RESEARCH ARTICLE OPEN ACCESS

Identification and Assessment of Social Subgroups or Community Structure on Social Networking Websites

Ravi Prakash Verma, Dr. Deepak Arora
Department of Computer Science & Engineering, Amity University, Lucknow, India

Abstract
Social Networking Analysis is gaining more popularity day by day due to its major contribution towards business marketing analytics and future decision policies for any large organization. These kinds of assessments have their potential advantages towards designing future business strategies for a particular product/object available in the market. The social media is now open for all kind of blogs, independent views and open communities. People are free to express their ideas or opinion for any specific product. Now social media has become a widely accepted platform to establish an opinion set for a particular product. The author have described various strategies to gather data from these social communities and try to find the specific interest group/subgroup for any product or person or any object. In this paper, the authors have used an open source tool NodeXL for analysis of twitter tweets and find out most effective group or community individually. This analysis is targeted on the factor of centrality and clustering coefficient analysis.

I. INTRODUCTION

Now a day's social network site have become very famous among the internet users. A user makes himself available and stay connected with other online users by using various mode of application. Among the available OSNs, it becomes the question of choice whether user is using a network as per the popularity and functionality of site. There exists a web of users by connecting people online. Due to the presence of online users, a huge number of transaction get executed over the internet on OSNs site at each single second, which is a very tedious task to manage on any site. In the scope of robustness, scalability, optimization Facebook one of leading OSN has posed several challenges.

Users of OSN have the prime objective to make new linkups. Every time users are using application, application provide some boundaries to restrict users accessing others private information that remain present in database. Every time with changing links an occurrence of any event due to users activity, there is an interaction with applications database to understand the boundaries with respect to different part of application such as user profile, wall post, group post etc. [1] Here the data has been collected and its structure has been analyzed with the help of graph model. Where node represent user and link represent the relationship boundary among the connected users.

This paper is represented as follows: Section II is described in brief about the representative of Social Network Analysis's related field. Section III discuss about tool to analyze and visualize the network. In section IV we will discuss about Experimental Setup, that how process is being carried and thereafter proceed to section V discuss about result analysis. And section VI contains the summary to conclude this paper and discussion about future work.

II. RELATED WORK

Salvatore et al. [2] presented the work about the data that describes the relationship among people connected with each other on social networking site and proposed method to collect the social network data, and its evaluation. Data was taken from Facebook, they crawled the two samples comprised of millions of peoples data, can be represented in graph form which is an undirected graph. They discussed about so many social network analyzing tools, which are capable to evaluate specific properties of social community graph (i.e. incident edges and vertices, edges incident to other, centrality values, and graph metrics). Social network on web are highly secured, Users on social network are not able to access others confidential information, and relationship data, so they have used the concept related to the graph theory for extraction of data. In graph theory they have notified users as node and relationship as ties called edges. They discussed about two sampling methodologies named of BFS and Uniform sampling. Emilio et al. [3] have proposed his worked on understanding of social activities directed by human like communication, and relationship between them. They worked on the specific properties of groups, community structure of social network, and the nature of communities to solve their purpose. They took two large samples of million people and relationship links between them. They discovered some features of communities describe some patterns, which are similar or dissimilar among

people, which are formed on the basis of different clustering techniques. They have given framework to evaluate social relations, helping to build a social structure. Mislove et al. [4] worked on most popular data sharing social media sites on web like Flickr, YouTube and Orkut. These sites enable users to share their thoughts, contact and media content. The study of these popular websites gives the understanding of social network and it's characteristic, its study can be used to improve the current topology of social networks by concerning the graph and its metrics. They have taken the sample data in their work of these popular websites, containing the 11.3 million peoples and 328 million links. They have examined the graph of social networks provides the facility to host user's content may belong to other users. They proposed a solution to understand, how a content gaining popularity on online social networking site. Daqing Zhang et al. [5] worked on social community intelligence (SCI), its purpose to investigate the social network activities and group behaviors. SCI enables to extract user's behavior and apply different mining methodologies to find the associated patterns among communities. It notices the daily interaction and behavioral activities among groups. SCI also works on some techniques like semantic analysis, clustering and graph matrices.

III. SOCIAL NETWORK ANALYZING TOOL OVERVIEW

In a market a number of tools are available for analyzing network and visualization. For e.g.:- UCINet, NetDraw which comes with GUI and give user a relief from extensive code writing. These tools can be used for creating the network. The NodeXL tool can be used easily by integrating it with Excel 97 – 2013. This tool provides an option of visualizing the network in spreadsheets. NodeXL comes with template workbooks through which graph can be obtained ones the data value is uploaded into spreadsheet. Its architecture contains three basic modules as follows:

1. Importing data: As discussed above NodeXL comes with template workbook that feeds the value given by user in order to generate network charts. In this data can be fed by a number of ways that include existing spreadsheets, files through value that are separated using comma. Best way to use spreadsheets as it can be easily exported to other system for reuse.

2. Analyzing Network: As considering then graph, prime concern is a network it represents. A network contains a number of nodes that are commonly act as entity and called as vertex in a graph, these vertex are connected by using links that is called as edges. The edges are responsible for indicating the relationship existing among the two connected nodes. With the help of analyzing tool included in NodeXL statistics regarding nodes can be calculated by using information such as in-degree & outdegree.

3. Layout of Graph: As canvas has been provided by NodeXL in order to manipulate the entered value and displaying the network charts. It is responsible for display of connecting links such as its color, thickness which depends on data attributes and user specified parameters.

IV. EXPERIMENTAL SETUP

The base of NodeXL is MS Excel sheet, in which it works being integrated with excel workbook 2007/2010/2013. That forms the data in structural form to analyze the network by visualizing data in graph form. It consist of some worksheets named of 'edge', 'vertices', 'cluster (the probability of a vertex to be clustered with its neighbor vertices)' and Overall Metrics. The advancement of NodeXL generally starts from extraction data from social media site like Twitter, YouTube, and Facebook etc. there are some of the steps are taken to result a graph of network's extracted data to gather useful information about network. Different phases to set the experiment are as follows:

Importing data from Social Networking Site: In social media site, the people often comment and share their views regarding each other, these information are stored in the data repository store in the form of text data. No matter that, from where these data was generated. As the data imported, the data is entered into the edge worksheet, with the representation of relationship among multiple users. The edges can be remarked by date stamp (i.e. from when the relationship was established).

Data cleaning: This process involves the removal of duplicate edges and redundant information from the data (if cleaning is required). It checks for the edges caused for multiple relationship redundancy, and remove them by aggregation into an edge.

Evaluation of graph metrics: NodeXL provides the framework to evaluate no. of network properties in term of node degree (In or Out), centrality (Closeness, Betweenness, and eigenvector will be discussed later in result Analysis section.), Page rank, Group Metrics, and overall metrics. Which help to have clear glance of network's insight detail.

Cluster formation: The nodes of network form cluster on the basis of attributes are being shared among them. The algorithm works behind in NodeXL to form the cluster between nodes together. Every cluster possesses its own attribute like color, shape and size.

Creation of sub graph images: NodeXL analyze the network data and form the graph, which is too large is made of number of small group's graph, this tool evaluate to each group and generate its respective graph called sub graph. And the sub graph images can

be saved separately in file by using sub graph image option given in NodeXL.

Expand the worksheet: Graphical attributes of a graph nodes are the color, size, opacity, shape can be auto filled to form graph, which is mapped with its relative attributes. NodeXL provide the number of shape can be used to change the appearance of nodes, and attributes are well scaled with attributes characteristics.

Show Graph: This function enable NodeXL to visualize the graph in the document action part. There are some of the function are also allied, which helps to alter or change layout of graph. NodeXL contains the number of algorithm to set lay out are following: Harel-Koren Fast, Fruchterman-Reingold, Circle, Spiral, Grid and Polar etc. Dynamic filter is one of the features to hide the nodes and edges, decided on the basis of attributes, only those will be visualized, whose attribute value is greater than the threshold value.

V. RESULT ANALYSIS

We experimented with data extracted from Twitter using "Foreign Affairs" as string in user's tweets by NodeXL. We gathered about 232 tweeting vertices (users) on the day of experiment. NodeXL has the capacity to extract 18000 tweets, the results are as discussed. There are about 30 groups are being formed and expressing their views about foreign affairs in their tweets are shown in figure 1, which shows SNA Map (Social Network Analysis Map) regarding foreign Affairs tweets laid out using the Harel-Koren Fast Multiscale layout algorithm, graph generated through NodeXL.

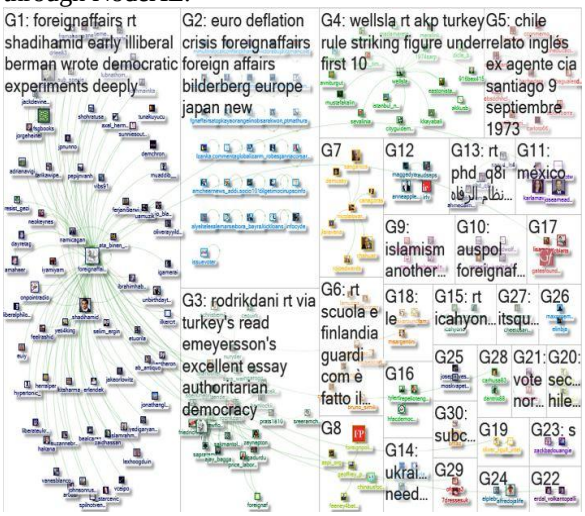


Figure 1: SNA Map of formed Groups discussing about 'ForeignAffairs' in their tweets

The edge colors are based on edge weight values. The edge widths are based on edge weight values. The edge opacities are based on edge weight values. The vertex sizes are based on follower's values. The vertex opacities are based on follower's values.

	Graph Metric	Value
2	Graph Type	Directed
4	Vertices	232
6	Unique Edges	278
7	Edges With Duplicates	28
8	Total Edges	306
10	Self-Loops	42
12	Reciprocated Vertex Pair Ratio	0.008
13	Reciprocated Edge Ratio	0.015873016
15	Connected Components	51
16	Single-Vertex Connected Components	31
17	Maximum Vertices in a Connected Component	119
18	Maximum Edges in a Connected Component	202
20	Maximum Geodesic Distance (Diameter)	6
21	Average Geodesic Distance	2.506148
23	Graph Density	0.004702194
24	Modularity	0.66361

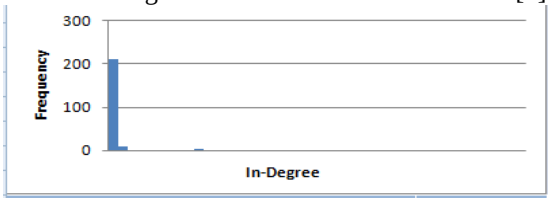
Figure 2 : Overall Matrices of Complete SNA Map

In Figure 2, overall Matrices shows the detailed information about the network formed regarding foreign affairs Tweets, in term of graph theory. The group G1 is the largest group among all thirty groups as shown in Figure 3.

	A	K	L	M	N
1		Graph Metrics			
2	Group	Vertices	Unique Edges	Edges With Duplicates	Total Edges
3	G1	71	93	12	105
4	G2	31	28	6	34
5	G3	25	59	6	65
6	G4	17	16	0	16
7	G5	10	10	0	10
8	G6	7	6	0	6
9	G7	7	6	0	6
10	G8	5	4	0	4
11	G9	4	4	0	4
12	G10	4	4	0	4
13	G11	4	3	0	3
14	G12	4	3	0	3
15	G13	4	4	0	4
16	G14	3	3	0	3
17	G15	3	1	2	3
18	G16	3	2	0	2
19	G17	3	2	0	2
20	G18	3	2	0	2
21	G19	2	1	0	1
22	G20	2	2	0	2
23	G21	2	2	0	2
24	G22	2	1	0	1
25	G23	2	1	0	1
26	G24	2	1	0	1
27	G25	2	1	0	1
28	G26	2	1	0	1
29	G27	2	1	0	1
30	G28	2	1	0	1
31	G29	2	1	0	1
32	G30	2	1	0	1

Figure 3: Number of Formed Groups

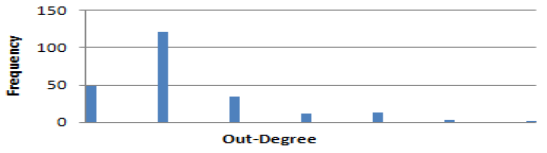
In group G1, it contains 71 vertices and unique edges are 93. Degree of a node is measured by the number of edges incident on a node from others [6].



Minimum In-Degree	0
Maximum In-Degree	85
Average In-Degree	1.254
Median In-Degree	0.000

Figure 4: In-Degree of 'ForeignAffairs' Vertex

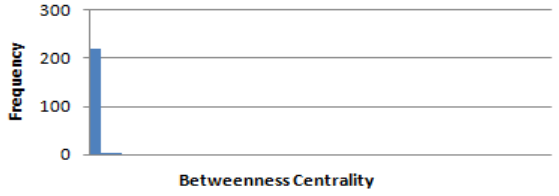
As shown in figure 4, the number of edges incidented on Foreign Affairs node from other, the In-degree garph shows Foreign Affairs have maximum In-Degree is 85 and Average In-Degree is 1.1254 and Out-Degree concern with the incident edges from source vertex to others.



Minimum Out-Degree	0
Maximum Out-Degree	6
Average Out-Degree	1.254
Median Out-Degree	1.000

Figure 5: Out-Degree of 'ForeignAffairs' Vertex

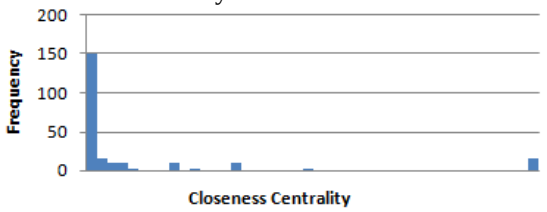
The fig. 5 shows the Out-degree garph of ForeignAffairs have maximum Out-Degree is 6, Average Out-Degree is 1.1254 and Median Out-Degree is 1.



Minimum Betweenness Centrality	0.000
Maximum Betweenness Centrality	12979.000
Average Betweenness Centrality	97.095
Median Betweenness Centrality	0.000

Figure 6: Betweenness Centrality of 'ForeignAffairs' Vertex

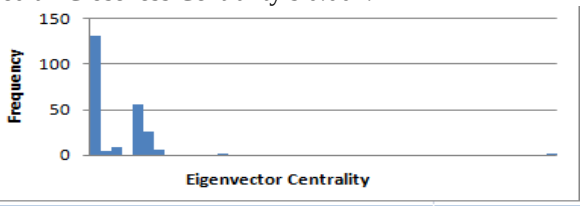
Betweenness centrality concern with number of times a vertex form a bridge between pair of vertices shortest path [7]. In above graph figure 6, shows the Betweenness Centrality of ForeignAffairs have maximum Betweenness Centrality is 12979, Average Betweenness Centrality is 97.095.



Minimum Closeness Centrality	0.000
Maximum Closeness Centrality	1.000
Average Closeness Centrality	0.113
Median Closeness Centrality	0.004

Figure 7: Closeness Centrality of 'ForeignAffairs' Vertex

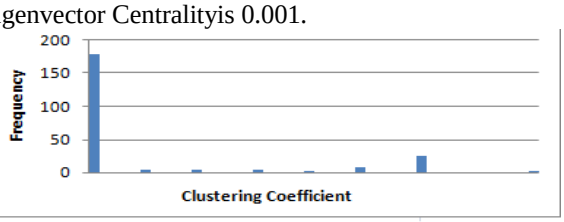
Closeness centrality deals with the closeness of vertex V to vertex H, it measures the length of path between one node to others [8]. In this graaph Figure 7, it shows the Closeness Centrality garph of ForeignAffairs have maximum Closeness Centrality is 1, Average Betweenness Centrality is 0.113 and Median Closeness Centralityis 0.004.



Minimum Eigenvector Centrality	0.000
Maximum Eigenvector Centrality	0.087
Average Eigenvector Centrality	0.004
Median Eigenvector Centrality	0.001

Figure 8: Eigenvector Centrality of 'ForeignAffairs' Vertex

Eigenvector centrality deals the influence of a vertex in the network and assign value to other vertices [9]. In this graph figure 8, it shows the Eigenvector Centrality garph of ForeignAffairs have maximum Eigenvector Centrality is 0.087, Average Eigenvector Centrality is 0.004 and Median Eigenvector Centralityis 0.001.



Minimum Clustering Coefficient	0.000
Maximum Clustering Coefficient	0.667
Average Clustering Coefficient	0.088
Median Clustering Coefficient	0.000

Figure 9: Clustering Coefficient of 'ForeignAffairs' Vertex

Clustering coefficient deals with the measurement of degree of a node have probability to be cluster with neighbor node. In figure 9, Clustering Coefficient of 'ForeignAffairs' node have maximum Clustering Coefficient is 0.667 and Average Clustering Coefficient is 0.088.

VI. CONCLUDING REMARKS AND FUTURE SCOPE

Social Networking Analysis is used for different range of applications and disciplines. In this paper, authors have focused on finding the biggest group, active on a specific day and also, finding out most influential node of the network on that particular day, this type of analysis is effective for business intelligence need. On searching a string value for different respects such as community, member and individual people using that string in their tweets. While it is also found how many users are using that string value. So, extraction of data for that given input in real time scenario has been analyzed in form of nodes and edges. By using this graphical representation it is clear how many users and

community are useful with respect to given input text. More specific and targeted analysis will help law agencies as well as business establishment in predicting and analyzing individual nodes.

VII. ACKNOWLEDGEMENT

We are very thankful to our respected Mr. Aseem Chauhan, Chairman, Amity University, Lucknow, Maj. Gen. K.K. Ohri, AVSM (Retd.), Pro-Vice Chancellor, Amity University, Lucknow, India, for providing excellent computation facilities in the University campus. We also pay our regards to Prof. S.T.H. Abidi, Director and Brig. U.K. Chopra, Deputy Director, Amity School of Engineering and Technology, Amity University, Lucknow for giving their moral support and help to carry out this research work.

REFERENCES

[1] Ching-Yung Lin (2012), '*Social Network Analysis in Enterprise*', IEEE, Vol. 100, PP. 2759-2776, 0018-9219.

[2] Salvatore A. Catanese, Pasquale De Meo, Emilio Ferrara Giacomo Fiumara, Alessandro Provetti (2011), '*Crawling Facebook for Social Network Analysis Purposes*', ACM.

[3] Emilio Ferrara (2012), '*A large-scale community structure analysis in Facebook*',Springer Open Journal 2193-1127.

[4] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Drusche, Bobby Bhattacharjee, (2007), '*Measurement and Analysis of Online Social Networks*' ACM, PP. 29-42.

[5] Daqing Zhang and Bin Guo, Zhiwen Yu (2011), '*The Emergence of Social and Community Intelligence*', IEEE Vol. 44 PP. 21 – 28, 0018-9162.

[6] Shahadat Uddin, Liaquat Hossain (2011), 'Time Scale Degree Centrality: A Time-Variant Approach to Degree Centrality Measures' in Proc. of the 2011 International Conference on Advances in Social Networks Analysis and Mining, PP. 520-524.

[7] Sanjiv Sharma, G.N. purohit (2012) *'A New Centrality Measure for Tracking Online Community in Social Network*', IJITCS Vol. 4 PP. 47-53, 2074-9015.

[8] Sungjoo Park, Minjae Park, Hyuna Kim, Haksung Kim, Wonhyun Yoon, Thomas B. Yoon, Kwanghoon Pio Kim (2013), '*A Closeness Centrality Analysis Algorithm for Workflow-supported Social Networks*' , ICACT 1738-9445.

[9] Yukiya Kato, Fumie Ono (2011), *'Node centrality on disjoint multipath routing*', IEEE, PP. 1 – 5, 1550-2252.