Discovering Communities in Social Networks Through Mutual Accessibility

Dr. M. Mohamed Sathik
Associate Professor in Computer Science,
Sadakathullah Appa College, Tirunelveli – 627011.
Tamil Nadu. INDIA.
mmdsadiq@gmail.com

A. Abdul Rasheed,
Research Scholar, Department of Computer Applications,
Valliammai Engg. College
Chennai. INDIA.
profaar@gmail.com

***Abstract* — Social network gains popularity due to its ease of use, as an application of Web 2.0 which facilitates users to communicate, interact and share on the World Wide Web. A Social network is a set of people or organizations or other social entities connected by a set of social relationships, such as friendship, co – working or information exchange. Social network analysis is the study of social networks to understand their structure and behavior. The study of networks is an active area of research due to its capability of modeling many real world complex systems. One such interesting property to investigate in any typical network is the community structure which is the division of networks into groups. Discovering communities in a social network environment is graph partitioning problem. None of the existing methods discussed about knowing the nodes of the network mutually. Hence, we propose a new approach called "mutual accessibility", to discover communities in a social network environment. We proved comparative study as results by taking both synthetic dataset and real datasets. There is a significant improvement in terms of accuracy and the number of communities discovered in the results obtained by this method.**

Keywords - Data Mining, Graph Partitioning, Community Discovery, Social Network, Mutual Accessibility

I. 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. Users of these sites form a social network, which provides a powerful means of sharing, organizing, and finding content and contacts. Sites, such as Twitter, allow bloggers to share thoughts and feelings instantaneously with friends and family and are much faster than e-mailing. 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.

Social Network Analysis (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 corresponding author only. 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 methodology to collect and analyze relational data. SNA facilitates for analyzing and comparing information flows in an organization as well as groups and individuals. SNA maps both formal and informal relationships that impedes the knowledge flow between interacting units such as who shares what information with whom by what communication media. 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 "community discovery". Discovering communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs.

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. 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. 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. The study of community structure in networks is closely related to the ideas of graph partitioning in graph theory. Finding an exact solution partitioning a graph into a number of sub graphs can be thought of as clustering a graph into number of sub graphs. Hence, the members (vertices) in the sub graph are denser than the members in another sub graph. In other words, we can say a cluster of nodes in a graph

should have many links among themselves, but few links to nodes outside the cluster, that is the community. Consider the figure shown in fig 1, which contains three groups of communities. This also shows the interaction level among the members of intra – community and also the interaction with inter – community members.

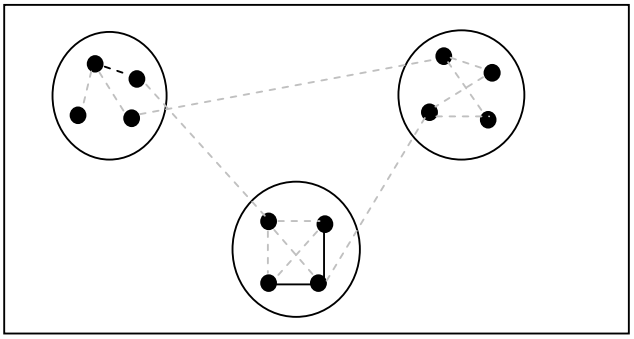

Figure 1 A group of three communities and the interaction among the members

On clustering a graph, we have to put vertices in a cluster if their attributes are similar while they also have a lot of links among themselves. Mathematically, we can formulate the problem of graph partitioning as follows: Given a graph G = (V, E), where V is the set of vertex and E the set of edges that determines the connectivity between the nodes. 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.

II. EXISTING 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 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.

III. STRONG CONNECTIVITY IN SOCIAL NETWORKING

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:

i. The vertices of G^{SCC} are the strongly connect components of the digraph G.
ii. 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.

Finding connected components is used in many diversified fields such as computer vision, where pixels in a two- or three-dimensional image are grouped into regions representing objects or faces of objects; spin models in physics; VLSI circuit design; communication networks; program analysis and implementation; neural networks and economics. The objective of discovering strongly connected components of a graph is to find path from every pair of vertices. "Connectedness" is a key property required for a community. As far as communities in social network are concerned, individual components are considered as one community. The members of the community are the connected vertices of the graph. The component digraph can also be considered as individual communities, in which there should not be any cycles. That is, the resultant graph would be a directed acyclic graph (DAG).

Algorithms for finding strongly connected components may be used to solve 2 – satisfiability problems. 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.For understanding purpose, we explained with the following digraph, as in Figure 2.

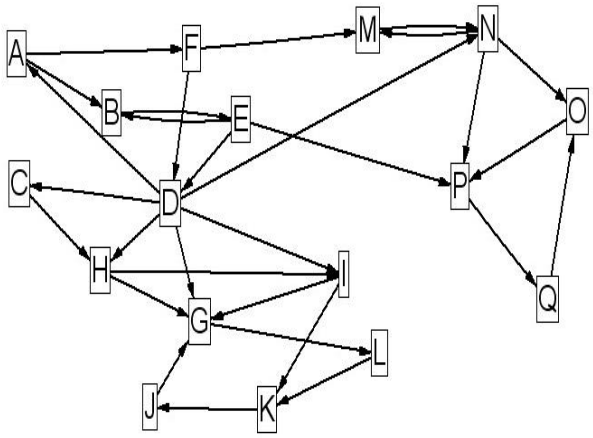

Figure 2 An Example digraph G

A strongly connected component is maximal subgraph of a directed graph such that for every pair of vertices u, v in the subgraph, there is a directed path from u to v and a directed path from v to u, denoted as u ~ v. We can say that u and v are reachable from each other. Tarjan has devised an O(n) algorithm for determining strongly connected components[11]. The algorithm's running time is therefore linear in the number of edges in G (i. e) O(|V| + |E|).

The outline of the algorithm is given as follows:

```
1: DFS(G) to compute finishing time of
each vertex f[v]
2: Compute G^T (Transpose of G)
3: DFS(G^T) in the order of decreasing
finish time of vertices f[v]
4: Output the vertices of each tree in the
DFS forest as a separate SCC
```

The following figure 3 shows the number of communities discovered for the given digraph as in figure 2.

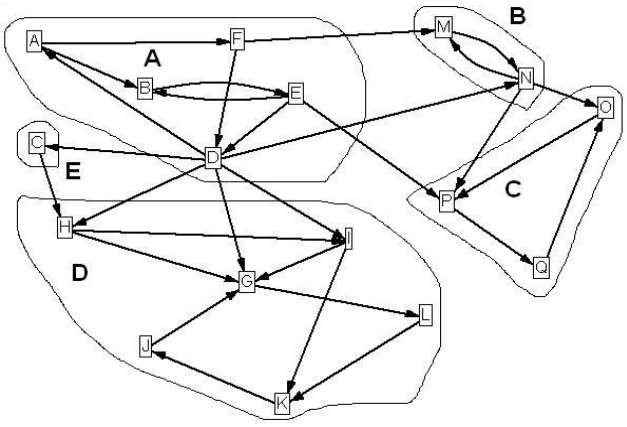

Figure 3 Communities Discovered of fig 2 by applying SCC

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

These above two properties are also satisfied once SCC is computed by using the algorithm. The final component graph is generated as a directed acyclic graph (DAG), as it satisfied the above two properties. Figure 4 is the equivalent component graph of Figure 3.

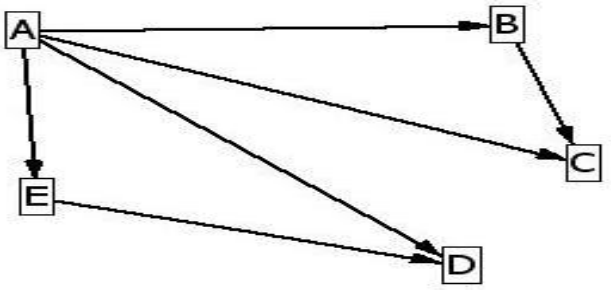

Figure 4 Component Graph of Figure 3

IV. PROOF

Strong connectivity among the vertices of a graph is an equivalence relation. When *a is strongly connected to b*, denoted as a ~ b, we allow the two paths to share vertices or even to share edges. With one vertex and no edges, there may be short paths so that any vertex is strongly connected to itself. So cycles in the graph can be thought of as strongly connected components, and hence a community.

A relation is another word for a collection of pairs of objects. An equivalence relation a # b satisfies the following three properties:

i. *Reflexive property*: For all a, a # a. Any vertex is strongly connected to itself, by definition.

ii. *Symmetric property*: If a # b, then b # a. For strong connectivity, this follows from the symmetry of the definition. The same two paths are looked in other order. That is, when there is a path form a to b, then another one may be from b to a. hence a ~ b and b ~ a.

iii. *Transitive property*: If a # b and b # c, then a # c. For strong connectivity: if a ~ b and b ~ c, we have four paths: a-b, b-a, b-c, and c-b. Concatenating them in pairs a-b-c and c-b-a produces two paths connecting a-c and c-a, so a ~ c, showing that the transitive property holds for strong connectivity.

For any equivalence relation a # b, we can define equivalence classes by the formula [a] = {b | a # b}. The equivalence classes for strong connectivity are called strongly connected components. These sets have the property that they partition the space of all vertices into disjoint subsets.

The key property that relates DFS to strong connectivity is that strongly connected components form subtrees of the DFS tree.

Proof of Correctness:

Let C and C' be two distinct Strongly Connected Components of the graph G = (V, E). Let u, v ∈ C and u', v'∈ C'. If there is a path from u to u' then there cannot be a path from v to v'. Hence the property of SCC is proved.

Proof by Contradiction:

Another property of Strongly Connected Component is that there should not be any cycles in the components. That is, the resultant components will be a Directed Acyclic Graph (DAG). Let us prove it by contradiction.

Suppose component graph of G = (V, E) was not a DAG and G comprised of a cycle consisting of vertices $v_1, v_2, \ldots, v_n$. Each v_i (i = 1 to n) corresponds to a strongly connected component (SCC) of component graph G. If $v_1, v_2, \ldots, v_n$ themselves form a cycle then each vi (i = 1 to n) should have been included in the SCC corresponding to v_j (j = 1 to n and i ≠ j). But each of the vertices is a vertex from a difference SCC of G. Hence, we have a contradiction. Therefore, SCC of G is a directed acyclic graph.

With respect to communities, the members of the community have a common interest or property. Therefore, the members of the community are grouped together to share the common interest. Hence, it is also proved that the discovered communities using strongly connected components are disjoint.

V. EXPERIMENTS

As a proof of the concept, we have taken different real – world network datasets for our study. In this section, we provided the comparative study of communities discovered by our method and it is shown that our results are better in terms of mutual accessibility among the members in the community. We used Python for implementation of the algorithm.

Intra - organizational network

This data set from a research team in a manufacturing company with 77 employees and provided by [15]. The network is based on the employees' awareness of each others' knowledge and skills. The dataset also contains information about the people. There are 77 vertices and 2326 ties (that means edges). We discovered two communities from this dataset. A community with 76 members and rest of the member is an isolated one. The result obtained for this dataset is shown in Fig 5.

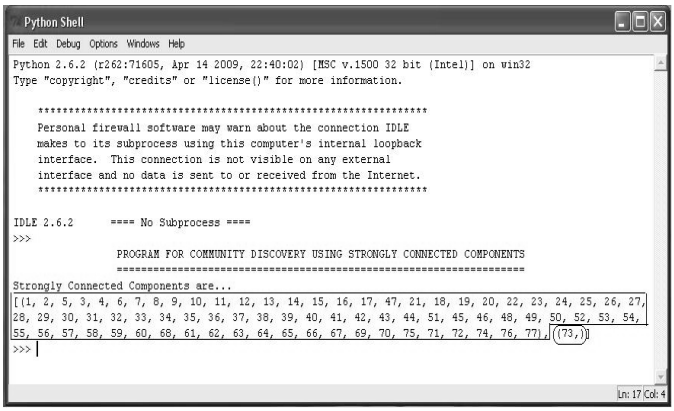

Figure 5 Discovered Communities for Intra – Organizational Network Dataset

Books About US Politics

This data set is provided by [14]. A network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. This dataset has 105 vertices and 441 edges. Edges represent frequent co – purchasing of books by the same buyers, as indicated by the "customers who bought this book also bought these other books" feature on Amazon. In this data set, no communities were discovered. That means, all the vertices are isolated and none of the nodes are mutually accessible by any other nodes in the entire network. The result obtained for this dataset is shown in Fig 6.

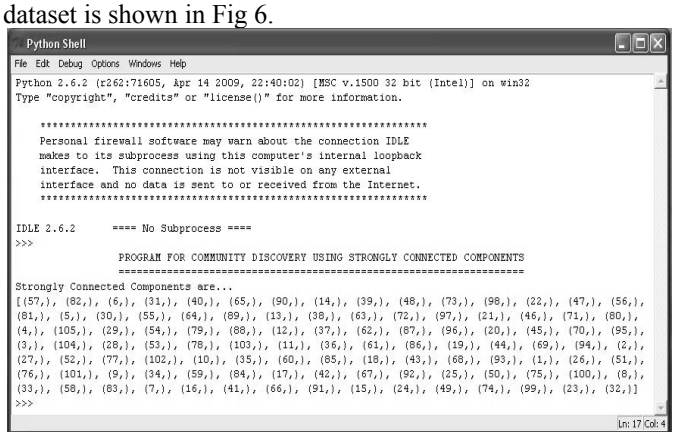

Figure 6 Discovered Communities for Books About US Politics Dataset

Online Social Network Data Set

We have taken another online social network dataset provided by [12]. The network originates from an online communication network among students at University of California, Irvine. The edgelist includes the users that sent or received at least one message over a period of time. There are 1,899 vertices. A total number of 59,835 online messages were sent among these over 20,296 directed ties, edges. There are 1899 vertices and 20296 edges in the entire network. We tested for mutual accessibility among these vertices with one among the other. This makes the members of the community to know them mutually. It was amazing to see that there is in existence of a single community with 1364 vertices out of 1899. These vertices known among themselves. The remaining vertices are isolated from the entire network. Fig 7 shows the result of the dataset.

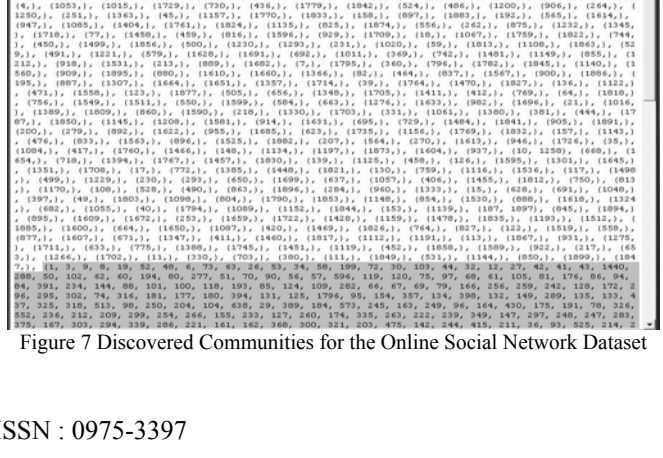

Figure 7 Discovered Communities for the Online Social Network Dataset

Computational Geometry Collaboration Network Dataset

This data set is provided by [17]. This dataset contains the network with 7343 vertices and 11898 edges. Author X wrote a joint work with author Y. The authors collaboration network in computational geometry was produced from the BibTeX bibliography [Beebe, 2002] obtained from the Computational Geometry Database geombib, version February 2002 [Jones, 2002]. Two authors are linked with an edge, iff they wrote a common work (paper, book, ...). Though the number vertices are in few thousands, there are no communities discovered for this dataset. Fig 8 is the result obtained for this dataset.

Figure 8 Discovered Communities for Collaboration Network Dataset

Synthetic Mobile Network Data Set

This dataset is provided by [18]. It is designed to examine how community finding techniques scale to large, sparse graphs, approaching the size of those occurring in real-world problems such as measuring churn in mobile subscriber networks. The dataset had 10000 nodes. As we obtained the result in this dataset, all the nodes in the entire network are not accessible mutually. All nodes are considered independent communities that mean, a community with single member. Figure 9 shows the result obtained by implementation.

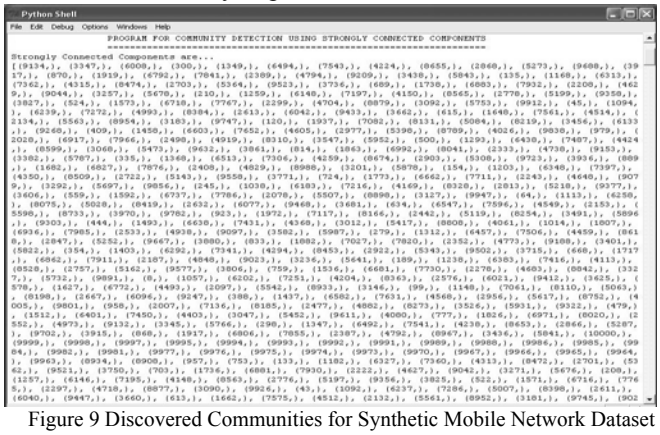

Figure 9 Discovered Communities for Synthetic Mobile Network Dataset

We provide the comparative study of number of communities discovered for all the above datasets as shown in table 1. When the number of communities discovered is equivalent to the number of nodes in the result, it means that the nodes are independent. The nodes are not mutually accessible from each other.

Table 1 A comparative study of the results obtained for different datasets

Data Set	No. of Vertices	No. of Edges	Communities Discovered
Intra Organizational Network	77	2326	2
Books about US Politics	105	441	105
Online Social Network	1899	20296	536
Authors' Collaboration Network	7343	11898	7343
Synthetic Mobile Network	10000	86621	10000

REFERENCES

[1] Shuzi Niu, Daling Wang, Shi Feng, Ge Yu, "An Improved Spectral Clustering Algorithm for Community Discovery", Proceedings of the Ninth International Conference on Hybrid Intelligent Systems , vol. 3, pp. 262 – 267, 2009

[2] Andrea Lancichinetti, Santo Fortunato, "Community detection algorithms: a comparative analysis", arXiv:0908.1062v1 [physics.soc-ph], 2009

[3] Clara Pizzuti, "Community Detection in Social Networks with Genetic Algorithms", Proceedings of the 10th annual conference on Genetic and evolutionary computation, pp 1137-1138, 2008[4] Daudin J. J, Picard F and Robin S, "A Mixture Model for Random Graphs", Statistics and Computing 18, pp 173—183, 2008

[5] Michael Chaua, Jennifer Xu, "Mining communities and their relationships in blogs: A study of online hate groups", Int. J. Human – Computer Studies 65, pp 57–70, 2007

[6] Shihua Zhang, Rui-Sheng Wang, Xiang-Sun Zhang, "Identification of overlapping community structure in complex networks using fuzzy c-means clustering", Physica A 374, pp 483–490, 2007

[7] Raghavan, U.N. and Albert, R. and Kumara S, "Near linear time algorithm to detect community structures in large-scale networks", Phys Rev E 76, 036106. 2007

[8] Newman M E J, "Finding community structure using the eigenvectors of matrices", Phys. Rev. E 74, 036104 2006

[9] Reichardt. J and Bornholdt S, "Statistical Mechanics of Community Detection", Phys. Rev. E, 74, 016110, http://arxiv.org/abs/cond-mat/0603718, 2006

[10] Pascal Pons, Matthieu Latapy, "Computing communities in large networks using random walks", http://arxiv.org/abs/physics/0512106 P. Yolum et al. (Eds.): ISCIS 2005, LNCS 3733, pp 284-293, 2005

[11] Robert Tarjan, "Depth-First Search and Linear Graph Algorithms", SIAM J. Computing, Vol. 1, No. 2, pp 146-160, 1972

[12] Opsahl, T., Panzarasa, P., 2009, "Clustering in weighted networks", Social Networks 31 (2), 155-163, doi: 10.1016/j.socnet.2009.02.002

[13] ICS 161: Design and Analysis of Algorithms Lecture notes available online at: http://www.ics.uci.edu/~eppstein/161/960220.html#sca

[14] V. Krebs, unpublished, http://www.orgnet.com/

[15] Cross, R., Parker, A., "The Hidden Power of Social Networks", Harvard Business School Press, Boston, MA, 2004

[16] M. Boguña, R. Pastor-Satorras, A. Diaz-Guilera and A. Arenas, Physical Review E, vol. 70, 056122, 2004

[17] Vladimir Batagelj and Andrej Mrvar, Pajek datasets, 2006

[18] A. Narasimhamurthy, D. Greene, N. Hurley, and P. Cunningham, "Scaling Community Finding Algorithms to Work for Large Networks Through Problem Decomposition", 2008

AUTHORS PROFILE

First Author Dr. M. Mohamed Sathik received his Ph.D., in Computer Science from Manonmaniam Sundaranar University, Tirunelveli, INDIA in 2006. He also received M. Phil., in Computer Science, MBA., M. Tech., in Compute Science and Information Technology. He has more than 25 years experience in teaching. He is a recognized supervisor for M. Phil., and Ph.D., in various universities. He published several papers in international journals. He is also a review member in several journals of international repute. He chaired international conferences. His research interest includes virtual reality, data mining and image processing.

Second Author Mr. A. Abdul Rasheed is graduated in MCA and M. E., with specialization in Computer Science and Engineering.. He has 13 years experience in teaching. He is pursuing his research in Computer Science and Engineering in Manonmaniam Sundaranar University, Tirunelveli under the guidance of Dr. M. Mohamed Sathik.