Twitter and Facebook Social Network Analysis

Harshita Singh, Likhita Navali, Prof. Bhaskarjyoti Das
PES University
BSK IIIrd Stage, Bangalore

I.Abstract

Social media has become an integral part of today's world. One of the reasons for social media gaining such popularity, is its ability to connect people worldwide with little to no hindrance. As more and more users begin interacting on social media, following other users, liking posts,etc., the network that represents the users and their interactions grows more complex. With this increasing complexity, increases the need to understand these networks and draw inferences. Facebook, particularly, gets an average daily traffic of 1.5 billion active users, constantly generating new relations and interactions. On an average, 6000 tweets are sent every second, 500 million a day and per year, an average of 200 billion tweets are sent among users. Analyzing all that data, though tiresome, can be extremely informative because social networks have a critical role to play in the economic, social, educational and health aspects of our day to day life and the manner in which we conduct ourselves in general.

In this paper, we look at twitter followers and interactions between twitteratis and facebook follower network and try various analysis techniques to better our understanding on the networks, identify important nodes and predict or suggest possible future connections among users. Here, we analyze three different datasets. First dataset looks at the follower network of one particular user, the second dataset looks at interactions between various twitter users, in the form of mentions and retweets, on the topic of elections. The third dataset depicts a facebook network of users and how they are interconnected.

II. Introduction

Twitter, a microblogging site, is ranked 10th world-wide. Twitter has often portrayed itself as a dominant supplier of facts for breaking news events. Twitter's rapidly growing impact and reach has encouraged many researchers to look further into the type of information that it captures.

Facebook is rapidly attracting multitudes of visitors every month instigating a shift in communication. This change consequently presents that societies are choosing to become part of the popular Facebook culture for various reasons, such as its renowned opportunities for keeping in touch with current social circles, reunifying long lost family and friends and broadening prospects of finding new companions. Facebook removes some of the barriers that may limit our regularity of communication with people, upholding the geographic differences, social class, busy lifestyles and economic factors that may usually discourage us from regular contact.

In this paper, we analyze twitter and facebook data to better understand people and their social networks in hopes of leveraging this information in multiple scenarios. Section III talks about previous research that has been conducted by scholars on the subject addressed by this paper. Section IV talks about the data and approach that was followed throughout the duration of this study and section V depicts the results obtained for the same.

III. Existing Work

The amount of data that social media generates can be computationally overwhelming for current hardware and software devices. GraphCT[1] helps analyze large quantities of unstructured social network data. In [1], GraphCT is used to analyze public data obtained from twitter. It detects clusters of conversations, detects and ranks actors to help focus on smaller subset of data.

It was found that twitter's public data stream forms a tree like structure which is analyzed using GraphCT to deduce interesting characteristics of interactions between the users. User interaction graphs are created based on mentions. Duplicate mentions are removed. The broadcast tree contains multiple nodes, some re-iterating information provided by other nodes (via retweets, etc.), while some nodes generate original data. The goal is to identify the nodes that generate new data so that analysts can focus on important interactions. The data is collected from a web and social media indexing service Spinn3r and the analyses are evaluated on Twitter updates aggregated by this site for H1N1 outbreak and Atlanta floods of September 2009, as shown in Fig 1. It was found that there was direct correlation between public health information and social communications.

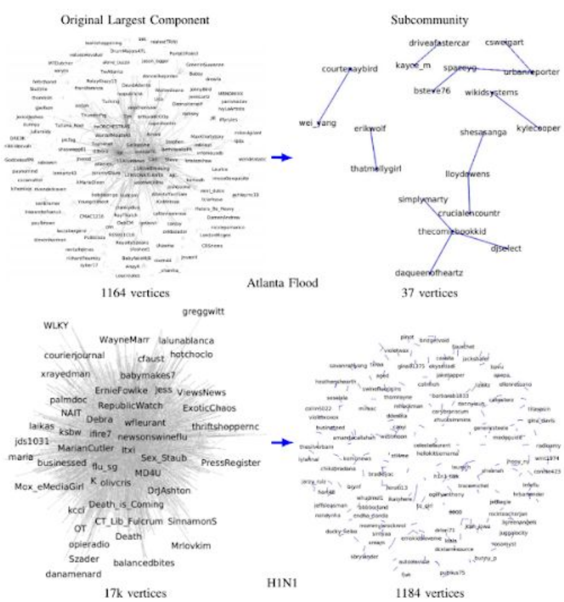

Figure 1.

An observation of these datasets shows that there are relatively few high-degree vertices. This follows the power-law distribution principle. One specific use case detected from the graphs is news dissemination.

This dataset however, is too small to completely scope out GraphCT's capabilities. The paper concludes that GraphCT and Cray XMTs facilitate analysis of datasets that were previously considered too massive.

Paper [2] describes a knowledge based framework for analysis detection of the critical events in disasters using information from social media.The two major modules are information analyzer module and event analyzer module.

Fig. 2 shows the framework of information analyser module.

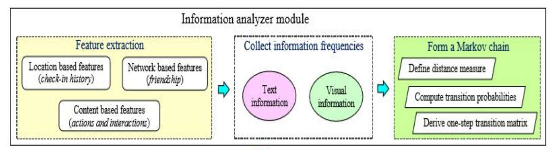

Figure 2.

It performs the tasks of extracting three different features of social network information such as location based feature, network based feature and content based feature. The location based feature contains the check-in history of users. The network based feature is concerned with social friendship activities information.

The number of text messages containing terms of interest are aggregated into hourly basis along with visual information containing the regions of interest.

Fig. 3 shows the framework of event analyser module.

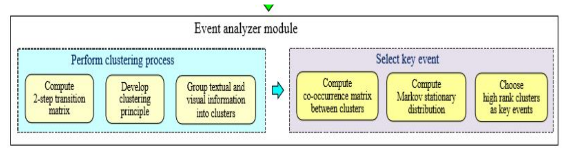

Figure 3.

This module performs the clustering process of the most frequent text messages and visual information by establishing a Markov chain based clustering principles.

The proposed framework has demonstrated its ability to provide useful situation awareness information by integrating two commonly used social networks Twitter and YouTube. In this experiment, pairs of visual and textual information from YouTube and Twitter are collected empirically. The model relates the visual image to the textual information.

Paper [3] deals with a graph representation of Twitter information and compare it with the Web graph. A follow edge from user u1 to user u2 exists if u1 follows the posts of u2. A publish edge from user u1 to post p1 indicates authorship of the post.

The paper deals with computation of the power-law exponents for the inlink and outlink distributions of retweet and follow links, using a nonlinear least-squares (NLLS) algorithm by Marquard Levenberg implemented as part of the gnuplot package.

Link semantics :

A retweet link is also expected to signify an endorsement of quality, however in different roles. User a will retweet the posts of user b if he either is interested in writing about the topic or expects his readers to be interested in this post. Thus a retweet edge signifies a connection from user a as a writer to user b as a writer. We expect this link to carry both an endorsement of quality and that of relevance, and thus carries a stronger topical signal.

IV. Methodology and Dataset

1.Twitter Data Analysis

a.Building Social Network

Twitter API is used to extract the data of ego-network i.e. the nodes followed by self, and first-degree 'Follows' of those nodes. This Graph of (Vertices, Edges) is used to build an adjacency matrix.

b.Identify influential friends using 'PageRank' formulation.

From the adjacency matrix, we can create a column-stochastic matrix (aka Markov transition matrix in random-surfer model) such that, a column with m outlinks will have 1/m as value in respective m cells.

On applying Transition-matrix transformation iteratively on PageRank vector, vector will eventually converge such that: Matrix.Vector = Vector. Equivalently, this is eigen-vector formulation with PageRank vector being the principal eigenvector corresponding to eigenvalue 1

c.Identify implicit clusters

Ideally, the number of clusters are decided using a plot of within-cluster sum of squares of distances vs number of clusters. Here for simplicity, we use a simple heuristic to fix the number of clusters in advance (~ 10 clusters)

d.Recommend new friends to follow

The clusters formed in previous section is used to recommend new friends. The top nodes in the network is also considered to recommend new friends.

2.Facebook Data Analysis

The dataset is found from [4].We read in the file and construct the Graph.The network consists of 4,039 nodes, connected via 88,234 edges.

The analysis includes calculating the betweenness centrality, eigenvector centrality, degree centrality.

V. Results and Analysis

The twitter data is extracted which involves getting a specified number of tweets. This dataset is used to analyze the number of inlinks, ie., the number of interactions that other users initiate with the user under observation, as shown in Fig 4

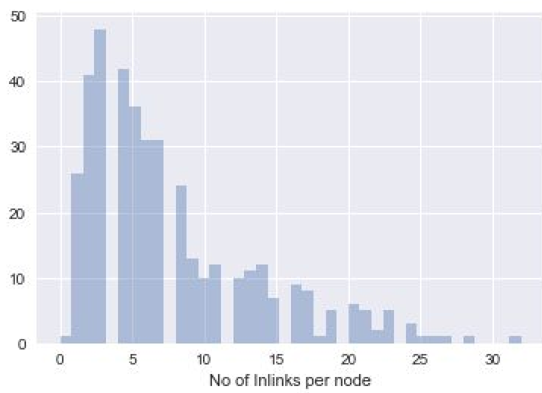

Figure 4.

Fig 5 shows details about the top 10 influential nodes in the network that we observed.

FullName	PageRank	UserName	Inlinks	Outlinks	Followers	Friends	Location	Created
Joe Gebbia	0.002475	jgebbia	1	0	88849	865	San Francisco	Dec-2007
Lada Adamic	0.002475	ladamic	1	0	10551	206		Mar-2007
Nathan Yau	0.002475	flowingdata	1	0	73688	242	California	Mar-2008
Chris Wanstrath	0.002475	defunkt	1	0	48916	5540	San Francisco	Jan-2007
John Foreman	0.002475	John4man	1	0	13762	372		Nov-2011
ML @ REDDIT	0.002475	mxlearn	1	0	27182	3855		Jul-2010
Steven Levy	0.002475	StevenLevy	1	0	108725	767	New York City	Mar-2007
Seth Godin	0.002475	ThisIsSethsBlog	1	0	664999	1	New York	Dec-2008
Clayton Christensen	0.002475	claychristensen	1	0	169198	165	Boston, MA	Jul-2009
John Hagel	0.002475	jhagel	1	0	32598	1585	Silicon Valley	Mar-2008

Figure 5.

The influential nodes are decided based on pagerank which is calculated using the formula

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

Where, d is the damping factor, PR(T1) through PR(Tn) are pageranks of all pages having a link to page A and C(T1) through C(Tn) is the count of all outgoing links for page 1 to n, respectively.

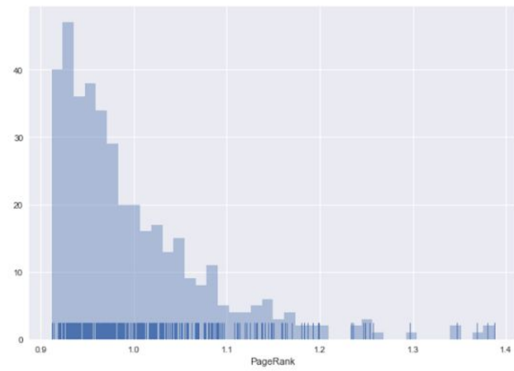

Figure 6.

We calculated PageRank for nodes in social-graph,then we calculate recommendations on the basis of top-ranked nodes in the graph. E.g. To get 20 recommendations, after looking at friends of top PageRank scoring nodes in my network

Three ways to discover new friends:
1.After looking at top nodes in full ego-network some of the friends suggested are as shown in Fig 6.

FullName	Freq	UserName	Followers	Friends	Location	Created
Barack Obama	13	BarackObama	103378627	617663	Washington, DC	Mar-2007
Stewart Butterfield	12	stewart	82392	2720	West coast	Sep-2006
Robert Scoble	11	Scobleizer	414852	53684	Campbell, CA	Nov-2006
Anil Dash □	11	anildash	590589	421	NYC	Dec-2006
Medium	11	Medium	2301868	94	San Francisco, CA, US	May-2012
Nat Friedman	10	natfriedman	38525	2257	San Francisco	Feb-2008
Farhad Manjoo	10	fmanjoo	171163	4333	California, USA	Mar-2007
jack	10	jack	4109510	3731		Mar-2006
Dave McClure	9	davemcclure	358417	18011		Jul-2006
Michael Arrington	9	arrington	228629	2017	Palo Alto, CA	May-2009
Alexis Ohanian Sr. 🚀	9	alexisohanian	245496	4423	Worldwide	Mar-2007
John Doerr	9	johndoerr	286903	383		Feb-2009
Esther Dyson	9	edyson	60608	1692	NYC, but usually elsewhere	Aug-2008
Jeff Bezos	9	JeffBezos	702023	0		Jul-2008
Clay Shirky	9	cshirky	343170	826	New York, NY	May-2007
michael_nielsen	8	michael_nielsen	35285	2723	San Francisco, CA	Jul-2008
Brad Feld	8	bfeld	305694	613	Boulder, CO	Apr-2007

Figure 6.

2. After looking at Data-Science clusters (Spectral clustering) some of the friends suggested are as shown in Fig 7.

FullName	Freq	UserName	Followers	Friends	Location	Created
aileenlee	14	aileenlee	54404	2639	palo alto, ca	Nov-2008
Techmeme	13	Techmeme	376862	892	San Francisco, CA	Mar-2007
Peter Fenton	13	peterfenton	44173	1319		Apr-2007
Aaron Levie	13	levie	2493593	435	Palo Alto	Mar-2007
M.G. Siegler	13	mgsiegler	188836	991	San Francisco, CA	Jan-2007
Alexia Bonatsos	12	alexia	173245	3280		Dec-2008
jack	12	jack	4109510	3731		Mar-2006
megan quinn	12	msquinn	54557	951	California	Jul-2008
John Doerr	12	johndoerr	286903	383		Feb-2009
Mitch Kapor	12	mkapor	120382	652	Oakland	Mar-2007
Liz Gannes	12	lizgannes	71400	2940	San Francisco	Feb-2007
Ryan Sarver	11	rsarver	251471	2953	San Francisco, CA	Feb-2007
Eric Ries	11	ericries	298231	1520	SF	Apr-2008
Dave McClure	11	davemcclure	358417	18011		Jul-2006
Christopher Mims 🕸	11	mims	82257	5815	Baltimore, MD	Mar-2007
Chris Messina	11	chrismessina	99547	5428	San Francisco, CA	Jul-2006
dick costolo	11	dickc	1611114	726	San Francisco	May-2007

Figure 7.

3.After looking at Design cluster (Spectral clustering) some of the friends suggested are as shown in Fig 8.

FullName	Freq	UserName	Followers	Friends	Location	Created
ashton kutcher	3	aplusk	18019499	778	Los Angeles, California	Jan-2009
jack	3	jack	4109510	3731		Mar-2006
Ron Conway	3	RonConway	96167	65		Jun-2009
Techmeme	3	Techmeme	376862	892	San Francisco, CA	Mar-2007
Square	3	Square	244551	840		Nov-2009
Erick Schonfeld	3	erickschonfeld	75292	1375	New York	Jan-2008
Dave McClure	3	davemcclure	358417	18011		Jul-2006
Kevin Rose ⛩	3	kevinrose	1639998	915	San Francisco, CA	Jan-2007
John Doerr	3	johndoerr	286903	383		Feb-2009
Peter Thiel	3	peterthiel	197086	0	San Francisco	Jun-2009
Aaron Levie	3	levie	2493593	435	Palo Alto	Mar-2007
Tim Cook	3	tim_cook	10929280	60	Cupertino	Jul-2013
Jeff Keni Pulver	3	jeffpulver	488553	39645	New York	Feb-2007
Greylock Partners	3	GreylockVC	195460	801	Silicon Valley	May-2009
Michael Arrington	3	arrington	228629	2017	Palo Alto, CA	May-2009
Farhad Manjoo	3	fmanjoo	171163	4333	California, USA	Mar-2007
marissamayer	3	marissamayer	1656617	350	San Francisco, CA	Nov-2008

Figure 8.

From the second dataset, that consists of a thousand recent most tweets regarding elections. This dataset is fetched using the tweepy library api and contains details such as, user handle, text, location, favourites count, hashtags used, etc.

These tweets are used for analysis. A directed graph is constructed. The edges in this graph represent the interactions between two nodes where the nodes are users. Interactions between users is represented in the form of 'retweets' and 'mentions' i.e., If a user retweets another user's tweet, or mentions a user's twitter handle in his tweet, an edge is introduced between the users.
All analysis like measuring the betweenness centrality, degree centrality, finding sub communities are performed on the data. Fig 9 represents the ego network of the twitter dataset.

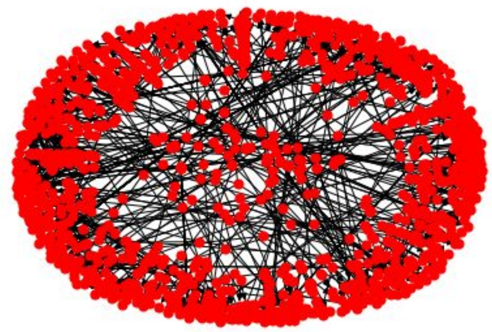

Figure 9.

The graph is further analyzed, the connected components are isolated and the number of connected components is plotted against the component size as shown in Fig 10. This plot follows a power-law degree distribution, ie., fewer number of components have a larger size and a larger number of components have a smaller size.

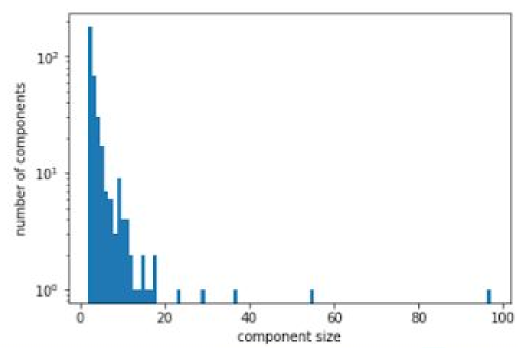

Figure 10.

The third dataset is from facebook, It holds nodes of people using facebook and the edges represents friendship between nodes. We were able to get the top ten most influential people among the people present in the dataset. Fig 11 shows the ego network for the used dataset.

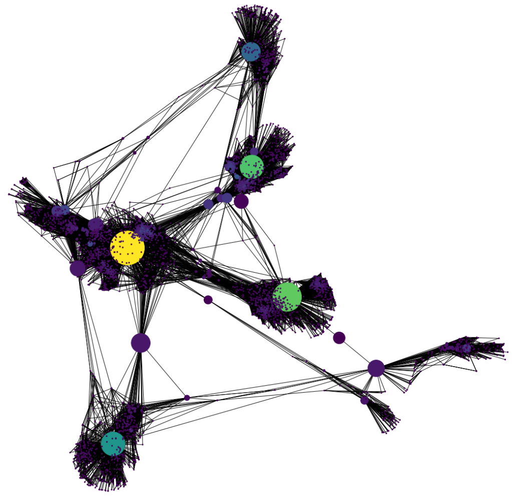

Figure 11.

The size of nodes in the figure depict the betweenness centrality of the nodes. Higher the betweenness centrality, larger the node. The color of the nodes changes as the degree of the nodes changes. Fig 12 shows the ego network of the node with the highest betweenness centrality.

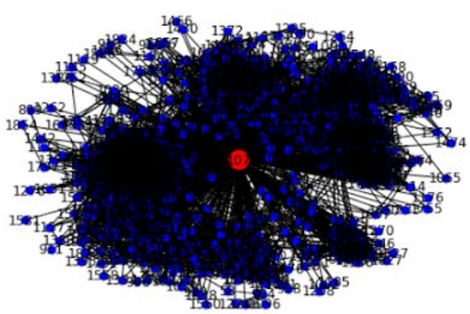

Figure 12.

VI. Conclusion and Next Steps

Centrality Measures can help us in identifying popularity, most liked, and biggest influencers within the network.

Eigenvector centrality is a measure used to decide whether a node is important if it is connected to other important nodes.

We performed a preliminary, empirical evaluation on a small data set to give some insight into the characteristics of the links.

In the future, the link prediction can be enhanced to provide more accurate suggestions. There is still a large amount of untapped information that can be gathered by further analyzing the data. Aggregating data from multiple sources may provide better insight into people's social lives and behaviors.

VII. References

[1] Massive Social Network Analysis: Mining Twitter for Social Good, David Ediger, Karl Jiang, Jason Riedy, David A. Bader, Courtney Corley, Rob Farber, William N. Reynolds; USA

[2]Knowledge based Social Network Applications to Disaster Event Analysis: Thi Thi Zin, Member, IAENG, Pyke Tin, Hiromitsu Hama and Takashi Toriu

[3] Topical Semantics of Twitter Links: Michael J. Welch, Uri Schonfeld, Dan He, Junghoo Cho.

[4]Snap.stanford.edu. (2018). *SNAP: Network datasets: Social circles*. Available at: https://snap.stanford.edu/data/egonets-Facebook.html