International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
National Conference on Advances in Engineering and Technology
(AET- 29th March 2014)

RESEARCH ARTICLE OPEN ACCESS

A Comprehensive Review of Overlapping Community Detection Algorithms for Social Networks

Ashish Kumar Singh[1], Sapna Gambhir[2]
[1](YMCA University of Science and Technology
Email: ashwebdeveloper@gmail.com)
[2](YMCA University of Science and Technology
Email: sapnagambhir@gmail.com)

Abstract—
Community structure is an interesting feature found in many social networks which signifies that there is intense interaction between some individuals. These communities have a tendency to overlap with each other as there are nodes that can belong to multiple communities simultaneously. Detection of such overlapping communities is a challenging task; it still remains a topic of interest for the researchers as it answers many questions about the behavior of the network and its operation as a function of its structure. This paper reviews overlapping community detection techniques proposed so far and points out their strengths and weaknesses. The paper also presents insightful characteristics and limitations of the existing state of art algorithms to solve the problem of overlapping community detection.
Keywords—Overlapping Community detection, Online Social Networks, Complex Networks, Community Structure.

I. INTRODUCTION

Online social networks have become a primary means of communication nowadays; they attract 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 this communication has attracted a lot of research focus in recent years under the domain named *Social Network Analysis.* 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. A very interesting feature in social networks is the formation of *Communities.* A *community* is a group of individuals in a social network who communicate more frequently with each other than with others outside the group. When a the social network is represented as a graph G (V, E), where V representing the individuals and E representing the connections among them, then a community $C \subset G$ such that the number of edges going outside from the vertices in C is far less than the number of edges with both vertices inside C. The detection of such communities is not trivial and is quite challenging as it is completely different from two similar and well studied problems in computer science namely *Clustering* and *Graph Partitioning.* The first most challenge in the domain of community detection is that there is no generally accepted definition of a community; still there are a large number of community detection algorithms available which produce effective results. Most of the community detection algorithms do not take in to account the *overlapping* between communities, which is a serious case in SNSs. Communities in social networks, tends to *overlap* with each other which means that a vertex which is a member of one community can also be a member of another community as shown in Fig 1. The idea of overlapping communities makes the problem of community detection tougher as the result of the algorithms would now be a *Cover*, a set of communities of which a vertex is a member. Most of the community detection algorithms start resulting in bad assignments of communities to vertices in the overlapping case as they generally merge two communities with dense overlaps into a single community.

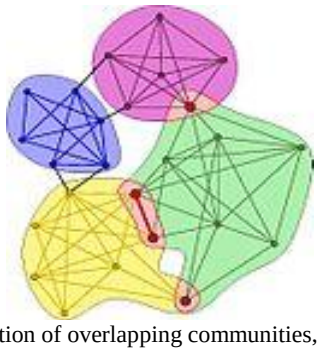

Fig 1: Illustration of overlapping communities, nodes shown in red color

This paper presents a systematic and organized study of overlapping community detection techniques. 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.
- To provide them with the existing state of art algorithms for the research problem.
- To make them aware of the challenges in the direction and the solutions proposed so far.

The paper is organized in to sections namely *Problem Formulation* which clearly states the problem of overlapping community detection, *Techniques* which explains the techniques used so far to solve the problem and their corresponding strengths and weaknesses, and at last Conclusion which sums up the work conducted and the future directions for work.

II. PROBLEM FORMULATION

Given a graph G (V , E), where V is the set of Vertices and E is the set of edges assign to each vertex $v \in V$ a cover C where C represents the set of communities of which v is a member.

$|V|$ =n, $|E|$ =m

For dense graphs m= O (n^2) and for sparse graphs m=O (n).

III. TECHNIQUES

The algorithms for overlapping community detection can be broadly classified into following categories:

a) Link Partitioning Algorithms
b) Clique Based Algorithms
c) Agent Based and Dynamic Algorithms
d) Fuzzy Algorithms
e) Local Expansion and Optimization Algorithms

We will explore each of them one by one, and will point out their relative strengths and weaknesses.

a) Link Partitioning Algorithms

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.

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 in [1] on the basis of edge similarity. If we are given a pair of links e_{ik} and e_{jk} , the edge similarity between these two links is calculated by Jaccard index as:

$$S(e_{jk}, e_{ik}) = |N_i \cap N_j| \,/\, |N_i \cup N_j|$$

N_i is the set of nodes which are in the neighborhood of node i including node i. After calculating edge similarities linkage clustering is done to find hierarchical communities. Generally single linkage hierarchical clustering is done because of its simplicity and efficiency which enables us to apply it on large networks. Other clustering methods include average and complete hierarchical clustering. Initially every node belongs to its own community, and then links with highest similarities are merged into a single community, this process is repeated until all the links belong to a single community. This whole merging process is stored in a *Dendrogram* which records the hierarchical community organization.

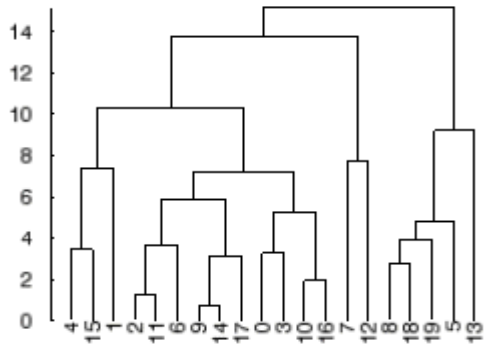

Fig 2: Illustration of a Dendrogram; y axis represents similarity (linkage distance) and x axis denotes node indices.

This Dendrogram is then cut at a threshold value of *partition density* to reveal communities as shown in Fig 2. Partition density attains a global maximum at some level in the Dendrogram; it is average at the top of Dendrogram and attains the lowest value at the leaves of the Dendrogram. The idea of link partitioning is quite natural and intuitive but it can't guarantee better results than node partitioning algorithms as it is also based on the ambiguous definition of community [2]. The complexity is O (nk^2_{max}), where k_{max} is the maximum degree of any node in the network.

b) Clique Based Algorithms

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

Step 1: Find all cliques of size k in the given network.

Step 2: Construct a clique graph. Any two k-cliques are adjacent if they share k-1 nodes.

Step 3: Each connected components in the clique graph form a community.

CPM (Clique Percolation Method) is based on the assumption that a network is composed of cliques which overlap with each other. CPM finds overlapping

communities by searching for adjacent cliques. As a vertex can be a member of more than one clique, so overlap is possible between communities. The parameter k is of utmost importance in finding communities via CPM, Empirically small values of k have shown effective results [3] [4]. An efficient implementation of the CPM method is *CFinder*. CPM is suitable for dense graphs where cliques are present. In case there are few cliques only CPM fails to produce meaningful covers.

Example 1: showing the working of clique percolation as shown in Fig 3.

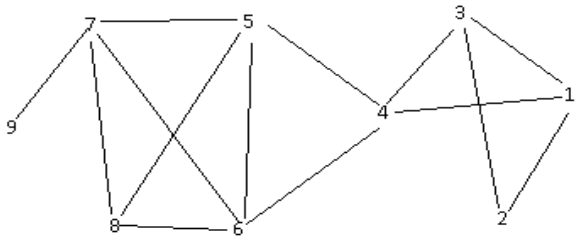

Fig 3: A network with 6 k-cliques where k=3

Clearly the network has 6 cliques of size 3, first of all these are identified by the CPM as {1, 2, 3}, {1,3,4}, {4,5,6}, {5,6,7}, {5,6,8}, {6,7,8}. After finding the cliques, a clique graph is formed in which two cliques are adjacent if they share k-1 nodes as shown in Fig 4.

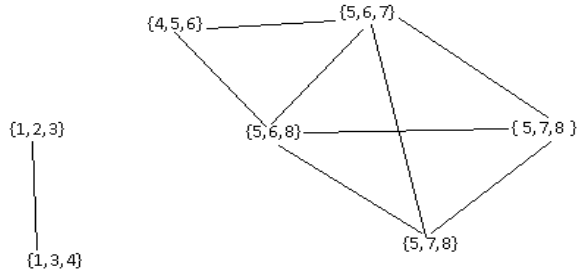

Fig 4: showing the clique graph for the network in Fig 3

Two communities will be shown as result and they are: {1,2,3,4} and {4,5,6,7,8}

The major disadvantage of CPM is that it fails to terminate for large networks. Some argue that CPM is more like a pattern matching algorithm as it searches for a particular pattern in the given network on the basis of input parameter k. Two other Clique based algorithms are CPMw [5] and SCPM[6]. CPMw is for weighted networks, it introduces the concept of intensity threshold. Only the cliques with subgraph intensity greater than the threshold are included in the community. CPMw produces communities with smoother contours as compared to CPM. SCPM is faster than CPM as it doesn't process the network for all values of k, instead it processes only for a given value of k. SCPM suits for weighted networks having hierarchical community structure.

c) *Agent Based and Dynamic Algorithms*

Three famous algorithms that come under this category are SLPA [7], COPRA [8], and Label Propagation algorithm [9].

SLPA is Speaker-Listener label propagation algorithm, 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 time complexity of SLPA is O(tm) where t is the number of iterations and m is number of edges. The best part of SLPA is that it doesn't require any prior knowledge 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. Whenever a node observes more labels in surrounding, it is more likely that it will spread those labels to other nodes. The Label Propagation algorithm described in [9] is extended to overlapping case by allowing a node to have multiple labels. 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. LPA was modified by Gregory [8], he introduced Community Overlap Propagation Algorithm (COPRA). 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. A node updates its belonging coefficient in a synchronous fashion by averaging the belonging coefficients of all its neighbors at each step. The parameter v controls the number of communities of which a node can be a member. The time complexity of COPRA is O (vmlog(vm/n)). According to benchmarks in [10] COPRA provides the best results for overlapping communities. An important optimization in LPA can be to avoid unnecessary updates, which will reduce the execution time.

d) Fuzzy Algorithms

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 b_{nc} 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.

$$\sum_{C} b_{nc} = 1 \text{ where } c \in C$$

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. In [10], authors proposed an algorithm with the combination of spectral clustering, fuzzy c means and optimization of a quality function. They propose a method to detect up to k communities by using the fuzzy c means clustering algorithm after converting the input network to a k- dimensional Euclidean space. The accuracy and computational efficiency of the algorithm is heavily dependent upon the parameter k. Nepsuz [11] model the problem of overlapping community detection as an optimization problem constrained nonlinearly which can be solved by simulated annealing methods. The objective function required to be minimized in this method is:

$$f = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(\ddot{x}_{ij} - x_{ij})^2$$

Where w_{ij} denotes a predefined weight and $\ddot{x}_{ij}$ denotes prior similarity between nodes i and j. Similarity x_{ij} is defined as:

$$x_{ij} = \sum_{c} a_{ic} a_{jc}$$

Where a_{ic} is the fuzzy membership of node i in community c, subjected to constraints of total membership and non empty community. To determine the value of k, it is increased repeatedly until the value of community structure doesn't improve as measured by modified fuzzy modularity function, Q defined as:

$$Q = \frac{1}{2m} \sum_{c} \sum_{i,j \in c} [A_{ij} - \frac{k_i k_j}{2m}] a_{ic} a_{jc}$$

A hybrid approach [18] based on Bayesian Non-negative Matrix Factorization to achieve soft partitioning of the network in computationally effective manner is proposed. The advantage of this approach is that it doesn't suffer from the problem of resolution limit. This approach is a mix of dimensionality reduction and feature extraction in machine learning. The problem with NMF approach is that it is computationally inefficient because of large matrix multiplications.

e) Local Expansion and Optimization Algorithms

These algorithms rely on a local benefit function which encodes the quality of densely connected subgraphs. The goal of these algorithms is to expand partial or natural communities so as to maximize the local benefit function. The quality of discovered communities heavily depends upon the quality of seed communities. A clique serves as a better seed than a single node. EAGLE [12] creates a Dendrogram by using agglomerative framework. First of all maximal cliques are identified, and initialized as seed communities, then similarity values are computed and communities with maximum similarity are merged. The optimal cut of the Dendrogram is calculated using extended modularity function defined in [13]. EAGLE is computationally expensive even without taking into account the time required for finding maximal cliques. GCE [14] also uses cliques as seed communities and expands them using a local fitness function. Communities are merged if they are found similar to previously detected communities. The similarity is computed using the distance function defined as:

$$1 - \frac{|c_1 \bigcap c_2|}{\min(|c_1|, |c_2|)}$$

If the distance is smaller than the value specified by parameter $\in$ then communities' c1 and c2 are merged. The time complexity of GCE is O (mh) where m is number of edges and h is the number of cliques. In [15], author proposed another two step technique in which nodes are first of all ranked according to some criterion, and then highly ranked nodes are removed until small disjoint cluster cores are formed. In the second step i.e Iterative Scan (IS) these cores act as seed communities which are expanded by adding or removing nodes until a local density function cannot be improved further. The density function used is given as:

$$f(c) = \frac{w_{in}{}^{c}}{w_{in}{}^{c} + w_{out}{}^{c}}$$

$w_{in}{}^{c}$ is total internal weight and $w_{out}{}^{c}$ is the total external weight of community c. The worst case complexity of this technique is $O(n^2)$. IS sometimes results in disconnected components as the algorithm allows removal of nodes during expansion, so CIS [16] was introduced in which connectedness is checked after each iteration. OSLOM [17] works by comparing the statistical significance of a cluster with the global null model (i.e. the random graph configuration model) during the expansion phase. To grow the community r value is computed for each neighbor, which is the cumulative probability of having more internal edges in the community than the number of edges from neighbors in the null model. If the cumulative distribution of smallest r value is lesser than a tolerance value, the node is considered significant and is added to the community otherwise second smallest r value is considered. The average time complexity is $O(n^{2)}$, it is dependent upon the underlying community structure of the input network. The main problem with OSLOM is

that it results in significant number of *singleton communities* or *outliers*. To detect both static and temporal communities iLCD [18] intrinsic longitudinal community detection was proposed which updates communities by adding nodes depending upon whether the number of their first, second robust neighbors is greater than an expected value or not. The algorithm depends upon two parameters one for adding nodes to the community and another for merging two communities. Recently there have been many improvements in the local optimization and expansion algorithms.

IV. CONCLUSION

Overlapping 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 network has emerged as a topic of growing interest as it shows the interplay between the structures of the network and its functioning. This paper tries to review all popular algorithms for overlapping community detection with their strengths and weaknesses. We have tried our 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.

REFERENCES

[1] Ahn, Y.-Y., Bagrow, J. P., and S. Lehmann, "Link communities reveal multiscale complexity in networks", *Nature -466, 2010, 761–764.*

[2] S.Fortunato, "Community detection in graphs", *Physics Report, 486(3), 2010, 75–174.*

[3] Palla, G., Derenyi,Farkas, I., and Vicsek, T., "Uncovering the overlapping community structure of complex networks in nature and society", *Nature-435, 2005, 814–818.*

[4] Lancichinetti, A., Fortunato, S., and Kertesz, J., "Detecting the overlapping and hierarchical community structure of complex networks", *New Journal Physics, Cornell University, U.S.A., 11, 2009, 33-41.*

[5] Farkas, I., Abel,D., Palla, G., and Vicsek, T. "Weighted network modules", *New Journal Physics, Cornell University, U.S.A., 9(6), 2007, 180-185*.

[6] Kumpula, J. M., Kivela "Sequential algorithm for fast clique percolation". *Physics Review, Cornell University, U.S.A., 78(2), 1-8*.

[7] Xie, J., Szymanski, B. K., and Liu, X., "SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process", *Proceeding of ICDM Workshop, IEEE, 2011, 344–349.*

[8] S.Gregory, "Finding overlapping communities in networks by label propagation", *New Journal Physics, Cornell University, U.S.A., 12, 2010, 1-13.*

[9] Raghavan, U. N., Albert, R., and Kumara, S., "Near linear time algorithm to detect community structures in large-scale networks". *Physics Review, Cornell University, U.S.A., 76, 2007, 33-42*.

[10] Zhang S., Wang R.S., Zhang X.S, "Identification of overlapping community structure in complex networks using fuzzy c-means clustering", *Elsevier, 374(1),2007, 483-490.*

[11] Nepusz Y., Petrocz A., Negyessy L., and Bazso F,Fuzzy , "Communities and the concept of bridgeness in complex networks", *Physics Review, Cornell University, 77, 2008, 1-13*.

[12] Shen, H., Cheng, X., Cai, K., and Hu, M.-B. "Detect overlapping and hierarchical community structure", *Physica A, 388, 2009, 1-7.*

[13] Shen, H., Cheng, X., and Guo, J., "Quantifying and identifying the overlapping community structure in networks", *J. Stat. Mech., 07, 2009, 9-16.*

[14] Lee, C., Reid, F., McDaid, A., and Hurley, N., "Detecting highly overlapping community structure by greedy clique expansion", *Proceeding of SNAKDD Workshop, 2010, 33–42*.

[15] Baumes, J., Goldberg, M., Krishnamoorthy, M., Magdon-Ismail, M., and Preston, N., "Finding communities by clustering a graph into overlapping subgraphs", *Proceeding of IADIS, 5(1), 2005, 97–104*.

[16] S.Kelley, "The existence and discovery of overlapping communities in large-scale networks", *Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, NY, 2009*.

[17] Cazabet R., Amblard F., and Hanachi C, "Detection of overlapping communities in dynamical social networks", *Proceedings of SOCIALCOM, 2010, 309–314.*

[18] Psorakis I., Roberts S. and Ebden Mark, "Overlapping community detection using Bayesian non-negative matrix factorization", *Physics Review E, 83, 2011, 1-9*.