Overview Of Various Overlapping Community Detection Approaches

Pooja Chaturvedi
Amity School of Engineering and Technology
Amity University, Lucknow
chaturvedi.pooja03@gmail.com

***Abstract*—With the advancement in technology, we are surrounded with huge amount of data, which need to be processed in some manner for further processing. For this purpose, we may combine the Data Mining Techniques with the concepts of Graph Theory. Real world systems may be represented as Graphs. For example: Social Networks may be depicted as graph with the people as nodes and their relationships represented as links. Community Structure has been found as an important property of these systems. In simple words we can define community structure as a connected sub-graph which is tightly connected within the subgroups and weakly connected within the subgroups. Various algorithms have been proposed till now to identify and detect these community structures. In this paper we will provide the overview of some approaches which identify the overlapping communities i.e. communities that may belong to more than one subgroups.**

Keywords-Complex Networks,Community Structure,Overlapping Community.

I. INTRODUCTION

Modern data processing and analysis tools have to often face the challenges of complex inter-relationships within the data. Identification and detection of community structure finds its applications in various fields such as in web-sites for the analysis of hyperlinks associated to a web-page, solving Biological queries resulting from the functional studies of metabolic and protein networks, technological problems resulting from large infrastructures etc. [1]

Real world systems depicts properties which are completely different than random graphs as proposed by Erdos and Reiny.[2] Many definitions of community exists in literature.[3]

Definition of Community in a strong sense: The sub-graph V is a community in a strong sense if $k_i^{in}(V) > k_i^{out}(V), \forall i \in V$.In strong sense each node has more connections within the community than with the rest of the graph.

Definition of a community in a weak sense: The sub-graph V is a community in a weak sense if $\sum_{i \epsilon V} k_i^{in}$ (V) > $\sum_{i \epsilon V} k_i^{out}$(V).In a weak community the sum of all degrees within V is larger than the sum of all degrees toward the rest of the network.

Community structure in various networks was studied by Girvan and Newman .[4]

II. DISJOINT COMMUNITY DETECTION APPROACHES

Community detection problem can be viewed as an extension to graph partitioning problem. The objective of graph partitioning problem is to divide the given graph into specific number of groups of predefined size. To obtain more than two partitions Minimal Bisection Method is adopted. Kernighan Li[5] is one of the most popular methods for graph partitioning. This method was based on the idea of the optimization of a benefit function Q which is calculated as the difference of the edges within the groups and between the groups. The problem with graph partitioning methods was that in these methods the number of partitions to be done should be known prior. This method was reasonably fast with the speed as O($n^2 \log n$).Another method was Spectral Bisection method which was based on the properties of the spectrum of the Laplacian Matrix. Other methods of graph partitioning can be found in [6].For graph partitioning cuts and normalized cuts have also been used which divides the graph in only two partitions.

The traditional method of graph clustering is Hierarchical clustering. Girvan and Newman proposed a method based on edge betweenness. [7] It was able to divide the graph in more than two clusters. This method was based on the fact of removing the edge with the maximum edge betweenness between the two edges. This process is repeated until all the edges have been processed. The methods proposed till now were unable to determine the quality of the partitions obtained. So Girvan and Newman proposed a new variable Modularity which was calculatedas follows:

$$Q = \frac{1}{2m} \sum_{vw} [A_{vw} - \frac{k_v k_w}{2m}] \delta(c_v, c_w)$$

Clauset, Newman and Moore [8] used this parameter in their algorithm for community detection which is commonly known as Fast Modularity. In this method a sparse matrix ΔQ_{ij} was maintained which stored the gain in modularity for each edge. In the next step we find the element with maximum modularity and the corresponding communities were merged. This process is iterated until all the edges have been visited. This method had a resolution limit problem i.e. it was unable to detect the communities smaller than a specified threshold. To overcome this problem Blondel [9] introduced a new algorithm which was able to detect the high modularity partitions in quick time. This algorithm was efficient for large networks consisting of millions of nodes too. Other community detection methods were Simulated Annealing[10],Spin Model[11],Random walks[12],Walk Trap[13],Markov Chain clustering[14],Synchronization[15], Stastical Interference[16], Bayseian Interference[17], Block Modeling [18],Label Propogation [19] etc.

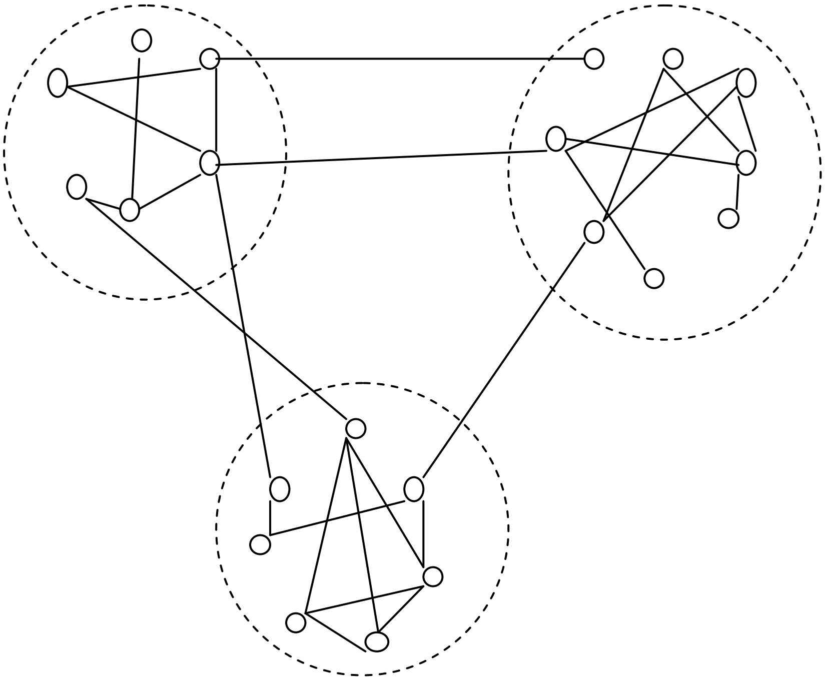

Fig 1: A network having community structure. This network has three communities which have dense connections within the community and sparser connections between the communities.

III. OVERLAPPING COMMUNITY DETECTION APPROACHES

The methods discussed above were able to detect the clusters in which nodes belonged to only a single group. But in real world scenarios the nodes may overlap i.e. they may belong to more than one cluster. Detecting these overlapping communities is still a challenging task for researchers and has gained a lot of attention in recent studies. In this section we will provide a brief overview of some of the approaches developed till now.

Clique Percolation Method: The most popular method for the overlapping community detection is CPM which was proposed by Pella et.al.[20].This method was based on the idea of using a template in the form of a connected sub-graph (commonly known as clique).This template was moved over the original network to detect the overlapping communities. Clique was defined as that connected sub-graph in which there is a link from a vertex to every other vertex. There are various algorithms available to find clique. For example: Born-Kerbosch algorithm.The condition for two cliques to be adjacent is that they should share atleast k-1 vertices. K-clique community is obtained by rolling over the adjacent k-cliques. Communities can easily be found and analyzed by using the software CFinder which was based on the implementation of CPM.

Sequential CPM[21]: This method is based on the idea of finding k-clique communities by inserting edges one by one starting from an empty graph. Whenever an edge is inserted we check for the formation of k-cliques by searching for k-2/-cliques in the subset of neighboring vertices of the endpoints of the inserted edge. The procedure requires building a graph in which the vertices are .k-1/-cliques and edges are set between vertices corresponding to k-1/-cliques which are sub-graphs of the same k-clique. At the end of the process, the connected components of graph correspond to the searched k-clique communities. It is considerably faster than CPM. It has the main advantage that it could be applied for the weighted graphs as well. The drawback of this method is that it assumes that original network has a large number of cliques, so it is unable to find the original community structure of the networks.

Agglomerative Hierarchical Clustering Based on Maximal Cliques (EAGLE) Algorithm[22]: This method is able to detect overlapping and hierarchical communities in the networks. EAGLE is an agglomerative hierarchical clustering method, which consists of two stages. In the first stage the dendogram is constructed and in the second stage the point on which the cut is to be made is determined. First stage process is defined as:

1. In the first step identify the maximum clique in the network and delete the sub-ordinate cliques. Maximum clique refers to the clique which is not the subset of any other clique and sub-ordinate clique refers to those cliques which are not maximal cliques. The remainder vertices constitute the initial communities. The similarity measure between the each pair of communities is obtained.
2. The pair of communities with maximum similarity is identified and is merged to obtain a new community.
3. Similar process is repeated until one community remains.
   The similarity between the communities is determined as:

$$\mathrm{M} = \frac{1}{2m} \sum\nolimits_{v\epsilon C_1, w\epsilon C_{2,v\neq w}} [A_{vw} - \frac{k_v k_w}{2m}]$$

Where A_{vw} is the element of the adjacency matrix. It takes value 1 if there is an edge between the vertex v and w and 0 otherwise. m= $\frac{1}{2}\sum_{vw} A_{vw}$.

Community Betweenness Algorithm: Steve[23] introduced an index named split betweenness,on the basis of which he proposed two algorithms—Peacock algorithm and CONGO algorithm. **CONGO algorithm** directly splits Vertices with high split betweenness in a specific way on the whole network to obtain overlapping Communities. The number of communities in the final result is provided by users. **Peacock** algorithm has two phases. In the first phase, a network is transformed to a new one by splitting vertices, based on split betweenness index. In the second phase, the transformed network is processed by a disjoint Community detection algorithm. Although both these algorithms were able to determine the meaningful overlapping communities, but the major drawback of these methods was its high computational complexity. To remove this drawback, a new index local betweenness proposed by Steve. Steve limits the lengths of the shortest paths to a threshold h. In-corporation of this betweenness index made these algorithms fast enough. But now the major concern was the quality of communities detected.

To remove the shortcoming of these methods a new index as Community Betweenness was introduced. This method was based on the idea of finding the initial community by using any of CDA's (either BGLL or Infomap). Community betweenness is calculated on a new small network SG (SV, SE), where SG contains all the vertices and edges in a pair of communities. In CBS algorithm, when a vertex is identified as an overlapping vertex or an edge is identified as a bridge-edge, we don't need to recalculate the community betweenness of left edges as GN algorithm, because the main community structures of the small network are known. Even if some bridge-edges don't have high community betweenness, they will still be processed later.

Greedy Randomized Adaptive Search Procedure (GRASP): GRASP[24] is an iterative process, which consists of two phases:
1. Construction Phase: In this feasible solutions are constructed iteratively one at a time. A restricted candidate list (RCL) is constructed with respect to a greedy function g: C-> R, which is ordered and elements are chosen randomly and added to the solution one element at a time. The randomness and greediness of the process is determined by randomly choosing a parameter ∝.
2. Local Search Phase: This phase consists of replacing the current solution with the optimal solution among the current solution and in its neighborhood. A solution is said to be as optimal solution, if there is no better solution exists in its neighborhood solutions.
GRASP procedure is used in finding the maximal quasi-cliques. The obtained cliques are used in the exploration of remaining portions of the graph.
Community detection based using the fitness function:[25] This method is based on the assumption that the communities are local structures, which comprise of the nodes of the modules themselves and the extension to the nodes in its neighborhood. This method identifies communities as subgraphs obtained by the maximaization of a fitness measure, which can be easily determined as:

$$f_G = \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^{\alpha}},$$

Where k_{in}^{G} and k_{out}^{G} are the total internal and external degrees of the nodes of the module G.α is the positive real valued parameter which controls the size of the communities.Internal degree of a module is twice the number of internal links of the module.External degree is defined as the number of the links from each node in the module to the rest of the network.We obtain initial subgraph having node A.We than calculate the fitness of each node and neighboring node of A.The nodes having the highest fitness is addded to the subgraph G.This process is repeated until all the nodes have been processsed or we obtain negative fitness.

Community Detection via local algorithm[26]: This method consists of two phases:

1. Finding the initial community.
2. Expanding the community.

Nodes in the finding communities are labeled as T, denoted by VT,$V_F = V - V_T$.All nodes are labeled as F initially. Nodes are ordered with the increasing values of node strengths. Nodes with the maximum node strength alongwith its neighborhood nodes constitute the initial community. For expanding the community we calculate the belonging degree for each node. If the belonging degree lies between the given threshold, it is added to the community. The node strength is determined as following:

$$k_u=\sum_{v\epsilon V} w_{uv}$$

The belonging degree is computed as follows:

$$B_{(u,C)}=\frac{\sum_{v\epsilon V} w_{uv}}{k_u}$$

To qualify the communities obtained following modularity measure is used:

$$Q_0=\frac{1}{2m}\sum_{c\epsilon C}\sum_{u,v\epsilon V}\alpha_{cu}\,\alpha_{cv}(A_{uv}\text{-}\frac{k_u k_v}{2m}).$$

Where $\alpha_{cu}=\frac{k_{cu}}{\sum_{c\epsilon C} k_{cu}}$ and $\alpha_{cv}=\sum_{v\epsilon C} w_{uv}$.

The constraint applied here are:

$$0 \leq \alpha_{cu} \leq 1, \forall c\epsilon C, u\epsilon V$$

$$\sum_{c\epsilon C}\alpha_{cu}=1$$

The computation complexity of this method in finding all the communities is $O(n^2)$.

Density based shrinkage algorithm for hierarchical and overlapping community detection (DENSHRINK):[27] This method consists of two phases:

1. In the initial phase each node belongs to its own community i.e. there are as many communities as the number of nodes. Then we find the initial micro community for each node on the basis of the structural similarity, which is calculated as following:

$$\sigma(u,v) = \frac{\sum_{x\epsilon\tau(u)\cap\tau(v)} w(u,x).\,w(v,x)}{\sqrt{\sum_{x\epsilon\tau(u)} w^2\,(u,x)}.\sqrt{\sum_{x\epsilon\tau(v)} w^2(v,x)}}$$

2. In the second phase communities found in the first phase are shrinked using the modularity gain.The communities with the maximum modularity gain are merged and process is repeated until we get the positive modularity gain.

$$Q_s=\sum_{i=1}^{k}\frac{IS_i}{TS}\text{-}(\frac{Ds_i}{TS})^2$$

Where k is the number of clusters, IS_i= $\sum_{u,v\epsilon C_i}\sigma$(u,v) is the total similarity of nodes within clusters C_i. $DS_i = \sum_{u\epsilon c_{i,v\epsilon V}} \sigma(u,v)$ is the total similarity between nodes in clusters C_i and any node in the network and TS=$\sum_{u,v\epsilon V}\sigma(u,v)$ is the total similarity between any two nodes in the network.

The features of this method are:

1. It is a parameter free method which not only detects overlapping hierarchical communities but also the hubs and outliers. The nodes which bridge the two communities are known as hubs and those nodes which are only marginally connected to the communities are known as outliers.
2. This method can detect communities with varying densities also, so it is efficient and effective.
3. It combines the advantages of density based methods with modularity optimization methods, so it overcomes the resolution limit problem as well.

CON (Condition for overlapping nodes) based community detection method:[28] This method is a two step process. In the first step we identify the overlapping nodes from the boundary node sets. Then we find the inner overlapping nodes. For two given communities C_i and C_j,$B_{ij} = \{v: v\epsilon C_{i,}\exists u \in C_j\}\ and$ (v, u)ϵE is the boundary node set of community i connecting to community j. Condition for overlapping nodes is imposed by:

$$p_{ij}=\frac{|N_{ij}^{no}(v)|}{N_{ii}^{no}(v)}$$

To find the overlapping nodes we check the condition for overlapping nodes for each node. We find the node with the maximum p_{ij}, If p_{ij}(v) is not less than φ,then the node v is added to the overlapping node set. In the next step we will determine whether some of the overlapping nodes have become non-overlapping nodes for the detection of new overlapping nodes. For this we find and delete the nodes with the minimum p_{ij}.This process is repeated until it is not less than the specified threshold. After the above procedure stops, CONA begins to find overlapping nodes from the inner node set of the two communities. This phase is very simple. Every inner node without non-overlapping neighbors will be inserted into the overlapping node set.

Conclusion and Future Scope: Real world systems depicts several properties which are highly inhomogeneous than those of random graphs. Complex systems have properties such as Power-law degree distribution, scale free property, small world effect etc. In this manuscript we have discussed some of the methods to detect overlapping community detection. The study of community structure in these systems plays a significant role as they provide the mesoscopic description of the graphs where the communities play an important role in the topology rather than the edges and vertices. Community structure may provide the facility to classify the vertices according to their roles in the network, which can aid in the study of individual property of the vertices. In real graphs, the topological roles can be related to functions of vertices: in metabolic networks, for instance, connector hubs, which share most edges with vertices of other clusters than their own, are often metabolites which are more conserved across species than other metabolites, i.e. they have an evolutionary advantage. If communities are overlapping, one can explore other statistical properties, like the distributions of the overlaps and of the vertex memberships. The overlap is defined as the number of vertices shared by each pair of overlapping clusters; the membership of a vertex is the number of communities including the vertex. Both distributions turn out to be skewed, so there seem to be no characteristic values for the overlap and the membership. Moreover, one could derive a network, where the communities are the vertices and pairs of vertices are connected if their corresponding communities overlap. Such networks seem to have some special properties.

Acknowledgment

I would like to thank Mr. Deepak Arora, Assistant Professor, Amity University for his valuable support and guidance throughout this work.Special thanks to Mrs. Mousumi Dhara,Research scholar,IT-BHU for her consistent guidance and suggestions without which this work could haven’t be done.

References

[1] Filippo Radicchi,Claudio Castellano,Federico Cecconi,Vittorio Loreto,and Domenico Parisi,Defining and identifing communities in netwroks,PNAS. March 2,2004

[2] Erdos, P.; Rényi, A. (1959). "On Random Graphs. I".

[3] Wasserman, S. & Faust, K. (1994) Social Network Analysis (Cambridge Univ. Press, Cambridge, U.K.).

[4] M.E.J.Newman and M.Girvan,Finding and evaluating community structure in networks,2003.

[5] B.W. Kernighan, S. Lin, An efficient heuristic procedure for partitioning graphs, Bell Syst. Tech. J. 49 (1970) 291-307.

[6] E.R. Barnes, An algorithm for partitioning the nodes of a graph, SIAM J. Algebr. Discrete Methods 3 (1982) 541-550.

Pothen, Graph partitioning algorithms with applications to scientific computing, Tech. rep., Norfolk, VA, USA, 1997.

[7] M.Girvan and M.E.J.Newman,Community structure in social and biological netwroks,April 2001.

[8] Aaron Clauset,M.E.J.Newman,and Cristopher Moore,Finding community structure in very large networks,Phys. Rev. E 70 (6) (2004) 066111.

[9] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. P10008 (2008).

[10] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (1983) 671-680.

[11] F.Y. Wu, The Potts model, Rev. Mod. Phys. 54 (1) (1982) 235-268.

[12] B.D. Hughes, Random Walks and Random Environments: Random Walks, vol. 1, Clarendon Press, Oxford, UK, 1995.

[13] M. Latapy, P. Pons, Lect. Notes Comput. Sci. 3733 (2005) 284-293.

[14] S. van Dongen, Graph clustering by flow simulation, Ph.D. Thesis, Dutch National Research Institute for Mathematics and Computer Science,University of Utrecht, Netherlands, 2000.

[15] A. Pikovsky, M.G. Rosenblum, J. Kurths, Synchronization : A Universal Concept in Nonlinear Sciences, Cambridge University Press, Cambridge, UK,2001.

[16] D.J.C. Mackay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, Cambridge, UK, 2003.

[17] R.L. Winkler, Introduction to Bayesian Inference and Decision, Probabilistic Publishing, Gainesville, USA, 2003.

[18] P. Doreian, V. Batagelj, A. Ferligoj, Generalized Blockmodeling, Cambridge University Press, New York, USA, 2005.

[19] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E 76 (3) (2007) 036106.

[20] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814-818.

[21] J.M. Kumpula, M. Kivelä, K. Kaski, J. Saramäki, Sequential algorithm for fast clique percolation, Phys. Rev. E 78 (2) (2008) 026109.

[22] Huawei Shen,Xueqi Cheng,Kai Cai,Mao-Bin Hu,Detect overlapping and hierarchical community structure in networks,Physics A 388(2009) 1706-1712

[23] Zhihao WU,Youfabg LIN,Huaiyu WAN,Shengfeng TIAN,An efficient methiod to find overlapping communities in networks,School of Computer and Information Technology,Beijing Jiaotong University,Beijing 100044,China

[24] James Abello,Mauricio G.C.Resende and Sandra Sudarsky,Massive Quasi Clique Detection

[25] Andrea Lancichinetti,Santo Fortunato and Janos Kertesz,Detecting the overlapping and hierarchical community structure in complex netwroks,NJ Physics 11(2009) 033015.

[26] Duanbing Chen,Mingsheng Shang,Zehua Lv,Yan Fu,Detecting overlapping communities of weighted networks via a local algorithm,Physics A 389 (2010) 4177-4187.

[27] Jianbin Huang,Heli Sun,Jiawei Hann ,Boqi Feng,Density based shrinkage for revealing hierarchical and overlapping community structure in networks,Physics A 390(2011) 2160-2171.

[28] Zhihao Wu,Youfang Lin,Huaiyu Wan,Shengfeng Tian,Keyun Hu,Efficient overlapping community detection in huge real-world networks,Physics A 13587,2011.

[29] Santo Fortunato,Community Detection in Graphs,Physics Reports 486(2010)75-174

AUTHORS PROFILE

Pooja Chaturvedi received the B.Tech. degree in Computer Science & Engineering from Ideal Institute of Technology,Ghaziabad(U.P. Technical University,Lucknow) India in 2010 and completed her M.Tech. in Computer Science and Engineering from Amity School of Engineering and Technology,Amity University,Lucknow in 2012