Information diffussion in Social Media

Andrew Dotey, Hassan Rom, Carmen Vaca
Final Project
CS224W
Stanford University

December 11th, 2011

1 Introduction

The proliferation of social networks in our daily lives has given rise to a plethora of opportunities to explore the flow of information and ideas throughout a network. In a network, information flows in the form of cascades. Analyzing cascade virality is an important, yet challenging, problem in social network analysis. An understanding of the properties of a viral cascade is crucial in order to understand exactly how information and ideas spread, as well as the differences in things that spread vs. things that don't. It can be used for tasks such as determining the most influential individuals within a network, which translates directly to various practical applications, such as maximizing marketing effectiveness or influencing public opinion.

However, cascade virality is a complex phenomenon with many different aspects, making it difficult to model and analyze. When analyzing the propagation of information throughout a network, there are many features of the network one must take into account. For example, consider structural and temporal network features: these two network features alone have a wide-reaching and profound impact on how viral a cascade is or is not, and influence how viral it can or can not be. Structural features of a network include its number of nodes and edges, the amount of clustering, the length of various paths, and its underlying distribution. Temporal features include the speed at which information propagates, the time to end for cascades, and the lifetime of cascades.

Other aspects one must consider include the probability of the information spreading from one node to the next, the amount by which it spreads at each level of propagation, the amount of nodes reached in the network, and what kind of environments stop or assist in the spread of cascades. Thus, depending on the context of the underlying network, a cascade's virality can be defined in many ways. Acquiring a full, coherent understanding and analysis of cascade virality appears daunting when faced with the task of accounting for these many intricacies.

With this problem in mind, our goal is to understand and analyze cascade virality and how information spreads in a social network. Specifically, we are interested in identifying the features that are most common for viral cascades and performing various analyses on them. In order to accomplish this, we have chosen to analyze the Twitter network. Twitter was a natural choice for the network to analyze because its sole purpose is spreading information, and the user-follower and retweet relationships give the potential for viral cascades over a massive number of nodes in the network.

Outline The aim of our project was to identify and analyze features that capture various structural and temporal information of cascades in networks, as well as general network statistics that are most useful for identifying a cascade's virality. The remainder of this paper is organized as follows: Section 2 presents the related work. Section 3 describes the datasets we used and

how we detected cascades. Section 4 outlines our structural and temporal analyses, as well as additional analyses we performed.

2 Related work

Much work has already been done on the analysis of viral cascades in social networks and the identification of some of their key features. [12] analyzes viral cascades in the Flickr social network. Here, the network consists of Flickr users, their friends, and photos. If one of the user's friends marks the photo as a "favorite", information is considered to have disseminated from one user to another. Using this network, Yu et.al discovered topological patterns in information propagation cascades. They explored how long it takes for content to spread to the first user and how long it takes to spread to subsequent users, giving intuition as to how fast content spreads. They also found that most cascades are in the shape of stars or are "shallow bursty cascades." Impressively, they were able to model the cascade network as the spread of a virus. The analyses done here were detailed and interesting, but they only scratch the surface on all the possible analysis and feature identification that can be done on a rich social network.

Another aspect of cascade virality to consider is influence. Individuals seem to naturally share the valuable things they learn, passing that information to their closest neighbors. Online social networks reduce the amount of effort needed to make interesting information public. However, this information doesn't spread through the network following the same pattern for every individual. There are some nodes in the network with a high influence over other nodes, enabling them to enact an effective information spreading mechanism that can reach a large number of users. This influence has been studied because of its importance in scenarios such as social tagging [1, 4] and viral marketing [6, 9]. Several studies have searched for a model to predict the influence of a given node. Intuition might tell us that nodes with a higher degree are more influential, but [11] found that users with the highest follower count are, in fact, not the most influential.

A broad set of fields stand to benefit from analyzing information cascades. In politics, it is possible to propose models for the formation of citizens' opinions about current issues [3]. For example, [8] analyzes the role of social media in massive protests. Sociologists can infer information about social behavior by analyzing cascade properties, as shown by [2], who presents a study on the role of information cascades in the adoption of the term "anchor baby." This term is generally considered racist and unpleasant, but nevertheless became popular, even in the blogosphere.

Online platforms are an excellent source of data for the study of information cascades. For example, studies have been conducted for platforms such as Digg [5], Blogs [7] and Facebook [10].

3 Datasets Preparation

The Twitter data we are using is taken from two different sources. The first dataset is a collection of tweets from the period between June and December of 2009. For each tweet in the dataset, information is available for the text of the tweet, the screen name of the user who made the tweet, and the time of the tweet. The data contains a total of 476 million tweets from 17 million unique users. The total size of the data in compressed form is 25.7 GB. To make it easier to work with, we ran a mapreduce over the dataset, grouped all the tweets by their user, and exported the data into our own format.

The second dataset we are using is from [4]. The dataset consists of two files. The first file contains a list of edges for a directed graph of the followers network in Twitter. The dataset consists of 40 million users with at least one follower and 1.5 billion directed followers' edges. The second file contains mappings from numeric user-ids to screen names. The second file is

used to create a mapping between the users in the first source with the users in the second source. We were able to join a total of 5.9 million users between the two datasets. For each of these users, we store the list of their followers in our own format.

From the exported data, we built two networks, which we then used in our analysis. The first network is a followers network in which the nodes are Twitter users, and an edge from node a to node b means a follows b. The second network we built is a collection of subgraphs, each representing a retweet cascade. In this network, the nodes are twitter users, and an edge from node a to node b means that node a retweeted from node b. For each edge, we store the time when node a and node b first tweeted, as well as the tweet's fingerprint. A tweet's fingerprint uniquely identifies an instance of a cascade.

To compute the cascades, we ran a mapreduce over the users' tweets and followers data and computed for each tweet of the user the fingerpring of the tweet. The fingerprint is computed by first stripping out the `RT` and the `@screenname` tokens in the tweet, and then computing a hash over the stripped tweet. Then for each tweet, we output to the reducer a key and value pair, where the key is the fingerprint and the value is the user's id, the user's tweet, the time of the tweet, and a list of the user's followers. If there are multiple tweets with the same fingerprint from the same user, we only output the earliest tweet. Algorithm 1 illustrates the mapper's algorithm.

Algorithm 1: The mapper of the mapreduce that computes retweet cascades.

```
// T_i is a set of user i's tweet.
// F_i is a set of user i's followers.
Input: T_i, F_i
// M is a map of tweet fingerprint to a set of (i, F_i).
Output: M
foreach t ∈ T_i do
    // ComputeFingerprint firsts strips off the 'RT' and '@screename'
    // tokens from the tweet text and then computes the hash over
    // the stripped tweet.
    f = ComputeFingerprint(t);
    M(f) = M(f) ∪ {(i, F_i)};
end
```

In the reducer, we then get a list of tweets (along with the user-id and followers of the user who made the tweet) with the same fingerprint. To compute a list of followers who retweeted, we first compute the set of user-ids who have the same tweet fingerprint, and then intersect that set with each of the user's followers. We also require that the follower's tweet came after the followee's tweet. At the end of the reducer phase, we output, for each fingerprint, a list of user-ids and his/her followers who retweeted the user's tweet. Each record outputted in the reducer phase is a retweet cascade. Algorithm 2 illustrates the reducer's algorithm. The number of cascades generated by our method is 1.2 million from 1.4 million unique users retweeting a total of 3 million times.

To evaluate the quality of the detected cascades produced by our approach, we randomly sampled 100 detected cascades and evaluated how likely each cascade is a retweet cascade. We also looked for any errors in how we generated the cascades. In our sample, we identified 20 cascades (20%) that either were probably not a valid retweet cascade, or have errors associated with them.

One common error we found in our approach was incorrect attribution of who the user was retweeting from. Since our approach does not take into account the screen names in the tweets when attributing sources of retweets, we simply assume that if: 1) two users A and B share the same tweet fingerprint, 2) B follows A, and 3) B's tweet came after A's, then B must have

Algorithm 2: The reducer of the mapreduce that computes the retweet cascades.

```
// f is a fingerprint of a tweet.
// L is a list of (i, F_i) that share the same fingerprint tweet f.
Input: f, L
// C(f) is a retweet cascade of tweets with fingerprint f.
Output: C(f)
// S is a set of user ids whose tweets share the same fingerprint.
S := ∅;
foreach i, F_i ∈ L do
    S = S ∪ {i};
end
// R is a set of edges from j to i such that j retweeted from user i.
R := ∅;
foreach i, F_i ∈ L do
    // z(i) is the time when i tweeted.
    R = R ∪ {(j, i)|j ∈ F_i ∩ S, z(j) > z(i)};
end
C(f) = (S, R);
```

retweeted A's tweet. While this assumption does correctly detect retweet cascades most of the time, there are some interesting false positives, such as shown in Appendix B.1.

Since our algorithm only uses the network of followers to infer the sources of a retweet, our algorithm may also incorrectly attribute a retweet of a user as being from multiple users he is following, when in actuality, the user only retweeted once from only one of the many users he is following. An example of such a case is listed in Appendix B.2.

Finally, another class of problems we found with our algorithm is that it can sometimes detect unrelated tweets as a retweet cascade. The unrelated tweets might be common phrases or marketing tweets, which are not necessarily retweets in the traditional sense. An example of such a case can be found in Appendix B.3.

4 Analysis

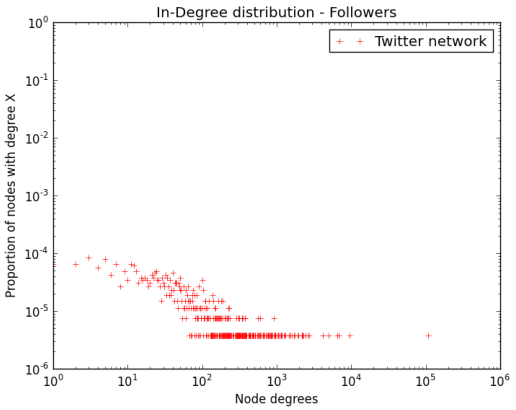
We used the datasets prepared as described in Section 3 to carry out various analyses. Specifically, we utilized the followers network, and the set of detected cascade networks. The analyses of the corresponding graphs can be broadly organized into three overarching categories: structural analyses, temporal analyses, and additional analyses. The applied methodology, as well as the numerical and qualitative results of such analyses, are presented in the following subsections.

4.1 Structural analysis

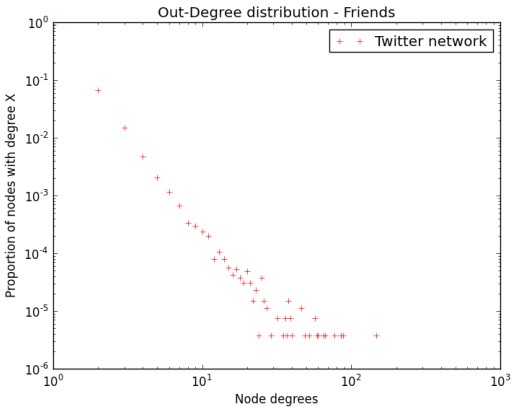
We utilized structural analysis techniques to measure properties of the followers and retweet networks in order to help us to draw conclusions about the patterns found when humans interact in the Twitter platform. We first measure connectivity using in-degree, out-degree, and clustering coefficient. We then analyze the properties of a cascade of information, such as its size and depth. We conclude this section by elaborating on the relationship between a user's status in the network and his ability to spread information. The structural analyses conducted and their corresponding results are as follows:

1. Degree distribution

   Figures 1(a) and 1(b) show the in-degree and the out-degree distributions, respectively, for the followers network. As expected, in Twitter, many people have very few followers

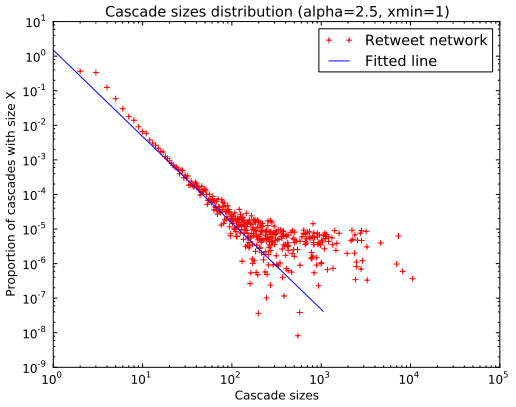


(a) In-degree distribution in the followers network

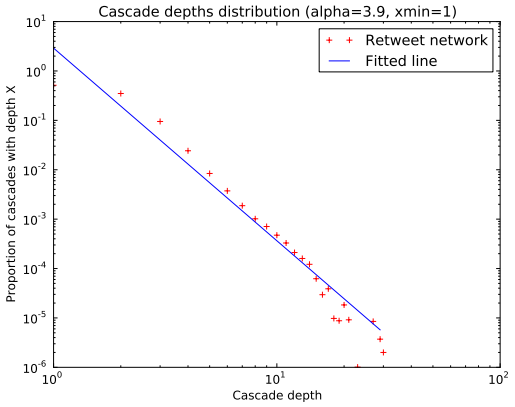


(b) Out-degree distribution in the followers network

Figure 1: In-degree and out-degree distribution of the followers network



(a) Distribution of cascade sizes



(b) Distribution of cascade depths

Figure 2: Distribution of cascade sizes and depths

and follow very few people. Empirically, the out-degree distribution appears to follow a power-law distribution.

2. Clustering coefficient

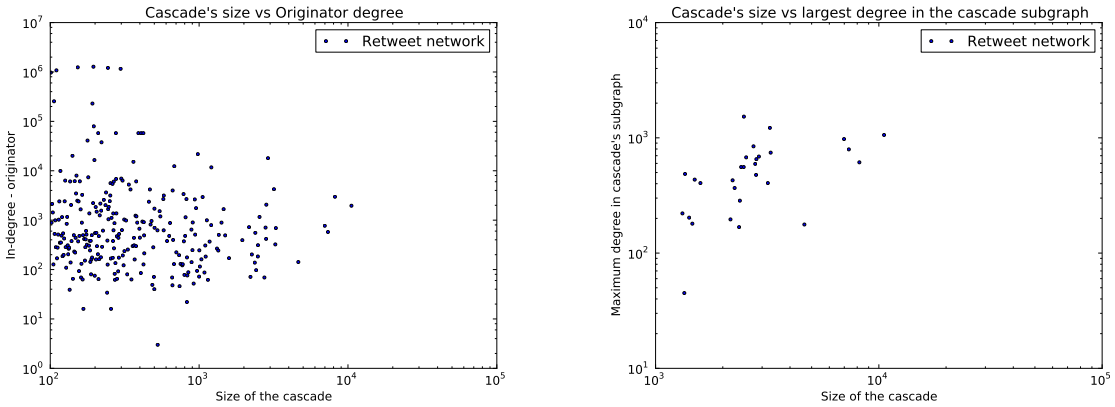
   We have run our analysis algorithm and found an average clustering coefficient of 0.0083. This is a very low measure for this parameter, which reflects the fact that, in Twitter, a given person might have followers that are not connected among themselves.

3. Cascade sizes and depths

   Figure 2(a) shows the Cascade sizes distribution and a fitted line, showing that it fits a power-law distribution well. We can conclude from the plot that most of the retweet cascades involve very few numbers of users. Figure 2(b) shows that the great majority of cascades are very shallow in depth, i.e., users will tend to retweet the original author of a post.

4. Relationship between a person's status within the community and their role in the diffusion of information

   We have attempted to understand if a user connected to many different communities is more likely to be retweeted than a user connected to a smaller number of communities.



(a) Cascade size vs Degree of the Cascade's originator. (b) Cascade size vs Degree of the User with larger number of retweets in the cascade.

Figure 3: Relation among cascade sizes and the status of a user in the community

Figure 3(a) shows a scatter plot of a cascade's size vs. the in-degree of the user at the root of the cascade, i.e., the user who posted the tweet at the earliest date. Only cascades of size greater than or equal to one hundred were considered, as we wanted to visualize the correlation between the user's status in the community and his ability to start a cascade that becomes viral. We ran a mapreduce job to process all the cascades in the dataset because constructing the graph for each cascade and determining its originators was computationally too expensive for the amount of data available.

The visual results in Figure 3(a) reveal that it is possible for large cascades to be started by a person with a small in-degree. We can see that even people with in-degrees around 100 can still produce cascades with sizes in the thousands. On the other hand, a cascade whose original post was by a user with a large number of followers might remain very small. People with 10,000 followers, for example, might start cascades joined by only ten or twenty other people. Similarly, [11] found that it is not always the case that a person with a higher degree will be the author of posts that propagate massively. The evolution of the cascade's size also depends on the hubs (users of large degree) that join the trend, which we address in the following paragraph.

To further measure the influence of a user in the virality of a cascade, the 30 largest cascade subgraphs were extracted, as well as the user who produced the largest number of retweets in each set. The number of followers of such a user was computed and plotted, along with the cascade size. Figure 3(b) shows the plot where we observe a direct relation between the two variables. This shows that, when users of large degree join a cascade, the cascade tends to become large. This result corresponds with our natural intuition of users of high degree having a high amount of influence.

As part of this analysis, the most popular tweets' posts were computed. It is interesting to see that there is little variation in the topics of the retrieved tweets. They primarily belong to 3 categories: how to get more Twitter followers, politics, and fighting cancer. Appendix A includes detailed information about these 30 cascades. Here are a few examples:

```
Format: cascade size, tweet
745, Sign up free and Get 400 followers a day using http://tweeteradder.com
1497, Have you tried the new and improved @mrtweet? Get great people
       recommendations with one click.http://mrtweet.com?v=20
7337, Show support for democracy in Iran add green overlay to your Twitter
```

avatar with 1-click -http://helpiranelection.com/

4.2 Temporal analysis

The cascade subgraphs were generated to include the date of each of the posted tweets. We converted the date of each tweet to epochs and determined the first and the last tweet posted to compute the lifetime of an information cascade. Using this information, we conducted the following analyses:

1. Distribution of cascade lifetimes broken down by cascade size

   Our analysis includes only the sizes for which at least 10 cascades have been found. This threshold was selected to reduce noise information in the plots, since there are a large number of cascades of small size. This frequency decreases as the size of the cascade grows. For the same reason, the plots were split into 4 different graphs for the sizes 1-10, 11- 20, 21-40, and 41-70, shown in Figure 4 a,b,c, and d, respectively. The x-axis represents the lifetime of a cascade in hours, and each line represents the distribution of lifetimes for cascades of a certain size.

   Visual analysis of the plot reveals that cascades of small sizes have a duration of less than a day. As the size of the cascade grows, the duration increases. Using the exact lifetimes for particular cases, we found, for example, that almost half of the cascades of size 51 have a duration spanning from one to six months. Figure 4 (d) supports this fact by showing a point at approximately 40% at the far-right end of the graph. In contrast, most of the cascades of size 3 have a duration of less than one hour.

2. Statistical properties of the cascades' lifetimes

   An alternative approach to analyzing the lifetime of retweets' cascades is comparing the mean and standard deviation for different sizes. Using the same threshold as before, the plot in Figure 5 shows an increasing pattern for the mean lifetime; it grows as the cascade size grows. Error bars have been used to show the standard deviation, which is very high in most cases. It is possible to find cascades of the same size that last as long as a couple of months, or as short as just a few hours. Cascades of size 51, for example, have a mean of 908.65 (around 1 month) and a standard deviation of 1299.33 minutes (around 2 months).

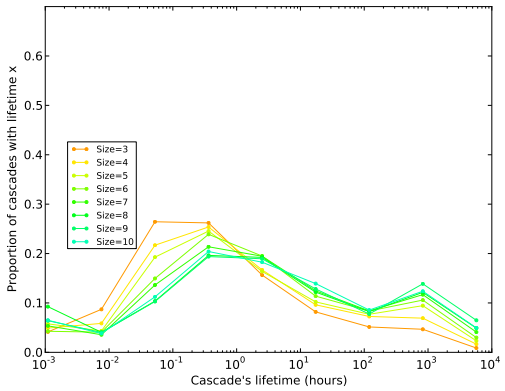
4.3 Additional analysis

The previous subsections included information about cascades of retweets in Twitter. However, we were also interested in analyzing cascades of hashtags. We can see that, in reality, cascades of hashtags form in Twitter. It is very common to see a hashtag's popularity grow and spread throughout the userbase. However, it is less clear how to detect this type of cascade than it is for retweets. It cannot be said, for example, that just because user A tweeted a hashtag before user B, there was a cascade from A to B. The difficulty lies in drawing causation. To get around this, we started with the retweet cascades from the previous sections and restricted the set to the tweets that contain a hashtag. We do our analyses on the resulting set. The analyses conducted on the hashtags retweet cascades are the following:

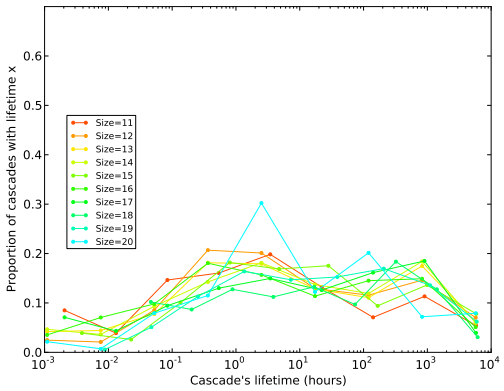
1. Distribution of hashtag retweets' size

   Figure 6 shows the hashtag retweets' size distribution. The distribution appears to follow the Power-Law distribution with $\alpha = 2.0$, which is lower than the $\alpha = 2.5$ in the overall cascade size distribution shown in Figure 2(a).

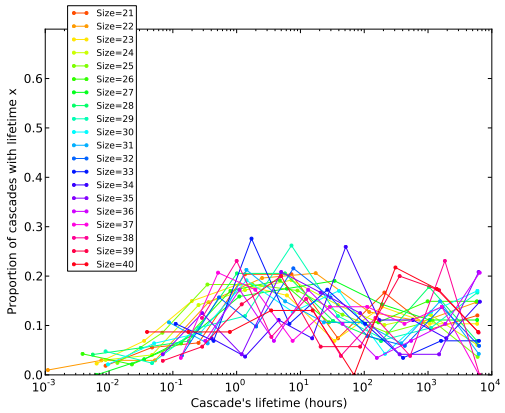
2. Distribution of the cascade sizes according to hashtags



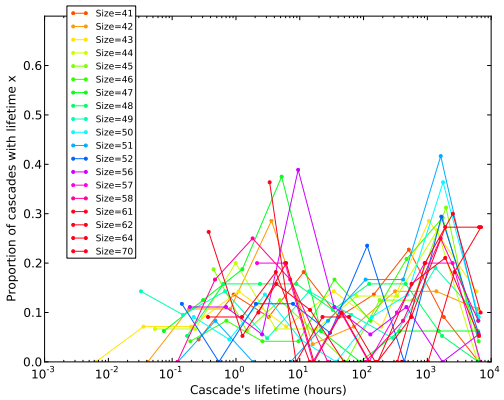
(a) Cascade sizes: 1-10



(b) Cascade sizes: 11-20



(c) Cascade sizes: 21-40



(d) Cascade sizes: 41-70

Figure 4: Cascades' lifetime distribution for different cascades' sizes

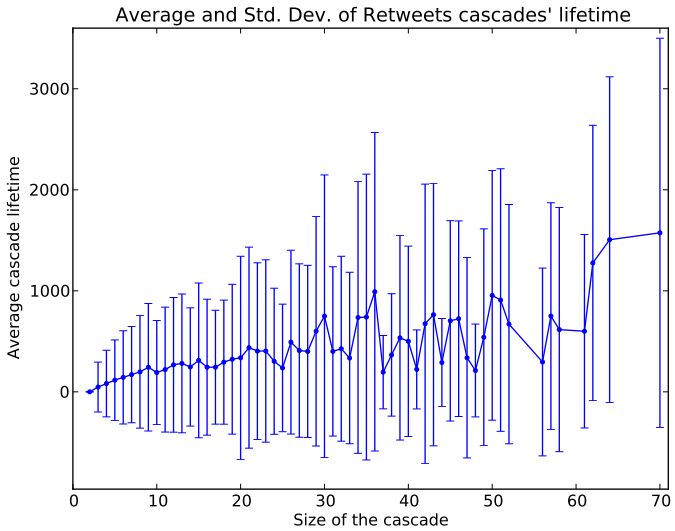


Figure 5: Average retweet cascade lifetime

Figure 7 shows the retweet size distribution of the top 10 most popular hashtags that appeared in the retweet cascades. From the figure, the hashtags' cascade size distribution

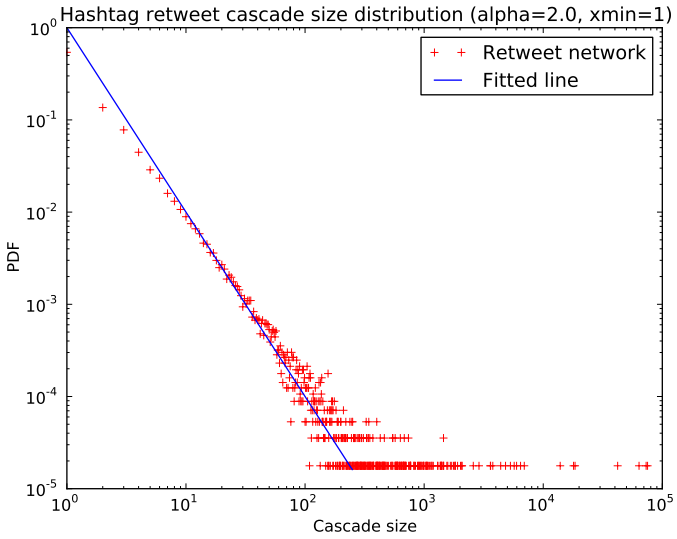


Figure 6: Distribution of hashtag retweet cascade sizes.

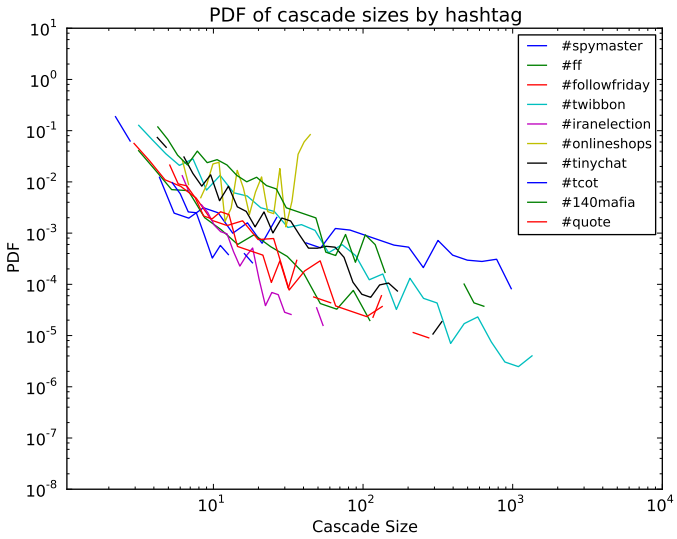


Figure 7: Distribution of cascade sizes broken down by most popular hashtags.

follows a similar distribution, except for the hashtag `#onlineshop`.

The `#spymaster` and `#140mafia` hashtags are typically used to mark a user's status message in the social game Spymaster and 140mafia. The cascades produced by the status messages in the two games are another example of a false positive in our cascade detection algorithm.

The hashtags `#followfriday` and its shortform `#ff` are used to suggest whom to follow. The tweets are typically made on Fridays. Understandably, these two hashtags are popular, and both of their cascade retweet size distributions are almost overlapping with one another.

Another interesting hashtag retweet cascade detected is `#iranelection`. The tweet data we used overlaps with the Iranian elections in June 2009. Twitter was heavily used during this time to protest against the election results.

Other hashtags that made it to the top 10 include: `#tcot`, a hashtag for following top US conservative politicians; `#twibbon`, a hashtag for twibbon.com, which is a service for users to overlay their twitter avatars with other images; and `#tinychat`, a hashtag for tinychat.com, which is a video chat application on the internet. The hashtag `#quote` is also popular in retweet cascades.

5 Conclusions

In this paper, we have developed a novel algorithm using the mapreduce programming paradigm that is capable of detecting retweet cascades over large datasets with an error rate of 20%. The classes of false positives produced by our algorithm can be easily fixed and are something that can be explored in future work.

We performed three types of analyses on the cascade data generated from our algorithm, in order to identify and analyze viral cascades: structural analyses, temporal analyses, and additional analyses.

In our structural analyses, we determined that cascades generally involve a low number of users and are shallow in depth. We also found that a large cascade can begin even from originators of low degree, though they tend to be larger when users of high degree get involved at some point of the cascade's lifetime.

Based on the results of the temporal analysis, we can confirm that most of the cascades of small size have a lifetime of a few hours. Larger cascades can last for a period of months, but we were also able to observe some large sets of retweets generated in a few hours. This leads us to conclude that the lifetime of a cascade depends more on the content and real-world context of the post itself, rather than the number of people who join the cascade.

The analysis of the most popular cascades in the retweet network showed us that there is not a direct relationship between the in-degree of the original post's author and the final size of a cascade. However, if an influencer joins the set of people retweeting, it will help increase the amount of participants in the cascade significantly.

In our additional analysis of hashtags, our algorithm was able to detect that the `#iranelection` hashtag was popular in retweet cascades, consistent with what we know was occurring in the media during that time. Another interesting finding in our hashtag retweet analysis is that the hashtag retweet cascade size distribution has a shorter tail than the overall retweet cascade size distribution.

While we have performed various analyses that help to further understand cascade virality and how information spreads in a social network, there are still many possibilities for future work. Additional analyses to explore that fall into the categories of analyses in this paper include analyzing the relationship between the number of followers and influence over a dataset, or analyzing the relationship between the total number of hashtag retweets and the number of cascades the hashtag retweets appear in. Furthermore, exploring new categories of analysis can build upon our own analyses.

References

[1] J. Huang, K.M. Thornton, and E.N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 173–178. ACM, 2010.

[2] G. Ignatow and A.T. Williams. New media and the anchor babyboom. *Journal of Computer-Mediated Communication*, 17(1):60–76, 2011.

[3] M. Kaschesky and R. Riedl. Tracing opinion-formation on political issues on the internet: A model and methodology for qualitative analysis and results. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10. IEEE, 2011.

[4] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue 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] K. Lerman and A. Galstyan. Analysis of social voting patterns on digg. In *Proceedings of the first workshop on Online social networks*, pages 7–12. ACM, 2008.

[6] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. *Advances in Knowledge Discovery and Data Mining*, pages 380–389, 2006.

[7] Jure Leskovec, Mary Mcglohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs. In *Proceedings of 7th SIAM International Conference on Data Mining (SDM)*, 2007.

[8] M. Lynch. After egypt: The limits and promise of online challenges to the authoritarian arab state. *Perspectives on Politics*, 9(02):301–310, 2011.

[9] M.R. Subramani and B. Rajagopalan. Knowledge-sharing and influence in online social networks via viral marketing. *Communications of the ACM*, 46(12):300–307, 2003.

[10] E. Sun, I. Rosenn, C. Marlow, and T. Lento. Gesundheit! modeling contagion through facebook news feed. In *Proc. of International AAAI Conference on Weblogs and Social Media*, 2009.

[11] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining*, pages 599–608. IEEE, 2010.

[12] B. Yu and H. Fei. Modeling social cascade in the flickr social network. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*, volume 7, pages 566–570. IEEE, 2009.

A Appendix - Tweets extracted from the most popular cascades

The 30 most popular retweet cascades found in our dataset, including: size, the most retweeted user (characterized by its in-degree in the retweet subgraph), the in-degree in the followers network, and finally the stripped tweet. We can observe possible false positives, such as the smiley face.

```
Cascade size, Max degree in the cascade's subgraph, Followers of the user, Tweet
================================================================================
2366 168 8406   Just changed my twitter background, check it out! Found it at
                http://www.TwitterBackgrounds.com
3182 405 68231  thanks for the RT!
2260 367 3486   I just become a member of this AWESOME site  that gets you
                TONS of followers:http://followersfree.com
2546 677 94609  #FollowFriday
1319 221 1600   WOW the best site online to gain more twitter  followers:
                http://followAdd.net
2828 653 6204   Get 400 followers a day using http://tinyurl.com/npfzt4
3250 1221 10566 Get 400 followers a day using http://www.tweeterfollow.com
8178 613 1657119  #FF
2385 285 54299  thanks for the
2745 844 14178  Sign up free and Get 400 followers a day using http://tweeteradder.com
1497 434 94609  Have you tried the new and improved @mrtweet? Get great people
                recommendations with one click.http://mrtweet.com?v=20
2213 428 9155   KAMI TIDAK TAKUT. We Are Not Afraid. Add Indonesia flag to your
                avatar with 1-click http://twcauses.com/iunite/ #indonesiaunite
2819 477 69265  #followfriday
1414 203 1645   WOW the best site online to gain more twitter followers for FREE:
                http://www.youradder.com
2902 688 89941  Thanks for the
7337 794 503675  Show support for democracy in Iran add green overlay to your Twitter
                 avatar with 1-click -http://helpiranelection.com/
10551 1059 154348
3277 744 8406   Support Yellow Ribbon for Pres. Cory, add a #twibbon to your avatar now!
                - http://bit.ly/2IsH3K
1355 486 14036  Support IndonesiaUnite, add a #twibbon to your avatar now!
                 -http://twibbon.com/join/IndonesiaUnite
2486 1528 39945 Support Lance's return to Tour and #LIVESTRONG global cancer fight.
                Add wristband to your Twitter with 1-click - http://twcauses.com/ls/
2417 558 786    I just wounded in an assassination attempt. #spymaster
                   http://bit.ly/playspy
1345 45 523     Finally found the BEST way to get tons of follow ers for FREE!
                http://www.youradder.com
6983 976 83510  Thanks for the RT!
2481 558 1719   I just become a member of this AWESOME site that gets you TONS of
                followers: http://vipfollowers.com
2795 594 1241331 lol
2165 196 47622  #ff
1588 405 1784   I just become a member of this AWESOME site that gets you TONS of
                followers: http://followersnow.com
4639 177 3299   Win a new Canon 5DMKII (or $2500 Gift Cert) from & Scott Bourne.
                Pls RT. Details here: http://bit.ly/BqU8N
1461 180 94609  :)
```

B Appendix - Examples of retweet cascades false positives

B.1 Incorrect retweet source attribution

In the example below, both users 14313400 and 21436876 retweeted from `@AlacraPulse`, but our algorithm detected that user 21436876 retweeted from user 14313400.

```
fingerprint: 0xde9106925599b7d
stripped_tweet: "Free Fitch report:Highly Leveraged Corp Borrowers to Compete for Limited Dollars http://bit.ly/1fdEvs via $$"
tweets <
  time: "2009-11-13 13:36:14"
  tweet: "RT @AlacraPulse: Free Fitch report:Highly Leveraged Corp Borrowers to Compete for Limited Dollars http://bit.ly/1fdEvs via @ResearchRecap $$"
  id: 14313400
  retweet_followers: 21436876
  num_followers: 399
>
tweets <
  time: "2009-11-13 13:53:47"
  tweet: "RT @AlacraPulse: Free Fitch report:Highly Leveraged Corp Borrowers to Compete for Limited Dollars http://bit.ly/1fdEvs via @ResearchRecap $$"
  id: 21436876
  num_followers: 8
>
```

B.2 Incorrect multiple retweets

In the example below, our algorithm detects that user 28188649 retweeted twice: once from user 45892886, and another from user 21454987. The user probably only retweeted from one of them, but it is not possible to conclude from which.

```
fingerprint: 0x8b62108b55722ae2
stripped_tweet: "\"Whenever you compliment anyone make sure it is accurate and sincere\" Zig Ziglar"
tweets <
  time: "2009-08-21 02:11:04"
  tweet: "\"Whenever you compliment anyone make sure it is accurate and sincere\" Zig Ziglar"
  id: 60327015
  num_followers: 40
>
tweets <
  time: "2009-08-21 02:13:24"
  tweet: "\"Whenever you compliment anyone make sure it is accurate and sincere\" Zig Ziglar"
  id: 45892886
  retweet_followers: 21354676
  retweet_followers: 21454987
  retweet_followers: 28188649
  num_followers: 2568
>
tweets <
  time: "2009-08-21 02:14:08"
  tweet: "\"Whenever you compliment anyone make sure it is accurate and sincere\" Zig Ziglar"
  id: 21454987
  retweet_followers: 21354676
  retweet_followers: 28188649
  num_followers: 14023
>
tweets <
  time: "2009-08-21 02:14:29"
  tweet: "\"Whenever you compliment anyone make sure it is accurate and sincere\" Zig Ziglar"
  id: 28188649
  retweet_followers: 21354676
  num_followers: 9554
>
tweets <
  time: "2009-08-21 02:14:38"
  tweet: "\"Whenever you compliment anyone make sure it is accurate and sincere\" Zig Ziglar"
  id: 21354676
  num_followers: 19545
>
tweets <
  time: "2009-11-20 15:42:58"
  tweet: "\"Whenever you compliment anyone make sure it is accurate and sincere\" Zig Ziglar"
  id: 34569915
  num_followers: 4117
>
```

B.3 Common tweets detected as retweet cascades

In the example below, the remaining tweet after stripping the `@screename` and `RT` tokens is `indico ~>`. From the cascade below, the phrase is considerably common, but it's not clear if any of the tweets are in any way related to each other, let alone retweets.

```
fingerprint: 0x8b735dcac43ec2b4
stripped_tweet: "indico ~>"
tweets <
  time: "2009-08-08 01:34:25"
  tweet: "RT @thiagaoo92: indico ~> @morning_bot @ReginaRamoss @ganjaboy74 @ygortischenko @Tiago783 @Lolo_peixoto @AnaCaroliiinah @gutuh"
```

```
  id: 40978340
  retweet_followers: 17086556
  num_followers: 812
>
tweets <
  time: "2009-08-24 17:45:58"
  tweet: "RT @djjadsonruslan indico ~> @ikaromennothy @beeelidalgo"
  id: 33386233
  num_followers: 2333
>
tweets <
  time: "2009-08-25 19:11:39"
  tweet: "indico ~> @andylorenzo RT @andylorenzo"
  id: 59113307
  num_followers: 8
>
tweets <
  time: "2009-09-04 03:25:24"
  tweet: "RT @taisa_naiara: indico ~> @PauloTakahashi @cih_ilhabela @Dinaardi @JefersonOliveir @ElderNask @kammys"
  id: 27578349
  num_followers: 539
>
tweets <
  time: "2009-09-06 17:16:43"
  tweet: "indico ~> @soll_ilhabela @LUANAPENZO @brennomotte @Guh_senna RT @Nessaenc @localMoSAC"
  id: 40694682
  num_followers: 43
>
tweets <
  time: "2009-09-06 18:38:33"
  tweet: "RT @katenasc indico ~> @mariliacaroline @GabrielSafari @ricky_henri @schneyder @novaispsirico @kaomore @EderPieroni"
  id: 55709125
  num_followers: 318
>
tweets <
  time: "2009-09-07 00:13:19"
  tweet: "RT @katenasc: indico ~> @danielccruz2 @schneyder @CaaiiQue @talita_ilhabela @rasnavalleria @ViniciusMyShopp @NatanUK @talita_ilhabela"
  id: 42202096
  num_followers: 654
>
tweets <
  time: "2009-09-14 08:53:48"
  tweet: "RT @lspearmanii: RT @myhmackenzie: indico ~> @roberto_ruiz @LocalMoSAC @Rafinha_Vas @RebecaDomingos @CrisL @lspearmanii @tictoc22 @kiferc"
  id: 21518886
  num_followers: 314
>
tweets <
  time: "2009-09-19 02:59:09"
  tweet: "RT @ricky_henri: RT @Enrique_Hall indico ~> @jujutvz @Smooth187 @Rosyka21 @WetNoses @Ariane_Siqueira"
  id: 21542738
  num_followers: 265
>
tweets <
  time: "2009-09-27 23:18:27"
  tweet: "RT @onblame indico ~> @_Alcoholic @nicolightning @isabela_b @mgramarim @suucarolina_"
  id: 38326772
  num_followers: 31
>
tweets <
  time: "2009-10-04 00:57:17"
  tweet: "RT RT @LocalMoSAC: indico ~> @SamuelLeite2 @abreeugustavo @Heavy_M4Tal @gui_gui_gui @elbiotwt @viniii93 @Kahx3Mah RT @kahrooL @localMoSAC"
  id: 53216906
  num_followers: 90
>
tweets <
  time: "2009-10-04 18:33:55"
  tweet: "RT @kahrooL: indico ~> @stefanbsp @thaciopassinho @renatoflavio"
  id: 23044598
  num_followers: 1061
>
tweets <
  time: "2009-10-04 18:54:21"
  tweet: "RT @official92: RT @kahrooL indico ~> @nintendo_will @abreumarlla @renatoflavio @stefanbsp @HenGuedes @ivsonsousa"
  id: 43144407
  num_followers: 49
>
tweets <
  time: "2009-10-04 23:39:23"
  tweet: "RT @kahrooL indico ~> @milberbernardes @samuelpretti @dttoficial @ichbinberry @monii32 @alliineb @Eliiiy_pereez @denisyuri @giu_ilhabela"
  id: 46283048
  num_followers: 208
>
tweets <
  time: "2009-10-07 05:18:27"
  tweet: "RT @kahrooL: indico ~> @Blaiiseee @LocalMoSA @jblm @danyel_boy @jduduc @renatoflavio"
  id: 15240444
  retweet_followers: 17086556
  num_followers: 10394
>
tweets <
  time: "2009-10-11 20:45:32"
  tweet: "RT @kahrooL: indico ~> @Thallinhos @Blaiiseee @UPfollowers @heynyne @Takhinha2 @RamonBarros @frankmagnifico @zoomarang"
  id: 28135772
  num_followers: 219
>
tweets <
  time: "2009-10-11 22:03:25"
  tweet: "RT @kahrooL indico ~> @lu_duzzi @emiliaaraujo @natanaelsacer @gustavoyngrid @Katia_OliverOps @gigeovane @raamonteiro @Victorialima_"
  id: 58854503
```

```
  num_followers: 59
>
tweets <
  time: "2009-10-13 00:21:43"
  tweet: "RT @kahrooL indico ~> @Priconstanti @Priconstanti @ItsBigRog @Blaiiseee @monndi @glngg @Itinho10 @Elona_ @naaay_k3 @ggalindo2 @glngg"
  id: 51918891
  num_followers: 8
>
tweets <
  time: "2009-10-13 17:56:26"
  tweet: "RT @kahrooL indico ~> @maira_fl @rodrigueshozana @AidenInc @gvalcorte @tarcioluciano @Blaiiseee @davidprado1 @DIOGO_LUCIANO"
  id: 36574520
  num_followers: 146
>
tweets <
  time: "2009-10-15 11:39:25"
  tweet: "RT @kahrooL: indico ~> @mechamoIrving @marcosquintal @odilei @RuthieNews @camalottllc @mikail6james @Marcus_Tav @nathanot @lidianamoret"
  id: 17086556
  num_followers: 2965
>
tweets <
  time: "2009-10-15 19:15:10"
  tweet: "RT @JCinqo RT @kahrooL: indico ~> @tieego @bellferrari @RuthieNews @sigamestre @cristianafranci @JCinqo @Daaniloh @beeccah @sigamestre"
  id: 45613253
  num_followers: 70
>
tweets <
  time: "2009-10-17 05:13:11"
  tweet: "RT @kahrooL indico ~> @Tondella @Marcus_Tav @ale_schuler @vinichichurra @Bismarckdepaula @JMYap @rockingto @sigamestre @roniyfonseca"
  id: 50927863
  num_followers: 134
>
tweets <
  time: "2009-10-17 05:39:45"
  tweet: "RT @kahrooL: indico ~> @Letizialuggo @P_Mendonca @ViihCarvalhoo @ricky_henri @jsaady @gvalcorte @ExpensiveGuy @mariinnahsouza @LunnaRibeiro"
  id: 42244258
  num_followers: 28
>
tweets <
  time: "2009-10-22 04:39:38"
  tweet: "RT @kahrooL: indico ~> @andersonreuel @marombeira2 @Inrocck @rafadmoura @jeemello @vinisqo @Frojack @andersonreuel @ThiiagoVieira"
  id: 58636171
  num_followers: 26
>
tweets <
  time: "2009-10-22 06:09:31"
  tweet: "RT @kahrooL: indico ~> @isaahmartiins @Ramonlima_ @brunolovesbrit @InRoll @brunolovesbrit @ValdirCardoso @allxclub888 @sigatodomundo"
  id: 17094167
  num_followers: 6912
>
tweets <
  time: "2009-10-24 18:47:51"
  tweet: "RT @davidprado1: RT @pedro_5 indico ~> @AnitaCaroline @perdomojr @thatah_zinha @ree_poosh @davidprado1 @aumentefollow"
  id: 36266034
  num_followers: 89
>
tweets <
  time: "2009-10-25 12:30:24"
  tweet: "RT @nicim_23: indico ~> @leeh_pinho @karenvegas @heyduds @nina_flor_ @lilianlameira @Brendha_valini @julbarion"
  id: 40087942
  num_followers: 142
>
tweets <
  time: "2009-10-29 19:12:57"
  tweet: "RT @pedro_5: indico ~> @follow11 @tweetforprofit @mariinnahsouza @FirojBD @lenzem @followubackall"
  id: 57638899
  num_followers: 1542
>
tweets <
  time: "2009-11-09 04:24:55"
  tweet: "RT @pedro_5 indico ~> @decentweet @marilynstannett @Camiila_ @mj_alive @Tchomperas @vi_brazil"
  id: 60828595
  num_followers: 1
>
tweets <
  time: "2009-11-27 00:41:38"
  tweet: "RT @pedro_5: indico ~> @CaaMiguel @jblm @NessaSama @ssnuness"
  id: 30773333
  num_followers: 38
>
tweets <
  time: "2009-12-28 17:25:43"
  tweet: "RT @LudmillaXP: indico ~> @LiveForLanza"
  id: 60424559
  num_followers: 17
>
```