

Sink Node Elimination to Enhance the Performance of Overlapping Detection Algorithms along with Comparison of Existing Algorithm

Er. Rohit Kumar* and Er. Harpreet Arora

Department of C.S.E., IKG Punjab Technical University, Ibban, Kapurthala – 144603, Punjab, India;
roh1986@gmail.com, er.harpreetarora@gmail.com

Abstract

Objectives: To eliminate Sink nodes so that rate of detection can improve within the community overlapping detection and this also increases the modularity. It also consumes less time. **Methods/Statistical Analysis:** Modified k-clique Algorithm is used. Clique algorithm considers Sink nodes. Sink nodes are those which do not have any connecting edge. The proposed algorithm (MKC) does eliminate these nodes and hence consider only those nodes which are connected in nature. To detect the Sink nodes adjacency matrix is used. MATLAB is used for the simulation. Community detection toolbox is used which provides several functions for graph generations, clustering algorithms etc., **Findings:** This approach produces better result in terms of community finding. More community are discovered with the proposed approach and also entropy is improved greatly. Result in terms of time consumption is reduced almost by 50%. **Application/Improvements:** The length and complexity of the cliques found is reduced considerably. The speed is almost enhanced by 5% which can further be increased by using hop count mechanism in addition to the already used Sink node elimination.

Keywords: Community Overlapping, k-clique, Sink Nodes

1. Introduction

In real world the most widely utilized social media is community overlapping or modular structure because it explains the whole functioning of the system. To detect overlapping between these communities the dataset are

required. The dataset that are used fetched from the Snap University. The retrieved data is transformed into adjacency matrix.^{1,7-10} The adjacent nodes are to be placed in adjacency matrix and the nodes are adjacent if there exist an edge between them. The nodes that are adjacent to each other 1 is placed in adjacency matrix correspondingly

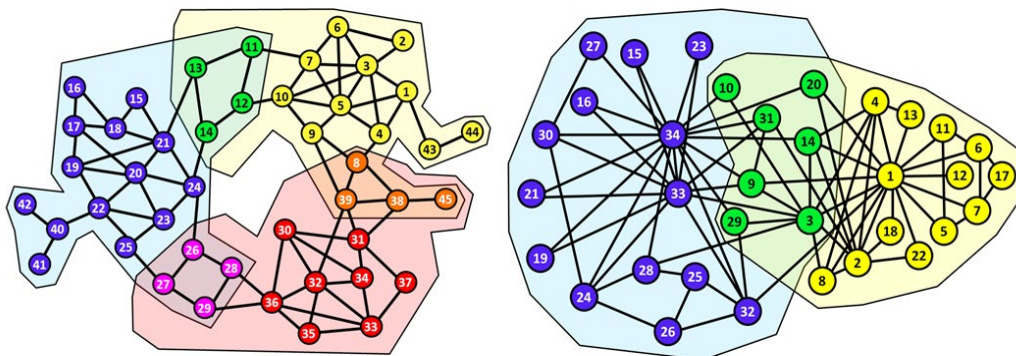


Figure 1. Community overlapping structure from social media.

*Author for correspondence

otherwise 0 is to be placed.² The overlapping community detection will be represented through Figure 1.

The liked resources can be retrieved by identifying the pages liked by users at most. To tackle such issues overlapping is reduced within the social networking pages. The real world environment can be understood by using Network communities, the group of nodes that are attached by some logical link, is known as a community. The people are multiple memberships associated with group are identified. A user can have many active areas like News, different rated movies. Including all stated active things are group to which a user may belong.

The set of groups which are not disjoint identifies by overlapping community exposure algorithm. The comparisons of various techniques are to be done in proposed paper that is used in overlapping community exposure.

2. Preliminaries

The basic definitions which are used throughout this paper are presented.³ Consider a Graph having set of nodes V and set of edges E and for network having n nodes complexity is denoted with $O(n)$.¹ The adjacency matrix will be used in order to determine which if elements are reachable from current node than adjacency matrix contains 1 otherwise adjacency matrix contains 0 node otherwise adjacency matrix will contains 0.

Nodes having adjacency corresponding to one node indicates that it has same number of liking and are grouped into common cluster.³ In the cluster nodes of same liking are grouped but node may or may not belongs to more than one cluster. Node belonging is judged with the help of AIC parameter. There exist constraints however. These constraints restrict overlapping nodes. There is following constraint to be satisfied in order for the node to be member of the clump

$$0 \leq a_{ic} \leq 1 \quad \forall i \in V, \forall c \in C$$

The existing work concentrates in the node detection which are overlapped is by the involvement of Sink nodes. The various algorithms which we are going to consider are as follows:

2.1 Clique Percolation

This method finds the overlapping cliques by determining strength of connection between nodes. Cliques with same liking are grouped together within common cluster. The size of the network is specified and critically

used for detection process.⁴ Once identification process is complete next subgraph with $k-1$ nodes is found out. Connected graph will be utilized to detect the overlapping community.³ Small values of overlapping factor k used to give good result.⁵ A small value of k also known as threshold value provide base for detection of overlapping community.³ Complexity of the network greatly depends upon the factor that graph is strongly connected or node. Node strongly connected within the graph with k nodes will have $k-1$ path associated with it. The nodes with low complexity will be judged on the basis of degree of inter-relationship. Nodes does not having any path from source to destination is known as Sink nodes. The primary task of proposed approach is to eliminate Sink nodes. This makes the overall process of detection easier and less complex. The nodes detected by the use of overlapping community detection mechanism are represented with the help of entropy. The entropy indicates degree of relationship between pixels.

The result for $k=4$ will be describe as shown in Figure 2.

The Figure 2 shows the findings of the cliques from the complex network like Facebook.

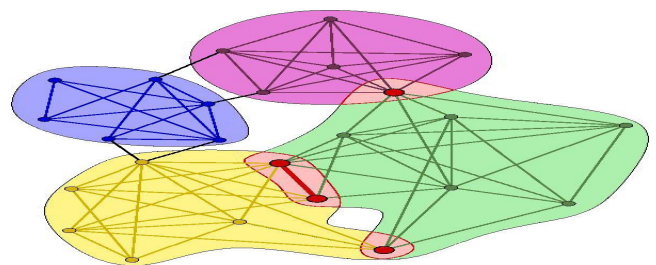


Figure 2. Showing the overlapping group of order 4.

Figure 3 shows all the communities which are present within the complex network. The complex network communities are determined by the use of Newman formula

$$Q = \frac{1}{4m} \sum_{ij} \left[\left(A_{ij} - \frac{K_i K_j}{2m} \right) (S_i S_j + 1) \right] = \frac{1}{4m} \sum_{ij} \left(A_{ij} - \frac{K_i K_j}{2m} \right) S_i S_j, \quad (1)$$

Q indicates the Modularity, A_{ij} indicates the adjacency Matrix, K indicates the node connection, m indicates constant factor whose value changes from 0 to 1 depending upon strong connection between graph and j .

2.2 Group Exposure

It is an important task in network analysis. A group (also referred to as a clump) is a set of cohesive vertices that have more connections inside the set than outside. In many

social and information networks, these communities naturally overlap. For instance, in a social network, each vertex in a graph corresponds to an individual who usually participates in multiple communities. In this paper, we propose an efficient overlapping Groupexposure algorithm using a seed expansion approach.¹¹ The key idea of our algorithm is to find good seeds, and then greedily expand these seeds based on a group metric. Within this seed expansion method, we investigate the problem of how to determine good seed nodes in a graph. In particular, we develop new seeding scenario for a personalized PageRank (PR) clumping scheme that optimizes the conductance group matches. The performance analysis indicates that Sink node elimination produces better result overlapping Groupexposure methods in terms of producing cohesive clumps and identifying ground-truth communities.

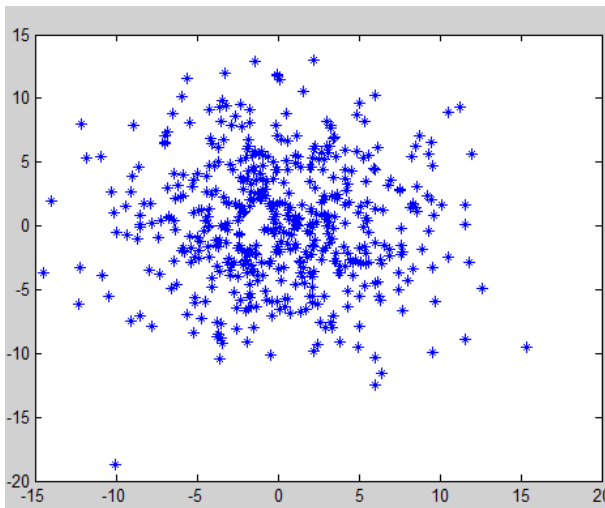


Figure 3. Showing k-clique mechanism.

2.3 Social Media

It is expanding day by day. With advancement in technology users can interact with each other through social media. Companies sell their products by publicising their product on social media. Users and vendors of social media are expanding with the rate of knots. Users are of varying intensions. Some users (consumers) and vendors are malicious in nature causing frauds. Fraud detection is objective of this literature.

2.4 Detecting Frauds

It is on social media is complex. The proposed literature uses modified k means clustering for detection of frauds on social media. Dataset derived from UCI is used for this

purpose. Simulation is conducted in MATLAB. Detection results indicate improvement in prediction by 20%.

2.5 Routing in Delay Tolerant Networks (DTN)

It adopts the store-carry-forward mode, and it requires nodes to forward data in a cooperative way. However, nodes may be not willing to help others in many applications and this behaviour can be called as individual selfish. On the other hand, nodes often can be divided into different communities, and nodes in the same group often have some social ties. Due to these social ties, nodes are more willing to help the one in the same group, but not others. This behaviour can be called as social selfish. Note that some nodes may belong to more than one community in the real world, and this phenomenon makes the network have overlapping communities. This paper proposed a theoretical model by the continuous time Markov process to describe the performance of Epidemic Routing (ER) in such network. Simulation result of this objective indicates enhancement in result by noticeable margin.

2.6 Social Media

It is providing resources for the user in order to make user interact with each other. It provides mechanism by which users interact with each other without physically moving from one location to other locations. Lack of time is causing more and more users to participate in the social media take over. The people now days do not have time. They communicate with each other by the use of internet. Internet provides number of mechanisms by which users can communicate with each other. Most common mechanisms which are used involve social media. Social media will help in establishing linkage between the different communities of users. The social media has allowed many users to share their views and also help the users around. But with the advent of the technology problems also start to appear. The main problem which is caused with the social media is deception. The deception model is then created in order to detect the problems with the online user accounts. Some users can have multiple accounts or some wrong information is provided by them. This paper describe that the deception is deliberate attempt to mislead the others. The deception will be such that the other user will not able to detect the falsifying information provided by the malicious users. The privacy of the users will be at stake if deception takes place.

3. Problem Definition

The Sink nodes have been considered in the existing system to detect group overlapping. Because Sink node consumes so much time so in our proposed work we will not consider them. We utilize K-Clique algorithm to detect overlapping. Overlapping community detection becomes complex as strongly connected graph is presented for work. Clique percolation method is commonly employed for this purpose.

The clump group can vary every time in the existing system it is very difficult to specify.⁶ We consider the adjacent node with value 1 to get rid of this problem instead of considering Sink node. In the whole network the adjacent nodes with value 1 will be considered and value 0 will be discarded.

4. Proposed Work

In the proposed work from the graph the adjacency matrix has been taken and by using this matrix we eliminate the nodes having 0 in row matrix. So time will be saved. The k-clique algorithm will be modified in this case. The proposed algorithm will be described as follows:

4.1 Algorithm

Algorithm begin by inputting number of nodes within the graph let it be n . Then algorithm detects cliques at each stage. Cliques detected should be more than threshold value. If at each stage cliques detected are more than k then algorithm terminates.

MKC(Graph G , Node n)

- a. Obtain Adjacency(A)= $Adj(G)$
- b. Set $i=1$
- c. Repeat while $i \leq n$
- d. Check Adj_i
- e. If($Adj_i > 0$)
- f. Accept the node(AC_i)= N_i
- g. Else
- h. Reject the node
- i. End of if
- j. Move to the next node
- k. $I=i+1$
- l. End of loop
- m. Calculate Modularity(Q) using Eq 1

The proposed algorithm will produce better result as compared to other existing algorithm.

5. Results

By the use of simulation and comparison Table 1, the proposed algorithm produces better result as compared to other algorithm that is described here. In this case the modularity is also calculated. The nodes are further divided in this and the modularity will be described in terms of the nodes. The complexity will be described as number of cliques within the network. The modules present within the system interact with each other are utilised to detect overlapping in group. The interaction can either be high or low. The complexity of the system can be decreased if the modules which are used are high. The numbers of modules which are used within the given system will decide the modularity. In other words modularity of the system should be high. In the existing system modularity is given as

K-Means	Hierarchal	K-Clique Modified
0.235	0.299	0.309

The comparison of various techniques in terms of the result will be as follows-

The performance of Modified Clique algorithm in terms of time is also better. The performance of K-clique algorithm without considering Sink nodes in terms of time is shown in the Figure 4.

From the above result it is clear that the Modified K-Clique without Sink nodes is better in every aspect.

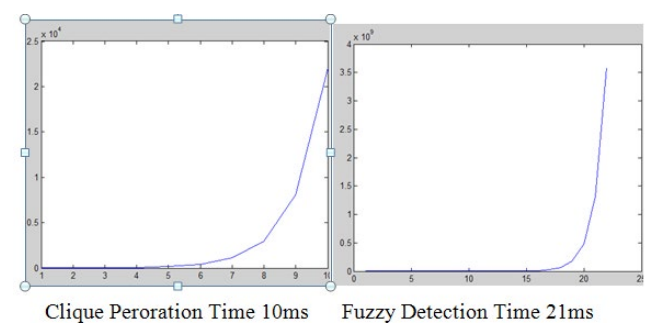


Figure 4. Showing time comparison of algorithm.

6. Conclusion

The improvement of results can be done by eliminating the Sink nodes from the graph in the given study of the Modified k-clique algorithm. In the adjacency matrix

where value greater than 1 is placed is considered. As the Sink node considered so fuzzy method will give least results in the comparison table. For detecting the clique we use much detection mechanism and also highlight the methods that are useful. In our proposed work we eliminate Sink nodes from the K-Clique method. The K-Clique is one of the simplest methods for the detection of the overlapping group detection.

Table 1. Describing the results of various algorithms regarding group overlapping

	K-Clique (Modified)	Fuzzy Detection	Hierarchal Clumping
Nodes Count	30	30	30
Size of Clique	4	4	4
Compared Nodes	10	21	26
Cliques	6	6	4
Consumed Time	10ms	21ms	26ms

7. References

1. Ghosh S, Dubey S. Comparative analysis of K-means and Fuzzy C-Means Algorithms. IJACSA; 2013. p. 35–9.
2. Hung WL, Yang MS, Chen DH. Parameter selection for suppressed fuzzy c-means with an application to MRI segmentation. Pattern Recognit Lett. 2006; 27(5):424–38. Crossref
3. Chatzis SP. A fuzzy c-means-type algorithms for clumping of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. Expert Syst Appl. 2011; 38(7):8684–9. Crossref
4. Ji Z, Xia Y, Sun Q, Cao G. Interval-valued possibilistic fuzzy C-means clumping algorithm. Fuzzy Sets Syst. 2014; 253:138–56. Crossref
5. Mingoti SA, Lima JO. Comparing SOM neural network with Fuzzy C-means, K-means and traditional hierarchical clumping algorithms. Eur J Oper Res. 2006; 174:1742–59. Crossref
6. Wang W, Zhang Y. On fuzzy clump validity indices. Fuzzy Sets Syst. 2007; 158(19):2095–117. Crossref
7. Wang X, Wang Y, Wang L. Improving fuzzy c-means clustering based on feature-weight learning. Pattern Recognit Lett. 2004; 25(10):1123–32.
8. Wu KL. Analysis of parameter selections for fuzzy c-means. Pattern Recognitt. 2012; 45(1):407–15. Crossref
9. Wu KL, Yang MS. Alternative c-means clustering algorithms. Pattern Recognitt. 2002; 35(10):2267–78. Crossref
10. Zhu W, Jiang J, Song C, Bao L. Clustering algorithm based on Fuzzy C-means and Artificial Fish Swarm. Procedia Eng. 2012; 29:3307–11. Crossref
11. Jierui X, Stephen K, Boleslaw KS. Overlapping Group Detection in Networks: The State-of-the-Art and Comparative Study. ACM Computing Surveys. 2013 Aug; 45(4):43.