Detecting and Analyzing Communities in Social Network Graphs for Targeted Marketing

Gautam Bhat, Rajeev Kumar Singh
Department of Computer Science and Engineering
Shiv Nadar University
Gautam Buddh Nagar, UP-201314 India

***Abstract*- As the size of an online social group or community increases, so does the complexity of its network and the cost of understanding its structure. In this paper, we propose a different method of forming the social graph of such networks based on past interactions and like patterns to understand dynamics of a community on the social web. We perform analysis on this graph by utilizing a modularity algorithm to detect clusters and computing various metrics of node ranking to determine influential nodes, which can be used for viral and targeted marketing in these modules.**

Keywords- social network analysis; social media marketing; network influence; community detection; network graphs;

I. INTRODUCTION

Social networks can be thought of as connections between individuals that capture relationships between them. Today, the ubiquity and scope of social network data is such that, intelligent computer programs are needed to reveal interesting patterns and answer questions about this data, or use it for certain purposes. One such application is marketing, and requires knowledge of various disciplines for successful advancement in research and implementation, such as Computer Science, Mathematics, Statistics, Psychology and Sociology. Marketing, with respect to this paper, can mean any of the following:

- Advertising a product
- Popularizing an idea
- Spreading and propagating news through a network

Research interest in the field has risen tremendously, as physical world societies and communities are steadily gaining a stronger presence online through social networking sites such as Facebook and Google+. These groups, representing a common belief, organization, or geographical location, can vary from a few hundred members to more than ten thousand, and the time and cost for understanding their structure and propagating through the network is proportional to the number of nodes.

Current graph theoretic techniques of analyzing and visualizing social graphs are based on studying connections among users in the network, such as friendships (undirected graphs) or follow relationships (directed graphs). While these techniques are able to give a better idea about a user's network and potential influence over it, such connections do not particularly reveal the true nature of the relationship between two users and their behavior in the network. For example, two acquaintances who just added each other as friends on social media and two close childhood friends that interact on daily basis are represented by similar links in the graph. Moreover, a user with the highest number of friends and a central position in the network is of no use to marketing if he or she is not active on the network. It is also observed that, in smaller groups most people are acquainted with each other, and hence there is a high probability that most people are friends with one another. This would lead to a high graph density, and hence a loss in accuracy and in the extent of community detection. In order to remedy this problem and overcome the previously mentioned shortcomings, in this research, we propose a different way of viewing relationships between nodes in a network, so as to get a better understanding of how nodes interact with each other, and what can be inferred from the nature of these interactions. The graph is constructed from a mathematical as well as statistical and social point of view. We utilize techniques from graph theory to detect communities within the network and also rank the nodes based on their importance and influence, for targeted marketing as well as viral marketing.

This paper is organized as follows. Section II gives a brief overview of the metrics used in the research, and presents the tools and methodology involved in carrying out the work. We report the results in Section III. Section IV describes recent and related work in the field, and Section V concludes with a summary and discussion of future research directions.

II. THEORY AND METHODOLOGY

A. DATASET

Historical data from our university community network on Facebook was extracted as a real-world representative model to analyze and predict properties of much larger and complex networks. This data was obtained by interfacing Facebook Graph API 2.0. Data was extracted in the form of users as actors, and likes on posts by other users as the relation that connects two nodes; this can be understood as an asymmetric relation.

Software tools R and NodeXL were used to clean the raw data and prepare it for further use. Data can be stored either in a database, with a table for nodes, and a table for edges between nodes and their weights, or as a weighted adjacency matrix, stored in a CSV or Excel format, which can be converted into .pajek or GraphML formats, depending on the software tool or program using the dataset. R and Gephi were used to compute various network and node metrics. The social graph is created as follows:

1. A node represents a user on the network
2. A directed edge is created between two nodes if one node has liked a post authored by another node.
3. For more likes by the same node on another node, add unit weights to the edge.

Therefore, if one user has liked three posts by another user, there would be a directed edge from the former to the latter, with an edge weight of 3.
The graph properties are as follows:

- Graph Type: Directed, Weighted
- Nodes: 875
- Edges: 15259
- Graph Density = |E| / (|V| * (|V| − 1)) = 0.02

B. NODE RANKING

There are various measures of centrality in graph theory and network analysis to study and determine the relative importance of a node within a graph. The following centrality measures were used to rank the nodes.

i. Weighted In-degree Centrality

This is the simplest measure of centrality, which calculates the total sum weight of all edges entering a node. A higher measure would mean more positive reception in terms of posts made by the node on the network.
The results after calculating the weighted in-degree of each node is shown in Fig. 1.

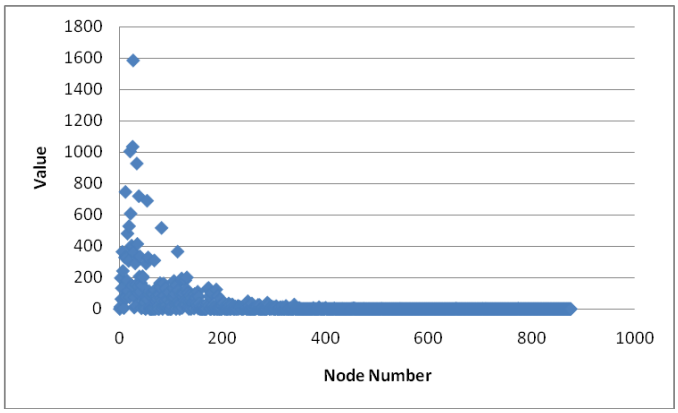

Figure 1: Weighted In-degree Distribution

ii. Betweenness Centrality

Betweenness centrality measures how often a node appears on shortest paths between other nodes. For a vertex v, and all other pairs of vertices (s,t), betweenness can be represented compactly as:[1]

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v. The measure proposed by authors of [2] takes into account edge weights for betweenness centrality and can also be applied to directed graphs. In our social graph, the betweenness of a node quantifies how central or influential it is in terms of the interactions other nodes have with respect to it.
The results are shown in Fig. 2.

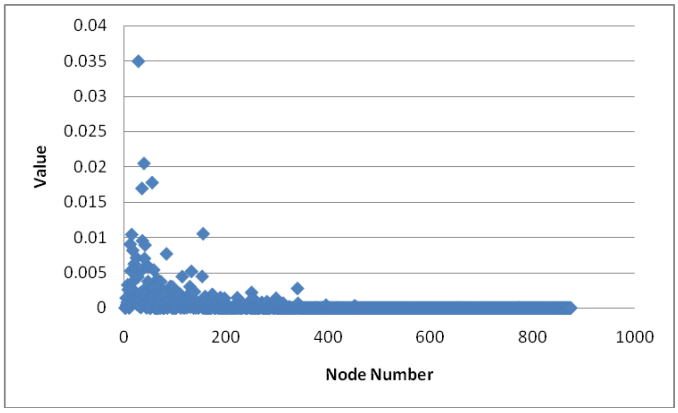

Figure 2: Normalized Betweenness Centrality Distribution

iii. Eigenvector Centrality

Eigenvector centrality is a measure of node importance based on its connections. It uses the adjacency matrix to calculate eigenvectors and assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. The centrality score of vertex v can be defined as:

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

where M(v) is a set of the neighbors of v and λ is a constant. With a small rearrangement this can be rewritten in vector notation as the eigenvector equation

$$\mathbf{Ax} = \lambda \mathbf{x}$$

The results are shown in Fig. 3.

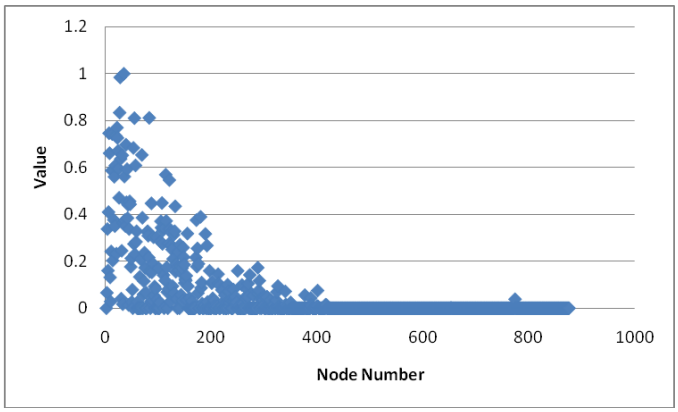

Figure 3: Eigenvector Centrality Distribution

Google's Pagerank is a variant of the eigenvector centrality measure in that calculates left eigenvectors, according to the algorithm described in [3]. It is a link analysis algorithm that assigns a numerical weighting to each node with the purpose of measuring its relative importance within the network. Although originally created to rank web pages, a generalization of it for weighted graphs is considered, when the edges of the graphs are appropriately weighted to provide additional information when the connections have different meaning, importance or quality.

The results are shown in Fig. 4.

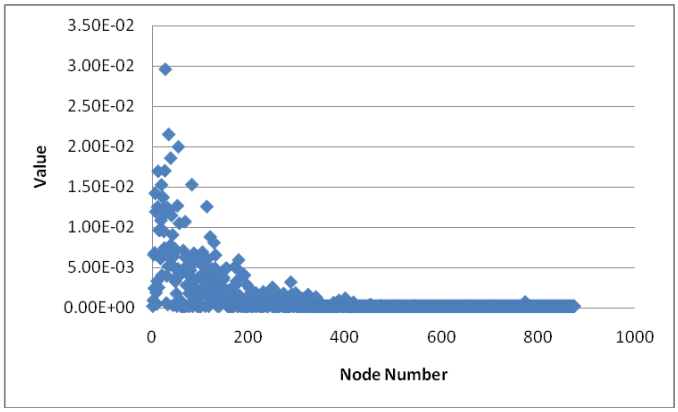

Figure 4: Pagerank Distribution

C. COMMUNITY DETECTION

To retrieve comprehensive information from large, complex networks, they are often partitioned into sub-units or communities, which are sets of highly inter-connected nodes. Modularity is one such measure which is used to detect and divide the network into modules, clusters, or communities.

We used a modularity algorithm proposed in [4] to detect communities in the network, in such a way that there are dense connections between nodes within modules but relatively sparse connections between nodes in different modules. By randomizing the algorithm and incorporating edge weights we were able to obtain 4 distinct communities with more than 40 nodes each. Both of the two largest modularity classes had over 300 nodes each. The pie chart in Fig. 5 illustrates the clustering of nodes, and Fig. 6 shows one of the modules visualized in Gephi.

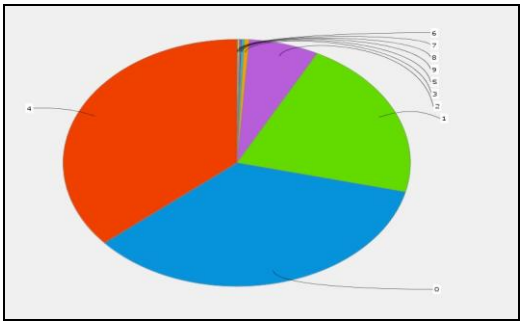

Figure 5: Distribution of nodes according to modules

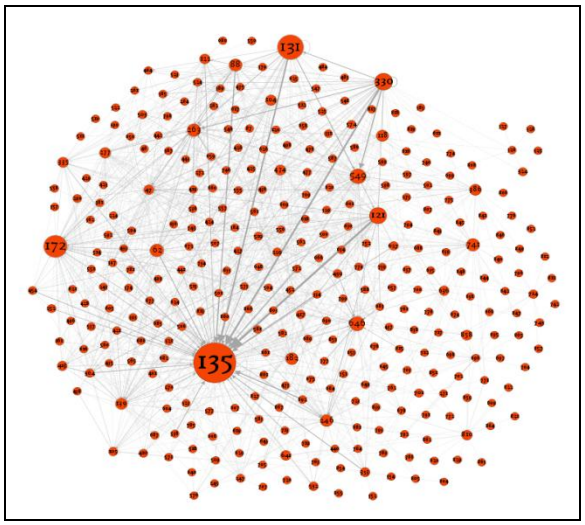

Figure 6: A sample community computed using modularity, and nodes ranked according to their betweenness centrality

III. RESULTS AND DISCUSSION

We were able to successfully run graph data mining algorithms on the newly-formed social graph, and detect communities using the modularity algorithm. The layout of the graph was performed partly manually and partly using Gephi's in-house Force Atlas 2 algorithm [5] to distance the nodes appropriately. On sorting and plotting the computed node metrics, it was observed that the distribution was heavy-tailed - most nodes had low centrality/degree, while a small number of nodes had very high centrality/degree. The degree distribution showed that the network followed power law behavior very closely. It was also observed that betweenness centrality, followed by weighted in-degree and pagerank, had more clear heavy-tailed distributions and gave better results than eigenvector centrality when determining undisputed influential nodes in such a large and complex network. If

multiple influencers are needed, then eigenvector centrality may be used as well. Since partitions were created in the graph using the community-detecting modularity algorithm, we were able to meet the objective of targeted and viral marketing to reduce cost and time of advertising, while also having some amount of control over selective advertising in a network.

The un-mined graph of the prepared dataset is shown in Fig. 7, and the final graph of the sample network after mining and clustering is shown in Fig. 8.

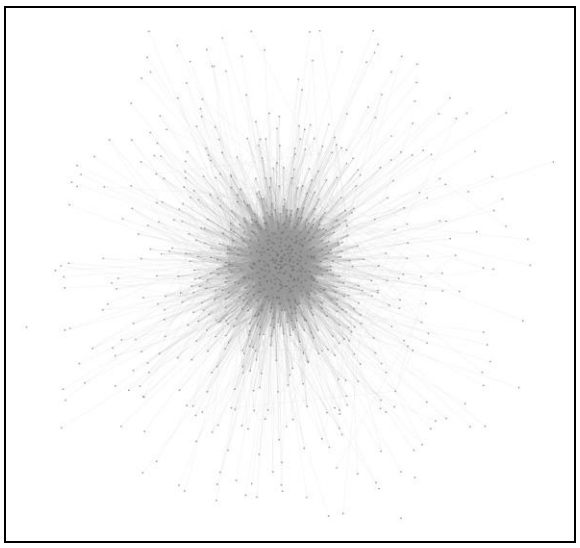

Figure 7: Visualization of prepared dataset in Gephi

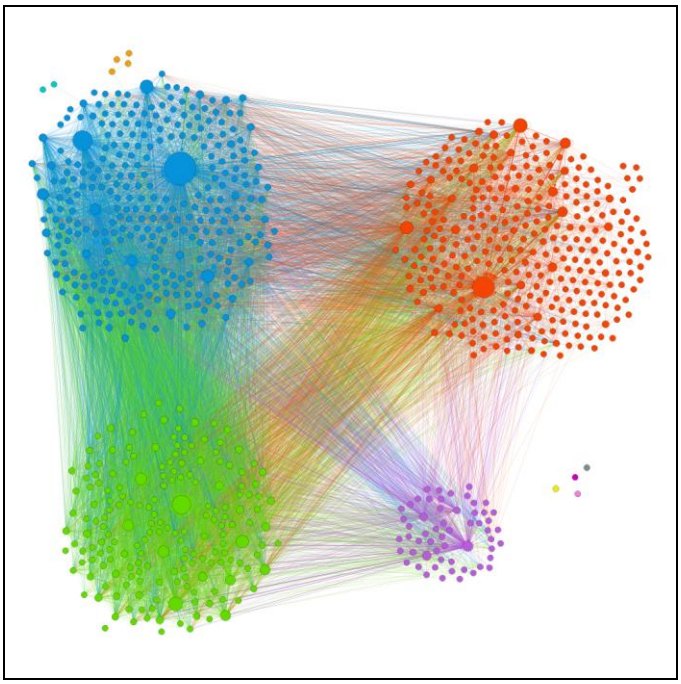

Figure 8: Final graph - Clustered communities and ranking of nodes by betweenness centrality

IV. RELATED WORK

Specialized data mining and machine learning algorithms have been developed for social networks, since techniques used for propositional data are generally unsuitable for the clustering of social-network graphs. Various researchers are working on community detection, and viral and targeted marketing. In Sharma and Shrivastava's work [6], two types of clusters were developed using data mining algorithms, one having strong tie and the other having a weak tie between members. The cluster with strong tie is used for discovering the highly influential node for viral advertising using target marketing to increase the profit with less advertising expense. Zhao et al. [7, 8] have addressed topic oriented community detection through social objects and link analysis in social networks. Wang and Chen[9] have worked on maximizing scalable influence for prevalent viral marketing in social networks.

Computational Advertising is a relatively new discipline within Computer Science that deals with algorithms of presenting the best advertisement displayed to a person, typically through an Internet browser.

V. SUMMARY AND FUTURE WORK

With our method of forming the social graph, we were able to study past interactions and like patterns of real-world data to predict how likely a node is to influence other nodes around it, and how likely are other nodes to take interest in posts made by a particular node in the future. From this we get a better understanding and overview of how nodes in the network actually behave on their connections, rather than how they are just connected. This method of forming the social graph is especially effective in smaller community counterparts on social networks, such as workplace or student community groups on Facebook, where most people are friends with each other, leading to an almost complete graph with high graph density, because of which modularity algorithms would be less successful in partitioning the friendship network into well-defined clusters. It was seen that betweenness centrality and weighted in-degree were the two most suited measures for finding influential nodes, as these metrics helped identify distinct important nodes for each module/cluster, which could be effectively used to propagate through these clusters with minimum hops and cost. These nodes can, hence, be used for cost- and time-effective targeted marketing during social campaigns, spreading news through the community, or marketing an idea or product.

The work done in this research can be taken forward to predict the reception a new post by a particular user on the network would receive, including the estimated number of likes and the likers. Further work can be done on this project by incorporating post comments into the data as well. Negative or positive edge weights can be assigned to the comments, with the help of sentiment analysis. Moreover, property graphs can be studied further, to target particular communities based on node properties and attributes. Looking ahead, the ideas and results from this project can be used to devise machine learning algorithms that work on dynamic, real world social

network data in the field of computational advertising, for viral and intelligent marketing.

REFERENCES

[1] Brandes, Ulrik (2001). "A faster algorithm for betweenness centrality" (PDF). Journal of Mathematical Sociology **25**: 163–177.

[2] Opsahl, T., Agneessens, F., Skvoretz, J., 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks 32 (3), 245-251

[3] Sergey Brin, Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", in Proceedings of the seventh International Conference on the World Wide Web (WWW1998):107-117.

[4] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, "Fast unfolding of communities in large networks", in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000.

[5] Jacomy M, Venturini T, Heymann S, Bastian M (2014) "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software". PLoS ONE 9(6): e98679. doi:10.1371/journal.pone.0098679

[6] Sharma, Shrivastava, "Viral Marketing in Social Network Using Data Mining", in International Journal on Recent and Innovation Trends in Computing and Communication, March 2013.

[7] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha, "Probabilistic Models for Discovering ECommunities", in Proc. of the 15th Int. Conf. on World Wide Web, 2006.

[8] Z. Zhao, S. Feng, Q. Wang, J. Z. Huang, G. J. Williams, and J. Fan, "Topic oriented community detection through social objects and link analysis in social networks", In Journal Knowledge-Based Systems 26, 2012, pp. 164–173.

[9] W. Chen, C. Wang, W. Wang, "Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks" July 2010