International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

Analysis of Graph Clustering Method

Ms. P.S. Boraste [1], Prof. S. M. Kamalapur[2]
[1]*M.E. Student* [2]*Associate Professor) Department of computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik Savitribai Phule Pune University, Maharashtra, India.*

Abstract: ***Network data clustering has vital importance in various domains such as social network analysis, epidemiology, World Wide Web analysis, etc. The clustering technique derives underlying structures present in the graph. Along with the cluster creation, vertices classification is also an important task. To detect hubs and outlier is very important task in graph mining. There are various graph clustering algorithm such as graph partitioning, density based, modularity based, etc. This work aims to study various techniques of graph data clustering.***
Keywords: ***graph partitioning, hub, outlier, structural graph clustering, structural similarity.***

I. INTRODUCTION

Various dataset such as social network data, worldwide web data, ecommerce network, biological significant network can be represented in terms of graph theory. Graph can be directed or undirected. Edges may have weight representing the connection among nodes. Network clustering can also be referred as graph partitioning. Graph clustering helps in recognizing hidden structures in a network. In a graph G, vertices with the dense edge connection are belonging to one cluster and few edges among vertices are partitioned into different clusters. The extracted structure may represent a hidden community structure in a network.

In social network analysis like facebook and twitter community analysis can be done using graph partitioning technique [11]. This technique is required to analyse complicated structures and schema-less data. Along with the community detection overlapping community detection [9] has importance in social network analysis. The overlapping communities can be categorized in two parts: node based and structure based.

With the identification of clusters, special vertices such as hub and outlier are also having significant importance in graph structure analysis. Hubs are the vertices connected to many clusters in a network. Outliers are the vertices weekly connected to a cluster or isolated nodes. Hub identification is useful in various domains such as epidemiology and world wide web network. Outlier detection is applicable in spam detection techniques for e.g. Stripping spam pages from web pages.

There are various techniques to generate clusters from the given graph such as: graph partitioning, density based partitioning, modularity-based method. All these techniques focus on only the cluster detection technique and are not able to find special vertices such as hub and outlier.

Structural graph clustering is the technique that utilizes the structural aspects of vertices and it is able to distinguish different role of vertices in the graph clustering technique. SCAN, SCAN++ and PSCAN are algorithm which uses the structural graph clustering technique. This algorithm calculates the structural similarity among nodes for clustering.

Fig: 1 shows the two different communities detected from the given graph structure. Community1: vertex 1 to 6. Community 2: vertex 6 to 13. Vertex 6 belongs to two communities and hence it is a hub vertex. Vertex 13 has single connection with cluster 2 hence it is an outlier vertex

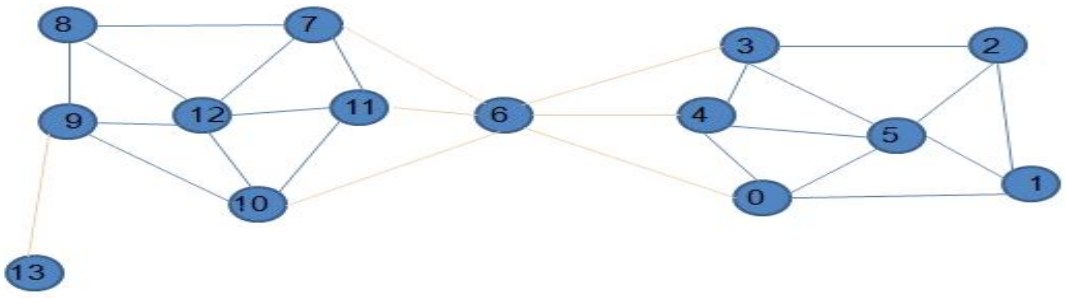

Fig: 1 Network Clusters and Special Vertices

Paper is organized as follows: Section I introduction about graph clustering. Section II gives the literature review. Section III concludes the paper.

II. LITERATURE WORK

In this section, various graph clustering techniques and comparative analysis of those techniques are studied. Graph partitioning, Density based partitioning, Modularity-based method are all technique used for graph clustering. All these techniques focus on only the cluster detection technique and are not able to find special vertices such as hub and outlier.

A. Other Graph lustering Methods

1) Modularity Approach: For finding clusters in large scale network Modularity Approach is proposed [7]. Modularity is the measure of structure in a network. In modularity approach, network is divided in communities with the help of greedy optimization technique based on the modularity strength of each community. But the outlier and hub like vertices are not detected in this method. Modularity approach works on undirected and un-weighted graph.

2) Graph Partitioning: This technique divides the graph in number of disjoint sub graphs called cluster. Min-max cut method [10] partitions the graph in two clusters A and B such that size of cluster A should be equal or similar to the size of cluster B. Cut defines the set of edges that need be removed from the graph to partition the graph in two sections. To achieve optimal solution for partitioning minimum number of edges need to be removed. The cluster size constraint proposed in min-max cut [4] algorithm is not valid for all the cases such as social network communities. In social network some communities are much larger than other communities. To overcome this problem Normalized cut [3] algorithm is proposed. By considering the total number of connections between the cluster and the rest of the graph connections cuts are normalized. Min-max cut and normalized cut generate two partitions in the graph. To generate k partitions this procedure need to followed recursively. But this solution may not lead to the optimal partitioning.

3) Dense Sub-graph Extraction: In social network analysis, network structure varies with respect to time. By imposing the temporal constraint over graph data, multiple graph snapshots can be generated. By identifying dense sub-graph over t snapshot as a maximal cliques and quasi-cliques, communities can be generated. [10]

4) Local Community based Clustering: In an online network, a local community can be detected by matching Q query keywords in a graph. Local community detection helps for identifying specific region of a graph with the help of specific metadata [5].

5) Density-based method: In Density based clustering, higher density area are forming one cluster. DBSCAN algorithm is used for density based clustering [1]. This algorithm evaluates core points, density-reachable points and outliers. To calculate core vertices, DBSCAN algorithm uses structural similarity computation.

SCAN, SCAN++ and PSCAN are extended versions of local density based algorithm techniques. Other techniques are not used for structural graph clustering. PSCAN aims to create clusters on weighed graph and also find special vertices like hub and outlier. It focuses on reduction of the structural similarity computations.

B. Structural Graph Clustering:

Various structural graph clustering algorithm are proposed in the literature. The basic three algorithms are: SCAN, SCAN++ and PSCAN. All these algorithms uses structural similarity technique but varies with respect to the performance aspects of clustering based on computational complexity and memory requirements.

1) SCAN: SCAN stands for 'Structural Clustering Algorithm for Networks' [6]. This technique computes the structural similarity among vertices and creates clusters of those vertices having nearly equal structural similarity.

Vertices u and v are structurally similar if $\delta(u,v) \geq \varepsilon$ and u is core vertex if it has at least μ structurally similar neighbors, where ε is the minimum similarity threshold and μ is the minimum cluster size threshold.

For algorithm execution ε and μ needs to be defined as a constant in the system. System dynamically computes the value of ε based on the given dataset and input parameter μ[7] [8]. Graph-skeleton [8] based clustering algorithm is proposed to generate Core-Connected Maximal Spanning Tree- CCMST. The value of ε is calculated using edge weights of the respective CCMST. System calculates the set of best ε values. Generated graph clusters using these ε values have highest modularity. But evaluating the value of ε is time consuming task. The results generated shows that value of ε do not vary with respect to the given input parameter μ and value of ε lies between 0.4 and 0.6 [8].

SCAN algorithm evaluates the structural similarity of all the vertices in the graph and hence SCAN is not applicable to the large networks.

2) SCAN++: To overcome the problem of high time complexity of SCAN algorithm, SCAN++ [12] is proposed. To overcome this problem directly two-hop-away reachable node set (DTAR) data structure is introduced. This system is based on the principle-'real-world graphs are expected to have high clustering coefficients' [9].
The topology of vertices and its neighboring vertices are likely to be clique if vertices having high clustering coefficients hence DTAR nodes with the given node are likely to be present in same cluster. Using this principle computational complexity is reduced.
3) pSCAN: Compared with SCAN algorithm SCAN++ requires less computation for structural similarity however the total computations are large. To overcome this problem new optimization approach is proposed in PSCAN algorithm [1]. This algorithm reduces the structural similarity computations. PSCAN provide optimized way for comparing structural similarity among vertices.
All these three algorithms worked on un-weighted undirected graphs.

III. CONCLUSION

Structural graph clustering is extended technique of Density based clustering. Other techniques such as Modularity Approach , Graph Partitioning Dense , Sub-graph Extraction, Local Community based Clustering are not used for structural graph clustering. Structural graph clustering is the only solution to find special vertices such as hub and outlier. Other techniques are not able to find such special vertices. The structural graph clustering algorithms works on un-weighted and undirected graph. There is need to develop a system for structural graph clustering applicable for weighted and directed graph.

IV. ACKNOWLEDGMENT

Authors would like to thanks Prof. Dr. K. N. Nandurkar, Principal and Prof. Dr. S. S. Sane, Head of Department of Computer Engineering, K.K.W.I.E.E.R., Nashik for their kind support and suggestions. We would also like to extend our sincere thanks to all the faculty members of the department of computer engineering and colleagues for their help.

REFERENCES

[1] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. Knowl. Discovery Databases Data Mining, 1996, pp. 226–231.
[2] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," Nature, vol. 393, pp. 440–442, 1998.
[3] J. Shi and J. Malik, "Normalized cuts and image segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 22, No. 8, 2000.
[4] C. H. Q. Ding, X. He, H. Zha, M. Gu, and H. D. Simon, "A minmax cut algorithm for graph partitioning and data clustering," in Proc. IEEE Int. Conf. Data Mining, 2001, pp. 107–114.
[5] R. Andersen, F. R. K. Chung, and K. J. Lang, "Local graph partitioning using pagerank vectors," in Proc. 47th Annu. IEEE Symp. Found. Comput. Sci., 2006, pp. 475–486.
[6] X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger, "SCAN: A structural clustering algorithm for networks," in Proc. 13th ACM SIGKDD Int. Conf. [2] Knowl. Discovery Data Mining, 2007, pp. 824–833.
[7] M. Kim and J. Han, "A particle-and-density based evolutionary clustering method for dynamic networks," Proc. VLDB Endowment, vol. 2, no. 1, pp. 622–633, 2009.
[8] H. Sun, J. Huang, J. Han, H. Deng, P. Zhao, and B. Feng, "gSkeletonClu: Density-based network clustering via structure connected tree division or agglomeration," in Proc. IEEE Int. Conf.Data Mining, 2010, pp. 481–490.
[9] J. Cheng, Y. Ke, S. Chu, and M. T. € Ozsu, "Efficient core decomposition in massive networks," in Proc. IEEE 27th Int. Conf. Data Eng., 2011, pp. 51–62.
[10] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community detection in social media," Data Mining Knowl. Discovery, vol. 24, no. 3, pp. 515–554, 2012.
[11] L. Chang, J. X. Yu, L. Qin, X. Lin, C. Liu, and W. Liang, "Efficiently computing k-edge connected components via graph decomposition," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2013, pp. 205–216.
[12] H. Shiokawa, Y. Fujiwara, and M. Onizuka, "SCAN++: Efficient algorithm for finding clusters, hubs and outliers on large-scale graphs," Proc. VLDB Endowment, vol. 8, no. 11, pp. 1178–1189, 2015.
[13] Lijun Chang, Wei Li, Lu Qin, Wenjie Zhang, Shiyu Yang, "PSCAN : Fast and Exact Structural Graph Clustering", Knowledge and Data Engineering IEEE Transactions on, vol. 29, pp. 387-401, 2017, ISSN 1041-4347.