An Analysis on Information Diffusion by Retweets in Twitter

by

Tomoaki Sakamoto

B.A., Economics, University of Tokyo (2004)
M.S., Applied Mathematics and Statistics,
State University of New York at Stony Brook (2012)

Submitted to the Computation for Design and Optimization
in partial fulfillment of the requirements for the degree of
Master of Science in Computation for Design and Optimization
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
February 2015

Signature redacted

Author .....
Computation for Design and Optimization
September 17, 2014

Signature redacted

Certified by .....
Roy E. Welsch
Professor of Statistics and Management Science
and Engineering Systems
Thesis Supervisor

Signature redacted

Certified by .....
Youssef M. Marzouk
Associate Professor of Aeronautics and Astronautics
CDO Reader

Signature redacted

Accepted by .....
Nicolas Hadjiconstatinou
Professor of Mechanical Engineering
Director, Computation for Design and Optimization (CDO)

An Analysis on Information Diffusion by Retweets in Twitter

by

Tomoaki Sakamoto

Submitted to the Computation for Design and Optimization
on September 17, 2014, in partial fulfillment of the
requirements for the degree of
Master of Science in Computation for Design and Optimization

Abstract

This dissertation examines retweeting activities as the information spreading function of Twitter. First, we investigated what kind of features of a tweet help to get retweets. We construct a model that describes peoples' decision making on retweets, and with related observation, we show that more retweeted tweets get retweeted more. In terms of specific features of tweets, it has been shown that the number of followers and the number of retweets are positively correlated, and hashtags attract more retweets than the tweets without hashtags. On the other hand, we also found that including hashtags and getting one or more retweets are statistically independent. Moreover, we showed including URLs or user-mentions in tweets and getting one or more retweets are statistically independent. In our results, including a picture is slightly effective to get this sense of retweetability. Second, we compare the retweeters of tweets including a picture and only text, especially focusing on distance from the original tweeters. Comparing the ratio of retweets by followers of the author of the original tweets among the initial 50 retweets, tweets with a picture have a slightly lower ratio, though there is no significant difference between the average for tweets with pictures and without pictures at the 95% significance level. We also investigate how many retweets are posted by users in followers' network connected to the original tweeter, and show that the depths of retweeters' network for tweets with picture have larger variance than that of tweets without pictures. This result implies that a tweet including picture can reach more people than a tweet without a picture potentially.

Thesis Supervisor: Roy E. Welsch
Title: Professor of Statistics and Management Science
and Engineering Systems

Acknowledgments

I am deeply thankful for all of the interactions with and help given by people at MIT. First, I would like to thank my supervisor, Professor Roy E. Welsch. I could not have completed this thesis without Prof. Welsch's guidance and support. I appreciate all of his helpful advice and patience. I am also very grateful to Professor Tauhid Zaman. He gave me the initial idea of my research. Without his advice, I couldn't start my work smoothly.

I am truly thankful to Professor Nicolas Hadjiconstantinou, the Director of CDO program, for accepting my status of thesis research in absentia. In addition, Barbara Lechner, Debra Blanchard, and Kate Nelson, the former and current CDO administrators always helped and encouraged me at every procedures. I could not have completed my work without them.

I would like to thank many of the students I've met at MIT for exciting and insightful discussions. They gave me plentiful knowledge and clues to the next step in my work. I am really grateful to have a chance to meet and talk with them.

Finally, I would like to thank my parents for their endless supports throughout my time at MIT, and long before then.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Information diffusion in the Internet has recently received broad attention due to its rapidly growing influence on the real world. Today, utilizing social network services and microblogging services effectively is the key to success in advertising and public relations. Twitter is one of the most popular microblogging services. It was launched in July 2006 and has grown rapidly. In an article published in "The Telegraph", on 23 Feb 2010, Claudine Beaumont wrote "In 2007, around 5,000 tweets were sent per day, with that increasing to 300,000 messages per day in 2008. The number of tweets sent last year grew by 1,400 per cent, to around 35 million per day, and that figure now stands at 50 million tweets sent per day." According to Lee et al. (2011), "as of June 2011, about 200 million tweets are being generated every day." These resources commonly state the rapid growth of Twitter. In terms of the number of users, 500 million accounts were registered in 2012 including many public figures and celebrities including the U.S. President and movie stars. Many global public presses also have their Twitter accounts. Twitter is also utilized for business by small and large enterprises. According to Alexa traffic ranks, Twitter is ranked as the 9th most-viewed website. Its large amount of users and their tweets enable us to observe people's behaviors macroscopically, and various research on Twitter has been made so far.

Twitter is not only a popular website, but also a type of media having interactions with the real world. Users can post a message, URLs, photos, and movies which reflect real-world incidents, and the information in the post can spread out, and let the viewers know something which they didn't notice by themselves. Additionally, sometimes it can stimulate viewers to react. In this sense, information diffusion in Twitter can affect people's decision-making.

Now, Twitter is a powerful tool for spreading information. In fact, Twitter is used for many types of announcements, advertisements, and breaking news. A public relations officer of a local government might want to make an announcement for some community event in that area, or some regional useful news. A museum curator can be interested in letting people know about a new exhibition. Some researchers notify their latest presentation documents through posting its URL on Twitter. Actors might try to spread their event notification as well.

Under a disaster, local and real-time information can be a key to be safe or to survive. There is possibility that confusing information also appears during crises from both the affected area and outside. Sakaki et al(2010) used Twitter as a sensor of earthquake, and developed a system which notifies people promptly of an earthquake.

Twitter changed news media or journalism in some sense. Today, traditional news media including CNN and BBC have their own Twitter accounts, and broadcast their news through Twitter. In addition to this, Twitter has become a catalyst letting people know breaking news. As famous instances which spread out via Twitter, Kwak et al.(2010) mentioned a case of an American student jailed in Egypt and the US Airways plane crash on the Hudson river. Hu et al.(2012) investigated how the news broke and spread on Twitter. They noticed that the news of Osama Bin Laden's death spread rapidly with a tweet created by U.S. President Barack Obama on May 1st in 2011. Moreover, Petrovic et al.(2013) pointed out that Twitter has an advantage of local information in comparison to newswires.

There are some reasons why information spreading in Twitter is important.

First, information on Twitter is mostly aggregated in one viewer that is called as "Timeline," though websites of traditional media have a lot of pages. In Twitter, users can see any information which is posted by other Twitter accounts that the user follows.

Second, one of the characteristics of Twitter is its real-time nature. This point is related with breaking news on Twitter.

Finally, Twitter is a community platform which is supported by users' connection. This means spread information can potentially become prevalent among those communities as a kind of common information.

For all these reasons, Twitter has become a strong tool for spreading information. In this work we focus on how to accelerate this function of Twitter.

Especially, retweet is a function to forward a tweet which is posted by another user. Tweets are usually seen by the followers of the user who posted those tweets. If those tweets are retweeted by another user, they are also seen by the followers of the user who did the retweets in addition to the followers of the original tweeter. Thus, retweets enable users to spread the original tweets. Obviously, a tweet which is retweeted hundreds of times will be exposed to hundreds of thousands of people. If a tweet is retweeted thousands of times, some other web-based media may cite it, and its information will spread more. In this work, we mainly focus on retweeting activity.

1.2 Literature Review

In fact, considerable attention has been paid to the research of retweeting behaviors on Twitter. Kwak et al. (2010) studied audience size of retweets, characteristics of retweet trees, and time-series observation of retweets. According to their observation in 2009, any retweeted tweet is to reach around 1,000 users on average. In terms of the length of retweet tree, most of them have height one, and , and no trees go beyond

11 distance, and the distribution of the users in a retweet tree follows a power-law. They also reported about 50% of retweets are posted in one hour after the original tweet was created, and 75% are posted in one day, while 10% are posted one month later. Note that those "retweets" are not exactly the same with built-in feature of retweets in the current user-interface of Twitter. These observations above are still useful to consider the characteristics of retweeting behavior, however, they might have changed today.

Boyd et al. (2010) focused on conversational aspect of retweeting, and investigated the practices of retweets. They qualitatively classified and listed the reasons why people retweet; to amplify or spread tweets to new audiences, to publicly agree with someone, to validate others' thoughts, and so on. They also examined what people retweet; for showing that they are the audience of the original tweet, and for encouraging social actions including raising funds and demonstration of collective group identity-making, and requesting help. Additionally, they reported 52% of retweets contain a URL, and 18% of retweets contain a hashtag. While their analysis above is insightful to consider retweet as a built-in function of Twitter, their main attention was paid for manually created retweets. For example, in built-in retweet syntax of a retweet is identified, and users are not able to retweet tweets posted by themselves.

Suh et al. (2010) investigated features that affect retweetability to understand why certain tweets spread more widely than others. They examined both manual retweets and built-in retweets. They found that URLs and hashtags have strong relationships with retweetability, and the number of followers and followees, and the age of the account also affects the retweetsbility, while the number of past tweets does not have influence on retweetability of a user's tweet.

Yang et al. (2010) investigated retweeting behaviors by focusing on features of users and contents of tweets. Based on their observations, most users retweet at a low frequency and only a few users are retweet-aholic. In their data, the average number of retweets of a user within seven days is 197, and only 3.13% of the retweets are

posted by users who retweet more than 1,000 times. Many users post far more tweets than retweets, but retweet-aholics post many fewer tweets than retweets. In terms of contents, they showed that users tend to retweet the messages that contain what they are interested in.

Nagarajan et al. (2010) discussed retweet behavior by connecting them to real-world events. By observing samples of popular tweets, they found that tweets calling for social action, crowdsourcing, or collective group identity-making tend to generate a sparse retweeters' network, i.e., the retweets are not necessarily connected to the original tweets. On the other hand, tweets sharing information generate a dense retweeters' network. In other words, the retweets are connected to the original author. They explicitly recognized this type of tweet as sharing information by containing URLs, images, and videos.

Zaman et al. (2010) utilized user IDs of the author of the original tweets and retweeters, the number of followers, and the words contained in tweets to predict future retweets.

Petrovic et al. (2011) conducted a human experiment on the task of predicting whether a tweet will be retweeted or not. They found that social features of the author of a tweet including the number of followers, friends, listed are very important information to predict whether the tweet will get retweeted, while tweet features including hashtags, user mentions, and URLs are useful information.

Luo et al. (2013) analyzed who retweets other users. They found that followers who retweeted or mentioned some other users' tweets frequently before, and have common interests are more likely to be retweeters.

In the observation by Myers et al (2014), there is partial relationship between following burst and retweets.

These studies above suggest to us insightful results. In this work, we try to review some of the results above, and try to provide some different viewpoints. The

knowledge of statistics in this work is covered in Rice (2007) unless we specify other references.

1.3 Overview

The remainder of this thesis is as follows.

In chapter 2, we formulate a model that describes individual users' decision making on tweets and retweets based on networks with preferential linking presented by Dorogovtsev et al. (2000). We discuss two versions of preferential behavior of retweeting; choosing a tweet to retweet randomly, and choosing a tweet to retweet depending on retweet count of the tweet. Our observation support the later assumption, and it implies that more retweeted tweets tend to get retweets more.

In chapter 3, we investigate the nature of retweetability by overviewing data. We start with explaining the way we construct our dataset, and provide some statistical reviews. As already pointed out by previous studies, the number of followers and hashtags tend to be related with the number of retweets. However, our experiment suggests that including hashtags, URLs, and user-mentions do not affect whether the tweet will be retweeted or not, although if those tweets have a seed of retweetability, they might get more retweets than the others. In our results, including a picture might effect retweetability slightly.

In chapter 4 we compare the tweets including pictures and those with plain texts focusing on the following relationship of retweeters. According to our observation, the variance of maximum distance of each tweet with pictures is larger than that of text-only tweet. This result implies that tweets with a picture potentially reach more distant people than text-only tweets, although tweets with a picture can stop in smaller networks.

In chapter 5, we explain our conclusion of this work. Our results imply that those who are interested in getting retweets should focus on getting initial retweets in addition

to utilizing specific features including hashtags, URLs, and pictures. By including pictures in a tweet, it might go further than tweets without pictures, while in some cases it stops in a narrower followers' network.

Chapter 2

A Decision Making Model for Retweeting

In this chapter, we build a retweet-network model which describes individual user's decision making on tweets and retweets. We consider two possible patterns of choosing which tweet to retweet; random choosing or weighted choosing. Our observation supports the weighted choosing pattern, and this result implies that more retweeted tweets tend to get retweeted more.

2.1 Definitions

In this model, we regard the relationship between tweets and retweets as a network and call this structure a retweet-network. First of all, the basic components of network structure need to be defined.

Definition 1 (Tweet) *A tweet is a node which is not a starting point of a directed edge.*

No tweet has an edge which starts from it. Each tweet has its unique ID which enables us to distinguish different tweets.

Definition 2 (Retweet) *A retweet is a node which is a starting point of a directed edge.*

Definition 3 (Retweet Edge) *A retweet edge is a directed edge which starts from a retweet.*

Figure 2-1 shows the relationship between an original tweet and a retweet to it.

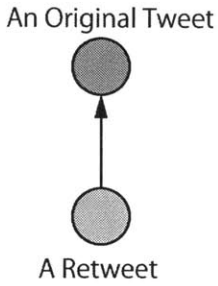


Figure 2-1: Relationship between a tweet and a retweet

Definition 4 (Retweet Network) *A retweet network is a network whose nodes are tweets and retweets, and whose edges are retweet edges.*

Definition 5 (Degree of a tweet) *Degree of a tweet is the number of retweet edges whose end point is the tweet.*

For example, in the case of figure 2-2, there is one original tweet and three retweets, and the degree of the original tweet is three.

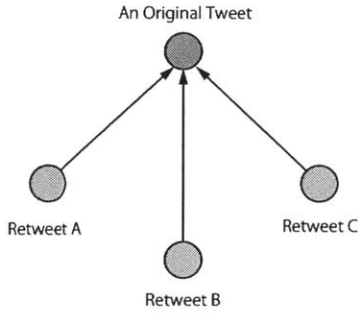


Figure 2-2: Simple case of a retweet network

2.2 Model Analysis

2.2.1 Description of a Retweet Network

A retweet network has two specific characteristics. First, a retweet network is not generally connected. Thus, in many cases, we cannot calculate the cluster coefficients and average length between nodes of a whole retweet network.

Second, a retweet cannot be retweeted. Here we didn't define the degree of retweet because even if we defined it, their degrees are always one, that is, the edge whose starting point is the retweet.

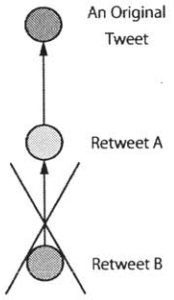


Figure 2-3: An invalid case of a retweet

2.2.2 Growth Dynamics of Retweet Network

In this analysis, we focus on the degree distribution here. The degrees of tweets are the indicators of the magnitude of the tweet influence.

We denote the tweet whose ID is i by v_i, and denote its degree by k_i. At each time, a user posts a tweet or retweet. Let's assume that v_i enters this network at time t_i. k_i grows as time passes, so we can denote $k_i(t)$, or at time t, the degree of v_i is k_i. In this analysis, we think of a unit time as an interval between a post (a tweet or a retweet) and the next post.

Here we assume that at the initial time $t_0 = 0$ there is one tweet as an initial condition, and a user chooses to retweet at probability q, and choose to tweet at probability $1-q$. If a user chooses to post an original tweet, one additional tweet which is a candidate of a target of a retweet is created, and if a user chooses to retweet, the degree of one of an existing original tweets increases by one. Potentially, different users have different q. However, to simplify this model, we assume common q here. Except for this point, this model describes directly the decision-making of tweet and retweet activity.

Now, k_i is the number of retweets of tweet ID i, and the distribution of k_i shows the tendency of retweets and their background. In order to know an approximate tendency, we make two different assumptions and test which one is more plausible. One assumption is that when a retweet is posted, the tweet which is retweeted is randomly chosen. The other assumption is that when a retweet is posted, more retweeted tweets tend to be chosen.

If the former one is true, every tweet equally get retweets. On the other hand, if the later one is true, more retweeted tweets tend to get more retweets, and then getting retweets right after it is posted is important for information spreading.

Here, we denote the probability of which v_i is retweeted at time t by $r(k_i)$, and investigate the distribution of k_i by defining $r(k_i)$ which correspond to these two assumptions. In the calculation of the degree distribution, we use the continuous approximation.

First, we consider the case in which each original tweet is randomly chosen. The number of tweets at time $t = 0$ is 1, and as t increases by 1, the number of tweets increases at the probability $1 - q$. Thus, the number of nodes is $N - 1$ when

$$(1 - q)t \approx N - 2 + 1 = N - 1.$$

under the continuous approximation. At this time t, the probability of which v_i is retweeted among $N-1$ tweets at the next retweet is approximately

$$r(k_i) \approx \frac{q}{N-1} \approx \frac{q}{(1-q)t}.$$

This approximated probability also stands for the expectation of a retweet which the user gets at the next time step. In other words, $r(k_i)$ is interpreted as the approximation of difference between a certain time step and the next time step. Hence, when assuming t and k_i are continuous, this approximation is expressed by the following differential equation.

$$\frac{dk_i}{dt} \approx r(k_i) \approx \frac{q}{(1-q)t}.$$

Here, as an initial condition, $k_i(t_i) = 1$, this equation is solved as following.

$$k_i = \frac{q}{1-q} \log \frac{t}{t_i} + 1.$$

By this k_i, the distribution function of k_i is calculated as following.

$$\begin{aligned} P(k_i(t) < k) &\approx P\left(\frac{q}{1-q} \log \frac{t}{t_i} + 1 < k\right) \\ &= P\left(\frac{q}{1-q} \log \frac{t}{t_i} < k-1\right) \\ &= P\left(\log \frac{t}{t_i} < \frac{1-q}{q}(k-1)\right) \\ &= P\left(\frac{t}{t_i} < \exp\left[\frac{1-q}{q}(k-1)\right]\right) \\ &= P\left(t_i > t \exp\left[\frac{1-q}{q}(1-k)\right]\right). \end{aligned}$$

Considering the number of tweets, it is the one at initial time $t = 0$, and added tweets at probability $1-q$ at each time unit, i.e., $(1-q)t+1$. Among these tweets, the ones which satisfy the condition

$$t_i > t \exp\left[\frac{1-q}{q}(1-k)\right]$$

are added after the time $t\exp\left[\frac{1-q}{q}(1-k)\right]$ has passed. Thus, the number of tweets which satisfy the condition above is

$$(1-q)\Big(t - t\exp\Big[\frac{1-q}{q}(1-k)\Big]\Big).$$

Hence,

$$P(k_i(t) < k) \approx \frac{(1-q)t}{(1-q)t+1}\Big(1 - \exp\Big[\frac{1-q}{q}(1-k)\Big]\Big).$$

By differentiating this equation by k, the density function of the degree is calculated as the following.

$$\begin{aligned} p(k) &= \frac{\partial P(k_i(t) < k)}{\partial k} \\ &\approx \frac{(1-q)t}{(1-q)t+1}\frac{\partial}{\partial k}\Big(1 - \exp\Big[\frac{1-q}{q}(1-k)\Big]\Big) \\ &= -\frac{(1-q)t}{(1-q)t+1}\frac{\partial}{\partial k}\exp\Big[\frac{1-q}{q}(1-k)\Big] \\ &= \frac{(1-q)t}{(1-q)t+1}\frac{1-q}{q}\exp\Big[\frac{1-q}{q}\Big]\exp\Big[\frac{1-q}{q}(-k)\Big] \\ &= \theta\exp(-\lambda k), \end{aligned}$$

where

$$\theta = \frac{(1-q)t}{(1-q)t+1}\frac{1-q}{q}\exp\Big[\frac{1-q}{q}\Big]$$

and

$$\lambda = \frac{1-q}{q}.$$

Therefore, by normalizing the coefficients for make $p(k)$ a probability density function, k follows exponential distribution.

On the other hand, the $r(k_i)$ which corresponds to the later assumption is considered as following. In this case, we apply a model which is modified version of networks

with preferential linking presented by Dorogovtsev, et al. (2000).

$$r(k_i) = \frac{q(k_i + k_0)}{\sum_j (k_j + k_0)},$$

where k_0 is a constant value. Under this assumption, the larger k_i becomes, the larger $r(k_i)$ becomes, i.e., more retweeted tweets tend to be retweeted even more. This assumption is intuitively natural because if a tweet is retweeted more times, more users should tend to notice the tweet. Similarly to the above, the number of tweets is $N-1$ when

$$(1-q)t \approx N-1,$$

and at this time,

$$t \approx \frac{N-1}{1-q}.$$

Then,

$$\begin{aligned} \sum_{j=1}^{N-1} k_j &\approx \left(\text{the number of all retweets at time } t \approx \frac{N-1}{1-q}\right) \\ &= qt. \end{aligned}$$

Hence,

$$\begin{aligned} r(k_i) &= \frac{q(k_i + k_0)}{\sum_{j=1}^{N-1}(k_j + k_0)} \\ &= \frac{q(k_i + k_0)}{\sum_{j=1}^{N-1} k_j + (N-1)k_0} \\ &= \frac{q(k_i + k_0)}{qt + (1-q)k_0 t} \\ &= \frac{q(k_i + k_0)}{q + (1-q)k_0}\frac{1}{t}. \end{aligned}$$

Similarly above,

$$\frac{dk_i(t)}{dt} \approx r(k_i) \approx \frac{q(k_i + k_0)}{q + (1-q)k_0}\frac{1}{t},$$

then under the initial condition $k_i(t_i) = 1$,

$$k_i(t) = (1+k_0)\Big(\frac{t}{t_i}\Big)^{\frac{q}{q+(1-q)k_0}} - k_0.$$

Thus,

$$\begin{aligned}
P(k_i(t) < k) &= P\Big[(1+k_0)\Big(\frac{t}{t_i}\Big)^{\frac{q}{q+(1-q)k_0}} - k_0 < k\Big] \\
&= P\Big[(1+k_0)\Big(\frac{t}{t_i}\Big)^{\frac{q}{q+(1-q)k_0}} < k+k_0\Big] \\
&= P\Big[\Big(\frac{t}{t_i}\Big)^{\frac{q}{q+(1-q)k_0}} < \frac{k+k_0}{1+k_0}\Big] \\
&= P\Big[\frac{t}{t_i} < \Big(\frac{k+k_0}{1+k_0}\Big)^{\frac{q+(1-q)k_0}{q}}\Big] \\
&= P\Big[t_i > t\Big(\frac{1+k_0}{k+k_0}\Big)^{1+\frac{1-q}{q}k_0}\Big].
\end{aligned}$$

At time t, the number of all tweets is $(1-q)t+1$. Among these tweets, the ones which satisfy

$$t_i > t\Big(\frac{1+k_0}{k+k_0}\Big)^{1+\frac{1-q}{q}k_0}$$

are the ones which added after time $t\Big(\frac{1+k_0}{k+k_0}\Big)^{1+\frac{1-q}{q}k_0}$ has passed. Thus, the number of them is

$$(1-q)\Big(t - t\Big(\frac{1+k_0}{k+k_0}\Big)^{1+\frac{1-q}{q}k_0}\Big).$$

Hence,

$$P(k_i(t) < k) \approx \frac{(1-q)t}{(1-q)t+1}\Big(1 - \Big(\frac{1+k_0}{k+k_0}\Big)^{1+\frac{1-q}{q}k_0}\Big).$$

Thus, the density function of k is

$$\begin{aligned}
p(k) &= \frac{\partial P(k_i(t) < k)}{\partial k} \\
&= \frac{\partial}{\partial k}\frac{(1-q)t}{(1-q)t+1}\Big[1 - \Big(\frac{1+k_0}{k+k_0}\Big)^{1+\frac{1-q}{q}k_0}\Big] \\
&= -\frac{(1-q)t}{(1-q)t+1}\frac{\partial}{\partial k}\Big(\frac{1+k_0}{k+k_0}\Big)^{1+\frac{1-q}{q}k_0}
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{(1-q)t}{(1-q)t+1}(1+k_0)^{1+\frac{1-q}{q}k_0}\frac{\partial}{\partial k}(k+k_0)^{-1-\frac{1-q}{q}k_0} \\
&= \frac{(1-q)t}{(1-q)t+1}(1+\frac{1-q}{q}k_0)(1+k_0)^{1+\frac{1-q}{q}k_0}(k+k_0)^{-2-\frac{1-q}{q}k_0} \\
&\propto k^{-2-\frac{1-q}{q}k_0}
\end{aligned}
$$

In this case, the degree distribution of k has power law, and its parameter is $2+\frac{1-q}{q}k_0$ in this setting.

From these calculations, if the tweets to retweet are randomly chosen, the distribution of k is nearly an exponential distribution, and if more retweeted tweets tend to get more retweets, the distribution of k is close to a distribution with power-law.

2.3 Observation of retweet network

For comparing the model above and our data, we counted how many times each tweet was retweeted in each trending topic, and observed the distribution of those degrees of tweets for each trend.

Each trending topic has its degree distribution. As examples, we show the degree distribution of 9 trending topics in Figure 2.4.

For each trending topic, the upper plot is the simple histogram of degrees, and the lower plot is the log-log plot of degrees and fitted line of a power-law.

These plots indicate that most of retweeted tweets are the retweeted just a few times, and only a few tweets are retweeted for hundreds of times. According to the calculation with a built-in function of the "powerlaw" module in python (Alstott et al.(2014)) with log-likelihood ratio, 2,496 in 2,508 examples are nearer the power-law distribution than exponential distribution.

Next, we notice the fitted parameters of the power-law. The average value of parameters is 3.822, the median is 2.664, standard deviation is 11.1018, the minimum value

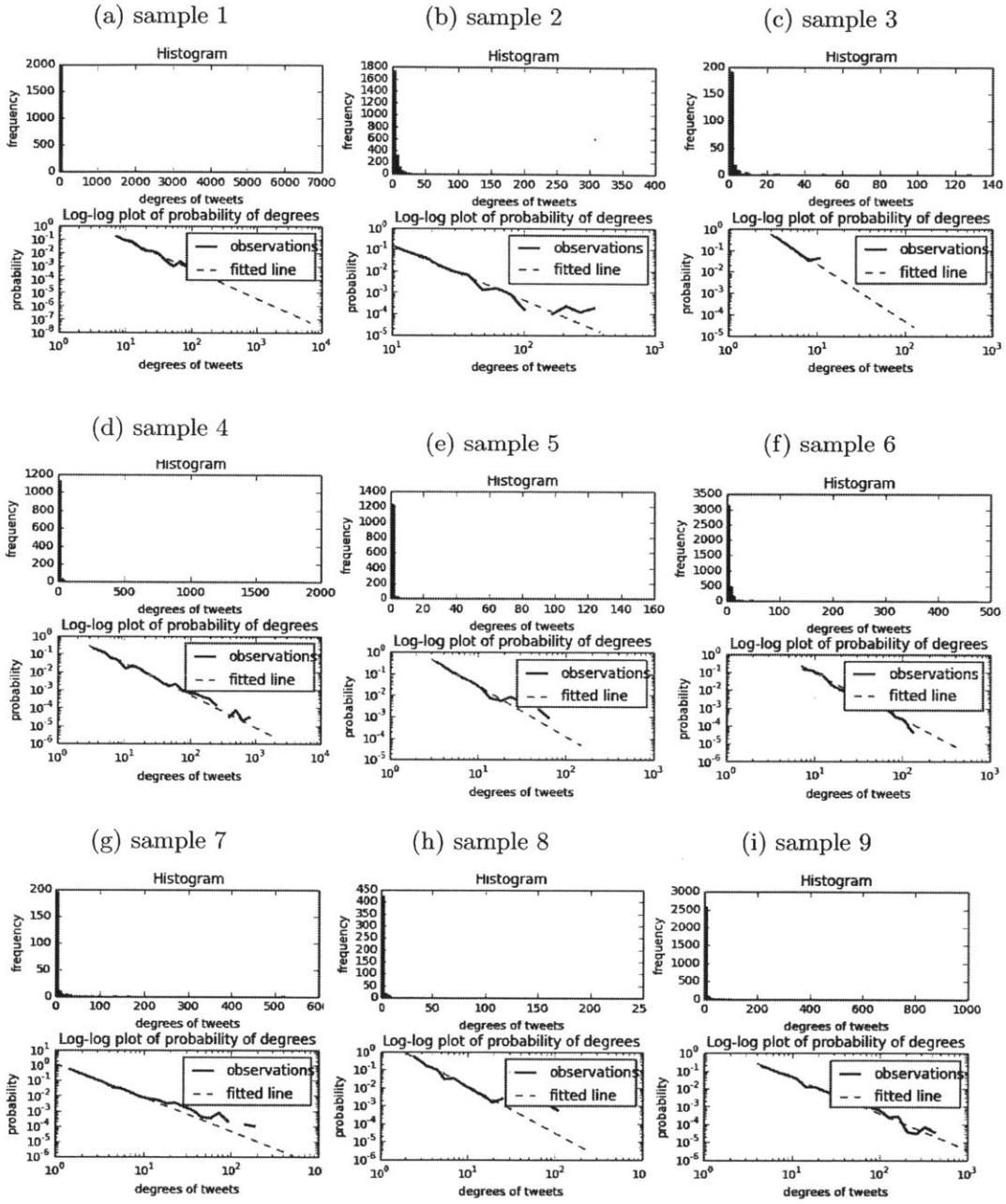


Figure 2-4: Degree distribution of 9 sample trending topics

is 1.264, and the maximum is 531.722.

In fact, the number of the trending topics whose parameters are greater than 10 is 87 in $2,508$ (3.47%). The number of topics whose parameters are in the range of $(2,4)$ is $1,750$ in $2,508$ (69.78%). These values do not contradict the calculation in the former section.

These results above support the statement that more retweeted tweets get retweeted more.

Chapter 3

Observations on Retweetability

In this chapter, we explain our data resources and observation of specific features of tweets mainly related to retweetability. As related work has already pointed out, the number of followers and the number of retweets are positively correlated, and hashtags attract more retweets than the tweets without hashtags. On the other hand, we found that including hashtags and getting 1 or more retweets are statistically independent. Moreover, we showed including URLs or user-mentions in tweets and getting 1 or more retweets are statistically independent. In our results, including a picture is slightly effective for increasing retweetability.

3.1 Data Overview

3.1.1 Data Resources

In order to build our dataset, we utilized two types of publicly open APIs provided by Twitter officially: Streaming API and REST API. [1] The available data with each API is as follows.

[1] More detailed information is available in Twitter API documentation.

Streaming API

Streaming API with "GET statuses/sample" returns a small random sample of all public statuses. This API provides three types of data: Tweet, Retweet, and Delete. Among them, we use tweets and retweets data, Both of Streaming API and Search API in REST API give us them with the same format.

Table 3.1 and 3.2 show the contents which are included in Tweet and Retweet data. The tweet (retweet) data includes the timestamp which indicates when the tweet or retweet is created, ID of the tweet or retweet, Hashtags which the message contains, and user information who posted the tweet or retweet. It also includes URLs and pictures which are linked with the post, geographic data, and language data, "favorite" information, reply information. If the post is a retweet, the data provides the information of the original tweet. A retweet also has its ID as a tweet. As of the format of data, the only difference between tweets and retweets is whether it includes the retweeted (original) message or not. In this work, we mainly use tweet ID, time stamp at which the tweet created, hashtags, URL, user-mention, retweeted status data and user information.

Table 3.1: Tweet Data provided via Streaming or Search API

Data Category	Contents
Created at	timestamp of its creation
Mention	User who mentioned the tweet
Hashtags	Hashtags which are included in the tweet
URL	URLs which are included in the tweet
Favorite_count	How many users favored the tweet
Geo	Area where the tweet is created
ID	ID of the tweet
Lang	Language of the text
Text	Message content
User	The data of user who posted the tweet

Table 3.2: Retweet Data provided via Streaming or Search API

Data Category	Contents
Created at	timestamp of its creation
Mention	User who mentioned the original tweet
Hashtags	Hashtags which are included in the original tweet
URL	URLs which are included in the original tweet
Geo	Area where the tweet is created
ID	ID of the retweet
Retweet status	Tweet data of the original tweet
Lang	Language of the original text
Text	Message content of the original tweet with "RT @-userID" at the head
User	The data of user who posted the retweet

REST API

REST APIs of Twitter provide various types of data. The data we used in the later half of chapter 2 was built by combining the Search API and the endpoint of "GET trends / place" in REST APIs.

The Search API returns a collection of relevant Tweets matching a specified query. This API is one of the REST APIs, and also utilized in related work. For example, Nagarajan, et al.(2010) and Asur, et al.(2011) used it. Before 2011, the version of API was 1.0. Thus, they wrote that Twitter search API gives only 1,500 tweets. This constraint has been changed and now we can access more than 1,500 tweets and retweets, though Search API doesn't necessary guarantee to collect all data of tweets which includes some specific words perfectly.

Twitter also provides 10 trending words every 5 to 10 minutes. Twitter API provides trending data for every location based on Yahoo! Where On Earth ID (WOEID), and we can choose the location. Global trend data is available by using 1 as the WOEID parameter, and we used it here. This API provides simply the names of 10 trending topics and the time stamps when the trends are created and collected. Unless Twitter service is down, we can get all trending topics technically. Table 3.3 shows the data contents of trends which we can get via Twitter REST API for trend with version

1.1.

Table 3.3: Trend Data provided via REST API for trend

Data Category	Contents
Created at	Timestamp at which the set of the trends is created
As of	Timestamp at which the set of the trends is collected
Trends	Names of 10 trending topics

REST APIs provide many other resources to collect data. The endpoint of "GET statuses/show/:id" provides the detailed information of a single tweet specified by the Tweet ID. By this API, we obtained the retweet count of a tweet after a certain period of time. The endpoint of "GET statuses/retweets/:id" provides the 100 most recent retweets of the tweet specified by the Tweet ID. If we access this endpoint before the tweet gets more than 100 retweets, we can obtain the information of initial retweets. By utilizing this API, we collected the information that we used in chapter 3 and chapter 4.

3.1.2 Data Description

Before examining the data, it is useful to overview how the users uses Twitter at present (in the first half of 2014). [2]

We collected the data of tweets and retweets during 7 days from May 1st(Thu) to May 7th(Wed). Table 3.4 shows the figures of the volume of tweets and retweets provided via Twitter Streaming API with version 1.1, and figure 3-1 to 3-4 show time series plots of them. We counted the volume of tweets and retweets that are posted in every 10-minute interval. Each plot consists of 2 areas. The bottom area stands for the volume of retweets, and the top area stands for the volume of tweets.

According to this data, one third of total posts are retweets, and around 15% are the tweets including hashtags. Every day around 7 a.m (UTC) the volume of tweets and

[2]Twitter sometimes changes its official user interface, and it might change users' actions in Twitter, though we don't focus on that effect in this work.

retweets shrinks, and growing up until around 3 p.m. This cycle looks stable with a few exceptions. In terms of the volumes of overall tweets and retweets, it doesn't seem there is clear "day of the week effect."

Table 3.4: Tweet and Retweet Volume (proportion) on each day

Date	Retweet	Tweet	Total
May 1 (Thu)	1,442,499	2,971,892	4,414,391
	(32.68%)	(67.32%)	(100.0%)
May 2 (Fri)	1,373,268	2,907,299	4,280,567
	(32.08%)	(67.92%)	(100.0%)
May 3 (Sat)	1,351,808	2,879,884	4,231,692
	(31.94%)	(68.06%)	(100.0%)
May 4 (Sun)	1,447,762	2,986,705	4,434,467
	(32.65%)	(67.35%)	(100.0%)
May 5 (Mon)	1,468,118	3,025,610	4,493,728
	(32.67%)	(67.33%)	(100.0%)
May 6 (Tue)	1,441,075	3,036,456	4,477,531
	(32.18%)	(67.82%)	(100.0%)
May 7 (Wed)	1,411,757	2,829,342	4,241,099
	(33.29%)	(66.71%)	(100.0%)

3.2 Retweet-related observation via Streaming API

Streaming data provides us more information about retweets. Here, we describe the relationship with the frequency, the number of hashtags, the number of unique users, the number of followers, and embedded other media including photos and URLs. Besides, for the tweets which have the largest number of retweets, we see the growth of retweets.

The relationship between retweet count and its frequency

Retweet count in Table 3.4 stands for the number of "retweeting" posts observed in Stream Data. On the other hand, Table 3.5 is for the "retweeted" side. In Stream Data, tweets which are retweeted for many times can appear repeatedly when the tweets which retweet the posts are caught. In table 3.5, those (retweeted) tweets are

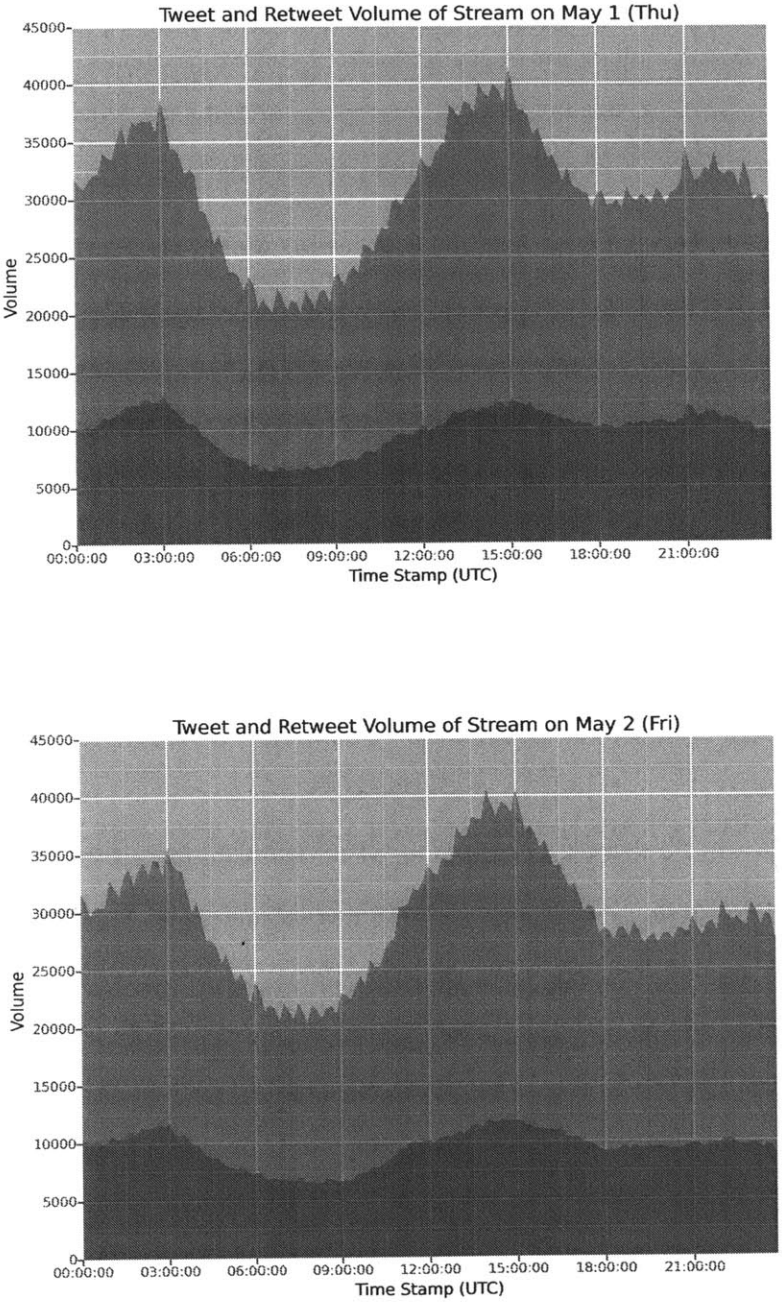


Figure 3-1: Time series plot of the volume of Tweets and Retweets (1)

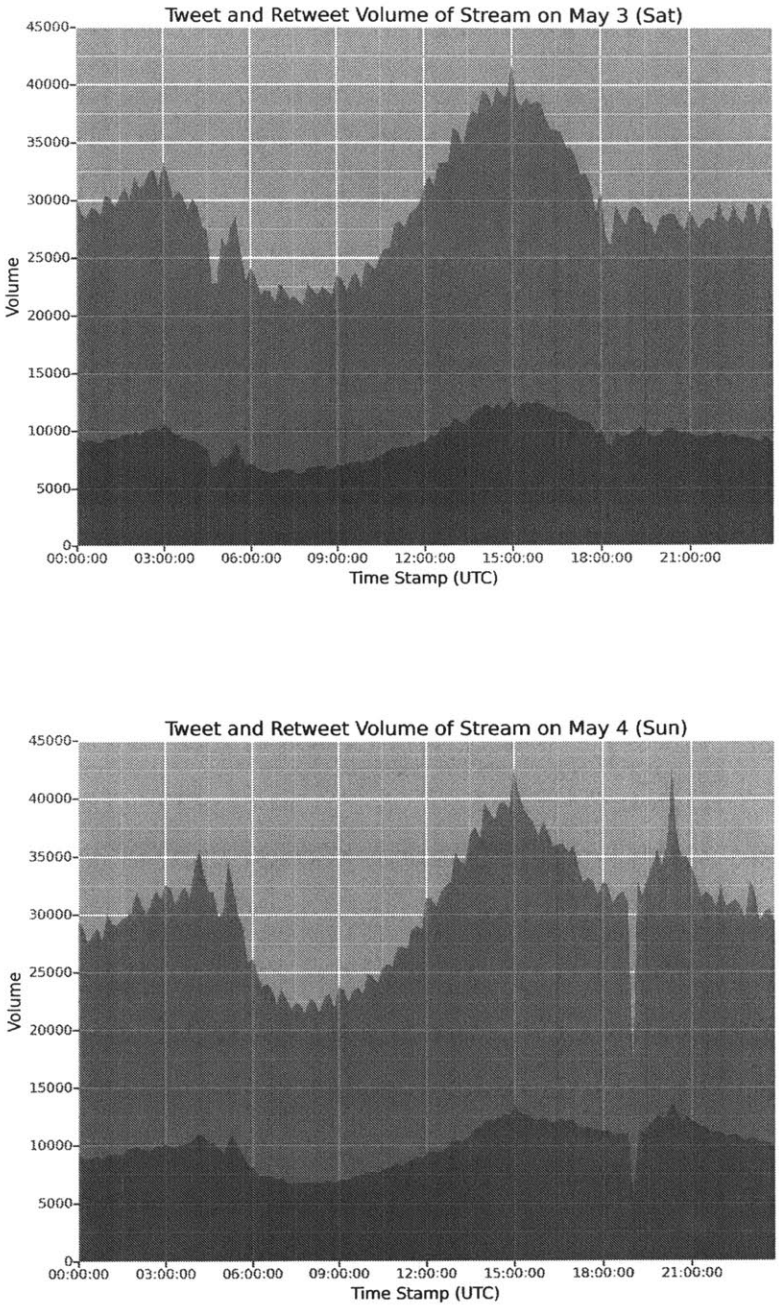


Figure 3-2: Time series plot of the volume of Tweets and Retweets (2)

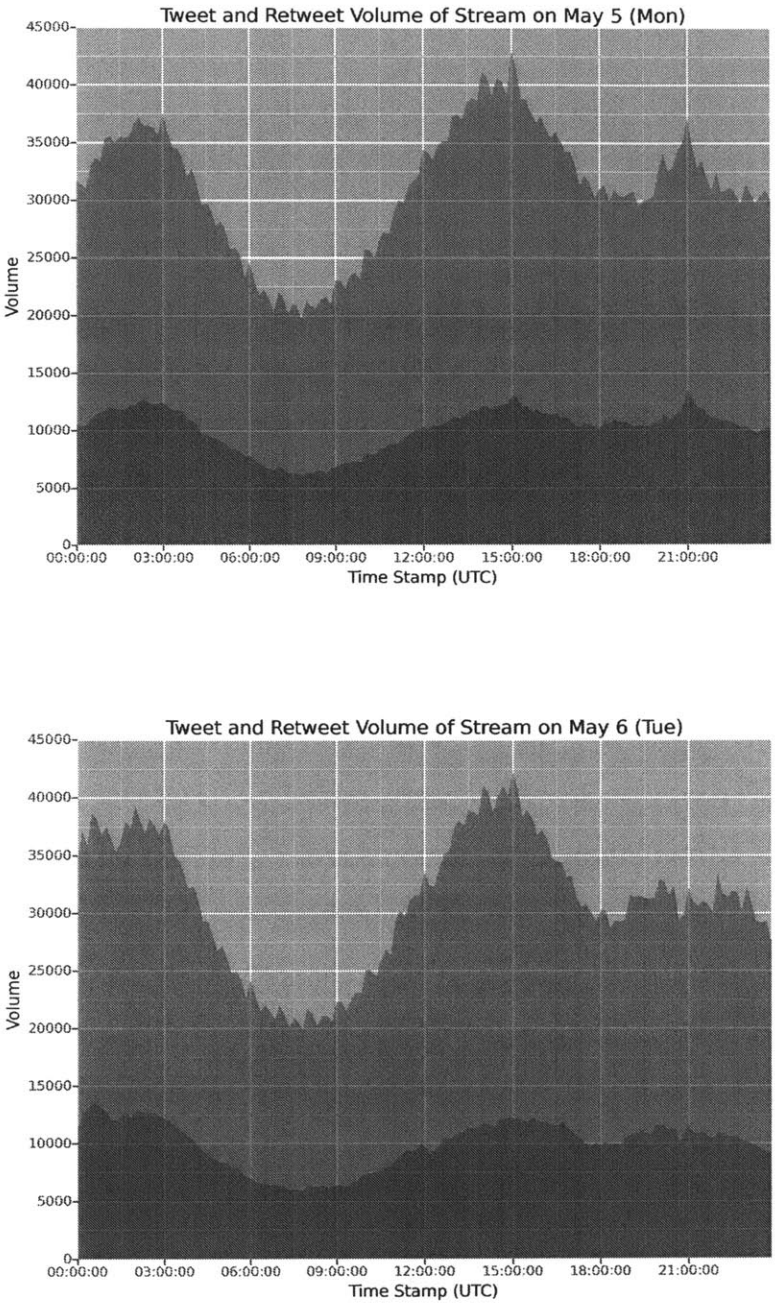


Figure 3-3: Time series plot of the volume of Tweets and Retweets (3)

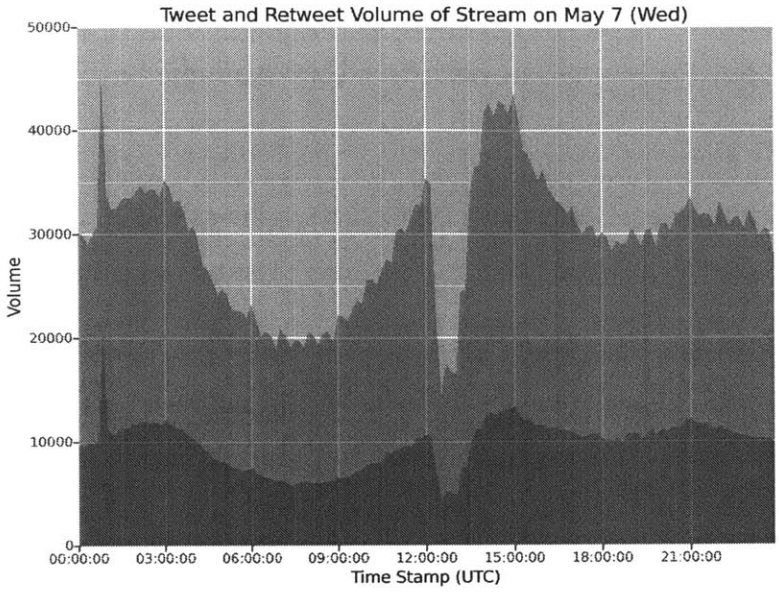


Figure 3-4: Time series plot of the volume of Tweets and Retweets (4)

counted only one time with the most retweet count among the tweets with the same ID. Table 3.5 shows that how many tweets are created and how many times they are retweeted. Here we focused on retweeting activities in one day. For example, if a post which was created on May 1 and was retweeted 5 times during May 1st, this post is added to the frequency in the category of "$1 \sim 10$ retweet count" on "May 1 (Thu)" in Table 3.5.

This table also shows that how many unique users are included in each class of retweeted. On May 1, the $533,659$ tweets which ended $1 \sim 10$ times retweeted are posted by $434,841$ unique users.

Looking at this table, the retweeting activities which are observed in Streaming Data are stable in this week.

Table 3.5: Retweet count and Frequency

Date	Retweet count range	Frequency	# of users
May 1 (Thu)	$1 \sim 10$	533,659	434,841
	$11 \sim 100$	166,970	78,328
	$101 \sim 1,000$	60,115	20,466
	$1,001 \sim 10,000$	5,444	2,050
	$10,001 \sim 100,000$	134	67
	$100,001 \sim$	1	1
May 2 (Fri)	$1 \sim 10$	507,453	414,596
	$11 \sim 100$	159,921	74,654
	$101 \sim 1,000$	58,492	19,633
	$1,001 \sim 10,000$	5,141	1,858
	$10,001 \sim 100,000$	119	65
	$100,001 \sim$	2	2
May 3 (Sat)	$1 \sim 10$	487,744	393,806
	$11 \sim 100$	157,101	73,582
	$101 \sim 1,000$	56,198	18,880
	$1,001 \sim 10,000$	4,852	1,844
	$10,001 \sim 100,000$	105	53
	$100,001 \sim$	0	0
May 4 (Sun)	$1 \sim 10$	529,463	424,583
	$11 \sim 100$	169,450	79,943
	$101 \sim 1,000$	61,336	20,561
	$1,001 \sim 10,000$	5,377	1989
	$10,001 \sim 100,000$	113	61
	$100,001 \sim$	2	2
May 5 (Mon)	$1 \sim 10$	541,062	440,614
	$11 \sim 100$	165,379	81,574
	$101 \sim 1,000$	62,571	21,473
	$1,001 \sim 10,000$	5,857	2,219
	$10,001 \sim 100,000$	141	62
	$100,001 \sim$	0	0
May 6 (Tue)	$1 \sim 10$	530,674	435,791
	$11 \sim 100$	166,282	81,119
	$101 \sim 1,000$	63,504	22,004
	$1,001 \sim 10,000$	5,301	2,150
	$10,001 \sim 100,000$	106	69
	$100,001 \sim$	0	0
May 7 (Wed)	$1 \sim 10$	513,368	415,275
	$11 \sim 100$	171,221	81,379
	$101 \sim 1,000$	62,722	22,420
	$1,001 \sim 10,000$	5,169	2,093
	$10,001 \sim 100,000$	121	68
	$100,001 \sim$	1	1

The growth of retweets for the most retweeted posts

Figure 3-5 to 3-8 are the histograms that show how many tweets reached how much % of the final retweet count of the day in one hour. For example, Figure 3.5 is for the tweets that reached more than 2,000 retweet count on May 1st. In this figure, around 550 tweets reached 100% in one hour, though other ones are moderately retweeted. We did not specify these rapidly retweeted and stopped tweets. In fact, the number of tweets that reached 100% and contains enough information is 530, and the other side which has no deficit information is $1,210$.

Looking at the Figure 3-6 to Figure 3-9, the shapes of the histograms are similar to Figure 3-5. This observation implies that the growth rate of retweets does not change depending on the level of final retweet counts as much. Even though the right most cases in which 100% of retweets are made in one hour might be exceptional tweets including spam tweets, our observation shows that in more than half (retweeted) tweets, they get more than 50% retweets in one hour. This result is consistent with the result presented in Kwak et al.(2010).

All users do not necessarily see their timelines always. Users can make a retweet only when they are watching their timelines. Thus, retweeting cascades happen only when real-time active users can see the original tweet. This implies that the number of real-time active users can be a key for information spreading by retweets. Especially, for the users who don't have a huge volume of followers, timely posts are important. If the user has millions of followers, the possibility that someone is on timeline increases naturally. However, if the user have thousands of followers, that possibility should be lower. In this case, if the users want to make their post spread as far as they can, they should be conscious of real-time active followers.

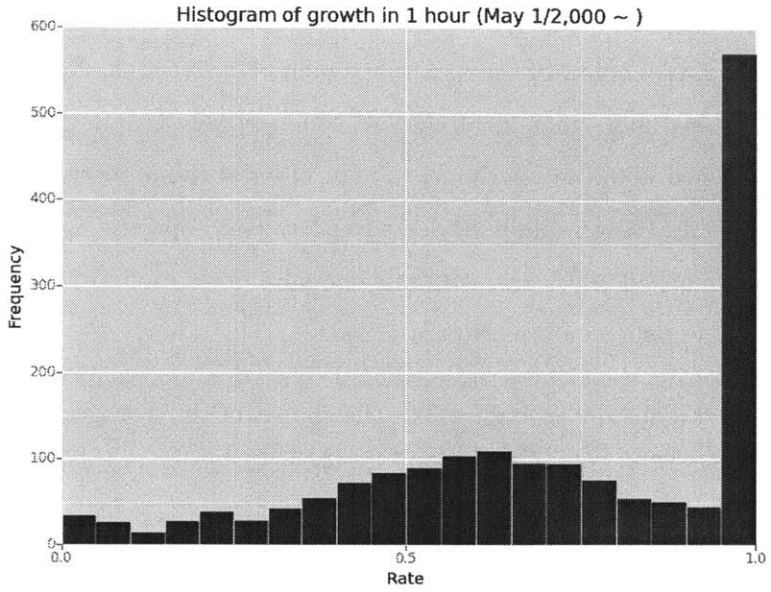


Figure 3-5: Growth rate of retweets in one hour (2000 over RT count)

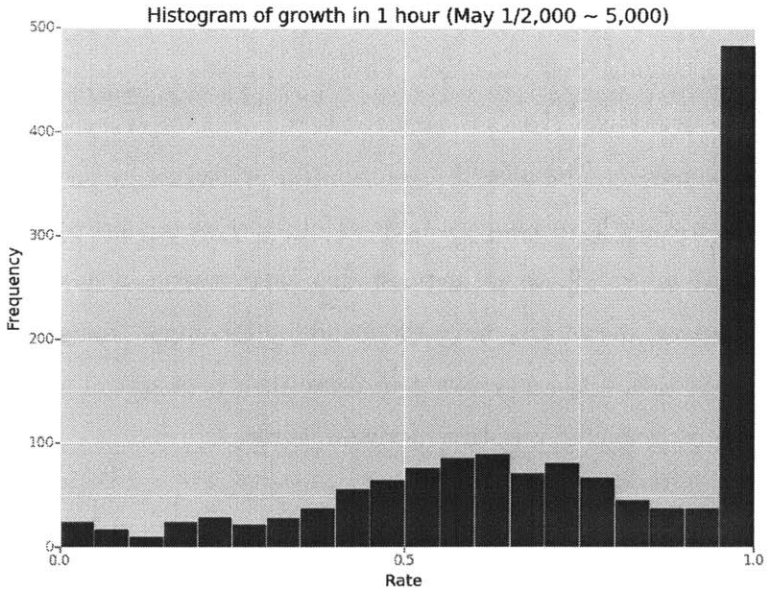


Figure 3-6: Growth rate of retweets in one hour (2000-5000 RT count)

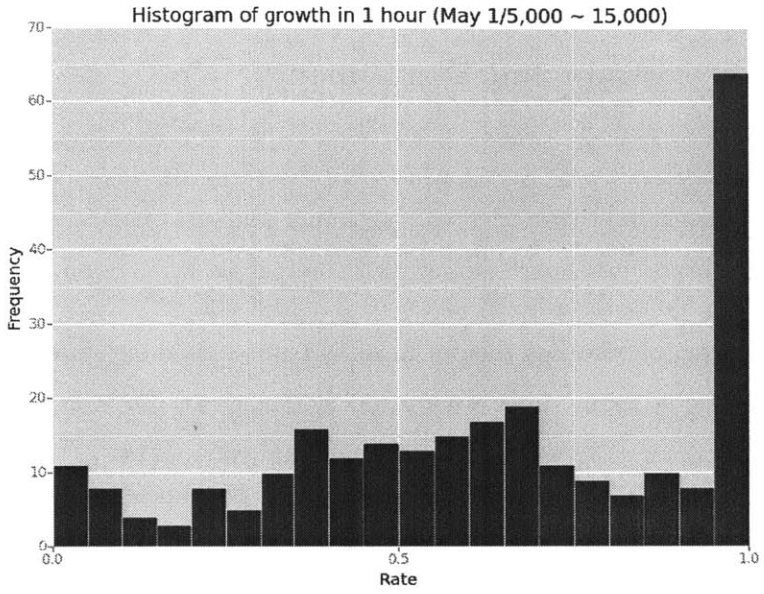


Figure 3-7: Growth rate of retweets in one hour (5,000-15,000 RT count)

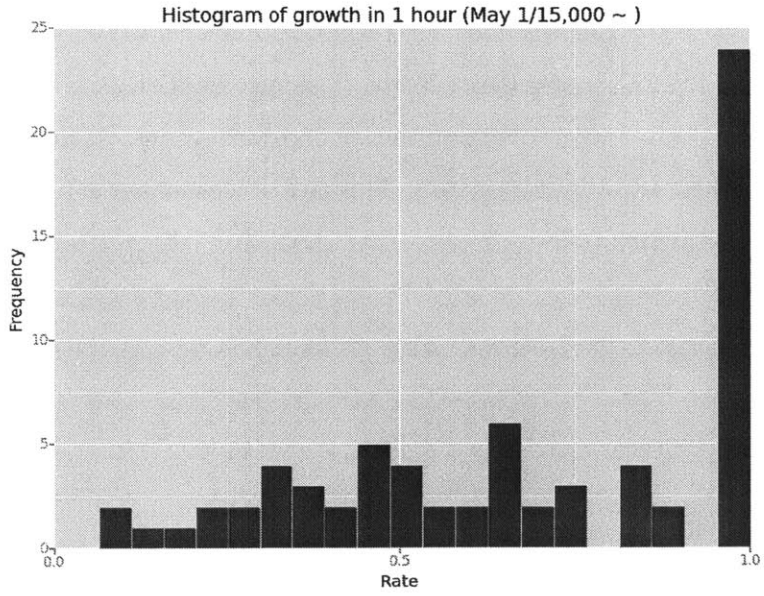


Figure 3-8: Growth rate of retweets in one hour (15,000- RT count)

3.3 Specific features of tweets and retweetability

3.3.1 Observation and Experiments

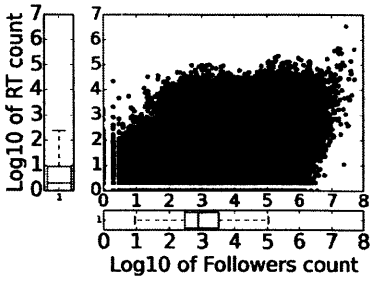
Followers

Intuitively, it is natural that a user who has many followers tends to get more retweets, because those users' tweets are viewed by many people who are potential retweeters.

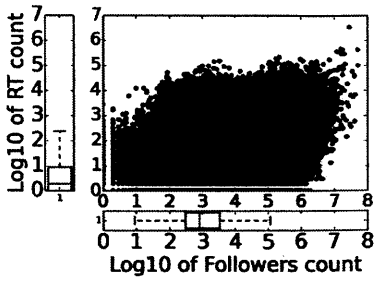
Table 3.7 roughly shows that this intuition is true. This table shows most of tweets which are retweeted many times are created by the "giant" accounts that have a huge number of followers. Even for the tweets whose retweet count is in the rage of $1 \sim 10$, the median of the number of followers is around 700.

Each original tweet data set includes its retweet count at the time that tweet was collected. The data also contains the user information about who posted the original tweet including the followers count.

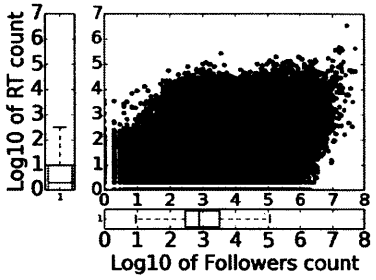
Figure 3-9 shows the relationship between the retweet count and the followers count. In this scatter plot, x-axis stands for the logarithm of follower count, and y-axis stands for the logarithm of retweet count. This plot implies the positive correlation between the log of followers count and the log of retweet count. In fact, the correlation coefficient of these two variables is 0.48. This means the number of followers positively affected retweetability.



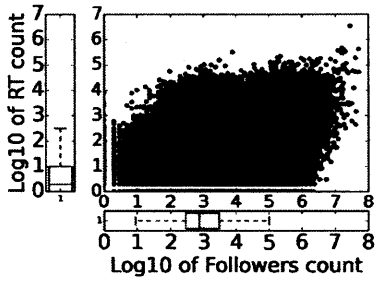
(a) May 1st



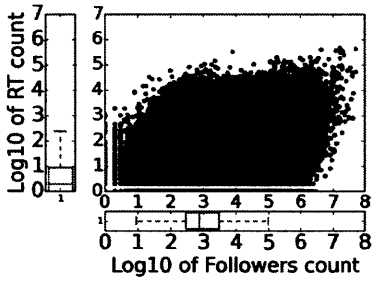
(b) May 2nd



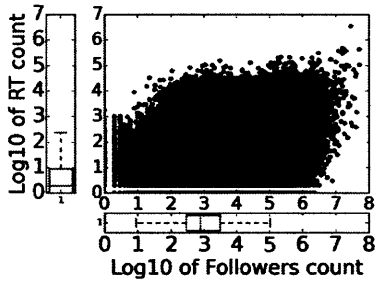
(c) May 3rd



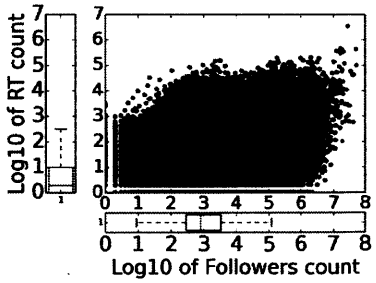
(d) May 4th



(e) May 5th



(f) May 6th



(g) May 7th

Figure 3-9: Log-log scatter plot with box-plot of followers count and retweet count

Table 3.6: Retweet count and Followers

Date	Retweet count range	min	median	mean	max
May 1 (Thu)	$1 \sim 10$	0	665	11,442.04	27,411,430
	$11 \sim 100$	0	5672.5	75,510.58	41,619,664
	$101 \sim 1,000$	0	25150.5	224,208.36	52,655,248
	$1,001 \sim 10,000$	0	110,576.5	762,358.52	52,657,637
	$10,001 \sim 100,000$	49	131.160	3,260,100.40	51,273,391
	$100,001 \sim$	16,498,576	16,498,576	16,498,576	16,498,576
May 2 (Fri)	$1 \sim 10$	0	682	11,736.13	30,113,411
	$11 \sim 100$	0	5,730	76,758.13	42,784,008
	$101 \sim 1,000$	0	21,904	225986.08	52,679,161
	$1,001 \sim 10,000$	0	75,252.5	689,851.38	52,682,116
	$10,001 \sim 100,000$	44	24,408	2,535,188.97	51,297,967
	$100,001 \sim$	3,183,938	9,850,031	9,850,031	16,516,124
May 3 (Sat)	$1 \sim 10$	0	664	11,341.61	24,003,107
	$11 \sim 100$	0	4994	70303.60	42,796,820
	$101 \sim 1,000$	0	21,584	212,077.51	42,800,892
	$1,001 \sim 10,000$	0	81,106	754087.80	52,708,712
	$10,001 \sim 100,000$	154	131,430	4115915.40	51,322,832
	$100,001 \sim$	–	–	–	–
May 4 (Sun)	$1 \sim 10$	0	649	10460.80	16,018,437
	$11 \sim 100$	0	4,766	66,434.67	41,693,380
	$101 \sim 1,000$	0	21,397	202807.38	42,815,384
	$1,001 \sim 10,000$	0	100,562	755,088.30	52,733,091
	$10,001 \sim 100,000$	44	39,790	2,266,124.90	51,349,948
	$100,001 \sim$	12,456,140	16,421,385.5	16,421,385.5	20,386,631
May 5 (Mon)	$1 \sim 10$	0	653	10,914.23	16,455,845
	$11 \sim 100$	0	4,801	70144.26	41,715,758
	$101 \sim 1,000$	0	20,213	210,255.36	42,831,974
	$1,001 \sim 10,000$	0	69,335	621627.90	52,755,069
	$10,001 \sim 100,000$	16	107,798.5	4,301,185.90	51,373,314
	$100,001 \sim$	–	–	–	–
May 6 (Tue)	$1 \sim 10$	0	667	11507.95	21,019,730
	$11 \sim 100$	0	4,824	71,828.07	28,816,317
	$101 \sim 1,000$	0	17,476	206,577.48	42,849,894
	$1,001 \sim 10,000$	0	61,793	681,515.05	52,778,927
	$10,001 \sim 100,000$	34	129,612	2,025,612.22	22,111,705
	$100,001 \sim$	–	–	–	–
May 7 (Wed)	$1 \sim 10$	0	688	12,118.77	21,047,502
	$11 \sim 100$	0	4825	73,097.15	41,763,478
	$101 \sim 1,000$	0	15,792	199,764.46	52,802,836
	$1,001 \sim 10,000$	0	78,711	739,920.52	52,802,140
	$10,001 \sim 100,000$	67	102,847.5	2,917,596.06	51,420,025
	$100,001 \sim$	16,585,689	16,585,689	16,585,689	16,585,689

Hashtag

Tweets containing hashtags are potentially viewed by the people outside of the followers of the tweeters. For this reason, tweets with hashtags are expected to get more retweetability than tweets without hashtags.

Table 3.7 shows the volume of tweets and retweets caught by Streaming API. In this table, "RT", "HT", and "Tw" stand for retweets, hashtags, and tweets respectively. "RT w HT" means the volume of retweets with hashtags, and "RT wo HT" means the volume of retweets without hashtags. Similarly, "Tw w HT" means the volume of tweets with hashtags, and "Tw wo HT" means the volume of tweets without hashtags.

Table 3.7: The volume of Tweet and Retweet (proportion) on each day

Date	RT w HT	RT wo HT	Tw w HT	Tw wo HT
May 1 (Thu)	301,665	1,140,834	369,303	2,602,589
May 2 (Fri)	289,216	1,084,052	352,877	2,554,422
May 3 (Sat)	279,350	1,072,458	336,110	2,543,774
May 4 (Sun)	301,978	1,145,784	348,318	2,638,387
May 5 (Mon)	307,767	1,160,351	373,193	2,652,417
May 6 (Tue)	305,854	1,135,221	376,571	2,659,885
May 7 (Wed)	314,606	1,097,151	351,522	2,477,820

By the observations in Table 3.7, the rate of the posts including hashtags can be calculated. The calculated values are in Table 3.8. This table shows that retweets include more hashtags than tweets. In other words, it indirectly suggests the interpretation that the retweets tend to target the tweets with hashtags.

Here, we propose another aspect of retweetability with hashtags, that is, whether including hashtags increases the possibility to get at least one retweet or not. In order to examine whether this intuition is true, we compare the two ratios; $P(R|H)$ and $P(R|\bar{H})$, where R stands for the observed number of the retweeted tweets, and H stands for the number of the tweets with hashtags. To calculate those ratios, we randomly chose $10,000$ tweets among the tweets which are collected via Streaming

API on each day, and checked those ex-post retweet counts. Table 3.10 to Table 3.12 show the results of this experiment. By the figures in Table 3.10, $P(R|H)$ and $P(R|\bar{H})$ on each day are calculated. Table 3.11 shows those ratios. On each day, the value of $P(R|H)$ and $P(R|\bar{H})$, are close to each other. In fact, Pearson's chi-squared test for independence doesn't reject the null hypothesis that states those two ratios are equal on each day. The results of this chi-squared test are in Table 3.12. Here, all $p-$values are more than 0.05.

Moreover, on 2nd, 4th, 6th, and 7th in May, $P(R|H)$ is less than $P(R|\bar{H})$, though the null hypothesis with 95% significance is not rejected.

Therefore, our experiment doesn't support the hypothesis that states including hashtags helps retweetability of this sense.

Table 3.8: Proportion of Tweet and Retweet including Hashtags on each day

Date	HT in RT	HT in Tw
May 1 (Thu)	20.91%	12.43%
May 2 (Fri)	21.06%	12.14%
May 3 (Sat)	20.66%	11.67%
May 4 (Sun)	20.86%	11.66%
May 5 (Mon)	20.96%	12.33%
May 6 (Tue)	21.22%	12.40%
May 7 (Wed)	22.28%	12.42%

Table 3.9: Tweet Volume table for calculating Chi-square statistic (Hashtag)

Date	$\lvert H \cap R \rvert$	$\lvert H \cap \bar{R} \rvert$	$\lvert \bar{H} \cap R \rvert$	$\lvert \bar{H} \cap \bar{R} \rvert$
May 1 (Thu)	132	983	800	6,733
May 2 (Fri)	121	1,052	814	6,625
May 3 (Sat)	119	953	789	6,648
May 4 (Sun)	123	998	819	6,635
May 5 (Mon)	105	1,008	843	6,647
May 6 (Tue)	133	1,112	794	6,524
May 7 (Wed)	111	959	819	6,779

Table 3.10: $P(R|H)$ and $P(R|\bar{H})$ Ratios (Hashtags)

Date	$P(R\|H)$	$P(R\|\bar{H})$
May 1 (Thu)	11.84%	10.62%
May 2 (Fri)	10.32%	10.94%
May 3 (Sat)	11.10%	10.61%
May 4 (Sun)	10.97%	10.99%
May 5 (Mon)	9.43%	11.26%
May 6 (Tue)	10.68%	10.85%
May 7 (Wed)	10.37%	10.78%

Table 3.11: Chi-square statistic (Hashtags)

Date	Chi-squre statistic	$p-$value
May 1 (Thu)	1.3759	0.2408
May 2 (Fri)	0.3492	0.5546
May 3 (Sat)	0.1888	0.6639
May 4 (Sun)	0.0013	0.9711
May 5 (Mon)	3.0942	0.0786
May 6 (Tue)	0.0159	0.8996
May 7 (Wed)	0.1213	0.7276

URLs

The tweets containing URLs are thought to include more information than the tweets only with text within 140 characters. Thus, including URLs possibly increases retweetability.

Here, we examine this hypothesis with the same method as above. In the table below, "U" stands for URL. In this case, $P(R|U)$ is less than $P(R|\bar{U})$ on 3rd, 4th, 6th, and 7th in May. Table 3.13 shows the tweet volume of each category, and Table 3.12 shows the results of Pearson's chi-squared test for independence. Only on 4th in May, the null hypothesis that states $P(R|U)$ is equal to $P(R|\bar{U})$ is rejected. However, on this day, $P(R|U) < P(R|\bar{U})$, i.e., including URL reduced retweetability.

This experiment doesn't support the hypothesis that states including URLs helps retweetability.

Table 3.12: Tweet Volume table for calculating Chi-square statistic (URLs)

Date	$\lvert U \cap R \rvert$	$\lvert U \cap \bar{R} \rvert$	$\lvert \bar{U} \cap R \rvert$	$\lvert \bar{U} \cap \bar{R} \rvert$
May 1 (Thu)	175	1,340	757	6,376
May 2 (Fri)	173	1,318	762	6,359
May 3 (Sat)	146	1,354	762	6,247
May 4 (Sun)	147	1,367	795	6,266
May 5 (Mon)	157	1,343	791	6,312
May 6 (Tue)	133	1,250	771	6,386
May 7 (Wed)	168	1,417	762	6,321

Table 3.13: $P(R|U)$ and $P(R|\bar{U})$ Ratios (URLs)

Date	$P(R\|U)$	$P(R\|\bar{U})$
May 1 (Thu)	11.55%	10.61%
May 2 (Fri)	11.60%	10.70%
May 3 (Sat)	9.73%	10.87%
May 4 (Sun)	9.71%	11.26%
May 5 (Mon)	9.90%	11.14%
May 6 (Tue)	9.62%	10.77%
May 7 (Wed)	10.60%	10.76%

Table 3.14: Chi-square statistic (URLs)

Date	Chi-squre statistic	$p-$value
May 1 (Thu)	1.0491	0.3057
May 2 (Fri)	0.9458	0.3308
May 3 (Sat)	1.5625	0.2113
May 4 (Sun)	2.9051	0.0883
May 5 (Mon)	0.4999	0.4796
May 6 (Tue)	0.0955	0.7573
May 7 (Wed)	0.0195	0.8889

User-mention

User-mention (replying) is basically thought to be a function for conversation with some specific users. In fact, including user-mention reduces the potential retweeters because the user-mentioning tweets are only visible on timeline of the people who follow the both accounts. Thus, intuitively, it is possibly expected that using user-mention reduces retweetability. Table 3.15 to 3.18 show the results. Here, "M" stands for the user-mention.

Intuitively, user-mention can reduce the possibility to get retweets because tweets with user-mentions seem like more private tweets than tweets without user-mentions. However, this results implies that using user-mention does not reduce the chance to get one or more retweets. At the same time, user-mention does not help to get retweetability.

Table 3.15: Tweet Volume table for calculating Chi-square statistic (User-mentions)

Date	$\lvert M \cap R \rvert$	$\lvert M \cap \bar{R} \rvert$	$\lvert \bar{M} \cap R \rvert$	$\lvert \bar{M} \cap \bar{R} \rvert$
May 1 (Thu)	333	2,748	599	4,968
May 2 (Fri)	321	2,549	614	5,128
May 3 (Sat)	313	2,667	595	4,934
May 4 (Sun)	346	2,692	596	4,941
May 5 (Mon)	349	2,622	599	5,033
May 6 (Tue)	316	2,587	611	5,049
May 7 (Wed)	305	2,623	625	5,115

Table 3.16: $P(R|M)$ and $P(R|\bar{M})$ Ratios (User-mentions)

Date	$P(R\|M)$	$P(R\|\bar{M})$
May 1 (Thu)	10.81%	10.76%
May 2 (Fri)	11.18%	10.69%
May 3 (Sat)	10.50%	10.36%
May 4 (Sun)	11.39%	10.76%
May 5 (Mon)	11.75%	10.64%
May 6 (Tue)	10.89%	10.80%
May 7 (Wed)	10.42%	10.89%

Table 3.17: Chi-square statistic (User-mentions)

Date	Chi-squre statistic	$p-$value
May 1 (Thu)	0.0011	0.9735
May 2 (Fri)	0.4283	0.5128
May 3 (Sat)	0.1096	0.7406
May 4 (Sun)	0.7213	0.3957
May 5 (Mon)	2.3374	0.1269
May 6 (Tue)	0.0082	0.9279
May 7 (Wed)	0.4028	0.5257

Including pictures

Generally, it is said that including pictures boosts getting retweets. For example,Twitter Official blog reports that photos boost retweets by 35%. In addition, intuitively including a picture might possibly affects retweetability. Thus we tried to test the hypothesis that states including a picture helps to get 1 or more retweets with the same method above. The results are shown in Table 3.18 to 3.20. The figures in table 3.17 reject the difference between tweets with a picture and tweets without a picture at 95% significance level. However, the values of $P(R|I)$ in table 3.16 tend to be relatively larger than the values of $P(R|H)$, $P(R|U)$, and $P(R|M)$. This result implies that including a picture slightly affects the initial retweetability. In the next chapter, we try to figure out how this occurs.

Table 3.18: Tweet Volume table for calculating Chi-square statistic (Pictures)

Date	$\lvert I \cap R \rvert$	$\lvert I \cap \bar{R} \rvert$	$\lvert \bar{I} \cap R \rvert$	$\lvert \bar{I} \cap \bar{R} \rvert$
May 1 (Thu)	55	392	788	6,450
May 2 (Fri)	52	432	872	7,162
May 3 (Sat)	43	422	835	6,944
May 4 (Sun)	58	411	875	7,150
May 5 (Mon)	62	415	880	7,203
May 6 (Tue)	71	402	851	7,165
May 7 (Wed)	56	413	866	7,265

Table 3.19: $P(R|I)$ and $P(R|\bar{I})$ Ratios (Pictures)

Date	$P(R\|I)$	$P(R\|\bar{I})$
May 1 (Thu)	12.30%	10.89%
May 2 (Fri)	10.74%	10.86%
May 3 (Sat)	9.25%	10.73%
May 4 (Sun)	12.37%	10.90%
May 5 (Mon)	13.00%	10.57%
May 6 (Tue)	15.01%	10.62%
May 7 (Wed)	11.94%	10.65%

Table 3.20: Chi-square statistic (Picture)

Date	Chi-squre statistic	$p-$value
May 1 (Thu)	0.7268	0.3939
May 2 (Fri)	0.0000	0.9997
May 3 (Sat)	0.8689	0.3513
May 4 (Sun)	0.8265	0.3633
May 5 (Mon)	1.8394	0.1750
May 6 (Tue)	8.4603	0.0036
May 7 (Wed)	0.6417	0.4231

Chapter 4

Retweets of picture tweets and the followers' networks

In this chapter, we compare the tweets containing pictures and the tweets which only contains text. Comparing the ratios of retweets by followers of the author of the original tweets among the initial 50 retweets, tweets with a picture have slightly lower ratio, though there is no significant difference between the average for tweets with pictures and without pictures at the 95% significance level. We also investigate how many retweets are posted by users in the followers' network connected to the original tweeter, and show that the lengths of a retweeters' network for tweets with picture have larger variance than that of tweets without pictures. These results imply that a tweet including picture can reach more people than a tweet without a picture potentially.

4.1 Concepts and Hypothesis

According to Twitter's official blog posted on March 10, including photos provides a 35% boost in retweets. However, the background of this phenomenon is not very obvious.

Intuitively, it is expected that text-only tweets tend to depend on the contexts of the

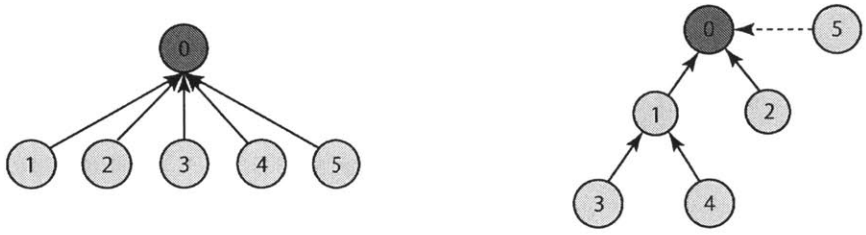
community more than tweets containing pictures. In other words, the retweeters of text-only tweets are expected to be nearer than the retweeters of tweets with pictures.

If this intuition is true, retweets of a tweet with a picture potentially expands to the people who doesn't know the tweeter originally.

Throughout this chapter, we focus on the relationship between retweeters and the author of the original tweet in order to compare the information diffusion by retweets of a text-only tweet and retweets of a tweet with a picture. Especially, we examine the followers' network connected to the original tweeter.

Figure 4-1 shows 2 imaginary cases of this network. In both cases, the node marked "0" stands for the original tweeter, and the node 1 to 5 stand for the 5 initial retweeters. In the case of (a) in figure 4-1, all of the 5 initial retweeters are the followers of the original tweeter. In contrast, in the case of (b), only 2 retweeters (user 1 and user 2) among the 5 initial retweeters follow the original retweeter. User 3 to 5 don't follow the original tweeter, while user 3 and user 4 follow user 1. In this case, it is probable that user 3 and user 4 noticed the original tweet through the retweet posted by user 1. User 5 is an outsider who don't follow the original tweeter, nor user 1 to 4. We don't specify how these outsiders noticed the original tweet, though they can notice tweets by a user whom they don't follow through searching Twitter, the tweeter's blog, or other web services.

In order to examine these networks from statistical viewpoint, we focus on three values; how many retweeters follow the original tweeter, how long these networks are, and how many outsiders retweet the tweets. For example, in the case of (a) in figure 4-1, all the retweeters follow the original tweeter, and the length of the network is 1, and no outsider retweets the tweet. In contrast, in the case of (b), 2 retweeters among 5 follow the original tweeter, and the length of the network is 2, and 1 outsider retweets the tweet. Comparing these 2 cases, the network of (b) is longer than that of (a), and (b) has more outsiders than (a). Hence we can interpret that the retweeters' network of the tweet of (b) is more open to people who are not connected



(a) Example 1 (b) Example 2

Figure 4-1: Examples of followers' network of retweeters

to the original tweeter directly than (a). Moreover, it is inferred that the tweet which is similar to (b) tends to spread more widely. In other words, if a tweet is more context-oriented, the retweeters' relationship should be more closed, and vice versa.

In the remainder of this chapter, we examine these values to compare the retweeting behavior of a text-only tweet and a tweet with a picture.

4.2 Data description

The process of data sampling is as follows.

First, we collect the retweets whose retweet count is more than 50 and less than or equal to 100. In order to fix the other conditions, we filtered the tweets as follows.

- excluding tweets with hashtags
- excluding tweets with URLs of other websites

- excluding tweets with user-mentions
- the original tweeter's follower count is between 800 and 1200
- the retweet count converges between 80 and 120
- posted during the 3rd and 4th week in August 2014

Second, we use the "GET statuses/retweets/:id" in REST API which provides the 100 most recent retweets of the tweet. From these retweet IDs, we gather the user-ID of the retweeters. Moreover, in order to see the following relationship among the retweeters of the same tweet, we collected the original tweeter's and the initial 30 retweeters' followers' user-IDs.

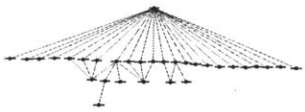
To compare the two types of tweets, we collected 100 text-only tweets and 100 tweets with a picture and investigated the retweeters' following network structures.

4.3 Results

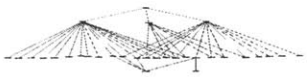
In this section, we show the result of analysis on the data collected with the method above. First, we demonstrate 6 actual examples of the followers' network connected to the original tweeter in figure 4-2.

The network of (a) in figure 4-2, 23 of the initial 30 retweeters are the followers of the original tweeter. The number of the outsider is 0, and the length of this network is 3. In this case, nearly 76.7% in the initial retweeters are the followers of the original tweeter, and 20% of them are the followers' follower.

The network of (b), the number of direct followers of the original tweeter among the initial 30 retweeters is 3. However, the number of retweeters with distance 2 from the original tweeter is 25. In other words, in this case more users who don't follow the original tweeter retweet the tweet than the direct followers of the original tweeter. The length of this network is 3, and the number of outsiders is 0.

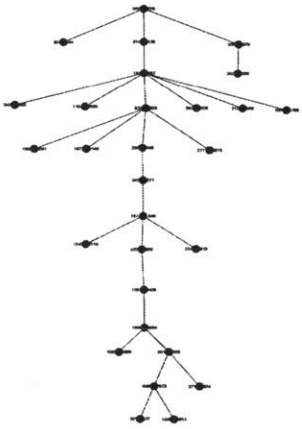


(a) Sample016 of text-only-tweet

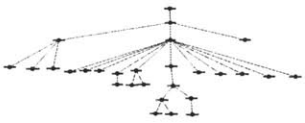


(b) Sample046 of text-only-tweet

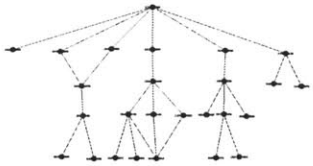
(c) Sample056 of text-only-tweet



(d) Sample010 of picture-tweet



(e) Sample037 of picture-tweet



(f) Sample057 of picture-tweet

Figure 4-2: Examples of actual followers' network of retweeters

Case (c) is one of the simplest cases of the followers' network connected to the original tweeter. In this case, all of the initial 30 retweeters are the direct followers of the original tweeter. Obviously the length of this network is 1, and the number of outsiders is 0.

(d) is the most deepest case in our data. The length of this network is 12. Three of the initial retweeters are the direct follower of the original tweeter, and the number of outsiders is two. Two of our sample cases for picture tweets have 12-length networks, while the length of the deepest network of text-only tweets in our data is eight.

In the case (e), the number of the direct follower of the original tweeter is only one, and the number of the followers of the direct followers is three. However, the number of the retweeters whose distance from the original tweeter is three is 14. In this case, one of the retweeter whose distance from the original tweeter is 2 was the trigger of 11 retweeters. The length of this network is seven, and the number of outsiders is three.

In the case (f), nearly equal number of retweeters are in each length of the network. The number of the direct followers of the original tweeter is six, and the number of the followers of the direct followers is five. Seven retweeters are in the length three, and other seven retweeters are in the length four. The length of this network is four, and the number of outsiders is five.

Table 4-4 and 4-7 located in the end of this chapter show that how many retweeters are at each length of the network for all our samples. We analyze these samples statistically below.

4.3.1 Retweets by the direct followers of the original tweeter

First, we observed the number of retweets posted by the direct followers of the original tweeters among the first 50 retweeters. Figure 4-3 shows the histogram of the number of retweets by the direct followers. The left figure stands for non-picture tweets, and

the right figure stands for picture tweets.

In both group, the rates of retweets by the direct followers are not very large. In most cases, the share of direct followers is less than 1/3.

This observation implies that at least in the case of tweets whose retweet count is around 100, retweets usually expand to outside of the direct followers.

Comparing the histograms for text-only tweets and for tweets with a picture in figure 4-3, the distribution of the number of direct followers for tweets with a picture is slightly located in the left to that of text-only tweets. In other words, tweets with a picture are more retweeted by the users who don't directly follow the original tweeter than text-only tweets. This observation is consistent with the assumption of which tweets with a picture tend to expand wider than text-only tweets, though there doesn't seem to see significant difference between the two groups.

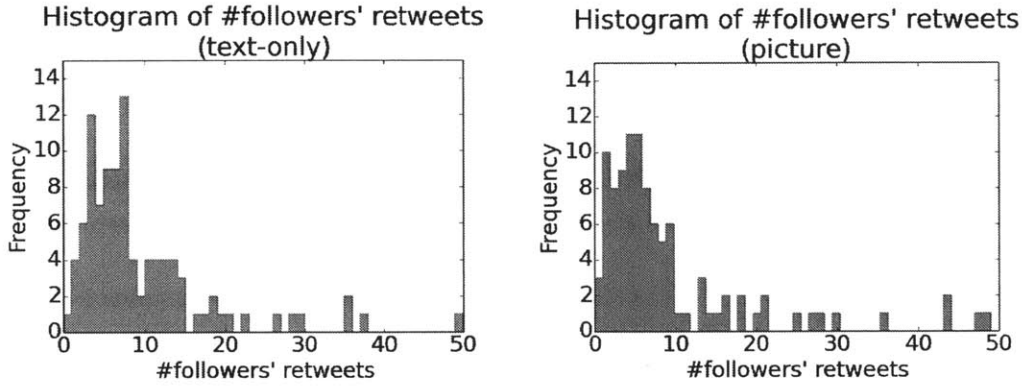
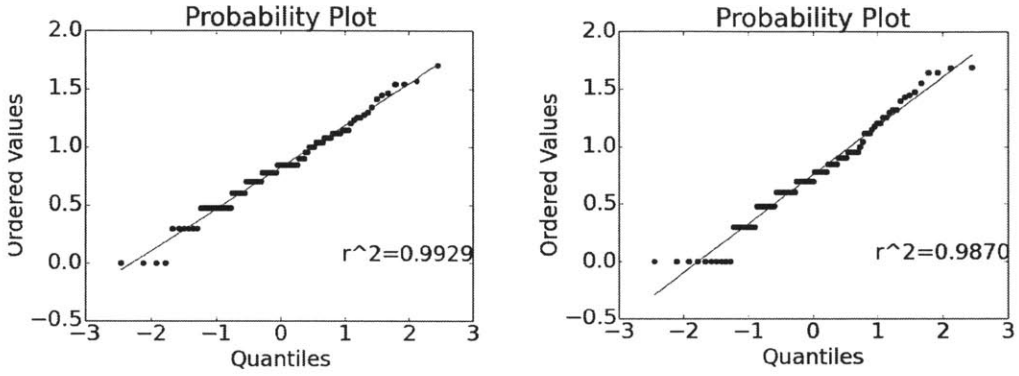


Figure 4-3: Retweets by direct followers of the original tweeter

In fact the average number of direct retweeters of non-picture tweets is 9.16 in 50, and that of picture tweets is 8.86. We tested the null hypothesis that assume there is no significant difference between the two averages, and the t-statistic was 0.22 and p-value was 0.82. Thus, this result does not reject the null hypothesis.

In terms of the shape of the distributions above we show the Q-Q plot for the logarithm of the number of the direct followers' retweets for these two types of tweets in figure

4-4. Comparing these two plots, the number of retweets by the direct followers of the original tweeter for text-only tweets is more likely log-normally distributed than that for picture-tweets. In fact, the Shapiro-Wilk test doesn't reject the log-normality for text-only tweets, while it rejects the log-normality for picture-tweets. Table 4-1 shows this result.



(a) text-only tweets

(b) picture tweets

Figure 4-4: Q-Q plot for the log of the number of retweets by the direct followers of the original tweeter

Table 4.1: Summary of the distribution of the number of retweets by the direct followers

	text-only	picture
Average	9.1600	8.8600
Variance	72.4344	106.1404
W of Shapiro-Wilk (log-)normality test	0.9839	0.9682
p-value for Shapiro-Wilk test	0.2695	0.0187

4.3.2 Retweets by followers' network

We observe retweeters' distance from the original tweeter. Figure 4-4 to 4-7 show histograms of the number of retweeters within each distance from the original tweeter. Figure 4-4 and 4-5 are for text-only tweets, and figure 4-6 and 4-7 are for picture tweets.

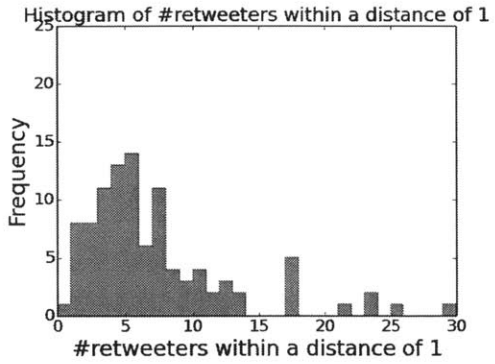
Figure 4-5(a) and 4-5(a) have similar shape of distribution as the histogram of Figure 4.3. By including more distant retweeters, the weights of distributions shift to the right naturally. Looking carefully at this shift, picture tweets moves slower than text-only tweets. Besides, the shape of the distribution for picture tweets is more moderate than that for text-only tweet.

Table 4.3 shows the average and variance of the coverage of each distance. The high average number of coverage within narrower networks implies that the network is more dense than the other type. In terms of this number, coverage numbers of text-only tweets is consistently larger than that of picture tweets. This implies that the number of retweeters within the followers' network connected to the original tweeter for text-only tweets are larger than that for picture tweets, though there are not statistically significant differences between them without the case for 2 to 5 distance coverages. Thus our observation implies there is a slight tendency that text-only tweets have a smaller retweeters network than picture tweets. The raw data are in Table 4.4 to 4.7. Table 4.4 and 4.5 are for the text-only tweets, and table 4.6 and 4.7 are for picture tweets.

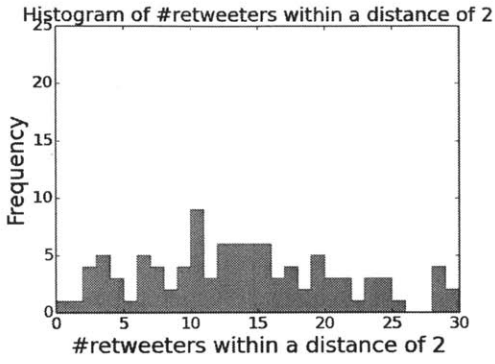
At the same time, the variances for picture tweets are consistently larger than that for text-only tweets. This implies that there are some picture tweets which have smaller network than text-only tweets, though other picture tweets have a wider network.

In both cases for text-only tweets and picture tweets, there are some retweeters who are not in the followers' network connected to the original tweeters. Figure 4-8 shows the histogram of those outsiders. This histogram shows that the distribution for text-only tweets has understandable shape. The case for more outsider retweeters is rarer. However, this is not necessarily true for picture tweets. In our observation, 11 picture tweets among 100 get more than 90% retweets from outsiders. Thus, this observation implies these two types of tweets possibly have different distributions for the number of retweeters outside of the followers network connected to the original tweeters.

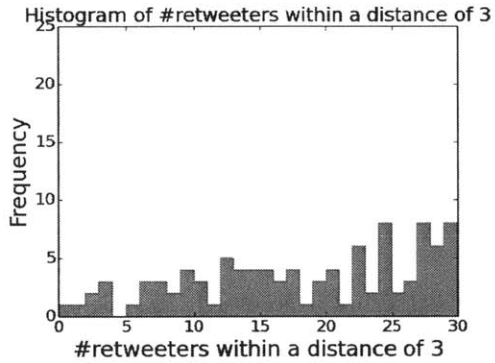
(a) within a distance of 1



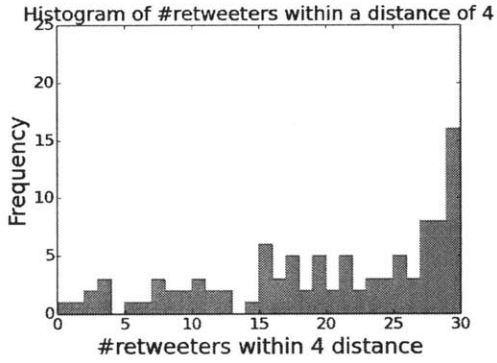
(b) within a distance of 2



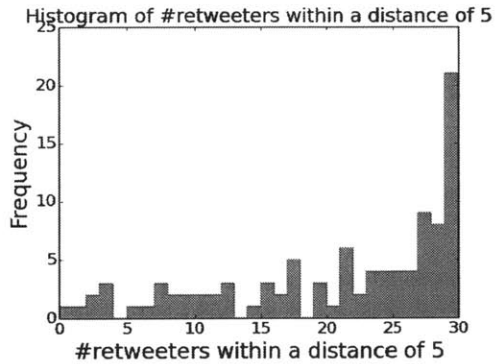
(c) within a distance of 3



(d) within a distance of 4



(e) within a distance of 5



(f) within a distance of 6

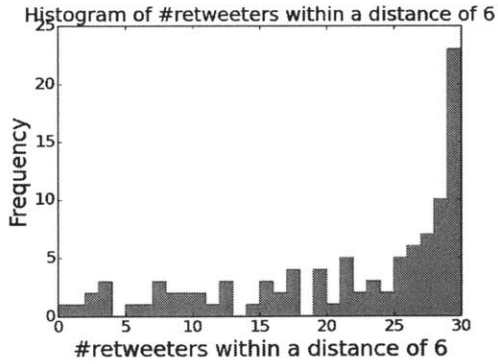
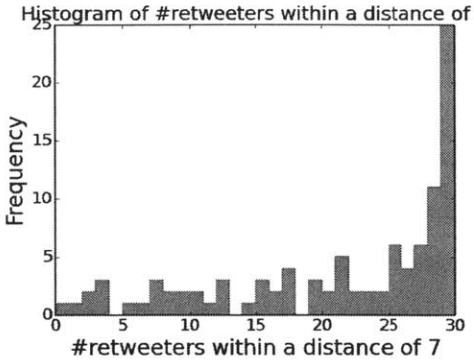
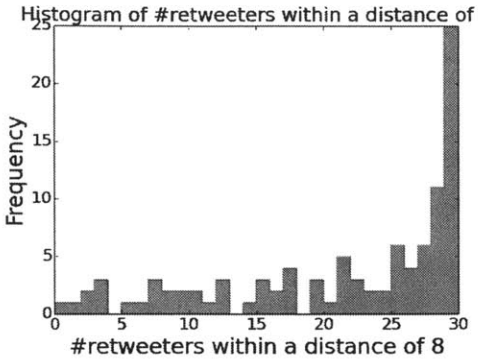


Figure 4-5: Histogram of #Retweeters within followers' network (text-only) (1)

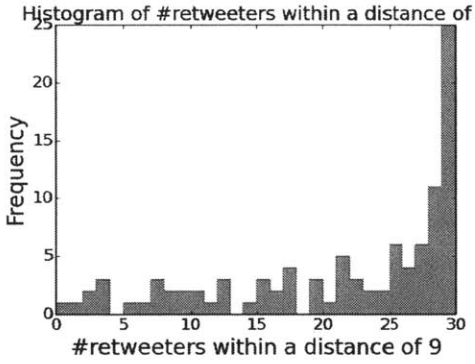
(a) within a distance of 7



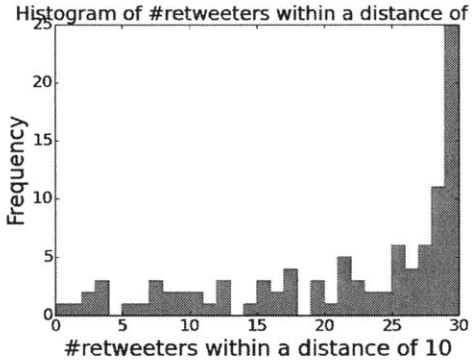
(b) within a distance of 8



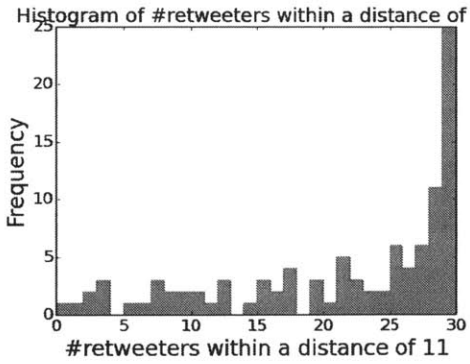
(c) within a distance of 9



(d) within a distance of 10



(e) within a distance of 11



(f) within 12 distance

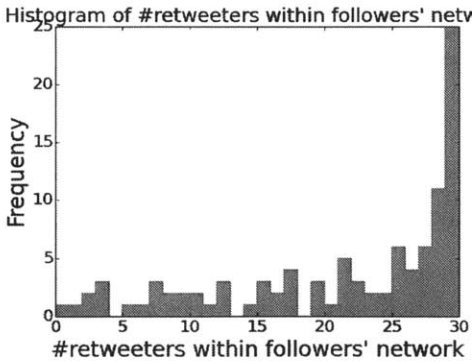
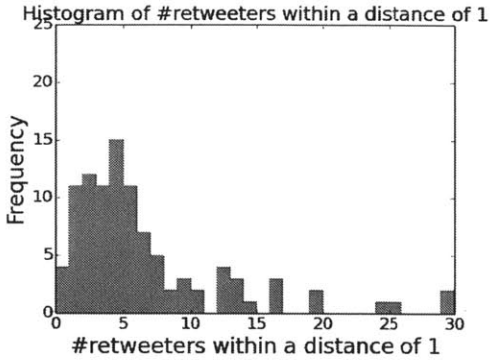
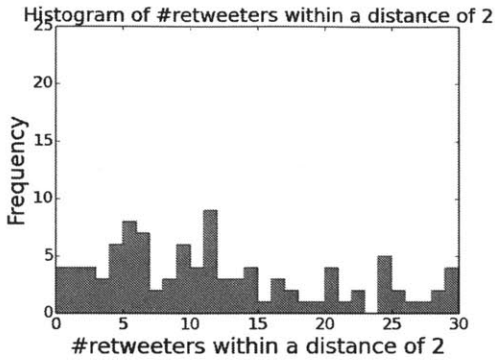


Figure 4-6: Histogram of #Retweeters within followers' network (text-only) (2)

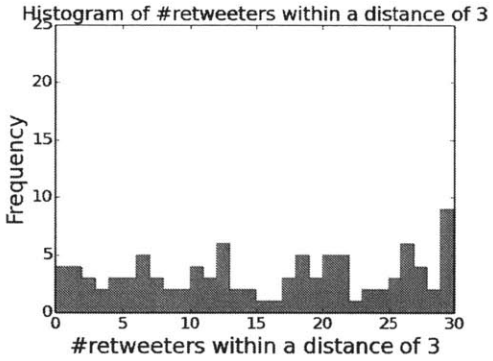
(a) within a distance of 1



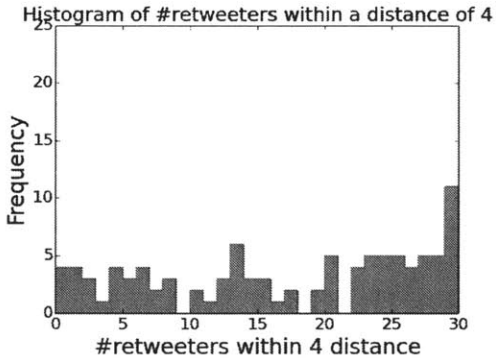
(b) within a distance of 2



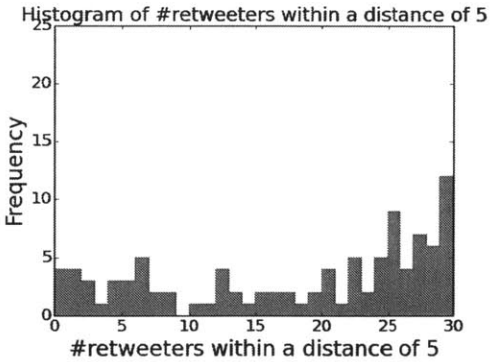
(c) within a distance of 3



(d) within a distance of 4



(e) within a distance of 5



(f) within a distance of 6

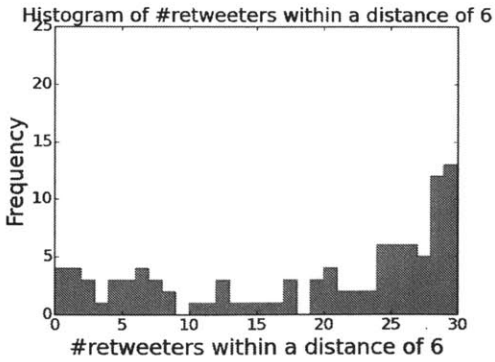
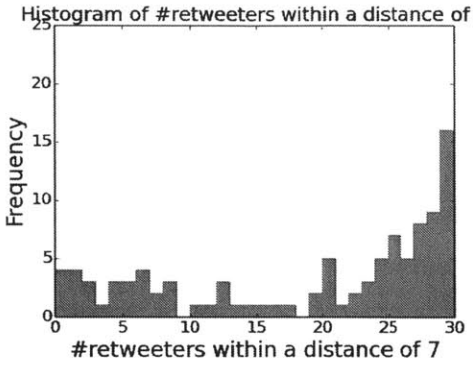
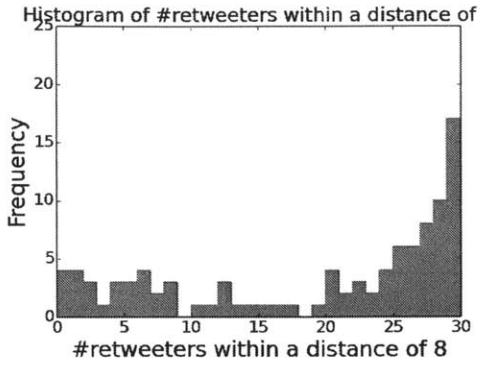


Figure 4-7: Histogram of #Retweeters within followers' network (picture) (1)

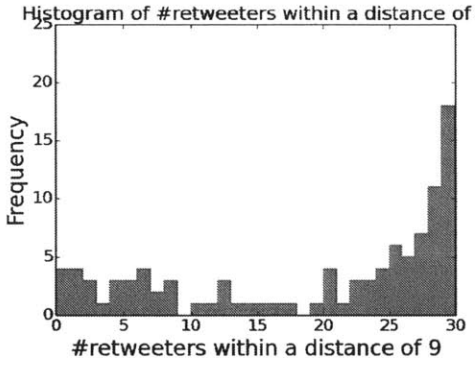
(a) within a distance of 7



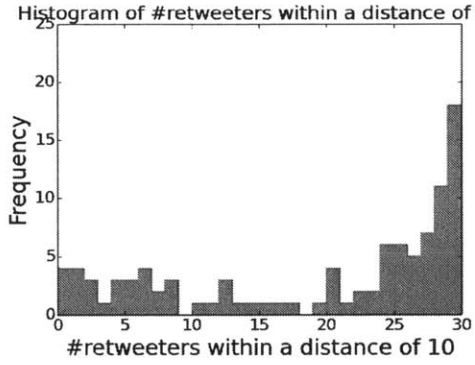
(b) within a distance of 8



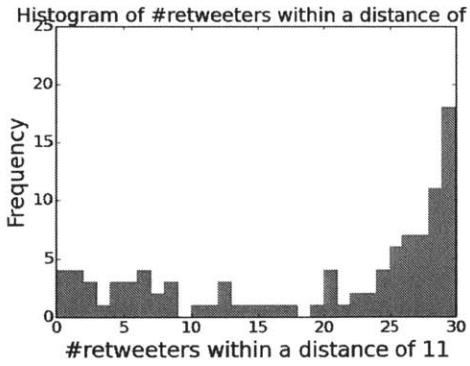
(c) within a distance of 9



(d) within a distance of 10



(e) within a distance of 11



(f) within 12 distance

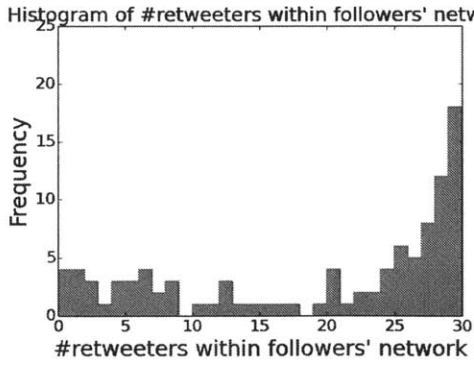


Figure 4-8: Histogram of #Retweeters within followers' network (picture) (2)

Table 4.2: Summary of retweeters' network

	Average		Median		Variance	
	text-only	picture	text-only	picture	text-only	picture
Within a distance of 1	6.82	6.19	5	4	31.81	36.79
Within a distance of 2	13.27	11.67	13	10	52.22	69.32
Within a distance of 3	17.90	15.38	19	16.5	70.07	87.06
Within a distance of 4	19.75	16.80	21	19	73.97	95.38
Within a distance of 5	20.65	17.81	23.5	21.5	77.53	99.09
Within a distance of 6	21.02	18.44	25	22.5	78.68	103.59
Within a distance of 7	21.12	18.73	25	23.5	79.67	105.76
Within a distance of 8	21.15	18.88	25	24	79.83	107.47
Within a distance of 9	21.15	18.95	25	24	79.83	108.49
Within a distance of 10	21.15	18.98	25	24	79.83	108.84
Within a distance of 11	21.15	19.02	25	24	79.83	109.22
Within a distance of 12	21.15	19.05	25	24	79.83	109.69
Outsiders	8.85	10.95	5	6	79.83	109.69

4.3.3 Lengths of retweeters' trees

The observation above can be explained by considering the lengths of retweeters' trees which start from the original tweeters.

Here, we assume the length of retweeters' network as the distance from the original tweeters to the most distant retweeters. Figure 4-9 shows the histogram of lengths. This figure implies that both types of tweets have similar distributions, and the variance of lengths for the picture tweets is larger than that for text-only tweets.

Table 4.3: Summary of the lengths of the retweeters' network connected to the original tweeter

	text-only	picture
Average	3.79	3.84
Median	4	3
Variance	2.3059	6.0544

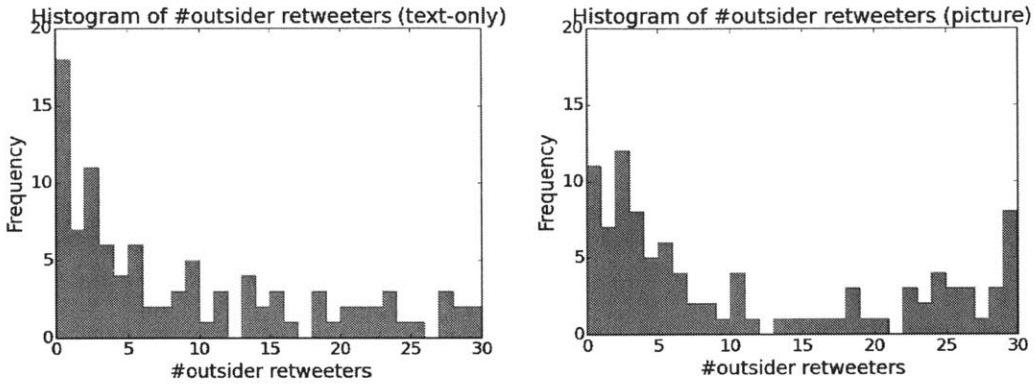


Figure 4-9: Outsider Retweeters

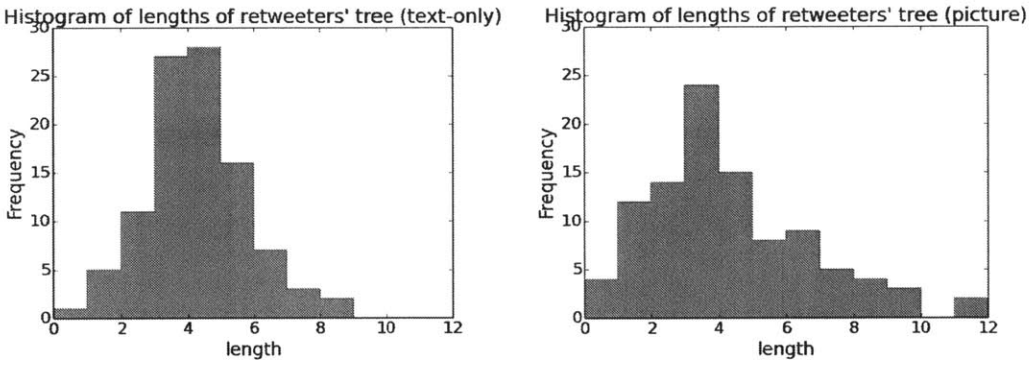


Figure 4-10: Lengths of network

4.4 Interpretation and Discussion

According to the observation above, the number of the followers of the original text-only tweeter is slightly larger than that of the original picture tweeter. In addition, the variance of the lengths of the retweeters' network of tweets with a picture is larger than that of text-only tweets. Moreover, the number of retweets by outsider of tweets with a picture is slightly larger than that of text-only tweets. To some extent, these results above support the hypothesis that tweets with a picture slightly tend to have opener retweeters' network than text-only tweets. In other words, our observation tends to support our hypothesis that text-only tweets tend to get their retweets from their community, and tweets with pictures can expand outside of the followers' network. In fact, retweeters' networks for some picture tweets reach further than those for text-only tweets.

However, in some cases of picture tweets have smaller networks than text-only tweets. In these cases, diffusion of tweets stops in a nearer area. For discussion, this observation is possibly interpreted as some pictures are more context-dependent than usual text-only tweets which get the same level of retweets. If this hypothesis is true, including pictures which are not context-dependent can be useful to diffuse the tweets further, although this observation does not contradict the fact that more picture tweets go further than text-only tweets. This point might be worth investigating in the future.

Table 4.4: The number of retweeters with each distance from the original tweeter (Text-only) (1)

		Distance												
	type	1	2	3	4	5	6	7	8	9	10	11	12	outsider
sample 001	text-only	5	8	13	4	0	0	0	0	0	0	0	0	0
sample 002	text-only	9	14	2	0	0	0	0	0	0	0	0	0	5
sample 003	text-only	8	5	11	3	1	0	0	0	0	0	0	0	2
sample 004	text-only	1	9	1	0	0	0	0	0	0	0	0	0	19
sample 005	text-only	7	14	7	2	0	0	0	0	0	0	0	0	0
sample 006	text-only	5	19	0	0	0	0	0	0	0	0	0	0	6
sample 007	text-only	5	9	6	4	2	1	2	0	0	0	0	0	1
sample 008	text-only	9	3	16	2	0	0	0	0	0	0	0	0	0
sample 009	text-only	1	2	6	0	0	0	0	0	0	0	0	0	21
sample 010	text-only	4	6	11	6	0	0	0	0	0	0	0	0	3
sample 011	text-only	4	2	2	2	1	19	0	0	0	0	0	0	0
sample 012	text-only	5	5	3	2	2	2	1	2	0	0	0	0	8
sample 013	text-only	5	10	3	3	0	0	0	0	0	0	0	0	9
sample 014	text-only	3	14	6	2	1	0	0	0	0	0	0	0	4
sample 015	text-only	5	11	13	1	0	0	0	0	0	0	0	0	0
sample 016	text-only	23	6	1	0	0	0	0	0	0	0	0	0	0
sample 017	text-only	4	11	9	3	0	0	0	0	0	0	0	0	3
sample 018	text-only	10	9	3	3	1	0	0	0	0	0	0	0	4
sample 019	text-only	5	9	1	0	0	0	0	0	0	0	0	0	15
sample 020	text-only	2	6	4	3	8	3	3	1	0	0	0	0	0
sample 021	text-only	17	4	6	1	1	0	0	0	0	0	0	0	1
sample 022	text-only	10	7	0	0	0	0	0	0	0	0	0	0	13
sample 023	text-only	4	8	11	4	2	0	0	0	0	0	0	0	1
sample 024	text-only	2	15	10	0	0	0	0	0	0	0	0	0	3
sample 025	text-only	1	1	3	0	0	0	0	0	0	0	0	0	25
sample 026	text-only	8	9	11	0	0	0	0	0	0	0	0	0	2
sample 027	text-only	4	11	12	3	0	0	0	0	0	0	0	0	0
sample 028	text-only	12	7	8	0	0	0	0	0	0	0	0	0	3
sample 029	text-only	2	2	25	0	0	0	0	0	0	0	0	0	1
sample 030	text-only	12	8	2	0	0	0	0	0	0	0	0	0	8
sample 031	text-only	0	0	0	0	0	0	0	0	0	0	0	0	30
sample 032	text-only	7	3	5	6	3	1	0	0	0	0	0	0	5
sample 033	text-only	17	6	2	0	0	0	0	0	0	0	0	0	5
sample 034	text-only	4	2	3	8	7	2	2	0	0	0	0	0	2
sample 035	text-only	5	11	6	4	1	2	0	0	0	0	0	0	1
sample 036	text-only	7	4	8	4	2	1	0	0	0	0	0	0	4
sample 037	text-only	17	8	3	0	0	0	0	0	0	0	0	0	2
sample 038	text-only	6	3	3	3	2	0	0	0	0	0	0	0	13
sample 039	text-only	2	2	2	0	0	0	0	0	0	0	0	0	24
sample 040	text-only	4	8	8	1	0	0	0	0	0	0	0	0	9
sample 041	text-only	7	8	9	2	1	0	0	0	0	0	0	0	3
sample 042	text-only	2	17	0	0	0	0	0	0	0	0	0	0	11
sample 043	text-only	3	11	10	4	1	0	0	0	0	0	0	0	1
sample 044	text-only	21	2	1	0	0	0	0	0	0	0	0	0	6
sample 045	text-only	3	11	5	1	1	0	0	0	0	0	0	0	9
sample 046	text-only	3	25	2	0	0	0	0	0	0	0	0	0	0
sample 047	text-only	8	11	8	3	0	0	0	0	0	0	0	0	0
sample 048	text-only	1	1	0	0	0	0	0	0	0	0	0	0	28
sample 049	text-only	4	6	6	3	11	0	0	0	0	0	0	0	0
sample 050	text-only	13	9	8	0	0	0	0	0	0	0	0	0	0

Table 4.5: The number of retweeters with each distance from the original tweeter (Text-only) (2)

		Distance												
	type	1	2	3	4	5	6	7	8	9	10	11	12	outsider
sample 051	text-only	1	1	4	11	0	0	0	0	0	0	0	0	13
sample 052	text-only	2	1	6	16	3	0	0	0	0	0	0	0	2
sample 053	text-only	17	7	3	1	0	0	0	0	0	0	0	0	2
sample 054	text-only	8	5	1	2	11	1	0	0	0	0	0	0	2
sample 055	text-only	7	4	6	4	0	0	0	0	0	0	0	0	9
sample 056	text-only	30	0	0	0	0	0	0	0	0	0	0	0	0
sample 057	text-only	6	1	2	0	0	0	0	0	0	0	0	0	21
sample 058	text-only	12	16	2	0	0	0	0	0	0	0	0	0	0
sample 059	text-only	4	5	4	6	8	1	0	0	0	0	0	0	2
sample 060	text-only	11	9	6	2	0	0	0	0	0	0	0	0	2
sample 061	text-only	7	8	2	1	1	0	0	0	0	0	0	0	11
sample 062	text-only	10	9	5	3	3	0	0	0	0	0	0	0	0
sample 063	text-only	3	4	0	0	0	0	0	0	0	0	0	0	23
sample 064	text-only	10	8	11	1	0	0	0	0	0	0	0	0	0
sample 065	text-only	5	4	5	5	4	2	0	0	0	0	0	0	5
sample 066	text-only	3	0	0	0	0	0	0	0	0	0	0	0	27
sample 067	text-only	4	2	1	0	0	0	0	0	0	0	0	0	23
sample 068	text-only	6	15	6	3	0	0	0	0	0	0	0	0	0
sample 069	text-only	6	4	0	0	0	0	0	0	0	0	0	0	20
sample 070	text-only	11	4	2	0	0	0	0	0	0	0	0	0	13
sample 071	text-only	3	0	0	0	0	0	0	0	0	0	0	0	27
sample 072	text-only	7	11	4	0	0	0	0	0	0	0	0	0	8
sample 073	text-only	1	23	2	0	0	0	0	0	0	0	0	0	4
sample 074	text-only	17	3	0	0	0	0	0	0	0	0	0	0	10
sample 075	text-only	13	1	2	0	0	0	0	0	0	0	0	0	14
sample 076	text-only	3	2	1	1	0	0	0	0	0	0	0	0	23
sample 077	text-only	5	2	1	0	0	0	0	0	0	0	0	0	22
sample 078	text-only	3	7	3	1	0	0	0	0	0	0	0	0	16
sample 079	text-only	1	0	0	0	0	0	0	0	0	0	0	0	29
sample 080	text-only	6	7	9	1	0	0	0	0	0	0	0	0	7
sample 081	text-only	5	3	5	2	0	0	0	0	0	0	0	0	15
sample 082	text-only	3	3	1	1	0	0	0	0	0	0	0	0	22
sample 083	text-only	9	3	0	0	0	0	0	0	0	0	0	0	18
sample 084	text-only	25	3	0	0	0	0	0	0	0	0	0	0	2
sample 085	text-only	4	3	3	1	1	0	0	0	0	0	0	0	18
sample 086	text-only	7	7	10	5	0	0	0	0	0	0	0	0	1
sample 087	text-only	7	6	1	2	0	0	0	0	0	0	0	0	14
sample 088	text-only	5	8	2	4	0	0	0	0	0	0	0	0	11
sample 089	text-only	5	6	4	3	3	0	0	0	0	0	0	0	9
sample 090	text-only	4	6	6	5	4	0	0	0	0	0	0	0	5
sample 091	text-only	7	2	1	0	0	0	0	0	0	0	0	0	20
sample 092	text-only	23	5	0	0	0	0	0	0	0	0	0	0	2
sample 093	text-only	6	6	2	1	0	0	0	0	0	0	0	0	15
sample 094	text-only	7	5	8	7	0	0	0	0	0	0	0	0	3
sample 095	text-only	1	1	0	0	0	0	0	0	0	0	0	0	28
sample 096	text-only	5	5	2	0	0	0	0	0	0	0	0	0	18
sample 097	text-only	4	2	6	5	4	2	2	0	0	0	0	0	5
sample 098	text-only	2	2	23	3	0	0	0	0	0	0	0	0	0
sample 099	text-only	3	0	0	0	0	0	0	0	0	0	0	0	27
sample 100	text-only	2	14	6	1	0	0	0	0	0	0	0	0	7

Table 4.6: The number of retweeters with each distance from the original tweeter (Picture) (1)

		Distance												
	type	1	2	3	4	5	6	7	8	9	10	11	12	outsider
sample 001	picture	1	0	0	0	0	0	0	0	0	0	0	0	29
sample 002	picture	1	0	0	0	0	0	0	0	0	0	0	0	29
sample 003	picture	2	4	4	3	6	6	2	1	0	0	0	0	2
sample 004	picture	5	0	0	0	0	0	0	0	0	0	0	0	25
sample 005	picture	7	5	11	1	0	0	0	0	0	0	0	0	6
sample 006	picture	2	2	5	4	8	7	1	0	0	0	0	0	1
sample 007	picture	9	11	6	0	0	0	0	0	0	0	0	0	4
sample 008	picture	0	0	0	0	0	0	0	0	0	0	0	0	30
sample 009	picture	25	3	0	0	0	0	0	0	0	0	0	0	2
sample 010	picture	3	2	6	4	1	1	3	1	1	2	2	2	2
sample 011	picture	12	12	5	1	0	0	0	0	0	0	0	0	0
sample 012	picture	1	12	7	3	2	1	1	0	0	0	0	0	3
sample 013	picture	6	5	7	4	5	3	0	0	0	0	0	0	0
sample 014	picture	3	2	0	0	0	0	0	0	0	0	0	0	25
sample 015	picture	5	4	12	3	0	0	0	0	0	0	0	0	6
sample 016	picture	12	3	2	2	1	0	0	0	0	0	0	0	10
sample 017	picture	4	4	0	0	0	0	0	0	0	0	0	0	22
sample 018	picture	16	6	5	0	0	0	0	0	0	0	0	0	3
sample 019	picture	2	8	7	5	0	0	0	0	0	0	0	0	8
sample 020	picture	5	15	6	0	0	0	0	0	0	0	0	0	4
sample 021	picture	10	6	7	3	2	0	0	0	0	0	0	0	2
sample 022	picture	1	13	7	2	0	0	0	0	0	0	0	0	7
sample 023	picture	19	5	0	0	0	0	0	0	0	0	0	0	6
sample 024	picture	0	0	0	0	0	0	0	0	0	0	0	0	30
sample 025	picture	9	2	1	0	0	0	0	0	0	0	0	0	18
sample 026	picture	2	4	0	0	0	0	0	0	0	0	0	0	24
sample 027	picture	8	6	5	0	0	0	0	0	0	0	0	0	11
sample 028	picture	2	3	1	0	0	0	0	0	0	0	0	0	24
sample 029	picture	14	11	2	0	0	0	0	0	0	0	0	0	3
sample 030	picture	2	4	3	4	4	2	5	3	2	0	0	0	1
sample 031	picture	1	0	0	0	0	0	0	0	0	0	0	0	29
sample 032	picture	4	5	11	4	1	0	0	0	0	0	0	0	5
sample 033	picture	4	2	0	0	0	0	0	0	0	0	0	0	24
sample 034	picture	3	6	3	1	2	2	2	3	1	1	2	1	3
sample 035	picture	13	11	5	0	0	0	0	0	0	0	0	0	1
sample 036	picture	6	1	0	0	0	0	0	0	0	0	0	0	23
sample 037	picture	1	3	14	4	2	3	0	0	0	0	0	0	3
sample 038	picture	1	2	3	0	0	0	0	0	0	0	0	0	24
sample 039	picture	3	2	2	3	3	8	4	1	2	0	0	0	2
sample 040	picture	8	0	0	0	0	0	0	0	0	0	0	0	22
sample 041	picture	5	19	3	0	0	0	0	0	0	0	0	0	3
sample 042	picture	4	7	7	5	4	1	1	0	0	0	0	0	1
sample 043	picture	5	5	3	0	0	0	0	0	0	0	0	0	17
sample 044	picture	6	5	5	4	6	2	1	0	0	0	0	0	1
sample 045	picture	1	2	0	0	0	0	0	0	0	0	0	0	27
sample 046	picture	7	13	2	1	2	3	0	0	0	0	0	0	2
sample 047	picture	9	2	2	1	0	0	0	0	0	0	0	0	16
sample 048	picture	19	6	0	0	0	0	0	0	0	0	0	0	5
sample 049	picture	4	4	18	2	0	0	0	0	0	0	0	0	2
sample 050	picture	4	6	10	5	0	0	0	0	0	0	0	0	5

Table 4.7: The number of retweeters with each distance from the original tweeter (Picture) (2)

		Distance												
	type	1	2	3	4	5	6	7	8	9	10	11	12	outsider
sample 051	picture	3	1	0	0	0	0	0	0	0	0	0	0	26
sample 052	picture	10	14	1	0	0	0	0	0	0	0	0	0	5
sample 053	picture	5	6	10	4	2	1	0	0	0	0	0	0	3
sample 054	picture	1	0	0	0	0	0	0	0	0	0	0	0	29
sample 055	picture	30	0	0	0	0	0	0	0	0	0	0	0	0
sample 056	picture	30	0	0	0	0	0	0	0	0	0	0	0	0
sample 057	picture	6	5	7	7	0	0	0	0	0	0	0	0	5
sample 058	picture	0	0	0	0	0	0	0	0	0	0	0	0	30
sample 059	picture	4	5	3	0	0	0	0	0	0	0	0	0	18
sample 060	picture	0	0	0	0	0	0	0	0	0	0	0	0	30
sample 061	picture	7	6	6	4	2	1	0	0	0	0	0	0	4
sample 062	picture	7	7	4	4	2	0	0	0	0	0	0	0	6
sample 063	picture	6	22	2	0	0	0	0	0	0	0	0	0	0
sample 064	picture	4	8	5	0	0	0	0	0	0	0	0	0	13
sample 065	picture	3	9	8	0	0	0	0	0	0	0	0	0	10
sample 066	picture	4	7	1	2	9	1	1	1	0	0	0	0	4
sample 067	picture	3	3	5	5	2	2	0	0	0	0	0	0	10
sample 068	picture	6	11	9	2	2	0	0	0	0	0	0	0	0
sample 069	picture	3	2	0	0	0	0	0	0	0	0	0	0	25
sample 070	picture	3	4	3	7	5	2	2	1	0	0	0	0	3
sample 071	picture	2	7	3	2	2	0	0	0	0	0	0	0	14
sample 072	picture	4	12	10	2	0	0	0	0	0	0	0	0	2
sample 073	picture	16	10	3	1	0	0	0	0	0	0	0	0	0
sample 074	picture	3	2	2	0	0	0	0	0	0	0	0	0	23
sample 075	picture	5	12	2	1	2	0	0	0	0	0	0	0	8
sample 076	picture	4	6	1	0	0	0	0	0	0	0	0	0	19
sample 077	picture	7	4	1	0	0	0	0	0	0	0	0	0	18
sample 078	picture	13	5	2	0	0	0	0	0	0	0	0	0	10
sample 079	picture	6	8	1	0	0	0	0	0	0	0	0	0	15
sample 080	picture	3	1	0	0	0	0	0	0	0	0	0	0	26
sample 081	picture	5	1	4	0	0	0	0	0	0	0	0	0	20
sample 082	picture	5	4	5	1	5	1	0	0	0	0	0	0	9
sample 083	picture	2	2	2	2	4	7	4	2	0	0	0	0	5
sample 084	picture	2	0	0	0	0	0	0	0	0	0	0	0	28
sample 085	picture	16	11	2	0	0	0	0	0	0	0	0	0	1
sample 086	picture	4	16	5	1	0	0	0	0	0	0	0	0	4
sample 087	picture	4	15	5	3	1	0	0	0	0	0	0	0	2
sample 088	picture	24	6	0	0	0	0	0	0	0	0	0	0	0
sample 089	picture	4	0	0	0	0	0	0	0	0	0	0	0	26
sample 090	picture	2	0	0	0	0	0	0	0	0	0	0	0	28
sample 091	picture	13	9	4	1	0	0	0	0	0	0	0	0	3
sample 092	picture	12	18	0	0	0	0	0	0	0	0	0	0	0
sample 093	picture	4	1	9	6	2	1	0	0	0	0	0	0	7
sample 094	picture	5	1	4	3	9	6	0	0	0	0	0	0	2
sample 095	picture	2	1	18	8	0	0	0	0	0	0	0	0	1
sample 096	picture	5	8	8	3	1	1	1	2	1	0	0	0	0
sample 097	picture	12	9	7	0	0	0	0	0	0	0	0	0	2
sample 098	picture	1	1	1	1	2	1	1	0	0	0	0	0	22
sample 099	picture	1	15	11	3	0	0	0	0	0	0	0	0	0
sample 100	picture	2	0	0	0	0	0	0	0	0	0	0	0	28

Chapter 5

Conclusion and Future work

5.1 Retweetability

We showed that the more retweeted tweets tend to get retweeted more. In terms of more specific characteristics of retweeted tweets, the tweets by a user who has many followers tends to get many retweets. However, our experiments show that including hashtags and URLs do not guarantee to get the first retweet, though they can help to get more retweets after the initial retweet. At the same time, according to our experiment, user-mention does not reduce the possibility of getting an initial retweet. On the other hand, our result imdicates that including a picture slightly helps to increase initial retweetability.

5.2 Comparison of text-only tweets and tweets with a picture

By investigation on the initial 50 retweeters' relationship with the original tweeter, the retweeters' network for tweets with picture is slightly open for the outside of tweeters' community than those for text-only tweets, thought clear statistical evidence of difference was not found at the 95% significance level. On the other hand, the distribution of the lengths of retweeters' network for picture tweets have larger variance

than that of text-only tweets. Comparing the distances of retweeters among the followers' network connected to the original tweeter, the distribution for picture tweets have a sharper curve than that for text-only tweets. In addition, the distributions of the frequencies of retweeters who are not connected to the followers' network of the original tweeter for those two types look different. For text-only tweets, frequencies of tweets monotonically decrease as the number of retweets from outsiders increases, while the number of picture tweets get retweets from outsiders at more than 90%. Overall, these results imply that a tweet including a picture can reach more people than a tweet without a picture, potentially.

5.3 Future work

The comparison between tweets with a picture and tweets without a picture might be worth doing further research. In this work, we put a limitation on the number of followers and the retweet counts in order to control factors other than the factor of including a picture or not. However, it is possible to get other results from different conditions. Other factors and different limitations are worth researching to get more plentiful insights to know the difference between a tweet with a picture and a tweet without a picture.

Bibliography

[1] Beaumont, Claudine, "Twitter users send 50 million tweets per day", The Telegraph 23 Feb. 2010 (http://www.telegraph.co.uk/technology/twitter/7297541/Twitter-users-send-50-million-tweets-per-day.html).

[2] Lee, Kathy, et al. "Twitter trending topic classification." Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on. IEEE, 2011.

[3] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[4] Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[5] Hu, Mengdie, et al. "Breaking news on twitter." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2012.

[6] Petrovic, Sasa, et al. "Can Twitter replace Newswire for breaking news?." ICWSM. 2013.

[7] Boyd, Danah, Scott Golder, and Gilad Lotan. "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter." System Sciences (HICSS), 2010 43rd Hawaii International Conference on. IEEE, 2010.

[8] Suh, Bongwon, et al. "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network." Social computing (socialcom), 2010 ieee second international conference on. IEEE, 2010.

[9] Yang, Zi, et al. "Understanding retweeting behaviors in social networks." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[10] Nagarajan, Meenakshi, Hemant Purohit, and Amit P. Sheth. "A Qualitative Examination of Topical Tweet and Retweet Practices." ICWSM. 2010.

[11] Zaman, Tauhid R., et al. "Predicting information spreading in twitter." Workshop on Computational Social Science and the Wisdom of Crowds, NIPS. Vol. 104. No. 45. 2010.

[12] Luo, Zhunchen, et al. "Who will retweet me?: finding retweeters in Twitter." Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013.

[13] Myers, Seth A., and Jure Leskovec. "The bursty dynamics of the Twitter information network." Proceedings of the 23rd international conference on World wide web. International World Wide Web Conferences Steering Committee, 2014.

[14] Dorogovtsev, Sergey N., Jose Fernando F. Mendes, and Alexander N. Samukhin. "Structure of growing networks with preferential linking." Physical Review Letters 85.21 (2000): 4633.

[15] Petrovic, Sasa, Miles Osborne, and Victor Lavrenko. "RT to Win! Predicting Message Propagation in Twitter." ICWSM. 2011.

[16] Wu, Fang, and Bernardo A. Huberman. "Novelty and collective attention." Proceedings of the National Academy of Sciences 104.45 (2007): 17599-17601.

[17] Alstott, Jeff, Ed Bullmore, and Dietmar Plenz. "powerlaw: a Python package for analysis of heavy-tailed distributions." PloS one 9.1 (2014): e85777.

[18] Twitter API Documentation, 2014 (https://dev.twitter.com/overview/documentation).

[19] Rogers, Simon, "What fuels a Tweet's engagement?", The Twitter Media Blog, 10 Mar. 2014 (https://blog.twitter.com/2014/what-fuels-a-tweets-engagement).

[20] Rice, John. Mathematical statistics and data analysis(Third Edition). Duxbury Press, 2007.