Affect of Varying the Similarity Threshold in Leader Similarity Based Community Detection

Er. Ridhi Bhatt[1] Er. Sunil Patel[2] Dr. Kumar Gaurav[3]
[1,2,3]Department of Electronics Engineering
[1,2,3]Harcourt Butler Technical University, India

Abstract— In the present scenario community detection is significant for extracting the community structure of a given network. It is very important in understanding and evaluating the formation of large and complex networks. It provides us the competency to look into group-level. There are numerous algorithms that have been put forward to study the community structure in a network and had been applied to various domains. They help in classifying the nodes on the basis of their functions in account of their positions in the communities. In this paper our focus is to study the sensitivity of the similarity threshold extending the leader similarity based community detection algorithm i.e. first of all leader is selected from the network using Eigen vector centrality than on the basis of the similarity threshold the followers are selected and we get the community in the network. We also analyzed how the variation of the similarity threshold helps us to evaluate the size of the community.

Keywords: Homophily, Similarity Threshold, Jaccard Similarity, Salton Similarity

I. INTRODUCTION

The theory of network is a dominant tool to model and interpret the complex networks in different disciplines. To extract community structure[2] in single layer network many theories have been developed but with the development of research, the scholars have realized that the only study of communities in a single layer network doesn't provide essential results in analyzing the structures and behavior in real life applications. As the complex networks in any field such as biology, economics, social sciences etc all consists of multiple type of relations among different entities. Community in a network can be defined as the set of nodes that can be classified into groups of nodes such that each group of nodes can be internally connected. The nodes correspond to the individuals, and the interactions among the individuals are represented by the edges. This interaction among the nodes can be explained with the Homophily i.e. the inclination of persons with alike ideas, behaviors, and fondness to get associated with each other that results in the formation of communities. Thus, with the help of community detection we can disclose the community structure of any network. It is used for studying biological network ,transport network that include community groups on the basis of using same route, same transport vehicle, same station etc which results in different layers of the transport system. It also helps in studying hidden relations and false relations among the nodes in any network. Analyzing communities is very helpful for studying information diffusion processes in the network like rumor spreading or epidemic spreading and to take steps to reduce it. It is often noticed that some links may not get observed or some links may enter into the data falsely. These types of problems are coped by community detection algorithm.

II. LITERATURE SURVEY

The study of community structure in the different types of network has been an important point of discussion and research. As a result a large number researcher has been contributing in this field and papers are being published so far.

In 1736 the problem of "Seven Bridges of Koinsberg" had completely revolutionized the research in the field of the network theory. It focused on the graph theory and helps to realize the network in reality. M. Girvan and M.E.J. Newman had conceptualized the edge betweenness[2] in the extraction of the community in the network. Their method can be applied to the network where we had no prior knowledge of the communities. It is applicable on both directed and weighted graphs.

Eytan Bakshy et.al had propounded the theory that in the social networks there is a significant role of the strong ties as well as the weak ties. It shows that for the influencing purpose the strong connections are needed while for the transferring of the information the weak connections are responsible.

M.E.J. Newman [2]had expounded the fast algorithm for extracting the community structure in the networks with the introduction of concept modularity irrespective of the natural existing community in the network.

Ruchi Mittal and M.P.S. Bhatia [8]depending upon the network they had classified the communities into eight types. They had also surveyed and categorized the algorithms for extracting the structure of the community on the basis of the technique used to detect the community structure.

Avani Kesarwani. et al proposed Leader Similarity Based Community Detection LSBCD approach to divide the community on the basis of the leader and followers. This algorithm provides accurate results.

In our paper we focus on varying the threshold value of similarity extending the LSBCD [1] and how the variation of its value helps us to evaluate the size of the community. It's an important factor in determining the nodes that forms the community by calculating the similarity factor between the leader node and the other nodes present in the community.

III. REPRESENTATION OF THE NETWORK

A single layer network [1] can be represented as the nodes connected by the edges that represent the connection among the nodes. The more realistic tool to represent a network is the graph. G= (V, E) where V represents the list of the nodes while E represents the list of the edges. From the graph adjacency matrix is obtained. To visualize a finite graph adjacency matrix is used. The mathematical representation of graph is done with the help of adjacency matrix A and elements are represented as A_{ab} with the condition

A_{ab}= {1, when there exists an edge between node a and b,

0, otherwise}

IV. PROPERTIES OF THE NETWORK

There are some important metrics that we should to know before analyzing the community and its structure. They are degree, eigen vector centrality, clustering ad similarity.

Degree- is examined for a node and is defined as the number of edges connected to the nodes.

Eigen vector centrality- It calculates the influence of a node in a network. If a node is connected by a large number of nodes that also have large value of eigenvector centrality then that particular node will have high eigenvector centrality.

Clustering- In a network the clustering is the tendency of the nodes to group together. For e.g. if p knows q and q knows r then there may be a high possibility that q knows r if members are selected randomly.

Similarity- It is the capability of people to share a bond with the persons that have similar habits , ideas and forms community in the network on the basis of same point of view. This is the salient feature that enables us to analyze people with same hobbies, behavior to interact with each other and form a community in a network.

V. PROPOSED ALGORITHM

In [1] Leader similarity based community detection approach is introduced. In this work the leader is selected from the network and its followers are calculated on the basis of the similarity. In our work we had varied the threshold value of similarity and how the variation of its value helps us to evaluate the size of the community. It's an important factor in determining the nodes that forms the community by calculating the similarity factor between the leader node and the other nodes present in the community.

The steps that we follow are:

1) From the graph of the network the adjacency matrix is obtained.
2) The centrality of the network is calculated from the eigen vector centrality method.
3) The more influential node is chosen as the leader.
4) Now the similarity index is calculated which plays an important role
   In selecting the followers of the leader i.e. more is the similarity index between the influential node and the other node it will be the part of that community.
   It directly affects the size of the community. More is the value of the similarity threshold less will be the community size and vice-versa.
5) We have iterated the similarity N number of times and the calculations are done for 300 iterations. Using the parameters of NMI and ARI the accuracy is calculated.
6) Similarity index is measured using the four similarity parameters:
   a) Jaccard

$$sim_{jaccard}(q,r) = \frac{|\Gamma(q) \cap \Gamma(r)|}{|\Gamma(q) \cup \Gamma(r)|}$$

   b) Salton

$$sim_{salton}(q,r) = \frac{|\Gamma(q) \cap \Gamma(r)|}{\sqrt{(kq * kr)}}$$

   c) HDI

$$sim_{HDI}(q,r) = \frac{|\Gamma(q) \cap \Gamma(r)|}{\max\{kq, kr\}}$$

   d) HPI

$$sim_{HPI}(q,r) = \frac{|\Gamma(q) \cap \Gamma(r)|}{\min\{kq, kr\}}$$

   where k_q is the degree of q and $\Gamma(q)$ is the set of neighbors of q and similarly k_r is the degree of r and $\Gamma(r)$ is the set of neighbors of r.
7) Now the nodes that are isolated are added within the community.
8) The community that is detected is deleted from the network.
9) Again the same algorithm is applied for the remaining network.

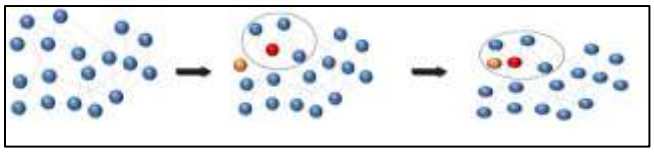


Fig. 1: Initial graph
Fig. 2: Community with isolated node
Fig. 3: Merging of nodes

In our study we use all the above mentioned four methods for measuring the similarity of leader node with the other node. All the similarity measures calculates the ratio of interaction of two sets in which one set is leader and other one is the node which is used to measure the similarity with the leader node. If the similarity of node is high it signifies that the interaction between the leader node and similar nodes is high and it leads to the formation of the community. With the similarity index we can conclude that the community size is similarity index sensitive.

VI. RESULTS

In this section the results of the proposed algorithm applied on the Zackary's karate club is compiled in which the ground truth is available.

A. Zackary's karate club network-

American university founded Zackary's karate club in 1977 consisting of 34 members and 78 interconnections i.e. edges between them. There occurred conflict between instructor and administrator of the club that results in the partition of club into two communities. After the conflict one group consists of 16 members while the other has 18 members. It is the undirected social network graph and is significant to test different community detection algorithms.

The dataset is significant to evaluate the accuracy of the proposed algorithm. The structure we get is tested by using two parameters –

Adjacent Rand Index (ARI)

Normalized Mutual information (NMI).

B. Adjacent Rand Index (ARI):

It is used to check that nodes are lies in same community as in ground truth community or not. Let us consider four unknowns w, x, y, z. now consider all possible node pairs. The value of w, x, y , z are respectively nodes lie in same

community in both Ground truth (G) and resultant community (R), nodes lie in same community in G but not in R, nodes lie in different community in G but in same community in R, nodes lie in different community in both G and R.

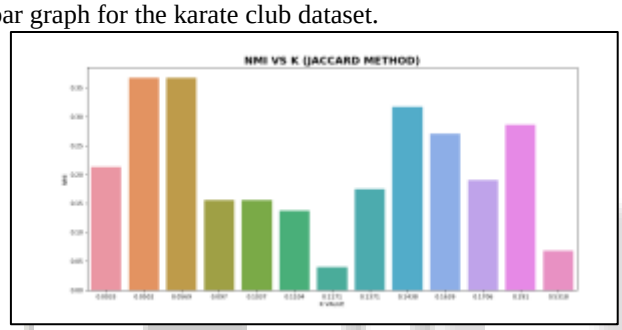
Then ARI is defined as ARI = n/2 (w+z)-[(w+x)(w=y)+(y+z)(x+z)]/(n/2)2-[(w+x)(w+y)+(y+z)(x+z)]

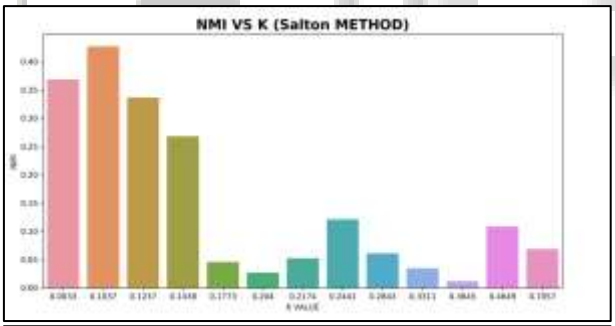
Normalized Mutual Information (NMI): This is also used to test the community measured by proposed algorithm. NMI for the community is measured as NMI (p,q) = 2I(p,q)/H(p)+H(q)

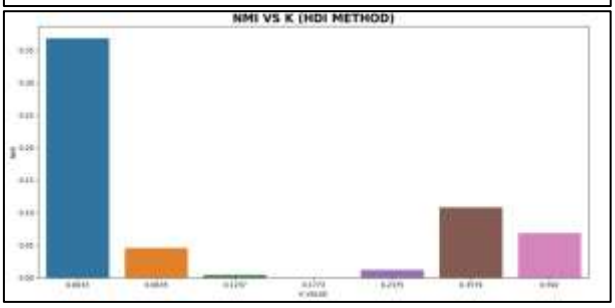
Where I(p,q) is mutual information of shared information by two variables, H(p) and H(q) are the corresponding entropy of p and q.

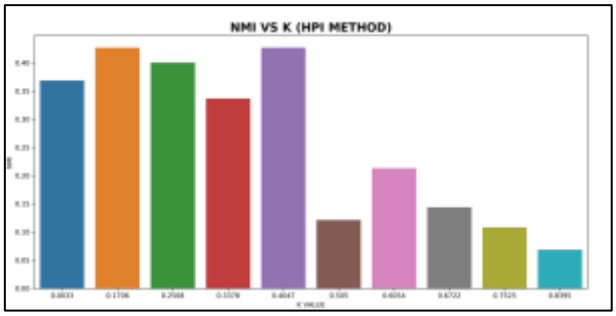
In this paper we use four similarity measure to sensitivity of the similarity threshold in the formation of the community and how its variation helps us to evaluate the size of the community. We have taken some set of values of k and NMI plotted the graph

The result of NMI v/s k for all the four methods of similarity i.e. jaccard, salton, hpi and hdi is shown by using bar graph for the karate club dataset.



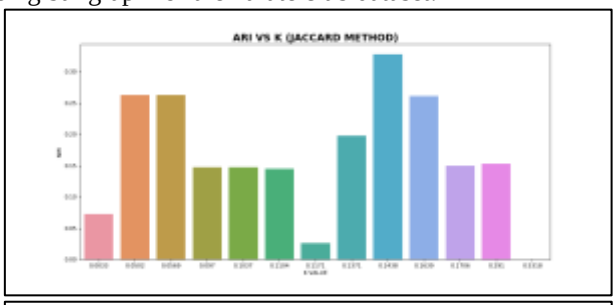


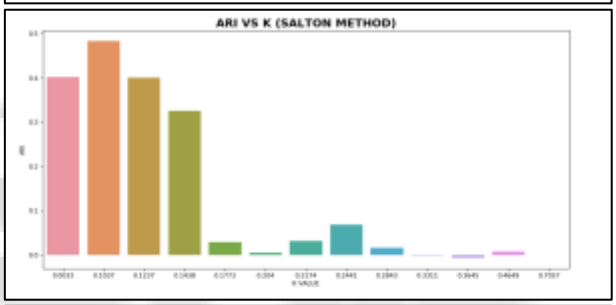


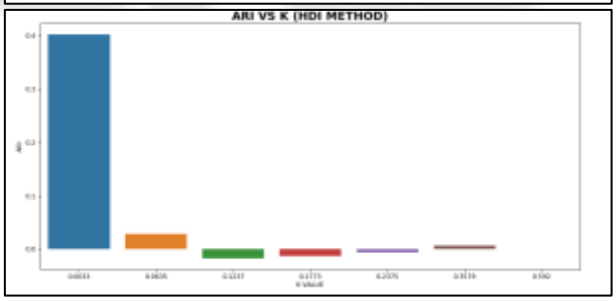


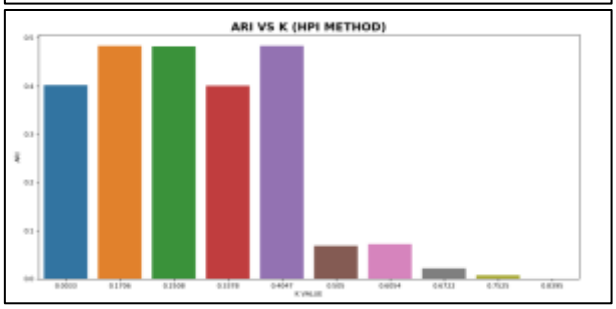
Again we have taken some set of values of k and ARI plotted the graph

The result of ARI v/s k for all the four methods of similarity i.e. jaccard, salton, hpi and hdi is shown shown by using bar graph for the karate club dataset.

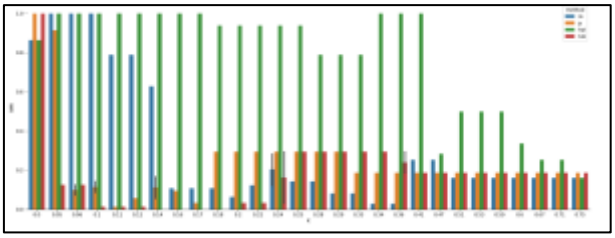




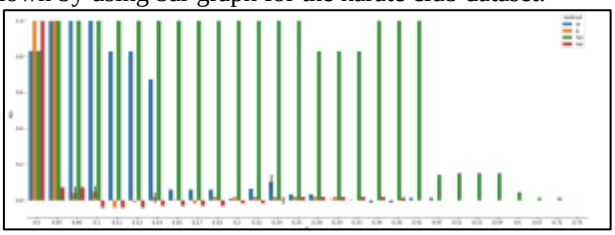




The combine bar graph of all the four methods of similarity i.e. jaccard, salton, hpi and hdi for NMI v/s k is shown by using bar graph for the karate club dataset.



The combine bar graph of all the four methods of similarity i.e. jaccard, salton, hpi and hdi for NMI v/s k is shown by using bar graph for the karate club dataset.



VII. CONCLUSION

By using HPI similarity we get more reliable community structure based on NMI and ARI. It gives more reliable result for small number of members in the network as compared to large members in the network. As in large members network the degree sometimes becomes zero that results in the error in the calculation of the eigen vector centrality.

REFERENCES

[1] A. Kesarwani, A. Singh, K. Gaurav and A. K. Shankhwar, "Leader Similarity Based Community Detection Approach for Social Networks," *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 2020, pp. 1-6, doi: 10.1109/INOCON50539.2020.9298371.

[2] Lee, Kyu-Min, Byungjoon Min, and Kwang-Il Goh. "Towards real-world complexity: an introduction to multiplex networks." *The European Physical Journal B* 88.2 (2015): 1-20.

[3] Lin, Shuyang, et al. "Understanding community effects on information diffusion." *Pacific-Asia conference on knowledge discovery and data mining*. Springer, Cham, 2015.

[4] Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." *Proceedings of the national academy of sciences* 99.12 (2002): 7821-7826.

[5] Ovens, Katie, B. Frank Eames, and Ian McQuillan. "Comparative Analyses of Gene Co-expression Networks: Implementations and Applications in the Study of Evolution." *Frontiers in Genetics* 12 (2021).

[6] Huang, Xinyu, et al. "A survey of community detection methods in networks." *Data Mining and Knowledge Discovery* 35.1 (2021): 1-45.

[7] Ovens, Katie, B. Frank Eames, and Ian McQuillan. "Comparative Analyses of Gene Co-expression Networks: Implementations and Applications in the Study of Evolution." *Frontiers in Genetics* 12 (2021).

[8] Mittal, Ruchi, and M. P. S. Bhatia. "Classification and comparative evaluation of community detection algorithms." *Archives of Computational Methods in Engineering* 28.3 (2021): 1417-1428.

