

Research Article

Network Clustering Algorithm Based on Fast Detection of Central Node

Ziruo Jia ¹ and Fuqiang Qi²

¹China-France Joint Research Center of Applied Mathematics for Air Traffic Management, School of Economics and Management, Civil Aviation University of China, Tianjin, China
²School of Economics and Management, Civil Aviation University of China, Tianjin, China

Correspondence should be addressed to Ziruo Jia; zrjia@cauca.edu.cn

Received 25 December 2021; Revised 31 December 2021; Accepted 11 January 2022; Published 25 February 2022

Academic Editor: Punit Gupta

Copyright © 2022 Ziruo Jia and Fuqiang Qi. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Based on scale-free and density-based complex networks and numerical clustering algorithm, a graph clustering algorithm based on fast detection of central nodes is proposed. Through the calculation of local density and comprehensive clustering of nodes in the network, the clustering center in the network can be found quickly and noncentral nodes can be divided into the clustering center according to the nearest neighbor principle, thus avoiding parameter limitations such as the number of clustering to be set in advance using conventional classic social network detection algorithm. The experimental comparison and analysis in the real network indicate that the graph clustering algorithm based on fast detection of the central node is highly effective and efficient.

1. Introduction

The detection of social structure is vitally important to understand topology structure and the interaction between nodes in the network. With the development of application of graph clustering algorithm in different scientific fields, researchers have attached great importance to graph clustering algorithm and conducted researches and improvements. In recent years, graph clustering algorithm has been extensively used as the module maximization and spectral clustering algorithm. Module maximization transforms graph clustering into a problem of realizing module maximization and spectral clustering algorithm applies spectrum analysis technology in graph theory to graph clustering, thereby minimizing cutting. Furthermore, agglomeration algorithm and splitting algorithm as two common hierarchical clustering techniques [1, 2] have their respective advantages, which both require prior information such as network clustering number under great constraints. Although density clustering algorithm is not restricted by the condition to test the clustering of any shape, the general density clustering

algorithm is only used for numerical clustering, not for network clustering [3–6].

Density-based clustering algorithm is the crucial direction of clustering analysis. The classic density-based clustering algorithm is DBSCAN algorithm, which divides high-density areas into multiple clusters. Although it can find clusters of arbitrary shapes in spatial data with noise, it cannot reflect changes in high-dimensional data and data density [7]. Therefore, researchers improved the algorithm and proposed OPTCIS algorithm, which can be used for multidimensional spatial data clustering. The current algorithm improvements are based on DBSCAN and OPTICS. In recent years, the CFSDP based on density clustering algorithm proposed by Alex et al. is popular with the main idea of adhering to the following two principles in the selection of clustering center: firstly, the density of the clustering center itself is greater than the density of adjacent points; secondly, the distance between the clustering center and other clustering center is relatively long. Based on CFSDP algorithm in combination with scale free of complex networks, the clustering algorithm for fast detection of central nodes proposed in this paper has the advantages of

high efficiency of density-based numerical clustering algorithm and no need to set clustering quantitative parameters in advance [8].

2. Methodology

2.1. Density-Based Clustering Algorithm. Density-based cluster is a dense region separated by sparse regions in the data space, and the density in any cluster is higher than that in other noisy regions. Each core point requires for no fewer than MinPts points adjacent to radius Eps , that is, the density of the neighborhood must exceed a certain threshold. Let D be the data space containing all data points, and density-based cluster definition assumes that the distance function of a pair of points is $\text{dist}(p, q)$, and $\text{NEps}(p)$, the neighborhood of p , can be expressed as $\text{NEps}(p) = \{q \in D | \text{dist}(p, q) \leq \text{Eps}\}$. If $p \in \text{NEp}(q)$ and $|\text{NEp}(q)| \geq \text{MinPts}$, p is directly accessible from q . If there is a set of sample points p_1, p_2, \dots, p_n , $p_1 = q$, $p_n = q$ so that p_{i+1} is directly accessible from p_i to p and q , and the density accessibility is given by the transitive closure of direct density accessibility, that is, the specification extension of direct density accessibility. Since the secondary relation is transferred, if there is a point o that makes both p and q accessible from the density of o , then the density of p is related to the density of q .

Density-based clustering is the maximum set of data points connected by density. If a cluster C is a nonvoid subset of D , then

- (1) $\forall p, q$, if $p \in C$ and p from q can access to density, $q \in C$ (maximum)
- (2) $\forall p, q \in C$, p , and q has density connection (connectivity)

Let $C_1 \dots C_k$ be the clustering of data space D . Noise is the set of points in D that do not belong to any clustering C_i , that is, $\text{noise} = \{p \in D | \forall i: p \notin C_i\}$.

Density-based clustering distinguishes three different types of points: the core points with dense neighboring points, the boundary points that belong to the same clustering but not densely neighboring points, and the noise points not belonging to any clustering.

2.2. Graph Clustering Algorithm Based on Fast Detection of Central Node. An unweighted network is represented by $G = (V, E, W)$. V is the set of vertices, E is the set of edges, W_{ij} represents the weight of nonitemized edges connected to node V_i and V_j , and W_{ij} represents the distance and proximity between node V_i and V_j in the graph clustering algorithm. According to the central idea of graph clustering algorithm based on fast detection of center node, the network is divided into k disjoint subgraphs, with dense edge connections inside and sparse edge connections between each other. In the k subgraph, each subgraph contains a cluster center. When clustering, the cluster center should be identified first, and then the noncluster central nodes are divided into appropriate clusters. Therefore, graph clustering algorithm based on fast detection of central node needs

to solve three core problems: (1) how to define the density and distance of a node; (2) how to determine the clustering center; (3) noncluster center clustering strategy.

In scale-free network, the degree distribution of nodes follows the power-law distribution, that is, the probability that a node is connected to k , other nodes $P(k)$ satisfy $P(k) \sim k^{-r}$ (r is a constant.) Obviously, the fewer the nodes, the higher the degrees. In addition, these nodes are surrounded by denser nodes, allowing the importance of nodes to be described and all clustering centers in the network to be found. In order to describe the importance of nodes, the concept of local density is introduced, and the local density ρ_i of node i is defined as

$$\rho_i = \sum_j X(d_{ij} - d_c), \quad (1)$$

$$d_{ij} = \min_{z=i}^{j-1} W_{z(z+1)}.$$

If $X \leq 0$, $X(x) = 1$, otherwise $X(x) = 0$, d_{ij} represents the shortest distance between node i and j , and d_c is the truncation distance. ρ_i is equal to the number of adjacent points whose distance from node i is less than d_c . However, since the local density of nodes cannot determine the clustering center, for quick identification of the clustering center, the comprehensive distance β_i of node i should be introduced to represent the shortest distance weighted value from node i to nodes with higher local density:

$$\beta_i = \delta_i (\varphi_i + 1), \quad (2)$$

$$\delta_i = \min_{j: \rho_j > \rho_i} d_{ij}.$$

δ_i represents the minimum distance between node i and any other node with high density, φ_i is the median centrality of node i , which is the weight factor used to control the influence of node, and C is the constant. If node i has the highest density, set δ_i equal to 0.3 times the diameter of the network.

Obviously, only the nodes whose local density is the local or global maximum value has a much larger comprehensive distance than their neighboring points. Therefore, the clustering center is the node with abnormally large ρ_i and β_i and the network graph can be depicted by the local density of node ρ and the comprehensive distance β , thereby identifying the network clustering center quickly.

2.3. Algorithm Description. In the network $G = (V, E, W)$, to calculate the local density ρ of each node in the network G and the comprehensive distance β , the network G decision graph should be drawn on the basis of the ρ and β to determine the clustering center rapidly. After the cluster center is determined, the number of network clustering can be determined, and the noncluster central node can be divided into the cluster where the cluster center is located.

Firstly, through Algorithm 1, graph clustering algorithm, based on fast detection of central node, the input is network G and the output contain k clustering C .

- (i) Input: undirected unweighted graph $G = (V, E, W)$
- (ii) Output: k clustering $C = \{C_1, C_2, \dots, C_k\}$
- (1) Calculate the local density ρ and the comprehensive distance β of the node in figure G
- (2) Detect the clustering center by identifying the clustering center algorithm
- (3) Divide the noncluster centers into appropriate clusters by nonclustering central node algorithm
- (4) Return the clustering result $C = \{C_1, C_2, \dots, C_k\}$

ALGORITHM 1: Graph clustering algorithm based on fast detection of central node (CFCN).

- (i) Input: undirected unweighted graph $G = (V, E, W)$, local density ρ , and comprehensive distance β
- (ii) Output: k clustering $C = \{c_1, c_2, \dots, c_k\}$
- (1) Determine the threshold $thrho$ and $thbeta$ by characterizing the local density ρ with the integrated distance β decision graph
- (2) $k = 0$
- (3) For each vertex i in G , perform
- (4) If $\rho_i > thrho$ and $\beta_i > thbeta$, then
- (5) $k = k + 1$
- (6) Set node i as the clustering center c_k
- (7) End if
- (8) End for
- (9) Return $Cen = \{c_1, c_2, \dots, c_k\}$

ALGORITHM 2: Clustering center algorithm.

Then, find the clustering center by Algorithm 2: the clustering center algorithm finds out the clustering center, specifically as shown in Algorithm 2.

In Algorithm 2, the input is network G and the local density ρ and the comprehensive distance β of node i in Algorithm 1. The output is the Cen of clustering central node. In Algorithm 2 (1), the thresholds $thrho$ and $thbeta$ of ρ and β are determined by characterizing the decision graph. From (3) to (8), find all nodes with local density greater than $thrho$ and comprehensive distance greater than $thbeta$. Through Algorithm 2, the clustering center algorithm can find the clustering center in the network only by traversing the nodes in the network once. Nonclustering center node clustering algorithm is used to cluster other nonclustering central nodes (Algorithm 3).

The input is network G , local density of all nodes and comprehensive distance β , nonclustering central node U , and clustering center Cen; the output is K clustering C . In Algorithm 3, from (4) to (6), divide v into the cluster with the closest clustering center in Cen. In (7), remove the non-clustering central node that has been clustered from U and determine whether the clustering is completed by judging whether U is an empty set.

3. Results and Discussion

3.1. Data. Through its application on real networks (Karate; Lesmis; Polbook; Netscience; and Metabolic) and comparing module maximization algorithm (GN), spectral clustering algorithm (Spectral), and early density-based graph clustering algorithm (SCAN), the network clustering algorithm based on the rapid detection of the central node can be evaluated, and the main comparison indexes are selection

method ($F1_score$), module degree (Q), and performance (Performance). The algorithm is implemented in Python and in order to avoid randomness, perform the runs 10 times. Details are as shown in Table 1.

3.2. Comparison and Analysis. Firstly, apply CFCN algorithm to Karate network to verify that CFCN algorithm does not need to set the number of clustering before clustering. Figure 1 shows the network topology and social network structure. Set the truncation distance $dc = 1$, calculate the local density ρ_i of each node i and the comprehensive distance β_i , and visualize it as the decision graph (Figure 2). Nodes i and 34 in the decision graph are outliers, which have greater local density and comprehensive distance compared to other nodes, so they can be determined as the clustering center. In the practical application of Algorithm 2, clustering center algorithm, the threshold values $thrho$ and $thbeta$ need to be determined. The threshold value range of clustering centers 1 and 34 and other nodes can be determined by the decision graph. Finally, Algorithm 3, nonclustering center node algorithm, is applied to cluster other nonclustering center nodes. As shown in Figure 3, the same shape (colour) has the same clustering. Through the above examples, it is found that the number of clustering is only related to the network topology, with no need for setting in advance.

Secondly, apply the CFCN algorithm, GN algorithm, and spectral algorithm to real social networks for comparative testing and evaluate clustering quality through the performance and module degree Q to verify the higher efficiency of CFCN algorithm. Set the parameter $dc = 1$; Table 1 shows the comparison of the modularity and performance of CFCN algorithm and other algorithms in social networks with different structures. It can be found that although the GN

Input: undirected unweighted graph $G = (V, W, W)$, local density ρ and comprehensive distance β , nonclustering central node U , and clustering center $\text{Cen} = \{c_1, c_2, \dots, C_k\}$

(ii) Output: clustering result $C = \{C_1, C_2, \dots, C_k\}$

- (1) Add c_k to C_k
- (2) While $U \neq \phi$, perform
 - (3) $v = a$ vertex in U
 - (4) For each $c \in \text{Cen}$, perform
 - (5) Join node v into the cluster to connect closer c
 - (6) End for
 - (7) Remove node v from U
- (8) Return C

ALGORITHM 3: Nonclustering central node clustering algorithm.

TABLE 1: Data statistics.

Network	Number of nodes	Number of edges	Average degree	Number of clusters
Karate	34	78	4.589	2
Lesmis	75	252	6.595	4
Polbook	106	442	8.413	3
Netscience	378	912	4.827	8
Metabolic	453	2041	9.005	11

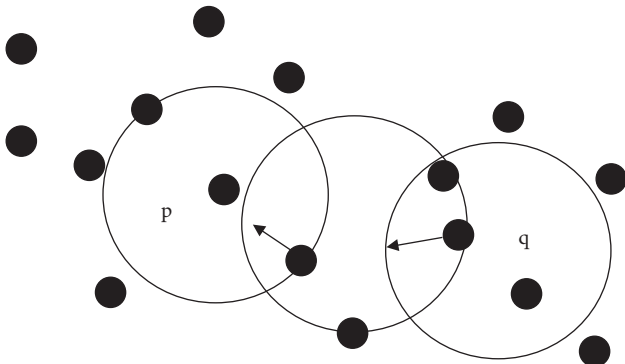


FIGURE 1: Clustering results.

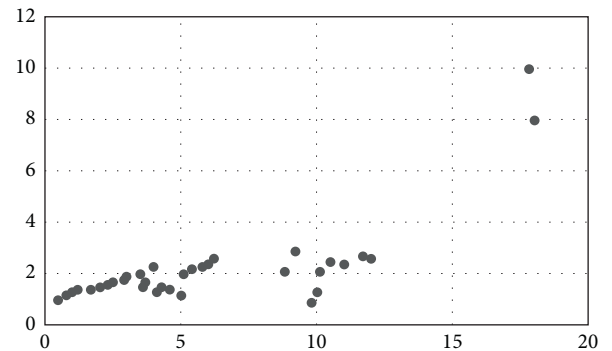


FIGURE 3: Density-reachable.

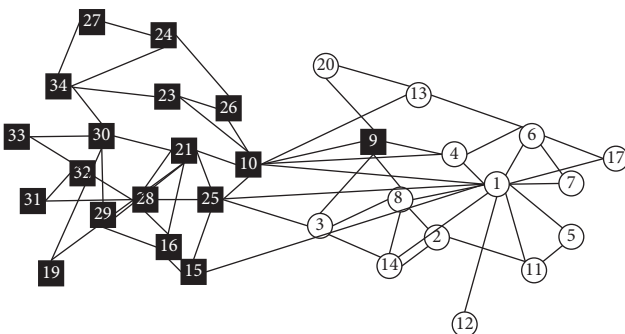


FIGURE 2: Decision-making.

algorithm in Polbook network module is better than CFCN algorithm, spectral algorithm in Metabolic network performance is better than CFCN algorithm, CFCN is superior to other algorithms overall.

4. Conclusion

This paper conducted an experimental research on graph clustering algorithm based on fast detection of central node. The following conclusions can be drawn from this research.

Through the calculation of local density and comprehensive clustering of nodes in the network, the clustering center in the network can be found quickly and noncentral nodes can be divided into the clustering center based on the nearest neighbor principle, thus avoiding parameter limitations such as the number of clustering to be set in advance in the conventional classic social network detection algorithm.

The experimental comparison and analysis in the real network indicate that the graph clustering algorithm based on the fast detection of the central node is highly effective and efficient.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under No. 51904313.

References

- [1] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, “The large-scale organization of metabolic networks,” *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.
- [2] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: an overview,” *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.
- [3] E. Ahmadi, M. Zandieh, M. Farrokh, and S. M. Emami, “A multi objective optimization approach for flexible job shop scheduling problem under random machine breakdown by evolutionary algorithms,” *Computers & Operations Research*, vol. 73, no. 9, pp. 56–66, 2016.
- [4] M. Aladeemy, S. Tutun, and M. T. Khasawneh, “A new hybrid approach for feature selection and support vector machine model selection based on self-adaptive cohort intelligence,” *Expert Systems with Applications*, vol. 88, no. 9, pp. 118–131, 2017.
- [5] I. Skrjanc, J. A. Iglesias, A. Sanchis, D. Leite, E. Lughofer, and F. Gomide, “Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: a Survey,” *Information Sciences*, vol. 490, no. 6, pp. 344–368, 2019.
- [6] J. J. Whang, D. F. Gleich, and I. S. Dhillon, “Overlapping community detection using neighborhood-inflated seed expansion,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 5, pp. 1272–1284, 2016.
- [7] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.
- [8] M. Hajjar, G. Aldabbagh, N. Dimitriou, and M. Z. Win, “Relay selection based clustering techniques for high density LTE networks,” *Wireless Networks*, vol. 25, no. 5, pp. 2305–2314, 2019.