Soc. Netw. Anal. Min. (2013) 3:341–357
DOI 10.1007/s13278-012-0082-8

ORIGINAL ARTICLE

# Modeling and evaluating information propagation in a microblogging social network

**Cheng-Te Li · Tsung-Ting Kuo · Chien-Tung Ho · San-Chuan Hong · Wei-Shih Lin · Shou-De Lin**

Received: 22 November 2011 / Revised: 2 August 2012 / Accepted: 11 August 2012 / Published online: 5 September 2012


**Abstract** Microblogging platforms, such as Twitter and Plurk, allow users to express feelings, discuss ideas, and share interesting things with their friends or even strangers with similar interests. With the popularity of microblogs, there are growing data and opportunities in understanding information propagation behaviors in online social networks. Though some influence models had been proposed based on certain assumptions, most of them are based on the simulation approach (not data driven). This paper aims at designing a framework to model, measure, evaluate, and visualize influence propagation in a microblogging social network. Considering how information contents are spread in a social network, we devise two influence propagation models from the views of messages posted and responded. Based on the proposed models, we are able to measure the influence capability of an individual with respect to a user-given topic. Our design of influence measures consider (a) the number of people influenced, (b) the speed of propagation, and (c) the geographic distance of the propagation. To test the effectiveness of our influence model, we further propose a novel evaluation framework that predicts the propagation links and influential nodes in a real-world microblogging social network. Finally, we develop an online visualization system allowing users to explore the information propagation with the functions of displaying propagation structures, influence scores of individuals, timelines, and the geographical information for any user-query terms.

**Keywords** Information propagation · Social networks · Influence analysis · Diffusion models

## 1 Introduction

Microblogs differ from traditional blogs in that their messages are typically shorter in terms of size. The usage of microblogging has become more popular with the emergence of Tumblr[1] and Twitter[2]starting from 2006. Since then, several online microblogging services had been established, such as Plurk,[3] Squeelr,[4] and Jaiku.[5] These services generated tremendous amounts of microblogging data, which have become important resources for social network researchers. Although each microblog service has its own interface and functions, they all share one important feature, namely, the embedded timeline, either explicit or implicit, that records each message's posting and replying time stamp. The update cycles of other online data sources, such as Web pages and blogs, are measured in days, weeks, or even months. In contrast, microblogs are updated every few minutes or even seconds, so that they can be regarded as real-time services. As a result, microblogging data have become a feasible resource for studying the dynamic nature of information, in particular how

C.-T. Li · T.-T. Kuo · C.-T. Ho · S.-C. Hong · W.-S. Lin · S.-D. Lin
Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan

S.-D. Lin (✉)
Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
e-mail: sdlin@csie.ntu.edu.tw

[1] Tumblr: http://www.tumblr.com.
[2] Twitter: http://www.twitter.com.
[3] Plurk: http://www.plurk.com.
[4] Squeelr: http://www.squeelr.com.
[5] Jaiku: http://www.jaiku.com.

information is propagated and distributed in a large social network.

The idea of social-based information propagation can be traced back to the concept of *viral marketing*,[6] which considers that ideas are spread by the "word of mouth" effect in the marketplace. Besides, Rushkoff (1994) and Scott (2011) claimed that information and influence can be propagated even more powerfully via the Internet than by traditional channels. The argument is even more pertinent now, thanks to the emergence of social networking services. Many advertisers and politicians take advantage of Face book, Twitter, and other social networking services to promote themselves and to enhance their influence.

In recent years, a number of microblogging services provide APIs that enable developers to extract the content of messages and obtain information about the social connections between users. Sun et al. (2009) use Facebook data to determine the correlations between different communities; and Kwan et al. (2010) use Twitter data to construct the relationships between people for analyzing user behaviors in online microblogging social networks. On the other hand, many recent studies have focused on devising models to simulate influence propagation behaviors and are evaluated using certain characteristics of the real propagation phenomenon (Gupte et al. 2009; Kempe et al. 2003; Kermack and McKendrick 1927; Ma et al. 2008). For example, the Independent Cascade and the Linear Threshold models (Kempe et al. 2003) are proposed and evaluated using the target-set-size to coverage curve, while the Greedy and the Courteous models (Gupte et al. 2009) are proposed and evaluated using the user-threshold to coverage curve (user-threshold is a parameter of Greedy and Courteous model). The performance of different influence propagation models can hardly be compared without a unified evaluation metric. Consequently, given certain type of diffusion data, it is difficult to decide which influence propagation model is the most suitable one with corresponding parameters.

This paper proposes a framework to model, measure, and evaluate the propagation of information in a microblogging social network. We also aim at developing a real-world online demo system that allows users to analyze and visualize the diffusion of certain topics of interest. Specifically, we consider the following four tasks. (1) *Modeling*: given a certain topic that has been discussed in a microblogging platform, how do we model the information propagation of such topic? (2) *Measuring*: how do we measure the propagation or influence capability of an individual to disseminate a certain topic in a microblogging social network and how do we determine the extent of diffusion for a particular topic in a microblogging social network? (3) *Evaluating*: how do we predict and compare the propagation capability and propagation links of different models given a specific topic to propagate? (4) *Visualizing*: how do we visualize information propagation in a microblogging social network? We believe these four tasks are essential because of several reasons, as listed in the following.

1. From a research perspective, academics in different disciplines (e.g., sociology and mass communications) are interested in how information is spread among people, and are urgently in need of a platform or tool to perform experiments. Microblogging services provide a feasible environment for such studies.
2. Being able to quantify how specific information is propagated in a social network can facilitate social science research, such as to study how information influences the evolution of a society and how a burst of popularity can arise for a particular product.
3. From the perspective of applications, knowing how to identify individuals with the ability to propagate certain ideas or influence can benefit the recommendation of products in online social networks. For example, companies could take advantage of influential people as seed candidates to perform marketing and advertising campaigns more effectively; governments can effectively broadcast emergency announcements.

We summarize the contributions of this work.

- Technically, we propose an information propagation model to capture how information is spread and measure the capability of users to propagate a certain topic in a microblogging social network. We define a loose and a rigid methods to model how influence is spread over individuals in a network. Our measures consist of three factors, the number of people being influenced, the speed of propagation, and the geographic range of propagation. Based on the proposed influence measures, we are able to determine the extent of spread of a query term in a microblogging social network. Such method allows us to quantify and rank influence scores for different query topics in microblogs.
- Empirically, we propose EPIC, a general evaluation framework, to assess the performance of different information diffusion models. We devise two prediction schemes, including propagation links prediction and propagation capability prediction, with the evaluation strategies of one-by-one and leave-one-out flows, in the EPIC framework.
- We develop a visualization framework with an online search-based service that implements the proposed

[6] Viral Marketing: http://en.wikipedia.org/wiki/Viral_marketing.

methods for demonstration purposes. Given a term as the query topic, our system automatically reports the top influential microbloggers who play critical roles in disseminating the topic in the network, as well as displays different kinds of propagation scores of each user. The demo video of our system is available at http://tinyurl.com/plurpagation.

## 2 Related works

*Model-based influence propagation* Richardson and Domingos (2002) proposed a probabilistic method for extracting information from a knowledge-sharing network and put forward a hypothesis about the most effective individuals for viral marketing. Subsequently, Kempe et al. (2003) proposed a model to maximize the influence of a social network. First, they showed that finding the most influential people is an NP-hard problem. Then, they proposed two models, the Linear Threshold Model and the Independent Cascade Model, and used them to simulate information propagation in a social network. Meanwhile, Gruhl et al. (2004) developed a propagation model for blogs based on the theory of infectious diseases; while Song et al. (2007) proposed a method to predict the target flow of information and how long it takes for a user to obtain new information. The major difference between the above works and our approach is that they focus on developing models to simulate, predict, or explain information propagation, rather than measuring and quantifying information propagation in a real-world microblog environment.

*Information propagation on real data* With the increasing availability of social network data in recent years, researchers have applied different models to analyze the data. Cha et al. (2009) exploited Flickr data to construct the relationships between photos and the photographers. They also tried to determine how widely information can be spread and what role word of mouth plays in such a network. Sun et al. (2009) investigated the propagation phenomenon of Facebook's News Feed, and created a social network based on users and fans of Facebook for analysis. Their objective was to observe the relationships between different kinds of propagation communities and determine if several short diffusion chains tend to merge together. They exploited zero-inflated negative binomial regressions to model the phenomenon. Kwak et al. (2010) used the relationships between the follower and following in Twitter to construct a social network for advanced analysis. Their results indicate that there exists a gap in influence inferred from the number of followers and that from the popularity of one's tweets. Goyal et al. (2010) propose a parameterized model to learn and predict the time required by a user to perform an action. They used Flickr data to verify the model's accuracy. Sakaki et al. (2010) proposed a real-time event detection system using Twitter data. They regard Twitter users as event sensors and use the messages posted by the sensor users to train the features set by their model and analyze the location and temporal information for some events. The authors apply their system to detect earthquakes and typhoons. Similarly, other works (Lampos and Cristianini 2010; Lampos et al. 2010) are done to track epidemics by monitoring the message flows in Twitter. The above works use microblogging data for certain kinds of analysis, but they do not consider the issues related to measuring the scale and speed of propagation. On the other hand, some works about information propagation attempt to identify influential users from real data (Cha et al. 2010; Song et al. 2007; Tang et al. 2010; Yang et al. 2010), select an initial user set with maximum influences given a diffusion model (Kempe et al. 2003, 2005; Kimura et al. 2010, discover the diffusion patterns (Leskovec et al. 2006), recognize diffusion sequences (Stewart et al. 2007), manage node disappearance (Sarr and Missaoui 2012), sample networks to identify influential users (Maiya and Berger-Wolf 2010), minimize the propagation budgets and time (Goyal et al. 2012; Zhang et al. 2012), find influential links (Bakshy et al. 2012), recommend propagation links to boost content spread (Chaoji et al. 2012), and construct the underlying diffusion network given historical diffusion records (Rodriguez et al. 2010). However, among these researches, few attempt to evaluate the capabilities of different influence propagation models. Furthermore, some works (Rodriguez et al. 2010; Snowsill et al. 2011; Bakshy et al. 2012) aim to analyze and infer the diffusion network given diverse kinds of propagation observations, for example, the plain text (Snowsill et al. 2011; Cha et al. 2012) and information entropy (Steeg GV Galstyan 2012).

## 3 Modeling and measuring information propagation

Ideas are shared and propagated frequently and widely in microblogs. In this paper, a person who posts messages about a topic is called the *topic propagator* and people who receive the information are called *topic receivers*. If a receiver then disseminates the information to his friends, he becomes the next propagator and we say that the information has been propagated to the second-step neighbors (usually denoted as $L_2$ neighbors in the literature). For example, in Twitter, one of the most popular microblog services, the function "retweet" can be regarded as a kind of information propagation clue because a user simply reposts an original message from someone else's page to his own page. Other social networking services, such as

Facebook and Plurk, provide similar functions "share", "reply", and "re-plurk". It is common for a person to post a friend's message on his own board in order to share it with other friends. Based on the above observation, we propose some simple yet intuitive methods to measure information propagation for a given user or a specified query term in a microblogging platform.

### 3.1 Model overview

Our system framework is shown in Fig. 1. First, the user inputs the query term Q and the time period of interest P. Given the input constraints, the system identifies a set of relevant posts and the corresponding replies, as well as the time stamps on them. Then, based on the above information, two kinds of inference trees are defined and constructed for each user: the rigid lower-bound influence (LBI) tree and a loose upper-bound influence (UBI) tree. For each tree, it is possible to produce three kinds of propagation values to characterize the diffusion capability: (1) the number of individuals influenced by the propagated information, (2) the speed of propagation, and (3) the spread geographic distance of the propagation. After aggregating the propagation scores of all the users, we can determine the extent of propagation for the query.

### 3.2 Influence tree construction

For a given query term Q, we construct a set of influence trees of users to model the propagation in a microblog platform. Before discussing the tree construction, we start from defining some key terms.

**Definition 1** (*microblog corpus*) A microblog corpus C comprises three kinds of entities: users, posts, and replies of posts. A corpus C contains a set of users U. Each user has a set of posts and each post might contain a set of reply messages. All posts and replies are associated with a time stamp.

**Definition 2** (*relevant microbloggers*) Given a query term Q and a specific time period P, we define relevant microbloggers with respect to Q as a set of users $U_Q = \{u_1, u_2,\ldots,u_n\}$ who satisfy the two requirements: (1) $u_i$ has a set of posts $A_i = \langle a_{i,1}, a_{i,2},\ldots,a_{i,k}\rangle$ ($k \geq 1$) which contains the query term Q, and (2) the time stamp of each $a_{i,k} \in A_i$ is within the specified period P. We also associate each $u_i \in U_Q$ with a time stamp considering the first message that $u_i$ posts about Q.

Note that to find those messages of posts related to the query term Q, we currently use simply term matching. However, the term matching would result in noisy collections of relevant microbloggers and posts and degrade the quality when the query term is ambiguous. There have already several well-known technologies about Name Entity Recognition (NER) to have some unambiguous name entities as query terms. Here, we leave such issue as an independent component of our system and focus on the modeling of information propagation.

To capture the information spread about a query term Q, we devise two propagation models, based on a rigid and a loose relationship of propagation.

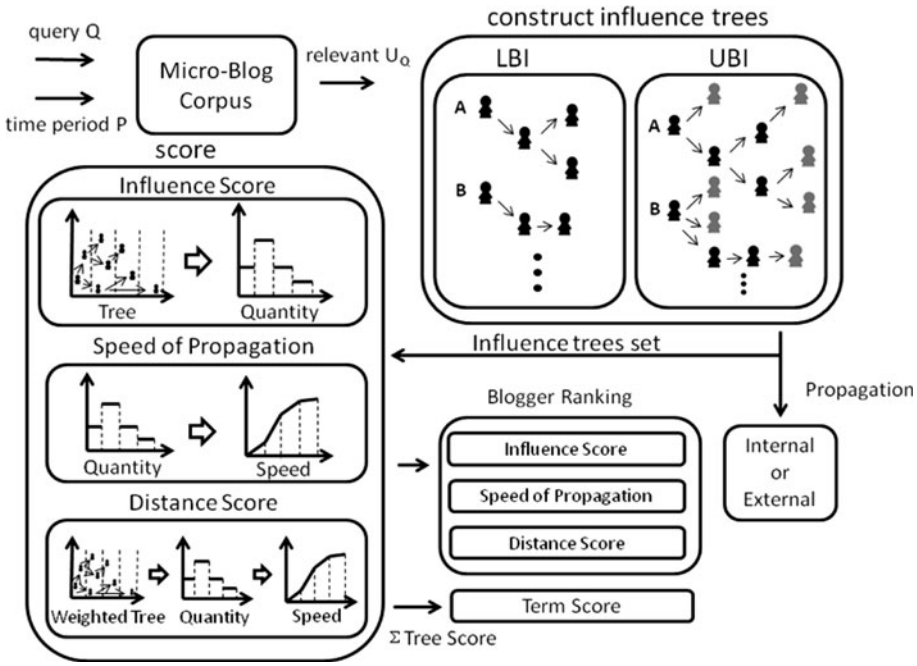


**Fig. 1** Overview of the proposed influence models and measures

- *Rigid-propagation relationship*: If X posts a message that contains Q, and an individual Y not only replies to the message and but also posts another message relevant to Q after his reply to X, we say the concept Q has been propagated from X to Y. Consequently, we assume that there is a rigid-propagation relationship between X and Y. If two individuals X and Y satisfy the above conditions, we assume that there is a rigid-propagation relationship between them. We consider that message propagation involves two actions: *receiving* and *distributing*. In other words, if Y replies to X, we consider that the message has been received by Y. If Y subsequently posts a similar message, then we say that the message has been distributed by Y. Note that if Y had posted about the same topic as Q before he replies to X, we will not consider that Y is influenced by X, because Y could be influenced by some others or external sources at an earlier time. This idea can be illustrated in Fig. 2a, b.
- *Loose-propagation relationship*: If X posts a message containing Q and an individual Y is the first to reply to the message, we can also consider that Q has been propagated from X to Y; and we assume that there is a loose-propagation relationship between X and Y. It is also possible that Y also replies to a post from Z on the same topic. In this case, the time stamps of the replies will decide who influenced whom. Assume that Y replies to X's post before he replies to Z's post; we will consider Y as being only influenced by X. In other words, for a certain topic, a person Y is only influenced by one single person X, depending on whose post receives an earlier response. Such an idea is illustrated in Fig. 2c, in which the solid red line indicates Y's replies to X at 03:00 AM (i.e., Y is influenced by X) and the dotted gray line indicates Y's replies to X at 09:00 AM (i.e., Y is considered to be not influenced by X). It is less rigid than the previous relationship of propagation because it does not need to satisfy the *re-post* condition. Here, we still assume that a propagation event can be decomposed into receiving and distributing actions. In some microblog systems, people can view the replies to posts; therefore, Y's reply to X's post can be viewed by Y's friends. To a certain extent, this can be regarded as propagating the information.

Based on these two kinds of propagation relationships, we define two influence trees to model the information propagation in a microblogging social network, as shown in Fig. 3. Both of the influence models are defined on one user, to characterize the information propagation about a certain topic starting from such a user. We will formally introduce two such definitions of influence flows in the following.

**Definition 3** *UBI* (*upper-bound influence*) *tree* An upper-bound influence tree of a person $u_i$ is a tree structure rooted at $u_i$. Each edge in the tree represents a loose-propagation relationship from the parent node to the child node.

**Definition 4** *LBI* (*lower-bound influence*) *tree* A lower-bound influence tree of a person $u_i$ is a tree structure rooted at $u_i$. Each edge in the tree represents a rigid-propagation relationship from the parent node to the child node.

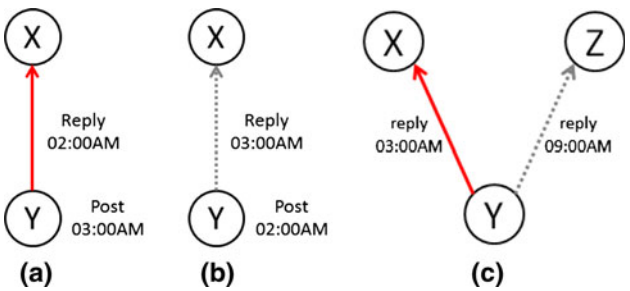


**Fig. 2** **a** Y replies to X and then posts the relevant message afterward. We consider Y is influenced by X in this case. **b** Y replies to X after Y had posted the relevant message. In such case, Y is not considered to be influenced by X. **c** Assume Y has no messages about the query topic, and Y has multiple relies (to X and Z with different time stamps of replies). In this case, we consider Y is influenced by X because it is the first time for Y to view such a topic (color figure online)

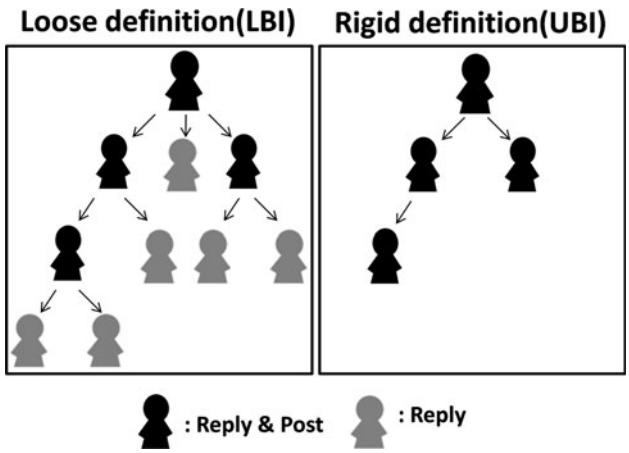


**Fig. 3** The proposed two influence trees. *Left* upper-bound influence (*UBI*) tree, which considers the propagation from one post to some responses with similar intents as a kind of information propagation. *Right* lower-bound influence (*LBI*) tree, which considers only the spread of information posts from one user to another as a kind of information propagation

Each tree represents a propagation structure starting from the root node. Clearly, the LBI tree is a subgraph of the UBI tree. We hypothesize that the true propagation pattern is bounded by these two trees. That is to say, the true propagation structure starting from the root is a *subgraph of the UBI tree* and *a super graph of the LBI tree*. In the realistic context, the information diffusion among users is a mix of the UBI and LBI propagations. Some microbloggers passively receive the information by simply replying the post of interest. But some actively disseminate what they feel interested to their friends. Therefore, we believe in the realistic information propagation, some influence will happen in LBI, while others will be modeled in UBI. Technically, the LBI and UBI trees can be constructed using BFS-like search method on the microblog data, as elaborated in Algorithm 1.

**Algorithm 1. Influence Tree Set Construction**

**Input:** A microblog corpus C; a query topic Q;
a specified influence relationship LBI or UBI; a time period P.
**Output:** an influence-propagation forest (i.e., a set of trees): $F=\{T_1, T_2,...,T_n\}$.

```
Based on query Q and time period P,
retrieve the set of relevant users U_Q from the corpus C.
for each u_i ∈ U_Q do
  T_i = T_i ∪ u_i.
  Enqueue(Queue, u_i).
  while size(Queue) ≠ 0 do
    u_x = Dequeue(Queue).
    for each u_y ∈ {U_Q \ T_i} do
        if u_x and u_y confirm the specified influence relationship do
            T_i = T_i ∪ u_y.
            Construct a link between u_x and u_y in T_i.
            Enqueue(Queue, u_y).
```

## 3.3 Measuring user informaiton propagation by influence trees

To measure the capability of propagation for a user, we consider three factors given an influence trees: (1) the number of people influenced, (2) the speed of propagation, and (3) the geographic distance of propagation.

### 3.3.1 *Scale of propagation*

To quantify the scale of propagation, we can count the total number of people (i.e., nodes) in the corresponding influence tree (Fig. 4). The higher the number, the greater will be the scale of the propagation, as shown by the two-dimensional diagram in Fig. 4. The horizontal axis represents the sequence of time stamps $\{t_1, t_2,\ldots,t_m\}$, and the vertical axis represents the number of people influenced during the specified time period. Note that each person should have two scores, one from the LBI tree and the other from the UBI tree. The LBI tree score is subject to tighter constraints; therefore, it represents the lower-bound value of the propagation. The UBI tree score represents the upper-bound value of the propagation.

### 3.3.2 *Speed of propagation*

Besides the amount of propagation, we are interested in the speed of propagation. A person who is capable of affecting a large number of people in a short time is considered as a strong candidate for disseminating information. Based on the constructed influence trees and the time stamp of each entity, we propose a method for estimating the propagation speed of a message sent by the person at the root node. We use Fig. 5 to illustrate our idea. The left-hand figure, which is similar to the one in Fig. 4, captures each propagation status over time. We aggregate the number of individuals influenced over time to produce the graph on the right-hand side of Fig. 5. Then, we use the area under curve to represent the speed of propagation. Intuitively, the higher the propagation speed, the larger will be the number of people influenced in a short time. This would cause the curve to reach a higher level earlier and therefore increase the area under the curve. Similarly, each root node (person) has the alternatives of LBI and UBI to measure the speed of propagation. Note that one may think the slope of the accumulated users should be a more suitable representation of measuring the speed. In fact, consider the fact that the

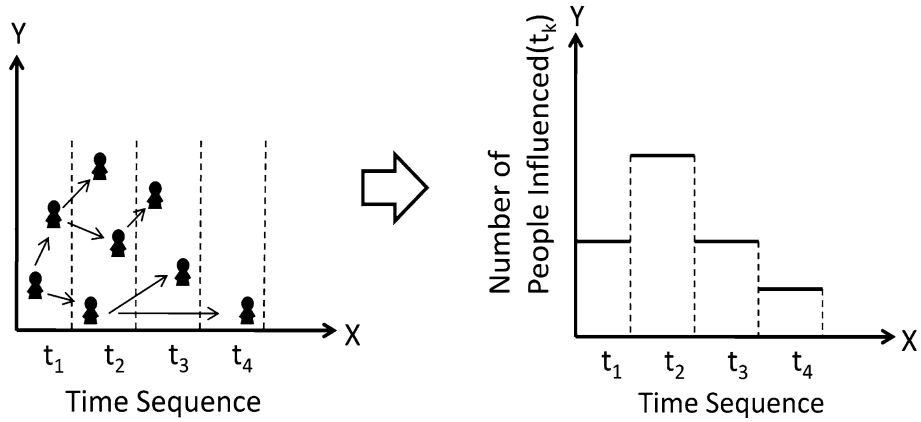


**Fig. 4** The distribution of people influenced during a specific period of time

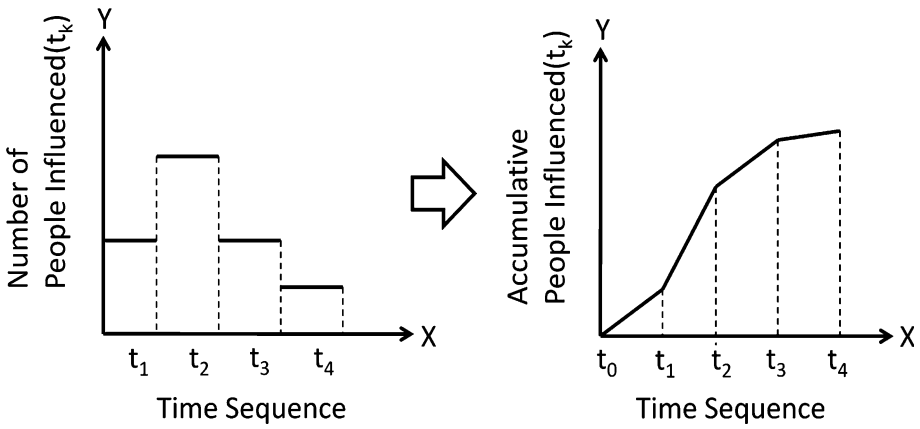


**Fig. 5** The distribution of accumulative people influenced during the time

slopes of the accumulated users change over time; it is hard to use one single slope to capture the rate. In addition, the area under the curve is essentially a weighted average of the slopes over time. Thus, it is positively correlated with the rate. To compare the effectiveness of two diffusion algorithms, we feel that the weighted slope over time is more suitable than a single slope at a specific time.

### *3.3.3 Distance of propagation*

Geographic information enables us to observe whether a message is disseminated globally or locally around the world. In this section, we propose a method for measuring the propagation in terms of the geographic distance. An intuitive way to determine the propagation distance is to calculate the total distance starting from the root user to those people he/she influences directly or indirectly. This can be achieved by a microblogging service as long as the location information, such as city and country of the users, is provided. We can pinpoint a location by its longitude and latitude coordinates and calculate the geographic distance between two geocodes. We use an *x*–*y* diagram, as shown in Fig. 6, to illustrate the geographical spread. The horizontal axis stands for the time periods, while the vertical axis represents the total distance starting from the root user to those people influenced in a specific time period. The distance between the root user and each influenced user is considered in this measure. For each time slot $t_i$ to $t_{i+1}$, we sum up the total geographical distance starting from the root user to each influenced user in such period, and plot the derived value on the right-hand side of Fig. 6. Consequently, we can sum up the geographical distance in each time period to be the resulting score of the distance of propagation. Note that either LBI or UBI model can serve as the constraint in the propagation measurement.

## 3.4 Measuring term information propagation by influence trees

Clearly, different kinds of message can be propagated differently in a microblog. Some topics (e.g., information about a natural disaster) are likely to be propagated more rapidly than others. In the following, we propose a method for measuring the level of propagation of a concept or term in a microblog. First, we define the concept seed users.

**Definition 5** (*seed users*) Given a query Q and a time period P, we define seed users as a set of users $U_S = \{u1,\ldots,us\}$ who satisfy the following requirements: (1) $u_i$ is a relevant user who posted a message about Q; and (2) $u_i$ is not the *topic receiver* in a loose- or rigid-propagation relationship.

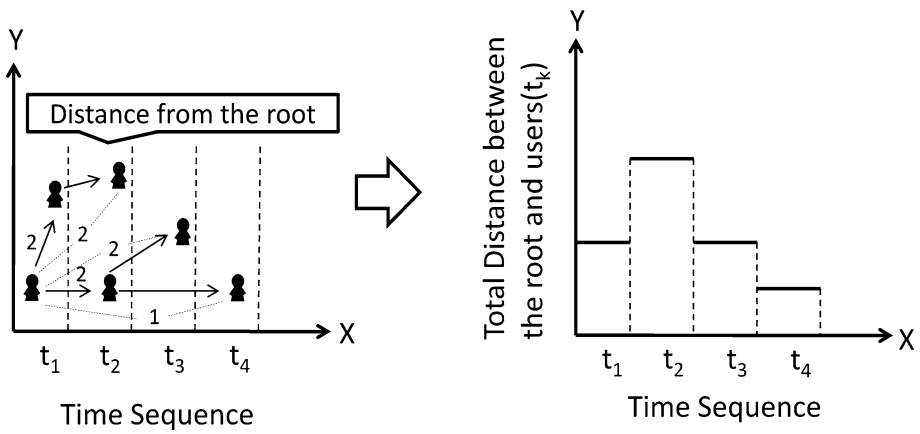


**Fig. 6** The distribution of the total geographical distance between the root user and influenced users

**Table 1** Formulas to compute six kinds of term influence scores

| | Lower-bound influence | Upper-bound influence |
|---|---|---|
| Propagation score | $\sum_{i=1 \cap i \in Seed}^{n} \mathrm{LBI}(\mathrm{PScore_i})$ | $\sum_{i=1 \cap i \in Seed}^{n} \mathrm{UBI}(\mathrm{PScore_i})$ |
| Propagation speed | $\sum_{i=1 \cap i \in Seed}^{n} \mathrm{LBI}(\mathrm{PSpeed_i})$ | $\sum_{i=1 \cap i \in Seed}^{n} \mathrm{UBI}(\mathrm{PSpeed_i})$ |
| Distance score | $\sum_{i=1 \cap i \in Seed}^{n} \mathrm{LBI}(\mathrm{DScore_i})$ | $\sum_{i=1 \cap i \in Seed}^{n} \mathrm{UBI}(\mathrm{DScore_i})$ |

Seed users play the role of message sources, i.e., they initiate the propagation of messages. They are the topic providers in the microblog. To quantify the level propagation of a concept, we sum the propagation scores of all seed users (i.e., topic providers). Table 1 lists the three components of the measurement. Each component can contain an upper-bound value and a lower-bound value for a term. A concept is regarded as well propagated if a large number of seed people share an idea with many other people, or propagate it over a long distance.

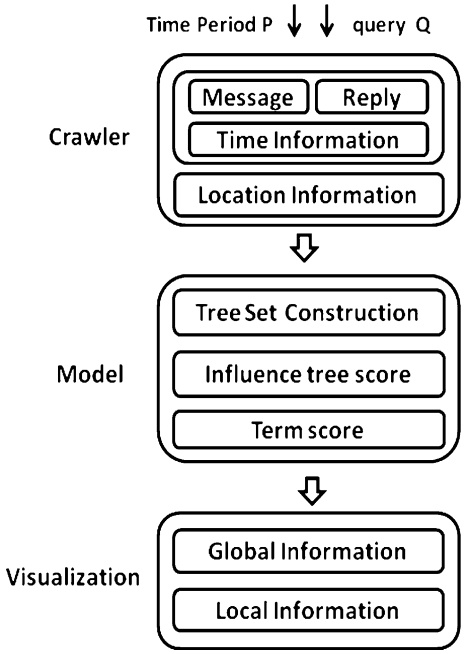


**Fig. 7** Architecture of our system

## 4 System demonstration

In addition to predicting the information propagation in a microblogging social network, we implement an online system to demonstrate the usage of our influence models and measures. We also propose some ways to display global and local information in the system.

### 4.1 Data description

We implement our system on Plurk microblog system, a popular microblog service in Asia. Unlike other social network services, Plurk implements a special feature called a dragable dynamic timeline to display messages. According to the analysis in Lai et al. (2009), more than 5 million users formed a giant graph in Plurk in 2009. Plurk also provides an API that allows developers to search its content using any given query.

### 4.2 System architecture

Our system serves as a search platform for users to observe and analyze the propagation of information in a microblogging social network. The system can also be regarded as a visualization tool for information propagation. In the system, we display the local propagation chart as well as some global statistics to facilitate further analysis. There are three stages in the system, as shown in Fig. 7. First, a crawler collects Plurk data using its provided API. Our system utilizes a topic term and the indicated time period to crawl the related posts and replies as well as the relevant users of Plurk. We crawl four kinds of content: (1) relevant posts; (2) the replies to each post; (3) the time stamps of the posts and the replies; and (4) the geographic information provided by the relevant users. Second, we use the crawled data to construct propagation trees for each relevant user. Then, we use the proposed measures to quantify the propagation capability of each relevant user. Finally, we display the information propagated by top-ranked users. We show the top five microbloggers for each measure. For each microblogger, users can click the icon to see the associated propagation paths and other relevant information, as will be shown later.

### 4.3 Case study

We use the term "FIFA 2010" as the query topic to demonstrate how our system works. FIFA 2010 began on 11 June 2010. It was one of the most popular sports events in 2010. We also set the time period as 11 June to 13 June 2010. In the following, we demonstrate information propagation from the global and local perspectives.

#### *4.3.1 Global information for propagation*

Given a query term (topic), the system displays four types of global information, as shown in Fig. 8: (1) histogram of repliers, (2) daily-post distribution, (3) Web content, and (4) influence table. We describe each type in detail below.


**Fig. 8** The global information page of our system

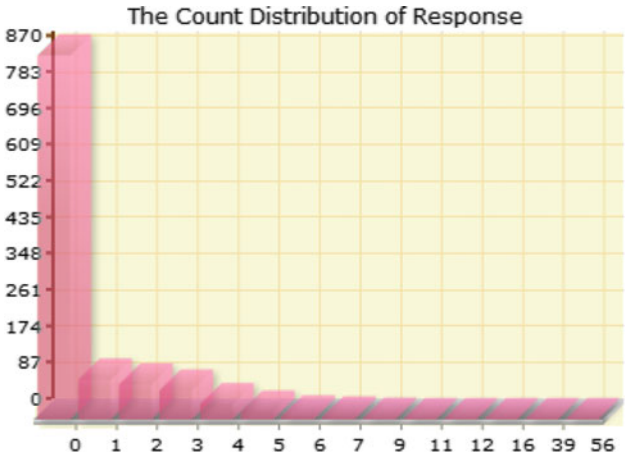


**Fig. 9** The distribution of "the number of responses"(*x*-axis) and "the corresponding counts" (*y*-axis)

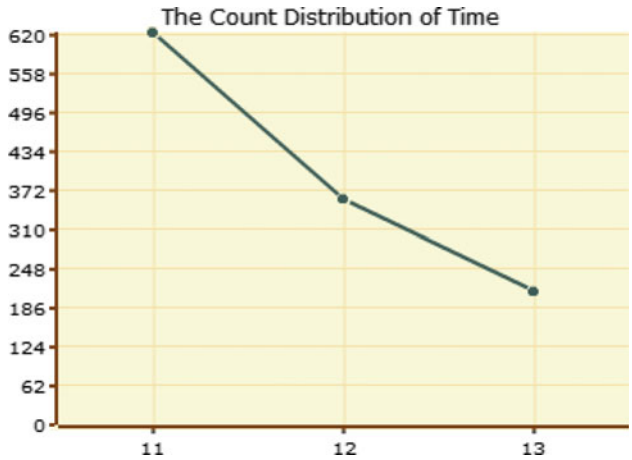


**Fig. 10** The distribution of "the day" (*x*-axis) and "the number of counts" (*y*-axis)

### *4.3.2 Histogram of repliers*

We gather the statistical information about the number of people that replied to each message in the given time period, and then use an open source, open flash chart,[7] to display the distribution. Figure 9 shows our example of 'FIFA 2010'. The horizontal axis represents the number of people that replied to each message, and the vertical axis indicates the frequency of each message. In this example, we can conclude that the distribution follows a power law distribution.

[7] Open Flash Chart: http://teethgrinder.co.uk/open-flash-chart/.

*4.3.2.1 Daily-post distribution* In the daily-post distribution diagram, we generate the statistics of the number of messages post with respect to time. The results are shown in Fig. 10, where the horizontal axis represents the specified time period (e.g., 11–13 June in this example) and the vertical axis indicates the total number of messages posted about the topic each day.

*4.3.2.2 Web information* We exploit the Yahoo Search API for a given query and display the top three related news items as well as the top image search results. The system can also serve as a search engine that provides some information about the term itself.

*4.3.2.3 Influence table* For each query term, we generate an influence table to show the top five users given each propagation measure (i.e., the number of people influenced, the speed of propagation, and the geographical distance). We use the LBI and UBI scores as the lower/upper bounds in Fig. 11. The system takes the average score of the two bounds to rank the influence of each user, as shown in Fig. 12. Each user has a hyperlink to display the local propagation information, which we discuss in the next section.

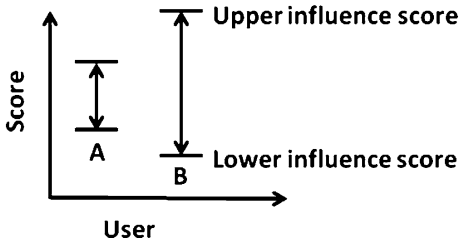


**Fig. 11** The illustration of boundary for LBI and UBI scoring

**Fig. 12** The influence table

**Top Five Records For Each Measure**

| Ranking | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Influence Score | briian (5~93) | daiiiLa (2~13) | nadianadiot (1~14) | IndahNovianty (1~11) | peby_bink (1~10) |
| Speed Score | briian (30~637) | dachuan (7~485) | IndahNovianty (7~474) | MOOa (7~472) | inga90611 (7~466) |
| Distance Score | briian (56~1612) | IndahNovianty (0~1128) | dachuan (0~1036) | inga90611 (0~1004) | sweetcody (0~976) |

(以上Score越高者代表傳播能力越強，Location越高者代表在地理位置上傳得越遠)

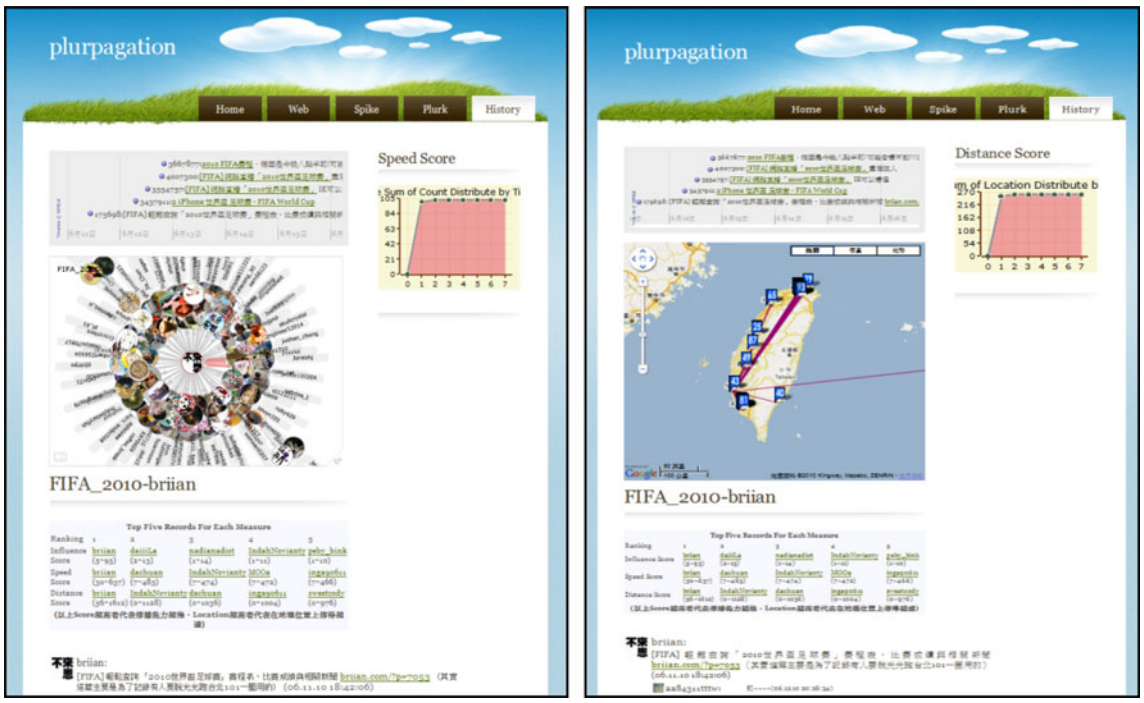


**Fig. 13** Two pages of local information

## 4.4 Local information for propagation

For each top-ranked microblogger, our system provides five kinds of local information: (1) a timeline, which displays the content of the messages and replies as well as the temporal relationship between them; (2) the structure of the influence tree as a visualization of the propagation; (3) a map to show the geographic propagation; (4) an aggregated spread count over time to indicate the speed of propagation; and (5) the content of the messages and replies in table format. Our system displays two kinds of pages, as shown in Fig. 13. Here we use the top blogger "briian" for demonstration.


**Fig. 14** The timeline of influence tree with "briian" as root

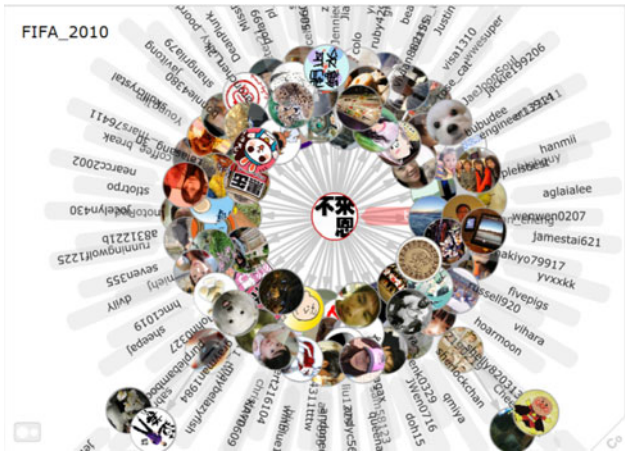


**Fig. 15** The influence tree by considering the user "briian" as the root

#### 4.4.1 Timeline

We use the tool Simile Widgets[8] to visualize information about the timeline of the related messages, which is root from "briian", as shown in Fig. 14. The timeline shows how quickly a message is spread within a given time period.

#### 4.4.2 Graph gear

The UBI influence tree rooted in "briian" is shown in Fig. 15. We use the open source tool Graph Gear[9] to visualize the tree. The images in the circles are the pictures of the bloggers in Plurk. The red circle indicates the root user. By following the directions of the edges, it is possible to identify the paths of information propagation. The display allows the users to determine whether the high influence of a person is caused by the long propagation paths, or simply the consequence of its high degree.

#### 4.4.3 Dynamic map marking for geographical information

We use Google Map to mark the geographic propagation location (i.e., the latitude and longitude coordinates of each user (node) in order to visualize the geographic influence of a UBI influence tree. The tree rooted in "briian" is shown in Fig. 16. The marks on the map appear in order of the posting time. Through the links, we can visualize the geographic propagation path. This feature also allows us to determine whether the query topic is propagated locally or globally.

#### 4.4.4 Aggregated spread count over time

The aggregated spread count chart represents an accumulation of propagated person count or total distance over time in an influenced tree, which is similar to the chart in Fig. 5. The feature allows users to determine the propagation speed of the root user.

#### 4.4.5 Message

Finally, the system displays the content of message posts and replies, as shown in Fig. 17. Here, we find that the bloggers "Cheese 0831" and "and one demon" have posted similar information.

### 4.5 Ranking on term scores

We evaluate how the results of our model compare to those generated by humans in terms of ranking the level of propagation of terms. We choose five well-known brands of 3C products, namely MAC, DELL, ACER, MSI, and LENOVO as query topics to generate term scores. The results are shown in Fig. 18. In Plurk, MAC has the highest level of propagation followed by Dell. We then find five humans to rank these five products based on how they believe the level of propagation of each idea in general. The results show that the Kendall tau rank correlation among our system's results and the human results is 0.60, which is considered fairly significant in statistics. The results show that our measures match the impressions of the human participants to a certain extent.

## 5 Application scenario: using our framework to evaluate diffusion models

To demonstrate the usage of the proposed measures of influence propagation, we propose to apply it to evaluate which models can accurately predict the propagation route and finding influential users in a microblogging social

[8] Simile widgets: http://www.simile-widgets.org/timeline/.

[9] GraphGear: http://www.creativesynthesis.net/recycling/graphgeardemo/.

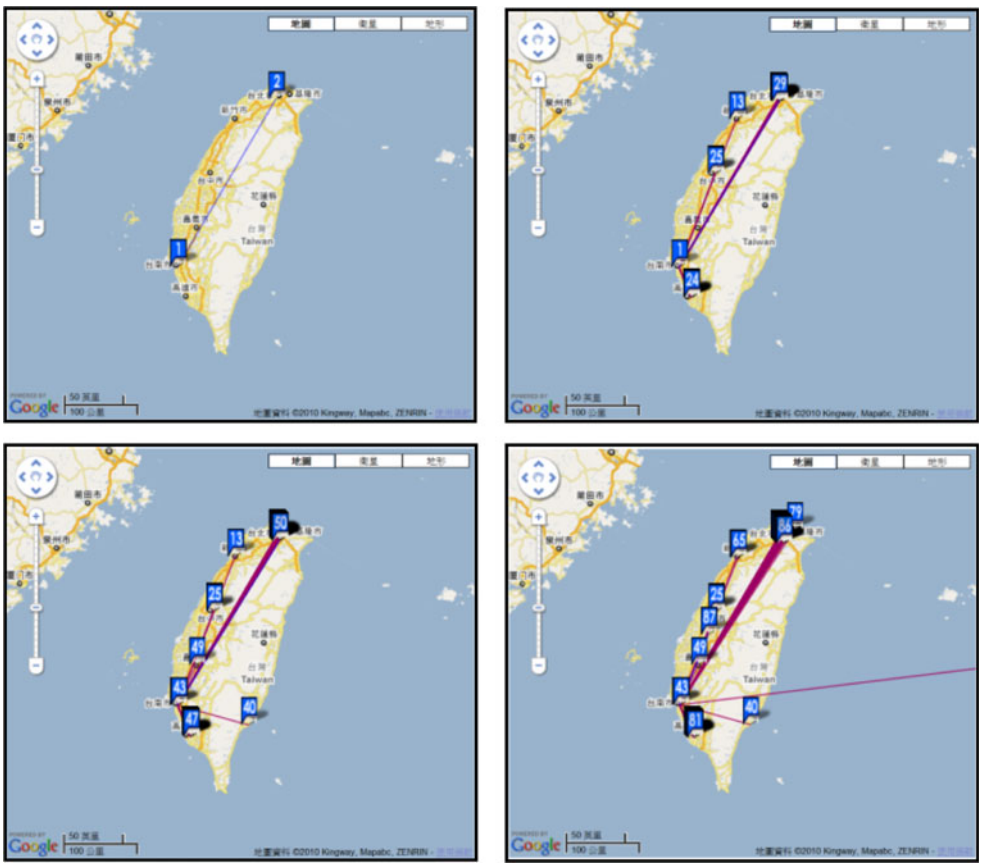

**Fig. 16** Dynamic visualization of propagation on Google Map


**Fig. 17** The system displays the content of message posts and replies

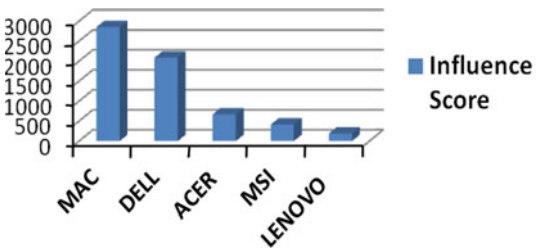


**Fig. 18** Term scores based on the five topics

network. An evaluation framework, Evaluation for Predicting Information Cascades (EPIC), is developed to fulfill this goal.

### 5.1 The EPIC evaluation mechanism

The EPIC framework is shown in Fig. 19. The input of EPIC is a set of influence propagation models, a social network through which the diffusion of information is conducted, and the propagation record for specific information. The propagation record is a set of {*time*, $node_1$, $node_2$} tuples which represents past diffusion ($node_1$ propagates a concept to $node_2$ at a specific *time*) behaviors. EPIC then determines the most suitable influence propagation model-given data. The EPIC framework consists of two evaluation schemes: propagation link prediction and propagation capability prediction.

The idea of EPIC is to exploit any user-specified influence measure (e.g., in-degree, retweet, or mention (Cha et al. 2010)) to construct the gold standard. In our experiment, we apply the *scale of propagation* measure proposed in the previous section as the influence measure. Then we can compare the gold standard with the diffusion outputs of different models to obtain the most suitable model. The data are divided into training and testing sets. The training set is used to find the best parameters for each model, while the testing dataset are used to determine the performance of each model. We have two different comparison schemas:

The evaluation on propagation link aims at comparing the links of predicted diffusions and the links of ground

diffusions. The performance of the diffusion models is determined using F-score of the predicted and the ground truth links. As in the example shown in Fig. 20, the recall is 2/3, precision is 2/4, and F-score is 4/7. This evaluation scheme is straightforward, but fails to consider nodes with higher impact (or diffusion capability) in propagation.

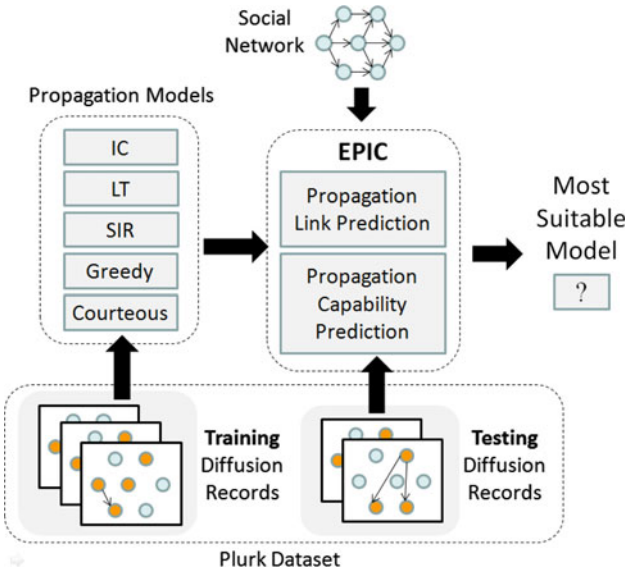


**Fig. 19** The EPIC framework. Given a set of simulating influence models, a social network, and diffusion records, we evaluate the models using propagation link prediction and propagation capability prediction to determine the most suitable one

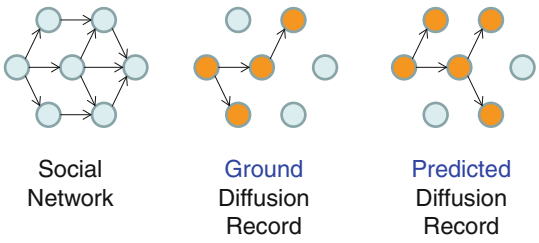


**Fig. 20** The example for the propagation link prediction scheme

In the propagation capability estimation, we first construct the upper-bound influence tree using BFS search method in the data. Then, for each seed node, we count the total number of the nodes affected in the corresponding influence tree as the propagation capability value of this node. For example, a tree constructed from the diffusion record is shown in Fig. 21a. In Fig. 21b, the scale of propagation of seed node "$a$" is $s_a = 4$("$c$", "$e$", "$f$", "$g$"); in Fig. 21c, the scale of propagation of seed node "$b$" is $s_b = 3$ ("$c$", "$d$", "$e$").

We then pick the top k users of the largest propagation capability values as the ground truth value. We compare this set with the top $k$ users obtained from a diffusion algorithm to be evaluated to learn how well the diffusion algorithm does in predicting the top-ranked nodes with the highest propagation capability.

However, evaluating the influences of the users one by one might *overestimate* the propagation power of each user. As shown in Fig. 21b, c, the nodes "$c$" and "$e$" are actually double counted for the scales of propagation of root node "$a$" and "$b$", thus we overestimated the propagation power of "$a$" and "$b$". Therefore, we design another flow, *leave-one-out*, to evaluate the prediction results. Suppose we want to calculate the influence of node $v$. First, we compute the scale of propagation given all seeds $s_{all}$. Then, we compute the scale of propagation *without* node $v$, which is $s_{all\text{-}v}$. Finally, the leave-one-out scale of propagation of node v is $s'_v = s_{all} - s_{all\text{-}v}$.

The intuition behind the leave-one-out flow is that we want to know "if a user were not present, how much influential power would be missing?". For example, to compute the influence of the root node "$a$" shown in Fig. 22a, we first compute the whole scale of propagation $s_{all} = 5$ (the size of the union propagation set from all seeds). Then, the influence without node "$a$" is $s_{all\text{-}a} = 3$, because the tree of 22(b) has to be removed. Finally, the leave-one-out scale of propagation $s'_a = s_{all} - s_{all\text{-}a} = 5 - 3 = 2$. Similarly, we can compute $s'_b = 5 - 4 = 1$, as shown in Fig. 22c. Note that while the one-by-one flow

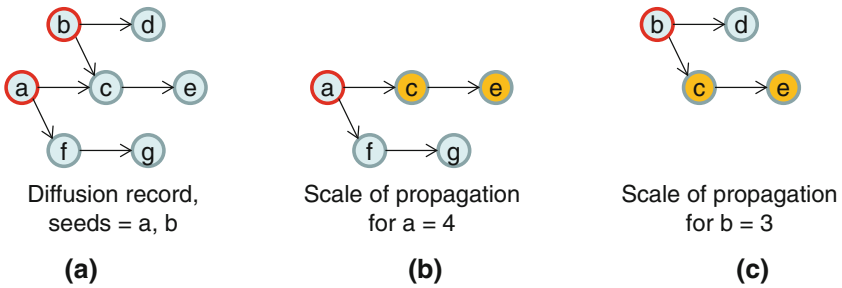


**Fig. 21** The propagation capability prediction scheme using one-by-one flow. **a** An example of the diffusion record, assuming node "$a$" and "$b$" are seed nodes. **b** Using node "$a$" as the seed node, the one-by-one scale of propagation is 4. **c** Using node "$b$" as the seed node, the one-by-one scale of propagation is 3. In addition, such prediction scheme tends to overestimate the diffusion capability. In this example, user $c$ and $e$ are double counted

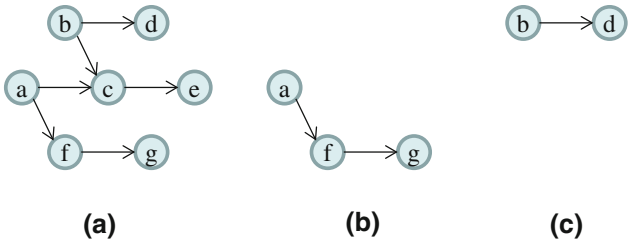


**Fig. 22** The scheme using leave-one-out flow. **a** The same example of the diffusion record. **b** Using node "*a*" as the seed node, the leave-one-out scale of propagation is 3. **c** Using node "*b*" as the seed node, the leave-one-out scale of propagation is 2

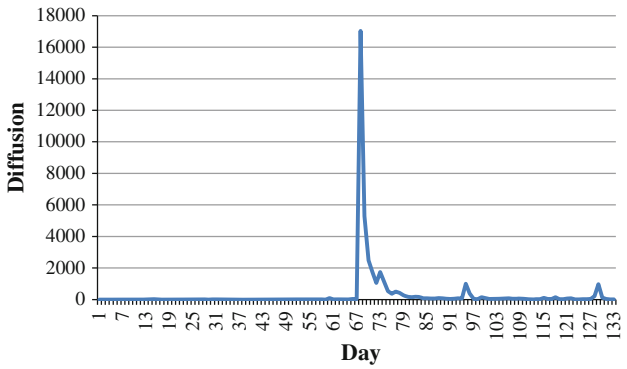


**Fig. 23** The diffusion-versus-day plot for the topic Earthquake dataset

tends to overestimate the influential power of the nodes, the leave-one-out flow tends to *underestimate* the influential power. We believe the real influential power lies between the scales of propagations computed using two flows.

## 5.2 Case study on EPIC

In this section, we would like to use the EPIC framework to compare the diffusion capability of several famous models. We use a machine with AMD Opteron 2350 2.0 GHz Quad-core CPU and 32 GB RAM to run the experiments.

### *5.2.1 Datasets*

In our experiment, we collect data from the Plurk microblog system, a popular microblog service in Asia. According to a previous analysis, more than 5 million users formed a giant graph in Plurk in 2009 (Lai et al. 2009). We first identify 100 hot topics from Plurk and then search the whole Plurk to collect the users who post or reply to related articles. Then, we collect the two-hop neighbors of the users. In this dataset, the number of nodes = 940,070 and the number of links = 7,660,770. The duration of the messages and responses is from 1 January 2011 to 15 May 2011. The most frequent topic is "Earthquake", which contains 80,336 messages and 303,523 responses. There are 38,453 diffusions, and among them there are 90 % diffusions for training (determining the best parameters and 10 % for testing).

Here, we select the top 50 users for capability prediction. Note that we removed the responses from users who were not part of the selected social network. The diffusion-versus-day plot for the Plurk dataset is shown in Fig. 23. It contains single spike diffusions (the spike is at the day after the disastrous Japanese earthquake).

### *5.2.2 Influence diffusion models for evaluation*

We tested five diffusion models in our experiments. For each model, we use the root nodes generated in our evaluation scheme as seed nodes. The values of seed nodes are initialized to 1.0. In the propagation link prediction scheme, all diffused links generated using the diffusion model are compared with actual links to determine the performance of the model. In propagation capability prediction scheme, we compute the scale of propagation for each seed node using the diffusion tree generated using the diffusion model. We then pick the top $k$ to compare with the top k in the group truth.

- *Independent cascade (IC)* (Kempe et al. 2003). When node $v$ becomes active, it has a single chance of activating any currently inactive neighbor $w$. The activation attempt succeeds with probability $p_{vw}$. Whether or not $v$ succeeds, it cannot make any further attempts to activate $w$ in subsequent rounds. The process runs until no more activation is possible. The single parameter of IC model in our experiment is $p_{vw}$.
- *Linear threshold* (*LT*) (Kempe et al. 2003). A node $v$ is influenced by each neighbor $w$ according to a weight $b_{vw}$ such that $\Sigma_{w \text{ neighbor of } v} b_{vw} \leqq 1$. Each node $v$ has a threshold $\theta_v$. A node $v$ is diffused if $\Sigma_{w \text{ neighbor of } v} b_{vw} \geqq \theta_v$. The process continues until no more activation is possible. In our experiment, we set $b_{vw} = 1/\text{degree}_v$, thus the single parameters of the LT model in our experiment is $\theta_v$.
- *Susceptible–infected–recovered* (*SIR*) (Kermack and McKendrick 1927). Each node can be one of the three states: susceptible (is healthy but can catch diseases if exposed to some infection), infected (has a certain disease and can pass it on), and recovered (has recovered from the disease and has permanent immunity). The birth rate $\beta$ is defined as the probability a node catches the disease from an infective one and the death rate $\gamma$ is defined as the probability that an infected individual recovers. The mathematical formulation of the SIR model at time $t$ is then defined as: $\mathrm{d}s/\mathrm{d}t = -\beta_{is}$, $\mathrm{d}i/\mathrm{d}t = \beta_{is} - \gamma_i$, and $\mathrm{d}r/\mathrm{d}t = \gamma_i$. In our experiment, we

set $\gamma_i = 0.01$, thus the single parameter of the SIR model in our experiment is $\beta_{is}$.

- *Greedy* (Gupte et al. 2009). A node would have its own preference on a specific topic. When a node is activated by its neighbor, then it will decide whether to propagate, depending on the faction $f_u$ of neighbors who like this topic. If $f_u$ is greater than a threshold $t$, the user will propagate the information. The propagate condition is $f_u = |Interested\ Neighbor\ (u)|/|Neighbor(u)| > t$. In our experiment, for each topic we randomly assign half the people who are interested in it. The only parameter of this model is $t$.
- *Courteous* (Gupte et al. 2009). The main difference between the courteous model and greedy model is that a node does not want to spam its friends. More precisely, the propagate condition is $f_u = |Interested\ Neighbor(u)|/|Neighbor(u)| > t$ and $|Seen\ Neighbor(u)|/|Neighbor(u)| \leqq c$. In our experiment, for each topic we randomly assign half the people who are interested in it and set $c$ to be 0.25. Therefore, the only parameter of also this model is $t$.
- Therefore, there is only one parameter for each diffusion model in our experiment. In the training stage, we use brute-force method to find which parameter value gives the best result. For each parameter, we apply the parameter value from 0.00 to 1.00, with increment = 0.01.

### 5.2.3 Evaluation results

The results of the propagation link prediction scheme for the best/conventional parameter values of five models are shown in Table 2. Comparing all models and parameter combinations, the SIR model with parameter $\beta_{is} = 0.91$ performs the best. In general, IC, LT, and SIR models perform better. The greedy model performs stably regardless of the parameter. Although the best F-score does not seem high (4.51 %), it should be noted that these models predict 3845 ground truth diffusions (which is unseen in the historical records) from a large number of candidate friendship links (about 7.7 M), which is an extremely difficult prediction task for any model.

**Table 2** F-score results for the propagation link prediction scheme

| Method | Parameter | | F-score (%) |
|---|---|---|---|
| IC model | Best | $p_{vw} = 0.90$ | 4.42 |
| | Conventional | $p_{vw} = 0.10$ | 3.45 |
| LT model | Best | $\theta_v = 0.05$ | 4.39 |
| | Conventional | $\theta_v = 0.50$ | 1.36 |
| SIR model | Best | $\beta_{is} = 0.91$ | 4.51 |
| | Conventional | $\beta_{is} = 0.50$ | 3.81 |
| Greedy model | Best | $t = 0.46$ | 2.76 |
| | Conventional | $t = 0.50$ | 2.43 |
| Courteous model | Best | $t = 0.67$ | 0.12 |
| | Conventional | $t = 0.50$ | 0.00 |

**Table 3** F-score results for the propagation capability prediction scheme

| Method | One-by-one (%) | Leave-one-out (%) |
|---|---|---|
| IC model | 6.00 | 12.00 |
| LT model | 4.00 | 4.00 |
| SIR model | 6.00 | 14.00 |
| Greedy model | 6.00 | 14.00 |
| Courteous model | 8.00 | 14.00 |

The one-by-one results of propagation capability prediction for best parameter values of five models are shown in Table 3. We do not show parameters in the comparison result because many parameter values give the same F-score. The LT model performs slightly worse (4.00 %) than other models for one-by-one flow, while SIR, greedy and courteous models perform better (14.00 %) for leave-one-out flow. It should be noted that the results for leave-one-out flow is significantly better than that for one-by-one flow. Because leave-one-out flow tends to underestimate a user's influence capability while one-by-one flow tends to overestimate it, we believe in the Plurk dataset the actual user influence can be better captured using leave-one-out flow.

## 6 Conclusion

In this work, we propose to mode land estimate information propagation in a microblogging social network. Instead of producing an exact quantification score, our model provides upper-bound and lower-bound values using the upper- and lower-bound influence scores. Besides proposing a simple yet intuitive way to measure information propagation by counting the number of people influenced, we present novel ways to measure the propagation speed and the geographical distances.

We further devise the EPIC evaluation framework to test the effectiveness of different influence propagation models by predicting influential nodes. Experimental results from two realistic datasets show the promising results of our methods, compared to existing IC, LT SIR, greedy, and courteous models. It should be noted that in the first part of our work, we tried our best to design some simple, intuitive, and data-driven measurements to produce the ground truth. Our idea is to use a 'data-driven' approach to evaluate the 'model-driven' approaches' for information propagation. In our data-driven approach, we do not

consider that idea of 'random establish of link' proposed by the IC model, or the 'node threshold' proposed by the LT model, or the 'random-walk' based concept proposed by random-walk and heat-diffusion models. Therefore we are confident that at least for the popular models we used here for comparison, our measurement does not impose apparent bias on any of them.

Furthermore, we develop an online system, called Plurpagation, to visualize how information is propagated in the Plurk social network. The system displays global information about topical propagation as well as local dynamic information allowing users to gain more insights into propagation patterns. Our system and model are for general purposes; so, these can easily be applied to other microblog services, such as Twitter.

## References


Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. In: Proceedings of ACM international conference on world wide web (WWW'12)

Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the flickr social network. In: Proceedings of ACM international conference on world wide web (WWW'09)

Cha M, Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user influence in twitter: the million follower fallacy. In: Proceedings of the 4th international AAAI conference on weblogs and social media (ICWSM'10)

Cha M, Perez JAN, Haddadi H (2012) The spread of media content through blogs. J Soc Netw Anal Min (SNAM)

Chaoji V, Ranu S, Rastogi R, Bhatt R (2012) Recommendations to boost content spread in social networks. In: Proceedings of ACM international conference on World Wide Web (WWW'12)

Goyal A, Bonchi F, Lakshmanan LV (2010) Learning influence probabilities in social networks. In: Proceedings of ACM international conference on web search and data mining (WSDM'10)

Goyal A, Bonchi F, Lakshmanan LVS, Venkatasubramanian S (2012) On minimizing budget and time in influence propagation over social networks. J Soc Netw Anal Min (SNAM)

Gruhl D, Guha R, Liben-Nowell D, Tomkins A (2004) Information diffusion through blogspace. In: Proceedings of ACM international conference on world wide web (WWW'04)

Gupte M, Hajiaghayi M, Han L, Iftod L, Shankar P, Ursu RM (2009) News posting by strategic users in a social network. In: Proceedings of international workshop on internet and network economics (WINE'09)

Kempe D, Kleinberg JM, Tardos E (2003) Maximizing the spread of influence through a social network. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)

Kempe D, Kleinberg J, Tardos E (2005) Influential nodes in a diffusion model for social networks. In: Automata, languages and programming, vol 3580, pp 1127–1138

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. In: Proceedings of the royal society of London

Kimura M, Saito K, Nakano R, Motoda H (2010) Extracting influential nodes on a social network for information diffusion. In: Data Mining and Knowledge Discovery (DMKD), vol 20, pp 70–97

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media. In: Proceedings of ACM international conference on world wide web (WWW'10)

Lai HC, Chen CW, Liu PS, Lin SD (2009) Exploiting cloud computing for social network analysis—exemplified in Plurk network analysis. In: Proceedings of international conference on technologies and applications of artificial intelligence (TAAI'09)

Lampos V, Cristianini N (2010) Tracking the flu pandemic by monitoring the social web. In: Proceedings of international workshop on cognitive information processing

Lampos V, Bie TD, Cristianini N (2010) Flu detector: tracking epidemics on twitter. In: Proceedings of ECML PKDD 2010

Leskovec J, Singh A, Kleinberg J (2006) Patterns of influence in a recommendation network. In: Proceedings of Pacific-Asia conference on knowledge discovery and data mining (PAKDD'06)

Ma H, Yang H, Lyu MR, King I (2008) Mining social networks using heat diffusion processes for marketing candidates selection. In: Proceedings of ACM international conference on information and knowledge management (CIKM'08)

Maiya AS, Berger-Wolf TY (2010) Online sampling of high centrality individuals in social networks. In: Proceedings of Pacific-Asia conference on knowledge discovery and data mining (PAKDD'10)

Richardson M, Domingos P (2002) Mining knowledge-sharing sites for viral marketing. In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (KDD'02)

Rodriguez MG, Leskovec J, Krause A (2010) Inferring networks of diffusion and influence. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD'10)

Rushkoff D (1994) Media virus: hidden agendas in popular culture. Ballantine books

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of ACM international conference on world wide web (WWW'10)

Sarr I, Missaoui R (2012) Managing node disappearance based on information flow in social networks. J Soc Netw Anal Min (SNAM)

Scott J (2011) Social network analysis: developments, advances, and prospects. J Soc Netw Anal Min (SNAM)

Snowsill T, Fyson N, Bie TD, Cristianini N (2011) Refining causality: who copied from whom. In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (KDD'11)

Song X, Chi Y, Hino K, Tseng BL (2007) Information flow modeling based on diffusion rate for prediction and ranking. In: Proceedings of ACM international conference on world wide web (WWW'07)

Song X, Chi Y, Hino K, Tseng BL (2007).Identifying opinion leaders in the blogosphere. In: Proceedings of ACM international conference on information and knowledge management (CIKM'07)

Steeg GV Galstyan A (2012) Information transfer in social media. In: Proceedings of ACM international conference on world wide web (WWW'12)

Stewart A, Chen L, Paiu R, Nejdl W (2007) Discovering information diffusion paths from blogosphere for online advertising. In: Proceedings of international workshop on data mining and audience intelligence for advertising (ADKDD'07)

Sun E, Rosenn I, Marlow C, Lento T (2009) Gesundheit! modeling contagion through Facebook news feed. In: Proceedings of AAAI international conference on weblogs and social media (ICWSM'09)

Tang J, Musolesi M, Mascolo C, Latora V, Nicosia V (2010) Analysing information flows and key mediators through temporal centrality metrics. In: Proceedings of international workshop on social network systems (SNS'10)

Yang CC, Tang X, Thuraisingham BM (2010) An analysis of user influence ranking algorithms on Dark Web forums. In: Proceedings of ACM SIGKDD international workshop on intelligence and security informatics (ISI-KDD'10)

Zhang W, Wu W, Wang F, Xu K (2012) Positive influence dominating sets in power-law graphs. J Soc Netw Anal Min (SNAM)