B.Sc. in Computer Science and Engineering Thesis

Community Detection Algorithm for Social Networking : FaNClust

Submitted by

Md Imraul Quyes Imrul
201114003

Syed Rijuan Rubaiyat Rahman
201114005

M. Nafeh Bin Mosharraf
201114021

Faisal Mahmud
200914044

Supervised by

Md. Mahboob Karim

Instructor Class-A

Department of Computer Science and Engineering

Military Institute of Science and Technology

Department of Computer Science and Engineering
Military Institute of Science and Technology

December 2014

CERTIFICATION

This thesis paper titled **"Community Detection Algorithm for Social Networking : FaN-Clust"**, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in December 2014.

Group Members:

Md Imraul Quyes Imrul

Syed Rijuan Rubaiyat Rahman

M. Nafeh Bin Mosharraf

Faisal Mahmud

Supervisor:

Md. Mahboob Karim
Instructor Class-A
Department of Computer Science and Engineering
Military Institute of Science and Technology

CANDIDATES' DECLARATION

This is to certify that the work presented in this thesis paper, titled, "Community Detection Algorithm for Social Networking : FaNClust", is the outcome of the investigation and research carried out by the following students under the supervision of Md. Mahboob Karim, Instructor Class-A, Department of Computer Science and Engineering, Military Institute of Science and Technology and Md. Shamsur Rahman, Doctoral Student (Ph.D. Candidate), Clayton School of Information Technology, Monash University, Clayton, Victoria, Australia .

It is also declared that neither this thesis paper nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Md Imraul Quyes Imrul
201114003

Syed Rijuan Rubaiyat Rahman
201114005

M. Nafeh Bin Mosharraf
201114021

Faisal Mahmud
200914044

ACKNOWLEDGEMENT

We are thankful to Almighty Allah for his blessings for the successful completion of our thesis. Our heartiest gratitude, profound indebtedness and deep respect go to our supervisor, Md. Mahboob Karim, Instructor Class-A, Department of Computer Science and Engineering, Military Institute of Science and Technology, for his constant supervision, affectionate guidance and great encouragement and motivation. His keen interest on the topic and valuable advices throughout the study was of great help in completing thesis.

We are thankful to our co-supervisor Md. Shamsur Rahman, Doctoral Student (Ph.D. Candidate), Clayton School of Information Technology, Monash University, Clayton, Victoria, Australia for his continuous guidance & support.

We are especially grateful to the Department of Computer Science and Engineering (CSE) of Military Institute of Science and Technology (MIST) for providing their all out support during the thesis work.

Finally, we would like to thank our families and our course mates for their appreciable assistance, patience and suggestions during the course of our thesis.

Dhaka
December 2014

Md Imraul Quyes Imrul

Syed Rijuan Rubaiyat Rahman

M. Nafeh Bin Mosharraf

Faisal Mahmud

ABSTRACT

The problem of community detection in social media has been widely studied in the social networking community in the context of the structure of the underlying graphs. Most community detection algorithms use the links between the nodes in order to determine the dense regions in the graph. These dense regions are the communities of social media in the graph. Such methods are typically based purely on the linkage structure of the underlying social media network. Community detection algorithms are fundamental tools that allow us to uncover organizational principles in networks. When detecting communities, there are two possible sources of information one can use: the network structure, and the features and attributes of nodes. Even though communities form around nodes that have common edges and common attributes, typically, algorithms have only focused on one of these two data modalities: community detection algorithms traditionally focus only on the network structure, while clustering algorithms mostly consider only node attributes.

In this paper, we explore a range of network community detection methods in order to compare them and to understand their relative performance and the systematic biases in the clusters they identify. We evaluate several common objective functions that are used to formalize the notion of a network community, and we examine several different classes of approximation algorithms that aim to optimize such objective functions. In addition, rather than simply fixing an objective and asking for an approximation to the best community of any size, we consider a size-resolved version of the optimization problem. Considering community quality as a function of its size provides a much finer lens with which to examine community detection algorithms, since objective functions and approximation algorithms often have non-obvious size-dependent behavior. And we propose a new algorithm Fast Network Clustering Algorithm (FaNClust) for better performance.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

CHAPTER 1
INTRODUCTION

Social networking is an important application in recent days. It enables social contact over the internet for geographically dispersed users. A social network can be represented as a graph. Here nodes represent users, and links represent the connections between users. The interest in the field of social networking has resulted a reinforcement of graph mining algorithms. So, many techniques have recently been designed for various graph mining and management problems.

1.1 Background

1.1.1 Social Network

In the area of social networking, community detection is an important problem. In this problem, the goal is to partition the network into dense regions of the graph. Such dense regions correspond to entities which are closely related. They are said to belong to a community. The determination of such communities is useful in the context of applications in social-network analysis, including customer segmentation, recommendations, link inference, vertex labeling and influence analysis. A considerable amount of research has been devoted towards algorithms for solving this problem.

Identifying network communities can be viewed as a problem of clustering a set of nodes into communities. Communities help us to discover groups of interacting objects and the relations between them. For example, in social networks, communities correspond to groups of friends who attended the same school, or who come from the same hometown. In protein interaction networks, communities are functional modules of interacting proteins. Identifying network communities allows us to discover functionally related objects.

Edges provide a pattern of community behavior. This models the characteristics of pair wise interactions rather than individual actors. In general, pair wise interaction content provides very specific information about the nature of the relationship between a particular pair of individuals. The different kinds of interactions of a single individual may be used in order to reflect their membership in different communities. Figure 1.1 illustrates an example of a social media network. The nodes represent users. The edges represent the favored images shared by the users. In this example, it is evident that the content information associated with the edges can be naturally categorized into two types, corresponding to the family and the folk music themes. This naturally induces two kinds of edge-based interaction groups. Thereby we can create interesting communities.

The nodes represent users while the edges represent the favored images shared by the users. It is intuitively evident that the nodes can be easily partitioned into the family and folk music groups.

When a community contains edges which are connected with similar content, and are also linked together, the interest area or expertise of a community may also be identified on this basis. This can be useful when it identifies subject matter that is most relevant to the community. Such an approach can be very useful in problems such as expertise search.

This paper will design a unique approach for community detection by tightly integrating the structural and content aspects of the network.

1.1.2 Protein complex and Functional Modules

Protein complex are groups of proteins that interact with each other at the same time and place, forming a single multimolecular machine [2]. These are the form of quaternary structure of proteins. Identified protein complexes include several large transcription factor complexes, the anaphase promoting complex, RNA splicing and polyadenylation machinery,

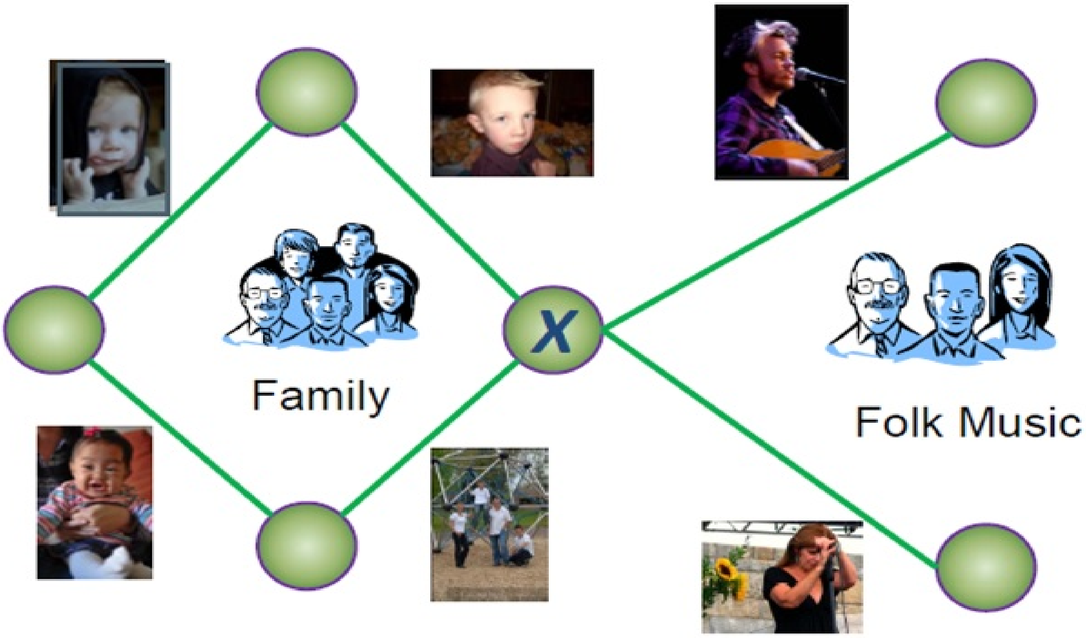

Figure 1.1: Illustration of a social media network

protein export and transport complexes etc. Protein complexes of Bakers yeast is shown in Figure 1.2.

Functional modules are consisted of proteins that participate in a common elementary biological process while binding each other at a different time and place (different conditions or phases of the cell cycle, in different cellular compartments etc.) [2]. Example of identified functional modules including the CDK/ cyclin module responsible for cell-cycle progression, the yeast pheromone response pathway, MAP signaling cascades etc. A 3D structural view of hyperclique pattern of functional modules within a protein complex is shown in Figure 2.1. It is very important to remember, functional modules contain multiple protein complexes [3, 4]. On the other hand, protein complexes carry out a specific task, but functional modules carry out a set of tasks which are carried out by individual protein complexes [4].

1.1.3 Related works

Empirical Comparison of Algorithms for Network Community Detection

A great deal of work has been devoted to finding communities in large networks, and much of this has been devoted to formalizing the intuition that a community is a set of nodes that has more and/or better links between its members than with the remainder of the network.

Very relevant to our work is that of Kannan, Vempala, and Vetta [5], who analyze spectral algorithms and describe a community concept in terms of a bicriterion depending on the conductance of the communities and the relative weight of between-community edges. Flake, Tarjan, and Tsioutsiouliklis [6] introduce a similar bicriterion that is based on network flow ideas, and Flake et al. [4] defined a community as a set of nodes that has more edges pointing inside the community than to the rest of the network. Similar edge-counting ideas were used by Radicchi et al. [**?**] to define and apply the notions of a strong community

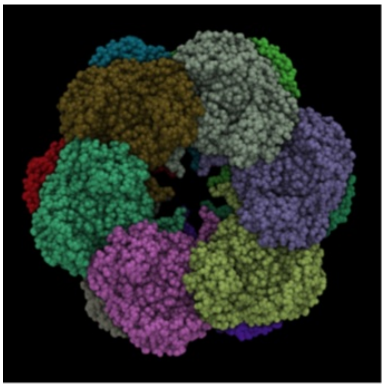

Figure 1.2: Protein complexes of Bakers yeast

and a weak community.

Within the complex networks community, Girvan and Newman [7] proposed an algorithm that used betweenness centrality to find community boundaries. Following this, Newman and Girvan [8] introduced modularity as an a posteriori measure of the overall quality of a graph partition. Modularity measures internal (and not external) connectivity, but it does so with reference to a randomized null model. Modularity has been very influential in recent community detection literature, and one can use spectral techniques to approximate it [9, 10]. However, Guimer, Sales-Pardo, and Amaral [11] and Fortunato and Barthlemy [12] showed that random graphs have high-modularity subsets and that there exists a size scale below which modularity cannot identify communities.

Community Detection with Edge Content in Social Media Networks

The problem of community detection has also been studied in the context of many graph-theoretic clustering algorithms. In its simplest form, a community may be considered as a group of nodes which are densely connected by edges. For example, a variety of node clustering algorithms for graphs with the use of shingling techniques, matrix co-clustering techniques, and tile determination in matrices [1,4] can be used for community detection in graphs. The problem is also related to that of finding dense cliques or dense regions in the underlying graph [2, 7, 13]. These techniques are designed for generic graphs rather than the specific case of social networks. The problem of community detection [11, 12, 14–16] in social networks has also been widely studied because of the increasing importance of social networking applications. A survey of a number of important algorithms for community detection is provided in [16]. Discussion of important statistical properties of web communities is discussed in [15]. A second related line of research is to use purely content based clustering methods [3,5,17,18]. However, such methods miss the rich information which is often encoded in the links in the underling network. Some recent work [19,20] uses a combination of relational attributes and link information for clustering purposes. However, this method is designed for the case when the attributes are associated with the nodes rather than the edges. Some research [21] has been performed for visualizing the social network when the content is associated with the edges. The technique is designed to provide an intuitive visual understanding, and provides a good understanding of how the different regions in the various modes of the network relate to one another. However, it is not specifically designed for determining communities in an automated way with clear objective criteria.

1.2 Objectives

The objectives of the thesis are as follows,

To design a hierarchical algorithm to improve the community detection processes for networks

No hierarchical method can solve the problem of classifying the vertices of degree one. In this thesis, we have proposed a new agglomerative approach of hierarchical method to solve the problem of classifying nodes containing one neighbor by using Relative vertex-to-vertex clustering value. As well as our proposed algorithm has produced/ discovered more dense subgraphs in networks than previous hierarchical algorithms.

To design a faster method for hierarchical approach

In 2011, Wang et al. [**?**] proposed a faster agglomerative hierarchical method for clustering PINs. The worst case time complexity of their algorithm is $O(\bar{d}^2m)$ where m is the number of interaction and $\bar{d}$ is the average degree of any network G. It is the fastest algorithm so far published. On the other hand, we have proposed an agglomerative algorithm which is known as FAC-PIN algorithm. The worst case time complexity of FAC-PIN algorithm is $O(\bar{d}^2n)$. In any protein interaction network, the number of proteins n is smaller than the number of interactions m.

1.3 Thesis organization

We organize our thesis into another four chapters. In Chapter 2, we discusse the preliminaries, representation related to the social network. We describe our proposed premetric relative vertex-to-vertex clustering values and new agglomerative algorithm: FaNClust in the Chapter 3. After designing the algorithm, we carry out computation experiments on several network datasets. We discuss the computation experiments and results in the Chapter 4. Finally in Chapter 5, we conclude our thesis with the discussion of social networking using FaNClust algorithms and its future works.

CHAPTER 2
PRELIMINARIES

2.1 Related Definitions

Network representation of proteins and their interactions are known as Protein Interaction Network [8]. In short it is called PIN. In PINs, proteins are represented as nodes or vertices and interactions are as edges. Maximum PINs are undirected networks with edge weight or not [8,12]. In Figure 2.2, an unweighted PIN of bakers yeast is shown. Girvan and Newman [22] and Fortunato [1] discuss about the five properties of protein interaction networks in their papers.

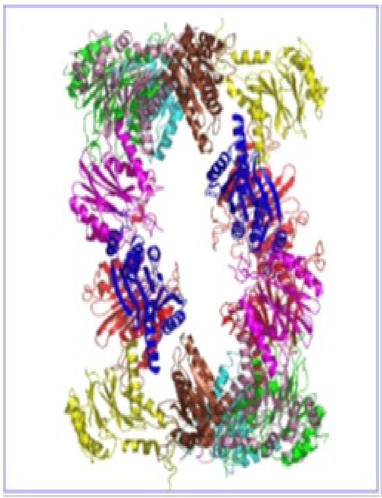

Figure 2.1: Hyperclique pattern of functional modules in a protein complex

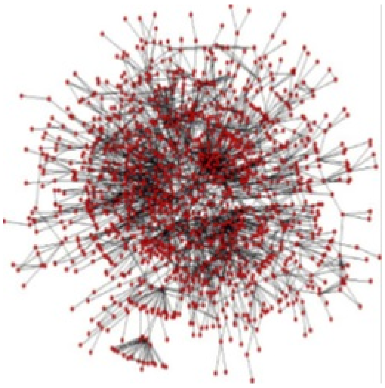

Figure 2.2: Protein Interaction Network of Bakers yeast

Small world effect which is the name given to the finding that the average distance between vertices in a network is small.

Power law degree distribution is a distribution where the number of the vertices with low degree is higher than the number of vertices with high degree.

Network transitivity is a property that two vertices that are both neighbor of same third vertex have a heightened probability of also being neighbor of one another.

Community structure is a property where intrales or both.-connectivity of a subset of vertices of graph G is higher than inter-connectivity between others.

Preferential attachment is a property where a new node u is likely to attach to a high degree node v than to a low degree node.

In PINs, all protein complexes and functional modules are strong subgraphs [8]. To identify the protein complexes or functional modules from PINs means strong subgraphs, authors of the algorithms were used any of five properties. Third and fourth properties are commonly used to discover protein complexes or functional modules. But unfortunately, fifth property have not still used by any authors which helps to identify the more significant strong subgraph having biological significance.

2.2 Community

A community is defined as a subgraph (a subset of vertices of graph G) within the graph G such that connections inside the subgraph are denser than connections with the rest of the network [19]. Luo et al [21] gave the more formal definition of community. Their definition is as followed-

Community U is a subgraph of a graph G in which in-degree of U is higher than out-degree and the ratio of in-degree and out-degree of U should be higher than 1.

In-degree of a community U is the number of edges connected between the vertices of community U and out-degree of a community U is the number of edges between other communities and U. From the formal definition of community, two properties of the community are revealed-

Homogeneity: Vertices of a community are highly similar or compact to each other. **Separability:** Vertices of different communities have lower similarity or compactness.

On the other hand, inhomogeneity or separability property suggests that the network has certain natural divisions within it. The communities are often defined in terms of the partition of the set of vertices, that is each node is put into either only one community just as in the

Figure 2.4 or into multiple communities. Depends on the distribution of the nodes among the communities, community can be classified into two groups-

Overlapping communities share one or more common nodes among them. In Figure 2.4, yellow, green and purple colored communities are sharing red colored vertices. These communities are the examples of overlapping communities.

Non-overlapping communities do not share any node between them. In Figure 2.4, blue and purple; blue and yellow colored communities do not share a single vertex between. So, these communities are the example of non-overlapping communities.

In the Figure 2.4, blue and green colored communities are not connected by any edge. These communities are known as disjoint communities. Moreover, Radicchi et al. [19] also classified the communities into two groups according their connectivities:

Strong community is a community U in which in-degree7 of all vertices are higher than out-degree.

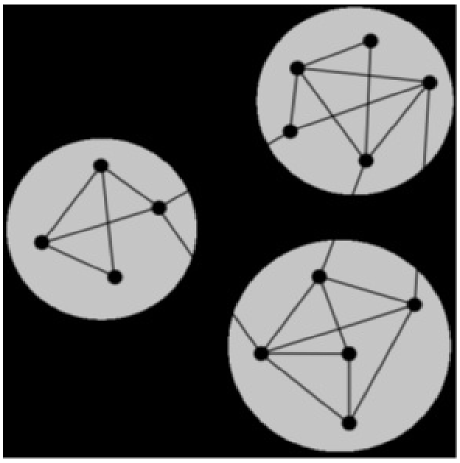

Figure 2.3: Community structure of a graph G

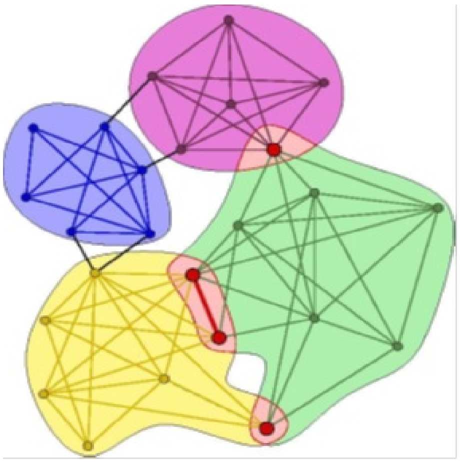

Figure 2.4: Overlapping and non-overlapping communities of a graph G

Weak community is a community U in which in-degree of some vertices are higher than out-degree.

In PINs, protein complexes and functional modules are formed by interacting proteins. PINs organize into densely linked complexes where interactions appear with high concentration among the proteins of the complex [23]. It indicates the protein complexes or functional modules are the communities in PINs in respect to the network and community definition. Generally, in PINs, the number of interactions are very large than the number of proteins, like Figure 2.2. It is not easy and simple to identify the protein complexes or functional modules. Some computational methods are required for detecting protein complexes or functional modules from PINs. Community detection algorithms are very common to identify the complexes or modules from PINs.

2.3 Representation

Network

A network is a group of two or more system or entities linked together. Network is represented by a graph.

Social Network

Social networking is the practice of expanding the number of ones business and social contacts by making connections through individuals. While social networking has gone on almost as long as societies themselves have existed, the unparalleled potential of the internet to promote such connections is only now being fully recognized and exploited, through web-based groups established for that purposes.

Node

In communication network node is either a connection point, a redistribution point or a communication end point. The definition of node depends on the network and protocol layer referred to.

Edge

In network, any connection between entities is represented by an edge. Edges connect two points together. Edge-based content is much more challenging, because the different interests of the same actor node may be reflected in different edges.

2.4 Community Detection Algorithm

Community detection in PINs is a computationally hard task. Conventional clustering algorithms are not well suited for this task [9]. Efficient, accurate, robust, and scalable methods are therefore required for mining large PINs. There are three approaches of community detection methods according to their working principles [1].

Density based technique finds the subgraphs in the network whose density is higher [1]. But this method cannot find the communities or clusters efficiently for scale free networks [1], see Figure 2.6). Moreover all PINs are scale free networks. For this reason, density based algorithms are not used in clustering of PINs [1].

Graph partition techniques find the bridge edges which connect the communities. By removing bridge edges, these algorithms discover the communities [7]. These algorithms are very efficient, but suffered by execution time.

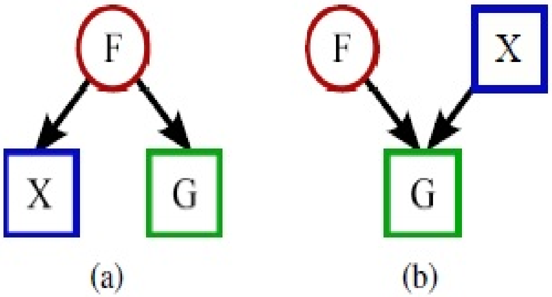

Figure 2.5: Two ways of modeling the statistical relationship between a graph G, attributes X, and communities F. Circles represent latent variables that need to be inferred and squares represent manifest (observed) variables

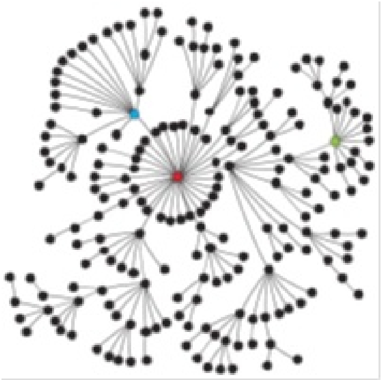

Figure 2.6: Scale Free Network G [1]

Hierarchical method finds the communities by calculating similarity or compactness between the nodes [1]. But this method cannot classify the vertices of degree one in same community with their neighbors which does not make sense biologically [1]. Time complexity is another problem of this method.

In this thesis, we put our emphasis on the problems of hierarchical method. We have designed a new algorithm which is known as FaNClust algorithm to solve the problems of hierarchical method.

2.5 Scoring Function

Scoring functions are used for evaluating the output of any algorithm. The results also allow us to compare the algorithms. All scoring functions build on the intuition that communities are sets of nodes with many connections between the members and few connections from the members to the rest of the network. We will use three most common scoring functions to compare the outputs of the algorithms.

2.5.1 Modularity

If a clustering result is represented by $\{P_k\} = \{C_1, ..., C_k\}$ with k clusters, we can use a popular scoring function which was introduced by Newman and Girvan, Modularity Measurement. It can be defined as,

$Q(P_k) = \sum_{i=1}^{k}\{e_{ii} - {a_i}^2\}$

where,

e_{ii} = Fraction of edges with both end vertices in the same community i

a_i = Fraction of edges with at least one end vertex in community i

Larger values of Q indicate greater distinctiveness in the community structures of the social networks. For simplifying the structure we considered unweighted connections between nodes in the network.

2.5.2 w-log-v

Modularity measure faces some resolution problem. Thus, we used another scoring function for evaluating our algorithm results, w-log-v, which was proposed in [24]. The equation can be represented as,

w-log-v = $\sum_{i=1}^{k}\{e_{ii} - log a_i\}$

where,

e_{ii} = Fraction of edges with both end vertices in the same community i

a_i = Fraction of edges with at least one end vertex in community i

2.5.3 Cut ratio

This is the fraction of edges leaving the community or cluster. The formula for this ratio is given by the number of edges on boundary of the community, c_s and number of nodes in community S, n_s. For a community, with a set of nodes S we can denote cut ratio as,

$f(S) = \frac{c_s}{n_s(n-n_s)}$

here,

n = Number of total nodes in the network

n_s = Number of nodes in the community S

c_s = Number of edges in the boundary of community S

The result of these scoring functions let us compare the algorithms and their outputs. We implemented FAC-PIN, FPNC and FanClust algorithms on three different dataset from [25]. The datasets are,

1. Copperfield Word Adjacencies

2. Dolphin Social Network

3. Zachari's Karate Club

CHAPTER 3
PROPOSED ALGORITHM

3.1 FaNClust Algorithm

Our proposed algorithm is based on the relative vertex-to-vertex clustering value of a network graph. The clustering value can be determined by the following formula,

$R_w(u \rightarrow v) = \frac{\sum_{k \in I^+_{u,k}} w(u,k)}{\sum_{s \in N^+_u} w(u,k)}$

Here,

$I^+_{u,k}$ = list of common neighbors between u and k vertices including u and k

N^+_u = list of neighbors of vertex iu including u

for unweighted graphs we can also use the following formula for $R_w(u \rightarrow v)$. That is,

$R(u \rightarrow v) = \frac{|N^+_u \cap N^+_v|}{N^+_u}$

Value of $R_w(u \rightarrow v)$ ranges from 0 to 1. The higher the value is, the higher the possibility for two nodes to be in the same community.

For FAC-PIN and FPNC algorithm designed by Rahman et al. [26] we have considered a random threshold value $\alpha and \Delta Q$. The operation reduces the speed of the algorithms. So we introduced an algorithm by getting rid of the threshold α and ΔQ.

In Algorithm 1 we show the proposed Fast Network Clustering algorithm or FaNClust.

Algorithm 1 The FaNClust Algorithm

```
for any vertex $u \in V$ do
    $cluster(u) \leftarrow u$
end for
for any vertex $u \in V$ do
    for any neighbor of $u, v \in N_u$ do
        Compute $R_w(u \rightarrow v)$
    end for
    find any neighbor v in which $R_w(u \rightarrow v)$ is maximum
    $cluster(v) \leftarrow cluster(v) + cluster(u)$
    $cluster(u) \leftarrow cluster(v)$
end for
```

Complexity of FAC-PIN is $O(nd_{max}^2)$ and FPNC is $O(nk_{max}^2)$ where the complexity of FaNClust is $O(nd_{max})$. So it is faster than both of the algorithms and gives an efficient result.

3.2 Previous Algorithms

We have tested our sample datasets with two previously devised algorithms by Rahman et al. [26], FAC-PIN and FPNC. Both of the algorithms are designed for protein-protein-interaction network. These networks are similar to social networks. So we can apply both of these on social network graphs and get valid output.

We have considered three very popular datasets available online and generated output and evaluated with the fore-mentioned scoring functions. Result shows us that, in most of the cases, our algorithm, FaNClust runs faster than the rest of the two algorithms.

CHAPTER 4
RESULTS, COMPARISONS AND DISCUSSION

For experiment we have used three popular network datasets from [ref website]. The datasets are -

1. Copperfield word adjacencies

2. Dolphin Social Network

3. Zachari's Karate Club

We have used FAC-PIN, FPNC and our proposed algorithm FaNClust for evaluating our result. The output of the algorithms were then evaluated and compared using the scoring functions described in Chapter 2.

Modularity:

$Q(P_k) = \sum_{i=1}^{k}\{e_{ii} - {a_i}^2\}$

where,

e_{ii} = Fraction of edges with both end vertices in the same community i

a_i = Fraction of edges with at least one end vertex in community i

From the table 4.1 we can see that for bigger network, FaNClust gives a higher modularity result. So the community detection will be much faster than FPNC and FAC-PIN for larger networks.

w-log-v:

w-log-v = $\sum_{i=1}^{k}\{e_{ii} - loga_i\}$

where,

Table 4.1: Modularity Comparison for the algorithms

	Copperfield Word Adjacencies	Dolphin Social Network	Zachari's Karate Club
FaNClust	0.384	0.298	0.128
FPNC	0.296	0.225	0.204
FAC-PIN	0.281	0.247	0.194

Table 4.2: w-log-v Comparison for the algorithms

	Copperfield Word Adjacencies	Dolphin Social Network	Zachari's Karate Club
FaNClust	0.882	0.891	0.673
FPNC	0.866	0.831	0.682
FAC-PIN	0.835	0.726	0.637

Table 4.3: Cut Ratio Comparison for the algorithms

	Copperfield Word Adjacencies	Dolphin Social Network	Zachari's Karate Club
FaNClust	0.514	0.728	0.323
FPNC	0.328	0.521	0.339
FAC-PIN	0.486	0.510	0.295

e_{ii} = Fraction of edges with both end vertices in the same community i

a_i = Fraction of edges with at least one end vertex in community i

Table 4.2 gives a clear view of the efficiency of FaNClust algorithm. It shows the best output for the considered datasets can be found by FaNClust algorithm.

Cut Ratio:

$f(S) = \frac{c_s}{n_s(n-n_s)}$

here,

n = Number of total nodes in the network

n_s = Number of nodes in the community S

c_s = Number of edges in the boundary of community S

For Cut-Ratio we can see that FaNClust gives better results larger networks. Thus it is efficient to use FaNClust for larger networks which can significantly make the process of community detection faster.

CHAPTER 5
CONCLUSION

In this thesis, we present an efficient algorithm for detecting communities in a social network FaNClust or Fast Network Clustering algorithm. In our test we could see that it can detect communities more accurately than previously devised algorithms, FAC-PIN and FPNC. The time complexity of FaNClust is also better than both FAC-PIN and FPNC. Complexity of FAC-PIN is $O(nd_{max}^2)$ and FPNC is $O(nk_{max}^2)$.

Although FAC-PIN and FPNC are devised as clustering algorithms for PPI (protein-protein-interaction) network, they can also be applied for detecting communities in social network as both networks can be represented as graphs. Our proposed algorithm, FaNClust can also be used for PPI networks with same complexity level and generate competitive results. The relative vertex-to-vertex clustering value dependency of the algorithm ensures better dense subgraphs for given networks. For larger networks FaNClust runs faster than FAC-PIN and FPNC. From the results of computing the evaluation of communities, validation we can say that our proposed FaNClust algorithm is faster and more efficient than FAC-PIN and FPNC in detecting communities in social networks.

The agglomerative approach of this algorithm also ensures that for any size of network the algorithm eventually decreases its computational. So it will give better time complexity for bigger and larger networks.

There are still a lot of scopes to expand our work on this FaNClust algorithm. We have listed them as follows:

- The algorithm can be modified to detect protein complexes or functional modules of protein.
- FaNClust has been designed for first order neighbor calculations. It can be improved to consider the second order neighbors in future.
- The algorithm has scope to be improved in community detection and characterize complex communities in today's social network.
- The algorithm can be used to make the search criteria faster.

REFERENCES

[1] S. Fortunato, “Community detection in graphs,” *CoRR*, vol. abs/0906.0612, 2009.

[2] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, “Development and implementation of an algorithm for detection of protein complexes in large interaction networks,” *BMC Bioinformatics*, vol. 7, p. 207, 2006.

[3] J. Chen and B. Yuan, “Detecting functional modules in the yeast protein protein interaction network,” *BMC Bioinformatics*, vol. 22, p. 18, 2006.

[4] A. Gingras, R. Aebersold, and B. Raught, “Advances in protein complex analysis using mass spectrometry,” *Journal of Physics*, vol. 563, p. 1, 2005.

[5] T. van Laarhoven and E. Marchiori, “Robust community detection methods with resolution parameter for complex detection in protein protein interaction networks,” in *Pattern Recognition in Bioinformatics - 7th IAPR International Conference, PRIB 2012, Tokyo, Japan, November 8-10, 2012. Proceedings*, pp. 1–13, 2012.

[6] L. Giot, J. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. Hao, C. Ooi, B. Godwin, and E. Vitols, “Vitols. a protein interaction map of drosophila melanogaster,” vol. 302, p. 1727 1736, 2003.

[7] E. C. Kenley and Y. Cho, “Detecting protein complexes and functional modules from protein interaction networks: A graph entropy approach,” vol. 11, pp. 1116–1121, 2011.

[8] V. Spirin and L. A. Mirny, *Protein complexes and functional modules in molecular networks*, vol. 100. 2003.

[9] S. Yook, Z. Olvai, and A. Barabsi, *Functional and topological characterization of protein interaction networks*, vol. 4. 2004.

[10] J. S. Richardson, *The anatomy and taxonomy of protein structure*, vol. 34. Academic Press, 1981.

[11] A. D. King, N. Przulj, and I. Jurisica, “Protein complex prediction via cost-based clustering,” *Bioinformatics*, vol. 20, no. 17, pp. 3013–3020, 2004.

[12] A.Gursoy, O. Keskin, and N. R, “Topological properties of protein interaction networks from a structural perspective,” *Biochemical Society Transaction*, vol. 36, 2008.

[13] E. Ravasz, A. Somera, and D. Mongru, “Hierarchical organization of modularity in metabolic network,” *Science*, vol. 297, 2002.

[14] H. N. Chua, K. Ning, W. Sung, H. W. Leong, and L. Wong, “Using indirect protein-protein interactions for protein complex prediction,” *J. Bioinformatics and Computational Biology*, vol. 6, no. 3, pp. 435–466, 2008.

[15] J. Hallinan, “Gene duplication and hierarchical modularity in intracellular interaction networks,” *Biosystem*, vol. 74, p. 3, 2004.

[16] M. Newman, “Fast algorithm for detecting community structure in networks,” *Physical and m review*, vol. 69, p. 066133, 2003.

[17] A. Bruce, A. Johnson, J. Lewis, M. Raff, K.Roberts, and P. Walter, “Molecular biology of the cell,” *Garland Science*, 2002.

[18] X. Li, S.Tan, C. Foo, and S. Ng, “Interaction graph mining for protein modulees using local clique merging,” *Genome Informatics*, vol. 16, 2006.

[19] F. Radicchi, C. Castellano, and F. Cecconi, “Defining and identifying communities in networks,” *Proceedings of Natural Academy of Science USA*, vol. 101, p. 9, 2004.

[20] L. Pauling, R. Corey, and H. Branson, “The structure of protein and two hydrogenbonded helical configurations of the polypeptide chain,” *Proceedings of Natural Academy of Science USA*, vol. 37, p. 4, 1951.

[21] F. Luo, Y. Yang, C. Chen, R. L. Chang, J. Zhou, and R. H. Scheuermann, “Modular organization of protein interaction networks,” *Bioinformatics*, vol. 23, no. 2, pp. 207–214, 2007.

[22] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*, vol. 99. 2002.

[23] H. Xiong, P.-N. Tan, and V. Kumar, “Mining strong affinity association patterns in data sets with skewed support distribution.,” in *ICDM*, pp. 387–394, 2003.

[24] T. van Laarhoven and E. Marchiori, “Robust community detection methods with resolution parameter for complex detection in protein protein interaction networks,” in *Pattern Recognition in Bioinformatics - 7th IAPR International Conference, PRIB 2012, Tokyo, Japan, November 8-10, 2012. Proceedings*, pp. 1–13, 2012.

[25] Various, “Uci network data repository.”

[26] M. S. Rahman and A. Ngom, “Fac-pin: Fast agglomerative clustering method for functional modules and protein complex identification in pins.,” in *ICCABS*, pp. 1–6, 2013.