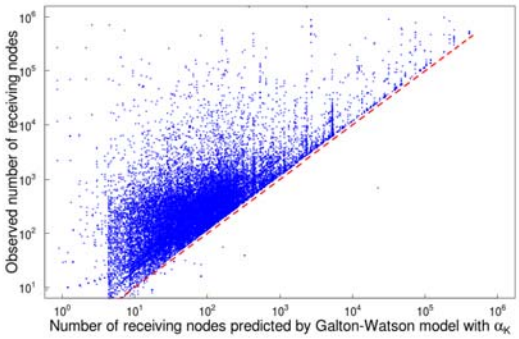

Fig. 5. Comparison of number of receivers in a tweet tree as predicted by the modified Galton Watson model with what observed.

validates the relevance of the GWK process for analyzing the propagation on microblog systems. In the following we present two possible applications demonstrating the usefulness of the GWK model.

VI. Applications

We present two applications of our GWK model in this section. The first application shows the use of the proposed GWK model in synthetic workload generation with similar statistical properties to empirical microblogs load, validating that our proposed is constructive. This opens way to implement microblog traffic simulators that can be used to stress microblog systems. The second application we will describe is relative to highly popular tweets. The GWK model and its parameters provide fine grain features that will be shown to be highly relevant to understanding the popularity of tweets.

A. Tweet load synthesis

The GWK model provides a way of generating tweet propagation trees by simulating the GWK model with parameters derived over an empirical dataset. The simulation can be easily implemented as it simply consists of beginning from one seed user and generating the first generation by choosing a number of receivers following the distribution $f(k)$ defined in Eq. 1. Recursively, each user of a new generation chooses its receivers number following the same distribution. At the end of each generation we check with probability π if the generation should be killed. The parameter π, and the distribution $\mathcal{D}(k)$ are obtained following the above described methods. The parameter α_i is chosen randomly from an empirical distribution obtained over the corresponding dataset. This can be implemented in a small program that generates trees similar to the ones generated by microblogs. We show in Fig. 6 the Complementary Cumulative Distribution Function (CCDF) of receivers number in the trees generated following the GWK model and we compare it with the empirical CCDF obtained over the dataset S_{user}. As can be seen there is a very good fit between the two distributions both in the head and the tail. We also show in Fig. 7 the distributions of the maximum retweeting hop number both in the tree synthesized by the GWK model and what observed empirically over the dataset S_{user}. Here the fit is also striking.

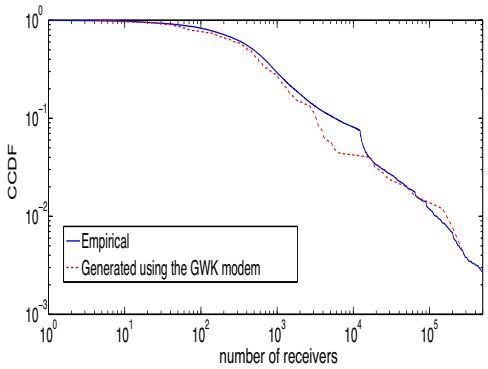

Fig. 6. Comparison of the CCDF of receivers in trees generated by the GWK model and the empirical CCDF over S_{user}

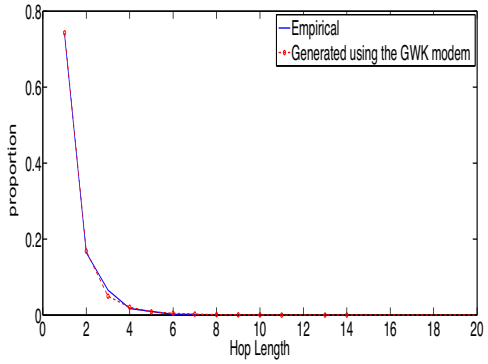

Fig. 7. Comparison of distribution of maximum retweeting hops in trees generated by the GWK model and the empirical distribution of maximum retweeting hops in S_{user}

The above analysis shows that the GWK can be used to synthesize retweeting trees that have realistic macroscopic distribution. This validates the use of the GWK model for microblog workload simulation.

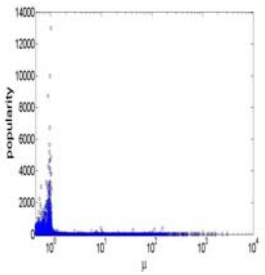

Fig. 8. Popularity against estimated μ in S_{user}

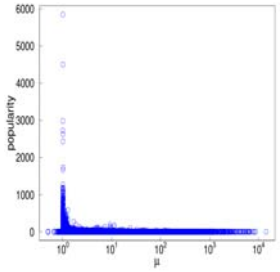

Fig. 9. Popularity against estimated μ in T_{user}

B. Popular tweets characteristics

The asymptotic analysis of the GWK shows the importance of the $\kappa = \frac{\log(1-\pi)^{-1}}{\log \mu}$ in predicting the tail behavior of popularity, measured by the number of retweets, and the audience, measured by number of receivers. This feature is interesting as it mixes in a single equation all the parameters of the GWK and weights the impact of each of these parameters. It is therefore interesting to look at the value of this feature for overall tweets. However a more precise look shows that as π is a parameter of the global microblog and is constant for all tweets, so that κ is directly related to $\mu = \alpha\delta$.

Previous works [25] showed that there are two paths for a tweet to become popular: a path that is endogeneous and involves having retweeters that have a large number of followers so that the tweet attains a large audience, and an exogeneous path that explains the popularity by the intrinsic interest of the message. Each one of these paths can be represented in $\mu = \alpha\delta$ where δ accounts for number of followers and α accounts for the intrinsic interest of the tweet, meaning that μ mixes these two aspects. The theorical analysis shows a clear distinction of the asymptotic behaviors between $\mu > 1$, where the tree is expected to become infinite in absence of killing, and the case $\mu < 1$ where the tree will surely extinct even without killing. It will be interesting to check if tweet audience and popularity, measured in term of number of receivers and number of retweets, are related to the value of μ.

We plot in Figs 8 and 9, the relation between the estimated μ using Eq.13 for each retweeting tree and the popularity evaluated as the number of retweets. While one would expect that large μ leads to large popularity, this is definitely not the case. Table III contains the characteristics of highly popular tweets. As can be seen all highly popular tweets in the two datasets are sharply concentrated around a value of $\mu = 1$. In fact we observe that what leads to a large diffusion and a large popularity is rather a balance between followers number and retweeting probability that results in μ being close to one. This is confirmed by looking at the CCDF of the estimated specific δ_i for each retweeting tree instead of the global constant δ which is shown in Fig. 10. Here we get δ_i from $\delta_i = \frac{\mu_i}{\alpha_i}$ where we also don't consider the $0th$ hop and $1st$ hop as section IV-A stated. As can be seen in S_{user} the popular tweets generally exhibit δ values that are generally larger than those observed over the whole dataset, but still these tweets have not very large δ. The situation is slightly different in T_{user} where all popular tweets have smaller δ's than other tweets. Nonetheless, either for Sina Weibo or Twitter dataset, popular tweets never happen for small δ's.

The above observations give an interesting characterization of highly popular tweets. These tweets have a δ relatively large (larger than 50 for Twitter and larger than 200 in Sina Weibo) and accordingly small retweeting probability (to end up with a μ close to 1). In particular no popular tweets resulting from several hops of small neighborhood diffusion has been observed, ruling out social rumors type of propagation. We also see few popular tweets with a large value of δ excluding the followers of publisher. This last observation is in accordance with [3].

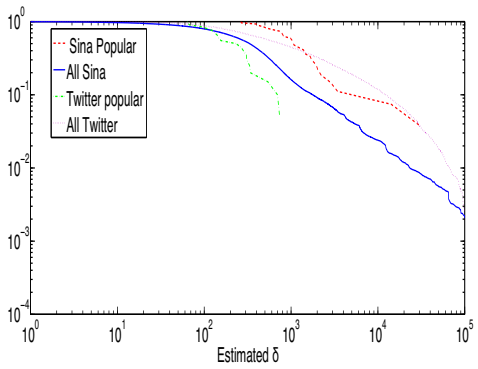

Fig. 10. CCDFs of δ estimated from popular and all tweets

VII. Conclusion and Future Work

In this paper we have analyzed information spreading patterns on microblogs and built a novel model for the information spreading. The model is based on the analogy with the Galton-Watson branching processes that describe the evolution of family names. We refine the model with the killing process and validate the applicability over two datasets from Sina Weibo and Twitter. We present two applications of our model, namely, microblog workload generation and popular tweet characterization. We show that the Galton-Watson with Killing model is useful not only for describing the information diffusion but also for providing the insights into popular tweets.

This work is leading us to the following new directions. The GWK incorporates time as a discrete generational index, and does not account for the temporal dynamics of tweet diffusion. We are considering a continuous time model that captures the temporal dynamics. In our model we have used the values of the mean number of followers, δ, the retweeting probability, α, and the killing probability, π, but have not addressed factors behind them. The user's social capital in the network as well as the history of retweets are likely to be correlated to the parameters of our model. In addition one particular observation of interest is that highly popular tweets all have a μ value

TABLE III
CHARACTERIZATION OF HIGHLY POPULAR TWEETS

Dataset	Popularity range	Popularity%	Mean of μ	Median of μ	1-percentile	99-percentile	Tweet%
T_{user}	All popularity	100%	3.691	1.247	1.000253	28.5956	99%
T_{user}	popularity $>$ 100	0.182%	1.053	1.002	1.000011	1.271943	51.13%
T_{user}	popularity $>$ 1000	0.002%	1.003	1.000	1.000006	1.028369	17.83%
S_{user}	All popularity	100%	3.596	1.000	0.984	34.1492	99%
S_{user}	popularity $>$ 100	3.06%	1.394	0.8583	0.7676	1.0262	66.8%
S_{user}	popularity $>$ 2000	0.09%	0.904	0.9376	0.8237	1.002	1.5%

close to 1. We leave the study of compounding factors for future work.

VIII. ACKNOWLEDGMENT

This work was supported in part by National Basic Research Program of China with Grant 2012CB315801, by National Natural Science Foundation of China (NSFC) with Grants 61133015 and 61272473, by National High-tech R&D Program of China with Grant 2013AA013501, by Strategic Priority Research Program of CAS with Grant XDA06010303, by the Instrument Developing Project of CAS with Grant YZ201229. The work was also supported by EC EINS project and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No.2012033242).

REFERENCES

[1] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, 2005.

[2] Y. Ahn, S. Han, H. Kwak, Y. Eom, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web*, 2007.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 2010.

[4] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: Analyzing the worlds largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference*, 2007.

[5] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in flickr. In *Proceedings of the 1st Workshop on Online Social Networks*, 2008.

[6] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World Wide Web*, 2009.

[7] M. Cha, J. Perez, and H. Haddadi. Flash floods and ripples: The spread of media content through the blogosphere. In *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*, 2009.

[8] M. Goetz, J. Leskovec, M. McGlohon, and C. Faloutsos. Modeling blog dynamics. In *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*, 2009.

[9] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

[10] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, 2004.

[11] A. Guille and H. Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. In *Proceedings of the 21st international conference on World Wide Web*, 2012.

[12] L. Guo, E. Tan, S. Chen, Z. Xiao, and X. Zhang. The stretched exponential distribution of internet media access patterns. In *Proceedings of the 27th ACM Symposium on Principles of Distributed Computing*, 2008.

[13] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform url sampling. In *Proceedings of the 9th international conference on World Wide Web*, 2000.

[14] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[15] H. Kwak, , H. Chun, and S. Moon. Fragile online relationship: a first look at unfollow dynamics in twitter. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, 2011.

[16] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web*, 2010.

[17] J. G. Lee, S. Moon, and K. Salamatian. An approach to model and predict the popularity of online contents with explanatory factors. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, 2010.

[18] K. Lerman and R. Ghosh. Information contagion: an empirical study of the spread of news on Digg and Twitter social networks. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 2010.

[19] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs patterns and a model. In *Proceedings of the 7th SIAM International Conference on Data Mining*, 2007.

[20] A. Mislove, M. Marcon, K. P.Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference*, 2007.

[21] S. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.

[22] W. Reed and B. Hughes. On the distribution of family names. *Physica A: Statistical Mechanics and its Applications*, 319:579–590, 2003.

[23] T. R.Zaman, R. Herbrich, J. van Gael, and D. Stern. Predicting information spreading in twitter. In *Proceedings of Neural Information Processing Systems*, 2010.

[24] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina. Correcting for missing data in information cascades. In *Proceedings of the 4th Annual ACM International Conference on Web Search and Data Mining*, 2011.

[25] D. Sornette, F. Deschâtres, T. Gilbert, and Y. Ageon. Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings. *Phys. Rev. Lett.*, 93:228701, Nov 2004.

[26] V. Vatutin and A. Zubkov. Branching processes. i. *Journal of Mathematical Sciences*, 39(1):2431–2475, 1987.

[27] K. A. Wojciech Galuba, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In *Proceedings of the 3rd Workshop on Online Social Networks*, 2010.

[28] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *Proceedings of the 10th IEEE International Conference on Data Mining*, 2010.

[29] S. Ye and F. Wu. Measuring message propagation and social influence on twitter.com. In *Proceedings of the 2nd International Conference on Social Informatics*, 2010.