IITP Summer Internship (2019) Report

Estimating Re-tweet likelihood using topological metrics

Debolina Mahapatra
mahapatra.debolina201@gmail.com
Dept. of Computer Science,
Silicon Institute of Technology, Bhubaneswar.

ABSTRACT – Since many years, Twitter has been an important platform in the study of social network analysis. Twitter supports "retweets", which is its most important feature and a means of information diffusion across the network. The study of the retweeting phenomenon has been done in a lot of literature in the past. However, in the recent years, the work of estimating the retweeting likelihood has received much attention. Some of the methodology adopted for this purpose incorporate analysis of the linguistic features of tweets, user information such as behavioural aspects, and studying information cascades using contagion models. Some of these works have considered using machine learning and deep learning methods for predictive modelling. In this work, we model the retweet prediction problem as a binary classification task. We have used the Higgs dataset consisting of 400k users and have studied both the social graph and the implicit retweet graph. This work focuses on the topological aspects. We use several underlying topological metrics as features in order to train a generalized linear model (GLM).

***Index Terms* — social network, re-tweet, Twitter, power law, generalized linear models**

I. INTRODUCTION

Over the last decade, online social networking sites have become very popular. These sites allow users to follow streams of posts generated by their friends and acquaintances. With the introduction of blogging services such as Wordpress and Blogspot and social networks such as Facebook, Twitter and Instagram, information sharing and consumption takes place online. Twitter is presently one of the most popular micro-blogging applications. Unlike some other social networking websites like Facebook which allow bi-directional links, Twitter allows directional links within the network structure among its users using the follower-followee connections. It supports one-way following relationships. Twitter is a blend of instant messaging, micro-blogging and texting, with brief content and mostly, a broad audience.

Primarily, Twitter relies heavily on micro-blogging for communication. Users post short messages called ‘tweets’ of length up to 280 characters in order to communicate with their followers. Twitter also includes other forms of communication such as “@ mentions” and “direct messages”. Twitter also supports the use of “#” hashtags to specify the subject of a tweet. People make connections by following other people’s twitter feeds. It is worth specifying that Twitter is a very public forum. Users may choose to lock their profiles so that only their followers can view their tweets. However, most of the users choose to keep their profiles open making their tweets accessible to the

public. Twitter allows users to broadcast their message all over the site. A Re-tweet is a re-posting of a Tweet. A follower can choose to 'retweet' a tweet from one of his followees and thus spread the tweet to her followers. Thus, through the retweet mechanism, a tweet from the original tweeter can propagate to many other users through the follower-followee connections. Retweeting has been studied as an important mechanism of information propagation on Twitter. Moreover, it has been observed that there is a spike in usage during major events.

Social media platforms allow rapid information diffusion, and serve as a source of information to many of the users. Particularly, in Twitter information provided by tweets diffuses over the users through retweets. So it is of great significance to study the characteristics of retweets. In this work, we have studied the Twitter social network using the SNAP Higgs Dataset. We have also studied the retweet network obtained from this social graph. The topological measures and metrics have been computed from the retweet graph and the social graph, and a comparative study has been done. Further, using the topological parameters, we have done predictive modelling using GLMs in order to predict whether a tweet will be retweeted or not.

II. LITERATURE REVIEW

Tweets generally travel in the network via the explicit social graph which has been well-studied in several works. Some main approaches for the retweet prediction task involves mining textual content, analyzing social network structure and studying the underlying mechanism of re-tweeting behavior. Zhang et. al [1] proposed a factor graph model to predict users' re-tweeting behaviours. Their proposed method achieved a precision of 28.81% and recall of 37.33% for prediction of the re-tweet behaviours. In their work, they have stated that the factors affecting re-tweeting are importance of content (eg. a trending topic) and interest of user. In some other works, re-tweeting has been examined as a way by which participants can converse and their retweeting practices (how, why and what they generally retweet) have been studied. Retweeting has been studied as a form of information diffusion and a means of participating in a diffuse conversation.

Remy et. al [2] have studied information propagation on Twitter in the time of crisis, by investigating the length of retweet chains of users. In their work, they have shown that the distribution of the probability of producing a tweet of a given retweet chain length can be represented as a power law for users of a given in-degree. The number of followers of a user has an influence on the average retweet chains of tweets. Their proposed model is accurate enough to generate realistic length of retweet chains on the network. Some other works have considered an epidemic model for the study of retweeting. Goel et. al [3] have studied retweet cascades on Twitter as an epidemiological process. Classical diffusion models have been used to study the dynamics of the cascade formation in Twitter. A model has been created using a multi-layer representation of tweet propagation in which tweets propagate via retweets in one layer, and via mentioned users in another layer.

The retweet likelihood problem has been solved using machine learning techniques and user behavior data. Individuals have been studied using their profiles and past retweeting behavior. Their interests and the active times of the day were also recorded. By searching for correlations in the data, the algorithm picks users who are most likely to RT on a particular topic. Bild et. al [4] have emphasized the study of the retweet graph. In their work, the re-tweet graph has been leveraged for spammer detection in the Twitter network by detecting spammers via their low connectivity in the retweet graph. The retweet graph analysis revealed structural differences from the followers graph that are more consistent with real world social networks.

The Twitter social graph has been extensively studied in most of the works. However, the implicit retweet graph has not received much attention. In this work, we have analyzed the subtle differences between the social graph and the retweet graph. The topology of the social graph has been considered for retweet likelihood in a number of literatures. We focus on the topological analysis of the retweet graph, and analyzing the metrics and their effect on the number of retweets.

III. DATASET DESCRIPTION

The information spreading processes on Twitter has been studied and The Higgs dataset [5] (available on SNAP website) has been built after monitoring the spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson on 4th July 2012. The messages posted in Twitter about this discovery between 1st and 7th July 2012 are considered.

The four directional networks made available here have been extracted from user activities in Twitter as:

i) re-tweeting (re-tweet network)
ii) replying (reply network) to existing tweets
iii) mentioning (mention network) other users
iv) friends/followers social relationships among user involved in the above activities
v) information about activity on Twitter during the discovery of Higgs boson

The user IDs have been anonimized, and the same user ID is used for all networks.

In this work, we have studied the retweet network and the social relationships network from the Higgs Twitter dataset.

a. RETWEET NETWORK

In a RT graph, if A has re-tweeted B, then there is a directed edge from A to B. The in-degree is the no. of unique users who RT a node. The out-degree is the no. of unique users re-tweeted by a node. This graph is small-world, scale-free, less disassortative and has much stronger clustering. It more closely models real-world social and trust relationships among users.

In our study, we have used the Re-tweet graph from the Higgs dataset. This network consists of 256,491 nodes and 328,132 edges.
The Re-tweet graph is in the form of a weighted edge-list, consisting of a source node, a destination node and the number of times the source re-tweets the destination (weight).

The RT Graph statistics, as computed by using Cytoscape, are as follows –

Clustering coefficient :	**0.008**	Number of nodes :	**256491**
Connected components :	**13199**	Network density :	**0.0**
Network diameter :	**23**	Isolated nodes :	**0**
Network radius :	**1**	Number of self-loops :	**0**
Shortest paths :	**156883071 (0%)**	Multi-edge node pairs :	**758**
Characteristic path length :	**7.937**	Analysis time (sec) :	**1686.557**
Avg. number of neighbors :	**2.553**		

b. SOCIAL RELATIONSHIP NETWORK

The Twitter social network graph consists of 456,626 nodes and 14,855,842 edges. The relationship between the nodes can be characterized as follower/following. The social network is present in the form of an edge-list.

The following are the network statistics as given on the official SNAP website –

Retweet Network statistics	
Nodes	256491
Edges	328132
Nodes in largest WCC	223833 (0.873)
Edges in largest WCC	308596 (0.940)
Nodes in largest SCC	984 (0.004)
Edges in largest SCC	3850 (0.012)
Average clustering coefficient	0.0156
Number of triangles	21172
Fraction of closed triangles	0.0001085
Diameter (longest shortest path)	19
90-percentile effective diameter	6.8

Social Network statistics	
Nodes	456626
Edges	14855842
Nodes in largest WCC	456290 (0.999)
Edges in largest WCC	14855466 (1.000)
Nodes in largest SCC	360210 (0.789)
Edges in largest SCC	14102605 (0.949)
Average clustering coefficient	0.1887
Number of triangles	83023401
Fraction of closed triangles	0.002901
Diameter (longest shortest path)	9
90-percentile effective diameter	3.7

IV. METHODOLOGY

In order to study the characteristics of the social graph and the retweet graph, and for a comparative study between the two networks, we consider some important topological measures [6] and metrics. We quantify the network structure by using the below listed measures.

a. CENTRALITY MEASURES

1. Degree Centrality – Degree is defined as the number of connections that a node has. The degree centrality for a node is the fraction of nodes it is connected to. Nodes having a higher degree are more central as compared to the other nodes. For a directed graph, we have two types of centrality measures: out-degree centrality and in-degree centrality. Out-degree is the number of ties that the node directs to others. In-degree is a count of the number of ties directed to the node.

2. Closeness Centrality – It measures how close a node is to all other nodes in the network. Closeness centrality of a node is the reciprocal of the sum of its distances to all other nodes. It gives a measure of the mean distance from a vertex to other vertices.

3. Eigenvector Centrality – It (also called eigen-centrality) is a measure of the influence of a node in a network. In many circumstances, a node's importance in a network is increased by having connections to other nodes that are themselves important. It gives each vertex a score proportional to the sum of the scores of its neighbours. In directed graphs, the right eigen-vectors are considered for computing the centrality.

4. Page Rank Centrality – PageRank is an algorithm that measures the transitive influence or connectivity of nodes. There are three distinct factors that determine the PageRank of a node: the number of links it receives, the link propensity of the linkers, and the centrality of the linkers.

5. Hubs and Authorities – HITS (Hyperlink-induced topic search) algorithm gives each vertex 'i' in a network, an authority centrality and a hub centrality. Nodes that contain useful information on a topic of interest are called authorities. Hubs are nodes that tell us where the best authorities are to be found. Authority centrality and hub centrality quantify vertices' prominence in the two roles.

b. CLUSTERING COEFFICIENT

The degree to which nodes in a graph tend to cluster together in tightly-knit groups is given by the clustering coefficient. The clustering coefficient is a measure correlated to a number of triangles in the network, i.e., sets of fully connected triplets of nodes. The local clustering coefficient of a vertex (node) in a graph quantifies how close its neighbours in order to become a complete graph.

c. AVERAGE PATH LENGTH

It is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. It is a measure of the efficiency of information transfer on a network. Most real networks have a very short average path length leading to the concept of a small world where each one is connected to the other through a short path.

d. DEGREE DISTRIBUTION

The degree distribution of a network P(k) is defined to be the fraction of nodes in the network with degree k. Degree distribution depicts the difference in the degree of connectivity between all the nodes in a network. It gives the probability distribution of the node degrees over the network.

e. ASSORTATIVITY

Degree assortativity measures the tendency of nodes to connect with others of similar degree. In directed networks, we consider all possible directional degree pairs as separate assortativity metrics. The assortativity coefficient is the Pearson correlation coefficient of degree between pairs of linked nodes. Positive values indicate a correlation between nodes of similar degree, while negative values indicate relationships between nodes of different degree. For a directed graph, we computer 4 types of assortativity: (in, in), (in, out), (out, in), (out, out).

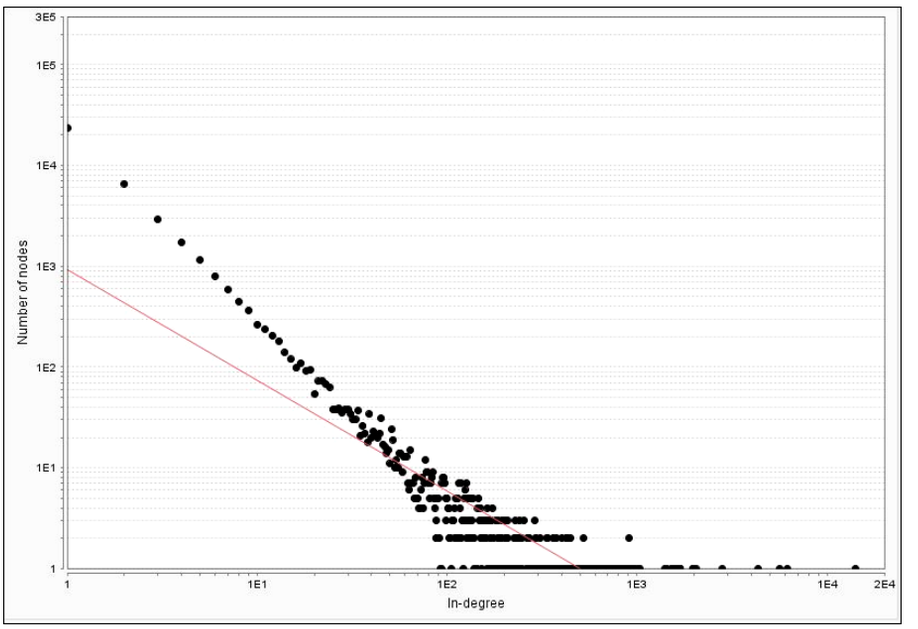
f. RECIPROCITY

Reciprocity measures the number of links that are bi-directional. It is given as the ratio of the number of bi-directional links to the total number of directed edges.

V. ANALYSIS OF RE-TWEET NETWORK

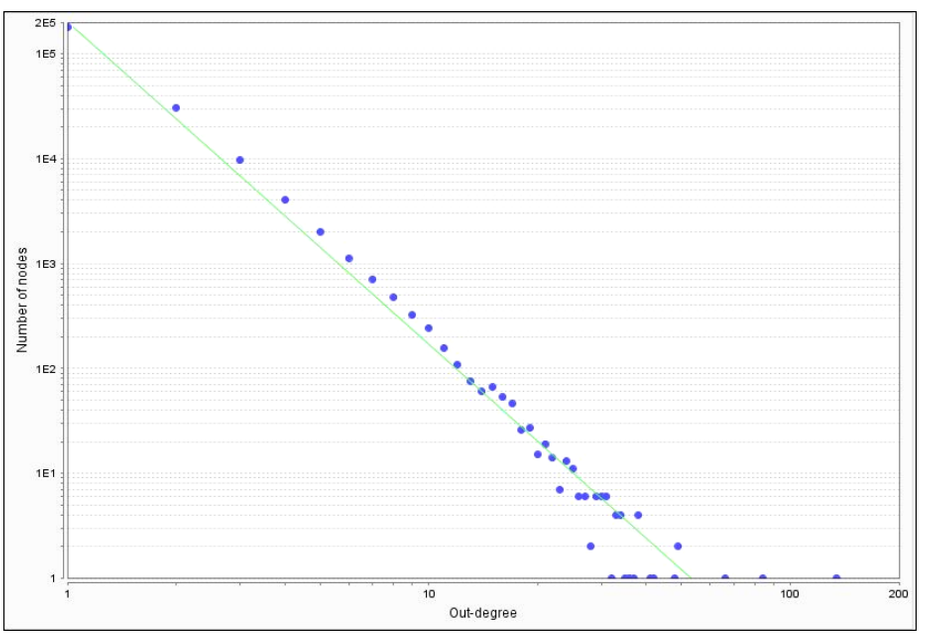
a. In-Degree Distribution: The graph represents the logarithmic degree distribution for the re-tweet graph. We tried to fit a power law to the in-degree distribution of the Re-tweet graph. The red line depicts the power law. It can be clearly seen that the in-degree distribution may seem to follow the power law, but it does not. It is a heavy-tailed skewed distribution.

In case of retweet graph, the in-degree depicts the number of times a node gets retweeted.



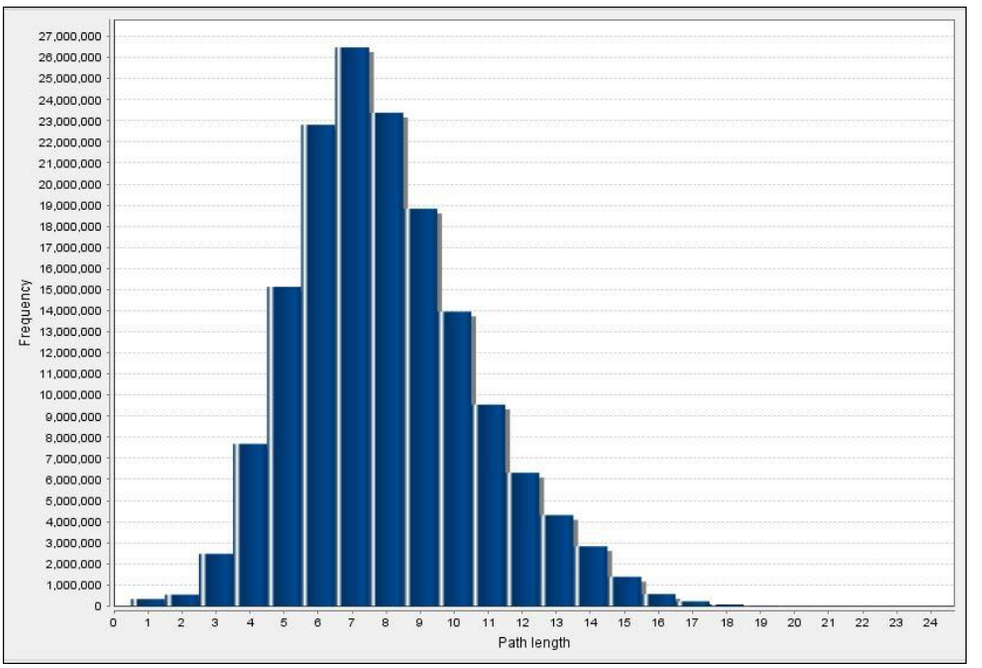
b. Out-degree distribution: The curve below depicts the logarithmic distribution for the average out-degree. The blue line is the power law fit to the distribution. It can be seen that the out-degree distribution tends to follow the power law. The in-degree variance is higher.

In case of retweet graph, the out-degree depicts the number of times a node retweets others.

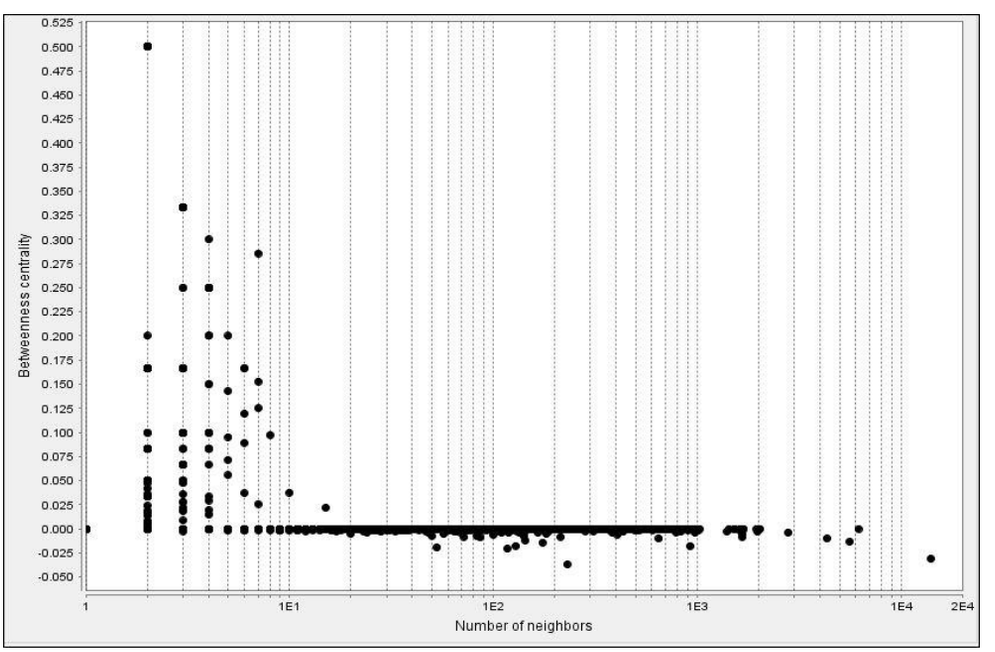


On Twitter, we define popularity as the number of users who retweeted an individual (given by in-degree) and prolificity as the number of users an individual retweeted (given by out-degree).
From the above distributions, we can conclude that popularity is more variable than prolificity [ ].

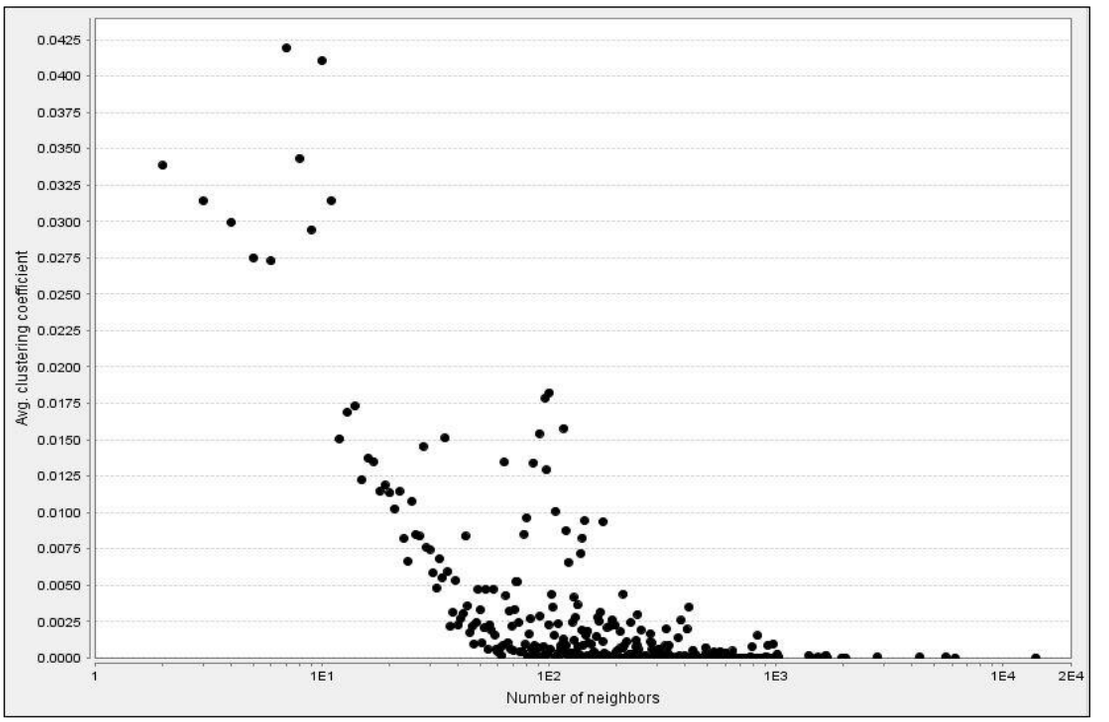
c. *Average Path Length distribution:* Average path length is a concept in network topology that is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. It is a measure of the efficiency of information or mass transport on a network. The average path length distribution is a right-skewed distribution.



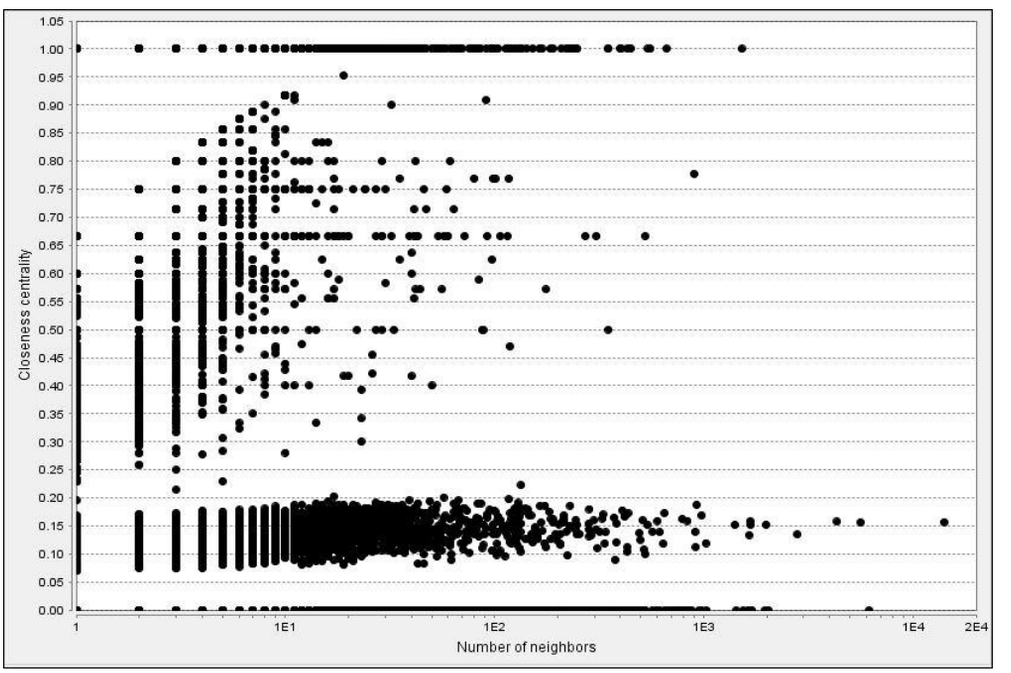
d. *Betweenness Centrality:* It measures the extent to which a vertex lies on the paths between other vertices. Vertices with high betweenness may have considerable influence within a network by virtue of their control over information passing between others. It can be seen from the graph that a only few nodes have a high betweenness centrality.



e. *Clustering Co-efficient:* Only few nodes can be seen as having a high average clustering coefficient.



f. *Closeness Centrality:* Closeness centrality indicates how close a node is to all other nodes in the network.



VI. EXPERIMENT

For all the computational purposes, the NetworkX [7] library of Python3 has been used in Spyder (Scientific Python Development Environment) in Windows10 64-bit Operating System, having an Intel®Core™i3-5015U (2.10GHz) processor and an 8 GB RAM capacity. The Higgs dataset, in the form of an edgelist, has been read by using NetworkX and the corresponding social network graph and the retweet graph was generated with the help of the DiGraph module.

a. Assortativities for retweet graph: In a directed network, we consider all possible directional degree pairs as separate assortativity metrics.

r (in, in)	-0.002
r (in, out)	0.034
r (out, in)	-0.011
r (out, out)	0.131

Assortativity gives a measure of the node degree correlation. Real world social networks are assortative in nature, that is, high degree vertices are surrounded by a periphery of lower degree ones. In an assortative network, the content easily propagates through connected components of tightly clustered, high-degree nodes that are resistant to node removal, but may not reach the low degree boundary of the network. Online social networks are generally dis-assortative in nature. These have larger connected components, so the content propagates further, but can be partitioned by the removal of a high degree node.

b. Reciprocity for the retweet graph is 0.0046, and for the social graph, its value is 0.316.

Taking the topological metrics into consideration, a dataset was constructed for the nodes involved in the retweet network structure.

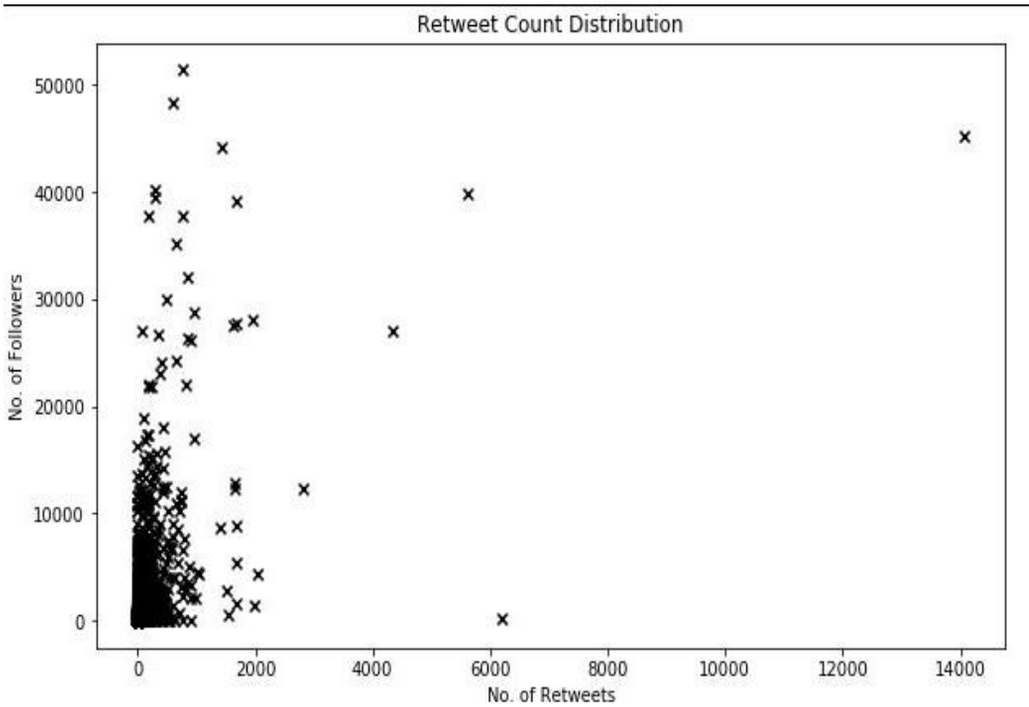
The following are the topological parameters in our dataset:

1. out-degree centrality
2. in-degree centrality
3. clustering coefficient
4. closeness centrality
5. authority centrality
6. hub centrality
7. eigen-vector centrality
8. page rank centrality

The follower-followee ratio was also included in the dataset.

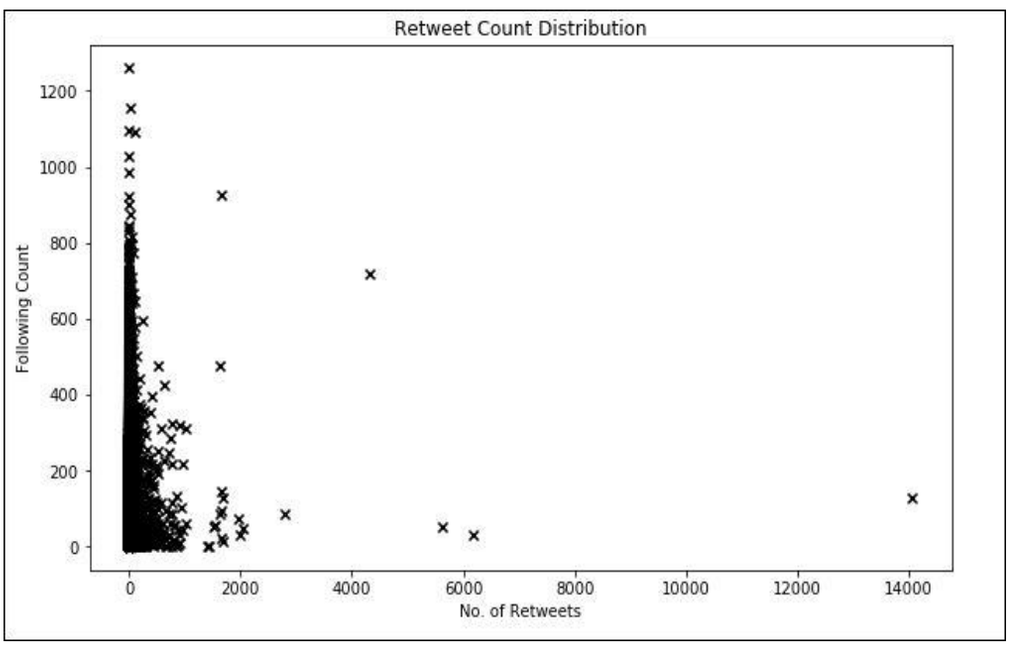
With the help of this Twitter social network graph, we find the number of followers for each node, and the number of connections each node follows. We reverse the original Re-tweet graph in order to analyse the direction of information flow. To draw additional insights, we use the information from the social graph, and merge it with that of the retweet graph.

We plot the distribution of number of retweets and find their correlations with the number of followers and followees.



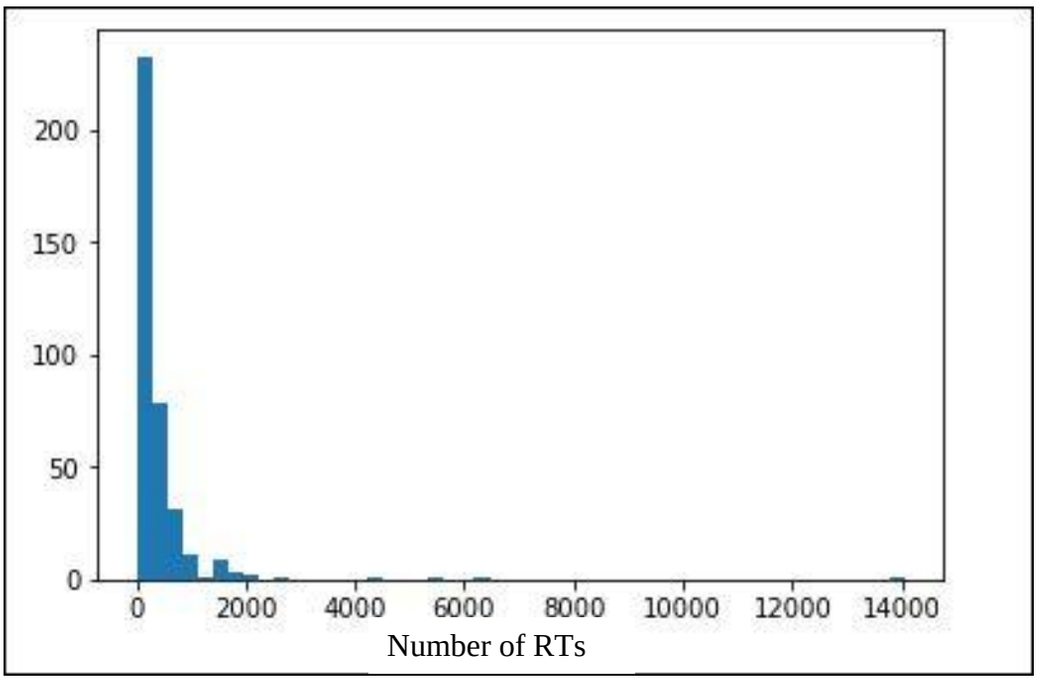
We cannot exactly say that the number of RTs is directly proportional to the number of followers. In some cases it is true; however, we can see clearly in the graph that the accounts having a large number of followers do not necessarily receive the highest number of RTs. While there exists some nodes that have lesser number of followers but they achieve a high number of RTs.

Next, we analyse the RT count distribution with respect to the number of accounts that a particular node is following.



The nodes which follow a large number of accounts cannot, however, attain a huge number of RTs. We can infer that merely following a large number of accounts cannot help to spread some information. On the other hand, we can see that the node that receives the highest number of RTs (around 14,000) has a huge number of followers but itself follows less number of nodes.

The histogram below represents the frequency distribution of re-tweets. We can see that only very few tweets receive a huge number of RTs.



We model the retweet prediction problem as a binary classification task.

Earlier works have considered content-based features (eg., presence or absence of hashtags), structure-based features (eg., friends count, statuses count), and multimedia-based features.
In this work, we primarily focus on the topological parameters.

For training the statistical learning models, we have used the h2o library in Python.

H2O is an open source, in-memory, distributed, fast, and scalable machine learning and predictive analytics platform that allows us to build machine learning models on big data.

A generalized linear model (GLM) is defined by the following three components:

a. An exponential family model for the response
b. A linear regression function or a linear predictor in the explanatory variables
c. A parameter transformation or a link function which relates the linear predictor to the mean

The most famous cases of GLMs are linear models, binomial & binary regression and poisson regression.

In our analysis, we have considered the Logistic Regression as a GLM. It models data as outcomes of a coin flip. Further, this model uses a Binomial Likelihood function, which perfectly fits our requirements.

VII. RESULTS AND DISCUSSION

The topological parameter dataset was divided into 70% training data, 15% validation data, and 15% test data. 10 fold cross-validation was applied. This model was evaluated as follows:

Evaluation metrics	Training data	Validation data	Test data	CV data
Accuracy	0.993	0.994	0.993	0.992
Mean squared error	0.033	0.034	0.034	0.035
Root mean squared error	0.184	0.184	0.186	0.189
Precision	0.991	0.992	0.993	0.987
Recall	0.967	0.971	0.966	0.963

The GLM based Logistic Regression model performs fairly for the binary classification task of retweet prediction.

VIII. CONCLUSION

The retweet graph is small-world and scale-free. It more closely models the real world social relationships. We have observed several topological measures for the retweet graph and have also compared it with the social graph of Twitter. The reciprocity (fraction of bidirectional links) is much less in the case of the retweet graph as compared to the social graph. This demonstrates that retweeting imitates real-world trust relationships. One not just listens to others' ideas, but actively forwards them to one's own followers. However, each node has only little control over the incoming links which why number of bidirectional links are so less.

We have also seen that the degree distributions are heavy-tailed power law. Preferential attachment occurs between exisiting nodes. Assortativity metrics also demonstrate that the prolific retweeters retweet each other. Moreover, the network is not tightly clustered as the average clustering is considerably low. The retweet graph is not very strongly connected. Also, there exists independence between one's own retweet behavior and the number of retweets one receives.

We further used a Logistic Regression model for the binary classification task of whether a given tweet will be retweeted or not. Based on our results, we can say that the retweeting likelihood depends a lot on the topology of the retweet graph. Future work can be done by adding some behavioural metrics on top of the topological measures in order to estimate the number of retweets a user is likely to receive.

REFERENCES

[1] Zhang et. al (2013), "Understanding Re-tweeting Behaviours in Social Networks"

[2] Remy et. al (2014), "Information Diffusion on Twitter: everyone has its chance, but all chances are not equal", IEEE

[3] Goel et. al, "A note on modeling retweet cascades on Twitter"

[4] Bild et. al (2015), "Aggregate Characterization of user behavior in Twitter and analysis of the retweet graph", ACM

[5] Domenico et. al (2013), "The Anatomy of a Scientific Rumour", Nature Open Access, Scientific Reports 3, 2980

[6] Freeman L (1978), "Centrality in social networks conceptual clarification", Social Networks, Vol. 1, No. 3, pp. 215-239

[7] NetworkX documentation (2015), "NetworkX documentation — NetworkX 1.10 documentation"