ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

Identification of Community in Social Network

Monica G Tolani*, Dr. M.Vijayalakshmi

M.E Student, Department of Information Technology, V.E.S.I.T, Chembur, India

Professor, Department of Information Technology, V.E.S.I.T, Chembur, India

ABSTRACT: Social networks have attracted much attention recently. Social network analysis finds its application in many current business areas. Different studies have been conducted to automatically extract social networks among various kinds of entities from the Web. Community detection is one of the most important and interesting research areas in social network analysis. Many works are dedicated to methods and algorithms for detecting communities in different kinds of social networks extracted on the Web. In this paper we give a brief description of a new method which can help to identify communities in networks.

KEYWORDS: social network; community detection; social network analysis.

I. INTRODUCTION

Online social networks have become center of attraction for a wide variety of audience. Nearly every person has a profile on Facebook, Google Plus, Orkut, Twitter etc which are collectively termed as Social Networking Sites (SNS). People usually communicate to others via these SNS and share their feelings, emotions, achievements, sorrows and nearly everything about their life.

A social network is a graphical representation of the communication among people, where people are represented as nodes and the edges between a pair of nodes represent some kind of communication between them [1]. It is a natural tendency of human beings to socialize and communicate with likeminded people. This selective nature of communication forms a typical structure in networks known as community. Inside communities people communicate more often to the members of the community where as they tend to talk less to the members outside the community.

The problem of detecting communities has gained huge interest from the researchers now days. When a social network is represented as an undirected graph G (V, E), where V represents the individuals and E represents the links among them, then a community C is such that the number of links going outside from the vertices in C is far less than the number of links with both vertices inside C. Communities are an important feature of complex networks [7].

Community detection algorithms answer a wide range of questions regarding the behavior and interaction patterns of people. Community detection is the process of finding such dense group of nodes which have high internal edge density. The first most challenge in the domain of community detection is that there is no definition of a community which is accepted generally; still there is a large of community detection algorithms available which produce effective results. The communities can be disjoint or overlapping [2]. In this paper we restrict our work to networks that are unweighted, and undirected, which form disjoint communities. Recent advances in SNSs provide quality data to the researchers which allow them to experiment and test their algorithms on real world data. This makes the testing quite rigorous and hence the algorithms produced are of high standards. Finding communities have been in focus of researchers because it helps them understand the communication patterns as a function of structural properties of the networks.

In social sciences, cliques are among the most visible and interesting types of groups in society[8]. In network theory, a clique embodies a basic community as it has the greatest possible edge density. However, the requirement that each pair of vertices be connected is too strict. Recently techniques have been developed where cliques are considered as the starting communities to be merged based on modularity or other metrics. However note that the task of finding cliques itself is computationally expensive. To tackle this problem, heuristics have been developed to obtain sub-optimal solution in a reasonable time.

This paper presents a systematic and organized study of overlapping community detection techniques and proposes a new clique based community detection algorithm. The strengths and weaknesses of each technique is also a matter of focus in this paper. The major contributions of this paper will be to serve as a base for those starting research in this direction and to provide them with the existing state of art algorithms for the research problem. The proposed algorithm with the advancements of existing algorithm in the domain proofs the validity of the algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

II. RELATED WORK

There are a numerous algorithms proposed for community detection, which are difficult to be reviewed one by one, so a classification of the algorithms is presented on the basis of the basic operating mechanism of the algorithms[3]. Link Partitioning Algorithms work by breaking the links and creating dendrograms whereas Clique Based Algorithms exploit the properties of cliques to detect communities. Agent Based Algorithms are based on propagation of labels between vertices of the graph and the Fuzzy algorithms are based on the mixed membership models.

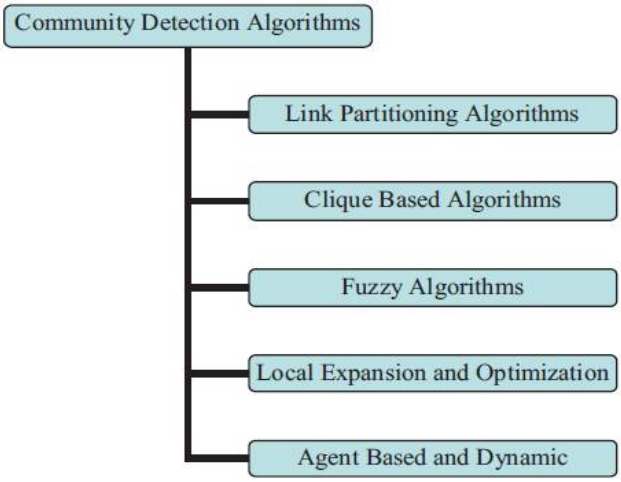


Fig 1: Classification of Community Detection Algorithms

Local Expansion and Optimization algorithms are based on some fitness measure to be optimized locally. Fig. 1 shows the classification of community detection algorithms. Following are the details of each category of the algorithms.

A. LINK PARTITIONING ALGORITHM

The basic idea of link partitioning algorithms is to partition links to discover the communities. Two steps of every link partitioning algorithms are:

Step 1: Construct the Dendrogram (a representation to store hierarchical structure of network). An illustration of dendrogram is shown in Fig 2.

Step 2: Partition the Dendrogram at some threshold.

A node will be identified as overlapping if the links to the node are present in more than one cluster. Links are partitioned by hierarchical clustering[1] in on the basis of edge similarity. If the only available information is the network topology, the most fundamental characteristic of a node is its neighbors. Since a link consists of two nodes, it is natural to use the neighbor information of the two nodes when we define a similarity between two links. If we are given a pair of links ejk

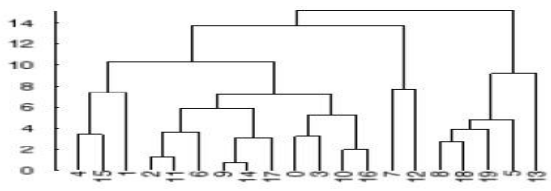


Fig 2 : Illustration of a Dendrogram

and eik , the edge similarity between these two links is calculated by Jaccard index as:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

$$S(e_{jk}, e_{ik}) = \frac{|(N_i \cap \tilde{N}_j)|}{|(N_i \square N_j)|}$$

Ni is the set of vertices which are in the neighborhood of vertex i including vertex i. After calculating edge similarities linkage clustering is done to find hierarchical communities. With this similarity, we use single-linkage hierarchical clustering to find hierarchical community structures. Generally single linkage hierarchical clustering is done because of its simplicity and efficiency which enables us to apply it on large networks.

B. CLIQUE BASED ALGORITHM

A clique is a maximal subgraph in which all nodes are adjacent to each other [4]. The input to Clique based algorithms is a network graph G and an integer k. Clique based algorithms have following steps in general:

a. Find all cliques of size k in the given network.

b. Construct a clique graph. Any two k-cliques are adjacent if they share k-1 nodes.

c. Each connected components in the clique graph form community.

The Bron–Kerbosch algorithm is an algorithm for finding maximal cliques in an undirected graph [3]. That is, it lists all subsets of vertices with the two properties that each pair of vertices in one of the listed subsets is connected by an edge, and no listed subset can have any additional vertices added to it while preserving its complete connectivity.

C. AGENT BASED AND DYNAMIC ALGORITHM

Three popular algorithms that come under this category are Speaker Listener Propagation Algorithms SLPA, Community Overlap Propagation Algorithm COPRA, and Label Propagation algorithm LPA. SLPA is Speaker- Listener label propagation algorithm [5], in which a node is called speaker if it is spreading information and is called a listener if it is consuming information. Labels are spread according to pair wise interaction rules. In SLPA a node can have many labels depending upon the underlying information it has learned from the network. The best part of SLPA is that it doesn't require any prior information about the number of communities in the network. In other two algorithms the node forgets the information it has learned in previous iterations but in SLPA each node has a stored memory in which it stores all the information it has learned about the network in form of labels.

Label Propagation algorithm is extended to overlapping case by allowing multiple labels for a node. Initially all nodes have their own unique label, labels are updated upon iterations depending upon the labels occupied by the maximum neighbors. Nodes with same labels form a community.

In COPRA each label consists of a belonging coefficient and a community identifier. The sum of belonging coefficients of communities over all neighbors is normalized.

D. FUZZY ALGORITHM

The overlap between communities can be of two types, one is the crisp overlap in which each node either belongs to a community or doesn't, the belonging factor is 1 for all the communities a node is a member. The other type of overlap is the fuzzy overlap in which each node can be a member of communities with belonging factor in the range 0 to 1.The membership strength of a node to a community is denoted by bnc and if we sum the belonging coefficients of a vertex for all the communities of which it is a member the result will be 1. In non-fuzzy overlapping, each vertex belongs to one or more communities with equal strength: an individual either belongs to a community or it does not. With fuzzy overlapping, each individual may also belong to more than one community but the strength of its membership to each community can vary. It is expressed as a belonging coefficient that describes how a given vertex is distributed between communities.

The major drawback of fuzzy based overlapping community detection methods is the need to calculate the dimensionality k of the membership vector; this value is generally passed as a parameter to the algorithms, while some algorithms calculate it from the data. Only a few fuzzy methods have shown good results.

E. LOCAL OPTIMIZATION ALGORITHM

These algorithms make use of a community quality score which needs to be optimized in order to reveal the community structure of a given network[6]. The popular algorithms that come under this category are Community Overlap Newman-

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

Girvan Algorithm (CONGA), LFM (Lanccichinetti and Fortunato), Connected Iterative Scan (CIS), Order Statistics Local Optimization Method (OSLOM), (intrinsic Longitudinal Community Detection)

Community Overlap Newman-Girvan Algorithm (CONGA) works by dividing the network into a number of seed communities and adding the vertices to those seed communities until the density of the community has increased to its maximum. LFM (Lanccichinetti and Fortunato) method was proposed to find both the hierarchical as well as overlapping community structure of the network. Connected Iterative Scan (CIS) which finds communities by utilizing the optimality principles and connectedness properties of the network. Order Statistics Local Optimization Method (OSLOM) is the first method which takes into account the edge directions and edge weights. iLCD (intrinsic Longitudinal Community Detection)detects both temporal as well as static communities.

III. PROPOSED WORK

It is observed that there are a lot of techniques are proposed for the detection of communities in social networks, very few of them used cliques. Most of the attention was on detecting disjoint communities, where as community overlap is a general phenomenon in social networks. Only a few overlapping community detection algorithms exist which can detect overlapping communities, the problem with these algorithms is that they fail when the community overlaps are denser.

So there is a need of new technique for clique based community detection which overcomes the major issues. This paper presents a new clique based community detection technique which modifies the present Bron-Kerbosch maximal clique finding algorithm.

The Bron–Kerbosch algorithm is an algorithm for finding maximal cliques in an undirected graph [2]. That is, it lists all subsets of vertices with the two properties that each pair of vertices in one of the listed subsets is connected by an edge, and no listed subset can have any additional vertices added to it while preserving its complete connectivity.

Bron-Kerbosch algorithm has two versions of it. The first version is a implementation of the basic algorithm that searches for all maximal cliques in a given graph G. The version 2 was modified version of 1, in which it makes a recursive call for every clique, maximal or not. To save time and allow the algorithm to backtrack more quickly in branches of the search that contain no maximal cliques, Bron and Kerbosch introduced a variant of the algorithm involving a "pivot vertex" u.

A. VERSION 1

Given graph G, given three sets R, P, and X, it finds the maximal cliques that include all of the vertices in R, some of the vertices in P, and none of the vertices in X. In each call to the algorithm, P and X are disjoint sets whose union consists of those vertices that form cliques when added to R. In other words, P ∪ X is the set of vertices which are joined to every element of R. When P and X are both empty there are no further elements that can be added to R, so R is a maximal clique and the algorithm outputs R.

The recursion is initiated by setting R and X to be the empty set and P to be the vertex set of the graph. Within each recursive call, the algorithm considers the vertices in P in turn; if there are no such vertices, it either reports R as a maximal clique (if X is empty), or backtracks. For each vertex v chosen from P, it makes a recursive call in which v is added to R and in which P and X are restricted to the neighbor set N(v) of v, which finds and reports all clique extensions of R that contain v. Then, it moves v from P to X to exclude it from consideration in future cliques and continues with the next vertex in P.

B. VERSION 2

A "pivot vertex" u, chosen from P (or more generally, from P ∪ X). Any maximal clique must include either u or one of its non-neighbors, for otherwise the clique could be augmented by adding u to it. Therefore, only u and its non-neighbors need to be tested as the choices for the vertex v that is added to R in each recursive call to the algorithm. Modification to the above two versions is done to get more accurate result in terms of large sparse social graphs and other real world graphs.

C. MODIFICATION OF THE ALGORITHM-PROPOSED ALGORITHM

The modification basically proposed the processing of all the vertices and to use the intelligent backtracking method. The pseudo code of algorithm proposed id as follows:-

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

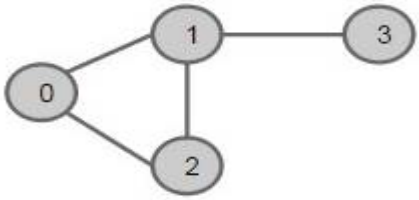
Pseudo code:-

input: a simple graph G.
output: all maximum cliques in G.

Step 1: proc BronKerboschWithProcess(P, R, X)
P={V} //set of all vertex in Graph G
R={}
X= {}
Step 2: for each vertex v in a processing ordering of G
Step 3:BronKerboschWithPivot(P ∩ {v}, R ∪ {V} , X \{v})
end for
End Proc
Step:4 proc BronKerboschWithPivot(P, R , X)
Step 5: if
P∪ X= {}
then
print set R as a maximal clique
end if
Step 6: Choose a pivot u from set P U X
Step 7: for each vertex v in P \nbrs(u)
do
BronKerboschWithPivot(P ∩ {v}, R ∪ {V} , X \{v})
P=P\{v}
X ∪ {v}
end for
End proc

IV. EXAMPLES AND RESULTS

I. Consider the Graph G



The input from file is in the form of :-

- *The first line of the file gives the total number of Graphs, T.*
- *For each graph G*
- *total Nodes*
- *total Edges E*
- *For the next E lines, each contains two space-separated integers Src End*

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

Example Input for the Graph above:

1

4

4

0 1

0 2

1 2

2 3

Output from version 1:-

Max Cliques

*******Star Graph 1**********

{}, {0,1,2,3}, {}

{0}, {1,2}, {}

{0,1}, {2}, {}

{0,1,2}, {}, {}---------- Maximal Clique : 0 1 2

{0,2}, {}, {1}

{1}, {2}, {0}

{1,2}, {}, {0}

{2}, {3}, {0,1}

{2,3}, {}, {} ---------- Maximal Clique : 2 3

{3}, {}, {2}

Output from version 2:-

Max Cliques with Pivot

*******Star Graph 1*****************

{}, {0,1,2,3}, {}

Pivot is 2

{2}, {0,1,3}, {}

Pivot is 1

{2,1}, {0}, {}

Pivot is 0

{2,1,0}, {}, {} --- Maximal Clique : 2 1 0

{2,3}, {}, {} --- Maximal Clique : 2 3

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

Output from Proposed methodology:-

Max Cliques with Processing Ordering

*******Star Graph 1*****************

Processing of graph is: (3,0,1,2,)

{3}, {2}, {}

{3,2}, {}, {} ---------- Maximal Clique : 3 2

{0}, {1,2}, {}

{0,2}, {1}, {}

{0,2,1}, {}, {} ---------- Maximal Clique : 0 2 1

{1}, {2}, {0}

{1,2}, {}, {0}

{2}, {}, {3,0,1}

II. Consider the Graph F, let it be facebook graph which shows the interaction between its users.

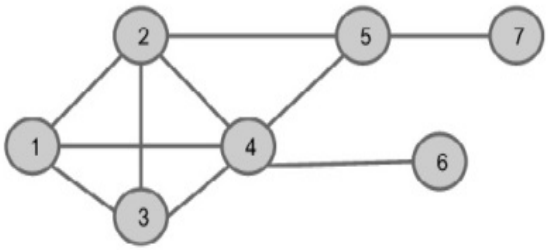


Fig 4 Graph F

Example Input for the Graph above:

1

7

10

1 2

1 3

1 4

2 3

2 4

2 5

3 4

4 5

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

4 6

5 7

Output from version 1:-

Max Cliques Without Pivot

*******Star Graph 1**********

{}, {1,2,3,4,5,6,7}, {}

{1}, {2,3,4}, {}

{1,2}, {3,4}, {}

{1,2,3}, {4}, {}

{1,2,3,4}, {}, {} ---------- Maximal Clique : 1 2 3 4

{1,2,4}, {}, {3}

{1,3}, {4}, {2}

{1,3,4}, {}, {2}

{1,4}, {}, {2,3}

{2}, {3,4,5}, {1}

{2,3}, {4}, {1}

{2,3,4}, {}, {1}

{2,4}, {5}, {1,3}

{2,4,5}, {}, {} ---------- Maximal Clique : 2 4 5

{2,5}, {}, {4}

{3}, {4}, {1,2}

{3,4}, {}, {1,2}

{4}, {5,6}, {1,2,3}

{4,5}, {}, {2}

{4,6}, {}, {} ---------- Maximal Clique : 4 6

{5}, {7}, {2,4}

{5,7}, {}, {} ---------- Maximal Clique : 5 7

{6}, {}, {4}

{7}, {}, {5}

Output from version 2:-

Max Cliques with Pivot

*******Star Graph 1**********

{}, {1,2,3,4,5,6,7}, {}

Pivot is 4

{4}, {1,2,4,5,6}, {}

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

Pivot is 2

{4,2}, {1,3,5}, {}

Pivot is 5

{4,2,1}, {3}, {}

Pivot is 3

{4,2,1,3}, {}, {} --- Maximal Clique : 4 2 1 3

{4,2,3}, {}, {1}

Pivot is 1

{4,2,5}, {}, {} --- Maximal Clique : 4 2 5

{4,6}, {}, {} --- Maximal Clique : 4 6

{7}, {5}, {}

Pivot is 5

{7,5}, {}, {} --- Maximal Clique : 7 5

Output from Proposed methodology:-

Max Cliques with Processing Ordering

*******Star Graph 1**********

Processing is: (6,7,5,1,3,2,4)

{6}, {4}, {}

{6,4}, {}, {} ---------- Maximal Clique : 6 4

{7}, {5}, {}

{7,5}, {}, {} ---------- Maximal Clique : 7 5

{5}, {2,4}, {7}

{5,4}, {2}, {}

{5,4,2}, {}, {} ---------- Maximal Clique : 5 4 2

{1}, {3,2,4}, {}

{1,4}, {3,2}, {}

{1,4,2}, {3}, {}

{1,4,2,3}, {}, {} ---------- Maximal Clique : 1 4 2 3

{3}, {2,4}, {1}

{3,4}, {2}, {1}

{3,4,2}, {}, {1}

{2}, {4}, {5,1,3}

{2,4}, {}, {5,1,3}

{4}, {}, {6,5,1,3,2}

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

The above all output shows interaction between users (1,2,3,4);(2,4,5);(5,7);(4,6) which leads to the formation of cliques.

The output of proposed algorithm shows the processing of the Graph which is considered, hence the backtracking method calls the processed output order for its computation leading to minimize the number of recursive calls made by the algorithm, the savings in running time.

V. CONCLUSION

Community detection approaches have attracted a lot of attention of researchers in recent years and there is a considerable increase in the number of algorithms published for solving the issue as it has applications in various domains like microbiology, social science and physics. Analyzing community structure in social networks has emerged as a topic of growing interest as it shows the interplay between the structures of the network and its functioning.

The paper tries its best to review all popular algorithms, but the study is by no means complete as there are newer algorithms discovered at a fast rate because of the growing interest of researchers in this domain. This paper describes nearly all the algorithms which exist for community detection, and also reviews their strengths and weaknesses. The basic concepts required for understanding the problem of community detection are described in great detail. The main goal was to come up with a technique which is better than the current state of art solutions. The proposed technique of maximum clique finding technique is described. The proposed technique performs well as compared to the classical algorithms. Tests on real world networks have proved that the community structure identified by the proposed technique is better than classical solutions proposed by other researchers.

In future the efforts will be concentrated on further optimizing the technique with color preprocessing technique of graph and using the recursive backtracking of the algorithm more efficiently.

REFERENCES

[1] Girvan, M., and Newman, M. E. “Community structure in social and biological networks.” Proceedings of the National Academy of Science USA 99, 12 ,7821—7826. (June 2002)
[2] Coen Bron & Joep Kerbosch,Algorithm 457,”Finding All Cliques of an Undirected Graph,Comm ACM(Sep 1973),Volume 16, Number 9.
[3] Ashish Kumar Singh, Dr Sapna Gambhir,” Greedy Social Algorithm for Overlapping Community Detection in Online Social Networks,2014-IEEE.
[4] Diana Palsetia, Md.Mosterfa Ali Patwary, William Hendrix, Ankit Agarwal, Alok Choudhary,”Clique Guided Community Detection”,2014-IEEE International Conference on Big Data.
[5] Xie, J., Szymanski, B. K., and Liu, X., “SLPA: Uncovering overlapping communities in social networks via a speakerlistener interaction dynamic process”. In Proc. of ICDM Workshop.344–349. (2011).
[6] Cazabet R.,Ambald C." Detection of overlapping communities in dynamical social networks". In Proceedings of the 2nd IEEE International Conference on Social Computing (SOCIALCOM’10).309–314.(2010).
[7] Andrea Lancichinetti, Filippo Radicchi, Jose´ J. Ramasco, Santo Fortunato, ”Finding Statistically Significant Communities in Networks” ,Plos One,April 2011,Volume 6,Issue 4,e18961.
[8] Suqi Zhang, Yongfeng Dong, Jun Yin, Jingjin Guo, “Improved Ant Colony Algorithm for finding the maximum clique in Social Network” , 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing.