WHO ARE INFLUENTIALS ON MICRO-BLOGGING SERVICES: EVIDENCE FROM SOCIAL NETWORK ANALYSIS

Caihong Sun, School of Information, Renmin University of China, Beijing, China, chsun@ruc.edu.cn
Lu Zhang, School of Information, Renmin University of China, Beijing, China, nanjingzhanglu@163.com
Qian Li, School of Information, Renmin University of China, Beijing, China, linda_lq@ruc.edu.cn

Abstract

User influence measurement on Micro-blogging can be classified into two types according to their concerns on following relationships or interactions among users. The former focuses on counting the audiences subscribe to a user’s tweets, neglecting the fact that many audiences do not read the tweets they subscribed to, while the latter emphasizes the audiences’ explicit actions such as retweeting or replying, excluding those who read the tweets but choose to be lurkers. Since user interactions indicate that audiences have read the tweets explicitly, and following network is used to estimate the audiences who read the tweets but implementing no explicit actions, to combine the advantages of these methods, an information diffusion approach is proposed to measure user influence in this paper. Simulation results demonstrate that our proposed approach can estimate the effective audiences in the process of tweet diffusion. In the view of information diffusion, the hybrid method gives a more accurate user influence measurement by integrating following network and user interaction.

Keywords: Micro-blogging, Information diffusion, effective audience, following network.
1 INTRODUCTION

Influential people in the society have been studied in the fields of sociology, communication, marketing and political science for their vital roles in both political and economic world (Katz and Lazarsfeld 1995, Lazarsfeld et al. 1968, Watts and Dodds 2007). For instance, many firms apply virus products’ marketing strategy through influential people and politicians want to get more support from these influential people. Finding influential people can help us better understand how certain opinions, trends or innovations are adopted and diffused and how we could help advertisers, politicians, marketers, and government to design more effective campaigns and policies (Cha et al. 2010). The emergence of social media is transforming the ways advertisers and politicians approach their consumers or supporters, Nielsen reports that 92% of consumers trust products recommended by people they know (Katona 2013).

Micro-blogging services, such as twitter, sina weibo (means micro-blogging in Chinese), develop very rapidly with a fast-growing user group. According to the Paris-based analyst group Semiocast, The number of twitter users exceeded 500M in June 2012. 200M of them are active users. Sina weibo, a popular micro-blogging platform in China has more than 0.3 billion users in June 2012 (announced by news report of sina media). Apparently, these users are not equally active and contributable, and in order to utilize the value of the vast consumer-to-consumer communication networks facilitated by social media, how to identify the influentials is becoming a hot and important research topic in social media marketing. Marketers have long mined data from social networks for “influencers,” people whose favorable tweets and posts can boost product sales.

Many researchers contribute their efforts to user influence measurement on micro-blogging services, especially on twitter. These researches could be classified into two main types according to their concerns on following relationships or interactions among users. Some researchers focus on following network measurement. For example, a most popular metric of user influence on Twitter is to measure the number of a user’s followers, it only considers one-step connections among users, ignoring contents, link structure, and interactions among users (Leavitt et al. 2009). Another similar and popular user influence measurement involves the ratio between the number of a user’s followers and the number of other people the user follows (Tunkelang 2009). Better than counting followers only, the ratio approach is still imprecise, because it neglects the ability of a user to interact with contents on the micro-blogging platform (Leavitt et al. 2009). TwitterRank, a PageRank-like algorithm, incorporating tweets information and following network, uses following link structure and weighted topic similarity to define transition probability matrix (Weng et al. 2010).

Since 2010, researchers started to concern user interactions for measuring user influence. Cha et al. give the concepts of retweet influence and mention influence (Cha et al. 2009). The former refers to the number of retweets containing one’s name, indicating a user’s ability to generate content with pass-along value; while the latter focuses on mentions containing one’s name, indicating a user’s ability to engage others in a conversation. These two influence measurements do consider the interactions of users from contents and conversations respectively. But the amount of retweets or mentions cannot represent the information diffusion network among users. For example, a user’s influence may increase if his tweet is retweeted by those who are more influential. To solve this problem, Bakshy et al. measure user influence in terms of the size of the entire retweet trees associated with each event (Bakshy et al. 2011). Zhang et al. propose an action-based network approach which captures explicit user interactions (such as retweeting and replying) based on action-based link structure and emphasizes user influence on explicit information diffusion tree (Zhang et al. 2011).

We argue that both following network and user interactions contribute to user influence. On one side, a user’s tweet will be passed to his followers; hence the followers, being called implicit audiences, have possibility to be influenced by his tweets. If the followers read a tweet they subscribed to without any explicit actions indicating their reading, we call them implicit effective audiences. On the other side, by considering the interactions among users, if a user makes explicit actions (such as retweet or reply) on a tweet, we call him an explicit audience, implying that he has been influenced by
the tweet. Furthermore, for those explicit effective users, their retweets could be forwarded to their followers which could bring new effective audiences.

This paper proposes an approach to measure user influence by considering both his implicit and explicit audiences and counting the number of effective audiences. The idea of measuring the number of effective users first appeared in Kwak’s Phd dissertation, and he assessed how many persons are exposed to information by each tweet and measured influence of tweets’ writer in term of aggregate number of exposed persons (Kwak 2011). He considered the effective audiences in one step, i.e. a user can influence only his followers. We argue that besides followers, a user’s tweet could be passed to others by retweeting. Hence besides the implicit number of exposed persons of a tweet, the size of explicit exposed persons arising from two kinds of explicit actions — retweeting and replying are considered recursively. A “reply” action indicates that the user has read the tweet, while a “retweet” action means that the user passes the tweet to his followers with information cascade after reading.

In this paper, we have three main contributions. The first one is that a tweet influence model is presented which includes influence on both explicit and implicit influenced audiences along the tweet’s information diffusion paths. It helps to identify the key players in an information diffusion process. Both following network structure and interactions among users play very important roles in information diffusion process. Hence, both following network and user interactions (retweet and reply actions) are considered to measure a tweet’s influence. There are three specific information diffusion cases in our model.

• Case 1: if user A publishes an original tweet $w$, all his followers could receive $w$. But due to information overload, for each user, he only read part of his received tweets. Hence, for A’s follower B, he could have a chance to read this tweet (with a probability), we say user B is influenced by user A with a probability $p$;
• Case 2: if user C (not necessary to be user A’s follower) employ “reply” action on the tweet, then user C is influenced by user A;
• Case 3: If user C (not necessary to be user A’s follower) employ “retweet” action on the tweet, then user C is influenced by user A, and moreover, user C’s retweet plays as a new tweet and could influence other users in the way of the above three situations, and influence of user C’s retweet is added into the influence of user A’s original tweet $w$ recursively.

In addition, to wipe off overlapped audiences, each user will be counted only once even if he is influenced repeatedly by the same tweet.

Our second contribution is that user influence is measured in each time period that could capture the ecology of user influence. Since micro-blogging is a dynamic system, user influence could change constantly, capturing user influence dynamically is quite important.

Finally, user influence could be different across topics, for instance a physical scientist could have more influence on physics topics than on financial topics. By classifying tweets into topics, our proposed user influence can be measured under different topics.

The rest of this paper is organized as follows: Section 2 gives the definition of user influence we considered and presents a basic model to quantify tweet influence; Section 3 presents four user influence measurements based on the basic tweet influence model; Section 4 demonstrates the simulations and our experimental results; Section 5 discusses the pros and cons of our proposed approach; and Section 6 draws the conclusions.

2 QUANTIFYING THE INFLUENCE OF A TWEET

Our basic idea of measuring user influence is to count the effective audiences of his published tweets. Hence, tweet influence measurement as a basic model is presented first.

2.1 The Definition of a Tweet’s Influence

It is difficult to give a universal definition of user influence embodying all aspects in various contexts. Many researchers, illustrate user influence on their own ways, such as number of followers and
number of retweets, resulting in different user influence rankings on micro-blogging platforms (Kwak et al. 2011, Cha et al. 2010 and Zhang et al. 2011). In this paper, micro-blogging platform is studied as a social medium. The posted tweets diffuse among the users on micro-blogging platform as time goes by. A tweet’s influence is defined as the number of audiences who have read the tweet during a period of time \( t \). Along the information diffusion paths, two kinds of effective audiences are distinguished: implicit effective audiences and explicit effective audiences.

Implicit effective audiences of a tweet \( w \) are those users who read tweet \( w \) without reposting or commenting on it. For example, user \( A \) posts a tweet \( w \), \( w \) is passed to all his followers, but it does not mean all of user \( A \)'s followers will read tweet \( w \), practically, only a fraction of his followers read tweet \( w \) due to information overload or time concerns. But it is hard to estimate the exact number of implicit effective audiences of a tweet. In our model we assume that a user reads a tweet he subscribes and becomes an implicit effective audience with a certain probability related to the number of users he follows.

Furthermore, we assume if a user retweets or replies a tweet, it implies that the user has read the tweet, and become an explicit effective audience. For example, if user \( A \) responds to (replies) user \( B \)'s tweet \( w \), it implies that user \( A \) has read \( w \), and user \( A \) is regarded as an explicit effective audience of \( w \). If user \( A \) cites or paraphrases (retweets) user \( B \)'s tweet \( w \), it implies that user \( A \) has read \( w \), and user \( A \) becomes \( w \)'s explicit effective audience. In a word, if a user takes “retweet” or “reply” actions of tweet \( w \), then he/she is \( w \)'s explicit effective audience. The thing is that “retweet” and “reply” actions are different, only “retweet” actions could bring new effective audiences. Since if user \( A \) retweets user \( B \)'s \( w \), then all of user \( A \)'s followers will receive this retweet, by which “retweet” actions could bring new effective audiences with information cascades.

As shown in figure 1, user 1 posted a tweet \( w \), user 5 replied \( w \), user 4 and 6 retweeted \( w \). User 2 and 3 are those who read tweet \( w \) without any explicit response, called the implicit effective users. Since user 5 replied tweet \( w \), he is an explicit effective user. User 4 and 6 are explicit effective users of \( w \). In addition, since \( w \) was retweeted by user 4 and 6, then their followers could become effective users of \( w \). In the case of figure 1, user 7, 9 and 10 became \( w \)'s explicit effective audiences by applying “retweet” actions, and user 8, 10 became \( w \)'s implicit effective audiences by reading the retweet of \( w \). The “retweet” actions could diffuse tweets and bring more effective audiences including implicit and explicit ones.

Micro-blogging service is an ecological system; therefore, the user influence evolves too. "Timeliness of user influence", being mentioned by researchers such as Kwak and Zhang et al., describes the time-frame characteristics of users (Kwak 2011 and Zhang et al. 2011 ). In our model, we discuss the effective audiences of a user in a time period to reflect the information diffusion effects and measure user influence.

![Information diffusion process](image.png)
2.2 Basic Model for Measuring Tweet Influence

A state variable (having state 1 and 0) is used to indicate whether a user is tweet w’s effective audience in a time period t. Initially all the state of the users for tweet w is set to 0, once a user becomes an effective audience of tweet w, then his state changes to 1. A user’s state of tweet w could not change from 1 back to 0. Since state \( S(v,w) = 1 \) means that the user v has read the tweet w. The state variable could help us fix the problem of overlapped audiences. For example, a user v could read tweet w several times from different information sources (for instance, user 3 in figure 1, he read w twice from his followee user 1 and user 7), by using state variable it was only counted as one effective audience.

Initially, for all users, we set the state of reading tweet w as 0.

\[
\forall v \in U, S(v,w) = 0 \tag{1}
\]

Where \( U \) is the set of all users and \( S(v,w) \) is the state of user v who has read tweet w.

Assume tweet w is posted by user u, and then tweet w’s explicit effective audiences are computed by the following steps. Firstly, all the explicit actions of tweet w are collected to build its information diffusion paths. A diffusion path of given tweet w is defined as follows. If user a published tweet w, and user b retweeted a’s tweet w, further user c retweeted user b’s retweet of w, then an information diffusion path should be \( a \rightarrow b \rightarrow c \). Secondly, tweet w’s retweet networks (w is published by user u) is constructed, denoted as \( \text{RetweetTree}(u,w) \). In addition, those users who replied to \( w \)'s retweets are explicit effective users of \( w \) too. To count these users who replied to user \( u \)'s tweet \( w \) indirectly, we traversed network \( \text{RetweetTree}(u,w) \) to find users who replied to \( w \)'s retweets. For each user \( v(v \neq u, v \in \text{RetweetUsers}(u,w)) \), we check those users who replied to \( v \)'s tweet (a \( w \)'s retweet). If user \( m \) replied to \( w \)'s retweet, it means that user \( m \) replied to tweet \( w \) indirectly. Let \( S(m,w) = 1 \), to indicate that \( m \) is an effective users of tweet \( w \). Thus, we compute all the effective audiences of tweet \( w \) by considering users who retweeted \( w \) or replied \( w \) directly or indirectly along the information diffusion paths.

Hence the total explicit effective audiences of tweet \( w \) originally written by user u, denoted as \( EEA(u,w) \), is a set of users who retweeted or replied the tweet \( w \) directly or indirectly.

Besides explicit effective audiences, our model also considers implicit ones based on following network. In the information diffusion process, since followers subscribe to his followees’ tweets, they will have chance to read the tweets and become effective readers even if they do not make any explicit responses to the tweets. It is very hard to know how many users read a tweet if they do not have explicit responses. Probability theory is used to estimate this kind of implicit effective readers. We assume users do not differentiate who post the tweets, and read each tweet they subscribe to with a same probability. From the view of information overload, we assume that this probability is relevant with the number of tweets a user receives in a period of time, i.e. \( p(i,u,w) \propto \frac{1}{\sum_{j \in \text{followee}(i)} T(j)} \), where \( T(j) \) is the number of tweets published by user j, and \( \text{followee}(i) \) is the set of users followed by user i. For simplicity, in our paper, we define that user i reads the tweet w posted or reposted by his followee u with a probability \( p(i,u,w) = \frac{\alpha}{\|\text{followee}(i)\|} \), where \( \alpha \) is a scaling parameter, and \( \|\cdot\| \) is the cardinal number.

Hence, implicit effective audiences (denoted as \( IEA \)), \( IEA \) of tweet \( w \) originally written by user u are defined as:
Then for $\forall v \in IEA(u, w), S(v, w) = 1$. In short, the number of implicit effective audiences is computed based on the retweet networks of tweet $w$ and following network. For every user in retweet networks, each of his/her follower $i$ who is not an explicit effective audience of $w$ has a chance $p$ to become an implicit effective audience. Implicit audiences are computed based on retweet networks and following network. In equation (2), $p$ is random variable drawn from uniform distribution $U(0,1)$.

In sum, the total effective audiences including the explicit and implicit ones of tweet $w$ written by $u$, denoted as $EA(u, w)$, are defined as:

$$EA(u, w) = EEA(u, w) \cup IEA(u, w) \quad (3)$$

i.e., $EA(u, w) = \{ v | v \in U, S(v, w) = 1 \}$

The effective audiences of tweet $w$ are those users who read tweet $w$ explicitly (replying or reposing $w$ directly or indirectly) or implicitly. Next section user influence is measured based on tweet influence.

### 3 Measuring User Influence

The basic model in section 2 measures the influence of a tweet by identifying all the effective audiences of a tweet. Then who are the influencers on a micro-blogging platform? As a basic measurement, tweet influence could be used to measure user influence in different ways. An intuitive idea of computing user influence is to count the number of effective audiences of all the tweets posed by a user. In addition, sometimes, a user is influential in an area does not mean he/she is also influential in another area. A topic-based user influence measurement is very useful in social media marketing and political elections. Besides, timeliness is an important factor to measure user influence. User influence can be measured under different topics dynamically.

Micro-blogging services always have tons of users, ranking all the users is not necessary. Since tweets are the most important ways on micro-blogging to exhibit user influence, in our paper, user influence is measured based on tweet influence under different topics and time periods. How to compute user influence based on tweet influence? For different purposes, six user influence measurements are presented here for selections.

**Measurement 1:** Total effective audiences, which considers all the effective audiences of all the tweets user $u$ published as user influence measure, $UI$:

$$UI(u) = \bigcup_{w \in T(u)} ||EA(u, w)|| \quad (4)$$

Where $T(u)$ is the total tweets posted by user $u$. The cardinal number of $UI$ can be used to computing ranking score. This measurement counts the effective audiences of all the tweets published by a user $u$. Each effective user is counted only once no matter how many times he influenced by user $u$’s tweets.

**Measurement 2:** Total tweet influence, which considers the influences of all the tweets user $u$ published as user influence measure, $UI$:

$$UI(u) = \sum_{w \in T(u)} ||EA(u, w)|| \quad (5)$$

This measurement sums up each tweet’s influence. Not like each effective user is only counted once in measurement 1, an effective user is counted $n$ times if he is influenced by user $u$’s $n$ tweets.
**Measurement 3:** Average tweet influence of a user, which measures the average tweet influence of a user, AUI:

\[
AUI(u) = \frac{\sum_{w \in T(u)} ||EA(u,w)||}{||T(u)||}
\]  

(6)

This measurement counts the average number of effective audience of user u’s each tweet.

**Measurement 4:** Topic-based user influence, which measures the influence of user u on a topic tp, TUI:

\[
TUI(u) = \bigcup_{w \in T(u,tp)} EA(u,w)
\]  

(7)

This measurement is similar to Measurement 1, the difference is that tweets are classified into different topics. Where T(u,tp) denotes user u’s tweets that belong to topic tp.

**Measurement 5:** Topic-based user influence, which measures the influence of user u on a topic tp, TUI:

\[
TUI(u) = \sum_{w \in T(u,tp)} ||EA(u,w)||
\]  

(8)

This measurement is similar to Measurement 2, the difference is that tweets are classified into different topics. Where T(u,tp) denotes user u’s tweets that belong to topic tp.

**Measurement 6:** Dynamic user influence, which considers influence dynamics over time, DUI:

\[
DUI(u,t) = DUI(u,t-1) \times \gamma + \beta \times I(u,t)
\]  

(9)

Where I(u,t) is user influence ranking score within time period t, it could be user influence measurement defined in equation (4) – equation (8); \( \gamma \) is the decay factor for the effect of user influence declines with time moving on and \( \beta \) is a constant parameter for scaling.

### 4 Empirical Analysis

To investigate the performance of our approach, we implement experiments on two data sets: empirical data set and artificial data set. In the artificial data set, the following network is real but the retweet and reply actions are generated artificially.

#### 4.1 Data Collections

We crawled the nearly entire sina micro-blog following network, the biggest micro-blogging platform in China, at that time by using the inner user ID which is a 10 digit number. A network with 41 million users and 2.29 billion edges is obtained. For simplicity, we extract a sample network which contains 133145 users and 2027048 edges by snowball sampling with some random seeds (in our case we use 5 seeds) from sina micro blog following network we crawled. We analyze the sample data and find our sample network has the similar degree distribution with the sina micro blog following network. Furthermore, we crawled all the tweets published by users in the sample network from April 18th to April 24, 2011. There are 25912 users who published 409732 tweets in this week. 36.76% tweets are original tweets, and 63.4% tweets are retweets. In average, each user published 5.81 original tweets and 10 retweets. It seems users are more likely to repost information, which conforms to the concept that micro-blogging service is a social medium.

In addition to using the empirical tweet data we collected, we also generate an artificial tweet data set. Since at the time we crawled the tweets, sina micro-blogging API did not provide methods to trace the tweets’ full diffusion paths. In order to examine the details and the key players along a tweet’s diffusion process, based on the sample network, we generate artificial tweets, “retweet” and “reply”
actions for testing effectiveness of our approach. Similar to retweet and reply mechanisms in Twitter, in the experiments, one user can read all the tweets published by his followees and he becomes an explicit effective audience only when he is in the artificial retweet or reply networks. Artificial tweets and their corresponding explicit actions are generated based on the following two assumptions.

**Assumption 1** The number of tweets being originally written by users in a time period follows power law distribution.

**Assumption 2** The following network structure is static during one time period.

Assumption 1 is inspired from empirical studies. The number of tweets posted by users is not evenly distributed. Twitter statistics reported by beevolve.com show that more followers more tweets. Empirical studies indicate that following network of micro-blogging services demonstrates power law characteristics (Kwak et al. 2010, Weng et al. 2010). In addition, our empirical study showed that the degree distribution of sina weibo following network is power law distribution.

Assumption 2 is for simplicity. Although the following network changes dynamically with time, but since our models mainly focus on the influence in a time period, it makes sense to assume the following network is static in that time period.

We randomly select 1182 users from the sample network, and assume that each user publish 5 tweets in average (in our empirical data, the average is 5.81), and by applying power law distribution for tweets generation, totally 5182 original tweets are generated. Furthermore, we generate “retweet” and “reply” actions for each tweet as follows: the retweeting probability of an original tweet \( w \) written by user \( u \) draws from \([0, 0.1] \) uniformly, and its replying probability draws from \([0, 0.05] \) uniformly. Moreover, the retweeting probability of a retweet of user \( v \) (where user \( v \) retweet \( w \) directly from user \( u \) ) decays with a factor 0.8, i.e., draws from \([0, 0.08] \). In another word, the retweeting probability of a retweet decays along with the distance from the original tweet through its information diffusion path (in retweeting networks). Replying probability follows the same decaying rule. In our experiments, scaling factor \( \alpha \) in \( p(i, u, w) \) is set to 0.1.

### 4.2 The Experiments on Tweet Influence

As shown in section 2, firstly a retweet network for a tweet can be built from the tweets data set we collected. An example of retweet network of a tweet is shown in figure 2. We can see this tweet is retweeted many times. It is easy to get replies of tweet \( w \), since these comments or replies are stored with tweet \( w \). By applying tweet influence algorithm presented in section 2, we can get the tweet influence results in Table 1 based on the empirical tweet data. For privacy, we recode the sina real 10-digit user id into an artificial id.

![Figure 2. The retweet network of a tweet](image)
<table>
<thead>
<tr>
<th>Tweets</th>
<th>User</th>
<th>User Type</th>
<th>No. Retweets</th>
<th>No. of followers</th>
<th>Effective Audience</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>U1</td>
<td>News media</td>
<td>4</td>
<td>3285</td>
<td>1488</td>
</tr>
<tr>
<td>T2</td>
<td>U2</td>
<td>Scholar</td>
<td>114</td>
<td>4279</td>
<td>1038</td>
</tr>
<tr>
<td>T3</td>
<td>U1</td>
<td>News media</td>
<td>38</td>
<td>3285</td>
<td>273</td>
</tr>
<tr>
<td>T4</td>
<td>U1</td>
<td>News media</td>
<td>40</td>
<td>3285</td>
<td>261</td>
</tr>
<tr>
<td>T5</td>
<td>U1</td>
<td>News media</td>
<td>23</td>
<td>3285</td>
<td>256</td>
</tr>
<tr>
<td>T6</td>
<td>U1</td>
<td>News media</td>
<td>73</td>
<td>3285</td>
<td>207</td>
</tr>
<tr>
<td>T7</td>
<td>U3</td>
<td>Entertainment</td>
<td>59</td>
<td>4803</td>
<td>187</td>
</tr>
<tr>
<td>T8</td>
<td>U4</td>
<td>Comedian</td>
<td>5</td>
<td>9664</td>
<td>161</td>
</tr>
<tr>
<td>T9</td>
<td>U5</td>
<td>Enterpreneur</td>
<td>23</td>
<td>3388</td>
<td>159</td>
</tr>
<tr>
<td>T10</td>
<td>U6</td>
<td>News media</td>
<td>24</td>
<td>3478</td>
<td>158</td>
</tr>
</tbody>
</table>

**Table 1. Top 10 influential tweets in empirical tweet data**

As shown in Table 1, we can see those tweets published by news media tend to have more effective audiences and so it is for celebrities. Our proposed tweet influence measurement does not strongly depend on number of retweets or the number of followers, but emphasize tweets’ diffusion paths. It is strongly related to the total number of followers along the information diffusion paths, since those followers are the main resources of the implicit effective audiences. Our main point of measuring tweet influence is how many users have read the tweet. The information diffusion process is most important for tweet influence rankings.

Figure 3 illustrates a tweet’s diffusion process. User002 is crucial to this tweet’s diffusion, since he can bring 1829 effective audiences although he has only three followers. The red dot line indicates the key information diffusion path of this tweet. From table 1 and figure 3, we can conclude that the number of retweets is not dominant to the influence of the tweet, but those who retweet these tweets are quite important, since they could bring more effective audiences, including implicit and explicit ones.

![Figure 3. A tweet’s main diffusion paths](image)

Note: The number before left parenthesis denotes user id, and the three numbers in the parentheses represents the number of new effective audiences the user can bring cascaded, the number of effective audiences the user can bring itself directly, and the number of followers of the user respectively.
Figure 4 presents an example to illustrate the key players in the process of information diffusion. It shows that sometimes who retweets the tweet are more important for tweet influence than the total number of retweets. In figure 4, there are two communities. User 1 posts a tweet and it is retweeted by user 2, 3, 4 and 5, but only the retweet of user 5 is crucial since he could pass the information to a new community which contains user 6, 7, 8, 9 and 10. Our proposed tweet influence approach could deal with this scenario.

4.3 The Experiments on User Influence

To investigate how our user influence measurement is effective in indentifying the influential user, we compare ranking results of our user influence measurement (measurement 1) with other 4 user influence measurements, i.e. number of followers, page rank on following network, number of retweets and page rank on retweet network. In addition, we also tested measurement 2 and 3, there’s only a marginal difference between them.

**Experiment 1:** Use sample following network, and artificial tweets data set.

Table 2 presents the comparisons of ranking results of experiment 1. The first row lists the most influential user ranked by our measurement, and this user is ranked as 13, 43, 5, and 333 by the other four user influence measurements: no. of followers, PageRank on following network, no. of retweets and PageRank on retweet network respectively. From Table 2, we can see ranking results of our proposed measurement are quite different from others (see the number with bold font). Our No. 4 influential user is ranked as 556 with PageRank retweets measurement. We find this user has many retweets (its no. of retweets ranking is 2). But most of his retweeters have high in-degree, in another word, his retweeters also retweeted many other users’ retweets, which abates his rankings of PageRank retweets. In row 9, the user is ranked as the 9th most influential user by our measurement 1. It has relatively fewer followers (its no. of following ranking is 285, its no. of PageRank following ranking is 152), but it has relatively moderate retweets (its no. of retweets ranking is 34). It indicates that our user influence could measure something new and integrates both retweets and following network together.

<table>
<thead>
<tr>
<th>Rankings of measurement1</th>
<th>Rankings of no. of followers</th>
<th>Rankings of PageRank followings</th>
<th>Rankings of no.of Retweets</th>
<th>Rankings of PageRank retweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>43</td>
<td>5</td>
<td>333</td>
</tr>
<tr>
<td>2</td>
<td>128</td>
<td>9</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>39</td>
<td>6</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>64</td>
<td>2</td>
<td><strong>556</strong></td>
</tr>
<tr>
<td>5</td>
<td>78</td>
<td>50</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>280</td>
<td>11</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>188</td>
<td>57</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>79</td>
<td>59</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td><strong>285</strong></td>
<td><strong>152</strong></td>
<td>35</td>
<td>34</td>
</tr>
<tr>
<td>10</td>
<td>73</td>
<td>80</td>
<td>12</td>
<td>66</td>
</tr>
</tbody>
</table>

Table 2: Top 10 users ranked by our user influence measurement (Measurement 1), and their rankings of no. of followers, PageRank on following network, no. of retweets and PageRank on retweet network respectively
Moreover, we use Spearman’s rank correlation coefficient to illustrate the relative influence ranking of our measurement 1 with other four measurements. Spearman’s rank correlation coefficient is a non-parametric measure of statistical dependence between two variables, and it is computed as $\rho = 1 - \frac{6 \sum (x_i - y_i)^2}{N^3 - N}$.

<table>
<thead>
<tr>
<th>Correlation with measurement 1</th>
<th>All users</th>
<th>Top 300</th>
<th>Top 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>no. of followers</td>
<td>0.92</td>
<td>0.30</td>
<td>0.17</td>
</tr>
<tr>
<td>PageRank followings</td>
<td>0.90</td>
<td>0.56</td>
<td>0.30</td>
</tr>
<tr>
<td>no. of retweets</td>
<td>0.99</td>
<td>0.83</td>
<td>0.76</td>
</tr>
<tr>
<td>PageRank Retweets</td>
<td>0.89</td>
<td>0.54</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Table 3.  Spearman’s correlation coefficient of rankings

Table 3 shows the correlation coefficients of rankings between our proposed user influence measurement and other four user influence measurements. Our user influence measurement has a very high correlation (almost above 0.9) with other classic user influence measurements for all users. But as discussed by Cha et al., the high correlation appears to be an artefact of the tied ranks among the least influential users. In the top 100 influential users, our model still has a high correlation with number of retweets measurement. We conjecture that the reason is that the retweet actions constitute information diffusion paths, and our model emphasizes information diffusion, so the two measurements have a strong correlation. But as shown in figure 4, our user influence model does not treat each retweet equally but takes the tweeters’ following network into account. Since a retweeter with more followers could have a higher probability to get more effective audiences (both implicit and explicit ones).

Table 4 gives user overlap in top 100, 200, and 500 influential users ranked by our model and the other four measurements. From Table 4, we can see the overlap of top 500 ranked by our model and no. of retweets is 99%, it mean only 4 users are different, but in top 100, there are 18 different users. Table 3 and Table 4 give descriptions of differences between our user influence measurement and the other four measurements, and the results show that our proposed user influence measurement (measurement 1) is highly correlated with the number of retweets measurement.

<table>
<thead>
<tr>
<th>Overlap with measurement 1</th>
<th>Top 100</th>
<th>Top 200</th>
<th>Top 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>no. of followers</td>
<td>0.41</td>
<td>0.58</td>
<td>0.90</td>
</tr>
<tr>
<td>PageRank followings</td>
<td>0.58</td>
<td>0.64</td>
<td>0.84</td>
</tr>
<tr>
<td>no. of retweets</td>
<td>0.82</td>
<td>0.86</td>
<td>0.99</td>
</tr>
<tr>
<td>PageRank Retweets</td>
<td>0.60</td>
<td>0.65</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 4.  Percentage of user influence ranking overlap between our proposed (measurement 1) and other four measurements

**Experiment 2:** Use sample following network and empirical tweets data set.

For comparisons, in experiment 2, we measure the influence of 1182 users selected in experiment 1. Because the retweet networks of empirical tweets are quite sparse, in Table 5, we omit PageRank retweets measurement. In addition, our empirical tweets data set does not include the information of replying actions.

As shown in Table 5, our model gives quite different rankings in comparison to number of followers and pagerank following networks measurements, but correlates with number of retweets measurement. Our approach concerns the process of information diffusion, and “retweeting” is a main information diffusion way on micro-blogging platforms, it makes sense that our user influence measurement correlates with number of retweets measurement. In addition, our approach takes diffusion paths into
account and weighs up the importance of retweets, which makes our approach different with that of counting number of retweets only. For example, in row 6, the user has a high ranking (no.6) by our measurement. Although he has relative little followers (ranked as 552), and relatively moderate retweets (ranked as 39), but along the retweeting information diffusion path, it has large potential audiences. In row 10, the user has relatively moderate followers and retweets, but it could be identified as an influential one by our measurement. In experiment 2, the top 100 influentials’ rank correlation coefficient between our user influence measurement and number of retweets measurement is 0.39.

<table>
<thead>
<tr>
<th>Our measurement 1</th>
<th>Rankings of no. of followers</th>
<th>PageRank on following network</th>
<th>Rankings of no. of Retweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>109</td>
<td>55</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>139</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>60</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>451</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>81</td>
<td>167</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>552</td>
<td>350</td>
<td>39</td>
</tr>
<tr>
<td>7</td>
<td>188</td>
<td>147</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>279</td>
<td>136</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>21</td>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>38</td>
<td>82</td>
<td>45</td>
</tr>
</tbody>
</table>

*Table 5. Top 10 users ranked by our user influence measurement (Measurement 1), and their rankings of no. of followers, PageRank on following network and no. of retweets respectively*

5 **FURTHER DISCUSSION**

We measure tweet influence in the view of information diffusion. Along tweet diffusion paths, both implicit audiences and explicit audiences are taken into account for measuring tweet influence. But it is very hard to estimate how many audiences have read the tweet without explicit actions. In this paper, a simple reading probability $p(i, u, w) = \frac{u}{|\text{followee}(i)|}$ is used to estimate the implicit effective audiences of $w$ published by user $u$. The reading probability has big impact on the number of implicit effective users and then affects tweet influence measurement a lot. In our future work we will try to apply different reading probability settings to examine its effects on tweet influence.

There are tons of tweets in micro-blogging, measuring every tweet’s influence is not possible. Since the basic building block of our user influence approach is tweet influence, our approach fits to examine a user’s influence in a specific topic or an event. Our approach could be used to estimate the effects of advertisements and find the diffusion patterns of different kinds of tweets.

Our approach for measuring influence takes following network structure and action-based network structure into considerations. Unlike TURank which consider both following relationship and retweet actions (Yamaguchi and Takahashi 2010), in which they built user-tweet graph, and used ObjectRank to compute user influence, our approach is not only present a ranking score, but a set of exact effective audiences. Our approach measures user influence in tweet level, which provides a good flexibility to study user influence according to different topics and time periods.

6 **CONCLUSION**

This paper proposes user influence measurements based on tweet influence along information diffusion paths. We argue that user does not read all the tweets he subscribes to, hence, a user probably cannot influence all his followers. Only measuring user influence through the number of
followers is not accurate. Moreover, any explicit actions (“retweet” or “reply”) do mean that user read the contents of the tweets, but there are some followers do read tweets without taking any explicit actions. So only measuring explicit actions could not reflect user influence comprehensively.

Our user influence approach considers both explicit effective audiences and implicit effective audiences, which gives a more accurate estimation of user influence. The main contribution of our paper is that we combine following network structure and action-based network structure together to measure tweet influence and user influence. In addition, micro-blogging service is an ecological system and user influence evolves as time goes by. And user influence could be different on different topics. Our proposed user influence measurements could be used to measure user influence on different topics (such as Measurement 4,5) and across the timeline (such as Measurement 6).

Experimental results show that our proposed model could identify some influential users who cannot be found by other influence measures, and our models could fix the overlapped audiences problem, find key players on the tweet diffusion paths. Our approach could provide deep influence investigations in the view of information diffusion.

Our future work includes: 1) applying our models to more empirical data sets and more experiments need to be done to evaluate influential tweets and users; 2) using the other popular information diffusion models to explore the implicit effective audiences; 3) measuring user influence on different topics over time; and 4) finding the characteristics of the most influential tweets and users.

In this paper, we focus on measuring user influence based on interactions among users, implicit effective readers without giving explicit actions and contents of micro-blogging platform. In future, we need to consider more scenarios and do more experiments to find out what factors could make a user or a tweet popular.

References


