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

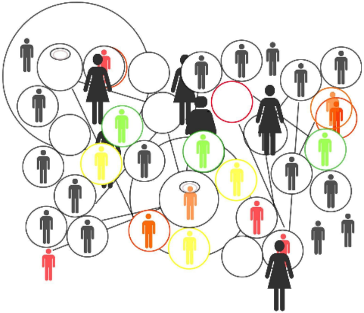
Diffusion in Social Networks

A fundamental process in social networks: behaviors that cascade from node to node

- News, opinions, rumors,...
- Virus, disease propagation
- Localized effects: riots

Microblogging Changes How People Discover and Consume Information Online

Post



Case Study: Gangnam Style

Katy Perry
@katyperry

Help, I'm in a gangnam style k hole: bit.ly/PVPJ4p

12K reposts

Britney Spears
@britneyspears

I am LOVING this video - so fun! Thinking that I should possibly learn the choreography. Anybody wanna teach me?! haha britspears.me/159zje

4K reposts

T-Pain
@TPAIN

Words cannot even describe how amazing this video is... youtube.com/watch?v=9bZkp7...

2K reposts

Motivation

Study the temporal dynamics of an information cascade in a microblogging system

- The number of users influenced at any given time

Related Work

- Epidemic model:
    - SIS model
    - SIIRP model
- Independent Cascade (IC) Model
- Linear Threshold (LT) Model
- Linear Influence Model

Data-driven Approach: Measurement Study

Tencent Weibo.

- 0.5B users - one of the largest social network services in China

A sample of video sharing in 20 days

- 1M users - social relation, behaviors
- 2M entries - each entry corresponds to one post or repost
- 350K video links - 5 video sharing websites, 14 categories

A Glance of Microblogging Diffusion

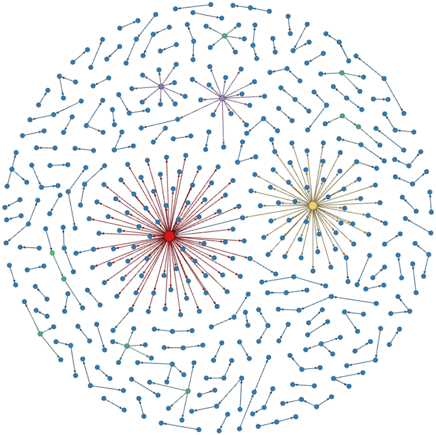


Figure: Example diffusion cascades in Tencent Weibo.

Power-law Distributions of the Number of Followers and the Number of Reposts

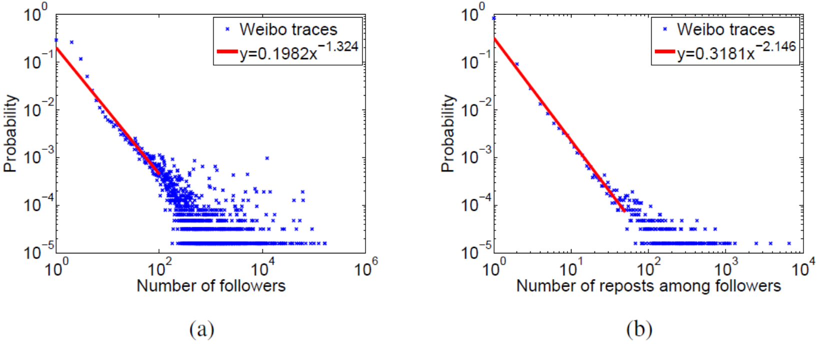


Figure: Distribution of the number of followers of users, and the number of reposts to their microblogs.

Evolution of Cascade Size

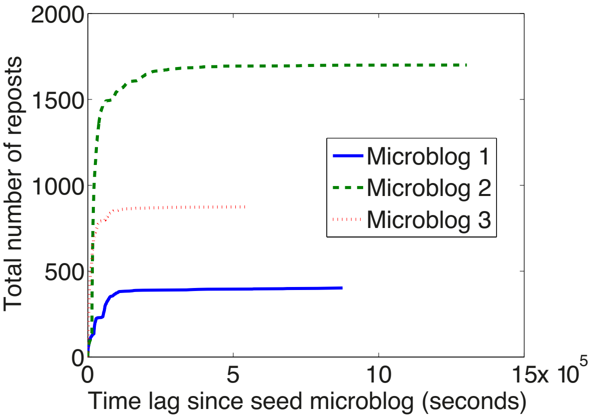


Figure: The total number of reposts versus the time lag since when the seed microblogs are posted.

Gamma Distribution of the Response Delays

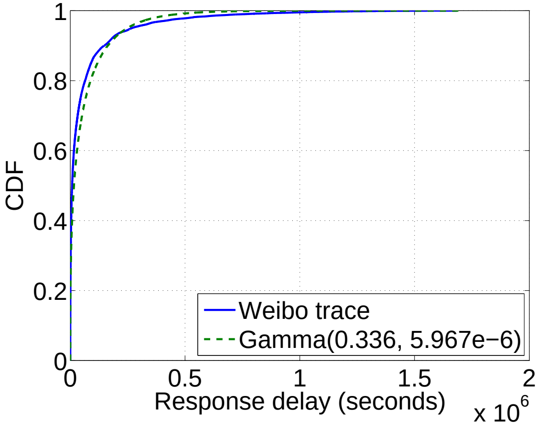


Figure: CDF of response delays of all reposts in our traces.

Objective

How many users in total are expected to have reposted the microblog after a certain time t?

Branching Process

Branching process

- Each individual gives birth to a random number of offsprings independently according to a certain distribution

Age-dependent branching process

- The lifetimes of individuals are considered based on a lifetime distribution

Mapping

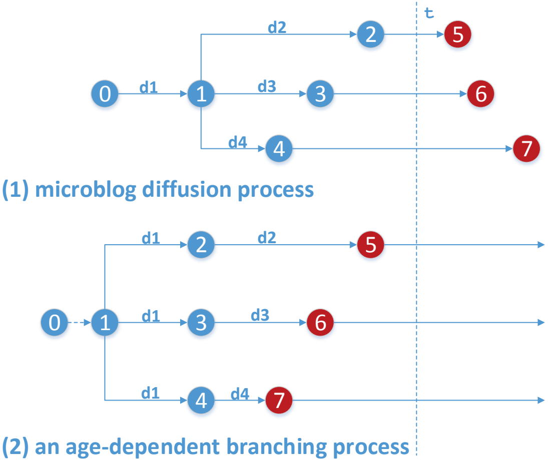
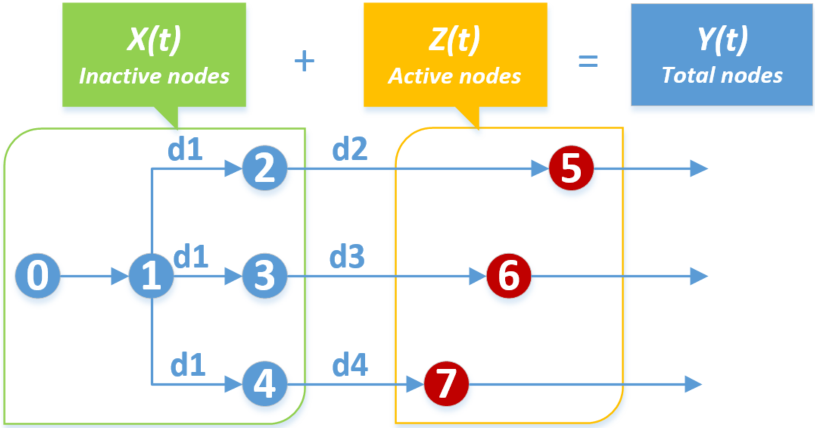
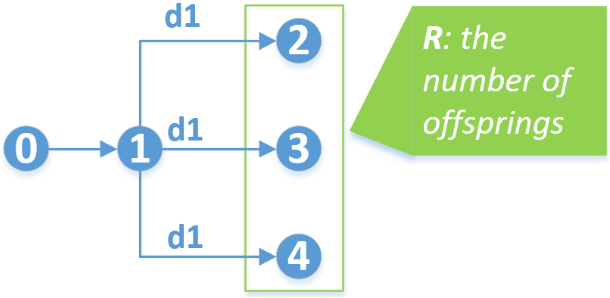


Figure: Mapping between the microblog diffusion cascade and an age-dependent branching tree.

Basic Notations

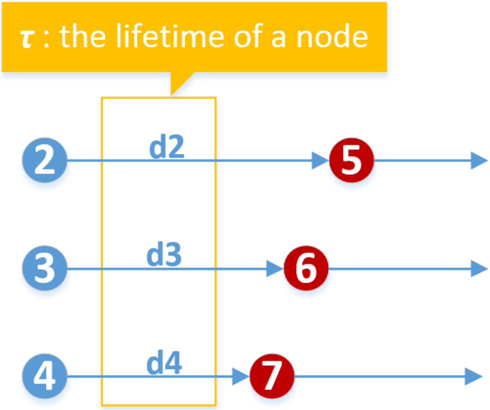


Degree Distribution



- $p(R = k) = p_k$: the probability density function of the number of offsprings of a node in the branching tree
  - $\mu = \sum_k^{\infty} p_k k$: the reproductive number of a node in the branching process

CDF of Lifetimes



- $G(\tau)$: the cumulative distribution function of the lifetimes of nodes in a branching process

Probability $P(Z(t) = k)$

- Probability density function of $Z(t)$

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

  - P^{*j} is the j-fold convolution of P
  - δ_{1k} is the Kronecker delta

Probability Generating Function

- Probability generating function of $Z(t)$

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

Partial Derivative

- Taking derivative

$$
\begin{aligned}
\frac{\partial F(s,t)}{\partial s} =& [1 - G(t)] \\
& + \int_0^t h^{'}[F(s,t-\tau)] \frac{\partial F(s,t-\tau)}{\partial s} dG(\tau)
\end{aligned}
$$

Bounded Limit

- Taking limit $s \to 1$

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

- Similarly

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

Expressions of $z(t)$ and $y(t)$

- Solutions using renewal theory

$$z(t) = [1 - G(t)] * U(t)$$

and

$$y(t) = U(t)$$

  - Renewal function $U(t) = \sum_{n=0}^{\infty} \mu^n G^{*n}(t)$

Expected Number of Inactive Nodes in a Branching Tree

- Solution of $x(t)$

$$x(t) = y(t) - z(t) = G(t) * U(t)$$

    - Analytic form of $x(t)$ via Laplace transform and inverse Laplace transform

Overall Size of a Microblog Cascade

- Overall size of a microblog cascade

$$\tilde{x}(t) = \sum_{k=0}^{\infty} p_k k x(t) + 1 = \mu x(t) + 1$$

  - 1 corresponds to the seed post

Experiment Setup

Simulate a microblogging network

- The number of followers and response delays follow the same distributions in the measurement study
- Run 10^4 times for every set of parameters

Evolution of Cascade Size

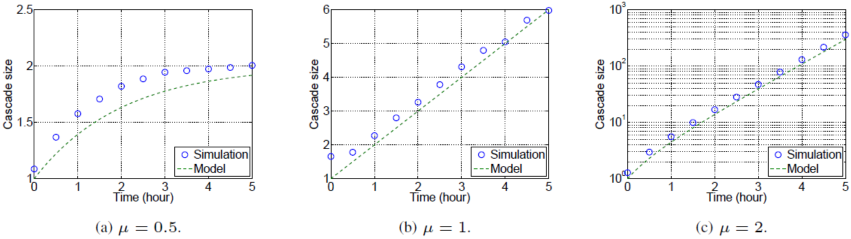


Figure: Comparison of the evolution of cascade sizes generated by simulations and our model.

Cascade Size over Time with Two-stage Degree Distributions

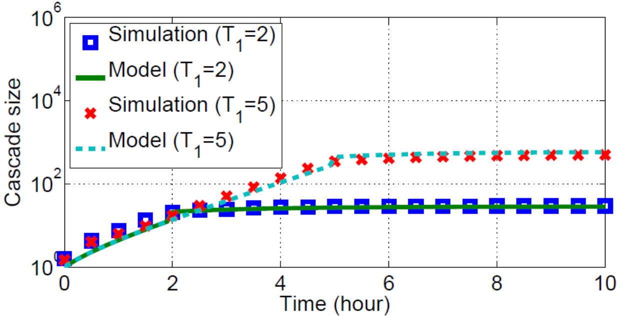
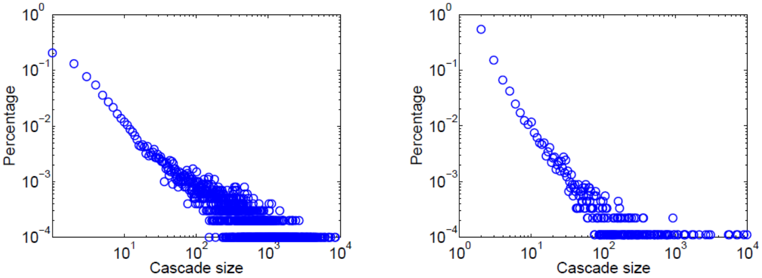


Figure: Comparison of the evolution of cascade sizes generated by simulations and our model: two-stage μ.

Final Cascade Size



(a) Simulation (b) Tencent Weibo

Figure: Distribution of final cascade sizes.

Summary

- A large-scale measurement study reveals several facts on microblog propagation
- Detailed mathematical derivation of the expected cascade size at any time during a microblog diffusion process is given
- Trace-based simulation experiments demonstrate the effectiveness of our model

Q & A

More information: http://i.cs.hku.hk/ cwu/papers/sshi-icc14.pdf