AN UPDATED LOOK AT SOCIAL NETWORK EXTRACTION SYSTEM A PERSONAL DATA ANALYSIS APPROACH

Preetha S[1], Ancy Zachariah[2]

[1]*Assistant Professor, Division of Computer Science and Engg., Cochin University of Science and Technology, Kerala*
[2]*Associate Professor, Division of Computer Science and Engg., Cochin University of Science and Technology, Kerala*

Abstract

With the rapid popularity of social networks, interactions between people are no longer limited to geographical boundaries. Social interactions between the different people are the basis for dynamic social networks. As a result of the interactions between different people they can influence one another. These interactions form the basis of community for e.g. scientific community. The communication between interconnected participants in a social network can be analysed and used to derive more information which can utilized in various ways. This paper studies the essential features of Facebook that is the friendship relation between participants and tries to find the influential user for a particular person. This has a huge impact on spreading new ideas and practises. In a marketing scenario the information is vital as identifying the influential person could be the turning point in sales promotion. As the dynamic network was studied over a period we could also analyze the structural changes in network that eventually changed the influential person. It is also observed that number of groups in the network tends to decrease as the size of network increases. Relationship is viewed in terms of network theory with actors as nodes and relation as links. An individual can bridge two networks that are not directly linked. The attributes of individual is less important than the relationship or ties with other nodes in and outside the network.

Keywords: *Social Network Analysis, Visualization, Influential User, DNA and Community Detection*

----------***----------

1. INTRODUCTION

A social network analysis maps relationship and flow between people or entities. The people in the networks can be represented by nodes and the relationship between the nodes is shown by the link. It allows both visual and mathematical analysis of human relationship. In the real world, relations can be studied as a general pattern. Relations are complex and their representation can be miscellaneous. Generally the relationship pattern are categorized as static and dynamic patterns. Static pattern do not change dramatically in nature while dynamic pattern always tend to change in a drastic manner.

Dynamic network analysis is a new approach that takes ideas from traditional SNA, Link analysis(LA) and multiagent system (MAS) which were formerly used. The statistical analysis of DNA data is of prime importance in this area. The alternate approach is the utilization of simulation to address issues of changes in network. DNA network are larger, dynamic and contain the various levels of uncertainty compared to SNA. The main difference of DNA to SNA is that DNA takes the domain of time into account. With social network there is a huge difference in the way we communicate with each other. Common issues can be highlighted and communicated by grouping the affected parties. Information and opinions get spread easily in social network making the world more global. Since theses data are time stamped we can study the spreading process at a system level. With tools like Netvizz and Gephi, gathering data and testing the models are very efficient compared to traditional methods. We can study any event on a time scale ranging from few days to several years. For example cultural events can be studied for a few days, whereas in online encyclopaedia sites changes occurs rarely, once in a few years.

Social interactions within network comprise an increasing event nowadays. Different aspects of societies and competitive market, such as social Medias through the internet and telecommunication environment hold a high correction with the social network analysis methodology. By understanding these social structures and interactions it might be possible to realize how the individuals and consumers related to each other and hence predict further social structure. However, the most of the current social network analysis project are in relation to static structure, not considering how the social networks evolve over the time. The dynamic approaches points out new perspectives in terms of social network analysis including prediction and simulation scenarios. In order to perceive the social network relation over the time it is crucial to collect the distinct snapshot of the structure, understanding not just how the social members related each other but in addition to that how these relationship evolves over the time. Measurement in relation to social network describes nodes and links by static matrices, depicting its strength, its overall distance to the other related nodes, and its amount of connections, among others. However in order to understand how the social interaction change over the time is important to follow all these measures.

2. BACKGROUND

The world we live in is very complex. The relationships between people, groups and organizations are on an increase. Social groups related to working organization, fans group of film, sports and other personalities who rise to fame, tech savvy groups are increasing. The researchers can study these systems and analyse the links. The attention can be focussed on the connections also referred to as links or edges, between various entities, that are referred to as nodes. Most of the technical systems are dynamic. Many changes occur as we grow from birth through our education period and as we age. This study over a period of time is focussed.

The understanding and evaluation of these systems are complicated as data on these systems is often incomplete, replete with errors and difficult to collect. Hence SNA and LA tools become obsolete to process these requirements. DNA has emerged as a new subfield of SNA. In DNA the proposal is combination of the methods and techniques in SNA and LA. This can be combined with MAS techniques which provide a set of techniques and tools for the analysis of dynamic system. In the study that follows we are identifying influential user, detecting community and studying the changes in the size of the community.

The pattern change is not very visible in static pattern whereas they change drastically in dynamic pattern. The example of a static change would be related to the change in printed characters over time. Hand written text tends to deform over time and recognizing such characters is a challenge hence can be classified as dynamic pattern. In the pattern recognition and recovery process, dynamic pattern are more challenging.

In social network analysis the main task is usually about how to extract data and relationships from different communication sources. The data used for building social networks is Relational data, which can be obtained and transferred from different resources including the web, email communication, internet relay chat, telephone communication, organization and business events, etc. For example, email communications are a rich source for interacting and constructing social networks. By measuring the frequency of email communication and studying the replies, we can study the relationship between email senders and receivers. This data can then be used for social network construction.

Social network can be generally defined as a group of individuals those are connected by a set of relationship. In network theory, link analysis is a data analysis technique used to evaluate relationship (connections) between nodes. Relationship may be identified among various type of nodes (objects) including organizations, people and transaction. Link analysis has been used for investigation of criminal activity (fraud detection, counterterrorism and intelligence), computer security analysis, search engine optimization, market research and medical research. This measures revealed that, as a difference with the classical Erds-Rnyl random graph model, real networks are characterized by a power low distribution of vertex degree, a high clustering coefficient or transitivity, and degree correlation between connected vertices. Yet, it is important to characterize up to which extend the measure provide the information about the studied networks. For instance it has been shown in some networks the degree correlations are consequence of the existence of large degree vertices and therefore the sequence of vertex degrees is sufficient to characterize those networks.

Counting the number of triangles in a graph is a beautiful algorithmic problems which has gained importance over the last year due to its significant role in complex network analysis. Matrices frequently computed such as the clustering coefficient and the transitivity ratio involve the execution of triangle counting algorithms. Furthermore several interesting graph mining applications rely on computing the number of triangle in the graph of interest. Social networks change with time. But the research in this area had always avoided the time of interaction in their study. Hence we can say all the studies were static. Another major area of social network analysis is visualization, and it is very appropriate to visualize a social network[1]. We can explore the properties of social network by visualizing social networks. The typical characteristics of social network can be easily understood. The structure of network, the node distribution, relationship between nodes and the groups in the social networks, can be more easily expanded. Other than extraction we can analyse the connectivity and degree of nodes and links using density measurement or measure clusters using cluster coefficient or measure closeness of social network using betweeness centrality.

A friendship graph is one which nodes collaborate with other nodes and nodes are entities. The friendship graph evolves over time. This friendship graph is analyzed[2]. People switch jobs, move to new places and their interactions pattern change. The entities move away from one another or as time passes entities will either come close together forming new groups. In social networks such pattern changes in interactions are very common. So we have to identify the changes that are happening in that network.

Application of DNA can be in identifying key individuals, locating hidden groups and estimating performance. DNA tools include software packages for data collection, analysis, and visualization. Identifying relationship between individuals and groups, discovering the key persons, the weakness points and comparing networks are the major tasks of data analysis. With the help of visualization tools analysts can graphically explore networks. The Fig-1 depict a typical view of a personal data .

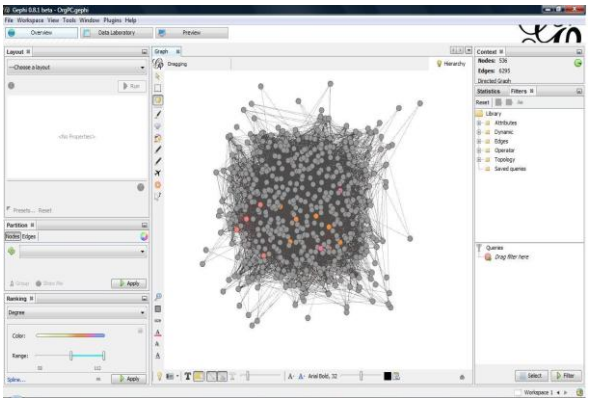

Fig-1: Personal data in Social Network

3. CASE STUDY

Based on the background we designed a system to solve the above problems. In this system we collect data from the Facebook, preprocess it and analyze it to be developed into a decision support system. Our experimental analysis mainly focusses on calculating SNA matrices and finally developing the results. The **Fig-2** depicts system architecture.

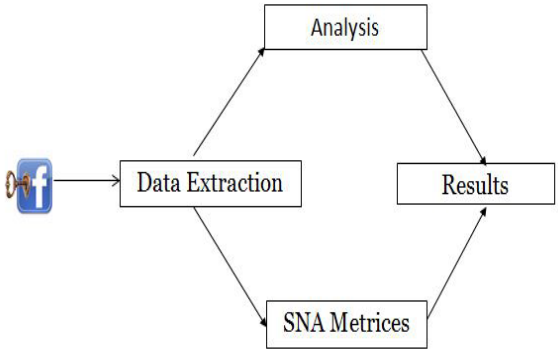

Fig-2: System Architecture

The first phase is data extraction were data is extracted from Facebook network. Then we have to analyse this network and calculate different metrics in social network. The relationships are visualized after extracting useful information from collected data.

3.1 SNA Metrics

SNA metric are the measures that we are using in social network analysis. The different metrics that we are using are and Betweeness centrality.

3.1.1 Degree Distribution

In the study of graphs and networks, the degree of a node in a network can be computed by counting the number of connections it has to other nodes. The probability distribution of these degrees over the whole network is called the degree distribution. In an undirected network we count the edges the node has to other nodes to compute the degree of the node. In a directed network with edges pointing in one direction there are two degrees called in-degree and out-degree . The in-degree is the number of incoming edges and the out-degree is the number of outgoing edges. The fraction of nodes in the network with degree k is called the degree distribution P(k) of a network. That is, in a network with n nodes if nk of them have degree k, Then degree distribution P(k) is computed as nk/n. The cumulative degree distributions, the fraction of nodes with degree greater than or equal to k are all referring to the same information.

3.1.2 Betweeness Centrality

Betweeness centrality is a measure of centrality of nodes in a network. To compute this we have to find the shortest paths that pass through that node. The total number of such paths is called betweeness centrality. Betweeness centrality not only indicates the importance of a node in network which is of local importance but it also indicates the load placed on the given node in the network which is of global importance.

4. EXPERIMENTAL RESULTS

4.1 Community Detection

The whole network of a person can be grouped into communities. A person may be connected to family, school and college friends, or colleagues. The work culture may induce other groups, for eg. teachers connected to students, in online marketing agents connected to clients. Thus the whole network can be summarized and visualized as communities. Without sacrificing the individuals confidentiality we can extract the information from community. Using clustering we can detect the evolving communities[4]. Shortest-path betweeness can be used for clustering. Another approach being clustering based on network modularity.

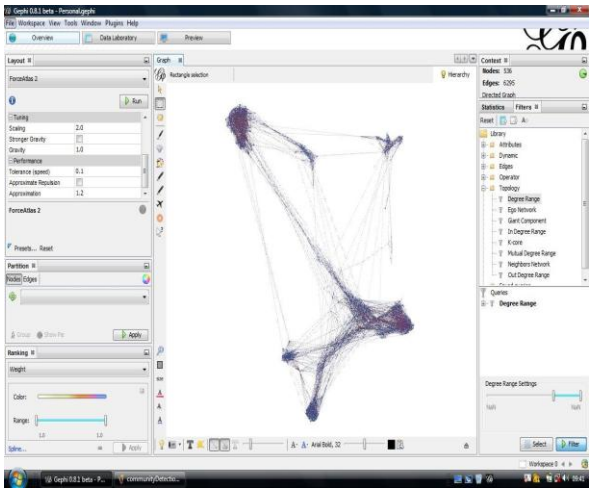

Fig-3 Community Detection

Community detection of a personal network is shown in Fig-3. We can see that there are four communities in the given graph. In that two communities are more clustered together than the other. It shows the strength of communication between users[4].

4.2 Celebrity Detection

Celebrity of a network can be identified based on the number of indegree of a node. The user with maximum number of indegree will be the celebrity of that particular network. The users with minimum indegree will be connected to the users having maximum indegree. In that way, the celebrity of a personal network influence the users who are not so active in social network activities.

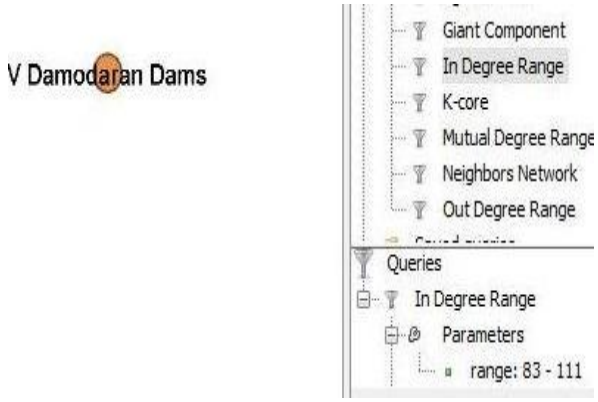

Fig-4: Celebrity Detection

The Fig-4 shows the celebrity of the personal network which we selected for demonstration of results. Here we can see a single user as celebrity. The same user can be a celebrity of more than one network. It depends on the number of indegree, that is the number of mutual friends. The users with minimum indegree will be connected to celebrity of the networks.

4.3 Most Influential and Least Influential User Detection

The users with minimum number of indegree are known as least influential users. The Fig 5 shows a network of least influential users.

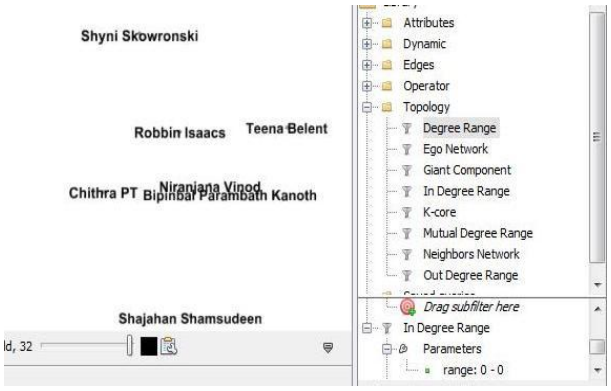

Fig- 5: Least influential user network

The least influential users will be connected to the most influential user. The user with maximum number of indegree is known as most influential use and fig 6 shows network of most influential users.

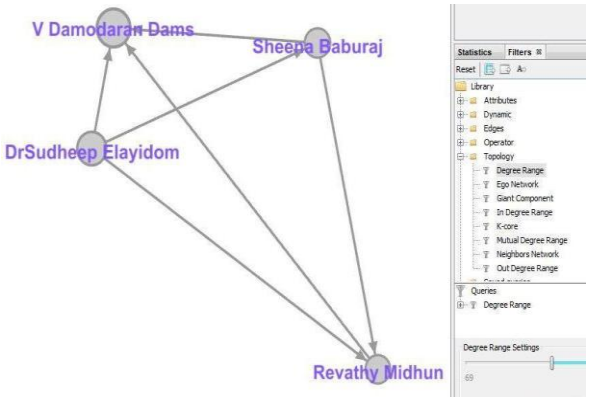

Fig 6: Most influential users

4.4 Ego Networks

Local circumstances influence the behaviour variation of individuals [1]. Hence we need to look into this aspect carefully. The goal of the analysis of ego networks is to describe and index the variation across individuals in the way they are surrounded in our social structures. An individual node when focussed will yield ego network. Hence there will be as many egos in a network as it has nodes. A persons or group can act as ego. A particular person's connection can be filtered with ego networks.

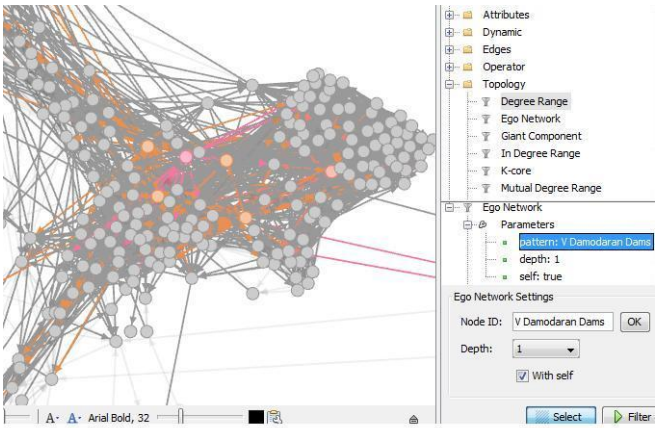

Fig 7: Ego Network

The Fig 7 shows ego network of the detected celebrity. The celebrity is represented with different colour from other user nodes for easy identification.

5. CONCLUSION

The work studied the dynamic change occurring in a social network during a period of time, detected influential user, detected possible communities and discovered the celebrity of a local network. The real world Facebook data set was used for this purpose and found the changes occurred during the period. By mining user interactions, important information can be retrieved In our work we identified the most and least influential user in a particular persons network. We were also able to detect various communities or groups. As network becomes larger the groups in the network decreased. It is because of the dynamic nature of the network. So from these results we can conclude that on

social network lots of interactions are happening which shows different patterns in network and patterns are changing with time. These results can be used in areas like marketing, advertisement, recommendation, political discussions etc. as influential users act as hub of information in a community. Smart organisations can augment their campaigns through proven media channels with social network analysis.

ACKNOWLEDGEMENTS

We appreciate Lima Johnson, Anju Shaji and Neethu K M for effective conversations and Sreelakshmi for help in editing.

REFERENCES

[1]. Kai-Yu Wang,I-Hsien Ting,Hui-Ju Wu,IEEE paper on 'A Dynamic and Task-Oriented Social Network Extraction System Based on Analyzing Personal SocialData', International Conference on Advances in social networks analysis and mining, August 2010.

[2]. Alice X. Zheng, Anna Goldenberg C "A Generative Model for Dynamic Contextual Friendship Networks" CMU-ML-06-107 School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213July 2005.

[3]. Ravi Kumar,Jasmine Novak ,Andrew Tomkins "Structure and Evolution of Online Social Networks" Proceeding of 12th ACM SIGKDD international conference on knowledge discovery and data mining, August 2006.

[4]. Jaewon Yang, Julian McAuley, Jure Leskovec "Community Detection in Networks with Node Attributes" in ICDM, 2013.

BIOGRAPHIES

Preetha S has completed her M.Tech from CUSAT in 2001. She is working as Assistant Professor in CUSAT since then. Her area of interest includes Big Data, Social Network Analysis and Computer Network.

Ancy Zachariah has completed her M.Tech from CUSAT in 2003. She is working as Associate Professor in CUSAT since 1999. Her area of interest includes Social Network Analysis, data mining and programming techniques.