Analyzing the short-term growth of online social networks subjectively

[Project Report]

Rishi Dabre*
Electrical Engineering and Computer Science Department
Syracuse University
Syracuse, NY
rrdabre@syr.edu

ABSTRACT

Online social networks like Twitter have been widely adopted as a medium for social interaction on the web. Study of this evolution can become a project extremely vast in scope with numerous potential aspects for evaluation. In this paper, we focus mainly on two such aspects, although coarse at the level of objective granularity, of these social graphs namely, structural changes and information propagation. Moreover, we analyze this growth of the Twitter network over a short period of not more than 21 days and subject the experiment to only certain information (categories). We conduct the experiment in two parts independently which comprises of data collection, processing and analysis, each covered in detail in the respective sections. Based on these results and our observations, we present our hypotheses about the apparent behavior with relevant arguments. We then talk about the limitations in our work and its future scope briefly. Finally, we conclude the discussion talking about the major takeaways from our analysis which includes- i) communities around popular nodes tend to grow and overlap significantly fast and proportionately with the popularity and ii) information pertaining to generic categories tends to spread faster and more against indigenous news. Our study presents a formally executed analysis of a seemingly trivial idea of the short-term growth of online social networks.

General Terms

Analysis, Growth, Online Social Network

Keywords

Short-term, Twitter, Hashtag, Streaming

1. PROBLEM STATEMENT

We desire to observe the changes in an online social network over a short period and determine- i) how much, how fast and to what extent does it grow and ii) how conducive is it for certain types of information. We pick one of the most popular online social networks- Twitter, for the interesting way (following) the connections are made in it and the way information is spread ultimately serving the purpose of providing an online platform for social interaction.

Within the period of a month, we collect few traces of this graph just enough large to be able to be visualized and processed upon for statistical computation with ease. We then perform a comparative analysis of these traces, specifically, we analyze the structural properties like the diameter, the clustering coefficient and the degree distribution. Further, we form a definite set of hashtags extracted from Twitter via the Streaming API to study the different extents to which the network was conducive for the specific information. We expect the results to be interesting and to vary considerably for the various kinds of information giving us valuable insight into the inclination tendencies of the social community towards the online network.

2. MOTIVATION

We are well aware of the steadily growing, if not diminishing, popularity of the online social networks available today. We believe these networks spanning across millions of people all over the world are bound to spread through their real life connections bringing in many more with time. While the timeline of this growth cannot be ascertained, we are interested in finding out how it manipulates, if at all, the network structure. Further, we intend to subject our analysis of the structural changes to only a select set of information spread and also restrict the apparently huge scope of our analysis therefore, by taking into account the growth over only a short-term period.

Such an analysis of the structural changes observed over a short period poses importance for it has not been done before, which we believe, is due to the triviality of the nature of such an analysis owing to the time period and the properties observed. We would further like to emphasize that an analysis like this can be of vital value to some significant areas as follows:

*Masters in Computer Science, May 2018

1. Advertisers and marketers - Promotions: An insight into the structural changes in an online social network can reveal useful trends to these observers which can be made sense of to determine efficient strategies of investing in promotions.

2. Industries - Anticipating data storage requirements: The online social network industries, like Twitter themselves, can anticipate and prepare if needed, the necessary measures for the growth of their network by observing its manner.

3. Government - Observing category of susceptible keywords: While far fetched, we believe that the Government can use the informational analysis to monitor and handle the spread of any sensitive content caused by the misuse of the social media.

3. RELATED WORK

A lot of work has been done around the evolution of social networks. In a recent study[6] from 2016, Efstathiades et. al. examined the changes in a social graph and the user behavior. But unlike our study, it was stretched over an extremely long term and the initial trace was not collected by the authors. In another work[7] (2016) by Zhang et. al., they studied how social groups evolve over time, which is very relevant to our definition of the problem except that their aim was to examine the temporal patterns to propose an evolution model.

Two of the interesting studies conducted in 2016 and 2015 dealt with social networks- Weibo[8] and Douban[9], respectively, which were specific to only a region. In the former, Ma et. al. examined the evolution of this real online social graph by sampled data. But a significant deviation in their method from that of ours was that they took into account the users' joining dates to build the evolving model instead of observing the network at time steps. Besides, Shan et. al. examined the evolution patterns of social and content networks in the latter study with the deviation being that of their focus on the User Generated Content (UGC) for the growth of network. Although one would argue that it aligns with our idea of analyzing the information spread across the network, it must be noted that they related the importance of the UGC directly with the evolution of the network while we aim to analyze the propagation of information per se, in the network.

Relatively early (2013), Tremayne studied a very interesting problem of observing how Twitter was used to organize and spread a real-world social movement[10]. The idea of analyzing how Twitter is conducive to different extents for certain information considerably aligns with their motive. While their study answers a more specific question of finding the central nodes and how the movement propelled, our goal is to present a precise picture of the conduciveness of Twitter to different kinds of information.

Furthermore, some studies from the past were relevant, lesser to our key idea but considerably to the notion of online social network evolution. In 2012, BrÃşdkaEmail worked on proposing and comparing a new algorithm to explore the evolution of social groups including all the changes- continuing, shrinking, growing, splitting, merging, dissolving, forming[11]. Lastly, two of the recent studies from 2015 and 2016 dealt with proposing a model for co-evolution of network structure of users and content[12] and revealing the leader and ordinary users from a specific microblogging site "Sina micro-blog"[13].

As discussed about earlier, our focus is on observing the structural changes over a short period of time while also confining the analysis to only specific information flowing across the network, none of which has been concisely answered before.

4. APPROACH

We perform the data collection and analysis in two parts to address the two distinct questions from our problem definition.

4.1 Part 1: Observing structural changes

In this part, we collected two traces of the Twitter users network falling 21 days apart. For the first sample, we obtained the trending 50 tags and filtered using the same, the top 5 then active (tweeting with the hash tags) users having the most followers with the help of Streaming API[3]. Next, we obtained top 40 followers having the most followers for each of these 5 users with the help of REST APIs[4]. Similarly, for each of these 40 followers of the 5 users, we obtained the top 40 followers having the most followers, again using REST APIs. Collection of each trace required consistent execution of the script for more than 2 days altogether.

After each data collection, we formed the data set conforming with the semantics of the ones for use with the snap.py[1]. Snap.py, which provides built-in functions for calculating the graph statistics and plotting different distributions, was used to compute the average clustering coefficient and the diameter of each graph trace and also to plot the clustering coefficient and the degree distributions. Further, we used Gephi[5] as the graph visualization tool and created visualizations of each trace using the Fruchterman-Reingold layout.

4.2 Part 2: Observing information propagation

For this part, we first obtained the top 50 trending tags immediately followed by obtaining 50000 tweets using these tags in real-time using Streaming API and its filtering facility which took approximately 20 minutes. 1 week later, at the same time, we attempted to collect 50000 tweets filtered using the same trending tags from the last week but could only reach the number 32437 in more than an hour after which, the machine's response deteriorated and it hung up. This was not surprising because i) it cannot be expected that the trending tags remain the same even after a week, precisely an implication of our requirement and moreover, ii) the sample of trending tags was naturally biased towards the first trace.

5. EXPERIMENTAL RESULTS

We now present our findings for each part of the experiment and our hypotheses based on them.

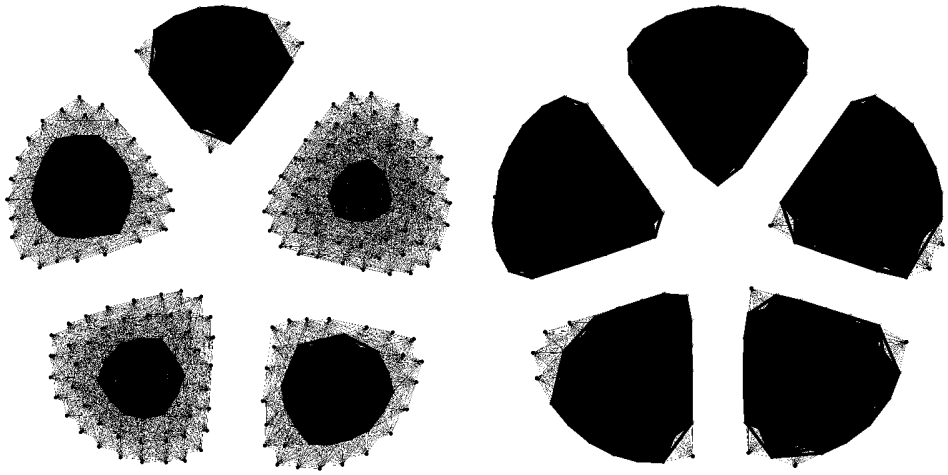

Figure 1: Visualized traces of Twitter user network (left- first, right- second)

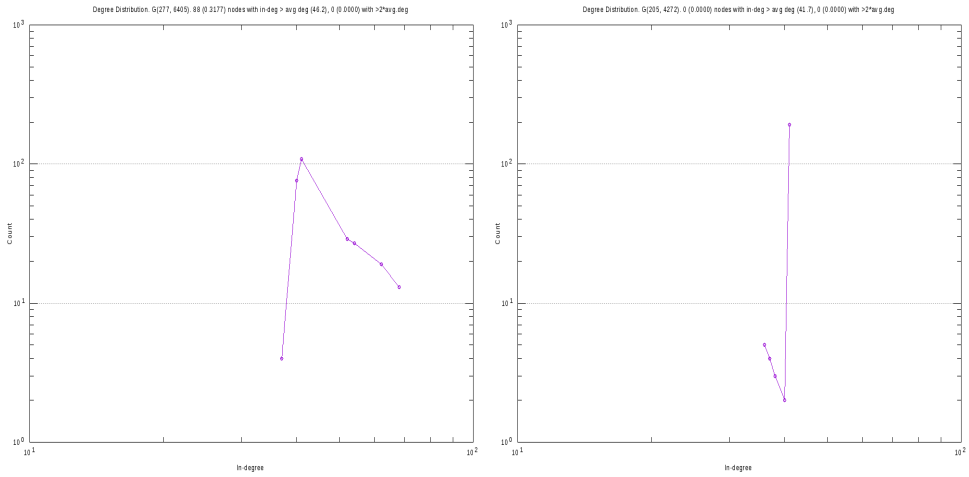

Figure 2: Degree distribution of Twitter user network traces (left- first, right- second)

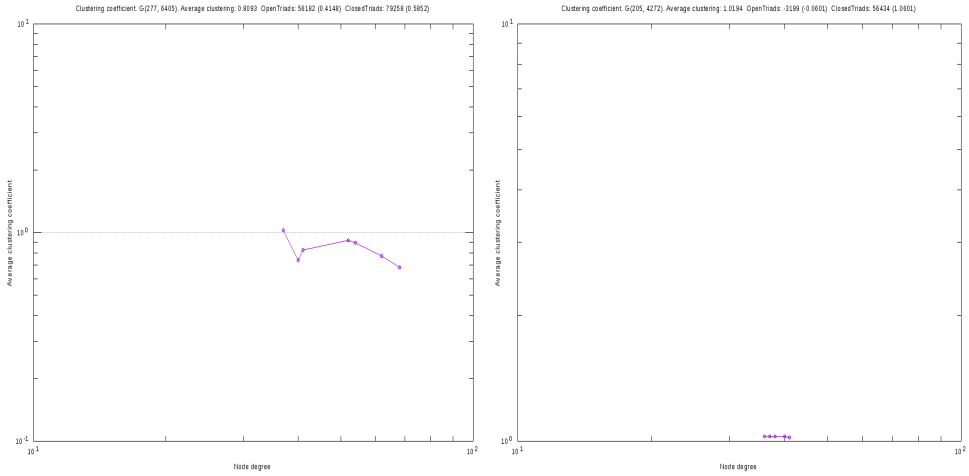

Figure 3: Clustering coefficient distribution of Twitter user network traces (left- first, right- second)

5.1 Part 1: Analysis of structural changes

Figure 1 shows the traces of the Twitter network as visualized with Gephi. Despite trying out different layouts, to our dismay, Gephi could render the graphs neither accurately nor completely. However, we would like to pinpoint the prominent difference between the two traces indicating a significant increase in the graph density over the period. Further, Figure 2 and Figure 3 show the degree and clustering coefficient distribution, respectively. It was observed that the average clustering coefficient for the first trace was 0.81 while that for the second one was 1.0, indicating the increased formation of triads in the network. Besides, the diameter of the largest connected component in both the samples remained 2 which reinforces the small-world nature of the network. More interestingly, we found the number of connected components to have gotten reduced from 5 to 1 which points towards the exhaustive formation of new connections among the users.

From the obtained results, we hypothesize that as a general tendency, communities around the most followed users grow and overlap rapidly. Intuitively, since the selected users are the ones who are both the most followed and the most active (filtered based on their tweets on the trending topics).

5.2 Part 2

In this part, we focused on discerning the different categories of hash tags based on their usage trend. Figure 4 shows the plot of tweet counts against the tags they were made using. We observed that 21 of the 50 tags were short-lived, meaning that they faded away quicker than the others. 25 others were seen to be consistent in that their usage did not deviate much over the week. 4 of them rose up pretty fast which we refer to as suddenly burst tags. We also observed that the maximum deviation in the usage of tags was 2925 for the ones getting outdated and 5008 for the ones gaining popularity. The minimum deviation for the same remained 0 (rounded up on the observation scale for values closer to 100). However, from among the list of 50 trending tags, 17 were not tweeted about at all in the first trace, which became 25 in case of trace 2.

Based on these results, we hypothesize that approximately half of the trending tags are retained over the period of a week which is a reasonable assumption to make. We also observed that the rate of the rising tags was around 5 times slower than that of the fading ones i. e. the ones getting outdated over the period of a week. Besides, the extent of these upcoming hash tags was almost twice that of the ones losing popularity. This makes sense as the usage of hash tags is driven by its spread while it requires no specific agent for the tags to be simply forgotten. Furthermore, we found a number of âĂŸnon-tweetedâĂŹ tags which again, given the limited time frame we collected the data in, can be reasonably assumed to have been used at other time of the day. Also, the increase in the number of such tags in the second trace is intuitive being in accordance with the diminishing popularity of the trending tags over the period. Finally, we note that the time taken for the collection increased significantly during the second trace which can be justified, again due to the set of tags we used to filter the data by.

6. LIMITATIONS

Despite the demarcated scope of our idea, some limitations arose during project execution.

Most importantly, the visualization tool, Gephi, was unable to render the graphs clearly which prevents the revelation of the vital trends in the graph.

Secondly, traces in the second part were collected for a very short time. We believe that this has a major implication in that the diversity of the data thus collected tends to deteriorate. Instead, it can be a better approach to, for comparatively fewer tags, stream the tweets and build the data set over a longer period of time. This will both ensure that the aforementioned diversity is preserved due to observation over a longer period and yield a picture with a higher resolution so as to reveal even the nuances in the graph structure.

Furthermore, time of the day for data collection was chosen arbitrarily for both the parts. This could have skewed the results not only as user activities tend to vary with different times of day but also due to the various time zones the collected data sources to. Without going to the extreme end of the spectrum and collecting data for every possible time slot of the day, samples of a select few periods can provide a more exhaustive idea of the distribution of hash tags.

Finally, it can be conspicuously established that the trending topics' list obtained, during the first collection, is biased towards the first trace. In other words, the list of tweets obtained in the first trace gets undue advantage of the filter tags being selected from the trending list from the time span overlapping with that of itself. As originally intended, we would prefer generating the list of filter tags manually in relevance to a future event that will eliminate this unnecessary bias and moreover, allow the analysis of information propagation to be more precise and subjective.

While all these limitations could have been kept away from, we chose to prioritize project completion over accuracy and integrity of the data, which would otherwise require both more time and more capable resources.

7. FUTURE SCOPE

While overcoming the limitations itself comprises a substantial future scope for the project, we would like to highlight a few important aspects required to be looked at in more depth. First, we believe measuring the other characteristics in addition to the ones computed in this study, including graph density, number of communities, etc., may provide more insights and may potentially even reinforce some of our hypotheses. Also, we encourage looking at other networks like Facebook and Google+ for that those are significantly different from Twitter as platforms for social interaction and seemingly involve more challenges in collecting data. A comparison of structural changes and information propagation across these three networks can highlight the peculiarities of each of them.

8. CONCLUSION

Based on the experimental findings, our inclination is to believe that the popularity of users catalyses the convergence of connections among the users of online social network. In general, the more popular a user is (here, the measure being

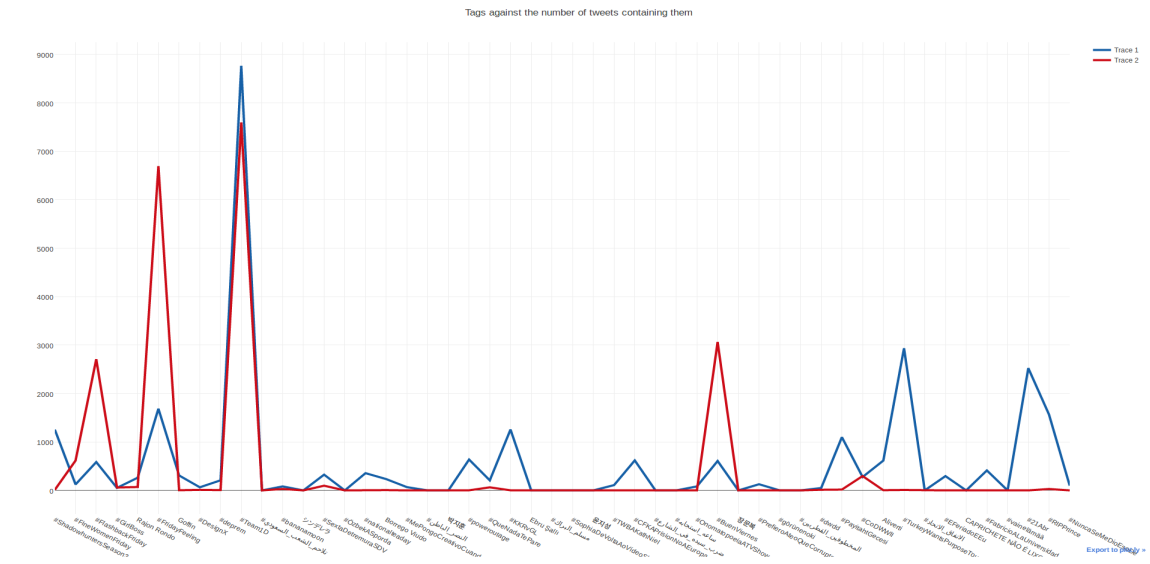

Figure 4: Count of tags tweeted

the number of followers of the user), the more likely are their followers to form connections among each other in turn resulting in a denser network structure implying structurally by the increased number of triangle formations. A period of approximately 3 weeks has been found to be enough for the growth and overlapping of communities for a Twitter network sample of this scale. The finding can however be generalized to other online social networks.

Our analysis of the information propagation revealed that generic terms (say, Friday) were more likely to be tweeted about than indigenous ones (say, RIPPrince), which is in conformity with ones general expectations. Also, on an average, the hash tag trend follows a 50% decay rate over a week i. e. almost 50% of the topics tend to disappear over a period of 7 days. We however expect this window to vary slightly depending upon the category of the hash tags being observed.

9. REFERENCES

[1] Snap.py Official Page `http://snap.stanford.edu/snappy/index.html`

[2] SNAP Project Page `http://snap.stanford.edu/`

[3] Twitter Streaming APIs `https://dev.twitter.com/streaming/overview`

[4] Twitter REST APIs `https://dev.twitter.com/rest/public`

[5] Gephi `https://gephi.org/`

[6] Hariton Efstathiades, Demetris Antoniades, George Pallis, Dept. of Computer Science, ZoltÃąn SzlÃąvik, Robert-Jan Sips. Online social network evolution: Revisiting the Twitter graph. In *2016 IEEE International Conference on Big Data (Big Data).*

[7] Tianyang Zhang, Peng Cui, Christos Faloutsos, Yunfei Lu, Hao Ye, Wenwu Zhu, Shiqiang Yang. Come-and-Go Patterns of Group Evolution: A Dynamic Model. In *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pages 1355-1364.*

[8] Lu Ma, Gang Lu, Junxia Guo. Observing the evolution of social network on Weibo by sampled data. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS).*

[9] Siqing Shan, Jie Ren, Cangyan Li. The dynamic evolution of social ties and user-generated content: a case study on a Douban group. In *ISSN: 1751-7575 (Print) 1751-7583 (Online) Journal homepage: http://www.tandfonline.com/loi/teis20.*

[10] Mark Tremayne. Anatomy of Protest in the Digital Era: A Network Analysis of Twitter and Occupy Wall Street. In *ISSN: 1474-2837 (Print) 1474-2829 (Online) Journal homepage: http://www.tandfonline.com/loi/csms20.*

[11] Piotr BrÃşdkaEmail, StanisÅĆaw Saganowski, PrzemysÅĆaw Kazienko. GED: the method for group evolution discovery in social networks. In *Social Network Analysis and Mining ISSN: 1869-5450 (Print) 1869-5469 (Online).*

[12] Prasanta Bhattacharya, Tuan Q. Phan, Xue Bai, Edoardo M. Airoldi. A Co-evolution Model of Network Structure and User Behavior in Online Social Networks: The Case of Network-Driven Content Generation. Available at *SSRN: https://ssrn.com/abstract=2703994 or http://dx.doi.org/10.2139/ssrn.2703994.*

[13] ZHANG Xin, CHEN Chao, HAN Dingding. An evolution model of online social networks based on "Sina micro-blog". In *Vol.45, No.3 Journal of Shanghai Normal University (Natural Sciences) Jun., 2016.*