Influence of Retweets

Haruki Oh & Chanh Nguyen

Group 35

Abstract

We define a new metric for measuring influence on Twitter using retweets. Existing metric focuses on the user's ability to pass-on high value content, while our new metric focuses on the ability to generate the original high value content. Experiments show that these two metrics produce completely different rankings of the most influential users. In addition, the influences of top influential users vary more over time with the new metric than with the old one.

Introduction

Identifying influential users in social networks has been an important task for different fields such as sociology and marketing. However, studying influence has been difficult because there has not been a clear quantitative definition of influence. One philosophical definition of influence is the "ability to persuade others" (Rogers 1962). Using this definition, viral marketing targets influential users, for example, by giving out free product samples to influential users in the hope that they can influence others to purchase product. Generally, there are two parts to influencing others: first, someone has to generate persuasive messages, and second, someone has to pass along that message to others.

Influence on Twitter

Cha et al. defines influence on Twitter in several ways: indegree influence as the number of followers of a user, directly indicates the size of the audience; retweet influence as the ability to generate content with pass-along value; and mention influence, as the ability of that user to engage others in a conversation. (Cha 2010)

Retweets

We focus on retweet influence. In Twitter, retweets are characterizes as "RT @username [Original Tweet]." This means the retweeting user saw [Original Text] at @username's page, and decided to retweet it, usually with a click of a button. Cha et al. measures the retweet influence of user as the "number of retweets containing one's name" (Cha 2010).

For example: If user A tweets "ABC", user B saw this tweet interesting and retweets, it would appear to user B's follower as "RT @A ABC." If one of B's follower, user C, retweets, it would appear "RT @B @A ABC" or "RT @B ABC", depending on the implementation of Tweeter client and if the user edits. Note regardless of user's editing, "RT @B" appears in the retweet, indicating that C saw the tweet on B's page.

In this example, A's influence is 1 or 2 depending on editing (since B retweeted), and B's influence is 1. As we continue the cascade of retweets, A's influence probably do not increase, and everyone involved in the cascade will have similar influence score.

In this influence measurement, users are given credit for passing along valuable content (the original tweet) to the next user. There are several problems with this approach. First is the inconsistency with the definition of influence which is the "ability to generate content with pass-along value". Cha et al. is measuring the ability to pass-along content with value. Second, it does not give additional credit to users whose tweets are retweeted many times through other users, because the measurement only counts one-level retweet. In practical applications of measuring influence such as viral marketing, influential users should be able to generate content with high value that influences many users.

In this paper, we define a new metric and study its usefulness and difference between the metric by Cha et al. The rest of paper studies the characteristics of retweet cascades, defines the new metric, explains how to compute the new metric, compares with the existing metric, and analyze its usefulness in practical applications.

Related Work

Cha et al. presented an in-depth comparison of different measures of influence. Our work is an extension to theirs by re-defining the meaning of one of the measures—retweets—that they used. Sadikov & Martinez defined a new metric for measuring external influences on the Twitter network and provided analysis of their interplay. This paper focuses on internal influences only, and analyzes the new metric comparing to the old one.

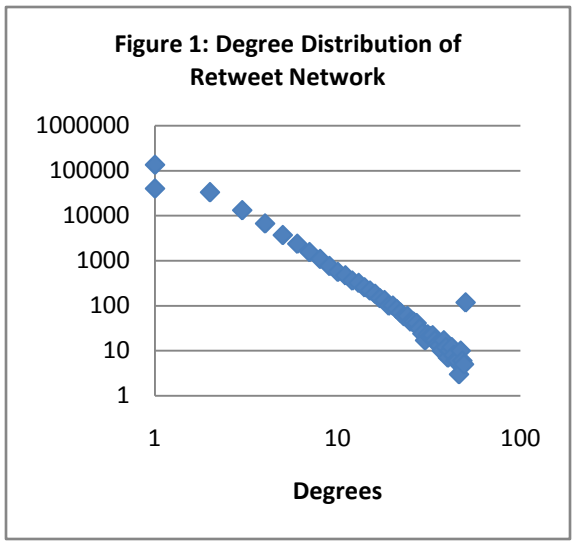
Characteristics of Retweet Cascades

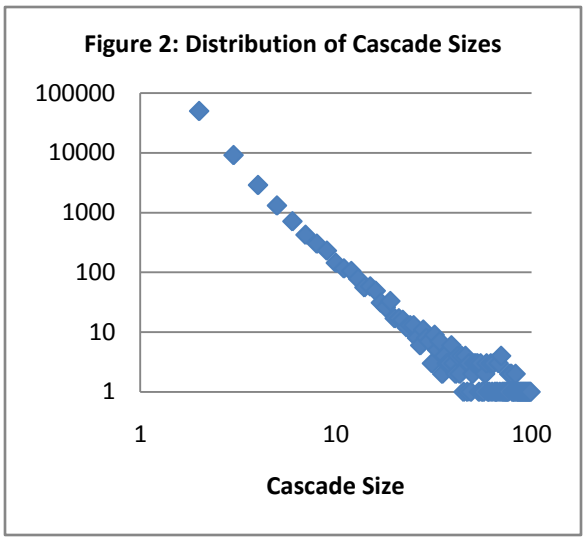
We define the Twitter cascade as follows. When there is a phrase "RT @userA content" in a tweet by userB, we construct a directed edge between nodes (userA, content) and (userB, content), indicating there was an influence from userA to userB about the content. A cascade of a content C is a connected graph whose nodes are pairs of (username, C). Retweets of different content appears in different cascades, and different cascades are not connected. A user may appear in multiple cascades, but each appearance is treated with a different node.

Twitter Data

We obtained tweets in June and July 2009 from the SNAP webpage. Data for subsequent months is no longer available. The data contained 476,553,650 tweets and of those, 71,835,017 (15%) are retweets. Each tweet is represented as a triplet of <username, time, content>.

To analyze the data and construct cascade structure, we scanned the data and put tweets into a hashmap of user and tweet (time, content, processed_or_not). Every time we see a retweet, we search through the tweets by the user where the retweet came from, we match the content by the smallest string edit distance (and where the edit distance is less than half of string size for missing data). We mark the tweet as processed. We parse the original





tweet to see if it is a tweet; if it is, repeat the process until we reach the original tweet or the earlier retweet cannot be found (eg. due to missing data). Now we have the nodes and edges from the origin of the cascade up to this point, and we can enter them into the JUNG network structure. The "processed" flag allows us to not input duplicates.

Figure 1 shows the degree distribution of retweet cascades in a log-log plot. The figure shows that the degree follows a power-law distribution and that most of the cascades are either chains or very thin trees.

Figure 2 shows the size distribution of retweet cascades. Most cascades are a few nodes, but there are a few large cascades as well. These two figures show that using the old metric defined by Cha et al., the original creator of the high-value content does not get more credit for its content cascading far away. For example users who tweet contents that created large cascades may be receiving the same credit as if the cascade size were a few nodes.

Redefining Retweet Influence

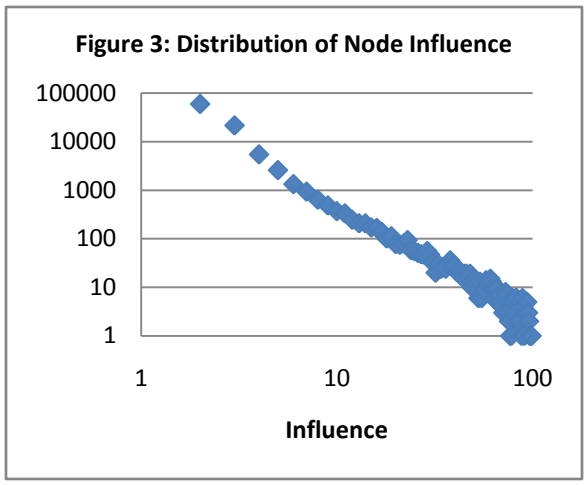
We now define a new metric for measuring retweet influence: influence of user X is the total size of all cascades where user X is the source. That is, a user has an influence only if the user generates an original content. A user does not have any influence by passing along other tweets.

Given the graph structure of cascades, finding the source is just finding the node where there is no incoming edge. The user of that node has an influence of the cascade size.

Figure 3 shows the distribution of node influences. As expected, there are many users who are not influential, and there are a few who are very influential.

Rank Differences between Metrics

Since we are interested in finding the most influential users rather than the exact value of influence measurement, it is meaningful to compare the rankings produced by two the metrics. We use the two metrics to compute influence for each user, and rank them according to who are the most influential users. To quantify how a user's rank change between the two



metrics, we used Spearman's rank correlation coefficient to determine if the two rankings are associated. If two metrics produce similar rankings, then they will have a high correlation.

Spearman's rank correlation coefficient is defined as:

$$\rho = 1 - \frac{6\sum(x_i - y_i)^2}{N^3 - N}$$

Where x_i is the rank of user i using metric x and y_i is the rank of the same user using metric y. N is the number of users in the ranking.

Because we are only interested in rankings of top influential users, we will compare the top 1% and top 10% influential users in the new metric. Any user who are not the top 1% and top 10% are removed from ranking using the old metric, and the remaining users in the ranking is compared against.

Top 1%	Top 10%
-0.204	-0.303

Table 1: Rank Correlation Coefficient between rankings using our metric vs. metric defined by Cha et al.

	Top 1%	Top 10%
New Metric	0.31	0.49
Old Metric	0.4	0.86

Table 2: Rank Correlation Coefficient between rankings of two subsequent months using our metric vs. metric defined by Cha et al.

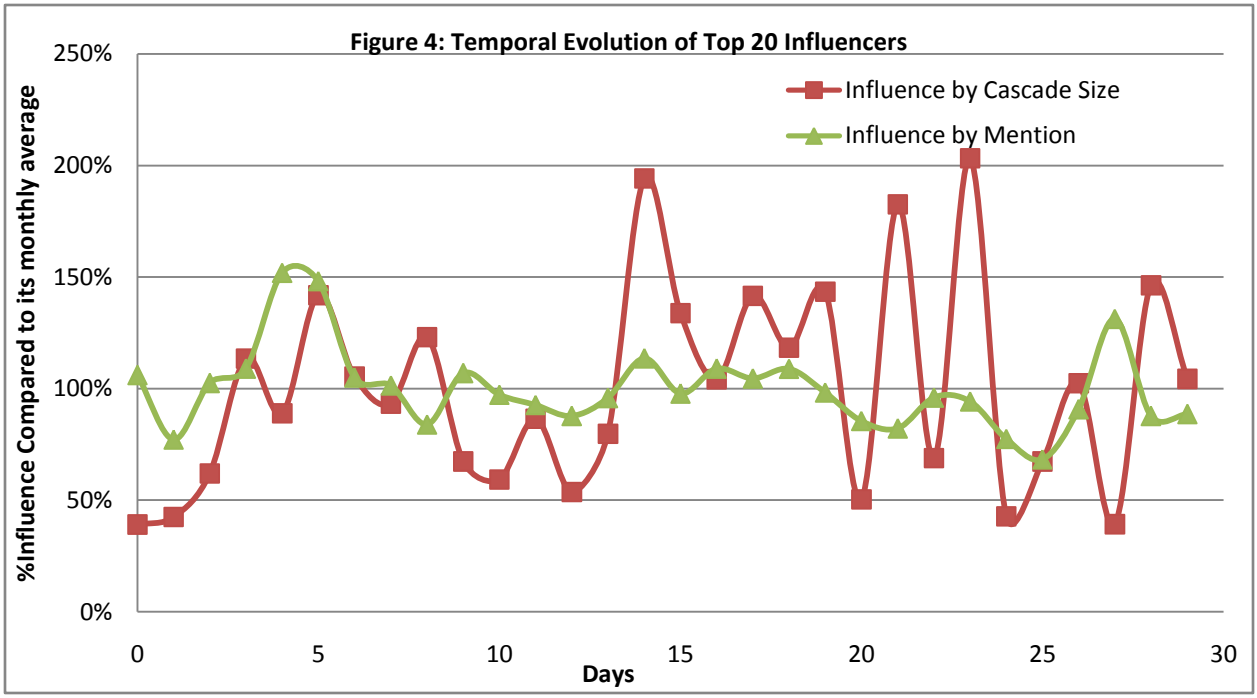
Table 1 shows that the two ranking metrics are not correlated (in fact, slightly negatively correlated). This means that being highly influential according to one metric does not mean the user is highly influential according to another metric.

Maintaining Influential Status

Identifying influential users is only useful if they remain influential. Therefore we analyze the change in the influence of users over time.

Figure 4 shows the % change in the total influence by the top 20 influential users over time using different metric. Since the two metrics produce different measurement in absolute scale, and since we are only interested in the change of influences, we plot the % of its average value.

The influence of top 20 influencers calculated with our metric varies significantly more than the influence of top 20 influencers calculated with the old metric. This is expected since the metric focuses on the ability of a user to generate content of high value. It is harder to sustain



generating high value content than just finding and retweeting high value content.

In order to see the change in ranking itself, we compute the ranking correlation between two months. Table 2 shows how two rankings using the same metric on tweets in June 2009 and July 2009 are correlated. The rankings generated using the old metric is more closely correlated to each other than the rankings generated using the new metric. This is probably an indicator that using the old metric (by Cha et al.) produces rankings that is more stable over time.

Conclusion and Future Work

We believe that influences and rankings calculated using our new metric varies more because the metric focuses on the user's ability to generate interesting content, which is unlikely to happen as frequently. In addition, Cha et al. argued that influence is gained through "concerted effort such as limiting tweets to a single topic." This suggests that high influence users are likely to tweet less often.

One weakness of our metric is that it does not give credit to passing interesting topics. While we still believe generating interesting content is an important aspect of being "influential", a new metric could combine the user's ability to pass along interesting content.

References

Sadikov, E. & Martinez, M. “Information Propagation on Twitter.” CS322 Project Report 2009.

M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In Proc. ICWSM, 2010.

Rogers, E. M. Diffusion of Innovations. Free Press. 1962.