Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 93 (2016) 1070 – 1076

Procedia Computer Science

6th International Conference On Advances In Computing & Communications, ICACC 2016, 6-8 September 2016, Cochin, India

Spanning Tree Based Community Detection using Min-Max Modularity

Ranjan Kumar Behera[a,*], S. K. Rath[a], Monalisa Jena[b]

[a]*Department of Computer Science and Engg., National Institute of Technology, Rourkela, Odisha, India, 769008*
[b]*Department of Information and Comm. Technology, F.M. University, Balasore, Odisha, India, 756019*

Abstract

Community refers to the group of entities which have similar behavior or characteristic among them. Usually community represents basic functional unit of social network. By understanding the behavior of elements in a community, one can predict the overall feature of large scale social network. Social networks are generally represented in the form of graph structure, where the nodes in it represent the social entities and the edges correspond to the relationships between them. Detecting different communities in large scale network is a challenging task due to huge data size associated with such network. Community detection is one of the emerging research area in social network analysis.

In this paper, a spanning tree based algorithm has been proposed for community detection which provides better performance with respect to both time and accuracy. Modularity is the well known metric used to measure the quality of community partition in most of the community detection algorithms. In this paper, an extensive version of modularity has been used for quality assessment.

Peer-review under responsibility of the Organizing Committee of ICACC 2016
Keywords: Modularity; Clustering; Community detection; Normalize Mutual Information

1. Introduction

Complex real world systems can be modeled into networks or structures of graph for analysis the behavior of components in the system. Usually graph are used for modeling social network, where the users are depicted as nodes and relationships between users are depicted as edges. Some of the users form groups based on the similar interest known as communities. Communities can be considered as a dense subgraph, where nodes inside the subgraph are strongly connected as compared to nodes outside the subgraph. Identifying such dense subgraph in the network is known as community detection. The process has proven to be effective in a number of research contexts such as biology, social sciences, bibliometrics, fraud detection, recommendation system, scientific collaboration analysis etc. However, social networks are more complex and dynamic in nature due to its heterogeneous media data. Community detection in such a complex network is a challenging task especially when data size is large.

* Corresponding author. Tel.: +91-943-959-2352
E-mail address: jranjanb.19@gmail.com

Peer-review under responsibility of the Organizing Committee of ICACC 2016
doi:10.1016/j.procs.2016.07.311

The objective of community detection is to identify the groups of entities which are corresponds to a functional components of social network. A number of methodologies are available in literature for community detection and most of which are based on structural analysis. In these algorithms an objective function is identified in order to optimize the features of the community structure.

In this paper, maximum spanning tree (MST) concept has been applied in order to explore the communities in large scale social network. It allows the community detection algorithm in reducing time complexity by preprocessing the dataset. The mapping of nodes to their communities can be done through maximum spanning tree. In most of the community detection algorithms, modularity has been chosen as a quality measure for community structure. But none of the algorithms in literature has considered the number of unrelated nodes present in the community. This quantity can be used to detect communities with more accuracy. In this paper a new modularity measure has been used in proposed algorithm that takes account of both dense connection and number of unrelated pairs inside the community.

The rest sections of the paper is organized as follows: In section 2, the related work in the field of community detection in large scale social network has been discussed. Section 3 brings out the description of min max modularity along with clustering coefficient. The proposed algorithm is presented in section 4. Implementation part is discussed in section 5. In section 6, a comparative study of different algorithms using evaluation metrics is presented. Conclusion and future work is presented in section 7.

2. Related Work

Community mining is one of the most emerging research areas in large scale social network analysis. There are quite a good number of challenges in social network analysis due to its exponential growth of data in recent past years. Link prediction, community detection, network evolution, influence analysis, keyword search, classification, clustering, transfer learning are the major research directions in social network analysis.

A good number of studies focusing on discovering communities are available in literature.[1 2 7 4] Several methods for community detection techniques have been developed and each has its own strength and weakness. An efficient community detection methods that used both local and global information about topological structure has been well explained by De Meo et al[9]. Global information about the network topology always give accurate community result, however it is not suitable for complex network, where as local information about network topology may lead to faster community detection, but are less accurate.

Pravin Chopade and Justin Zhan have discussed the structural and functional characteristics for community detection process in complex social network in their paper[3]. Community detection based on structural parameter of the network topology has more interest of research as compared to community based on functional parameter of the network.

Community detection using random walk has been extensively discussed by authors Pons,Pascal et al[14]. Walktrap algorithm for community detection is similar to the random walk model. The intuition of the walktrap algorithm is that a walker more likely to gets trap inside the dense region if it moves randomly inside the network. Girvan and Newman have employed edge-betweenness concept in their algorithm for community detection[2]. Edge-betweenness value of an edge can be measured by calculating all possible shortest path that pass through the edge. Edges exist between the communities seem to have more edge-betweenness value as compared to edges within the communities. Identifying edges with high edge-betweenness value may help in discovering community in large scale social network.

Steve Gregory proposed label propagation algorithm for community detection in linear time complexity[10]. The main idea behind the algorithm is that a node is more likely be a part of that community, where its maximum neighboring nodes belong to. Label of a node is propagated through its neighboring nodes in multiple iteration until a label is confined to a group of nodes. It is the fastest available community detection method, which has the linear time complexity. Community detection algorithm spends most of time in measuring the similarity between pair of nodes especially in case of unweighted graph.

2.1. Vertex Similarity

Two vertices are said to be more similar if they share large number of neighbors. The strength between two vertices can be calculated based on the similarity measures. One of the suitable measure, used to calculate the similarity is

based on Jaccard coefficient[11], which can be defined below:

$$S_{uv} = \frac{|neighbour(u) \cap neighbour(v)|}{|neighbour(u) \cup neighbour(v)|} \tag{1}$$

where $neighbour(u)$ and $neighbour(v)$ is a set of neighboring nodes of 'u' and 'v' respectively.

3. Methodology Adopted

3.1. MinMax Modularity

Modularity is the difference between number of edges exist within the community to the expected number of edges that would be present in a random assignment. In a random graph having 'n' nodes and 'm' edges, the expected number of edges between any two nodes 'i' and 'j' having degree 'd_i' and 'd_j' respectively is $d_i d_i/2m$. The actual number of edges between nodes 'i' and 'j', can be obtained by adjacency matrix 'A_{ij}'. The traditional modularity of a graph is given by the following equation[5]:

$$Q = \frac{1}{2m} \sum_{c} \sum_{i \in C, j \in C} A_{ij} - d_i d_j / 2m \tag{2}$$

For a better community structure, it is not enough to have strong connection inside the community. It is also desirable to have less number of unrelated node pairs within the community. If two nodes are connected by an edge, they are certainly related, however if link does not exist between pair of nodes, they may or may not be related to each other. In this work Jaccard similarity measured has been used to detect if they are related or unrelated. It is assumed that if the similarity value between a disjoint pair is greater than 0.5, they are said to be related otherwise they are considered to be unrelated.

In this paper, a new measure has been used to quantify the community known as MIN-MAX modularity. This modularity not only gives score to densely connected nodes, but also penalizes to unrelated node pairs within the community. The objective of this measure is to both maximize the connection and minimize the unrelated pairs of nodes within the community. It is observed that maximizing the number of edges within the group does not automatically minimize the unrelated pairs. The MIN-MAX modularity measure is defined as follows:

$$Q_{MIN-MAX} = Q_{Edge-density} - Q_{Unrelated-pair} \tag{3}$$

where $Q_{Edge-density}$ is the modularity value based on link density inside the group. It may be noted that it is same as described in equation 7. $Q_{Unrelated-pair}$ is the modularity value based on number of unrelated pair of nodes inside the community. In this paper the new objective value $Q_{MIN-MAX}$ is to be maximized by maximizing $Q_{Edge-density}$ and minimizing $Q_{Unrelated-pair}$. The first part of the equation can be calculated by following equation:

$$Q_{Edge-density} = \frac{1}{2m} \sum_{1 \le i,j \le n} (A_{ij} - E_{ij}) \delta(C_i, C_j) \tag{4}$$

where $\delta(C_i, C_j)$ is function that returns 1 if 'i' and 'j' corresponds to the same group and 0 if they belong to different group. E_{ij} is the expected number of edges between 'i' and 'j' and it can be defined as follows:

$$E_{ij} = \frac{d_i d_j}{2m} \tag{5}$$

The second part of equation 3 ie. $Q_{Unrelated-pair}$ can be calculated by transforming the graph into its complement form G^c where, edges between two nodes exist only if they are unrelated to each other. The adjacency matrix of the complement graph A^c_{ij} is defined as follows:

$$A^c_{ij} = \begin{cases} 1 & \text{if } A_{ij} = 0 \ \ and \ \ s(i,j) < 0.5 \\ 0 & otherwise \end{cases} \tag{6}$$

and

$$Q_{Unrelated-pair} = \frac{1}{2(n_{C_2} - m - \alpha)} \sum_{1 \le i,j \le m} (A^c_{ij} - E^c_{ij})\delta(C_i, C_j) \tag{7}$$

where

$$E^c_{ij} = \frac{d^c_i d^c_j}{2m'} \tag{8}$$

Here d^c_i and d^c_j are the degree of nodes 'i' and 'j' respectively in G^c. 'm' is the total number of edges in G^c. 'α' is the number of related pairs, between which there exist no edge.

3.2. Clustering Coefficient

Clustering Coefficient is another useful metric that defines probabilities of a group of node to make a community. It is associated with every node of the network. High clustering coefficient of a network indicates the presence of community structure. Strength of the community structure is affected by mean value of clustering coefficient in the network. The clustering coefficient of a node 'i' in an graph is defined as the ratio between number of edges exist to the total possible edges among the neighboring node of 'i'. It is given by the following equation[6]:

$$CC_i = \frac{2\,|\{e_{kl} : v_k, v_l \in N(i) \text{ and } e_{kl} \in E\}|}{N(i)(N(i) - 1)} \tag{9}$$

4. Proposed Algorithm

Spanning Tree Based Algorithm (STBA) for Community Detection

Input: The social network dataset in the form of graph $G = (V, E)$
Output: Partitioned network with multiple communities,
where 'V' and 'E' represent set of vertices and set of edges of the network respectively.

Step 1: The network is converted to adjecency matrix(A) where :

$$A_{ij} = \begin{cases} 1 \; if(i,j) \in E \\ 0 \; if(i,j) \notin E \end{cases} \tag{10}$$

Step 2: For every edge $(u, v) \in E$, the strength or weight can be obtained using following equation:

$$w(u,v) = \frac{C_{neigh}(u,v) + C_{edges}(u,v)}{T_{neigh}(u,v) + C_{neigh}(C_{neigh} - 1)/2} \tag{11}$$

where $C_{neigh}(u, v)$ and $T_{neigh}(u, v)$ is the number of common and total number of neighboring nodes between 'u' and 'v'. $C_{edges}(u, v)$ is the number of edges existing between common neighbors of u and v.

$$C_{neigh}(u,v) = |neighbour(u) \cap neighbour(v)| \tag{12}$$

$$T_{neigh}(u,v) = |neighbour(u) \cup neighbour(v)| \tag{13}$$

Step 3: After calculating weight of each edge in step 2, all the edges may be arranged in nondecreasing order.
Step 4: The maximum spanning tree of the given graph G(V,E) can be identified using Kruskal methods.
Step 5: The value of MIN-MAX modularity of the network is calculated using the following equation:

$$Q_t = \left[\frac{1}{2m} \sum_{i,j \in [1,n]} (A_{ij} - E_{ij}) - \frac{1}{2(n_{C_2} - m - \alpha)} \sum_{i,j \in [1,n]} (A^c_{ij} - E^c_{ij}) \right] \delta(C_i, C_j) \tag{14}$$

Step 6: An edge from the maximum spanning tree in the order obtained in step 3 is removed.
Step 7: The modularity value Q_{t+1} is then calculated after removal of the edge.
Step 8: ΔQ is calculated with the help of following equation:

$$\Delta Q = Q_{t+1} - Q_t \tag{15}$$

if $\Delta Q > 0$, step 6 and 7 are repeated otherwise the community partition of the network as found in step 6 is returned.

5. Implementation

5.1. Evaluation Metrics

Following evaluation metrics have been used for measuring performance of the proposed spanning tree based algorithm.

(i) **Normalized Mutual Information (NMI):** NMI is a measure used to access quality of community partition, if ground truth about community is available. It can be evaluated with the help of confusion matrix (CM) where each row corresponds to the community, present in the real partition and each column corresponds to community, detected through the proposed algorithm. Each element in the confusion matrix CM_{ij} represents the number of vertices in i^{th} real community, which are also present in j^{th} detected community. NMI of the detected partition can be defined as:

$$NMI(X,Y) = \frac{-2\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} CM_{ij} log(\frac{CM_{ij}CM}{CM_i CM_j})}{\sum_{i=1}^{n_X} CM_i log(\frac{CM_i}{CM}) + \sum_{j=1}^{n_Y} CM_i log(\frac{CM_j}{CM})} \tag{16}$$

where X and Y are the community partition structure corresponding to ground truth and detected structure respectively. CM_i and CM_j indicates the communities in true and detected community partition respectively.

(ii) **MIN-MAX Modularity:** MIN-MAX Modularity is the proposed measure, which has already been discussed in section 3.1.

(iii) **Accuracy:** In this paper accuracy of the detected community structure has been evaluated. It has been measured by calculating the percentage of vertices whose predicted community and ground truth community are same.

$$Accuracy = \frac{\sum_{i=1}^{n}\left(1 \cdot N_{l_{tv}=l_{pv}} + 0.N_{l_{tv}\neq l_{pv}}\right)}{n} \tag{17}$$

5.2. Datasets Used

The data sets for social network analysis is in the form of graph, which consists of several nodes and edges. The nodes depict as actors and edges depict as relationships among the actors in the network. For measuring the performance of proposed algorithm, following social network datasets has been taken into consideration.

- Zachary Karate Club: This dataset consists of members of a Karate club, collected from University Karate club by Wayne Zachary [12]. It consist of friendship of 34 members, where some of them have higher influence factors than others and average clustering coefficient is found to be 0.256.
- DBLP Citation Network [13]: This network consists of set of papers based on high energy physics collected between January 1993 to April 2003. It is a directed network, where there is an edge directed from node 'a' to node 'b' if paper 'a' cited paper 'b'.
- Amazon [13]: Amazon is the one the most popular online shopping network. When a customer buys a product he/she most likely purchases another co-product. There is an edge betwween product 'i' with product 'j' if they are frequently co-purchased by a customer.
- Youtube [13]: Youtube is the most popular online video sharing social network. In Youtube group of people with common interest form community. Ground truth about community is being mentioned in the Table I.

Details of the datasets is listed in TABLE 1.

Table 1: Datasets Used for Experiment

Datasets	No. of Nodes	No. of Edges	Clustering Coefficient	Communities
Zachary Karate Club	34	78	0.256	4
DBLP Citation Network	317080	1049866	0.6324	13477
Amazon	334863	925872	0.3967	75149
Youtube	1134890	2987624	0.0808	8385

6. Experimental Result

The experiment has been carried on a machine with i7 processor with 3.4Ghz clock speed and 4GB RAM. 'R' language has been used for measuring performance of the algorithms. Gephi tool[8] has been used for visualization of graph. Jaccard similarity index has been used to identify the unrelated pairs. The threshold value for similarity for disconnected pairs has taken to be 0.5. The proposed algorithm ie. STBA with MIN-MAX modularity has been compared with following community detection algorithms, available in literature:

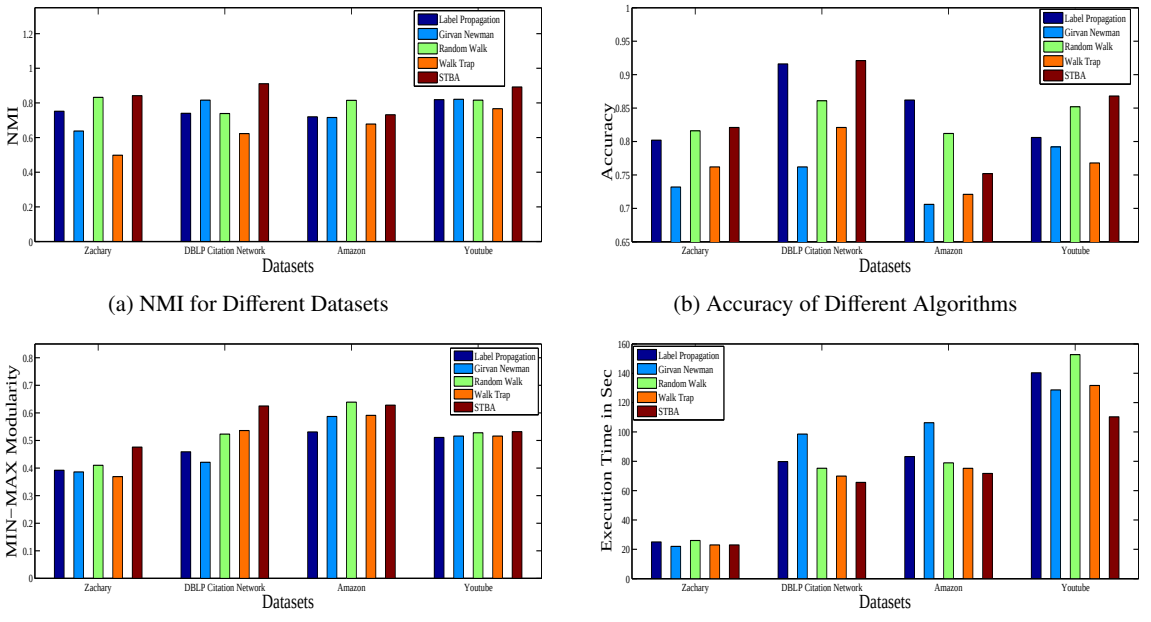
- Girvan Newman Community Detection (GN)[2]
- Label Propagation Community Detection (LP)[10]
- Walktrap Community Detection (WT)[14]
- Random Walk Community Detection (RW)[14]

NMI for different real world social network datasets has been calculated using different adaptive algorithms. The community structure obtained using STBA has better NMI values in all datasets except Orkut. Random Walk community detection algorithm provides better community partition, which are very close to structure provided by STBA. Accuracy of STBA is high in all datasets except Orkut, due to presence of large number of overlapping communities. From Figure 1b it can be easily identified that accuracy obtained in STBA and label propagation algorithm is higher as compared with other three algorithms. MIN-MAX modularity value has been measured for community partition structure of all datasets obtained through different algorithms. Higher the modularity value, better is the community partition. The MIN-MAX modularity value obtained through STBA is found to be large.

Social network is one of the largest source of data in Internet. The number of entities and their relationships are increasing exponentially in social network. Community detection in large scale network in reasonable amount of time is still a challenging task. For this reason, in this paper an effort has been made in measuring execution times for different algorithms in community detection. Figure 1d shows the comparative study of execution time for different community detection algorithms. The execution time of different algorithms for Zachary datasets has been normalized with a scale of 10 units in y-axis being represented as 1 second for better visibility. Although STBA does not have much less execution time for smaller datasets, it provides better performance for larger datasets.

7. Conclusion and Future Work

Community detection is one of the challenging problem in social network analysis. In this paper, an efficient and fast community detection algorithm has been proposed which is based on maximum spanning tree. A new modularity measure, which is based on both maximizing inter community density and minimizing unrelated node pairs inside the community has been used in the proposed algorithm in order to have better accuracy. From experimental analysis, it has been shown that the proposed algorithm ie. STBA provides better accuracy as compared to other well known community detection algorithms taken into consideration. It also provides better performance in terms of time complexity and modularity value, especially in case of large scale network.



(a) NMI for Different Datasets

(b) Accuracy of Different Algorithms

(c) MIN-MAX Modularity value for Different Datasets

(d) Execution time of Different Algorithms in Sec

Fig. 1: Comparative Study of different Community Detection Algorithms

The proposed algorithm can be extended to dynamic social network, where a large number of nodes along with their relationships are added more frequently. In future, distributed system like Spark or Hadoop can be considered for parallel processing of nodes and their edges in order to achieve better performance when data size is large.

References

1. Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.
2. Mark EJ Newman. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):321–330, 2004.
3. P. Chopade and J. Zhan, "Structural and functional analytics for community detection in large-scale complex networks," *Journal of Big Data*, vol. 2, no. 1, pp. 1–28, 2015.
4. Diane J Cook and Lawrence B Holder. *Mining graph data*. John Wiley & Sons, 2006.
5. Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
6. Peng Zhang, Jinliang Wang, Xiaojia Li, Menghui Li, Zengru Di, and Ying Fan. Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and its Applications*, 387(27):6869–6875, 2008.
7. Mark EJ Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.
8. Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. *ICWSM*, 8:361–362, 2009.
9. P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Mixing local and global information for community detection in large networks," *Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 72–87, 2014.
10. Steve Gregory. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10):103018, 2010.
11. Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In *Data mining (ICDM), 2013 ieee 13th international conference on*, pages 1151–1156. IEEE, 2013.
12. Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.
13. Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.
14. Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005*, pages 284–293. Springer, 2005.