A Survey on Overlapping Community Detection

P.Lakshmi Bhavani[1], C.Harshitha[2], R.Sandeep Kumar[3]
[1, 2, 3] Department of Computer Science and Engineering
[1, 2, 3] G.Pullaiah College of Engineering and Technology, Kurnool.

Abstract- *Network Communities represents basic structures for understanding the organization of real world networks .A Community is defined as group of nodes those have more connections among their members than between members and remainder of the network. Overlapping of Communities in network has been occurred as nodes belongs to multiple networks clusters at once.*

In this paper ,we discussed the brief description of community and its detection. And also describes the algorithms that detect the overlapping community with their merits.

Keywords- Community detection, Overlapping Community detection

I. INTRODUCTION

Community:

Community is a group of nodes that have some common properties and have common role in organization. Group of nodes are more densely connected if they belongs to the same community and less likely to be connected if they are not the members of same community. [1]

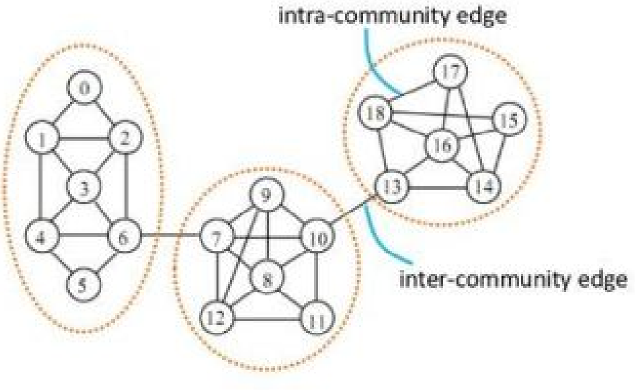


Figure 1. Community

Community Structure:

Community structure is also named as cover of community. Community structure is a set of communities present in network. It is represented as F={f1,f2,f3,f4,....fk}. Here F is the community structure and f1,f2,f3,...fk are communities. For example there are two communities f1={1,2,3} and f2={3,4,5}. Thus community structure is F={f1,f2}.

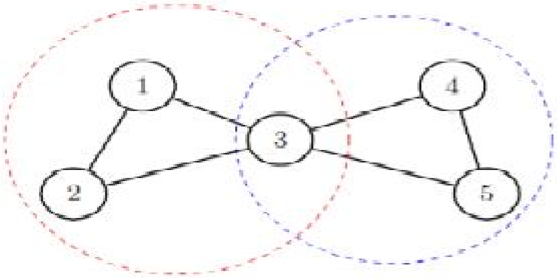


Figure 2. community structure having communities c1 and c2.

Types of communities:

Community can be of two types:

- Disjoint Community:

In disjoint community a node belongs to single community. Disjoint community is also known as crisp assignment, where binary relationship is being held between a node and a community. A node can belong to at-most 1 community and atleast 0 community[3].

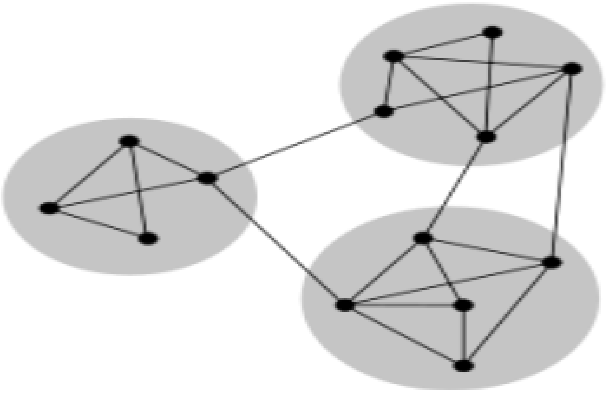


Figure 3. Disjoint communities.

- **Overlapping Community:**

In Overlapping community a node may belongs to more than one community [4]. This is known as fuzzy assignment of nodes where, a node may belong to more than one community.

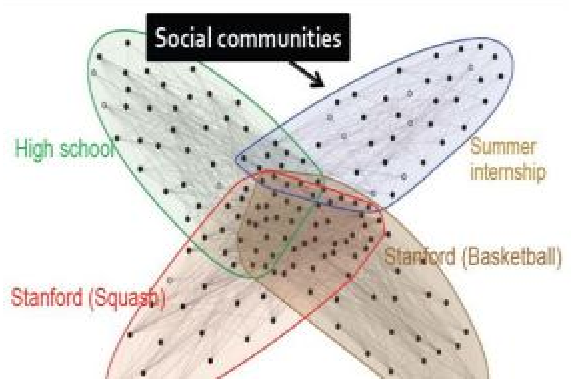


Figure 4. Overlapping Community

II. RELATED WORK

Community Detection:

According to M. E.J. Newman, Anatol Rapoport Distinguished University Professor of Physics at the university of Michigan, “Loosely stated, [community detection] is the problem of finding the natural divisions of a network into groups of vertices [called communities or clusters] such that there are many edges within groups and few edges between groups. What exactly we mean by “many” or “few”, however, is debatable, and a wide variety of different definitions have been proposed, leading to a correspondingly wide variety of different algorithms for community detection[1].

Community detection-Parameters:

- Propinquity measure: For a subset M of the graph G, propinquity measure gives nearness of the inner-connections.
- Revelatory Structure: Organize the graph elements and get the community structure from the intertwined connections among them.

Use of Community Detection:

“The most common use for community detection,” says Newman, “is as a tool for the analysis and understanding of network data. For instance, the community structure in social networks” can give us clues about the nature of the social interactions within the community represented”. Clusters of nodes in a web graph for instance might indicate groups of related web pages. Clusters of nodes in a metabolic network might indicate functional units within the network.

The use of social network analysis in political science, to study political participation and conservation, has made visible the polarization in U.S.politics.

Example: The Human Disease Network

In biology, networks are being used to gain insight into the mechanism that underlie disease. In his new book, Network Science, Albert-Laszlo Barabasi, the Robert Gray Dodge Professor of Network Science at Northeastern University, observes:”communities play a particulary important role in understanding human diseases. Indeed, proteins that are involved in the same disease tend to interact with each other.”

In the “Disease Gene Network” shown in fig, “each node is a gene, with two genes being connected if they are implicated in the same disorder. The number of disorders in which the gene is implicated is proportion to size of each node.Only nodes with at least one link are shown.”

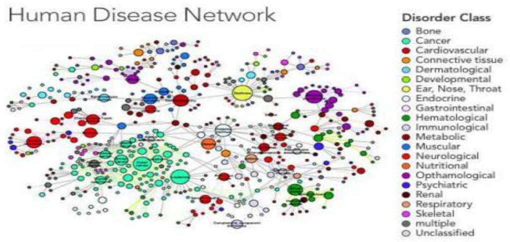


Figure 5. Communities of genes associated with the same disorder.

Community detection with hierarchical Clustering Algorithms:

A variety of community detection methods have been developed (e.g., clique-based, graph portioning, and modularity maximization detection method) in order to meet the assumptions of different definitions. It is important to note that the brute-force approach of finding all possible partitions of a network and choosing the one that best fits a particular definition of community is not feasible as it is an NP-hard problem. Using brute force to solve even the simplest community-finding problem, graph bisection, is NP-hard.

Community detection with Agglomerative Hierarchical clustering Algorithms

Agglomerative clustering is a clustering algorithm used to construct hierarchy of clusters from nodes of a network that is given. Every cluster is assigned by a node initially.The two nearest clusters are merged into same cluster.This process will be repeated until one cluster is left.

It is also called as Agglomerative Neting(AGNES).It is based on distance. To carry out this clustering, it requires two decisions:(1)distance function’s choice, d, for measuring the distance d(i,j) between two nodes, i and j in the network and(2)making the decision of how to extend distance function, d, to define the distance d(A,B) between any two clusters of nodes A and B.

Some of the commonly used ways to mesure the distance between two nodes iand j in undirected network as follows:

- Euclidean distance, di,j, is defined as:

$$d_{l,j}=\sum_{k=1}^{N}\left(A_{ik}-A_{kj}\right)^2$$

- Cosine-similarity measure,σ_{ij},which uses the dot product of vectors that are formed by rows i and j in the adjacency matrix.

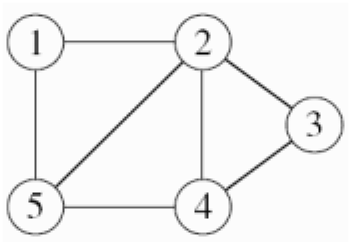
$$\sigma_{ij}=\frac{\sum_{k=1}^{N}A_{ik}A_{kj}}{\sqrt{\sum_{k=1}^{N}A^2_{ik}}\sqrt{\sum_{k=1}^{N}A^2_{jk}}}$$

- the standard Pearson correlation coefficient r_{ij},which uses rows i and j variance in adjacent matrix:

$$r_{ij}=\frac{cov(A_i,A_j)}{\sigma_i\sigma_j}=\frac{\sum_{k=1}^{N}(A_{ik}-<A_i>)(A_{jk}-<A_j>)}{\sqrt{\sum_{k=1}^{N}(A_{ik}-<A_i>)^2}\sqrt{\sum_{k=1}^{N}(A_{jk}-<A_j>)^2}}$$

Example:

Let us take an example of this graph



Now, we will look out single-linkage, hierarchial clustering using squared Euclidean distance formula $d_{ij}=k_i+k_j-2n_{ij}$,for measuring the distance between nodes i and j.

The degrees of five nodes are $k_1=2,k_2=4,k_3=2,k_4=3,k_5=3$

Next,we arrange those values n_{ij}(i.e., the number of neighbours of nodes i and j) are as follows:

$$[n_{ij}]=\begin{bmatrix}2&1&1&2&1\\1&4&1&2&2\\1&1&2&1&2\\2&2&1&3&1\\1&2&2&1&3\end{bmatrix}$$

Applying the formula $d_{ij}=k_i+k_j-2n_{ij}$,for all i and j,1<=i,j<=5,we can get the following matrix:

$$D=[d_{ij}]=\begin{bmatrix}0&4&2&1&3\\4&0&4&3&3\\2&4&0&3&1\\1&3&3&0&4\\3&3&1&4&0\end{bmatrix}.$$

And now we need to work out on hierarchical clustering:

i. At first, we have 5 clusters1,2,3,4,5. Distance matrix gives the distance between each pair of clusters. The minimum distance between clusters is 1.

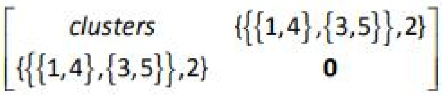
$$\left[\begin{array}{cccccc}clusters&1&2&3&4&5\\1&0&4&2&1&3\\2&4&0&4&3&3\\3&2&4&0&3&1\\4&1&3&3&0&4\\5&3&3&1&4&0\end{array}\right]$$

.

ii. Here, we need to merge cluster 1and cluster 4 in one cluster {1,4} at a distance of 1;we also merge cluster 3 and cluster 5 into another cluster{3,5}.The minimum distance between clusters is 2.

$$\left[\begin{array}{cccc}clusters&\{1,4\}&\{3,5\}&2\\\{1,4\}&0&2&3\\\{3,5\}&2&0&3\\2&3&3&0\end{array}\right]$$

iii. And later we merge cluster{{1,4},{3,5}}and cluster 2 into one cluster{{{1,4},{3,5}},2} at distance of 2.The final cluster contains all objects, thus agglomerative algorithm is concluded.



$$\begin{bmatrix} \textit{clusters} & \{\{\{1,4\},\{3,5\}\},2\} \\ \{\{\{1,4\},\{3,5\}\},2\} & \mathbf{0} \end{bmatrix}$$

Overlapping community Detection:

In present research scenario many researchers in their field of interest results in increasing number of interdisciplinary publications due to spreading of multiple research directions. So it is mandatory to detect overlapping communities for relevant publications.

Here we discuss some of the algorithms which detect overlapping communities like 'OverCite' and 'Online Community Detection Algorithm'.

OverCite [7]:

OverCite algorithm detects overlapping communities of papers, authors and venues with the help of publication hypergraph and the citation information simultaneously. OverCite follows three step procedures.

In First step, this algorithm converts the hypergraph H to its weighted line-graph H' where the hyperedges in H become nodes in H'. Nodes named as ei and ej in H' will be linked with non negative weight in terms with the help of three factors:

- Hypergraph Neighbourhood Similarity(HNS)
- Co-citation Strength(CCS)
- Bibilographic Coupling Strength(BCS)

The similarity between ei and ej is measured finally by combinig HNS, CCS and BCS linearly:

Similarity(ei,ej) = α.HNS + β.CCS + γ.BCS

If the weighted line graph H' is constructed from the hypergraph H, any community detection algorithm for weighted unipartie graph can be applied to cluster the nodes in H', which results in production of communities of hyperedges in H.[5]

Finally, the community structure decided in H', each hyperedge in H is assigned to a single community, which results in assigning multiple overlapping communities to nodes in H.

III. Metrics

We define the following metrics for measuring similarity between hyperedges, that capture citation information based similarity and hypergraph neighbourhood similarity to calculate the weight.

Hypergraph Neighbourhood Similarity (HNS):

Hypergraph Neighbourhood Similarity computes the relative overlap between common neighbours of end vertices of two hyperedges.

Co-citation Strength (CCS):

Co-citation Strength is measured by the number of times two papers are cited together in the subsequent literatures the higher the co-citation is, the more citations the two papers have in common.

The relative measure of co-citation strength of two hyperedges ei and ej is calculated by the ratio of actual and maximum citations received by two end-points in paper partition.

Bibilographic-coupling Strength (BCS):

BCS is defined as the number of common citations. It is another way of determining the similarity between related works of two papers.

Online community detection algorithm[6]:

A straight forward way to make online communityis to take sequence of edges as input,optimize modularity at each step for network based on partition.But this greedy algorithm shows poor performance. Brandes et.al[2] proved that a partition with maximum modularity doesnot have community that consists of single node having degree 1.

To avoid poor performance, our algorithm optimizes expected modularity for final network.

Our algorithm processes a network edge by edge in the order that the network is fed to the algorithm. It does not optimize modularity but expected modularity has been done to avoid poor performance.

Advantages:

The two major advantages of our algorithm are:

- The update uses knowledge about network's local structure relates to new edge.
- In constant time, updates can be done.
- The algorithm can be processed in large networks efficiently in real time.This algorithm has been applied to public real world large network data sets.

IV. CONCLUSION

In this paper, we discussed two algorithms on overlapping community detection namely 'OverCite' and 'Online Community Detection Algorithm'.In 'OverCite'algorithm,we construct a publication hypergraph that consists of authors,venues and publications in three partitions and converted into a weighted line graph by including the related information.Later communities are detected. And this 'online community detection algorithm' is used to reduce poor performance and optimize the modularity.

REFERENCES

[1] p.357 Networks: An Introduction, M.E.J Newman, Oxford University Press(2010)

[2] Brandes U,Delling D, Gaertler M, Gorke R, Hoefer M,et al (2008) On modularity clustering, IEEE Transactions on Knowledge and Data Engineering 20:172-188.

[3] S.Gregory, "Finding overlapping communities using disjoint community detection algorithms,in complex Networks", pp 47-61,Springer,2009.

[4] J. Xie, S.Kelley and B.K.Syzmanski ,"Overlapping community detection in networks: the state of the art and comparative study", arXiv preprint arXiv:1110.5813,2011.

[5] S. Fortunato and A. Lancichinetti, "Community detection algorithms: a comparative analysis: invited presentation, extended abstract", in Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools. ICST 2009.

[6] Tanmoy Chakraborty, Abhijnan Chakraborty, "OverCite: Finding Overlapping Communities in Citation Network",India,2013.

[7] Gang pan, Shijian Li, Zhaohui Wu, "Online Community Detection for Large Complex Networks",Spain,2014.