ANALYSIS OF LARGE SOCIAL NETWORKS TO INVESTIGATE INTRIGUING PATTERNS

A report submitted to

M S RAMAIAH INSTITUTE OF TECHNOLOGY

Bengaluru

IS715 Project Work I

as partial fulfilment of the requirement for

Bachelor of Engineering (B.E) in Information Science and Engineering

by

AJITESH JAYANTH (USN- 1MS12IS011)
AKSHAY RAO AK (USN- 1MS12IS013)
ANKITH MOHAN (USN- 1MS12IS014)
G AKASH ROHIT (USN- 1MS12IS037)

under the guidance of

KRISHNARAJ P M

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

M S RAMAIAH INSTITUTE OF TECHNOLOGY

Dec 2015

Department of Information Science and Engineering

M S Ramaiah Institute of Technology

Bengaluru - 54

CERTIFICATE

This is to certify that Ajitesh Jayanth (USN- 1MS12IS011) , Akshay Rao A K (USN- 1MS12IS013), Ankith Mohan(USN- 1MS12IS014) and G Akash Rohit (USN- 1MS12IS037) who were working for their **IS715 Project Work I** under my guidance, have completed the work as per my satisfaction with the topic **Analysis Of Large Social Networks To Investigate Intriguing Patterns**. To the best of my understanding the work to be submitted in dissertation does not contain any work, which has been previously carried out by others and submitted by the candidates for themselves for the award of any degree anywhere.

(Guide)
Krishnaraj P M
Assistant Professor, Dept. of ISE

(Head of the Department)
Dr. Vijay Kumar B P
Professor & Head, Dept. of ISE

(Examiner 1)

Name

Signature

(Examiner 2)

Department of Information Science and Engineering
M S Ramaiah Institute of Technology
Bengaluru - 54

DECLARATION

We hereby declare that the entire work embodied in this **IS715 Project Work I** report has been carried out by us at M S Ramaiah Institute of Technology under the supervision of Krishnaraj P M. This Project report has not been submitted in part or full for the award of any diploma or degree of this or any other University.

AJITESH JAYANTH (USN- 1MS12IS011)
AKSHAY RAO AK (USN- 1MS12IS013)
ANKITH MOHAN (USN- 1MS12IS014)
G AKASH ROHIT (USN- 1MS12IS037)

Abstract

The focus of this study is social network analysis. The work described here involves concepts such as social influence analysis, time series analysis, link prediction, cluster analysis, small world phenomenon and power law analysis. These are illustrated with analysis of certain large social networks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid computerization and availability of high computing power have led to the emergence of social-networks. A brief insight about the topological and behavioral aspects of complex networks can be achieved by thorough understanding of social networks. It is considered a very important topic for research due to the widespread success of social networking websites such as facebook, twitter and linked-in.

Social networks are structures where nodes represent people or entities and edges represent interaction, influence or collaboration between entities. They grow and change quickly over time through the addition of new edges, signifying the appearance of new interactions in the underlying social structure. Social network analysis involves identifying mechanisms by which they grow and evolve. The major application of social networks are point to point overlay networks, security systems in ad-hoc networks, hybrid sensor networks, prediction and control of an epidemic, cellular telephone systems and it also helps us to predict links in the world wide web or any given computer network.

Social network analysis is our primary focus wherein we follow practices such as design, analysis, coding, testing and maintenance so that the evolution of the network can be observed over time. Huge data-sets are collected and assimilated which signify interaction between two or more social entities thus providing insights regarding how closely related they are, at what instance of time they interact and the duration of the interaction. The underlying concept or the phenomenon has to be proved by various methodologies such as algorithmic analysis

either in a iterative or in a regressive fashion followed by modeling, visual representation , animation and detailed presentation.

Obtaining of preliminary results sheds some light on the very fact that whether this approach gives us comprehensive and satisfying results in the near future. The underlying concept is proof-tested by applying on various data-sets obtained from heterogeneous sources. The results and finding from this experimental set-up helps to identify relationship among social entities, derive patterns and accurately predict the future outcomes. Detailed results are obtained thus enabling to conclude its significance in a real world environment.

The interests of this paper lie in social networking concepts such as social influence analysis, time series analysis, link prediction, cluster analysis, small world phenomenon and power law analysis. In order to satisfy the processing of real world type networks work in a parallelized environment or in a distributed computing paradigm must be conducted.

Chapter 2

Literature Review

2.1 Weight Assignment

It is necessary to obtain a clear perspective of the networks to carry out effective analysis. In general, networks can be weighted or un-weighted. The drawback of using un-weighted networks is the lack of real world representation. Therefore the analysis of these networks is difficult. Hence it is necessary to assign weights to the ties among the nodes in a network. Weights are assigned to the edges by considering the features of the network and properties of graph theory. The network features include timestamps, influence, interactions, relationships and emotions. The properties are centrality, betweenness, clustering coefficient and power-law behavior. In some cases it is useful to come up with alternative definitions of centrality, cohesiveness and affinity. (1)

Weight assignment involves the following steps: Initially a feature-rich dataset is obtained and converted to useful form. Score of the network is computed which is based on similarity feature and the score is associated with each edge to obtain the final weight. Advantages of using weighted networks are they emulate the structure of real social networks thus enable to display community structures with weak and strong internal links connecting the communities: facilitation of better understanding of complex networks by providing a complete view of the network and effective results are obtained from rigorous analysis. (8; 11)

Feature or attribute weighting finds applications in content-based recommender systems and profile matching in social networks. Here, weighted network

data is used to perform influence analysis and link prediction. (5)

2.2 Influence Analysis

An attempt is made using social network analysis to answer the question "Who are the most influential individuals in a network?". Influential individuals are said to be those who are capable of convincing other individuals to adapt their attitude, behaviour, or belief to that of the former. Identifying the key influential individuals in a network can play a key role in the introduction, longevity, and fidelity of program implementation.

Given a set of weighted, directed relations among the individuals in a large network, clustering based on the fast greedy algorithm (3) is initially performed on this network to identify closely related sub-networks of individuals. Next, these clusters are analysed one by one as independent organizations. The PageRank (2) of all the individuals in the cluster are computed in order to determine their local standing (relative to the cluster). For each individual A, the weights that A has assigned to every other individual in the cluster is multiplied by the PageRank of A in order to incorporate the influence of A. With these updated weights, the in-degree centrality (also known as in-ties) is computed and then used as a measure of the total communication directed at each individual. For each individual, the in-tie measure indicates how well other individuals in the cluster weigh this individual. Individuals who receive higher scores are considered more influential in the network.

In order to visualize the distribution of the in-degree measure within an organization, the in-degree scores for all individuals can be sorted in descending order and then graphed, resulting in a "scree" plot. This technique allows a researcher to get a sense of the distribution of in-ties for all the actors in the organization. Now the focus is on establishing a defensible threshold for identifying the most influential individuals in an organization. Visual inspection of the scree plot can make identifying influentials an onerous task. As an alternative, four reproducible methods are investigated for categorizing influential individuals in an organization. A detailed explanation of these methods follows. These explanations and further details can be found in (4).

Method 1 - Absolute Cut Score

The simplest and most intuitive method for determining a cut score is to set a predetermined absolute value above which individuals are deemed influential and below which they are not. Graphically, this can be accomplished by superimposing a horizontal line over the in-degree scree plot. Those individuals whose influence scores are above the horizontal line are then categorized as influentials. However, since this method is based entirely on a single point and is determined independent of variation in the distribution of in-ties, it can result in the situation where every individual (or no individual) in a network can potentially be deemed influential since the criterion is absolute, not relative.

Method 2 - Fixed Percentage of Population

An alternative method of identifying influentials is to select a fixed percentage of the population as influential. If the top 20% of individuals in an organization are to be categorized as influentials, this is equivalent to selecting the leftmost 20% of the individuals in the graph. Those individuals to the left of a vertical line superimposed over the scree plot are categorised as influential. As with the Absolute Cut Score, this method identifies individuals as influential independent of the variation of in-ties. It ensures that a given percentage of individuals in the organization are identified as influential and their identification is based upon their performance relative to the performance of other individuals in the organization.

Method 3 - Standard Deviation

Unlike the first two approaches, the Standard Deviation method focuses on the variation in the distribution of ties. This procedure requires calculation of the mean and standard deviation of the number of in-ties. Then we create a horizontal line two standard deviations above the mean, which can be superimposed over the scree plot. This horizontal line approach is similar to the Absolute Cut Score (Method 1), however, the Standard Deviation method does not choose a cut score a priori, instead it utilizes the observed data in determining where to set the cut point. Under this method, those individuals whose in-degree scores are above this line are marked "influential."

Method 4 - Random Permutation

Through the use of random permutations, Method 4 produces results which identify those individuals who received significantly more in-ties than would have

occurred by chance alone. This method capitalizes on the creation of a sampling distribution of potential networks that could have occurred, conditional on the fixed row marginals or (out-tie distribution). In order to obtain a sampling distribution of influence for the network, the graph (nodes and edges) is modeled as an exponential random graph (6). Then the edges are randomly reassigned to individuals in the network keeping the out-degree distribution fixed. By performing one thousand such random permutations the sampling distribution of influence (that is, in-degree distribution) is derived under conditional independence.

The ties are not completely independent, as we restrict their new random locations to only emanate from their original sources in the actual data (i.e. the row marginals are fixed). However, in forcing this restriction, we are able to create a sampling distribution of influence that is comparable to our actual data. The result is the distribution that would arise by random chance, given the set of survey responses, and therefore can be used to identify those individuals whose influence is statistically greater than random chance. Once this is completed, individual influence scores are recalculated according to the in-degree measure described earlier. If an individuals actual influence score is higher than his/her ranked counterpart for 95% of the random iterations, then the individual is labeled a significant influential at the p ¡ .05 level.

2.3 Link Prediction

The problem of predicting links that are either missing or may appear in the future is addressed by link prediction. The methods proposed to cope with the link prediction problem are divided into two categories: unsupervised and supervised methods. Unsupervised methods assign a score for each pair of nodes using either local or global neighborhood information. Experimental results on different social networks show that these methods are time consuming and are not feasible for large social networks. Supervised methods on the other hand consider link prediction as a classification problem. It uses different network information such as structural properties of the network, node attributes to determine the link existence between a pair of nodes. However, it does not use other information such as behavior of nodes. It concentrates only on the existence of the links and

not the properties of individuals that the nodes represent. In order to overcome such drawbacks, hybrid methods are developed which considers local information as well as community information.

Given a network, the probabilities of link existence and non existence of all the links in the network are computed. To determine the probability of a non-existent link, a hybrid method is established which uses Bayesian theory (12). For the weighted relationships among the nodes in a large network, the probabilities of existence of any link in the network are computed. Next, clustering is performed to obtain the different communities present in the network. For pairs of nodes which are not linked, the number of common neighbors between the nodes within and outside of their communities are determined.

From the parameters calculated so far, the posterior probabilities of link existence and non existence for the pairs of nodes are computed. A similarity score for the pairs of nodes is calculated as the ratio of posterior probability of link existence to the posterior probability of non existence. The computed scores are sorted in decreasing order to show that links with higher probability will come into existence earlier than those will lower probabilities.

2.4 Time Series Analysis

Time series analysis is an approach where data or datasets are analysed over discrete intervals of time such that processed data for a period of time say t0 serves as an input to time slot t1. The efforts here are to use novel methods so as to analyze the network, establish relations among nodes, predict and forecast the behavior of social networks over a given period of time. With the help of temporal analysis methods and the underlying algorithms supporting them temporal distances and metrics are determined. (7; 9)

Time series analysis consists of various steps namely estimation, prediction, query processing, result analysis and stimulation in order to get precise and aesthetically better results. It is used in engineering, business activities and sees application in various fields. Here, the attempt is to try and explore how a given social network reacts over time and how it behaves with the addition of new users,

or which users tend to interact among each other, estimation of likelihood that a person X may collaborate with another person Y in future. (7; 10)

Chapter 3

Preliminary Results

3.1 Weight Assignment

Considered here is a dataset consisting of email communications between a large network of individuals. For each email communication, information on sender(from) and receiver(to) is available. From this information, scores for edges are computed based on similarity feature and the score is associated with each edge to obtain the final weight. Listed in the table given below are 10 of the total 59835 edges along with their corresponding weights.

Table 3.1: Weights for edges

from	to	weight
1	2	0.0315
3	4	0.0315
5	2	0.0315
6	7	0.0315
8	7	0.0315
9	10	0.0315
9	11	0.0315
12	13	0.0315
9	14	0.063
9	15	0.0315
⋮	⋮	⋮

3.2 Influence Analysis

For the network given in (3.1) the clustering algorithm has identified 25 clusters. For each of these clusters, the scree plot is generated and the four approaches for influence analysis described in (2.2) are performed. The results of this analysis for the first two clusters are shown in the figures given below.

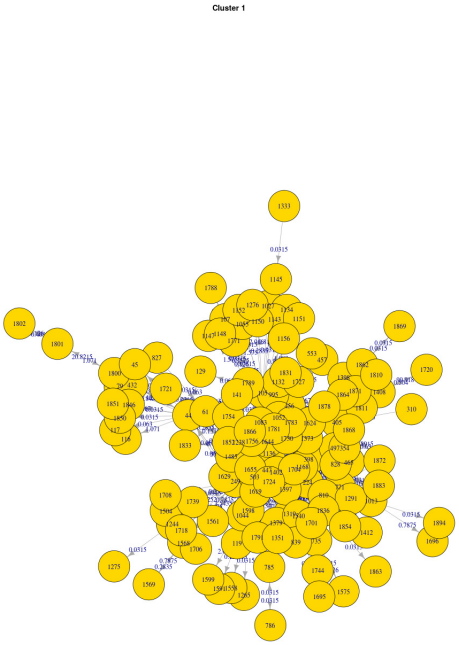

Figure 3.1: Cluster 1

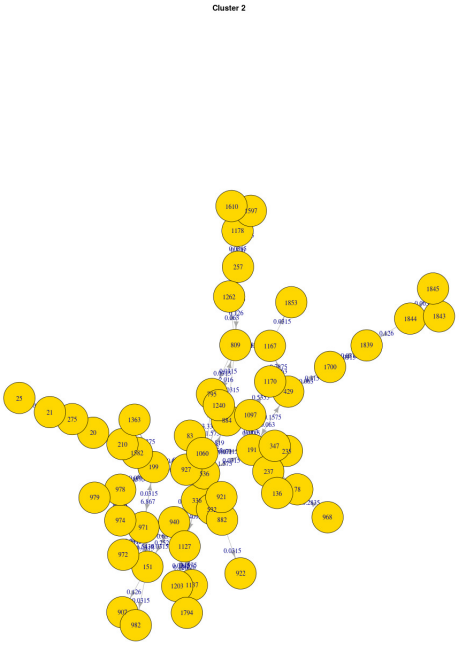

Figure 3.2: Cluster 2

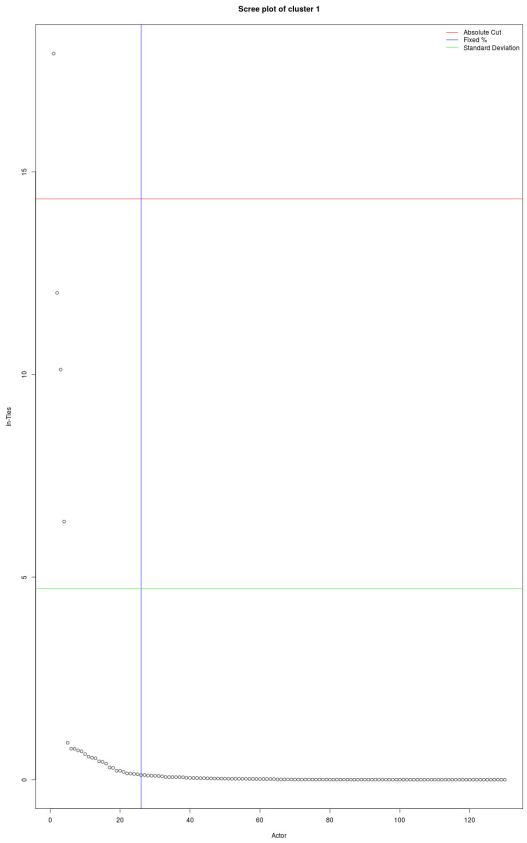

Figure 3.3: Scree Plot 1

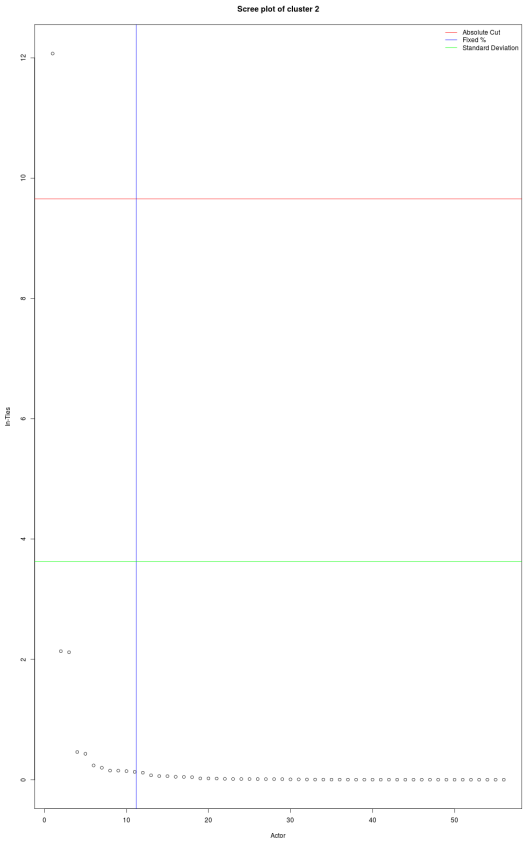

Figure 3.4: Scree Plot 2

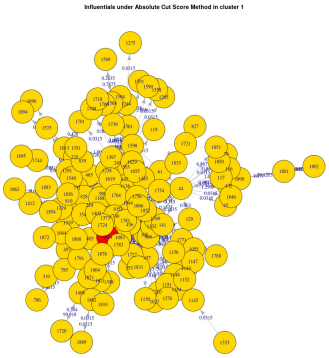

Figure 3.5: Absolute cut score for cluster 1

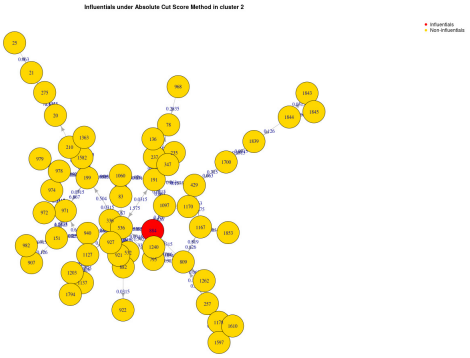

Figure 3.6: Absolute cut score for cluster 2

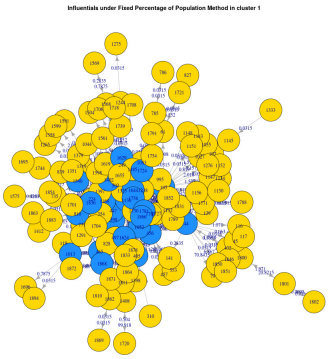

Figure 3.7: Fixed percentage of population for cluster 1

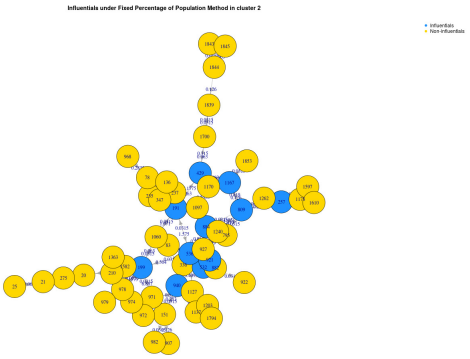

Figure 3.8: Fixed percentage of population for cluster 2

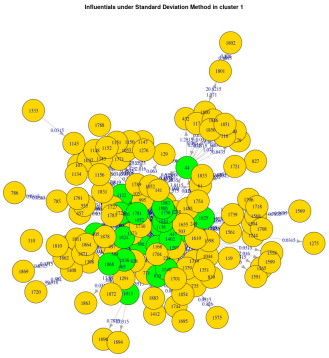

Figure 3.9: Standard deviation for cluster 1

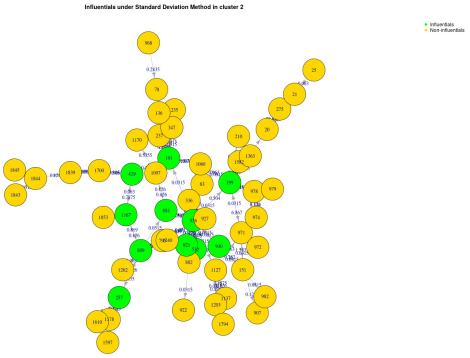

Figure 3.10: Standard deviation for cluster 2

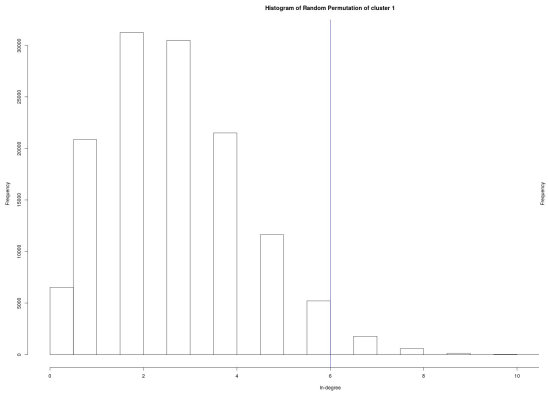

Figure 3.11: Random Permutation Histogram for cluster 1

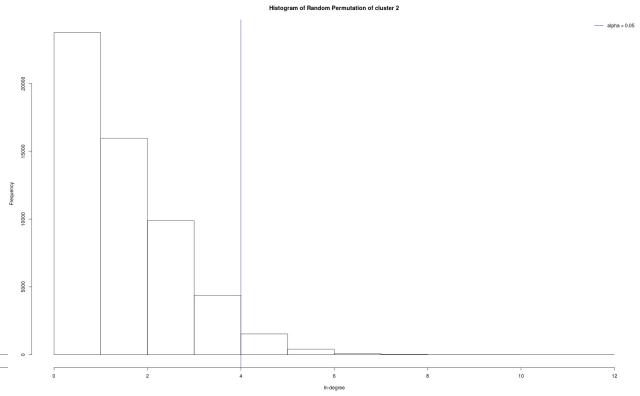

Figure 3.12: Random Permutation Histogram for cluster 2

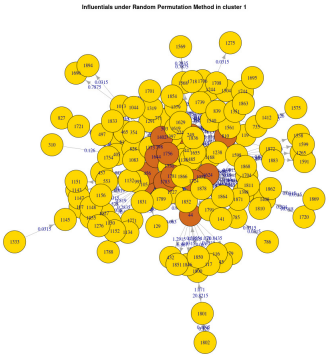

Figure 3.13: Random Permutation for cluster 1

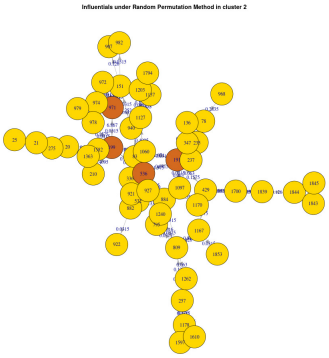

Figure 3.14: Random Permutation for cluster 2

3.3 Link Prediction

For the network given in (3.1) the probability of existence of an edge between any pair of nodes is computed as given in (2.3). The result of this analysis is shown in the table given below.

Table 3.2: Probabilities for edges

from	to	probability
11	14	0.07272727
11	15	0.07272727
14	11	0.07272727
14	15	0.07272727
15	11	0.07272727
15	14	0.07272727
11	13	0.04848485
11	44	0.04848485
11	66	0.04848485
13	11	0.04848485
⋮	⋮	⋮

3.4 Time Series Analysis

For a social media dataset collected over time, time series analysis is performed to try and explore how this network reacts over time and how it behaves with the addition of new users, or which users tend to interact among each other.

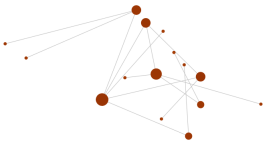

Figure 3.15: At first time instance

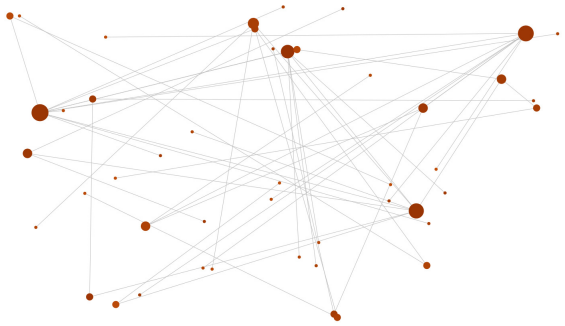

Figure 3.16: At second time instance

Chapter 4

Conclusion and Future Work

Novel concepts such as Small World Phenomenon, Power Law, Link prediction and Cluster analysis have enabled the understanding of the importance and significance of social networks by successfully processing datasets and deriving results. The same has also been implemented as a RShiny application. The challenge is that complex datasets in terms of its features and size need to be investigated so that the relevance and behavior of social network can be thoroughly understood.

Weighted networks serve as a suitable input for rigorous analysis techniques such as Influence analysis, Link prediction and Time series analysis. Influence analysis has enabled the identification of influential individuals based on statistically significant measures with greater efficiency. Improved link prediction methodology taking into account the structural properties of the network has improved the accuracy with which links are predicted not only in local communities but also in the network as a whole. Time series analysis has provided us with aesthetically better comprehension of the rate of growth of the social network and the properties that a network should posses so as to be considered a strong social entity.

The aim now is to analyse other algorithms in a similar fashion and perform a comparison in its implementation and determine which of these is the most suitable and efficient approach. Better simulation techniques are researched so that the influential social entities present in the social network can be identified with greater ease. These approaches will be developed such that networks containing millions of nodes can be analysed in a parallelised environment.

Bibliography

[1] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004. 3

[2] S Brin and L Page. Anatomy of a large-scale hypertextual web search engine. 7th intl world wide web conf. 1998. 4

[3] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004. 4

[4] Russell Cole and Michael Weiss. Identifying organizational influentials: Methods and application using social network data. *Connections*, 29(2):45–61, 2009. 4

[5] Souvik Debnath, Niloy Ganguly, and Pabitra Mitra. Feature weighting in content based recommendation system using social network analysis. In *Proceedings of the 17th international conference on World Wide Web*, pages 1041–1042. ACM, 2008. 4

[6] David R Hunter, Mark S Handcock, Carter T Butts, Steven M Goodreau, and Martina Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software*, 24(3):nihpa54860, 2008. 6

[7] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010. 7, 8

[8] Elie Raad, Richard Chbeir, and Albert Dipanda. User profile matching in social networks. In *Network-Based Information Systems (NBiS), 2010 13th International Conference on*, pages 297–304. IEEE, 2010. 3

[9] Nicola Santoro, Walter Quattrociocchi, Paola Flocchini, Arnaud Casteigts, and Frederic Amblard. Time-varying graphs and social network analysis: Temporal indicators and metrics. *arXiv preprint arXiv:1102.0629*, 2011. 7

[10] John Tang, Mirco Musolesi, Cecilia Mascolo, and Vito Latora. Temporal distance metrics for social network analysis. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 31–36. ACM, 2009. 8

[11] Riitta Toivonen, Jussi M Kumpula, Jari Saramäki, Jukka-Pekka Onnela, János Kertész, and Kimmo Kaski. The role of edge weights in social networks: modelling structure and dynamics. In *SPIE Fourth International Symposium on Fluctuations and Noise*, pages 66010B–66010B. International Society for Optics and Photonics, 2007. 3

[12] Jorge Carlos Valverde-Rebaza and Alneu de Andrade Lopes. Link prediction in online social networks using group information. In *Computational Science and Its Applications–ICCSA 2014*, pages 31–45. Springer, 2014. 7