**Anatomy of Twitter social graph**

566 Project Report Supervisor : Satya Siva Katragadda

**Dhana Lakshmi(dxv3492)**

**SaiSree Bhagi (sxb8547)**

## ABSTRACT

Today's computerized world has been simply erasing the distance between two individuals from anywhere on the globe through social networking sites like Twitter, Facebook etc. For identifying the proximity between different groups of users in Twitter, we are viewing the Twitter network system as a "Graph" defining users as nodes and connectivity between them by edges as their relationships. The main issue here is analyzing the structure of social graph of Twitter users, and drawing conclusions based on the results obtained after analyzing the graph.

## INTRODUCTION

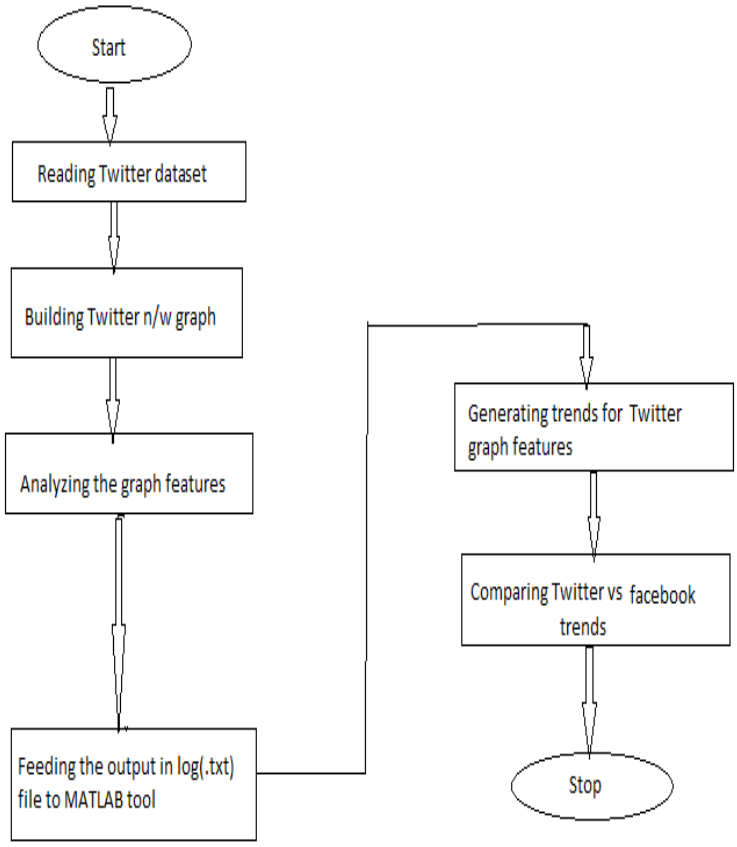
Twitter is an online social network used by millions of people around the world to stay connected with their friends, family members and co–workers through their computers and mobile phones. Dataset contains nodes and edges of graph which represents users and their relationship in the community. A study of social interactions within Twitter reveals that the driver of usage is a sparse and hidden network of connections underlying the “declared” set of friends and followers. We are dealing with numerous features of the graphs that includes the number of users and friendships which can be considered for studying "degree distributions" , minimum hop distance between any two users in the social network, average number of users and number of pairs of users per hop distance, detecting outliers, and finally the global structure of the Twitter network graph, which shows the connectivity between two individuals on a global scale. All these properties help to determine the structural patterns of Twitter network graphs and for correlating these results with the existing results of Facebook network system.

## WORKING MODEL

Implementation of the project is diagramatically shown in the figure and pseudocode below helps to understand implementation. Data is collected from Twitter for analyzing it's community structure. In the dataset, we are treating each node in the node list(.csv file) as user and edge in the edge list(.csv file) as a link between two users. After reading the dataset, we build the social network graph.

In the analysis we tested certain features of twitter community like average number of users per hop distance, degree distribution, fraction of pairs within one hop distance, outliers detection. "Degree distribution" indicates fraction of users who are associated at different outdegrees. Shortest path between two users or hoplength between two nodes is determined by using Dijkstra's shortest path. Through degree distribution we can determine the average number of users per hop distance. The final results of these features are taken as an input to MATLAB tool for generating twitter community trends. For making future proximity decisions and for building effective community structure, these analyzed results will be considered.

**Working Model talks about the anatomy of Twitter network**



**Input:** Nodes and edges of twitter network graph

**Output:** Twitter network graph and various metrics derived from it.

**Pseudocode**

Read nodelist and edgelist in dataset

Graph generation:

Graph<Integer, Integer> g; // Graph declaration

while(content != null)

{

for(each i) //nodes

{

add each node to g

}

while(content != null)

{

for(each j) //edges

{

Tokenize(edges) // stringtokenizer is used for separating two nodes

reading nodes into id2,id3;

if(id2 &&id3!=null)

{

add each edge to g

}

Degree distrubution

Iterate until vertices count starting with 0

{

Get vertices into a collection

}

Iterate until vertices count starting with 0

{

Get outdegree of all vertices into a collection

}

Iterate until the collection size

{

Now calculate collection frequent by:

Collection1.frequency(i);

Path length

Iterate until the collection size

{

DijkstraShortestPath<Integer,Integer> alg = new DijkstraShortestPath(g);

// detecting shortest path between the nodes

getPath(i, j);

Add paths into collections say (p)

Add size of p into another collection

}

Numbers of pairs within the hopdistance

Iterate until the collection size

{

DijkstraShortestPath<Integer,Integer> alg = new DijkstraShortestPath(g);

// detecting shortest path between the nodes

getPath(i, j);

Add paths into collections say (p)

Add size of p into another collection say(k3)

}

for ( int m1=0;m1<k3.size();m1++)

{

k4.add(Collections.frequency(k3, m1));

}

Average number of friends

Iterate until vertices count starting with 0

Get vertices into a collection

Iterate until vertices count starting with 0

Get outdegrees of all vertices into a collection

for(int n1 = 0; n1<collection.size();n1++)

{

sum += ((ArrayList<Integer>) k).get(n1);

}

Get collection size

Divide the collection sum by size

End

## Tools used:

JUNG AND GEPHI

Most of analysis is done in JUNG that provides all Graph API's . For visualization purpose gephi is used. Because Gephi is an open-source tool containing all JAVA packages that provides ease for visualization.

## METHODOLOGY

We used outdegree calculations for finding degree distrubutions and used betweeness centrality and dijkstra's algoritm for implementing number of user pairs that fall within particular hop distance. We calculated average of all the outdegrees for finding average number of friends a user in the twitter network and for detecting outliers.

### Steps for implementation:

1. **Reading data:** Reading the data from the csv files

**2. Generation of graph from the given data:** This is one of the important task in the project and it can be done using the JUNG API in java

**3. Visualization of Graph:** This is done by extracting subgraph or part of graph from the main graph and it can be visualised in either gephi or jung

4. **Performing operations on graph:**

1. **Degree distrubution:** A fundamental quantity measured repeatedly in empirical studies of networks has been the degree distribution pk . The degree k of an individual is the number of friends that individual has, and pk is the fraction of individuals in the network who have exactly k friends. We computed the degree distribution of twitter users

2.**Calculating number of pairs that fall under that particular hop distance :**

When studying a network's structure, the distribution of distances between vertices is a truly macroscopic property of fundamental interest. Formally, analyzing the graph describes the number of pairs of vertices ( u, v ) such that u is reachable from v along a path in the network with h edges or less. Here we measure them and plot what percentile of vertex pairs are there in any hop distance.

3.**Calculating average number of friends for all users:** We calculated the average number of friends for all the users in the network graph.

4.**Calculating the pathlength:** Pathlength determines the distance between two users.Through this, we can understand how much distance a user can be reached from every other users.

5. **Using Matlab for Twitter trend analysis:**

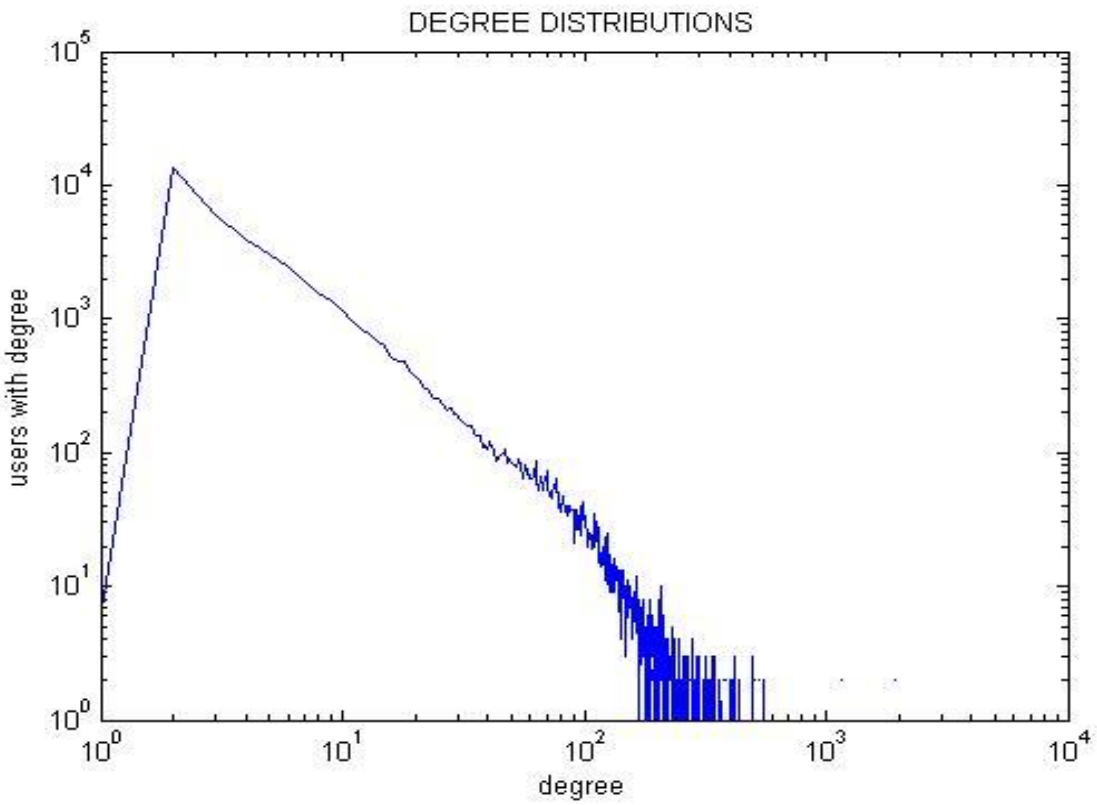
Output generated from the above features are taken in the form of log files. These are then given as an input to the matlab and different graphs like number of people with particular number of friends (degree distrubutions) and number of pairs that fall in particular hop distance are analyzed using graphs.

6. **Compare Results with facebook:** After analyzing all the structural patterns of Twitter network graphs, we compared those trends with facebook trends for making future proximity decisions for Twitter network.

## RESULTS:

### 1. Degree distrubutions:

A fundamental quantity measured repeatedly in empirical studies of networks has been the degree distribution .The degree k of an individual is the number of friends that individual has, and pk is the fraction of individuals in the network who have exactly k friends. We computed the degree distribution of twitter network system for 50,000 users. The degree distributions are shown in Fig. 1, displayed on a log-log scale. With x-axis showing the degrees and y-axis showing number of individuals with that degree The distribution is nearly monotonically decreasing which is similar to facebook, except for a small anomaly near ‘2 ‘friend count which increases rapidly. This might be due to analyzation of small number of users. This is similar to facebook that facebook has this small anomaly at 20 friends to encourage low friend count individuals in particular to gain more friends until they reach 20 friends. The distribution shows a clear cutoff at 5000 friends, a limit imposed by Facebook on the number of friends at the time,but we cannot give any cutoff for the twitter in our case as only 50,000 users are analysed.. The results of twitter are also similar to that of Facebook in case of degree distribution . In twitter we have lot of users with single digit friends but there are some of the users who reported exceptionally high number of friends like 4485. Reflecting most observed social networks, our social relationships are sparse. Indeed, most individuals have a moderate number of friends on Facebook, less than 200, while a much smaller population have many hundreds or even thousands of friends




. Figure1

**2.Number of pair of vertices with in particular hop distance**

When studying a network's structure, the distribution of distances between vertices is a truly macroscopic property of fundamental interest. The graph describes the number of pairs of vertices ( u, v )such that u is reachable from v along a path in the network with h edges or less.

After analysing the twitter network graph we have the result in form of a graph showing hop distance on x-axis and count of pairs of vertices that can be reached with the hop distance on y-axis and

By analysisng this graph we can say that , highest number of users can reach other users with in a hop distance of 2 We find that the average distance between pairs of users was 4.7 for Facebook users and path lengths between individuals, the so-called "six degrees of separation" foundby Stanley Milgram's experiments on a global scale are seen in facebook at that time of computing metrics Here in twitter we an say that the degree of separation to the maximum is 6 and any user can reach other within 6 edges of seperation to the maximum ,In this way the results correlate with that of facebook.

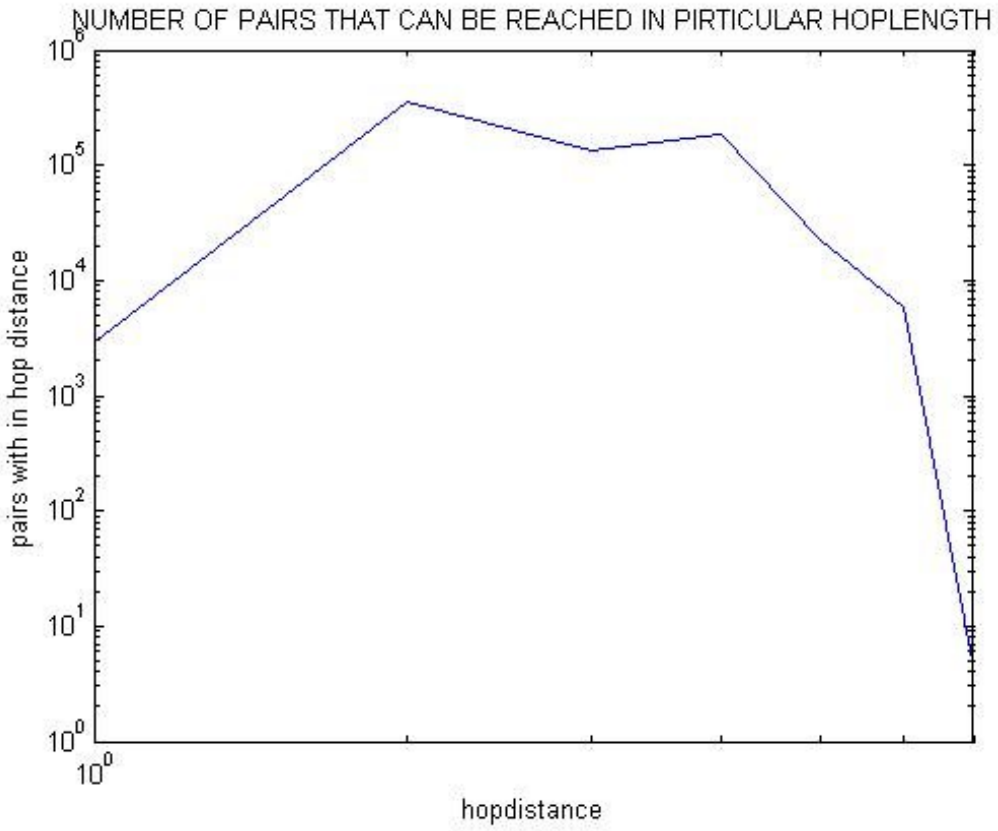



Figure2

3.**Average number of friends:**

Here by analyzing the twitter graph for 50000 users we have average number of friends for users as 23 and for facebook it is nearly 200 .By analyzing the number of friends for each user we have an interesting results like some of users have exceptionally high number of friends we can call these users as outliers and example for that is number of friends count of pirticular user in twitter graph is '4485' which is nearly equal to highest number of friends count in facebook.

**4. Path length between two users in a sub graph:**

**Sub-graph :** A "sub-graph" is a graph that is extracted from the main graph. In this case, we have generated a sub graph for 1000 nodes from the entire dataset.

**Path length:** It is defined as a distance between any (two nodes )that is any two users in a twitter social network.It is found using dikstras algorithm.

Suppose in our case, we generated a sub-graph with 1000 nodes and edges. The distances between every two users within this sub-graph is shown in the (diki.txt) file attached with this

report. If there is no path between any two nodes, then path length will be "0". Based on output in the (diki.txt) file, we can say that the most frequent path-length between two nodes in this sub-graph be 2. This value might vary from sub-graph to sub-graph.

**Average path-length of sub-graph:** It is defined as the average of all the path lengths between all the users

The average path length for the above mentioned sub-graph be 1.9955035 which is nearly equal to 2. This represents that distance from one user to other is nearly 2,which says that on an average any users can reaach other user within a distance of 2 if the path exits between them.

**Sample output:** This represents the path length of "user 1000" to reach any of his previous 20 users within his vicinity in a particular region(sub-graph).

path length from n1000 to n981 is:
0
path length from n1000 to n982 is:
2
path length from n1000 to n983 is:
4
path length from n1000 to n984 is:
0
path length from n1000 to n985 is:
2
path length from n1000 to n986 is:
2
path length from n1000 to n987 is:
2
path length from n1000 to n988 is:
2
path length from n1000 to n989 is:
2
path length from n1000 to n990 is:
2
path length from n1000 to n991 is:
2
path length from n1000 to n992 is:
2
path length from n1000 to n993 is:
2
path length from n1000 to n994 is:
2
path length from n1000 to n995 is:
2
path length from n1000 to n996 is:
2

path length from n1000 to n997 is:
0
path length from n1000 to n998 is:
2
path length from n1000 to n999 is:
2
path length from n1000 to n1000 is:
0

For this sample output, most frequent path length be 2 and least frequent path length be 4

## FUTURE WORK

In future work,different metrics like degree correlation,Site engagement correlation, Login correlations and more advanced techniques like mixing patterns, normalized country adjacency matrix can be computed for twitter network graph.The final statistical results of the Twitter network system we got are then correlated with the existing Facebook network results for making future proximity decisions about the community's distribution and users communication relationship in Twitter network system.

## Note:

Due to huge dataset, we are getting memory issue when testing our model for the entire dataset. So results analysed for 50,000 nodes are provided.

## REFERENCES

1.boyd dm, Ellison NB (2007) Social network sites: definition, history, and scholarship. Journal of Computer-Mediated Communication 13: 210–230.

2. http://arxiv.org/abs/1111.4503

3. http://web.ibs-b.hu/research

4. http://research.microsoft.com/pubs/122433/YardiICWSM.pdf

5.Wasserman S, Faust K (1994) Social Network Analysis. Cambridge: Cambridge University Press.

6. Adar, E. (2006). GUESS: a language and interface for graph exploration. In Proc. SIGCHI '06. Montreal, Canada.

7.The Anatomy of the Facebook Social Graph by Johan Ugander, Brian Karrer, Lars Backstrom, Cameron Marlow

8. http://en.wikipedia.org/wiki/Social_network_analysis

9. http://en.wikipedia.org/wiki/Social_graph

10. http://jung.sourceforge.net/

11. https://gephi.org/

12. https://www.youtube.com/watch?v=Ges_y5lHgs ) (part1,part2)

13.Bonacich P (1987) Power and centrality: a family of measures. American Journal of Sociology 92: 1170–1182.

14. Kumar R, Novak J, Tomkins A (2010) Structure and evolution of online social networks. Link

15. http://web.ibs-b.hu/research

16. http://research.microsoft.com/pubs/122433/YardiICWSM.pdf

Distance dead or alive, Online Social Networks from a geography perspective by IBS

17. http://www.mathworks.com/help/matlab/ref/importdata.html