HOW DO USER BEHAVIORS AFFECT INFORMATION PROPAGATION IN TWITTER

By

YU-SONG SYU

A THESIS PRESENTED TO THE GRADUATE SCHOOL OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

UNIVERSITY OF FLORIDA

2012
I dedicate this thesis to my dearest mother. Wish her the best in heaven.
ACKNOWLEDGMENTS

Thanks to all the help I have received in writing and learning about this thesis.

Thanks to Dr. Thai for the advices during the two years. Thanks to Yilin, who always helps me when discussing, doing simulation, and writing articles. Also thanks to all the lab members.

Finally, greatest thanks to my family. Their support is my best power to keep going on the way pursuing the degree in the States.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>INFORMATION DIFFUSION IN TWITTER</td>
<td>12</td>
</tr>
<tr>
<td>2.1</td>
<td>Datasets</td>
<td>12</td>
</tr>
<tr>
<td>2.2</td>
<td>The Operating Mechanism and Information Diffusion Model</td>
<td>12</td>
</tr>
<tr>
<td>2.3</td>
<td>Building Cascades of Information</td>
<td>14</td>
</tr>
<tr>
<td>2.4</td>
<td>The Selection of Topics</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>THE EFFECT OF MULTIPLE RETWEETS IN TWITTER</td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>Longer Range Than Expected</td>
<td>20</td>
</tr>
<tr>
<td>3.2</td>
<td>Every Retweet Matters in Twitter</td>
<td>22</td>
</tr>
<tr>
<td>3.3</td>
<td>Timeliness of News in Twitter</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>TIME INTERVAL BETWEEN MULTIPLE RETWEETS</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>TIME INTERVAL BETWEEN CASCADE LEVELS</td>
<td>28</td>
</tr>
<tr>
<td>5.1</td>
<td>Distribution of Time Interval between Cascade Levels</td>
<td>28</td>
</tr>
<tr>
<td>5.2</td>
<td>Informed Users at Each Level</td>
<td>30</td>
</tr>
<tr>
<td>5.3</td>
<td>Time interval between Cascade Levels</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>USER ACTIVENESS</td>
<td>32</td>
</tr>
<tr>
<td>7</td>
<td>DISCUSSION AND FUTURE WORK</td>
<td>34</td>
</tr>
<tr>
<td>7.1</td>
<td>More on Time Needed to be Retweeted</td>
<td>34</td>
</tr>
<tr>
<td>7.2</td>
<td>Probability of Being Retweeted</td>
<td>34</td>
</tr>
<tr>
<td>7.3</td>
<td>Future work</td>
<td>35</td>
</tr>
<tr>
<td>8</td>
<td>RELATED WORK</td>
<td>37</td>
</tr>
<tr>
<td>9</td>
<td>CONCLUSION</td>
<td>38</td>
</tr>
<tr>
<td>REFERENCES</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>BIOGRAPHICAL SKETCH</td>
<td></td>
<td>43</td>
</tr>
</tbody>
</table>
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Dataset Details</td>
<td>12</td>
</tr>
<tr>
<td>2-2</td>
<td>Details of Topics</td>
<td>17</td>
</tr>
<tr>
<td>5-1</td>
<td>Time Interval between the $1^{st}$ and $2^{nd}$ Level (hr)</td>
<td>31</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>The UIIS Model</td>
<td>14</td>
</tr>
<tr>
<td>2-2</td>
<td>An Example of a Cascade</td>
<td>15</td>
</tr>
<tr>
<td>2-3</td>
<td>The Direction of Information Diffusion</td>
<td>16</td>
</tr>
<tr>
<td>2-4</td>
<td>Number of Times to Retweet and to be Retweeted of a User</td>
<td>19</td>
</tr>
<tr>
<td>3-1</td>
<td>Distribution of Range</td>
<td>21</td>
</tr>
<tr>
<td>3-2</td>
<td>The Average Influence of Each Retweet Per User</td>
<td>23</td>
</tr>
<tr>
<td>3-3</td>
<td>The Newly Influenced Followers by Each Retweet</td>
<td>24</td>
</tr>
<tr>
<td>3-4</td>
<td>Time Needed to be Retweeted</td>
<td>24</td>
</tr>
<tr>
<td>4-1</td>
<td>Time Interval between Retweets</td>
<td>27</td>
</tr>
<tr>
<td>5-1</td>
<td>Distribution of Time Interval Between Levels</td>
<td>29</td>
</tr>
<tr>
<td>5-2</td>
<td>Accumulative Size of Audience by Each Hop</td>
<td>30</td>
</tr>
<tr>
<td>6-1</td>
<td>Activeness of Users</td>
<td>33</td>
</tr>
<tr>
<td>7-1</td>
<td>The Distribution of Time Needed to be Retweeted</td>
<td>35</td>
</tr>
<tr>
<td>7-2</td>
<td>The Probability of Being Retweeted</td>
<td>36</td>
</tr>
</tbody>
</table>
Abstract of Thesis Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Master of Science

HOW DO USER BEHAVIORS AFFECT INFORMATION PROPAGATION IN TWITTER

By

Yu-Song Syu

December 2012

Chair: My T. Thai
Major: Computer Engineering

Online social networks (OSNs) have become an imperative channel for information propagation and influence such as advertising products in viral marketing or convincing voters in a political campaign. Knowledge of the intrinsic understanding of this phenomena in OSNs provides us not only the user behaviors and their mutual impacts, but also key insights into designing better advertisement strategies. However, most of related works have only focused on the analysis of specific propagation mechanism in certain networks and largely ignored two important factors: the propagation behaviors between users in general social networks and the crucial factors affecting information propagations.

In this paper, we have demonstrated that: 1) multiple retweets have an unignored effect on enlarging the size of diffusion cascades and extending the propagation chains. If we extract cascades with respect to topics rather than messages based on the operating mechanism of Twitter, we can find a considerable number of cascades with ranges larger than 10 hops. Multiple retweets also keep the news fresh by allowing people sharing new different information about one topic, so that a news propagation is dynamic process rather than a static story; 2) the time interval between two consecutive retweets of one user clearly indicates the tendency that whether this user will retweet about the same topic many times; and 3) the time interval between the first two levels in a cascades helps us forecast the this cascade will conduct a long lifetime period or not.
The rapid growth of Online Social Networks (OSNs), such as Facebook, Twitter, and Google+ [1, 2, 6], has made them become one of the most important media for fast information propagations [15, 20]. According to the statistics, there are 2.9 million followers following Britney Spears in Google+ and more than 526 million users log in Facebook everyday [7, 8]. Using this popular channel, many individuals and companies can spread their own messages or advertise their products by leveraging the power of others' influences. For instance, Twitter drew 456 tweets per second about the dead of Michael Jackson [5]; the 2009 president election in Iran reached 221,744 tweets per hour on June. 16\textsuperscript{th} [4]. Therefore, it becomes an urgent need to exploit an intrinsic understanding of the information propagations in OSNs and how they are affected by user behaviors.

Recently, many works have been proposed to investigate the phenomena of message spreading in OSNs, focusing on the characteristics (e.g., range, scale, and speed) of information diffusion and its relations to network structure and traffic characteristics. Among which, Gómez et al. [16] studied the reactions in Slashdot social networks. Based on the radial tree generated by the nesting of comments, some interesting properties were found such as self similarity within the different nesting levels of a discussion. However, a user can rate a post only once in Slashdot, which does not necessarily happen in other social networks. Later on, Cha et al. [13] explored the properties of information diffusion in Flickr, which concluded that the propagation cascade is “neither wide nor tall”. That is, both the number of audience and the number of propagation hops are small. Yet, their constructions of cascades does not guarantee the information flow. In Digg network, Steeg et al. [29] further analyze the reason why the propagations will stop, in which the concept “saturation” was first introduced. Particularly, when a user knows the information through a lot of friends, the network is
called saturated and, therefore, no further information spreadings are needed. Again, the assumption that a user can only forward a topic once made the results unrealistic and less convincing. Therefore, the above works only focused on some mechanisms of message disseminations in specific networks, yet neglected the propagation behaviors between users in other but more general social networks (i.e., Twitter and Facebook) and the crucial factors in user behaviors which heavily affect information propagations.

In this paper, we focus on the analysis of the message traces in Twitter, in which user behaviors are more general and practical. In particular, after reading a post, a user can either use the retweet or mention mechanism to further spread it out to other users. Also, a topic can be posted multiple times and a post can retweeted or mentioned more than once. In order to propose a successful campaign strategy, our analysis focuses on the impact of user behaviors to information propagations, mainly addressing the following three questions: (1) how can we predict the propagation tendency of a tweet in social networks at its early stage? (2) how possible will a post be retweeted? If so, how long will it take? (3) will a user retweet a specific topic frequently if he is usually not active in social networks?

First, we compile the traces in Twitter by using two datasets: one is collected by Yang et al. in [32] consisting of 476 million tweets (including contents) by 17 million users in 2009; and the other one from Kwak et al. in [20], which describes the network topology with 41.7 million users and 1.47 billion social relations (the “following” relationships between users). To understand the information diffusion in Twitter, we select two hot topics from the “Top Tweeter Trends in 2009” [5] to demonstrate our observations.

Our findings can be summarized as follows:

- The influence of multiple retweets: The distribution of cascades range doesn’t follow the “2 to 6 hops” assumption, which has been widely proved and admitted when we extract cascades with respect to topics rather than messages. Multiple retweets about one particular topic won’t cause the information saturation in Twitter
but enlarge the influence and extend cascades size. Besides, multiple retweets can shorten the time cost to be retweeted by other users.

- **The interval** of multiple retweets one user made and its relation to the information propagation: Contrary to our intuition, if the interval between first two retweets is small, we find that this user is very likely to make more retweets regarding the same topic. That is to say, the time interval between the first two retweets indicates the tendency about whether the user would like to make more retweets or not. In addition, after the third retweet, the time intervals among the rest of retweets are almost the equal.

- **The interval** of each level among the cascades and its relation to the information propagation: The quicker the first level audience pick up the news from the seeds and forward it out, the larger the cascade range will be. This finding tells us that the time interval between the first two levels can be served as a indicator to predict the whether this seed can generate a large cascade or not.

- The influence of the **interest** of users: considering a post with a specific topic, a user will retweet it only if he is active and keeps retweeting in all kinds of topics in OSNs. That is, there if very few such users who is only interested in one topic and retweet only the posts containing this topic.

The rest of the paper is organized as follows. We describe our measurement methodology, introduce the dataset we use, and explain the selection of the two popular news for analysis in Chapter 2. We describe our analysis and findings in Chapter 3, 4, 5, and 6. Chapter 7 discusses the insights by the our findings and possible future work. We summarize related work in Chapter 8 and conclude in Chapter 9.
CHAPTER 2
INFORMATION DIFFUSION IN TWITTER

In this chapter, we describe the datasets we use, delineate the operating mechanism and diffusion model in Twitter, and then introduce our methodology to construct the information cascades during the diffusion process.

2.1 Datasets

In this paper, we use two datasets [20, 32] for the analyses of user behavior and information diffusion in Twitter. The first dataset in [32] is collected by Yang et al., which consists of 467 million tweets posted by 17 million users from June to December in 2009. This dataset provides the time, author, and content of each tweet, but it does not provide any information about the relationship between users, i.e., who follows whom on Twitter. Thus, we also use another dataset in [20] in order to construct the network topology. In [20], Kwak et al. collected the 1.47 following relationships between 41.7 million users on Twitter.

2.2 The Operating Mechanism and Information Diffusion Model

Here we first delineate the operating mechanism in Twitter, and then introduce the information diffusion model that we use in this paper.

On Twitter, people can post a paragraph (within 140 words) on his own page or reply other users’ tweets. Besides tweeting and replying, there are three important user-user interactions that one user may take on Twitter, namely follow, mention, and retweet. These are the three important actions that makes information propagated in Twitter. Users interact in Twitter by first becoming “followers” of one another, after which they can then see all the posts made by whom they follow on their own Twitter page. When users read a piece of news, they may want to further forward to their followers.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Tweets</th>
<th>#Users</th>
<th>#Edges (Followings)</th>
<th>Directed/Undirected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang</td>
<td>467 million</td>
<td>17 million</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kwak</td>
<td>106 million</td>
<td>41.7 million</td>
<td>1.47 billion</td>
<td>Directed</td>
</tr>
</tbody>
</table>
By typing *RT @username* or *via @username* followed by the content, they can indicate the source of this post; similarly, the post will show on all their followers’ page. A user can specify who to see the post by using “mentioning”. Note that a tweet can be both a retweet and a mention. See the second tweet in Figure 2-3 as an example.

Twitter’s operating mechanism is different from some other OSNs (e.g., Slashdot and Digg). After reading one tweet, a user can retweet it for several times. Such mechanism is used in other OSNs, such as Facebook and Google+. In fact, this is more general not only it is more widely used in trending OSNs, but also because we can consider the mechanism that Slashdot and Digg use is a special case of Twitters. Such generality is important because it largely affects the propagation model and performance would be different.

Based on the operating mechanism, we would like to analyze users’ behaviors in Twitters. Because a user can forward (retweet) on a topic more than one time, it is not applicable to use the widely used *inactive-active* and *Susceptible-Infected-Recovered* in which a user has only one chance to influence his followers to prolong the diffusion process for one hop further. Thus, we propose the *Unknown-Informed-Influential-Stopped* (UIIS) model to represent user behaviors in Twitter. During the diffusion process on one topic in Twitter, we find that a user has four status:

1. *Unknown*: the initial state, at which the user does not know the news yet.
2. *informed*: the user knows the news by reading a tweet or being mentioned by other users, but has not further retweet the news yet.
3. *influential*: the state that the user has further forwarded this news out.
4. *stopped*: the final state that the user has stopped retweeting, and will not retweet anymore.

Figure 2-1 demonstrates the transitions between these four states. It is important to have knowledge of this transition model because it is what our analyses in the following sections rely on.
Figure 2-1. The UIIS model. This represents the four possible states in which a user in Twitter can be. At the beginning, the unknown user has not known the news yet. He then becomes informed when he knows the news by reading a tweet or being mentioned by others. If he finds this topic interesting and further retweets it out, he then becomes influential. The user switch to the final state stopped once he decides not to forward anymore about this topic.

2.3 Building Cascades of Information

After a user tweets about one topic, the information diffusion begins when another user reads this tweet and then retweets it, and continues when there are some new users who see the retweets and then further retweets them. We define a cascade by collecting all users that have ever participating the diffusion process and the flows of the information, starting from the very first user that posts the earliest tweet. To illustrate the flow of each cascade, initialized by a single user, we build a hierarchical tree based on [17] and modify it to fulfill our need. We call such a tree a “cascade”. In a cascade, all nodes are connected and each node has only one parent, except the root, which does not have a parent. In such a tree, we name the root as the initiator, who is the first user in this cascade. Starting from the initiator, the existence of each edge $i \rightarrow j$ represents the information flow from $i$ to $j$. We then name the leaves as the receivers, and all other nodes as the spreaders, as Figure 2-2 illustrates.

The methodology that we use to construct a cascade is as follows. First we have to identify the tweets related to one topic. In Twitter, users are allowed to explicitly identify topics when tweeting by using hashtags, a word starting with “#” (e.g., #H1N1,
An example of a cascade of a piece of news. An arrow from node $i \rightarrow j$ means that the information flows from $i$ to $j$. We mark the initiator as a star, the spreaders as circles, and the receivers as squares. The dashed arrow $2 \rightarrow 5$ indicates that although 2 tells 5 about the news, this has no effect because 5 has already known it from 1. This cascade terminates when no node further spreads out the topic. In this example, the range of this cascade is 3

This makes the topic identification applicable. However, the majority of users do not use hashtags (only about 1/10 tweets have hashtags [10]). Thus, we then need to determine topics for tweets without hashtags. We first identify the set of keywords by surveying news websites and the tweets in our dataset. We then identify relevant tweets by searching for the keywords in the tweets.

For one topic, the next step is to find out all the cascades. We first find out all initiators who “have never retweeted any other users or been mentioned by other users, before posting the first tweet about the topic”. Starting from each initiator, we then find the edges to another users. In our construction of cascades, the existance of an edge $i \rightarrow j$ represents: 1) $i$ mentions $j$, 2) $j$ retweets $i$, or 3) $i$ makes a post and $j$ follows $i$. Thus we have the first level of users in this cascade. We then find the edges from such users to another users by the same method, until we cannot find a new participating user anymore.

One may ask: “Why is the direction of diffusion different between retweeting and mentioning?” When $i$ mentions $j$, it explicitely indicates that $i$ posts an article and hopes
that twitter would make $j$ see that, so the information flows from $i$ to $j$; when $i$ retweets $j$, it implicitly means that $i$ reads a previous post on $j$’s page and then forward it out (to $i$’s followers), so the information flow is from $j$ to $i$, as Figure 2-3 shows. One may then ask: “What if there are multiple users that provide information to the same user about the same topic?” Our answer is, under this circumstance, we select the earlist edge for that it is the earlist.

Note that, despite that a hierarchical tree has a single root (initiator), there may be multiple initiators who share the same news concurrrently. We then build a tree for each initiator. For the completeness of the analysis on the tree structure and the behaviors of the initiaotrs, we make it possible that a node appears in multiple trees. Moreover, in this paper, the range indicates the maximum level of a hierarchical tree, and the audience of one topic means all the users that appear in the hierarchical trees about this topic, except the initiators.

![Figure 2-3. The direction of information diffusion. The information about the concert flows from A to B when B retweets A; The information about the video by D flows from D to C when D mentions C.](image)

### 2.4 The Selection of Topics

We then choose represenatative topics for case study. In 2009 that our datsets cover, there are 70 most popular topics discussed on Twitter $[5]$. Among which we find us able to roughly divide them into two genres. One is the more serious “hot-news” genre (e.g., iranelection, michaeljackson, swineflu), and one is the more leisure
“entertaining” genre (e.g., Transformers 2, A-Rod, Lakers). We choose one topic from each genre, and first make preliminary observations on their characteristics w.r.t. information diffusion. The chosen topics are: the election in Iran and the 2009 National Basketball Association (NBA) champion - L.A. Lakers. For simplicity, in this paper, we use iranelection and lakers to represent these two topics respectively.

These two topics are different from many aspects. First, iranelection is a serious political topic; lakers is a leisure sports topic. Second, iranelection happens in Asia and spreads all around the world; lakers happens in the States and is mainly discussed in the areas where basketball is popular. Third, iranelection has new events updated consequently, including the result, protests, and arrests; lakers happened on June 14th, and then no major issues comes out because of the end of the season.

We first summarize the numbers of tweets, retweets, and mentions about both topics in Table 2-2. Besides the difference in their scales, their user behaviors during the diffusion process are also different. iranelection draws much more retweets than mentions (3.2 times), while for lakers, the number of mentions is moderately (24%) larger than retweets. This can be explained by the nature of the topics: for more serious news, one may intend for simply retweeting an article (rather than identifying a specific user) so that all followers can see it; while for a more leisure topic, because the interested audience varies with different topics, one may want to “mention” a particular friend, who are known to be interested in that topic.

For both topics, the number of “retweet” is a innegligible measure (around 1/4 of the total tweets). Thus, we then would like to understand the retweeting pattern of users in Twitter. We show the distribution of the number that users retweet and be retweeted in

<table>
<thead>
<tr>
<th>Topic</th>
<th>#Tweets</th>
<th>#RTs</th>
<th>#mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>iranelection</td>
<td>408347</td>
<td>97685  (23.9%)</td>
<td>30163 (7.3%)</td>
</tr>
<tr>
<td>lakers</td>
<td>74051</td>
<td>20918 (28.3%)</td>
<td>25989 (35.1%)</td>
</tr>
</tbody>
</table>
Figure 2-4. In both 2-4A and 2-4B, the distributions roughly follows a long-tail shape. That is, most users retweet and be retweeted very few times about one topic, and most of the retweets refer to and are posted by very few users. Moreover, these figures also show how the topic #iraneelection and #lakers differ from each other in the aspect of the traffic amount of retweets (by roughly an order of magnitude).

In this paper, we focus on the information diffusion process about popular topics on Twitter. Specifically, we study the relation between the range of cascade and user behavior during the diffusion process. We believe our analysis demonstrates how user behaviors affect information spread through Twitter. We then present our observations in the following sections.
Figure 2-4. Number of times to retweet and to be retweeted of a user, in terms of Complementary Cumulative Distribution Function (CCDF). Both roughly show long-tail shapes.
CHAPTER 3
THE EFFECT OF MULTIPLE RETWEETS IN TWITTER

As mentioned in Section 2.2, multiple retweets made by one user on a specific topic features the operating mechanism of Twitter. In this chapter, we represent the effect of such mechanism on the information propagation in Twitter.

3.1 Longer Range Than Expected

We first observe the length of information propagation in Twitter. It is widely admitted and applied as an assumption in previous work that information mostly propagations for 2 to 6 hops in OSNs [16]. We first check whether such phenomenon can be observed in Twitter.

We use the method described in Section 2.3 to extract cascades for the two topics #iranelection and #lakers selected in Section 2.4 and see how their ranges distribute. The range of a cascade is denoted by the longest path from the initiator to the receivers in it (i.e., from root to leaf). We then use Figure 3-1 to show the distribution of cascade ranges about these two topics. To our surprise, the range distributions follow “2 to 6 hops distribution ” assumption [16]. When the cascade ranges are less than 8, the frequency drops with the increase of range; however, the frequency rises with the increment of range when the range is larger than 10 hops. This is especially apparent in topic #iranelection as shown in Figure 3-1A. Although #lakers does not apparently show this phenomenon, we still can find a considerable number of cascades with range of up to 15 in Figure 3-1B.

Observing this curious phenomenon, we are interested in the reason why this happens in Twitter. After comparing with other OSNs such as Flickr, Digg, and Slashdot, we find out that the operating mechanism of Twitter could be one explanation.

Rather than having only one chance to make a vote for a picture or story, users in Twitter can execute “retweet” and “mention” for several times to make consecutive discussions in terms of one topic. This mechanism is also used in many other social
networks such as Facebook and Google+, where the users can respond to a message many times as long as he is still interested in it. One may then wonder whether such sequential responses really make a difference on information diffusion process in the underlying social network.

Intuitively, we may assume that if a user retweets more about one topic, it would be more possible that his followers are influenced and then extend the propagation chain, thus the phenomenon above is explained. However, one has to be aware of another problem –information **saturation**, which is defined by Steeg et al.: “despite multiple opportunities for infection within a social group, people are less likely to become
spreaders of information with repeated exposure” [29]. To examine whether this would happen along with multiple retweets in Twitter, we further present our investigation in the next section.

### 3.2 Every Retweet Matters in Twitter

We first see the relation between the total number of retweets and the average number of influenced followers by a user when he posts a tweet about one topic. This is to see how differently do users influence their followers, when they retweets more about the same topic. As Figure 3-2 shows, the influence does not increase with the number of retweets of a user as we assumed. That is, we cannot find a “leftbottom-righttop” trend in both topics. In fact, a “lefttop-rightbottom” trend is observed.

Then, we wonder whether the followers’ willing to become influential drops instead as the increment of a user’s number of retweets. If this case, we can expect that it drops to zero as the number of retweets reaches a certain value. To see this, we then count the total number of retweets that each user has made, and then calculate the average number of **newly influenced followers** caused by every retweet. As presented in Figure 3-3, a user’s first retweet influences nearly 2.5 followers with respect to the topic about #iranelection. (Note the difference between Figure 3-2 and 3-3) Such influence drops slightly after the second retweet, and oscillates between 1 and 2 from the third retweet on. The same phenomenon exists w.r.t. the topic #lakers.

Therefore, we find that multiple retweets explains the longer ranges of cascades observed in Twitter. Furthermore, although the influence of the first two retweets differ from each other on a certain scale, the “saturation” does not obviously exist since the number of newly activated users does not drop to zero with more retweets.

### 3.3 Timeliness of News in Twitter

Besides the range, multiple retweets in Twitter also affects the timeliness of the propagation of news, which is one of the most important features to propagate the
Figure 3-2. The average influence of each retweet per user. A user’s average followers influenced by per retweet decreases when this user retweets more times.

news because quicker propagation could help news keep fresh and thus influence more people during the diffusion process.

To demonstrate the effect, we then group the users by the total number of retweets that they have made, and show the relation between the number of retweets and the time that a tweet needs to be retweeted by another user in Figure 3.3, where people with more retweets can been retweeted in a relatively shorter time for both of the topics. This results show that multiple retweets shorten the time to be retweeted by other, i.e., a quicker propagation can be expected.
Figure 3-3. The newly influenced followers caused by a user’s each retweet. The first retweet of a user plays the most important role when influencing followers to further retweet the topic out. The repeating exposure (more than 3 times) does increase the followers’ willingness to become a spreader at the influenced state, but the increment converges to a certain range, rather than to zero or a large number.

Figure 3-4. Time Needed to be Retweeted. A retweet takes longer time to be retweeted if the author does not retweet much about the same topic.
CHAPTER 4
TIME INTERVAL BETWEEN MULTIPLE RETWEETS

In Chapter 3, it is shown that multiple retweets have an unignored effect on enlarging the size of diffusion cascades and extending propagation chains. Then, we would like to figure out the factors that raise users’ willingness of making multiple retweets. As mentioned perviously, timeliness plays an important role in information diffusion, so we try to find a predictor to forecast who would be more likely to retweet more by checking and comparing interval between consequential retweets that a user made. This predictor should be determined in the early stage of the propagation process.

For each user, we examine how the time interval between each two consecutive retweets affects his performance in terms of retweeting. We first define the time interval $l_R^n$ as: the time interval between a user’s $n^{th}$ and the $n + 1^{th}$ retweets about one topic. Then we choose the first 15 retweets for each user to study (i.e., calculating $l_R^1, l_R^2, \ldots, l_R^{15}$), for the reason that: 1) although the number of retweets one user made ranges widely from 1 to 350 regarding one topic, 98% of which are less than 15; and 2) the majority of the other 2% of the users are robots, which automatically retweets on random or specific topics – we have to eliminate these accounts. Additionally, to alleviate the impact of personal difference, we group the users by the number of retweets. This is similar to what we did in the previous section, but this time we divide the users into five groups: the 2 retweets group, 3-4 retweets group, 5-10 retweets group, 11-20 retweets group, and 21+ retweets group, to conduct more detailed analysis. Results are shown in Figure 4-1. Note that for topic #lakers we combine the last 2 groups together because there are only a few users in the 21+ group (as Figure 2-4A shows).

Two important findings are listed in the following. The first finding is that with the decrement and convergence of time interval as the number of retweets increases.
For the topic about #iranelection, we find that the time interval between the first three retweets, i.e., $l_1^4$ and $l_2^3$, are different from the later ones. The time interval of making more retweets drops apparently before the fourth retweet, but later on, the time needed to make the more retweets oscillates within a certain range. This phenomenon truly reflects the diffusion process in reality. When a fresh news story comes out, people may spend time verifying its authenticity: whether this news is related to their lives, etc. Once they feel interested in it, they tend to forward more frequently, and once the users are fully engaged in this topic afterwards, they will continue retweeting news about it in a stable pace. However, we do not see the phenomenon on the topic #lakers in terms of decrement, while the convergence is still observed. This may because of its smaller scale and the earlier saturation. We will discuss the latter conjecture in the next subsection.

Secondly, we find that the relationship between the number of a user’s retweets and the time interval between these retweets. By comparing the time that these five groups spend on making each next retweet, the observation suggests that the time one user spend on making the first two retweets $l_1^4$ can be served as a clear predictor to forecast the potential of making multiple retweets in an early stage. That is to say, if users spend less time on sending the first several retweets, then they enjoy a relatively high chance of talking about the specific topic my makeing multiple retweets in a short time. We attribute this to users’ interests, since users would like to quickly identify and pick up news that they are interested in. Once they are fully engaged in this topic, the development of the events would draw their attention, so that they have more chances to make more retweets.
Figure 4-1. Time Interval between retweets. This figure shows the time that a user needs to make the next retweet. We find two phenomenons here: 1) People retweets more frequently when they gain interests in a topic. This is apparent in *iranelection*, but not in *lakers*; 2) users tend to retweet more if their first few retweets are more frequently made.
CHAPTER 5
TIME INTERVAL BETWEEN CASCADE LEVELS

So far we have observed the attributes of users. In this chapter, we continue discussing the impact of time intervals at each level in a cascade. We find that: 1) The behaviors of the direct friends of the initiators are different from those of other users; 2) The information reaches to majority of the users that it can reach at the early stages; and 3) The interval spent by the direct friends of initiators determines the range of the cascades.

5.1 Distribution of Time Interval between Cascade Levels

Since the timeliness plays a significant role in diffusion unfold process as we have found in the previous sections, and also because the cascades are built based on the order of informed time, it is knee-jerk to assume that users who retweet earlier and quicker will gain larger audience and thus enlarge the whole cascade as a result; that is, the time that the information needs to be propagated from one level to the next should be longer at the higher level (i.e., farther from the initiator) of the cascade. To see the correctness of this assumption, we collect all cascades for topic #iranelection and #lakers, and draw the distribution of the interval between each two consecutive levels. We use $I_n^k$ to represent the interval between the time that a user at level $n$ is informed and the time that he retweets about this topic, so that his followers at level $n + 1$ are informed. The distribution is shown in Figure 5-1.

From Figure 5-1 we have two observations. First, we can see the difference between the first few levels and the others. Take iranelection for example, the first three levels can still be distinguished from one another, but from the fifth level on, the curves are all interleaved together. Topic lakers has similar trend, yet only the first and second levels are distinguishable from others.

Second, by looking at the “most steep section” of each curve representing level 4 to 10, we can still point out that $I_n^k$ increases with $n$ to verify the previous assumption, which
Figure 5-1. Distribution of time interval between levels. Only the curves of the first few levels can be distinguished from one another. After the fifth level, the time difference distribution does not change a lot from level to level. This explains why only the interval between the first few levels more clearly indicates the cascade range.

suggest that it takes longer time to propagate the information if the user is at the higher level in a cascade. However, the differences between these curves are so small that one cannot tell clearly. What leads to such findings? We then take one more step deeper into this problem.
5.2 Informed Users at Each Level

To this end, we calculate the cumulative percentage of informed users in each level to the overall informed users ever, as Figure 5-2 shows. For #iranelection, 80% people in the cascades are already informed about this topic before the fifth level; for topic about #lakers, it informs 80% of the people that can be eventually reached by this topic even within 3 levels of propagation. Hence, we explain why information propagation tends to slow down after be propagated for certain hop (because of the difficulty to inform new friends); we also explain why the curves in Figure 5-1 tend to be indifferent after certain hops (because of the little difference between newly informed users between different levels).

5.3 Time interval between Cascade Levels

As we can see in Figure 5-1, the time interval at the first level $l_1$ diverges greatly. Thus, we wonder whether its value affects the propagation cascade on its range. Here we divide the cascades into groups based on the range of cascades. For a cascade in each group, we would compute the average time that a user first retweets the news
Table 5-1. Time Interval between the 1\textsuperscript{st} and 2\textsuperscript{nd} Level (hr)

<table>
<thead>
<tr>
<th>Cascade Range</th>
<th>&lt;5</th>
<th>5 to 9</th>
<th>10 to 14</th>
<th>&gt;15</th>
</tr>
</thead>
<tbody>
<tr>
<td>hlineiranelection</td>
<td>24.29</td>
<td>6.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lakers</td>
<td>21.08</td>
<td>13.82</td>
<td>7.40</td>
<td>2.21</td>
</tr>
</tbody>
</table>

from the time he was firstly informed\textsuperscript{1} about this topic, i.e., the averaged value of all $I_1^L$s. After that, we find that the time interval between the first and second level is a very important indicator to the range of a cascade: the sooner the first level audience forwards the news, the larger the cascade range will be. This is shown in Table 5-1.

\textsuperscript{1} See the definition in Chapter 2
In this chapter, we discuss the effect of “activeness” of a user in Twitter. In Twitter, we define the activeness of a user as the average number of tweets that he posts hourly (#tweets/hr). Besides, as we can assume, users’ interests could make contribution to the diffusion process on the underlying network. Thus, people would like to broadcast more news about the topics that they are interested in, compared with other news.

In order to see this, we calculate the activeness of a user in general, and compare it with the interest of a user about a specific topic. We measure the user’s interest about a topic by the total number of retweets by the user about the topic. We then observe the relation between these two measures to see their relations in Figure 6-1.

In Figure 6-1, by observing the relations of activeness and interest under different value of x-axis and y-axis, we summarize our observations as follows:

1. The lower parts of both figures are dense, and they do not apparently tend to the left corner or right corner. This shows that a lot of users help propagate the popular topics despite the activeness, if they are not very interested in that topic.

2. More specifically, some active users retweet about a topic that they are not interested in. They retweet simply because it is popular. The right-bottom corners show this.

3. If a user is interested in one topic, he will be active in general. This is shown by the empty left-top corner of both subfigures of Figure 6-1. That is, a user would not be interested in a specific topic only; he must be also active in general.
Figure 6-1. Activeness of Users. The scattered plots reflect the relation between users’ interest on specific topics and their activeness in general.
CHAPTER 7
DISCUSSION AND FUTURE WORK

In this paper we discuss a lot of measures for users, tweets, and cascades in Twitter, and discuss their relationships between one another. In this section, we show two interesting phenomenon of one factor, and discuss two possible explanations of it. Then, we discuss about our future work.

We would like to see the effect of the selection of topic. Thus, we find it’s influence on two measures: time needed to be retweeted, and probability to be retweeted.

7.1 More on Time Needed to be Retweeted

In section 3.3, we discuss one factor that inculuces the time needed to be retweeted made by a user: the total number of retweets of that user. We would like to see whether the selection of topic affect this measure, too. Thus, despite the authors, for all the retweets made about each topic, we show the distribution of the time needed to be retweeted in Figure 7-1.

As shown in the figure, about 80% of the tweets are retweeted by the followers within 48 hours (2 days), and there are only few tweets that being first retweeted after a week. Another interesting finding is the very similar distributions on this measure for both topics, i.e., the selection of topic does not affect the tendence of a tweet to be early retweeted.

7.2 Probability of Being Retweeted

The above analysis is on the articles that are retweeted. One may then prompt another question: how many of the articles are retweeted by others? We then answer this question as follows.

We then would like to see whether frequent retweets and different topics lead to higher chances to be retweeted. Thus, we group the users by the number of retweets about one topic, and see each user’s average probability of being retweeted, as shown in Figure 7-2. As one can see, the probability increase when a user retweets more about
Figure 7-1. The Distribution of Time Needed to be Retweeted. The two topics show very similar distributions: once a tweet is retweeted, the tendency of the time being retweeted will not be affected by the selection of topics.

#iranelection, but for #lakers we cannot find such a correlation at all. We then discuss two possible explanations as below.

The first reason could be the topics’ difference in nature. Comparing to #iranelection, #lakers is of smaller scale and more “interested-oriented”, as argued in Section 2.4. Thus, not only the overall probability of being retweeted is smaller, but also people tend to retweeted based on their interested in #lakers, rather than on the influence by others.

We then argue that this is because of their different need of expertise. After scanning on the most frequently retweeted articles, for #iranelection, such tweets are usually written by journalist or media columnists. However, #lakers is not the case: it is relatively easier to write a comment in depth, despite of the number of retweets about the same topic (which reflects the user’s expertise).

7.3 Future work

Although there have already been a lot or efforts put on the information diffusion on Online Social Networks (OSNs), the previous work either have too strong assumptions to fulfill a generous diffusion pattern of OSNs, or have a limited performance fitting
The relationship between number of retweets of users and the probability of being retweeted. The existence of the relationship depends on the topic. They are positively correlated for #iranelection, but nearly irrelevant for #lakers.

real-world data. To the best of our knowledge, a general and perfectly fitting information diffusion model is still lacking.

In this paper, we proposed a UIIS model to represent the users’ states during the diffusion process; however, we still need to take individual differences into consideration in order to qualitatively fit the model. The transitions between the four states are crucial and in need of clear identifiers: a tweet about the topic posted by whom a user follows indicates the transition from unknown to informed; a further retweet indicates the follower’s transition from informed to influential; however, the only unclear transition is that between Influential and Stopped. The observations we have presented provide us an insight of monitoring the transition between the last two states, yet need further mathematical evaluations.

Moreover, to determine whether two tweets are about the same topic, like [12], here we use a simple method that determines whether they contain common keywords. This can be improved by “outsourcing” to the OpenCalais service [3, 10], which extracts tags from input articles. We leave these for the future work.
CHAPTER 8
RELATED WORK

Throughout the paper, we have discussed the references that closely relate to our work. To have a thorough view of the whole background of the literature, we briefly review related work on the network structure of online social networks, investigation information diffusion, and modeling of the diffusion process.

In the last decade, a number of research efforts have been put on understanding the network structure, user interactions, and traffic characteristics [9, 11, 14, 18, 19, 22–26, 28]. Several work have studied the information diffusion over OSNs using empirical approaches [12, 13, 16, 21, 27], and some used statistical methods to predict the characteristics (e.g., range, scale, and speed) of the information diffusion process on OSNs [33].

Recently, some researchers have been involved in analyzing the process of information diffusion on OSNs [29–31]. Sun et al. [30] argued that the assumption that “a few nodes start long chain reactions, resulting in large-scale cascades” does not hold for social media networks. Besides, Wang et al. [31] proposed a diffusive logistic model to predict information diffusion over a time period in Digg.

The most related work to ours is by Steeg et al. [29]. They raised arguments to oppose the widely used models with real-world data. The information diffusion in OSNs is fundamentally different from other contagion processes: people do not become more likely to further spread information out with repeated exposure. Such “Saturating” phenomenon was discussed as the main reason that stops the information diffusion. By the selection of diffusion model, however, they neglected the user behaviors in more general social networks (e.g., Twitter, Facebook, and Google+). In this work, we have studied the diffusion process on a more general diffusion model, and found some crucial factors that heavily affect information propagations, yet neglected in the literature.
In this paper, we have demonstrated that multiple retweets have an unignored effect on diffusion process in Twitter. There are considerable number of cascades with range larger than 10 hops when we extract them with respect to topics rather than messages. Multiple retweets enable people keep talking and sharing information about the development of one events, thus allow us to see a dynamic news rather than a static story. In addition, time interval among multiple retweets plays an important role in predicting the tendency that one user would make more retweets or not. Furthermore, we can also evaluate the time interval between the first two levels in the cascades to forecast whether a source can generate larger cascades or not.
REFERENCES


URL http://doi.acm.org/10.1145/1401890.1401897


URL http://doi.acm.org/10.1145/1644893.1644900


URL http://doi.acm.org/10.1145/1629890.1629907


URL http://doi.acm.org/10.1145/1367497.1367585


URL http://faculty.ucr.edu/~hanneman/nettext/


URL http://doi.acm.org/10.1145/1879141.1879190


URL http://doi.acm.org/10.1145/1401890.1401945


URL http://scholar.google.de/scholar.bib?q=info:ITFPFHlz1aAJ:scholar.google.com/&output=citation&hl=de&as_sdt=0&ct=citation&cd=32

URL http://doi.acm.org/10.1145/1367497.1367591

URL http://dx.doi.org/10.1073/pnas.0708471105

URL http://doi.acm.org/10.1145/1452520.1452527

URL http://doi.acm.org/10.1145/1644893.1644901

URL http://doi.acm.org/10.1145/2068816.2068852

URL http://doi.acm.org/10.1145/1644893.1644899


URL http://ilpubs.stanford.edu:8090/984/


BIOGRAPHICAL SKETCH

Yu-Song Syu got his bachelor’s degree and master’s degree from the Department of Computer Science in National Tsing Hua University in Taiwan. He then came to the United States for his second master’s degree in the Department of Computer Information and Science Engineering in University of Florida.

Yu-song’s research interests are broadly distributed among DTV interface, vocal signal processing, and social networking. Among his technical skills, he is used to coding with Java the most.

Yu-Song’s extracurricular activities include cooking, exercise, and music. He loves Chinese and Japanese food the best. He plays softball and basketball casually and listens to all kinds of music.