Comparative Analysis of Community Discovery Methods in Social Networks

Dr. M. Mohamed Sathik
Associate Professor
Dept. of Computer Science
Sathakathullah Appa College
Tirunelveli. INDIA.

Dr. K. Senthamarai Kannan
Professor & Head
Dept. of Statistics
Manonmaniam Sundaranar University, Tirunelveli. INDIA

A. Abdul Rasheed
Assistant Professor
Dept. of Computer Applications
Valliammai Engineering College, Chennai, INDIA.

ABSTRACT
The study of networks is an active area of research due to its capability of modeling many real world complex systems. Social Network gains its popularity due to its ease of use. It enables people all over the world to interact with each other with the advent of Web 2.0 in this Internet era. Online Social Networking facilitates people to have communication nevertheless of considering geographical location over the globe. Social Network Analysis is the field of research that provides a set of tools and theoretical approaches for holistic exploration of the communication and interaction patterns of social systems. A common pattern among the group of people in a network is considered as a community which is a partition of the entire network structure. There are few existing methods for discovering communities. We introduced a method called "mutual accessibility" for community discovery. This article compares such three different methods including the one that we introduced. The results of those methods are also shown by taking various datasets as an analysis.

General Terms
Data Mining, Pattern Matching

Keywords
Graph Clustering, Social Networks, Social Network Analysis, Community Discovery.

1. INTRODUCTION
Social networks gained popularity recently with the advent of sites such as MySpace, Friendster, Orkut, Twitter, Facebook, etc. 133 million blog records indexed by Technorati since 2002 and 900000 blog posts in 24 hours. By June 2008, Technorati tracked blogs in 81 languages and there are 77.7 million unique visitors in the US by August 2008. The number of users participating in these networks is large, e.g., a hundred million in these and growing.

Social network represented a graphical representation of people who are connected by relationships, groups connected by any relations, and organizations connected by relations. A social network N consists of a collection of nodes such as people, organizations, or groups A, B, C,… together with a collection of link sets L(A;B) which generalize the idea of a link from A to B. Network analysis is the study of social relations among a set of actors. Social Network Analysis (SNA) provides a spectrum of tools and theoretical approaches for holistic exploration of the interaction patterns among individuals, groups and even organizations. SNA is a field of research that provides a set of tools and theoretical approaches for holistic exploration of the communication and interaction patterns of social systems.

Social network analysis techniques have been applied to a variety of problems and they have been successful in uncovering relationships not seen with any other traditional method. A goal is to study the factors which influence relationships and to study the correlations between relationships. A fundamental problem related to these networks is the discovery of clusters or communities. One of the most important research and review questions in social networks is the "identification of communities".

Communities can be defined as collections of individuals who interact unusually frequently. The identification of communities often reveals the properties shared by the members such as occupations, social functions, or some other common hobbies like dating. These properties include related topics or common view points, which has led to a large amount of research on identifying communities in the web graph.

Community detection in complex networks has attracted a lot of attention in recent years. Detecting communities can be a way to identify substructures which could correspond to important functions. One of the most relevant features of graphs representing real systems is community structure. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. A community is a densely connected subset of nodes that is only sparsely linked to the remaining network. A community is a subset of nodes on the network. Suppose the network was represented as a graph G = (V, E) with V vertices to say people and E edges to say relationship that exists between two people. The graph partitioning problem consists on dividing G into k disjoint partitions. The goal is to minimize the number of cuts in the edges of the partition. Community discovery is generally considered as a clustering problem in which nodes in same community (Intra – Community) are more like to be connected than nodes in different communities (Inter – Communities). Communities can be discovered using graph partitioning. Communities of different kinds are also possible and in existence. For example, Communities in a citation network might represent related papers on a single topic and communities on the web might represent pages of related topics. Fig. 1 shows three groups of communities and their interaction level within community and outside community.

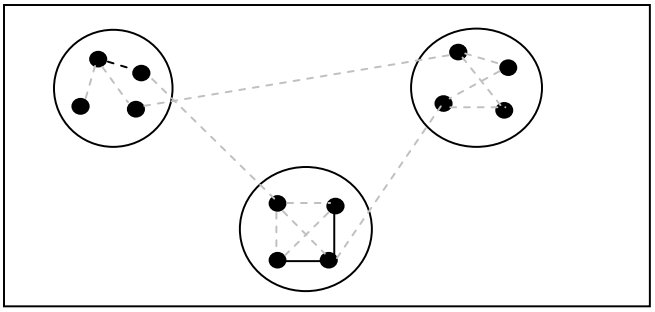

Fig 1: A group of three communities and the interaction among the members

Community discovery is basically a graph clustering problem, which decomposes the entire graph into sub-partitions. Graph clustering is the task of grouping the edge structure of the graph in such a way that there should be many edges within each cluster (intra-cluster) and relatively few between the clusters (inter-clusters). Due to its complexity of clustering the edges in to sub-partitions, graph clustering is considered as NP-hard problem. The graph partitioning problem consists on dividing G into k disjoint partitions. The goal is to minimize the number of cuts in the edges of the partition.

2. RELATED LITERATURE

Community detection in complex networks has attracted a lot of attention in recent years. The researchers are putting their effort by applying different methodologies to discover such communities. In this section, we provide some of the existing methods which are reviewed in the past decades. Through the existing literature, we came to know that no such existing method talks about how one person (vertex) knows the other (vertex). That means there should be a strong tie between the two vertices in the entire graph. This purpose can be solved by using Strongly Connected Components (SCC), as it identifies the paths between any two vertices involved. The communities are formed in such a way that when there is a path from a vertex u to v, then there should also be a path from v to u. Hence, the intermediate vertices can also have the similar kind of relationship, equivalence relationship, to form strong components, and hence communities.

An improved spectral clustering method for discovering communities in social network is presented in [1]. To make full use of the network feature, the core members are used in this method for mining communities. The authors utilized Page Rank method for discovering communities. In this work, the authors proved that their method is better in terms of time and accuracy.

A good survey on various community detection algorithms can be found in [2]. This gives an elaborate description about different algorithms along with the results that are obtained by those algorithms. In this paper, the authors tested several methods against a recently introduced class of benchmark graphs, with heterogeneous distributions of degree and community size and the results produced in the form of charts. Biologically inspired algorithms are applied for wide variety of problems. Community discovery is no way exempted from this phenomenon. Hence, a genetic algorithmic approach is applied by [3]. The algorithm uses a fitness function able to identify groups of nodes in the network having dense intra – connections, and sparse inter – connections.

A random graph is a graph that is generated by some random process. A random graph is a graph in which properties such as the number of graph vertices, graph edges, and connections between them are determined in some random way. The random graph is defined by the join distribution of the presence or absence of vertices. The inclusion of vertices can be combined to form communities. This method is introduced by [4], as a method of discovering communities in networks. In this paper, the authors used block structures model for the purpose in the context of social sciences, using a Bayesian approach.

Communities are emerging in various types both in good and bad groups. One such ideal way to identify hate group through blogs are done by [5]. The authors proposed a semi-automated approach to analyze virtual communities and to monitor for activities that are potentially harmful to society. The authors used blogs as their data source for this work.

Community discovery is basically a clustering problem, in data mining perception. As inter – cluster members may either be included in one or more clusters, which is so called overlapping of communities. Identifying overlapping of communities is done by [6]. The authors devised a novel algorithm to identify overlapping communities in complex networks by fuzzy c – means clustering approach.

A simple label propagation algorithm for community discovery is done by [7]. The authors used the network structure alone as its guide for the work. This work didn't require any pre-defined objective function or prior information about the communities.

The concept of modularity matrix for community detection is introduced by [8]. In this paper, the authors defined the maximization process that can be written in terms of the eigenspectrum of a matrix, called the modularity matrix, which plays a role in community detection. The algorithms and measures proposed are illustrated with applications to a variety of real-world complex networks.

[9] Showed how community detection can be interpreted as finding the ground state of an infinite range spin glass. In this paper, the community structure of the network is interpreted as the spin configuration that minimizes the energy of the spin glass with the spin states being the community indices.

Random walks has several important advantages like it captures well the community structure in a network, it can be computed efficiently, and it can be used in an agglomerative algorithm to compute efficiently the community structure of a network. This approach for community discovery is used by [10]. The authors proposed a measure of similarities between vertices based on random walks for community discovery.

An extremal optimization method for community discovery was proposed by [11] which is a divisive algorithm for graph partitioning. It optimizes the modularity using a heuristic search based on the extremal optimization EO algorithm. The authors produced the results by taking computer-simulated and real networks and compare them with other approaches.

Community detection using modularity was proposed by [12]. It is an agglomerative hierarchical clustering method. The basic idea of the algorithm was modularity. The author produced the results by taking various applications to prove the efficiency of the proposed method, as it is faster than other previous algorithms.

Problem decomposition to discover communities was applied by [13]. As per this approach, the network is decomposed into manageable sub networks using a multilevel graph partitioning procedure.

We introduced the method of mutual accessibility by using strongly connected components. The description of the method was given in [14]. Our method provides the stability and enhancement of the members of the community, when compared with all other existing methods.

3. METHODOLOGY

Suppose a graph G has V vertices and E edges, mathematically G = (V, E). A strongly connected component of a directed graph G is a maximal set of vertices $C \subseteq V$ such that for every pair of vertices u and v, there is a directed path from u to v and a directed path from v to u. A directed graph is called strongly connected if there is a path from each vertex in the graph to every other vertex. Two vertices are "strongly connected" if they are mutually reachable. The strongly connected components (SCC) of a directed graph G = (V, E) are its maximal strongly connected sub graphs. Two vertices of directed graph are in the same component if and only if they are reachable from each other.

Strong connectedness is an equivalence relation on vertices, and the resulting equivalence classes are called the strongly connected components of the graph. Within a strongly connected component, any vertex can be reached from any other. We can more formally generalize the strongly connected components as follows: Given a graph G = (V, E), where V is a set of vertices (say size n) and E is a set of edges (say size m), the connected components of G are the sets of vertices such that all vertices in each set are mutually connected (reachable by some path), and no two vertices in different sets are connected. Given a strongly connected digraph G, we may form the component digraph G^{SCC} by the following two properties:

- The vertices of G^{SCC} are the strongly connect components of the digraph G.
- There is an edge from v to w in G^{SCC}, if there is an edge from some vertex of component v to some vertex of component w in digraph G.

Algorithms for finding strongly connected components may be used to solve 2 – satisfiability problems. A 2-satisfiability is the problem of determining whether a collection of two - valued variables with constraints on pairs of variables can be assigned values satisfying all the constraints. A 2 – satisfiability instance is unsatisfiable if and only if there is a variable v such that v and its complement are both contained in the same strongly connected component of the implication graph of the instance.

There are two properties of Strongly Connected Components of a directed graph:

1. There should be at least a path from each vertex in the graph to every other vertex
2. There should not be a cycle or loop in the resultant SCC

Tarjan has devised an O (n) algorithm for determining strongly connected components [15]. The algorithm's running time is therefore linear in the number of edges in G (i.e.) O (|V| + |E|). The basic idea of the algorithm is to apply a depth-first search (DFS) begins from a start node. The strongly connected components form the sub trees of the search tree, the roots of which are the roots of the strongly connected components. The nodes are placed on a stack in the order in which they are visited. When the search returns from a sub tree, the nodes are taken from the stack and it is determined whether each node is the root of a strongly connected component. If a node is the root of a strongly connected component, then it and all of the nodes taken off before it form that strongly connected component. The algorithm in pseudo code is given as follows:

```
procedure scc(v)
{
 lowlink(v) := number(v) := ++scc_number
 push(v)
 for all successors w of v do
 if w is not visited then -- v->w is a tree arc
 scc(w)
 lowlink(v) := min( lowlink(v), lowlink(w) )
 elsif number(w) < number(v) then -- v->w is cross link
 if in_stack(w) then
 lowlink(v) := min( lowlink(v), number(w) )
 end if
 end if
 end for
 if lowlink(v) = number(v) then -- next scc found
 while w := top_of_stack_node; number(w) >=
number(v) do
 pop(w)
 end while
 end if
}
```

Fig. 2 is used to explain a digraph and the number of components in it. The vertices of the digraph are numbered 1 through 12. There are four different communities of variable in size for the given digraph. There is a single member community indexed as A, two member community mentioned by C, and three member communities is specified as B and six members community is represented as D. The outline boundaries are used to draw the number of components as communities. The final digraph is also satisfying the properties and 2-satisfiability of SCC.

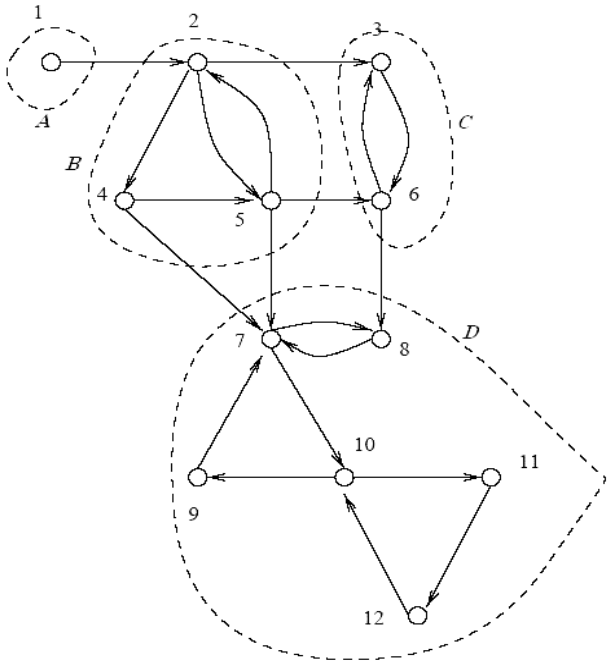

Fig 2: A Sample digraph and its subcomponents as communities

4. MATERIALS AND RESULTS

In this article, we tried to compare the results obtained by our method with the other two prominent methods. We introduced the method of mutual accessibility which provides the enhancement of how the members of the network can be accessible themselves. The other two methods that we have taken for comparison are: i. extremal optimization, proposed by [11] and ii. modularity, introduced by [12]. The results obtained by the methods for various datasets are summarized in table 1.

Table 1. Comparative study of results obtained in different methods for various datasets

Network Type	No. of Nodes	No. of Edges	No. of Communities discovered by method		
			[12]	[11]	[14]
Zachary Karate Club [16]	34	77	2	4	1
American College Football Team [17]	115	613	6	--	1
Jazz Musicians Network [18]	198	2742	4	5	1
E – mail Network [19]	1133	5452	13	15	1
Online Social Network [20]	1899	13838	--	--	4
Synthetic Mobile Network [21]	10000	86619	--	--	1
PGP Network [22]	10680	24316	80	965	1
Co-authorship Network [23]	40421	175693	--	--	954
Facebook New Orleans Network [24]	63565	809212	--	--	146

Some of the datasets are tested by exclusive methods. For example, American College football team dataset was analyzed by [7]. The results are unknown for the same dataset for the method described by [11]. Hence, we represented by hyphens. We tested the betterment of our method by taking additional datasets like co-authorship dataset, which is a test-bed dataset, because this dataset has 40421 vertices and 175693 edges. Our method provides enhancement among the members in the network. We also tested our method by having a large network provided by [24]. This dataset is a well-known social networking site Facebook – New Orleans network provided by [24]. It has 63565 vertices and 809212 edges. Our method discovered 146 communities (clusters).

5. CONCLUSION AND DISCUSSION

Community is a special group formed from an existing network that has specialized property. Community discovery is a phenomenon in which the entire graph can be partitioned into smaller sub partitions, called clusters. In this article, we compared two prominent algorithms that are used for community discovery with the one that we introduced. The results obtained by all the three methods are produced in a tabular format. Our results show that there is an enhancement among the communities discovered.

In the proposed method, communities are discovered among the mutually accessible members in the network. Mutual accessibility is a 2-satisfiability problem. Hence, this enhances that the members knows among themselves within their network. We have to provide visualization technique for the communities discovered. This is taken as a future direction for our research.

6. REFERENCES

[1] Shuzi Niu, Daling Wang, Shi Feng, Ge yu, 2009, An improved spectral clustering algorithm for community discovery, Ninth Intl. Conf. on Hybrid Intelligent Systmes, Vol. 3, 262-267.

[2] Andrea Lancichinetti, Santo Fortunato, 2009, Community detection algorithms: a comparative analysis, arXiv: 0908.1062vl physics soc-ph.

[3] Clara Pizzuti, 2008, Community detection in social networks with Genetic Algorithms, Proceedings of the 10th annual conference on genetic and evolutionary computation, 1137-1138.

[4] Daudin J. J , Pichard F and Robin S, 2008, A mixture model for random graphs, statistical computing 18, 173-183.

[5] Michael Chaua, Jennifer Xu, 2007, Mining communities and their relationships in blogs: a study of online hate groups, Int. J. human – computer studies 65, 57-70.

[6] Shihua Zhang, Rui-Sheng Wang, Xiang-Sun Zhang, 2007, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Physica A374, 483-490.

[7] Raghavan U N, Albert R and Kumara S, 2007, Near linear time algorithm to detect community structures in large – scale networks, Physical Review E76, 036106.

[8] Newman M E J, 2006, Finding community structure using the eigenvectors of matrices, Physical Review E74, 036104.

[9] Richardt J and Bornholdt S, 2006, Statistical mechanics of community detection, Physical Review E74, 016110.

[10] Pascal Pons, Matthieu Latapy, 2005, Computing communities in large networks using random walks, LNCS 3733, 284-293.

[11] Jordi Duch an dAlex Arenas, 2005, Community detection in complex networks using extremal optimization, Physical review E72, 027104.

[12] Newman M E J, 2004, Fast algorithm for detecting community structure in networks, Physical Review E69, 066133.

[13] Narasimhamurthy A, D. Greene, N. hurley and P. Cunningham, 2008, Community finding in large social networks through problem decomposition, 19th Irish conference on Artificial Intelligence and cognitive science (AICS'08).

[14] Dr. M. Mohamed Sathik, A. Abdul Rasheed, 2010, discovering communities in social networks through mutual

accessibility, Intl. Jnl on computer science and engineering, vol. 02, no. 04, 1423-1428.

[15] Robert Tarjan, 1972, Depth – first search and linear graph algorithms, SIAM J. Computing, Vol. 1 , No.2, 146-160.

Dataset References:

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

[17] Girvan M, Newman M. E. J, 2002, Proc. Natl. Acad. Sci, USA 99, 7821-7826

[18] Gleiser P, Danon L, 2003, Adv. Complex Syst 6, 565

[19] Guimera R, Danon L, diaz-Guilera A, Giralf F, Arenas A, 2003, Self-similar community structure in a network of human interactions, Physical Review E, vol 68, 06503

[20] Opsahl T, Panzarasa P, 2009, Clustering in weighted networks, Social Networks 31 (2), 155-163.

[21] Narasimhamurthy A, Greene D, Hurley N, Cunningham P, 2008, Scaling community finding algorithms to work for large networks through problem decomposition, 19th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'08), Cork, Ireland.

[22] Guardiola X, Guimera R, Arenas A, diaz-Guilera A, Streib D, Amaral L. A. N, 2002, Macro- and micro-structure of trust networks, arXiv: cond-mat/0206240v1

[23] Newman M. E. J, 2001, the structure of scientific collaboration networks, Proc. Natl. Acad. Sci., USA98, 404-409.

[24] Minas Gjoka, Macief Kurant, Carter T Butts, Athina Markopoulou, 2010, Walking in Facebook: A case study of unbiased sampling of OSNs, Proceedings of IEEE INFOCOMM '10.