‘Me Too’ 2.0: An Analysis of Viral Retweets on the Twittersphere

Rio Akasaka
Stanford University
Department of Computer Science
rio@cs.stanford.edu

Patrick Grafe
Stanford University
Department of Computer Science
pgrafe@stanford.edu

Makoto Kondo
Stanford University
Management Science & Engineering
makondo@stanford.edu

ABSTRACT

We propose an analysis of retweets and their propagation through Twitter. A retweet allows Twitter users to rebroadcast specific tweets of interest by incorporating all or part of the original tweet into their own. We hypothesize that retweets play a role in introducing new connections to a network, so that individuals who do not follow an individual may see retweets of their tweets and eventually become followers. We characterize the retweet trees as defined strictly by RT: @ tags. We also perform cursory analyses on social networks that give rise to popularly-retweeted individuals and compare them to popular individuals (by follower count). Analyzing the nature of the network represented by active tweeting/retweeting can give us additional insights to how users use the service and how ideas and information spread across the network.

Author Keywords

information cascade, viral network diffusion

ACM Classification Keywords

H.2.8 Database Management: Database applications Data mining (Algorithms; Experimentation)

INTRODUCTION

Nearly 200 million people worldwide currently use Twitter to broadcast messages of up to 140 characters to anyone else who is interested. The uses of tweeting are diffuse and numerous including: updating one’s current status or activity, distributing the latest news, social discussion among friends, product marketing and advertising, etc. Twitter’s “retweet” capability allows a user’s message to be rebroadcast by other users to a much wider audience. It also commonly serves as a way of saying ‘me too’, in response to a user’s tweet. The common syntax of a retweet begins with the label of “RT @”, which is followed by a username of an original tweeter, followed by the tweets contents. The following are examples of retweeets using both the “RT @” format and the “via @” format.

```
RT @fastcodesign:  Can a Christmas Makeover Solve
Sears’s Problems?  http://ow.ly/3m6Qc
Facebook and Twitter Suspend Operation
Payback Accounts http://j.mp/h9CbT8
http://techme.me/AOkv (via @techmeme)
```

Both retweets represent common methods of retweeting; however, the methods and syntax of retweets are inexact and numerous. Often times, there may be multiple RT @ sequences indicating encapsulated retweets. Users can and do alter the content of the original tweet in various ways including adding their own comments as well as abbreviating the original tweet to fit in the 140 character limitation. Lastly, retweets can occur between two users, even if they don’t formally follow each other. This indicates that the information flow on Twitter is not limited to friend and follower connections.

Following a user allows one to see the tweets and retweets of that user; however, the influence of a user’s tweet can extend farther than just to his followers. As we will see, while following a user allows for easy access and viewing of content, some retweeters don’t actually follow the person who generated the original tweet. Furthermore, depending on who retweets the original tweet, the tweet can reach a number of users orders of magnitude larger than the original user’s set of followers. We will be examining the relations between retweeting and the social graph as well as temporal aspects of retweets.

RELEVANT WORK

Dana Boyd has published relevant work on retweets [1] as part of a larger conversation that occurs on Twitter, and characterized their content and general practices adopted in retweets. Recent work by Kwak et al [4] has attempted to define tweets and retweets using trending analyses and examine how they contribute to information diffusion. Our attempts here are to expand upon that analysis, specifically on retweets. Cha et al. [2] determined that popular twitter users in terms of follower count did not correlate with retweet influence, but instead more strongly with the content of the tweets themselves. Golder and Yardi [3] analyzed triadic closure and reciprocity using a survey of active Twitter users. They showed that if attention flowed from A to X to B, then there was a high likelihood that A would be interested in following B; however, if any of those edges were reversed such as attention flowing between A and X and flowing to X from B, then there was actually a decreased likelihood of A following B, especially in the presence of a reciprocal relationship. They suggested that in the presence of a reciprocal relationship, the decreased likelihood of triadic closure may be due to differences in status. Lastly, Romero and Kleiberg [5] proposed models based on communities and preferential attachment to

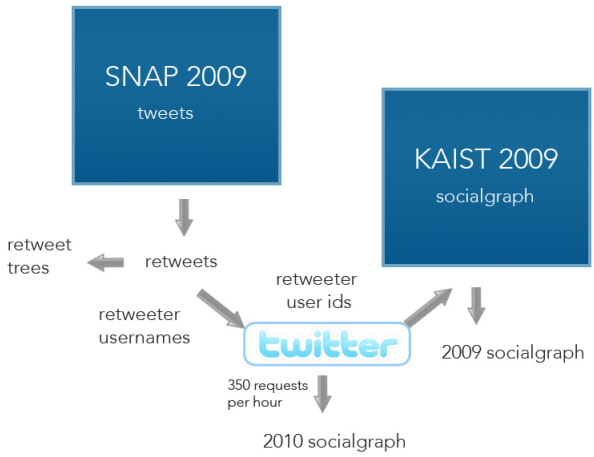

Figure 1. Data retrieval and analysis workflow using KAIST and SNAP data in order to build social graphs and retweet trees

explain the formation of directed closure on Twitter.

METHODOLOGY

Data collection

We used the Stanford Network Analysis Platform (SNAP) data set [7] which includes 476 million tweets from the latter half of 2009, as well as the KAIST social graph [4] compiled in 2009 containing 41.7 million friend/follower relationships. We generate social graphs based on retweet streams, or the collection of retweets spawned from the same original tweet. In order to map the individual retweeters to the social graphs from 2009, we map user names in SNAP to user IDs in KAIST using Twitter API calls. Figure 1 shows the complete workflow for our analyses.

Levenshtein Distance

Generating accurate retweet streams from static data presents an interesting challenge because of the variety in form a retweet can take:

```
@noukieluv: God I love RevRun! RT @ RevRunWisdom:
True love is like a good pair of socks,, it takes 2
and theyve gotta match :-) luv yall.. gud nite
@sdenise: RT @ RevRunWisdom: True love is like a
good pair of socks,, it takes 2 and theyve gotta match
```

In order to handle these, Levenshtein distance is used to denote similarity. Retweets of similar length were compared using a simple Levenshtein distance and an empirically determined threshold. Retweets of a different length required a more complicated comparison because users often add their own comments (including `RT` and `via` tags) to the beginning or end of a retweet. Thus we perform a Levenshtein distance comparison between the shorter retweet and the beginning and end of the longer retweet. If either of these comparisons result in an almost zero Levenshtein distance, then we assume the retweets are part of the same retweet stream. If neither prove to be identical, then we apply a loose empirically derived Levenshtein distance to the entire retweets in case the user added or changed content in the beginning, end, or even middle of a retweet.

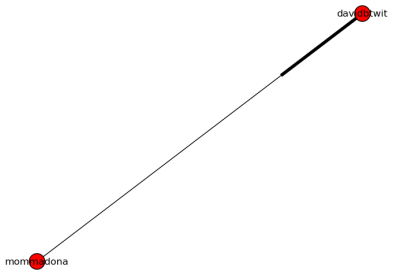

Figure 2. A retweet tree of depth one

RESULTS

There were 56,838,024 tweets in November 2009 and 9,163,490 retweets (16.12%). To ensure both depth and breadth of analysis a representative subset of that data is employed.

- 20 most popularly retweeted *(popular)*
- 20 most followed *(mostfollowed)*
- 20 random celebrities *(random)*

This is motivated by the idea of status in which we hypothesize that popularly-retweeted individuals have high status that is similar to but not identical to individuals who are most followed. Furthermore, we attempt to highlight differences between individuals who are popularly followed and those who are popularly retweeted.

Retweet Trees

To analyze information propagation through Twitter, we generate retweet trees that characterize cascading behavior in retweets based on the `RT @username` syntax. These trees are separate from friend/follower relationships, which are discussed later. The presence of multiple `RT @` or `via` provide a strong definition of retweet path. Approximately 80% of retweets we examined were one level deep, meaning a single user directly retweeted the celebrity as in Figure 2.

The following is an example of an encapsulated retweet.

```
RT: @LuvMEorHATEonME RT @RevRunWisdom Ladies Gentlemen::
It is BETTER to be with no one then to be with the
wrong one.. real tlk
```

Encapsulated retweets demonstrate some very interesting patterns that occur in retweet propagation. One of the most common patterns occurs when an influential individual retweets another popular Twitter user. The following examples demonstrate this pattern clearly.

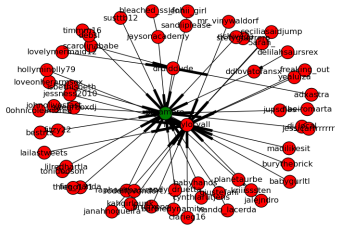

Figure 3. A sample retweet tree from a popularly-retweeted individiual (Paramore @paramore, in green), with many retweets coming through Paramore's rhythm guitarist (@itstayloryall).

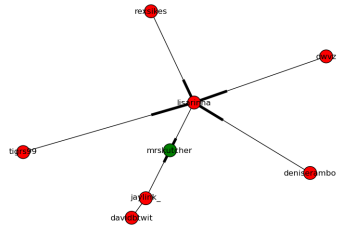

Figure 4. A sample retweet tree from a most-followers (Demi Moore @mrskutcher, in green). Here again we note the presence of an influential individual, @lisarinna, who is also an actress.

Figure 5 is a graph of frequency of retweet counts for the retweets analyzed. All of the subsets showed similar tendencies - the graphs are shifted based on the total number of retweets for the subset.

A power law fit to the distribution provides x_{min}= 1 and scaling parameter α = 1.5. This is slightly out of the expected bounds for normal power law distributions on empirical data, which may be explained due to a relatively small n.

Triadic Closure

Undirected triadic closure occurs when an edge connects two nodes that have a common neighbor [5]. For directed networks closure is expected between two nodes if there is a neighbor who is part of a path between the two, as shown in Figure 6.

One of the main points of investigation is the differences between retweeting and following. We examine the most popularly retweeted tweets from celebrities and note that a surprising number of individuals retweet those they actually do not follow. Furthermore, many of the users who were not following have become followers within the year following. This indicates a strong correlation between retweeting and triadic closure. Table 3 displays a sample of some retweets demonstrating this correlation.

The overall increase in follower count is notable in light of

Table 1. Retweet tree depth distribution in percentages

Tree depth	popular	mostfollowers	random
1	93.40	87.17	91.07
2	6.11	11.94	8.34
3	0.45	0.79	0.59
4	0.031	0.079	0
5	0.0044	0.022	0
6	0.00161	0.0055	0
7	0.0004	0	0

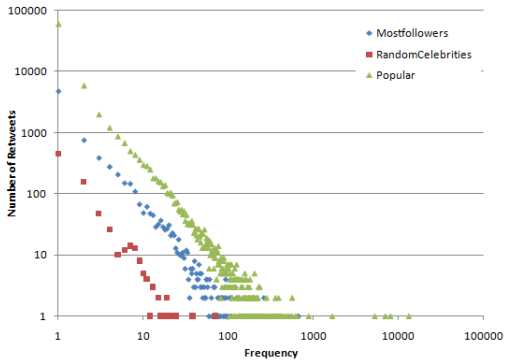

Figure 5. A graph of retweet count against frequency

the fact that there are individuals who choose not to follow. As an example, 9 individuals who retweeted one of KimKardashian's tweets and were not followers of her eventually became followers. At the same time, 6 of those who were followers and retweeted in 2009 had unfollowed by 2010.

The fact that some users retweet but do not follow the original tweeter demonstrates a couple of scenarios. Many users may view a twitter channel periodically, but choose not to follow it for some reason (some users limit the number of users they choose to follow). Retweeters also may observe and retweet tweets indirectly through individuals they follow that have retweeted the original tweeters.

Network properties

Some characteristics of the social graph (friends/followers) of the retweeters for a few select celebrities (excluding the celebrity themselves) are explored. The results that follow

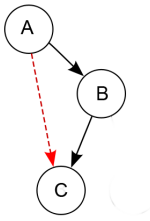

Figure 6. Directed triadic closure. We expect a tie between A and C if there is a two-step path to it through B.

Table 2. Power-law fit for retweet frequency

	N	x_{min}	α	log-likelihood
mostfollowers	22	1	1.5	-860.71
random	113	1	1.5	-364.47
popular	269	1	1.5	-74.884

Table 3. Evidence of closure in retweets

	IDs collected	Following in 2009	Following in 2010
kimkardashian	39	23	26
luansantanaevc	6	1	5
nytimes	70	67	46
revrunwisdom	440	102	270
shakira	29	29	24
techcrunch	176	61	116
thenewyorkpost	3	0	0

indicate very weakly-connected graphs with short distances found between users if a path is found at all.

Table 4. Social graph for retweeters (excluding original tweet author)

	Nodes	Edges	Transitivity	Average shortest path length
kimkardashian	2602	3512	0	2.71
luansantanaevc	70	77	0	0.66
shakira	2614	3220	0	2.23
nytimes	11491	15655	0.00005	3.33
thenewyorkpost	7893	8867	0	0.982
techcrunch	45128	79703	0.000007	3.39

K-Core Decomposition of Retweet Social Graphs

In order to better understand the characteristics of the social network of retweeters, we examined the social graph of retweet trees including the users that retweeted along with their friends and followers. The dropoff in the size of K-cores between 1 and 2 indicates that we are looking at very weakly connected graphs. This indicates that the retweeters do not represent a tightly knit community which is to be expected for celebrities and news sources. Interestingly though there are some very small communities within these social graphs, especially for TechCrunch as we see that there are some non-zero number of cores of degree 10. This is not surprising as some users, in this case tech enthusiasts, may exhibit clustering due to shared interests.

Degree Distribution

Figure 9 shows the degree distribution of the social graph of all retweeters along with their friends and followers. Given the results from K-core decomposition, it is not surprising to see the majority of nodes in the graph being of degree 1 or 2. The frequency of node degrees drop off considerably after one and two; however, the most popularly retweeted individuals represented here, RevRunWisdom and TechCrunch, have several retweeters with degrees in the thousands and even tens of thousands. The widespread popularity of these

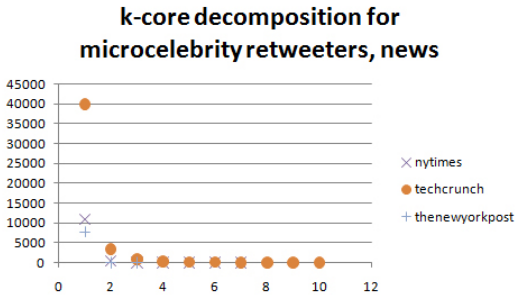

Figure 7. K-core decomposition for retweeters of news tweets

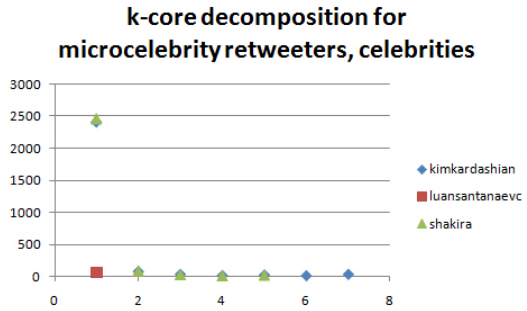

Figure 8. K-core decomposition for retweeters of celebrities

users' tweets can likely be at least partially attributed to the large number of minor celebrities rebroadcasting their tweets.

Lifetime of Retweets

Intuition suggests that retweets occur shortly after the original tweet and frequency diminishes rapidly over time. We analyzed across our sets of celebrities, the average lifetime of their retweet trees as well as the average time between the first and last retweet in a retweet cascade as indicated by encapsulated retweets. Our results indicated that the lifetime of the retweet tree spans on average 1 day for most celebrities, but closer to three days for the most retweeted users. Interestingly, the time between the first and last tweets in a retweet cascade are on the order of hours rather than days. Table 5 shows our results.

Table 5. Lifetime of Retweets and Retweet Cascades

	popular	mostfollowers	random
Average time between first and and last retweet in a cascade	6.41 hours	4.93 hours	1.14 hours
Average time between first and and last retweet	62.24 hours	29.72 hours	19.90 hours

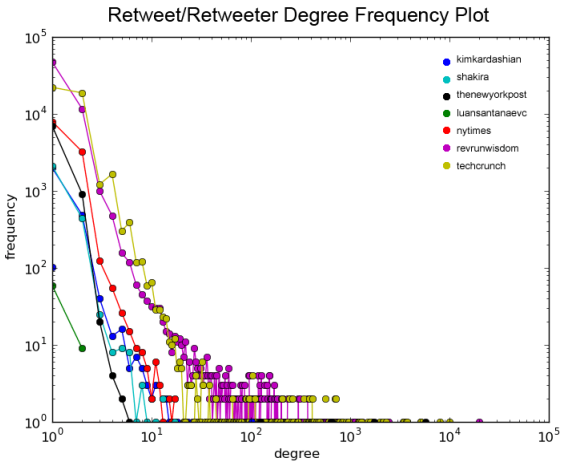

Figure 9. A graph of retweeter degree distribution for seven micro-celebrities

CONCLUSIONS

We have observed a number of interesting characteristics of retweets, specifically in the context of celebrities. Qualitatively, frequently retweeted tweets tend to fall under a number of categories: charities or causes, product promotion, spreading breaking news, and even just words of wisdom. Retweets typically have longevity on the order of 1 to 3 days, but retweet cascades are relatively shorter in lifespan on the order of hours. Furthermore, retweet count frequency follows a power-law distribution with a significant majority of retweet trees being relatively small and far fewer viral retweets. We also determined that influential individuals and celebrities often help spread retweets to a much larger audience of users. As a result of our analysis covering celebrities and news agencies, the graph of retweeters does not appear to demonstrate much clustering. Future analysis may show different results for less popular more localized users. Finally an analysis of individuals who retweet but did not follow an individual has shown that often they eventually do follow, indicating a positive influence on triadic closure formation. This confirms our intuitions that retweets not only serve to spread information, but also introduce users to individuals they are interested in following.

FUTURE WORK

These analyses are as yet fairly superficial and much remains to be analysed with this data. Finding retweet streams and generating the social graph is a very time consuming process and our analysis was thus limited. In the future, we would like to expand our analysis to a larger and more comprehensive survey of the Twittersphere. Ideally the results can be improved by obtaining whitelisting from Twitter, so that retweets streams can be more robustly analyzed in real time.

One aspect we'd like to explore in the future is the effect of geographic location on retweet trees, expanding upon the topic of "self presentation" as highlighted by Golder and Yardi. Given existing semantic analysis [6] on what tweet terms encourage retweeting, expanding on retweet structure research would help determine the extent to which Twitter can be used for marketing and promotion. We have seen national and worldwide musicians effectively use Twitter to announce new releases, and their loyal followers proceeded to spread the news far and wide through retweeting. Can retweets also be used to effectively market products at the local level?

REFERENCES

1. D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

2. M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

3. S. A. Golder and S. Yardi. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *The Second IEEE International Conference on Social Computing (SocialCom2010)*, 2010.

4. H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

5. D. M. Romero and J. M. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *ICWSM*, 2010.

6. B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. *Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust, 2010 IEEE International Conference on*, 0:177–184, 2010.

7. J. Yang and J. Leskovec. Temporal variation in online media. In *ACM International Conference on Web Search and Data Mining (WSDM '11)*, 2011.