A SEED-CENTRIC ALGORITHM TO DETECT OVERLAPPING COMMUNITIES IN NETWORKS

By

DEEPTHI KASUVAJJALA

Bachelor of Technology in Electronics & Computer Engineering

Jawaharlal Nehru Technological University

Hyderabad, Telangana, India

2012

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December, 2016

A SEED-CENTRIC ALGORITHM TO DETECT OVERLAPPING COMMUNITIES IN NETWORKS

Thesis Approved:

Dr. Cristopher Crick

Thesis Adviser

Dr. Johnson Thomas

Dr. David Cline

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude towards my Adviser Dr. Christopher Crick for his guidance, valuable advice at every phase of the problem formulation, and constant support. I would also like to thank my other committee members, Dr. David Cline and Dr. Johnson Thomas for their valuable comments and advices. I would also like to thank my lab mates for their advices and help, any time I needed. Lastly, I would like to thank my parents for their love and support. I dedicate this thesis to my family.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: DEEPTHI KASUVAJJALA

Date of Degree: DECEMBER, 2016

Title of Study: A SEED-CENTRIC ALGORITHM TO DETECT OVERLAPPING
COMMUNITIES IN NETWORKS

Major Field: COMPUTER SCIENCE

Abstract: Community Detection is a trivial task in Network Analysis and Data Mining. A widely accepted definition of a Community or a Cluster is the group of vertices that are densely connected within and sparsely connected to the other communities. Uncovering the community structure of the network provides useful insights about the structural and functional characteristics of the network of any kind, in consideration. Community Detection is a simpler and computationally less-challenged task, if communities are assumed to be disjoint and non-overlapping. On the contrary, most of the real-world networks exhibit the overlapping community structure rather than the assumed disjoint community structure. Hence the problem of Disjoint Community Detection is now being perceived as Overlapping-Community Detection.

The state-of-the-art is rich in the way it offers various classes of algorithms, yet there is a lot of scope for further exploration of the problem and devising different algorithms based on different metrics. This thesis is inspired by this idea and we propose a new technique to detect the overlapping communities in a network. We further devise a multi-stage algorithm which starts with detecting the seed nodes in the initial stage and converges with the detection of overlapping communities in the network, in the last stage.

Our approach is unique in its choice of the graph/network metric, which is taken into consideration to detect the seed nodes of the network. Or contributions are two-fold. First being, the ability to determine highly-important nodes of the network based on the Betweenness Centrality metric of a node. Second being, detecting densely overlapping communities, thereby optimizing the density measure of a community. Experiments show that our algorithm s successful in scaling to large, real-word networks and is efficient in detection of overlapping communities. The algorithm is compared to other state-of-the-art algorithms and its performance is evaluated.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

1.1 Overview of Community Structure

In many real-world networks, such as Computer and Information networks, Social, Biological, Collaboration, Co-purchasing networks, many properties such as small-world phenomena represented by Fig 1, power-law degree distributions represented by Fig 2 and network transitivity are common to the networks. Apart from these properties, another most common characteristic is the community structure. Though there is no unique definition provided for a community, the most widely accepted definition among the research community is that, in the context of a network, a community refers to the group of nodes that are densely connected internally and are sparsely connected with the other groups of the network. Fig 3 represents the community structure in a network.

Community structure arises from the natural grouping phenomena that occurs subjective to the network's activities. It is a dynamic property of the network and the community assignments change as the network evolves over time. Community structure is not only a structural/topological aspect of a network but is also related to the functionality, behavior of the network.

Community structure is very subjective to the type of network in consideration. Thus, understanding the dynamics of the structure and correlating it to the network functions is often an important and challenging task. In the present era of big data, it is practically not possible to cater to every individual in a network. Hence, grouping the nodes helps in overcoming this problem. Community structure in social and biological networks is discussed in [1].

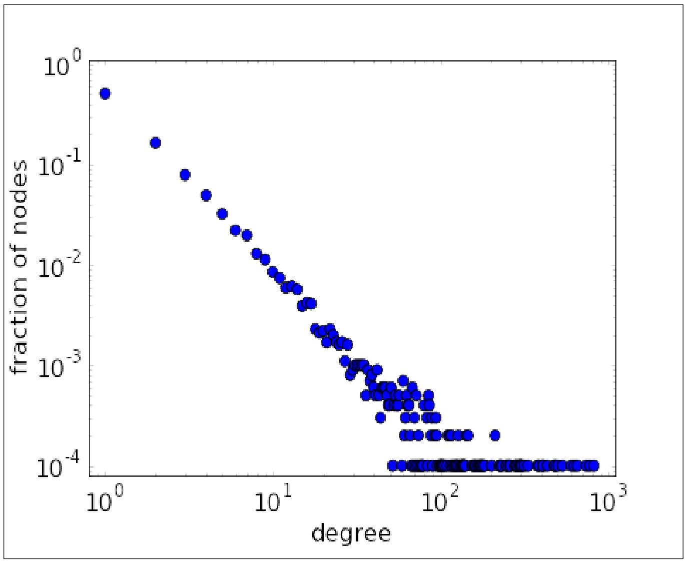


Figure 1: Power-law Degree distribution of nodes.

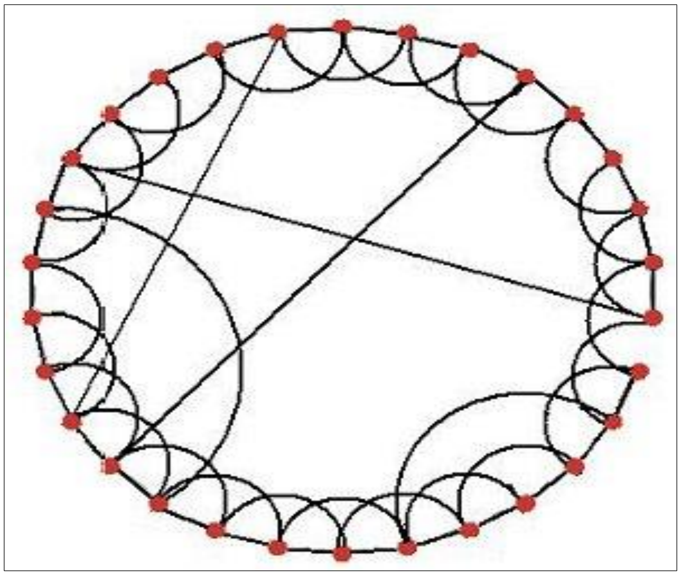


Figure 2: Small- World Phenomena.

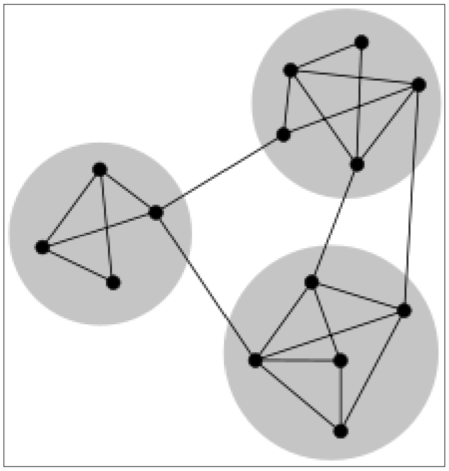


Figure 3: Community Structure in a Network.

1.2 Applications of Community Detection

In many practical settings, applications, understanding the structure of the network is very important to be able to design a marketing strategy, a recommendation model, to predict the network activities etc. For ex: a marketing strategy has to have its targeted set of nodes/people in a network so that it succeeds in marketing their company, product or an idea.

In social settings, which are represented by the online social networking websites like Facebook, Google+, Twitter, Friendster, Myspace etc., to study every individual's properties and behavior is practically not possible. Hence grouping people into communities helps in understanding the activities, interests, needs of the people on a global level.

In ecological, food, and biological networks, community structure detection helps to understand the growth patterns of certain species, the protein-protein interactions etc.

In collaboration networks, a community structure provides an understanding of the world-wide collaboration patterns among the researchers. It also helps in determining the widely studied research problems at a certain period of time.

Community structure in Citation Networks, provides an understanding of the referencing patterns of a certain research problems and also determines the most cited publications, papers in a particular research problem.

Thus, different networks have different kinds of applications that require an understanding of the community structure of those networks. Hence, the problem is of utmost practical importance.

1.3 Disjoint Communities vs. Overlapping Communities

In the initial studies, communities were often defined as the disjoint set of nodes, which means that every nodes strictly belongs to one community. Fig 3 depicts the disjoint community structure where no nodes belong to various communities at the same time. This is often an assumption, useful simplification that is helpful to make the community detection task, less computationally-challenging.

Most of the community detection methods find this type of community structure of a network. [2]- [5]. [1] was one of the very first attempts in understanding the community structure in networks. It proposes a method to detect communities in food, collaboration networks. It is considered as a benchmark algorithm for community detection. There are several other algorithms that were proposed later and are also widely accepted and are the benchmark techniques, mentioned in [13].

However, in real-world, it is very common that a node performs more than one activity and more likely to have multiple interests. For example, a person in a social network usually has multiple interests and connects to several social groups like family, friends, colleagues etc. A researcher may collaborate with researchers in several research areas and at various geographical locations. Hence, disjoint community assignments may mislead us in understanding the complete user-profile and results in partial knowledge acquirement. Hence, the better representation is the one in which nodes belong to more than one community. Fig 4 depicts the aforementioned difference between the two types of communities.

Even in the case of overlapping community assignments, there are two types of assignments that are possible: crisp (non-fuzzy) or fuzzy assignment [6]. In the case of crisp assignment, each vertex belongs fully to each community for which it is a member. With crisp assignment, the relationship between a node and a community is binary. This means, a node either belongs to the cluster or does not. Whereas, in fuzzy assignment, each node belongs to a community to a different extent. This extent is determined by the belonging factor (soft assignment).

However, by setting a threshold, a fuzzy assignment can be converted to a crisp assignment. In general, most of the algorithms, of any class, output crisp overlapping community assignments [7]-[10].

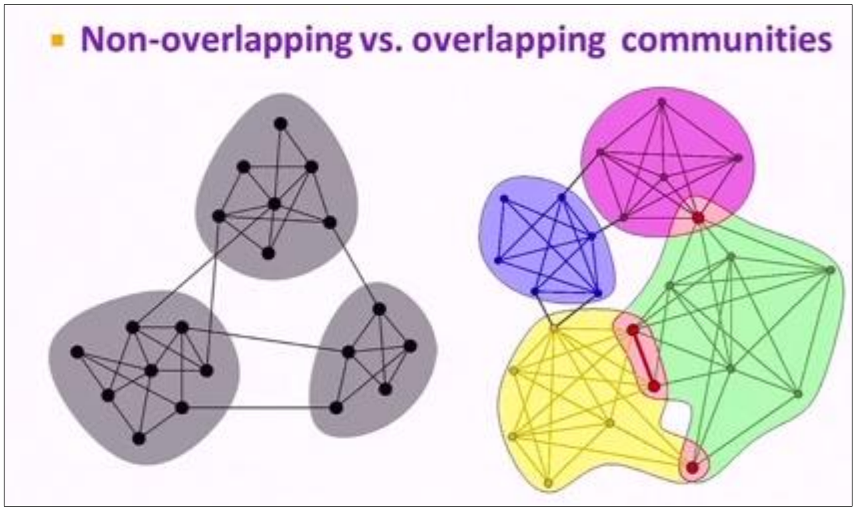


Figure 4: Difference between Disjoint and Overlapping Communities.

1.3.1 Importance of Overlapping Nodes

Detecting Overlapping Communities inherently involves detecting the nodes that are responsible for the overlap to occur. The main aim of Overlapping Community detection is to get to know these nodes which represent multiple communities at once and are quite complex and important user profiles of any network.

Building successful user-profiles for a social, organizational, information network requires the knowledge of the nodes that have these characteristics. Particularly, in Information Networks, these nodes play a crucial role in information diffusion and generation, since they are highly active in the network's activities.

The overlapping nodes have both structural and functional significance in a network. Thus, a high-level clustering often misses out on the functional significance of a node in the network. Hence overlap is indeed a significant feature of many real-world networks. For all these reasons, there is growing interest in overlapping community detection algorithms and techniques.

1.4 Seed Detection based on graph/network metrics

Seeds are the initial set of nodes that are used to determine the community structure in a network. A major class of overlapping community detection algorithms are seed-centric approaches. A seed node is prominent both structurally and functionally. Fig 5 indicates the seed nodes of a graph, determined according to a certain node-level metric.

Seeds are especially useful for local community detection and expansion strategies. They are often associated to distinct graph/network metrics. The importance of the metric in the context of a community, is also associated with the seed.

Each graph metric when semantically correlated with the network in study, brings out a significant insight to understand the network's functionality. For example, a seed node with high out-degree provides us an understanding that the seed node is crucial in information generation and its outward flow within the community/network.

Likewise, a seed node with more centrality within the community is the most desirable node to represent and coordinate the activities of the community. [11] provides a good overview of the various graph metrics that are useful to determine the metric of importance in the process of devising the overlapping community detection algorithm.

The choice of the seed nodes determines the quality of the clusters/communities identified by the algorithm and thus seeding phase is very crucial. The main aim of seeding and overlapping community detection in the context of structure of a graph or community is that the graph metric needs to be optimized by the algorithm. For example, a community detection needs to be optimize any of the metrics such as conductance, modularity, density etc. and new technique should preserve these qualities pertaining to the community structure.

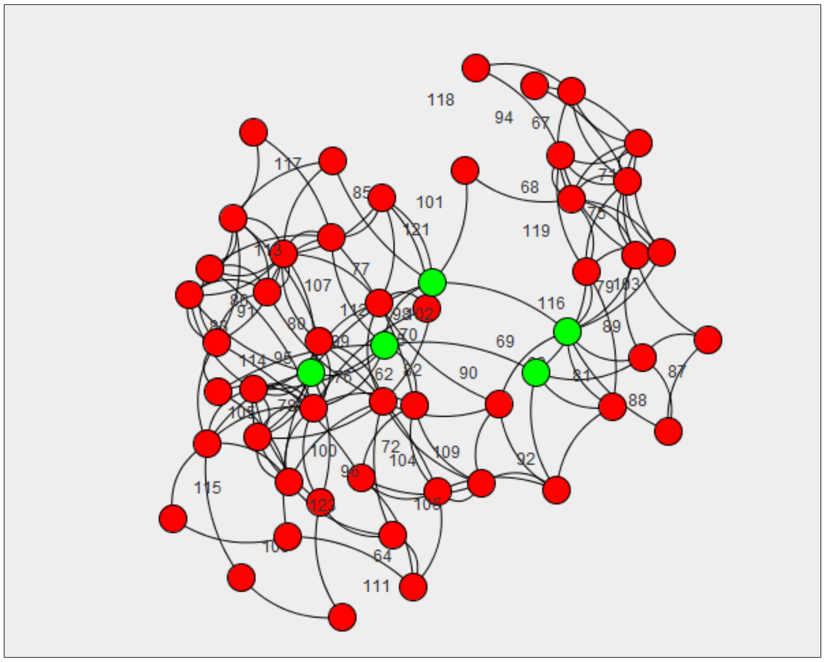


Figure 5: Seeds in a network (nodes highlighted in Green Color)

1.4.1 Introduction to our metric-Betweenness Centrality

Betweenness Centrality is one among the various centrality indices of a node. Centrality indices are helpful in characterizing the important vertices of a network. Different centrality indices depend on different definitions of "importance". The importance can be either the type of flow across the network or the involvement the nodes have in accounting for the cohesiveness/ closed group structure of the network.

Betweenness Centrality is a metric that is characterized by the walk structure in a network. It is further considered under the class "medial centralities" that count the walks which pass through the given vertex.

Betweenness is a centrality measure of a vertex within a graph . Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Accordingly, vertices that have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have a high Betweenness. Fig 6 represents the nodes with low and high betweenness values in a network.

The betweenness centrality of a vertex v in a graph G = (V, E) with V vertices and E edges is computed as follows:

- For each pair of vertices (s, t), compute the shortest paths between them.
- For each pair of vertices (s, t), determine the fraction of shortest paths that pass through the vertex in consideration (here, v).
- Sum this fraction over all the pair of vertices (s, t).

The computation can be mathematically represented using the following summation.

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

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.

Normalization of the metric value is done by diving through the number of pairs of vertices not including v, which for undirected graphs is (n-1) (n-2)/2.

[12] ,[13] provide more elaborative discussion on the metric.

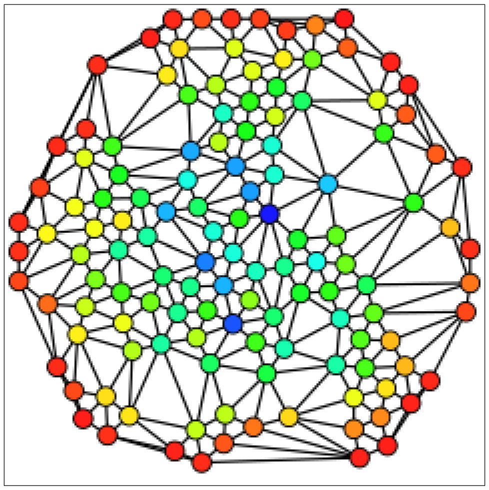


Figure 6: Nodes characterized by the Betweenness metric, Red= 0, Blue=max Betweenness.

Betweenness Centrality not only accounts for the centrality of a node, which is the most desirable quality of a seed node, but also accounts for the information flow, connectivity of a community. Thus, we chose this metric to develop our seeding algorithm.

The dual role of the metric is depicted in the following figures, Fig 7 and Fig 8. Fig 7 denotes that nodes with high betweenness centrality are central to the network. Fig 8 denotes that the nodes with high betweenness centrality are also prominent in the flow of information between two communities by acting as a bridge between them.

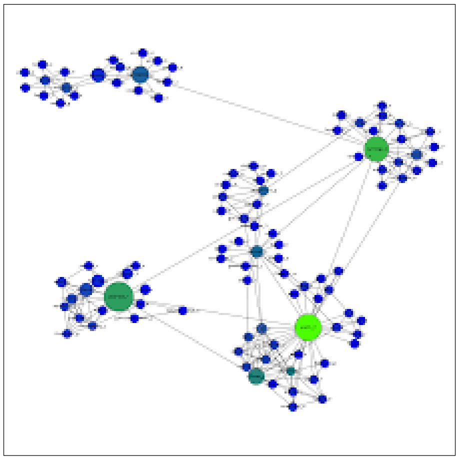
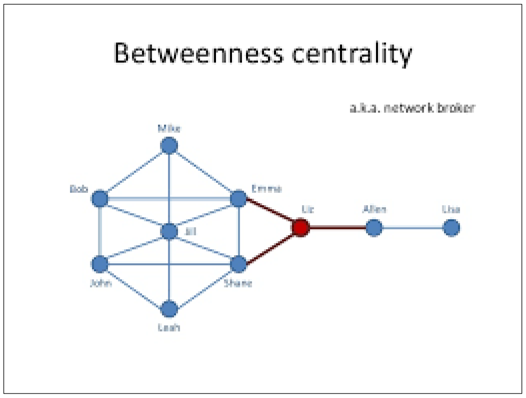


Figure 7: Betweenness Centrality accounting for centrality.



Betweenness centrality

a.k.a. network broker

Figure 8: Betweenness Centrality accounting for connectivity between two communities.

CHAPTER II

REVIEW OF LITERATURE

2.1 Community Structure and its Detection Algorithms - an Overview

Community structure is being studied ever since the networks are chosen to represent various types of systems. For example, an Organization, is best represented as a network of people ordered in a certain hierarchy and preference. Thus, networks, are a way to represent/model systems that contains interacting entities. [1] is the most studied literature when one attempts to understand the community structure in a network and devise an algorithmic technique to detect the communities.

[14] provides a good overview on the various Community Detection Algorithms (Disjoint). It provides a comparative analysis of the various algorithms and strikes a good trade-off between them. The analysis is based on the well-designed strict tests and variety of benchmark, synthetic networks with heterogeneity in the distributions of groups in the network.

Community Detection problem in networks is considered analogous to the Graph Clustering problem in a graph. This is because graphs are the best suitable data structures to represent the network. Hence, community detection is essentially a graph clustering problem and can be solved using various graph clustering techniques. [15] is a survey on the various graph clustering techniques and reviews them on performance criteria. The authors review both global and local approaches and provide a conclusion remarks that can be considered by a researcher to develop new technique to solve the problem of community detection.

Most of the algorithms and representations assume undirected edges and no edge weights to have incur no loss in generality. However, in reality, the edges are directed and may also have weight assigned to them. So, techniques that work for these types of networks are also needed and are developed in recent research works. [16] is a survey conducted on the Clustering and Community Detection techniques in directed networks. It is the first review of its kind, dedicated to directed networks and networks with weighted edges. The authors provide an in-depth review of the existing methods to detect communities in directed networks. Towards the end, they also provide metrics and techniques to evaluate the obtained results and interesting future research directions.

2.2 Overlapping Community Detection Algorithms- an Overview

The current research pertaining to the Community Detection problem has drifted apart from the disjoint Community Detection to Overlapping Community Detection. Due to the increased possibility of the phenomena of overlap, and with the advent of massive online networks, the overlapping community detection has become a widely studied problem. Several new algorithms and techniques have been developed to accommodate the overlap phenomena in the process of assigning nodes to communities. [17] reviews the existing and widely accepted algorithms to detect overlapping communities. It also reviews the quality measures and the benchmarks. The authors have thoroughly evaluated the node-level and community-level performance metrics and strike a trade-off among them.

A large amount of work is devoted in devising the techniques to detect overlapping communities in a network. Existing algorithms for detecting the overlapping communities can be classified into many classes.

A brief classification is:

Class 1: Algorithms that are specifically designed to be able to scale to the massive datasets. [9] [18],[19],belong to this class. For example, in [19], the authors develop a highly scalable and efficient solution for a stochastic gradient Markov chain Monte-Carlo algorithm that detects overlapping communities in graphs. Specifically, they discussed how the algorithm was structured to facilitate its parallelization.

So the algorithms in this class are computationally efficient, and distributed algorithms. The underlying approaches in the algorithms of this class are statistical modelling, parallel programming, high-performance computing.

Class 2: Algorithms that are based on Topological properties of graphs, such as, cliques, n-cliques, connected components, node attributes, edge weights etc. These algorithms semantically correlate the properties of graphs to the functional role, the property can play, with the network in consideration. [20], [21] belong to this class.

For example, in [20], they propose an overlapping community detection algorithm based on node convergence degree which combines the network topology with the node attributes. PageRank algorithm is used to get the importance of each node in the global network. Finally, the overlap communities can be identified by the Spectral Cluster based on node convergence degree.

This class of algorithms have a limitation of being able to scale to massive datasets, though they can scale to huge datasets.

Class 3: Another major class of algorithms for overlapping community detection, is the seed-centric approaches. Seed-centric approaches are the typical local-objective based community detection approaches which in their later stages, expand their communities and thereby become equivalent to global-objective based clustering/partitioning of the graph/network. [22]-[25] belong to this class. This is our class of interest since our newly developed approach belongs to this class.

They can be further classified into approaches that opt for different ways of selecting the seeds (random, informed), the type of seeds (single, set), expanding the seed to determine their communities etc. Algorithm in [22] is a seeding algorithm which is similar to a voting with several rounds. In each round, the seeding algorithm selects the node with the most votes as seed and locally updates the supports of nodes to their neighbors.

The design of the process flow of our approach is heavily based on [26], the way the graph is filtered in the initial stage and then the filtered components are added back in determining the final overlapping communities. The Algorithm uses seeding techniques that are based on metrics like centrality, node degree. It has two variants of seeds obtained using two different techniques. The expansion method is based on a metric called conductance.

The Algorithm contributes majorly through the expansion technique which is based on “Neighborhood Inflation” phenomena. It plays a crucial role in the success of the algorithm. It is the latest algorithm that considerable outperforms the existing state-of-the-art overlapping community detection methods in terms of scalability, quality measures and identifying overlapping communities that are very close to the ground-truth of the networks.

2.3 Review of various Graph/Network metrics

An exclusive study of the graph and networks metrics is essential to determine the metric of importance in the context of a community. [11] is a survey of the measurements that characterize the networks. The survey is the first one of its kind. It presents and discusses a comprehensive set of network measurements. In addition to reviewing the various measurements it also addresses the other important issues related to visualization of the networks to represent these measurement characteristics.

[27] orders and classifies the wide range of graph metrics which characterize a graph. The article provides a brief, a quick review on the various metrics and classifies them into topological and

services metrics and also provides the correlation among the topological metrics. This is extremely useful in making a choice in the process of determining the metric of importance.

2.4 Review of various Graph/Network Analysis platforms

There are various Graph and Network analysis platforms which have different capabilities and specialties. The listing in [28] is worth-referring to, in the process of making a choice to determine the tool to be used to implement our Algorithm. Apart from these interactive tools, there are also language-based support APIs that help to develop a customized and new algorithm using the built-in and basic functions provided by the API. [29] is one such software library provided by the JUNG Framework Development Team.

2.5 About JUNG

JUNG - the Java Universal Network/Graph Framework, is a software library that provides a common and extendible language for the modeling, analysis, and visualization of data that can be represented as a graph or network. It is written in Java, which allows JUNG-based applications to make use of the extensive built-in capabilities of the Java API, as well as those of other existing third-party Java libraries.

It provides Visualization support through its Algorithms which helps in interactive exploration of the graph data. There is a great scope for customization of the algorithms and implementing a new method using the existing library. Hence we chose JUNG as our supporting library and JAVA to be the language platform of implementation.

Having reviewed the rich state-of-the-art literature in the research area and determining the technical tools/platforms for implementation, we go about implementing our proposed algorithm.

CHAPTER III

METHODOLOGY

3.1 Preliminaries

We present here, the basic definitions that are used throughout the further description of the methodology. The terms Graph, Network and Cluster, Community are used interchangeably throughout the discussion. Given a network, G = {E, V}, V is the set of n vertices and E is the set m edges of the network. Without loss of generality, we assume that our networks are undirected, and no edge weights are considered.

3.2 Problem Formulation

The aim of the traditional graph-clustering problem is to divide the network into k disjoint Clusters, C_1, C_2 …C_k such that C_1 U C_2 ….U C_k = V. whereas, the overlapping community detection problem is to find C_1, C_2 …C_k such that C_1 U C_2 ….U $C_k \leq$ V. This means that the overlapping communities do not necessarily include all of the vertices of the graph.

3.3 Our Algorithm-Process Flow

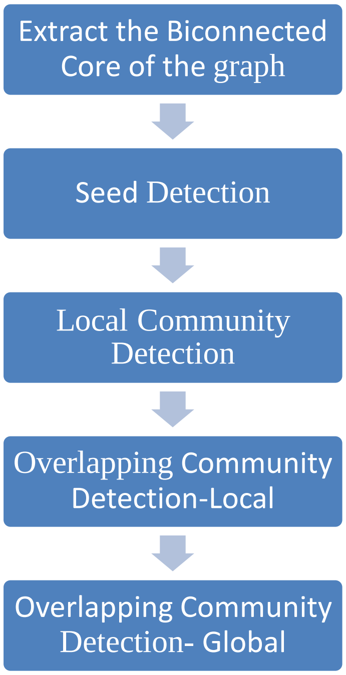

Extract the Biconnected Core of the graph
Seed Detection
Local Community Detection
Overlapping Community Detection-Local
Overlapping Community Detection- Global

Figure 9: Process Flow of our Algorithm

Our Algorithm - Various Stages

- The algorithm we propose includes several stages.
- Stage 1: Extracting the Biconnected Core from the Original Graph.
- Stage 2: Seed Detection
- Stage 3: Local Community Detection
- Stage 4: Local Community Expansion to obtain overlapping communities
- Stage 5: Adding back the connected components, single nodes removed in Stage 1

The convergence of various stages is defined by different convergence criteria based on the motive behind that stage.

3.3.1 Stage 1- Bi-connected-core extraction

Algorithm 1: Extracting the Biconnected Core

- Biconnected Core G_C = (V_C, E_C) is the maximum size connected sub-graph of G”= (V, E\Es).
- Compute the biconnected core of the given input graph, by removing the sub graphs connected to the biconnected core of the graph.

In this stage of the algorithm, we essentially filter all the components that do not contribute to the Community structure of the network. These nodes can be considered as outliers and are not necessary. These are added back in the last stage of the Algorithm, stage-5 so as to recover the original graph structure.

A Biconnected core of a graph is the maximum size connected sub-graph of the maximal induced Sub graph. A Biconnected core contains substantial fraction of the edges and is dense enough to consider the presence of Communities, which we are looking for.

3.3.2 Stage 2- Seed Detection

Algorithm 2: Seed Detection

```
➢ Sort the vertices of the graph in decreasing order of their degree, set D.
➢ Input: Graph G= (V, E), the number of seeds k, determined by the known-information of
  the graph or using the Modularity Report.
➢ Output: the seed set S
Initialize S= null set.
    while |S| < k do
        for each x of D do
            Compute Betweenness Centrality (BC) for x with all its unmarked neighbors
                if BC(x) > threshold then
                    S= {x} U S
                    Mark x and its neighbors as visited.
                end if
        end for
    end while
```

In this stage of Algorithm, we determine the seed nodes of the community. It is the most crucial step in the Algorithm, since detection the final communities is heavily depended on the quality of the seed nodes detected in this stage of the Algorithm.

The number of seeds to be chosen, represented by the parameter, k, is determined in two ways.

1. Using the Ground-truth of the network.
2. Using the Modularity Report generated by the tool [30]. It uses a high-level clustering Algorithm [31], to find the number of modules/ communities in the network. We use this as our input to roughly estimate the number of communities in the network.

We later modify the value of k, if the communities do not seem to be containing quality overlap.

3.3.3 Stage 3- Local Community Detection

Algorithm 3: Local Community Detection

- This is the phase of local community detection based on the seed nodes detected in the previous stage.
- Input: Seed Set S, Graph G= (V, E).
- Output: Local Communities set LC

```
for each s of S
    Set NS = Compute_Neighbors(s)
    Set LC= {s} U {NS}
end for
```

In this stage of the Algorithm, we compute the local communities based on the seed nodes obtained in the previous step of the Algorithm.

3.3.4 Stage 4- Overlapping Community Detection-Local

Algorithm 4: Overlapping Community Detection - Local

- This is the phase of overlapping/ global community detection based on the local communities detected in the previous stage.
- Input: Local Communities set LC, Graph G= (V,E).
- Output: Set of sparsely Overlapping Communities SOC, Set of Densely Overlapping Communities DOC.

```
Identify vertices not a part of LC set, name it LC'.
for each lc in LC, compute the leaf nodes set LN(lc).
end for
for each pair (i,j) of leaf node set
    Compute betweenness for the pair of vertices, through every node of lc' of LC'.
    if average betweenness > threshold, then dense overlap set, DOC.
        else add to sparse overlap set, SOC.
    end if
end for
```

In this stage of the Algorithm, we compute the overlapping communities by expanding the neighbors obtained in the previous step of the Algorithm. The Algorithm computes the overlap of communities using the threshold set for the betweenness metric of the algorithm. This results in classifying the overlapping communities into Dense and Sparse overlapping communities.

The densely overlapping communities denoted by the DOC set, is of major interest since lot of insights could be drawn using the community structure detected.

The sparsely overlapping communities are also useful but comparatively provide less information about the network activities and properties.

The algorithm converges by detecting the local Overlapping Communities of the network.

3.3.5 Stage 5- Overlapping Community Detection-Global

Algorithm 5: Overlapping Community Detection - Global

- This restores the input graph back and converges the process the identifying the overlapping communities.
- Input: Graph G= (V, E), Core with communities identified, from the previous stage.
- Output: Restored Graph with further additions to the communities identified.

```
for each C of DOC
    Identify the biconnected components of C, B.
C = C U B
end for
```

In this last stage of the Algorithm, we add the filtered components of stage 1, back to graph so that the original graph structure is restored and our algorithm converges by detecting the global Overlapping Communities of the network.

3.4 Datasets

We have considered various real-time networks to perform our experiments. The benchmark datasets are used to test every new algorithm to verify its correctness.

Dataset Name	Number of Nodes	Number of Edges	Category of Dataset
Zachary's Karate club	34	78	Social
Dolphin Social Network	62	159	Social
U.S. College Football	115	616	Social
Facebook	4,039	88,234	Social

Table 1: Statistics of the Datasets.

Zachary's Karate Club Dataset [32]

This network is drawn from the well-known "karate club" study of Zachary. In this study, relations between 34 members of a karate club over a period of two years are observed. During the study, a disagreement developed between the administrator and the teacher of the club, which eventually made the club split into two smaller ones, centering on the administrator and the teacher, represented by node 34 and node 1. Zachary was able to construct a network of friendships, using a variety of measures to estimate the strength of ties between members.

US College Football Network [33]

This dataset is the schedule for 787 games of the 2006 National Collegiate Athletic Association (NCAA) Football Bowl Subdivision. In the NCAA network, there are 115 universities divided into 11 conferences. Additionally, there are 4 independent schools, namely Navy, Army, Notre Dame and Temple, as well as 61 schools from lower divisions. Each school in a conference plays more often with schools in the same conference than schools outside. Independent schools do not belong to any conference and play with teams in all conferences, while lower division teams play very few games.

Dolphin Social Network [34]

The social network of a community of 62 dolphins in Doubtful Sound, New Zealand was as described by Lusseau et al (2003). The community was stably structured with close and long-lasting associations among members.

Facebook Dataset [35]

This dataset consists of 'circles' (or 'friends’ lists') from Facebook. Facebook data was collected from survey participants using this Facebook app.

CHAPTER IV

PROCESSING AND FINDINGS

The chapter discusses the stage-wise findings as a result of executing the Algorithm on the Datasets.

4.1: Stage-wise processing of the Network

Initially the Network datasets are read by the Algorithm and visualized as follows.

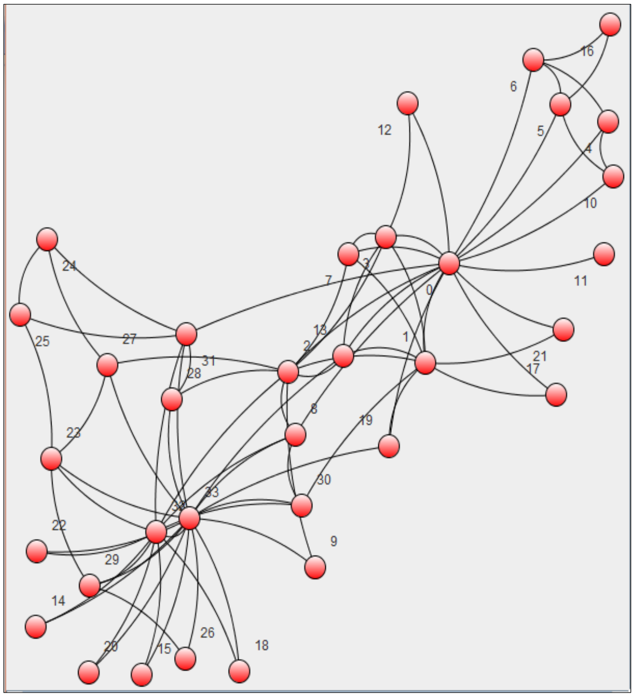


Figure 10: Zachary’s Karate Club Network

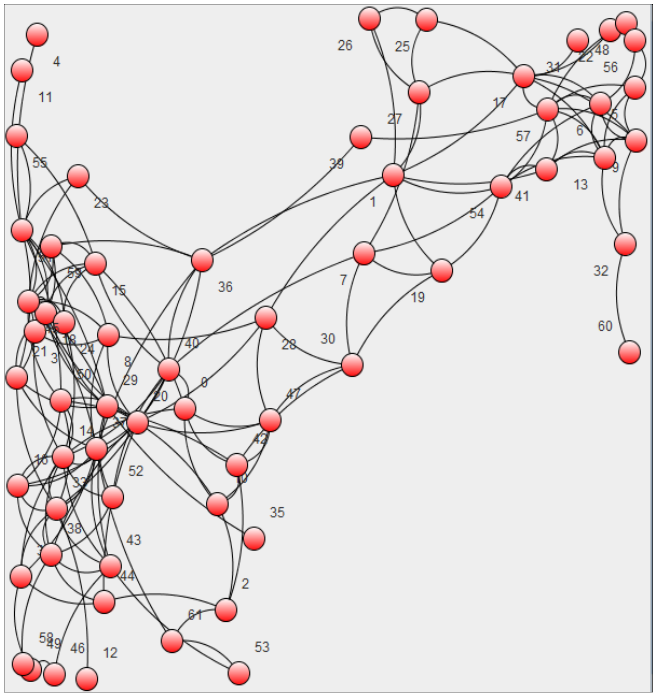


Figure 11: Dolphin Social Network

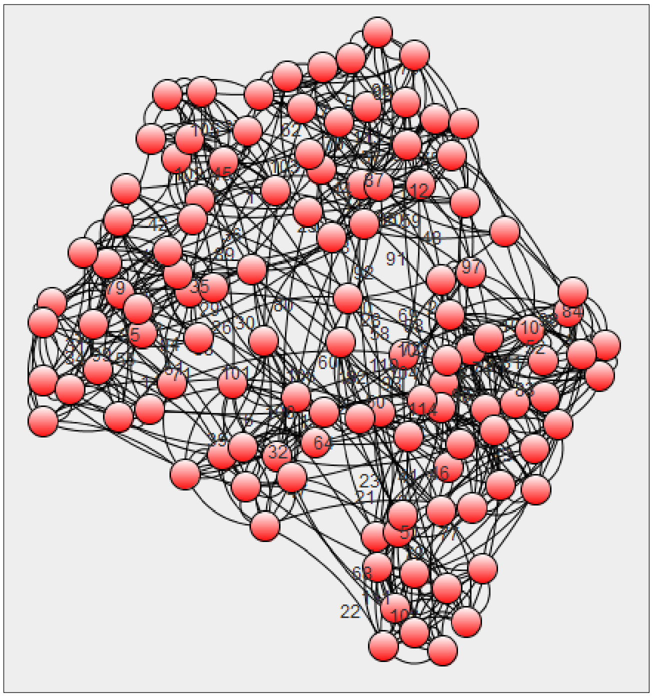


Figure 12: US College Football Network

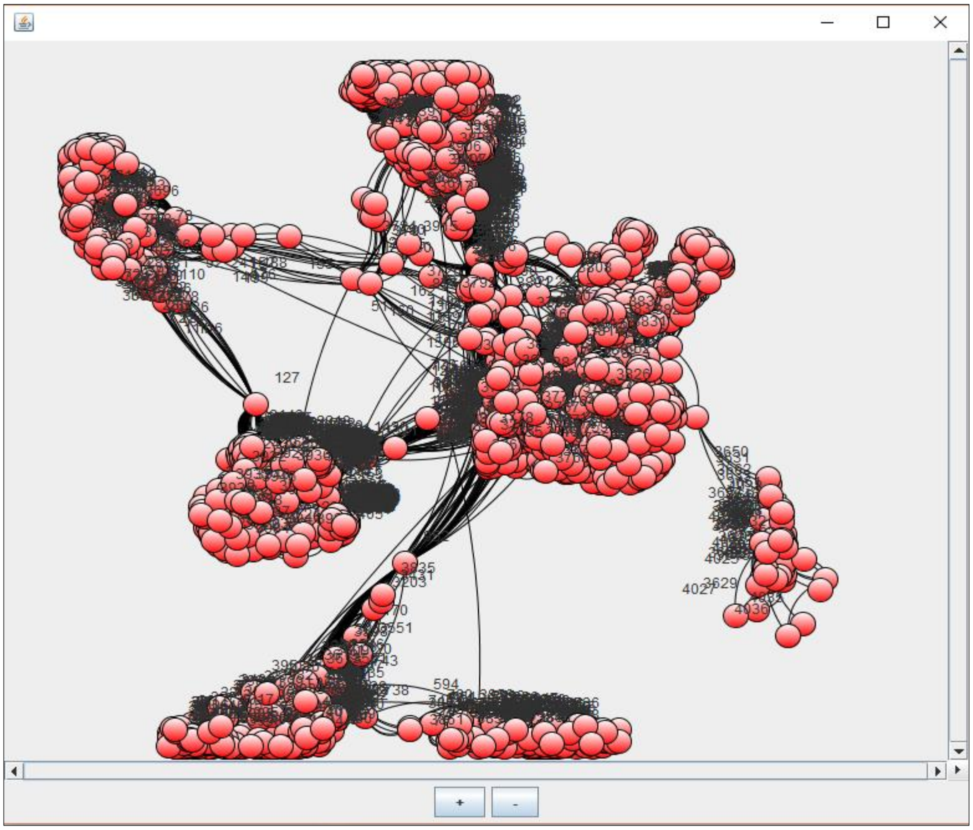


Figure 13: Facebook Dataset representing social groups.

Network Representation after Stage 1- Dolphin Social Network

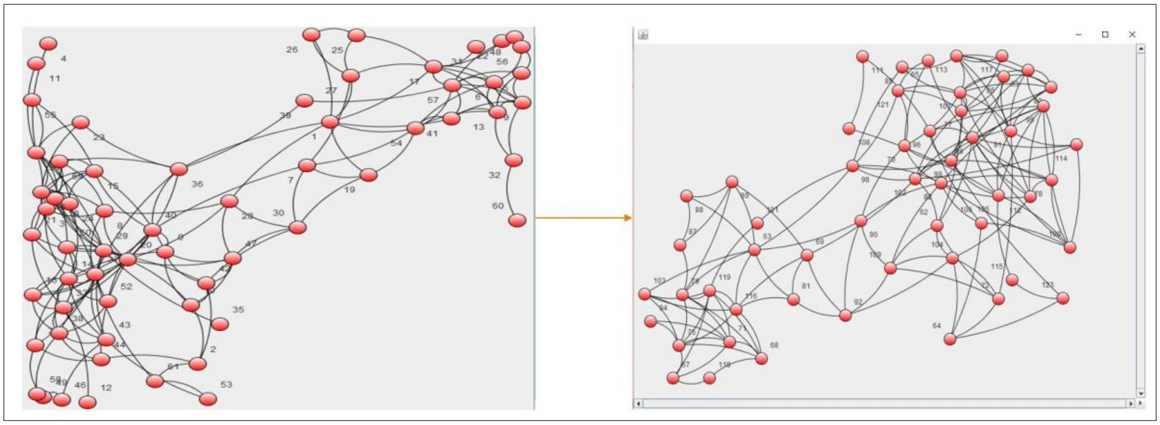


Figure 14 : Processing the input network in Stage 1

In transformation of the network shown in Fig 14 represents the filtering process we perform in the Stage-1. The Biconnected sub graph components are removed from the original graph and the Biconnected core of the graph is obtained as shown.

Network Representation after Stage 2- Dolphin Social Network:

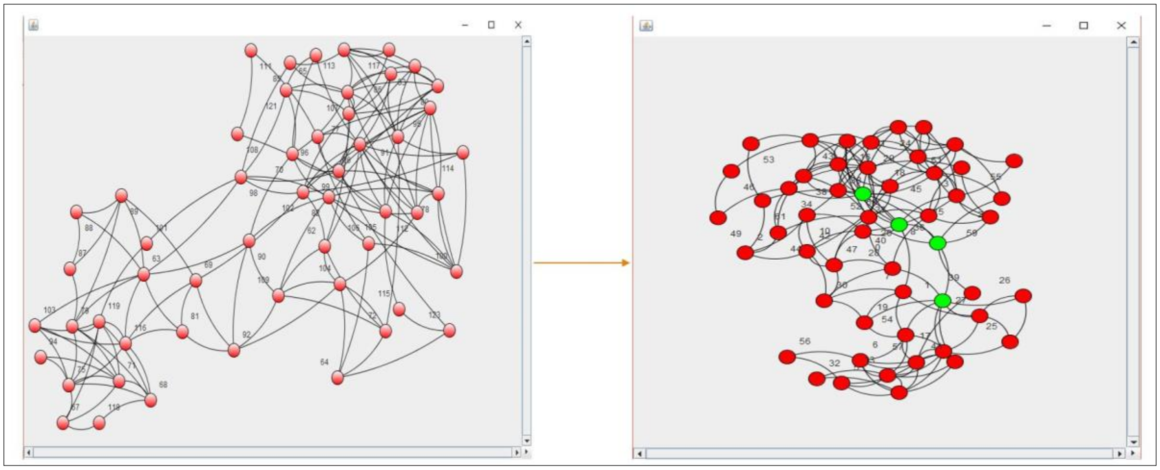


Figure 15 : Determining the seed nodes in Stage 2

The nodes highlighted in Green color represent the seed nodes of the network. The number of seed nodes is determined using the Modularity report generated using [30]. The following Figure, Fig 16 shows the modularity report, based on which the number of communities in the graph are determined to be 4 for the Dolphin Social Network.

After deciding on the number of seed nodes to be determined in the network, we chose the seeds based on our metric, Betweenness Centrality score of a node. Fig 17 represents the Betweenness Centrality score distribution of the nodes of Dolphin Social Network.

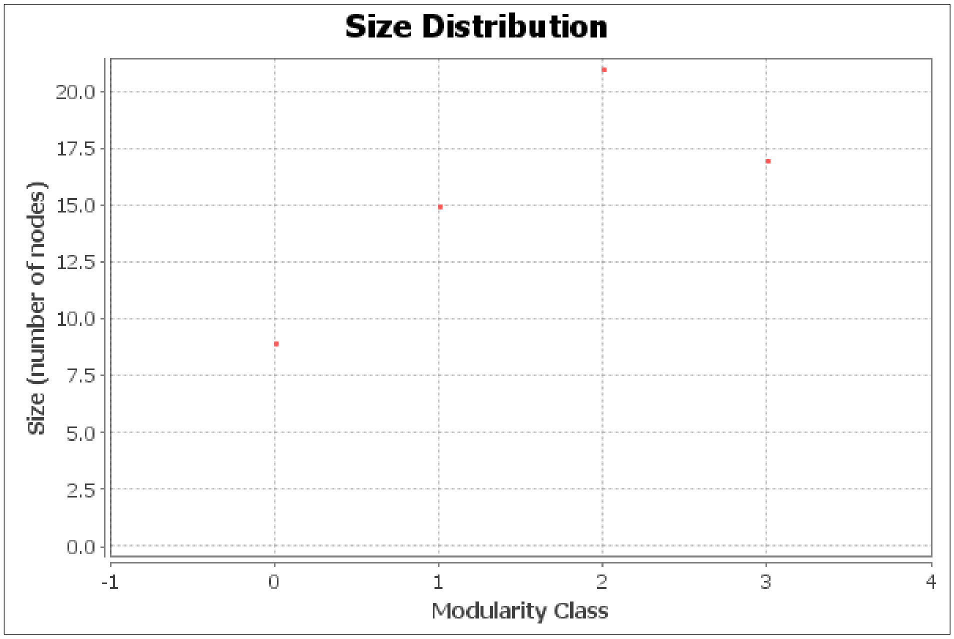


Figure 16: Modularity Report of Dolphin Social Network.

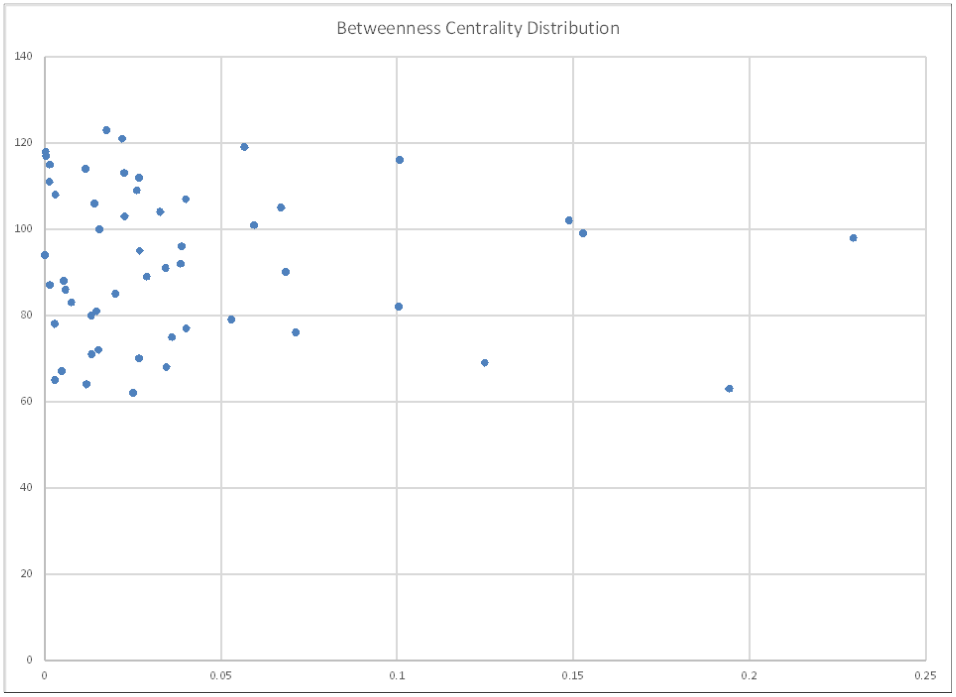


Figure 17: Graph representing the Betweenness Centrality Distribution.

The top 4 values according to the distribution are chosen to be the seed nodes of the network. The seed nodes has high Betweenness Centrality scores.

Network Representation after Stage 3- Dolphin Social Network

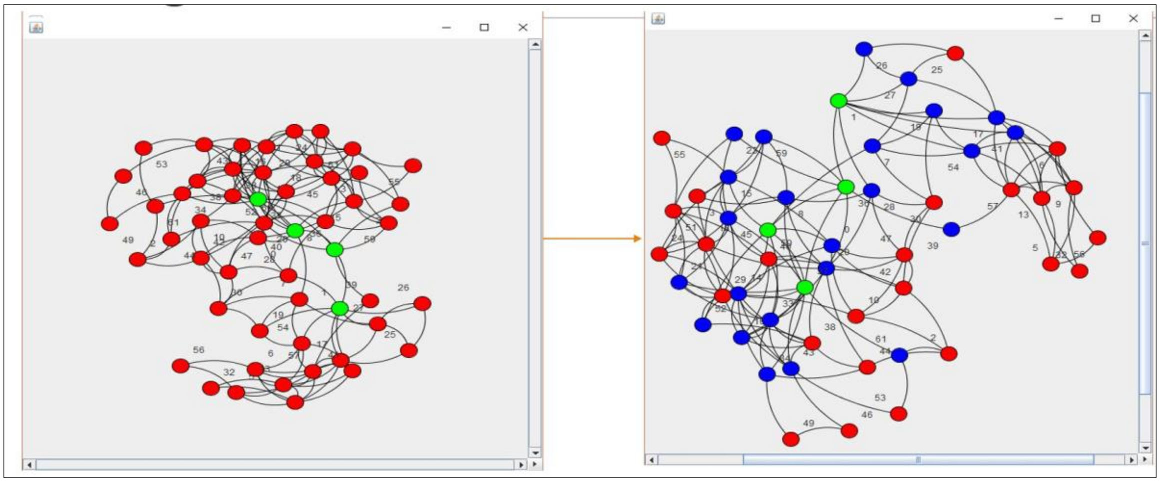


Figure 18: Local Community Detection.

Fig 18 represents the local communities of the network (nodes highlighted in blue color), which are determined using the seed nodes (highlighted in green color). The immediate neighbors of the seed nodes form the local communities of the network. These local communities when expanded, bring out the overlap feature of the network.

Network Representation after Stage 4- Dolphin Social Network

In this stage, we obtain the pairs of overlapping communities (Dense & Sparse) by further expanding the local communities determined in the previous stage. Figs 19, 20 & 21 represent the Dense Overlapping Communities, pair-wise. Fig 22 represents the pair of sparsely overlapping communities. The overlapping communities are represented in Blue and Green colors. The nodes in Pink color are the nodes that overlap between the Blue and Green communities. The nodes highlighted in Black are not only the overlapping nodes between the two communities, but also are the seed nodes of the network.

This infers the following about the importance of our metric:

- The seeds nodes are not necessarily central to the communities.
- They are a part of the significant overlap feature of the communities and are equally important in acting as bridges between the two communities.
- Our metric plays a role in determining the significant overlaps but distinguishing between the communities and classifying them as dense or sparse based on the threshold limit set.
- Hence, the choice of our metric, Betweenness Centrality is justified.

The sparsely overlapping communities are of less importance to the analysis, since the significant feature are aiming for in this research is the overlap of the communities.

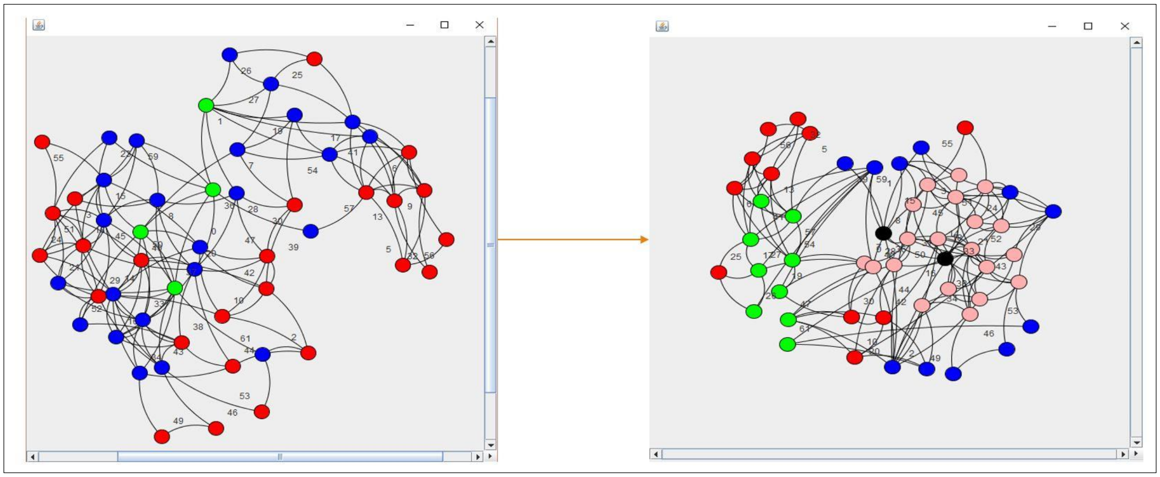


Figure 19: Overlapping Community 1- Local

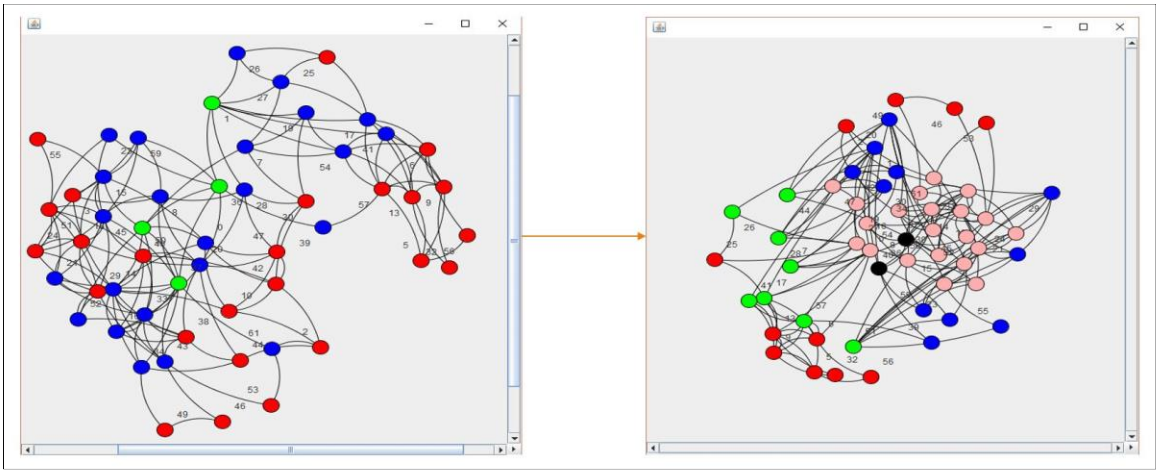


Figure 20: Overlapping Community 2- Local

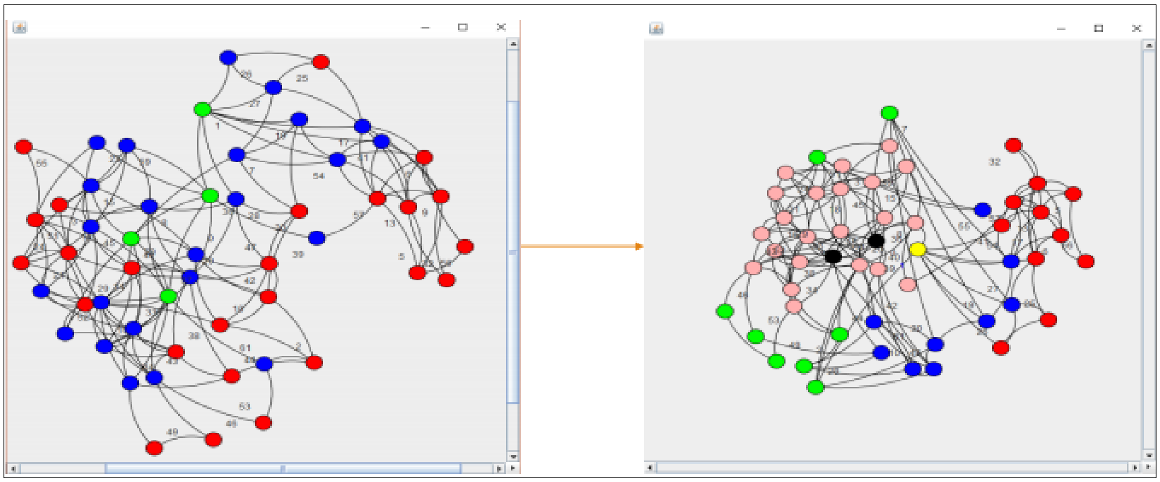


Figure 21: Overlapping Community 3- Local

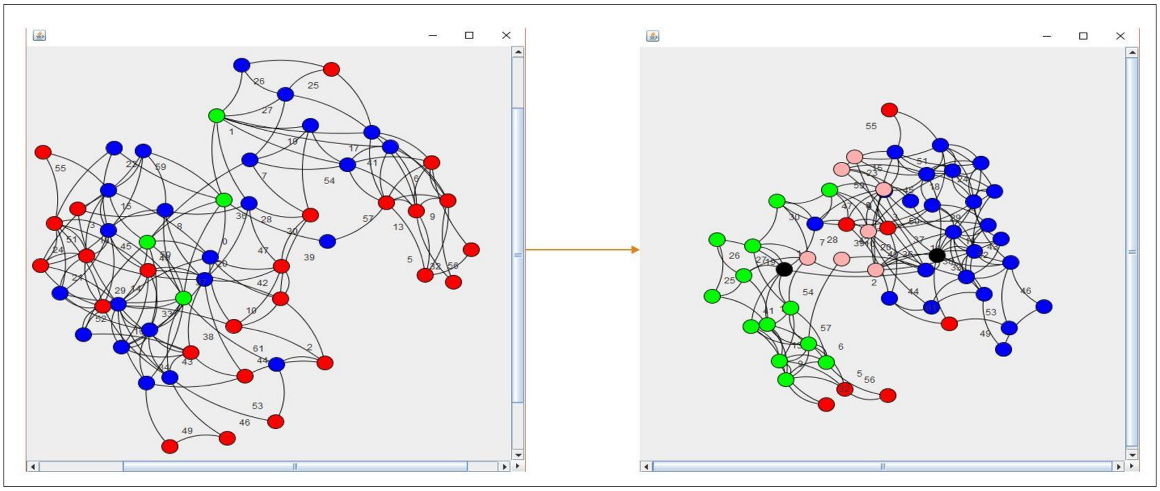


Figure 22: Overlapping Community 4- Local

Thus, we obtain the set of Overlapping Communities (Dense & Sparse) of the network at the convergence of Stage 4.

Network Representation after Stage 5- Dolphin Social Network

Fig 23 represents the transition of the filtered, classified network attaining its original structure, by restoring the filtered components to the network. The filtered Biconnected components are added back to the Biconnected core so that the original network is restored, in terms of total number of nodes and edges. This step does not alter or disturb the overlapping communities detected, or the classification of different nodes. This post-processing step just resets the counts of number of nodes and edges to their original values.

The algorithm converges with this stage and we successfully determine the overlapping communities of the Dolphin Social Network. We also experiment on Facebook Dataset and summarize the statistics of the computation.

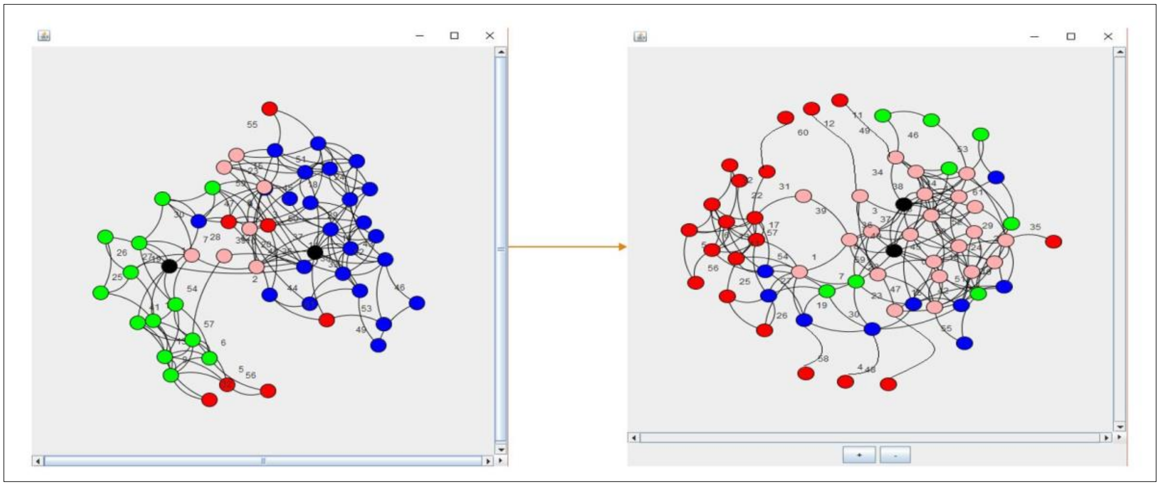


Figure 23: Overlapping Community - Global

4.2 Analysis of the resultant Overlapping Communities

We infer that the Dolphin social network is qualitatively divided into ‘k-1’ number of communities which are dense. The sparse communities are also of good quality, which could be considered based on our application. We infer that our seed nodes play a crucial role in both the local community and the expanded (overlapping) community. The seed quality is high both structurally and functionally. Almost all of the vertices are assigned to any of the local communities. Very few of the nodes remain unassigned, leading to a good graph coverage percentage. Similar analysis applies to Facebook Communities as well. Applications of the overlapping community detection in Dolphin social network could be studying the associations between different groups of dolphins which are characterized by their doubtful sounds. Likewise, in Facebook, the overlapping communities signify the multiple activities of the users and help in determining the most important people in the community who are responsible for the information diffusion within the network.

4.3 Comparative Analysis against different metric choice

We compare our metric Betweenness Centrality with another metric choice, the Distance Centrality, which is very close in nature to our metric, yet is different. The Distance Centrality is a structurally significant metric .It is equal to distance of a node i to each other vertex in the graph.The closer the node, the more centered it is, in the community. The results obtained show that the distance centrality network does not categorize the nodes to have more individual features. Our metric is more suitable than this centrality metric when modularity of the detected communities is considered. The extent of overlap observed greatly affects the modularity feature of the communities as seen in Fig 25. Fig 24 represents the stages 2 & 3 of the Algorithm. Fig 25 represents network structure after stage 4 of the Algorithm, i.e., after detecting the overlapping communities.

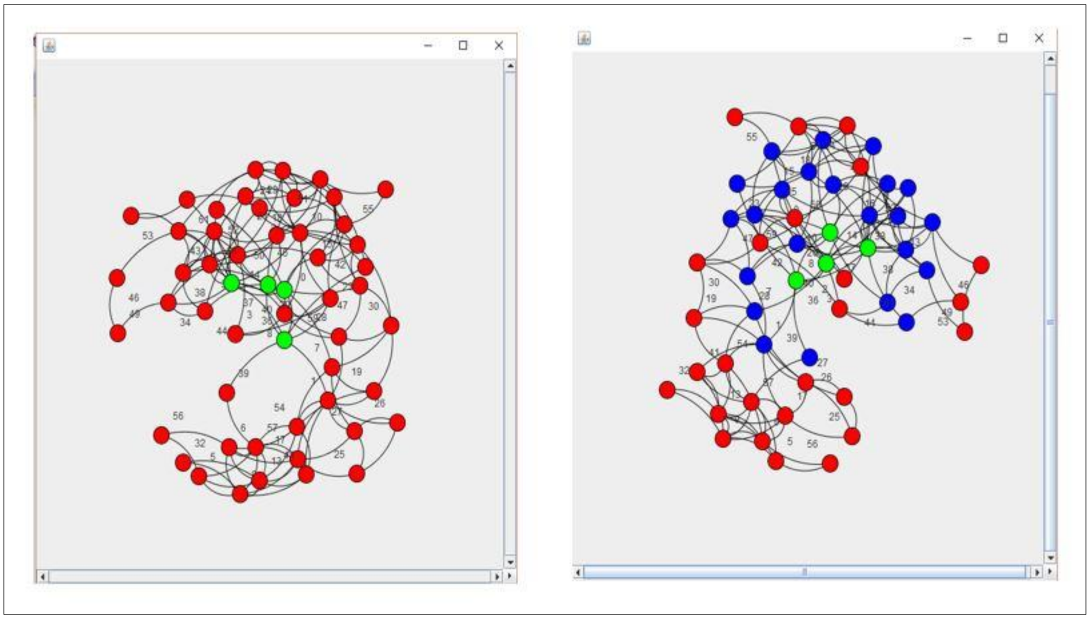


Figure 24: Stages 1 & 2, using Distance Centrality metric

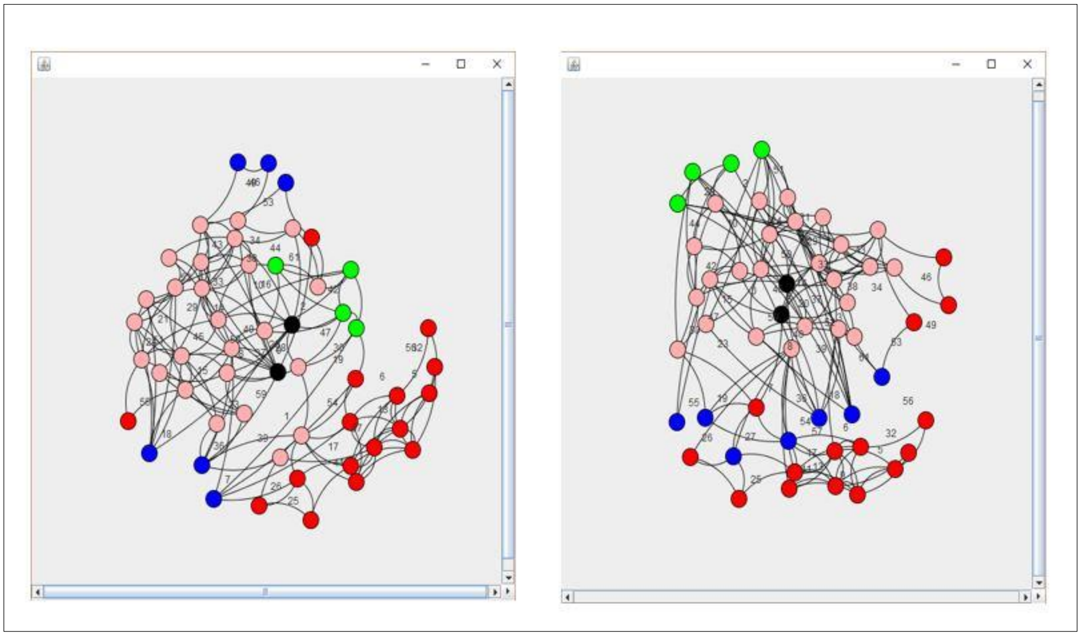


Figure 25: Overlapping Communities, using Distance Centrality metric

4.4 Evaluation of the Algorithm- comparison with the state-of-the-art

We have compared our algorithm with other state-of-the-art algorithms, CFinder [36] and OSLOM [37] and found that our algorithm produces nearly comparable results.

Algorithm	Network	Coverage %	No of Communities
CFinder	Dolphin	85.83	4
OSLOM	Dolphin	100	4
Our Algorithm	Dolphin	84.37	3(dense) +1(sparse)

Table 2: Comparison of Statistics for Dolphin Dataset.

Algorithm	Network	Coverage %	No of Communities
cFinder	Facebook	90	11
OSLOM	Facebook	82.5	14
Our Algorithm	Facebook	85.04	5(dense) +2(sparse)

Table 3: Comparison of Statistics for Facebook Dataset.

Thus, we state that our algorithm performs fairly well in comparison with the existing state-of-the-art algorithms of same class. Our metric choice is completely justified in all stages of the algorithm and is proven to be successful in determining the dense overlapping communities of a given network.

The coverage percentages also signify that our algorithm spans the entire network and is able to provide a good classification of the network. Additionally, we provide two categories of Overlapping Communities, Densely overlapping communities and sparsely overlapping communities instead of providing a generic classification. Thus, our algorithm is quite on par with the existing and accepted state-of-the-art algorithms.

CHAPTER V

CONCLUSION

In this thesis we proposed a new seed-centric algorithm to detect overlapping communities of a network. We discussed on the importance of the problem. i.e., overlapping community detection, and its applications in various practical settings. We thoroughly analyzed the various graph/network metrics that are significant in the context of a community structure in the network.

Our choice of metric is very well justified by the results we obtained. Our multi-stage algorithm, is computationally effective and represents full-cycle of the overlapping community detection process. Every stage has its own significance and contributes to the main aim of the thesis.

Our contributions are two-fold. First being, the ability to identify most prominent nodes on the networks as seeds. Second being, obtaining dense and sparse overlapping communities without the loss of modularity of the communities. Our algorithm is also effective in scaling to large, real-world networks. It is proven effective in preserving the important features of the community such as modularity, cohesiveness etc. of a community.

The performance of our algorithm is compared with two of the state-of-the-art techniques and our algorithm performs fairly well. It outperforms other metric, Distance Centrality, in determining quality overlapping communities. Hence, we conclude that our algorithm detects quality overlapping communities that have structural as well as functional significance in the network.

REFERENCES

[1] Girvan, M., and M. E. J. Newman. “Community Structure in Social and Biological Networks.” Proceedings of the National Academy of Sciences of the United States of America 99.12 (2002): 7821–7826. PMC. Web. 5 Dec. 2016.

[2] Fortunato, Santo. "Community detection in graphs." Physics reports 486.3 (2010): 75-174.

[3] Mahmood, Arif, and Michael Small. "Subspace Based Network Community Detection Using Sparse Linear Coding." IEEE Transactions on Knowledge and Data Engineering 28.3 (2016): 801-812.

[4] Khorasgani, Reihaneh Rabbany, Jiyang Chen, and Osmar R. Zaïane. "Top leaders community detection approach in information networks." 4th SNA-KDD Workshop on Social Network Mining and Analysis, Washington DC. 2010.

[5] Duch, Jordi, and Alex Arenas. "Community detection in complex networks using extremal optimization." Physical review E 72.2 (2005): 027104.

[6] Gregory, Steve. "Fuzzy overlapping communities in networks." Journal of Statistical Mechanics: Theory and Experiment 2011.02 (2011): P02017.

[7] J. Yang and J. Leskovec, “Overlapping community detection at scale: A nonnegative matrix factorization approach,” in Proc. 6th ACM Int. Conf. Web Search Data Mining, 2013, pp. 587–596.

[8] Kelley, Stephen, et al. "Overlapping communities in social networks." International Journal of Social Computing and Cyber-Physical Systems 1.2 (2011): 135-159.

[9] Gopalan, Prem K., and David M. Blei. "Efficient discovery of overlapping communities in massive networks." Proceedings of the National Academy of Sciences 110.36 (2013): 14534-14539.

[10] Yang, Jaewon, and Jure Leskovec. "Overlapping communities explain core–periphery organization of networks." Proceedings of the IEEE 102.12 (2014): 1892-1902.

[11] Costa, L. da F., et al. "Characterization of complex networks: A survey of measurements." Advances in physics 56.1 (2007): 167-242.

[12] Freeman, Linton (1977). "A set of measures of centrality based upon betweenness". Sociometry. 40: 35–41. doi:10.2307/3033543.

[13] Brandes, Ulrik (2001). "A faster algorithm for betweenness centrality" (PDF). Journal of Mathematical Sociology. 25: 163–177. doi:10.1080/0022250x.2001.9990249. Retrieved October 11, 2011.

[14] Lancichinetti, Andrea, and Santo Fortunato. "Community detection algorithms: a comparative analysis." Physical review E 80.5 (2009): 056117.

[15] Schaeffer, Satu Elisa. "Graph clustering." Computer science review 1.1 (2007): 27-64.

[16] Malliaros, Fragkiskos D., and Michalis Vazirgiannis. "Clustering and community detection in directed networks: A survey." Physics Reports 533.4 (2013): 95-142.

[17] Xie, Jierui, Stephen Kelley, and Boleslaw K. Szymanski. "Overlapping community detection in networks: The state-of-the-art and comparative study." ACM Computing Surveys (csur) 45.4 (2013): 43.

[18] El-Helw, Ismail, Rutger Hofman, and Henri E. Bal. "Towards Fast Overlapping Community Detection." Cluster, Cloud and Grid Computing (CCGrid), 2016 16th IEEE/ACM International Symposium on. IEEE, 2016.

[19] El-Helw, Ismail, et al. "Scalable Overlapping Community Detection."

[20] Wen, Xuyun, et al. "A Maximal Clique Based Multiobjective Evolutionary Algorithm for Overlapping Community Detection." IEEE Transactions on Evolutionary Computation (2016).

[21] Li, Weimin, et al. "Overlap Community Detection Based on Node Convergence Degree." Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive

Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2016 IEEE 14th Intl C. IEEE, 2016.

[22] Hu, Yanmei, et al. "Voting based seeding algorithm for overlapping community detection." Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2015 International Conference on. IEEE, 2015.

[23] Moradi, Farnaz, Tomas Olovsson, and Philippas Tsigas. "A local seed selection algorithm for overlapping community detection." Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. IEEE, 2014.

[24] Xu, Bingying, et al. "Local community detection using seeds expansion." Conference Anthology, IEEE. IEEE, 2013.

[25] Lee, Conrad, et al. "Seeding for pervasively overlapping communities." Physical Review E 83.6 (2011): 066107.

[26] Whang, Joyce Jiyoung, David F. Gleich, and Inderjit S. Dhillon. "Overlapping community detection using neighborhood-inflated seed expansion." IEEE Transactions on Knowledge and Data Engineering 28.5 (2016): 1272-1284.

[27] Hernández, Javier Martın, and Piet Van Mieghem. "Classification of graph metrics." Delft University of Technology, Tech. Rep (2011).

[28] http://www.kdnuggets.com/2015/06/top-30-social-network-analysis-visualization-tools.html

[29] http://jung.sourceforge.net/

[30] https://gephi.org/

[31] 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.

[32] W. W. Zachary, “An information flow model for conflict and fission in small groups, Journal of Anthropological Research” 33, 452-473 (1977).

[33] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002).

[34] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, “The Bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations, Behavioral Ecology and Sociobiology” 54, 396-405 (2003)

[35] J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.

[36] Adamcsek, Balázs, et al. "CFinder: locating cliques and overlapping modules in biological networks." Bioinformatics 22.8 (2006): 1021-1023.

[37] Lancichinetti, Andrea, et al. "Finding statistically significant communities in networks." PloS one 6.4 (2011): e18961.

VITA

DEEPTHI KASUVAJJALA

Candidate for the Degree of

Master of Science

Thesis: A SEED-CENTRIC ALGORITHM FOR OVERLAPPING COMMUNITY DETECTION IN NETWORKS

Major Field: Computer Science

Biographical:

Education:

Master of Science in Computer Science at Oklahoma State University, Stillwater, Oklahoma in December, 2016.

Bachelor of Technology in your Electronics & Computer Engineering at Jawaharlal Nehru Technological University, Hyderabad, Telangana, India in 2012.

Experience:

Graduate Research Assistant at Department of Management, Spears School of Business, OSU from June 2016- December 2016.