Modeling of information diffusion on social networks with applications to WeChat

Liu, Liang; Qu, Bo; Chen, Bin; Hanjalic, Alan; Wang, Huijuan

DOI
10.1016/j.physa.2017.12.026

Publication date
2018

Document Version
Accepted author manuscript

Published in
Physica A: Statistical Mechanics and its Applications

Modeling of Information Diffusion on Social Networks with Applications to WeChat

Liang Liu[a,b], Bo Qu[b], Bin Chen[a], Alan Hanjalic[b], Huijuan Wang[b,*]

[a] *College of Information System and Management, National University of Defense Technology, Changsha, China, 410073*
[b] *Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Mekelweg 4, Delft, The Netherlands, 2628 CD*

Abstract

Traces of user activities recorded in online social networks open new possibilities to systematically understand the information diffusion process on social networks. From the online social network WeChat, we collected a large number of information cascade trees, each of which tells the spreading trajectory of a message/information such as which user creates the information and which users view or forward the information shared by which neighbors. In this work, we propose two heterogeneous non-linear models, one for the topologies of the information cascade trees and the other for the stochastic process of information diffusion on a social network. Both models are validated by the WeChat data in reproducing and explaining key features of cascade trees.

Specifically, we apply the Random Recursive Tree (RRT) to model the growth of cascade trees. The RRT model could capture key features, i.e. the average path length and degree variance of a cascade tree in relation to the number of nodes (size) of the tree. Its single identified parameter quantifies the relative depth or broadness of the cascade trees and indicates that information propagates via a star-like broadcasting or viral-like hop by hop spreading. The RRT model explains the appearance of hubs, thus a possibly smaller average path length as the cascade size increases, as observed in WeChat. We further propose the stochastic Susceptible View Forward Removed (SVFR) model to depict the dynamic user behavior including cre-

*Corresponding author
Email address: h.wang@tudelft.nl (Huijuan Wang)

ating, viewing, forwarding and ignoring a message on a given social network. Beside the average path length and degree variance of the cascade trees in relation to their sizes, the SVFR model could further explain the power-law cascade size distribution in WeChat and unravel that a user with a large number of friends may actually have a smaller probability to read a message (s)he receives due to limited attention.

Keywords:
Information cascade, Stochastic model, Social networks, WeChat, Random recursive tree

1. Introduction

The rapid development of the Internet, smart phones and information technology has facilitated the boost of online social networks, such as Facebook, Twitter, Flickr, Digg and Sina Weibo, each of which support the spread of information, behaviour and opinion [1–17]. Data such as information diffusion trajectories recorded online allows us to further identify the spreading patterns and the underlying spreading process on a social network. Such understanding is crucial for businesses to promote products and for governments to predict and even regulate public opinion [18–20].

In this work, we consider the information diffusion trajectories recorded on a social network. The spreading trajectory of each information content can be represented by a cascade (tree), where the root is the source node that creates the information and the links represent the information transmitting paths between users. First, we aim to model the topologies of the information cascade trees with few parameters. Such a topology model of a group of cascade trees with few parameters would allow us, for example, to quantify to what extent information spreads viral-like (via hop by hop propagation) or broadcast like (via hubs), to compare various online social networks, and possibly to distinguish and/or identify the spread of a certain type of information such as misinformation [21]. Second, we aim to develop a dynamic model of the information diffusion process on a social network with few parameters, that could capture several key features observed in the cascade trees. Such discovery of the first-order spreading process/mechanisms is essential to design optimisation strategies e.g. how to select the source node to publish the information such that more users could be reached.

Topologies of cascade trees have so far been characterised by the average

path length of a cascade tree[1], also called structural virality, in relation to the size of the tree[9, 22, 23]. The size (number of nodes) distribution of cascade trees has been shown to be highly skewed [24]. Consider the class of cascade trees collected from an online social network. If the average path length of a cascade tree does not increase much with the size (number of nodes) of the tree, hubs may exist in relatively large cascade trees. In this case, information propagates via star-like broadcasting and large cascade trees are relatively shallow [9, 25, 26]. If the cascade trees' average path lengths increase dramatically with their sizes, large cascade trees tend to be deep without large hubs and information spreads viral-like, hop by hop. However, we lack a systematic method to quantify the extent of the shallowness or deepness of a group of cascade trees. In this work, we propose to use of the generalised random recursive tree (RRT) [27, 28] with a single parameter to model a group of cascade trees (possibly of a given type of contents) with diverse sizes in an online platform. The RRT, a growth tree model, could well capture two features of WeChat cascade trees: the average path length and the degree variance, as a function of the cascade size. The identified parameter in the RRT model quantifies how deep or shallow the cascade trees are and indicates the possible growing mechanisms of cascade trees.

Stochastic models, such as *cellular automata* [29], *Threshold models* [30–33], *Susceptible Infected Recovered (SIR)* [14, 34–36], and *Linear Influence* [37] have been studied to understand how the dynamics of information diffusion such as the spreading rate and the social network topology could influence a key feature of the diffusion process such as cascade size. However, we still insufficiently understand whether such first-order models with few parameters could quantitatively reproduce several key features of real-world information diffusion. Moreover, does a user with a large number of friends have a lower probability to view a message it receives, according to earlier evidence found in [14]? Correspondingly, we propose the heterogeneous Susceptible View Forward Removed (SVFR) model, which allows users to have different probabilities of viewing a message, depending on their degree (the number of neighbors) in the underlying social network. Interestingly, our SVFR model could well explain the power-law distributed size of cascade trees, the degree variance and the average path length of a cascade tree in

[1]The average path length of a cascade tree is the average number of links in the shortest path between two nodes. The shortest path between any two nodes in a tree is unique.

relation to the tree size.

Our modeling methods have been illustrated and verified by the information diffusion trajectories recorded in WeChat, a social network with 800 million monthly active user accounts in 2016 [38]. We choose WeChat also because we understand far from sufficiently WeChat, a semi-closed social network where information is shared mainly via strong social ties (i.e. friends that mutually agree to share information) [39].

Our characterisation and modeling of WeChat in this work is a starting point to explore the difference between semi-closed social networks and open social networks like Twitter. Does information spread more viral-like in semi-closed networks? To answer such questions, we need to collect the diffusion trajectories of the properly selected type of contents among the properly selected population on the two types of networks for comparison purpose, beyond our modeling approach.

The remainder of this paper is organized as follows: Section 2 describes the WeChat information diffusion data and how to construct the cascade trees. Section 3 and 4 present the RRT and SVFR model to capture the topology of the cascade trees and the dynamics of the information diffusion respectively. Section 5 summarizes our findings and points out interesting future work.

2. Dataset Description

We will use the information diffusion dataset of WeChat to validate the two models that we are going to propose. We focus on the diffusion of web pages, in the WeChat social network. A user may react to a web page forwarded/shared by his/her friend, as such appearing in his/her WeChat with a title in three ways: (i) *View* the web page, meaning that the user clicks the link of the web page and views the content, (ii) ignore the web page without a click to view the content, and (iii) *Forward* (or share) the URL of the web page to all or subgroup of his/her friends after viewing the content. An example of the diffusion of a web page in WeChat is shown in Figure 1. First, a user being at the root of the tree initially forwards a web page to his friends in WeChat. Then his friends may ignore, view or forward the web page after seeing the web page appearing with a title. The forwarding of the information (web page) allows its friends to further view, forward or ignore the information. The users who have received and ignored the web page, or equivalently to whom a webpage has been shared, can not

be detected. Our topology and stochastic models to be proposed aim to capture the features of the observed (view) cascade trees composed of users that have created, viewed and forwarded the messages[2].

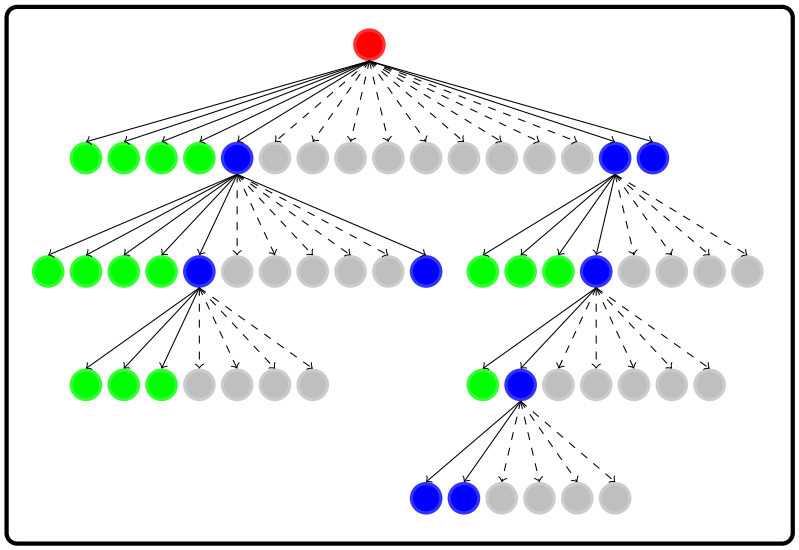


Figure 1: Schematic diagram of the diffusion of a web page in WeChat. Colors differentiate between the users showing different behaviors regarding what they do with the information forwarded to them. The green circles represent users who have viewed the message. The blue circles stand for users who have shared the message after viewing it. The grey circles are those users who have not viewed the content. A view cascade tree is composed of the source node that initially forwards the message, the nodes that have viewed the message, thus both the blue and green nodes and the black solid arrows among them. The view cascade tree of each web page is well recorded in the data.

We obtained the web page spreading dataset in WeChat Moments from

[2]It is possible that a user views/clicks the same content multiple times, forwarded by one or different friends, to read the content completely or more than once. A user in WeChat may share a content to all his/her friends or share the contents several times to several groups of friends. We aim to understand users' two levels of perceiving information: read or share the information but not more detailed behaviours such as reading a content in one time or not and sharing a content once to all friends or several times to sub-groups of friends. Hence, we construct the cascade trees by taking into account only the first view and sharing actions for each user per content. If we taken into account all the view actions, the information diffusion trajectories are not necessarily trees. Our collection of the cascade trees to be described in detail below ignores 8 percent links.

a third-party service company[3]. The service company helps users create HTML5 format web pages (e.g. news and advertisements) to share on WeChat. Spreading trajectories of these web pages have been recorded. The dataset includes all user activities from January 14 to February 27 in 2016, such as viewing and forwarding, and their corresponding time stamps related to all the web pages created with the format support from the service company. A user must first view a web page before (s)he forwards it. Whenever a user views a web page shared by a friend, the index of both the user who views the web page and the friend who shares the web page are recorded in the dataset, allowing us to construct the view cascade tree for each web page. We aim to select the web pages whose diffusion starts and ends within the period of 45 days. We assume that a web page starts to diffuse within the 45 days' observation window if the page is not viewed nor forwarded on day 1 but later and the first view of the page is a view at the page shared by the root, the user who publishes the page. We assume that the diffusion of a content stops within the observation window if there is no view nor forward action of the content on day 45 [25]. The precise identification of contents whose diffusion starts and ends within a period is challenging because the diffusion of a content could recur after a long period without being viewed/shared [40]. For example, we identify the pages that start the diffusion within period $[11, 45]$ under our assumptions and 8.8 percent of these identified pages have actually started their diffusion within day $[1, 10]$. Both the content of the web pages and users are anonymised by web page indexes and user indexes, respectively.

As a result, we obtain $229,021$ web pages, whose life span is approximately within the considered time window. More than 5 million users are involved in the diffusion of these web pages. For each web page, we construct its view cascade tree, in which nodes represent the users who have viewed the web page and some of these nodes may have forwarded the web page. A user seldom views/forwards the same content more than once. If, in the rare case a user views (shares) a web page more than once, we consider only the first time when the user views (shares) the page. Hence, each information cascade is a tree without cycles. The users who have received and ignored the web pages, thus the underlying social network, are unknown.

[3] http://www.fibodata.com/

3. Modeling of Information Cascade Tree Topology

In this section, we focus on the modeling of the topologies of the information cascade trees, without considering the underlying dynamics of users. We aim to propose a tree model that could construct trees that share similar properties of the cascade trees observed in WeChat. We will analyse two fundamental properties of the information cascade trees in WeChat, that we would like our model to reproduce, namely the *average path length* and *degree variance*. Afterwards, we propose to use the Random Recursive Tree (RRT) to model information cascade trees and illustrate to what extent this model could capture the two key features of the information cascades in WeChat.

3.1. Cascade Structure in WeChat

Two basic properties of a generic tree are the *average path length* and the *degree variance*. The average path length, also known as "Wiener Index" or "Structural Virality", is the average of the number of links H_{ij} in the shortest path between any two nodes i and j. Hence, in a tree with N nodes we can formulate it as

$$E[H] = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} H_{ij}, \tag{1}$$

The *degree variance* is the variance of degrees of all the nodes in a tree,

$$Var[D] = \frac{\sum_{i=1}^{N} (d_i - E[D])^2}{N}, \tag{2}$$

where the degree d_i of the node i tells how many links a node i has and $E[D]$ is the average degree of all the nodes. The degree variance can be equivalently characterized by the standard deviation $\sqrt{Var[D]}$ of the degree, which is used later in our data analysis and model validation.

Both properties can depend on the size of the tree. Hence, we propose to characterize how deep/shallow the class of observed cascade trees is by these two properties as a function of tree size. As shown in Figure 5, the sizes of the cascade trees collected from WeChat follow approximately a power-law distribution. Hence, we group the cascades trees according to their sizes that are slitted uniformly in logarithmic scale. We consider cascading trees that have more than 100 nodes in the dataset, which corresponds to the web pages that could propagate to a certain extent. Both properties are explored

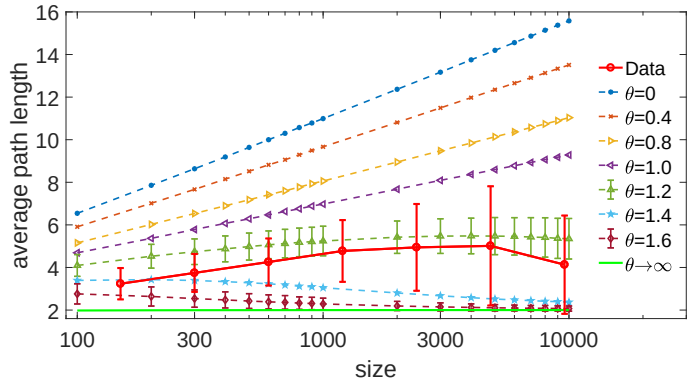
for each group of trees. Figure 2(a) and 2(b) show the average path length and degree variance of a cascade tree as a function of the size of the tree, respectively. The average path length increases first and decreases afterwards as the size of the cascade tree increases. The decrease of the average path length with the cascade size when the size is above 10^4 is due to the hubs in the cascade trees, i.e. high degree nodes, which is reflected in the large degree variance of large cascade trees.

3.2. The Random Recursive Tree Model

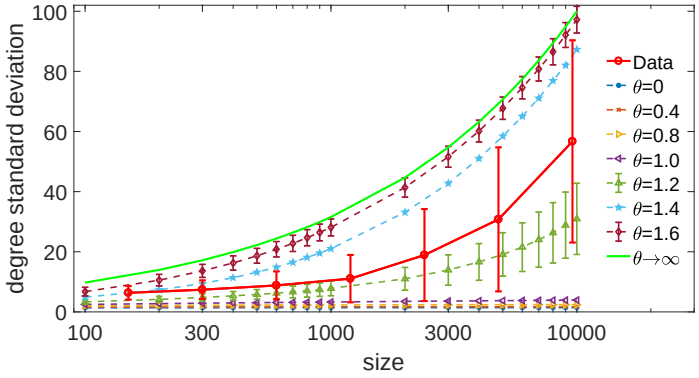
We propose to use the Random Recursive Trees (RRTs) to model the cascade trees. The RRT [27, 28, 41] is a growth tree model with a single preferential attachment parameter. It starts with the root node at $t = 0$ and adds a node at each time step t to an existing node selected as follows: each existing node i with its degree $d_i(t)$ at time t has the probability $\frac{d_i^\theta(t)}{\sum_{i=1}^t d_i^\theta(t)}$ of being connected to the newly added node. Hence, the probability that an existing node is connected to a newly added node is proportional to the degree of this node of power $\theta, \theta \in [0, \infty)$. We denote a RRT with N nodes and the scaling parameter θ by $T(N, \theta)$. Specifically, $T(N, 0)$ corresponds to a uniform recursive tree (URT) where at each time step, a randomly selected existing node is connected to the newly added node [42, 43]. $T(N, 1)$ is a scale-free tree where at each time step, the probability for an existing node to be connected to the new node is proportional to the degree of this node [44]. When $0 < \theta < 1$ ($\theta > 1$), the probability that an existing node is attached to a new node is sub-linear (super-linear) of the degree of the existing node. When $\theta \to \infty$, the RRT approaches a star topology, whose average path length is $2 - 2/N$ for a star with N nodes.

We conduct 10^4 independent realizations of each RRT class $T(N, \theta)$ with size N and scaling parameter θ, and obtain for each class the average as well as the standard deviation of the two key topological features, i.e. the average path length and the degree variance. As illustrated in Figure 2, a small (large) θ suggests a relative deep (shallow like a star) tree with a large (small) average path length, that corresponds to the viral (broadcast) type of information diffusion.

Figure 2 shows that the average path length and degree variance (standard deviation), in WeChat cascade trees as a function of the tree size can be well captured by the RRT model with the scaling parameter θ around 1.2 if we look at the mean of these two properties. When the standard deviation



(a)



(b)

Figure 2: The average path length and degree standard deviation of the cascade trees in WeChat and the RRT models as a function of tree size. The cascade trees in WeChat are grouped according to their sizes: [100,200), [200,400), [400, 800), [800,1600) etc. The average and standard deviation (error bar) of these two properties are obtained for each group and plotted as a function of the medium size of each group. For a given size of the trees and a given θ, 10^4 RRTs are generated independently and the average and standard deviation (error bar) of the average path length and degree standard deviation are obtained from the 10^4 realizations. The error bar for the two properties are shown for the RRT model with $\theta = 1.2$ and $\theta = 1.6$.

of these properties, i.e. error bar, is taken into account, the WeChat cascade trees can be well described by the RRT model with $\theta > 1$, suggesting that the WeChat cascade trees may follow a growth rule where a high degree node in the tree has a high probability to attract the connection to new nodes. When $\theta = 0$, the average path length $E[H] \sim logN$ scales linearly with the logarithmic of the network size [45]. When θ is positive, the average path

length of RRTs increases first and decreases afterwards as the size of the tree increases. This can be observed evidently in the RRTs when $\theta = 1.2$ in Figure 2. Such transition is due to the fact that as a RRT grows in size with a positive θ, hubs tend to form and have a higher chance to be connected to newly added nodes. Such dominant growth of the hubs or local stars reduces the average path length and increases the degree variance. The average path length starts to decrease at a small tree size when θ is large thus hubs form faster as a tree grows. The average path length in WeChat cascade trees indeed increases first and then decreases as the cascade tree size increases, which can be thus well captured by the RRT model.

The RRT model could be used to model the cascade trees, not limited to WeChat, that have diverse sizes. The parameter θ that best fits the data reflects quantitatively how deep the tree is and how diverse the degrees of the nodes in the tree are. In this way, we could compare different online systems with respect to in which system information propagates more via hubs/broadcasting or viral-like spreading.

4. Modeling of Information Cascade Process

In this section, we aim to develop a stochastic model of the information diffusion process based on our understanding of the WeChat information diffusion mechanisms that is able to reproduce three key features of cascade trees as observed in the WeChat dataset: the distribution of the sizes of the cascade trees, the average path length and the degree variance of a cascade tree in relation to the size of the tree.

4.1. The Susceptible View Forward Removed Model

We propose the Susceptible View Forward Removed (SVFR) model to describe the information diffusion process on a social network. This model is based on classic viral spreading models such as SIR model but more general and practical with respect to the definition of the possible states of a user and the possible non-liner and non-homogeneous probability for a user to view a message shared by its friend.

In the SVFR model, each node can be in one of the following four states at any time step:

- Susceptible (S) - the user has the potential to read a message/content, but has not yet read it,

- View (V) - the user views the message,
- Forward (F) - the user forwards the message,
- Removed (R) - the user ignores the message either because (s)he does not want to read the message or has already viewed or forwarded the message.

For a given message, all the nodes are initially susceptible, except for the source node that publishes/shares this message thus is in state F at step $t = 0$. The state transition diagram has been shown in Figure 3. For any node that is in state F at any time step t, each of its susceptible neighbours in the social network has a probability β to view the message at step $t + 1$. Moreover, each neighbour that views the message has a probability γ to forward the message immediately after reading, and thus transits to state F at step $t + 1$. In other words, each S neighbour has a probability $\beta * (1 - \gamma)$ of being in state V (view but not forward) and a probability $\beta\gamma$ of being in state F (view and forward) and probability $1 - \beta$ of being in state R (ignore the message without reading the content) at time step $t + 1$. For any node in state V or F at any given time, this node will be in state R at the next time step. The diffusion process of a message stops when all the nodes are either in state S or R, thus when the system reaches the stable state.

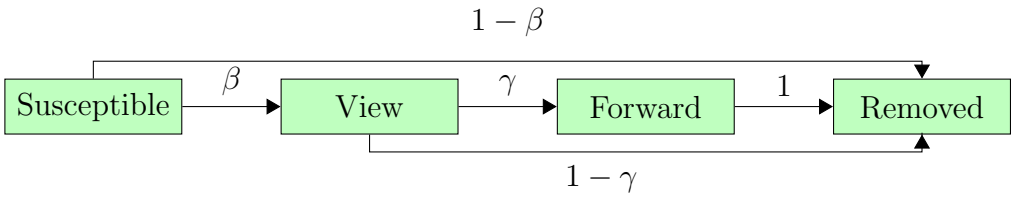


Figure 3: States transition diagram of the SVFR model.

Furthermore, we generalise the SVFR model to be a heterogeneous stochastic model where the probability β that a user reads a message shared by its friend may depend on the degree of this user in the underlying social network. This is motivated by the fact that a node has a large number of friends tends to have a low probability to read a message shared by his/her friend due to the large number of messages he/she is exposed to and his/her limited effort in reading messages [46, 47]. Without loosing generality, we assume that the probability for a node i to read a message shared by a neighbour may depend

on the degree d_i of this node, and we denoted the probability as $\beta_i = cd_i^{-\alpha}$, where the power exponent α is assumed to be positive and the constant c is determined by the given average probability β to view a message over all the nodes [4]:

$$\beta = c \sum_{k=d_{min}}^{d_{max}} k^{-\alpha} Pr[D = k], \tag{3}$$

As observed in the data and assumed in our model, users seldom reads or share a message more than once. The average view probability β suggests how infectious/interesting a message is for users to view it. When $\alpha = 0$, all nodes have the same view probability. Similar homogeneity has been usually assumed in previously proposed information diffusion models [9]. Our heterogeneous model takes into account the possibility that the view probability of each node may be inversely proportional to the degree of the node, characterized by the degree scaling parameter α. Evidence has been found in [14] that the probability a node shares a message may be inversely proportional to the degree of the node thus $\alpha = 1$. Our model is more generalised with respect to its polynomial scaling α and realistic states of user activities, aiming to reproduce several key features of cascades observed in real-world data. In the proposed stochastic model, we did not take into account a realistic and possibly heterogeneous time delay, e.g., between the time when a node shares a message and the time a neighbour reads or shares the message.

We assume that the probability γ that a user forwards a message after viewing it, the so-called forward probability, is a constant, which is a simple start for the model study. Given the underlying social network and given the parameters α, γ and β to be calibrated, the SVFR model could iterate the stochastic propagation of a message, each resulting in a cascade tree composed of users that have created, viewed and forwarded the message.

4.2. Model Validation

The (average) forward probability in a cascade tree can be obtained as the number of nodes that forward the message over the total number of nodes in the cascade. Figure 4 shows that the forward probabilities of the WeChat cascade trees follow approximately a Gaussian distribution where forward

[4] Each node may view a message maximally once.

probabilities are close to the average. Hence, we consider the average forward probability $\gamma = 0.091$ observed in the data as the forward probability in our SVFR model.

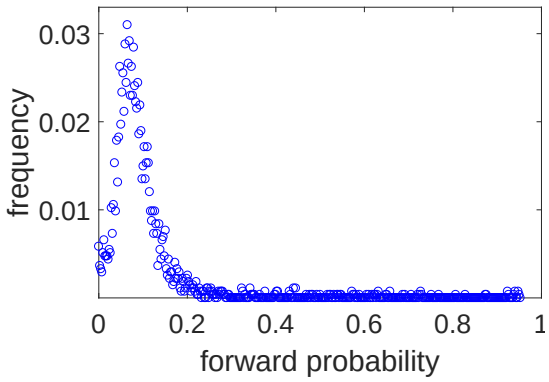


Figure 4: Distribution of the average forward probability in a cascade tree. This distribution is obtained from the WeChat cascade trees that have a size larger than or equal to 100.

The WeChat social network topology is unknown. Hence, we cannot derive directly from the data the two parameters related to the degree dependent view probabilities: the average view probability β and scaling parameter α. Instead, we will explore whether the SVFR information diffusion on a social network model with tunable parameters β and α could reproduce the three key features of the WeChat cascade trees: the size distribution, the average path length and degree variance in relation to the tree size. The distribution of the sizes of the cascade trees is a crucial feature for a online social network, characterizing the distribution of the prevalence or popularity of the information propagated on the network.

We assume that the underlying social network is a scale-free network with a power law degree distribution $Pr[D = k] = ck^{-\phi}$, as observed in many real-world networks [48]. We use the configuration model [49–51] to construct the random scale-free networks with a power exponent of the degree distribution $\phi = 2.5$, a minimum degree $d_{min} = 10$ as in [9] and a cutoff of the maximum degree $d_{max} = N^{1/(\phi-1)}$ [52], where N is the network size. When the network size is $N = 10^5$, the average degree $E[D] \approx 26.7$.

For each given pair of β and α, we generate independently 100 scale-free networks and on each generated network, we carry out the information spread of 100 messages independently according to the SVFR model where the initial

node that creates/shares the message is chosen uniformly at random. In total, we obtain 10^4 cascade trees for the given β and α.

First, we explore the distribution of the sizes of the cascade trees in both the WeChat dataset and in our SVFR model. As shown in Figure 5, the distribution of the sizes of the observed WeChat cascade trees is approximately a power-law distribution. Since we are interested in the cascade trees with a size larger than 100, that corresponds to the messages that could propagate to a certain extend, we fit the tail part of the distribution when the size is larger than or equal to 100. The power exponent is approximately $\lambda = 2.17$. The power-law cascade size distribution has also been observed in other social networks, such as Twitter [7, 8, 24], Flickr [11], Digg [12] and Sina Weibo[13].

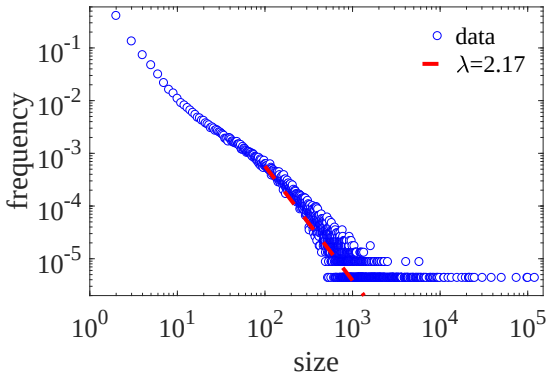


Figure 5: Distribution of the size of the WeChat cascading trees with the curve fitting for the tail where the size is larger than or equal to 100.

We take as an example the SVFR model with the average view probability $\beta = 0.3$ whereas the degree scaling parameter α varies. Figure 6 illustrates how the size distribution of the cascade trees generated by our SVFR model changes as the degree scaling parameter α increases.

When $\alpha = 0$ or α is small , i.e. all the nodes have a similar probability to view a message, the cascade size distribution has a peak in the tail, thus a significantly higher probability to be large. When the view probability β or the network size N increases, the separation between the power law decrease and the peak in the size distribution becomes even more apparent. As α increases, the cascade size distribution becomes a power-law distribution, the same as observed in WeChat. The hubs play a key role in such a change in the size distribution. First, a hub (a high degree node in the underlying scale-

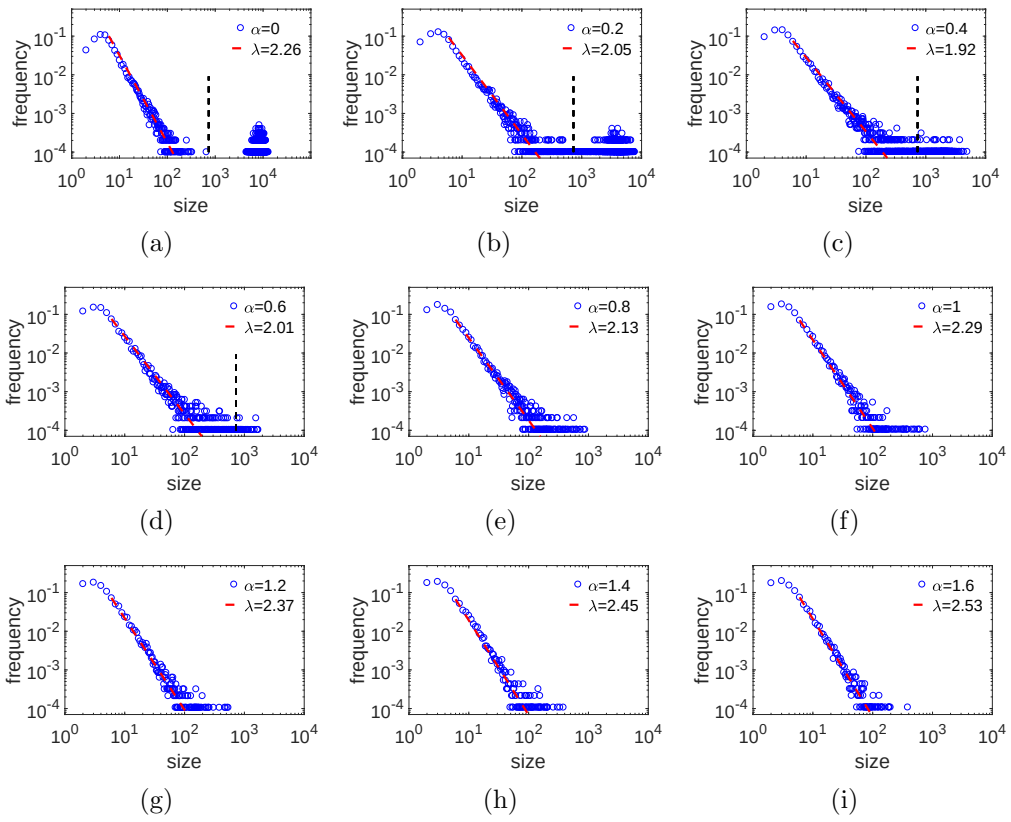
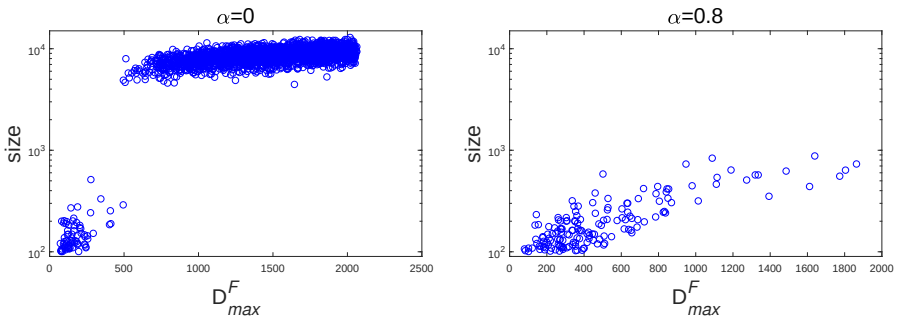


Figure 6: Cascade size distribution of the SVFR model for different degree scaling parameter α. The underlying scale-free network size is $N = 10^5$ and the average view probability is $\beta = 0.3$. The power-law part of the tail has been fitted. Each figure is obtained by 100 independent realizations of the SVFR process on each of the 100 independently generated underlying scale-free networks.

free network) has a higher probability that one of its neighbours forwards the message than low degree nodes. Second, a hub has a higher probability to view thus forward a message when α is smaller, given the same average view probability β. Third, the forwarding of a message by a hub allow its large number of neighbours to further view and forward the message, leading potentially to a large cascade. Hence, hubs facilitate the appearance of large cascades, especially when α is small. This explains as well why the largest possible cascade size decreases as α increases. Figure 7 further supports our explanation. We look into the maximal degree (in the underlying social network) D^F_{max} of nodes that have forwarded the information in a cascade

tree in relation to the size of the cascade. As the D^F_{max} increases, i.e. a higher degree node involves in the forwarding of the message, an abrupt jump occurs in the cascade size, when $\alpha = 0$. Hence, the bulk in the size distribution $\alpha = 0$ corresponds to the large cascades where hubs involve in forwarding the information. When $\alpha = 0.8$, the increase of the cascade size with D^F_{max} is relatively continuous.



(a) (b)

Figure 7: The size of a cascade tree generated by the SVFR model versus the maximum degree D^F_{max} in the underlying social network of the nodes that have forwarded the message in the cascade tree when (a) $\alpha = 0$ and (b) $\alpha = 0.8$. Cascade trees larger than 100 in size are considered.

Figure 6 suggests that α should not be small in order to capture the power-law size distribution in the WeChat dataset. Furthermore, we explore how the power exponent/slope λ of the power-law cascade size distribution generated by the SVFR model is influenced by the size N of the underlying network, the average view probability β and the degree scaling parameter α. As shown in Figure 8, the exponent λ is obtained via the power-law curve fitting of the power-law decreasing part of the size distribution [50].

As shown in Figure 8, power exponent λ is insensitive to the size N of the underlying networks, though the average cascade size may depend on the size of the underlying network. We will focus on the underlying network size $N = 10^5$, which is large as well feasible for simulations. A smaller α and a large average view probability β contribute to a smaller power exponent λ, thus large cascade trees with a higher probability. The power exponent $\lambda = 2.17$ observed in WeChat can be approximated by our SVFR model when $\beta = 0.3$ and $\alpha = 0.8$ or $\beta = 0.4$ and $\alpha = 1.2$ or $\beta = 0.5$ and $\alpha = 1.6$.

Finally, we investigate the average path length and the degree variance of the cascade trees in relation to the cascade tree sizes produced by our SVFR

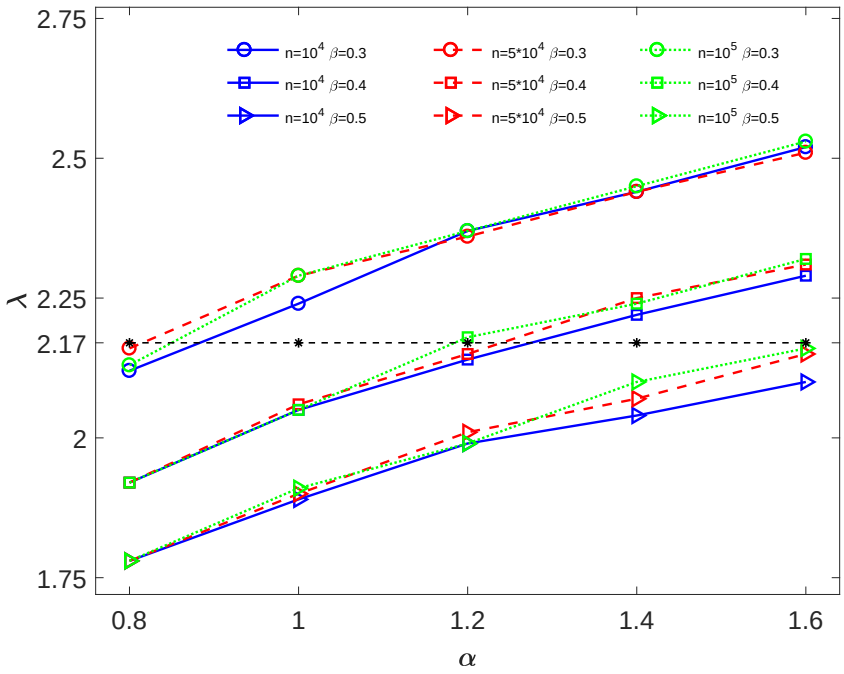


Figure 8: The power exponent λ of the power-law cascade size distribution generated by the SVFR model as a function of the size N of the underlying network, the average view probability β and the degree scaling parameter α. For each set of parameters, the cascade size distribution is obtained from the 100 iterations of the SVFR information spread on each of the 100 independently generated underlying social networks.

model with the aforementioned three sets of parameters that could already well capture the cascade size distribution of WeChat.

Figure 9 shows that the cascade trees generated by the SVFR model with $\beta = 0.3$ and $\alpha = 0.8$ well approximate the cascade trees in WeChat with respect to their average path length and the degree variance/standard deviation. The cascade trees generated by the SVFR, the same as the WeChat cascade trees, are also well bounded by the RRT models with $\theta = 1.2$ and $\theta = 1.6$ and closer to RRT models with $\theta = 1.2$, verifying the consistency of the RRT and SVFR models.

Our SVFR model could well explain the cascade size distribution including the power-law decay exponent, the average path and the degree variance of the cascade trees in WeChat and suggests that a user with a large number of friends may have a lower probability to view the message shared by a friend.

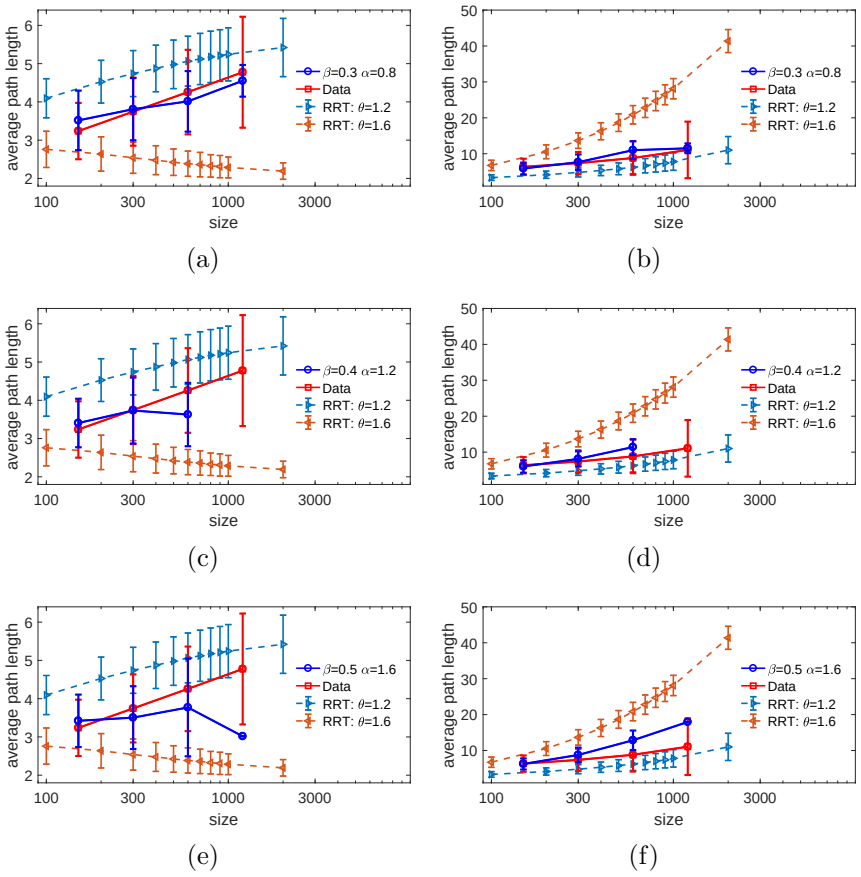


Figure 9: The average path length and degree standard deviation of the cascade trees in WeChat, of the RRT structural model and of the cascade trees generated by the SVFR model. We consider the SVFR model with the three sets of parameters β and α that could well capture the WeChat cascade size distribution. The underlying networks of the SVFR model are scale-free with size $N = 10^5$. Given the parameter β and α, we perform 100 realisations of the SVFR model on each of the 100 independently generated underlying networks leading to 10^4 cascade trees. These cascade trees generated by SVFR are grouped according to their sizes: [100,200), [200,400), [400,800) and [800,1600]. The average and standard deviation of the two key properties are derived for each group and plotted as a function of the medium size of the group. When $\beta = 0.4$ and $\alpha = 1.2$, the cascade trees generated by SVFR model are all smaller than 800 in size. Given the parameter θ and tree size, we carry out 10^4 iterations of generating the cascade trees using the RRT model and obtain the average and standard deviation (error bar) of these two properties.

5. Conclusion

The cascade trees that describe the information spread trajectories in social networks have been widely studied. In this work, we rely on the data extracted from the WeChat social network as a test bed to further advance the information diffusion analysis methods from two aspects.

We propose to model the cascade tree topology by random recursive trees RRTs. The RRT model could well reproduce the tendencies of two fundamental properties of the cascade trees in the WeChat network, i.e. the average path length and the degree variance in relation to the tree size. The identified single parameter θ in the RRT model, allows us, for the first time to quantify how deep (viral like spread) or shallow (broadcast type spread) a class of cascade trees are. Hence, we could compare or classify different online networks regarding to that the information spread on each network is more broadcast or viral like. The RRT model also unravels some interesting phenomena in the cascade-tree growth, like the emergence of hubs.

We introduced the SVFR stochastic model to capture the information diffusion process on a network. The model encodes three types of user reactions to a message they receive: ignore, view and forward the message. We have shown that this model is able to capture three main properties of the WeChat cascade trees: tree size distribution, the average path length and the degree variance of a tree in relation to the size of the tree. The identified model parameters based on the dataset of WeChat cascade trees suggests that a WeChat user with a large number of friends tends to have a low probability to view a message shared by his/her friends. This finding can be supported by the cognitive and biological constraints of users as predicated by Dunbar's theory [46, 47].

The WeChat dataset served as excellent test bed enabling the above mentioned contributions due to the rich user actions it captures and related to the way how users react to the message forwarded to them. We believe, however, that our contributions can serve as a starting point to systematically explore the structure and dynamics of information diffusion in general social networks, not limited to WeChat.

The proposed SVFR stochastic model can be applied to other online social networks as well to explore e.g. whether other types heterogeneity may exist. For example, the view or forward probability of a content may depend on the content. Another promising future research direction is to explore the time delay in the information diffusion model in order to explain e.g. how fast a

message could reach a certain number of users.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Conceived and designed the experiment: LL, BQ, BC, HW. Performed the experiment: LL. Analysed the data: LL, BQ, AH and HW. Wrote the paper: LL and HW. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Lingnan He (The School of Communication and Design, Sun Yat-sen University) and Yichong Bai (Fibonacci Consulting Co. Ltd.) for providing the WeChat dataset. This study is supported by National Key Research & Development (R&D) Plan under Grant No. 2017YFC0803300 and the National Natural Science Foundation of China under Grant Nos. 71673292, 61503402 and Guangdong Key Laboratory for Big Data Analysis and Simulation of Public Opinion.

References

[1] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, B. S. Silvestre, Social media? get serious! understanding the functional building blocks of social media, Business horizons 54 (3) (2011) 241–251.

[2] A. Guille, H. Hacid, C. Favre, D. A. Zighed, Information diffusion in online social networks: A survey, ACM SIGMOD Record 42 (2) (2013) 17–28.

[3] J. A. Obar, S. S. Wildman, Social media definition and the governance challenge-an introduction to the special issue, Available at SSRN 2663153.

[4] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, Y.-C. Zhang, Dynamics of information diffusion and its applications on complex networks, Physics Reports 651 (2016) 1–34.

[5] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media?, in: Proceedings of the 19th international conference on World wide web, ACM, 2010, pp. 591–600.

[6] E. Bakshy, J. M. Hofman, W. A. Mason, D. J. Watts, Everyone's an influencer: quantifying influence on twitter, in: Proceedings of the fourth ACM international conference on Web search and data mining, ACM, 2011, pp. 65–74.

[7] R. A. Baños, J. Borge-Holthoefer, Y. Moreno, The role of hidden influentials in the diffusion of online information cascades, EPJ Data Science 2 (1) (2013) 1.

[8] I. Taxidou, P. M. Fischer, Online analysis of information diffusion in twitter, in: Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, pp. 1313–1318.

[9] S. Goel, A. Anderson, J. Hofman, D. J. Watts, The structural virality of online diffusion, Manage Sci 62 (1) (2015) 180–196.

[10] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, The role of social networks in information diffusion, in: Proceedings of the 21st international conference on World Wide Web, ACM, 2012, pp. 519–528.

[11] M. Cha, A. Mislove, 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, ACM, 2009, pp. 721–730.

[12] R. Ghosh, K. Lerman, A framework for quantitative analysis of cascades on networks, in: Proceedings of the fourth ACM international conference on Web search and data mining, ACM, 2011, pp. 665–674.

[13] P. Bao, H.-W. Shen, W. Chen, X.-Q. Cheng, Cumulative effect in information diffusion: empirical study on a microblogging network, PloS one 8 (10) (2013) e76027.

[14] L. Feng, Y. Hu, B. Li, H. E. Stanley, S. Havlin, L. A. Braunstein, Competing for attention in social media under information overload conditions, PloS one 10 (7) (2015) e0126090.

[15] Y. Li, M. Qian, D. Jin, P. Hui, A. V. Vasilakos, Revealing the efficiency of information diffusion in online social networks of microblog, Information Sciences 293 (2015) 383–389.

[16] R. Wang, S. Rho, B.-W. Chen, W. Cai, Modeling of large-scale social network services based on mechanisms of information diffusion: Sina weibo as a case study, Future Generation Computer Systems.

[17] B. Zhang, Z. Qian, S. Lu, Structure pattern analysis and cascade prediction in social networks, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2016, pp. 524–539.

[18] A. L. Hughes, L. Palen, Twitter adoption and use in mass convergence and emergency events, International Journal of Emergency Management 6 (3-4) (2009) 248–260.

[19] A. M. Kaplan, M. Haenlein, Users of the world, unite! the challenges and opportunities of social media, Business horizons 53 (1) (2010) 59–68.

[20] H. H. Khondker, Role of the new media in the arab spring, Globalizations 8 (5) (2011) 675–679.

[21] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, W. Quattrociocchi, The spreading of misinformation online, Proceedings of the National Academy of Sciences 113 (3) (2016) 554–559. doi:10.1073/pnas.1517441113.

[22] H. Wiener, Structural determination of paraffin boiling points, Journal of the American Chemical Society 69 (1) (1947) 17–20.

[23] C. Li, H. Wang, W. de Haan, C. J. Stam, P. V. Mieghem, The correlation of metrics in complex networks with applications in functional brain networks, Journal of Statistical Mechanics: Theory and Experiment 2011 (11) (2011) P11018.
URL http://stacks.iop.org/1742-5468/2011/i=11/a=P11018

[24] S. Goel, D. J. Watts, D. G. Goldstein, The structure of online diffusion networks, in: Proceedings of the 13th ACM conference on electronic commerce, ACM, 2012, pp. 623–638.

[25] A. Anderson, D. Huttenlocher, J. Kleinberg, J. Leskovec, M. Tiwari, Global diffusion via cascading invitations: Structure, growth, and homophily, in: Proceedings of the 24th International Conference on World Wide Web, ACM, 2015, pp. 66–76.

[26] G. Bounova, O. de Weck, Overview of metrics and their correlation patterns for multiple-metric topology analysis on heterogeneous graph ensembles, Physical Review E 85 (1) (2012) 016117.

[27] A. Rudas, B. Tóth, B. Valkó, Random trees and general branching processes, arXiv preprint math/0503728.

[28] P. L. Krapivsky, S. Redner, Organization of growing random networks, Physical Review E 63 (6) (2001) 066123.

[29] J. Goldenberg, B. Libai, E. Muller, Talk of the network: A complex systems look at the underlying process of word-of-mouth, Marketing letters 12 (3) (2001) 211–223.

[30] D. J. Watts, A simple model of global cascades on random networks, Proceedings of the National Academy of Sciences 99 (9) (2001) 5766–5771–559.

[31] M. Granovetter, Threshold models of collective behavior, American journal of sociology (1978) 1420–1443.

[32] Q. Li, L. A. Braunstein, H. Wang, J. Shao, H. E. Stanley, S. Havlin, Nonconsensus opinion models on complex networks, Journal of Statistical Physics 151 (1) (2013) 92–112. doi:10.1007/s10955-012-0625-4.

[33] B. Qu, Q. Li, S. Havlin, H. E. Stanley, H. Wang, Nonconsensus opinion model on directed networks, Phys. Rev. E 90 (2014) 052811. doi:10.1103/PhysRevE.90.052811.
URL http://link.aps.org/doi/10.1103/PhysRevE.90.052811

[34] H. W. Hethcote, The mathematics of infectious diseases, SIAM review 42 (4) (2000) 599–653.

[35] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Physical review letters 86 (14) (2001) 3200.

[36] M. E. Newman, Spread of epidemic disease on networks, Physical review E 66 (1) (2002) 016128.

[37] J. Yang, J. Leskovec, Modeling information diffusion in implicit networks, in: 2010 IEEE International Conference on Data Mining, IEEE, 2010, pp. 599–608.

[38] Tencent, Tencent announces 2016 second quarter and interim results (2016).
URL http://www.tencent.com/en-us/ir/news/2016.shtml

[39] Z. Li, L. Chen, Y. Bai, K. Bian, P. Zhou, On diffusion-restricted social network: A measurement study of wechat moments, arXiv preprint arXiv:1602.00193.

[40] J. Cheng, L. A. Adamic, J. M. Kleinberg, J. Leskovec, Do cascades recur?, in: Proceedings of the 25th International Conference on World Wide Web, WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2016, pp. 671–681. doi:10.1145/2872427.2882993.
URL https://doi.org/10.1145/2872427.2882993

[41] J. Kunegis, M. Blattner, C. Moser, Preferential attachment in online networks: Measurement and explanations, in: Proceedings of the 5th Annual ACM Web Science Conference, ACM, 2013, pp. 205–214.

[42] C. Su, Q. Feng, Z. Hu, Uniform recursive trees: Branching structure and simple random downward walk, Journal of mathematical analysis and applications 315 (1) (2006) 225–243.

[43] P. Van Mieghem, Performance analysis of complex networks and systems, Cambridge University Press, Cambridge, United Kingdom, 2014.

[44] G. Szabó, M. Alava, J. Kertész, Shortest paths and load scaling in scale-free trees, Physical Review E 66 (2) (2002) 026101.

[45] S. Wagner, On the wiener index of random trees, Discrete Mathematics 312 (9) (2012) 1502 – 1511, recent Trends in Graph Theory and Combinatorics. doi:https://doi.org/10.1016/j.disc.2011.05.008.
URL http://www.sciencedirect.com/science/article/pii/S0012365X11002159

[46] R. Dunbar, Neocortex size as a constraint on group size in primates, Journal of Human Evolution 22 (6) (1992) 469 – 493. doi:http://dx.doi.org/10.1016/0047-2484(92)90081-J.
URL http://www.sciencedirect.com/science/article/pii/0047248492900 81J

[47] B. Gonçalves, N. Perra, A. Vespignani, Modeling users' activity on twitter networks: Validation of dunbar's number, PLOS ONE 6 (8) (2011) 1–5. doi:10.1371/journal.pone.0022656.
URL http://dx.doi.org/10.1371%2Fjournal.pone.0022656

[48] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, science 286 (5439) (1999) 509–512.

[49] M. E. Newman, Power laws, pareto distributions and zipf's law, Contemporary physics 46 (5) (2005) 323–351.

[50] A. Clauset, C. R. Shalizi, M. E. Newman, Power-law distributions in empirical data, SIAM review 51 (4) (2009) 661–703.

[51] J. M. Hernandez, T. Kleiberg, H. Wang, P. V. Mieghem, A qualitative comparison of power law generators, in: International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2007), 2007.

[52] R. Cohen, K. Erez, D. Ben-Avraham, S. Havlin, Resilience of the internet to random breakdowns, Physical review letters 85 (21) (2000) 4626.