Logical Clustering of Similar Vertices in Complex Real-World Networks

Md A. Rahman and Natarajan Meghanathan

Abstract We show that vertices part of a physical cluster (determined per the edges that connect the vertices) in a complex real-world network need not be similar on the basis of the values incurred for node-level metrics (say, centrality metrics). We adapt a recently proposed approach (based on unit-disk graphs) to determine logical clusters comprising of vertices of similar values for node-level metrics, but need not be physically connected to each other. We use the Louvain algorithm to determine both the physical and logical clusters on the respective graphs. We employ the Silhouette Index measure to evaluate the similarity of the vertices in the physical and logical clusters. When tested on a suite of 50 social and biological network graphs on the basis of neighborhood and/or shortest path-driven centrality metrics, we observe the Silhouette Index of the logical clusters to be significantly larger than that of the physical clusters.

Keywords Logical clusters · Physical clusters · Centrality metrics · Silhouette index · Complex network analysis

1 Introduction

Clustering (also referred to as community detection) is a critical component of complex network analysis. In the traditional sense, a cluster (community) in a network comprises a group of vertices that are more connected to each other than to vertices outside the cluster [1]. We refer to such clusters as physical clusters. Several clustering algorithms (like Girvan–Newman algorithm [2], Louvain algorithm [3], and neighborhood-overlap-based greedy algorithm [4]) are available in the literature to determine physical clusters of vertices of larger modularity. A physical cluster

M. A. Rahman · N. Meghanathan (✉)
Computer Science, Jackson State University, Jackson, MS 39217, USA
e-mail: natarajan.meghanathan@jsums.edu; nmeghanathan@jsums.edu

M. A. Rahman
e-mail: md.a.rahman@students.jsums.edu

G. Ranganathan et al. (eds.), *Inventive Communication and Computational Technologies*, Lecture Notes in Networks and Systems 89,
https://doi.org/10.1007/978-981-15-0146-3_86

is considered to be more modular [1] if it has high intra-cluster density and low inter-cluster density.

With the objective of maximizing edge-based intra-cluster density, it is difficult to expect a clustering algorithm to group vertices that are similar to each other. Similarity of vertices is typically assessed on the basis of node-level metrics that could be topology or domain-driven. Centrality metrics are classical examples for topology-based metrics that quantify the extent to which a node is important on the basis of its location in the network [1]. The centrality metrics are used as the node-level metrics for our analysis in this paper. We consider the following centrality metrics: neighborhood-driven degree (DEG) [1] and eigenvector (EVC) [5] centrality metrics and the shortest path-driven betweenness (BWC) [6, 7] and closeness (CLC) centrality metrics [8, 9]. For more details on the centrality metrics, the interested reader is referred to [10]. The analysis presented in this paper could be seamlessly extended to any combination of domain-driven or topology-driven metrics as well.

In Fig. 1, we show a motivating example graph wherein the tuple next to a vertex represents the degree (DEG) and eigenvector centralities (EVC) of the vertex. Any well-known clustering algorithm in the literature would determine the two physical clusters in sub Fig. 1a that have high modularity (larger intra-cluster density and lower inter-cluster density). However, a closer look at the tuples for the vertices within a physical cluster would indicate less similarity among the vertices on the basis of their DEG and EVC values. On the other hand, in sub Fig. 1b, we show two clusters each of which comprises vertices that are exactly similar on the basis of their DEG and EVC values. The two clusters in sub Fig. 1b are not very modular (the inter-cluster density is even larger than the intra-cluster density), but each of these clusters would be more cohesive (i.e., comprised of vertices that are similar) on the basis of their DEG and EVC values.

We propose the following approach (more details are in Sect. 2) to determine such cohesive clusters of similar vertices in real-world network graphs. We distribute the vertices in a coordinate system whose coordinates are the normalized values of the node-level (centrality) metrics of the vertices. For such a logical topology,

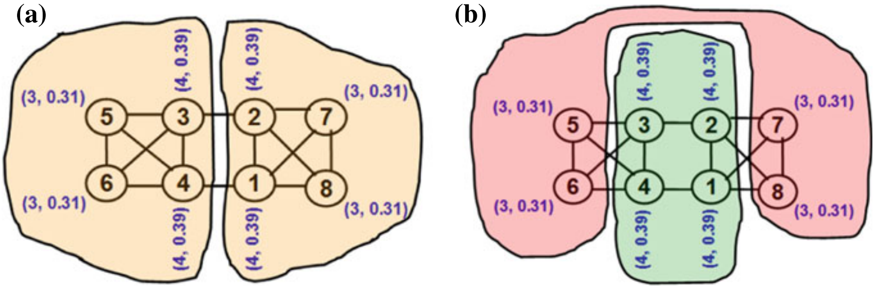


Fig. 1 Example graph: physical modular clusters versus logical clusters of similar vertices. **a** Physical clusters (larger intra-cluster density, but lower vertex similarity). **b** Logical clusters (lower intra-cluster density, but larger vertex similarity)

we iteratively attempt to connect the vertices together (an edge exists between two vertices if the Euclidean distance between their normalized coordinate values is within a threshold) in the form of a unit-disk graph. We use a binary search approach to identify the minimum value for the threshold distance that would connect together the vertices in the unit-disk graph. We run the Louvain clustering algorithm [3] on the connected unit-disk graph to determine one or more (logical) clusters whose member vertices are more similar compared to the vertices in the physical clusters obtained by running the Louvain algorithm on the corresponding real-world network graph. We use the Silhouette Index [11] measure to assess the similarity of the vertices in the logical clusters vis-a-vis the physical clusters with respect to the centrality metrics considered. In a recent work [12], we have successfully used the unit-disk graph and binary search-based approach to quantify the similarity between any two vertices in a network (in the form of a metric called the *node similarity index*). Our hypothesis in this research is that for a set of centrality metrics, the Silhouette Index of the logical clusters would be larger than the Silhouette Index of the physical clusters.

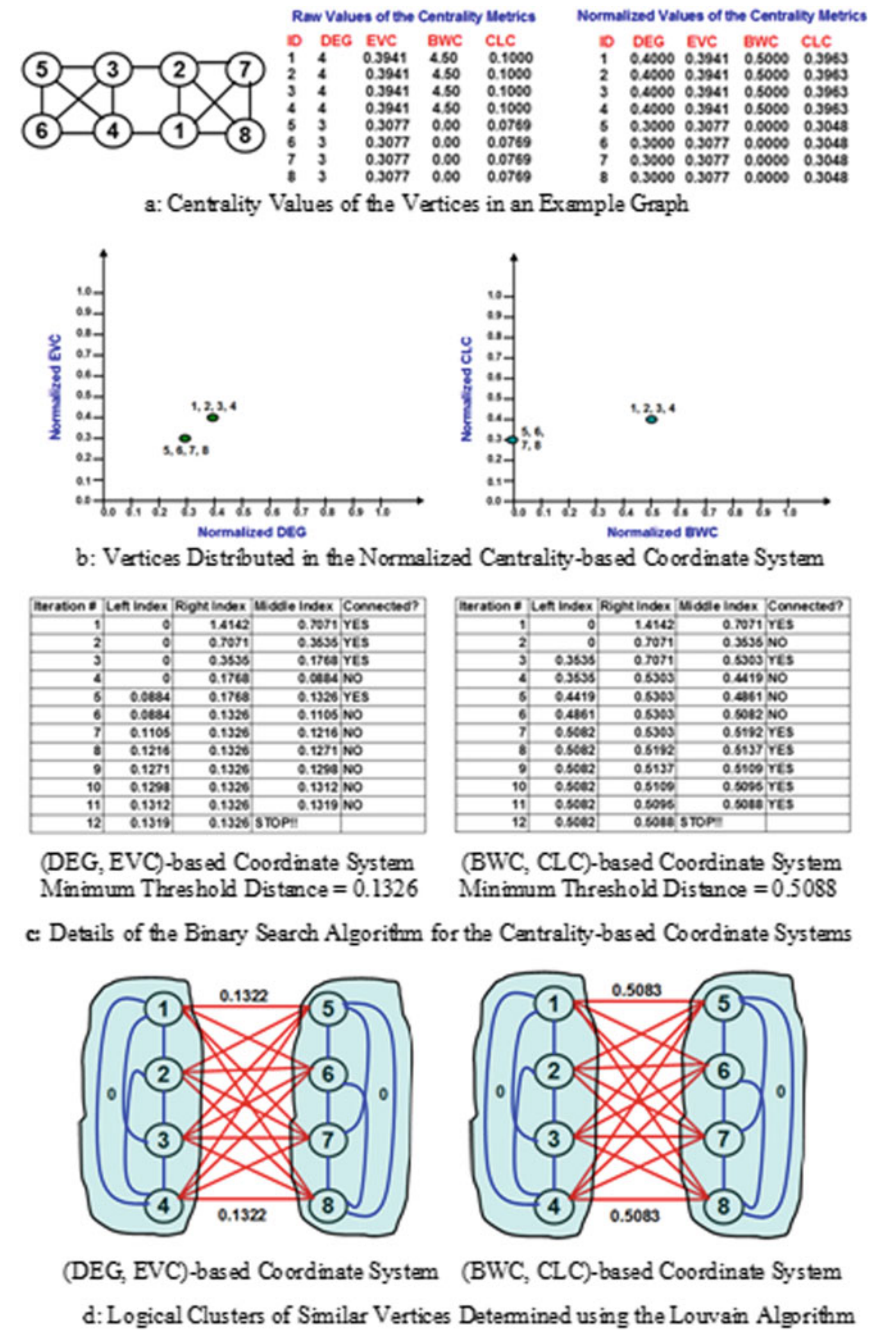
The rest of the paper is organized as follows: In Sect. 2, we explain the approach to determine logical clusters of similar vertices on the basis of node-level (centrality) metrics. In Sect. 3, we explain the formulation for the Silhouette Index measure. In Sect. 4, we test our hypothesis on a suite of 25 biological networks and 25 social networks on the basis of three sets of centrality metrics: (DEG, EVC); (BWC, CLC), and (DEG, EVC, BWC, CLC). We present and analyze the Silhouette Index results for the physical clusters and logical clusters obtained for the different combinations of centrality metrics. Section 5 presents our conclusions and outline plans for future work.

2 Logical Clusters of Similar Vertices

We describe the sequence of steps (see Fig. 2) involved in determining logical clusters of similar vertices (for k centrality metrics) in a real-world network graph.

Step (i): Construction of a Logical Topology: Let the number of centrality metrics considered for logical clustering of a real-world network graph be k. To get started, for each of the k metrics, we determine their raw centrality values (see Fig. 2a) and then independently normalize them (using the square root of the sum of the squares of the raw values). Following this, we build a logical topology (k-dimensional coordinate system) of the vertices wherein the coordinates of a vertex are its normalized centrality values (ranging from 0 to 1). See Fig. 2b.

Step (ii): Binary Search Algorithm to Deduce a Connected Unit-Disk Graph: We attempt to deduce (through a sequence of iterations) a unit-disk graph that would connect all the vertices in the logical topology at a minimum threshold distance for an edge to exist. For a set of k centrality metrics, the threshold could range from (0,…, $\sqrt{k}$); if the threshold distance is $\sqrt{k}$, we will have a unit-disk graph that is sure to be connected (completely connected indeed!), but not connected at a threshold



Raw Values of the Centrality Metrics

ID	DEG	EVC	BWC	CLC
1	4	0.3941	4.50	0.1000
2	4	0.3941	4.50	0.1000
3	4	0.3941	4.50	0.1000
4	4	0.3941	4.50	0.1000
5	3	0.3077	0.00	0.0769
6	3	0.3077	0.00	0.0769
7	3	0.3077	0.00	0.0769
8	3	0.3077	0.00	0.0769

Normalized Values of the Centrality Metrics

ID	DEG	EVC	BWC	CLC
1	0.4000	0.3941	0.5000	0.3963
2	0.4000	0.3941	0.5000	0.3963
3	0.4000	0.3941	0.5000	0.3963
4	0.4000	0.3941	0.5000	0.3963
5	0.3000	0.3077	0.0000	0.3048
6	0.3000	0.3077	0.0000	0.3048
7	0.3000	0.3077	0.0000	0.3048
8	0.3000	0.3077	0.0000	0.3048

a: Centrality Values of the Vertices in an Example Graph

b: Vertices Distributed in the Normalized Centrality-based Coordinate System

Iteration #	Left Index	Right Index	Middle Index	Connected?
1	0	1.4142	0.7071	YES
2	0	0.7071	0.3536	YES
3	0	0.3536	0.1768	YES
4	0	0.1768	0.0884	NO
5	0.0884	0.1768	0.1326	YES
6	0.0884	0.1326	0.1105	NO
7	0.1105	0.1326	0.1216	NO
8	0.1216	0.1326	0.1271	NO
9	0.1271	0.1326	0.1298	NO
10	0.1298	0.1326	0.1312	NO
11	0.1312	0.1326	0.1319	NO
12	0.1319	0.1326	STOP!!	

(DEG, EVC)-based Coordinate System
Minimum Threshold Distance = 0.1326

Iteration #	Left Index	Right Index	Middle Index	Connected?
1	0	1.4142	0.7071	YES
2	0	0.7071	0.3536	NO
3	0.3536	0.7071	0.5303	YES
4	0.3536	0.5303	0.4419	NO
5	0.4419	0.5303	0.4861	NO
6	0.4861	0.5303	0.5082	NO
7	0.5082	0.5303	0.5192	YES
8	0.5082	0.5192	0.5137	YES
9	0.5082	0.5137	0.5109	YES
10	0.5082	0.5109	0.5096	YES
11	0.5082	0.5096	0.5088	YES
12	0.5082	0.5088	STOP!!	

(BWC, CLC)-based Coordinate System
Minimum Threshold Distance = 0.5088

c: Details of the Binary Search Algorithm for the Centrality-based Coordinate Systems

(DEG, EVC)-based Coordinate System (BWC, CLC)-based Coordinate System

d: Logical Clusters of Similar Vertices Determined using the Louvain Algorithm

Fig. 2 Example to illustrate the computation of the logical clusters and their evaluation

distance of 0 (unless all the vertices are co-located). We use this observation as the basis to run a binary search algorithm to determine the minimum possible value for the threshold distance to obtain a connected unit-disk graph [12]. We start the binary search algorithm with the left index set to 0 and the right index set to $\sqrt{k}$ and go through a sequence of iterations (see Fig. 2c).

Across all the iterations, the following invariant is maintained: If the threshold distance value corresponds to the left index, the unit-disk graph is not connected; if the threshold distance value corresponds to the right index, the unit-disk graph is connected. In each iteration, we first determine the middle index as the average of the left index and right index, and seek to construct a unit-disk graph with the threshold distance value corresponding to the middle index. If such a unit-disk graph is connected, we set the right index to the value of the middle index; otherwise, we set the left index to the value of the middle index. We continue the iterations until the right index and left index differ not more than a cutoff parameter ($\in$). We use $\in = 0.001$ for all the analysis conducted in this paper. We set the minimum threshold distance to correspond to the value of the right index in the last iteration of the algorithm.

Step (iii): Logical Clustering of the Connected Unit-Disk Graph: On the connected unit-disk graph obtained for a minimum threshold distance, we run the Louvain community detection algorithm [3] to determine (logical) clusters of vertices that have a larger intra-cluster density (and a lower inter-cluster density) in the unit-disk graph. The vertices within a logical cluster are expected to be more similar to each other. The Louvain algorithm (a hierarchical community detection algorithm) is designed to identify highly modular communities. To determine the logical clusters using the Louvain algorithm, the weight of an edge in the connected unit-disk graph is the Euclidean distance between their corresponding normalized coordinate values. Note that the edge weights for the real-world network graphs are "1" when we run the Louvain algorithm to determine the physical clusters.

3 Silhouette Index

We use the Silhouette Index [11] measure (ranges from −1 to 1) to evaluate the extent of similarity among the vertices of the physical clusters and logical clusters with respect to the normalized centrality values of the vertices. The larger the values for the Silhouette Index for a cluster, the more similar are the vertices within the cluster with respect to the centrality metrics in consideration. The Silhouette Index for a cluster is the average of the Silhouette Index values of its member vertices. The Silhouette Index for a network is the weighted average of the Silhouette Index values of its clusters. The Silhouette Index for a vertex i in a cluster C_k is calculated as per formulation (1). Here, $\overline{d_{i,\min}}$ is the minimum of the average of the Euclidean distances for vertex i to vertices in the other clusters; $\overline{d_{i,Ck}}$ represents the average of the Euclidean distances for vertex i to vertices in its own cluster. A negative Silhouette Index for a vertex is an indication that the vertex is not in the appropriate cluster.

A negative Silhouette Index for a cluster is an indication that its member vertices should have been in other cluster(s) for better cohesiveness.

$$\text{Silhouette Index}(i) = \frac{\overline{d_{i,\min}} - \overline{d_{i,Ck}}}{\max\left\{\overline{d_{i,\min}}, \overline{d_{i,Ck}}\right\}} \tag{1}$$

4 Evaluation on Real-World Networks

In this section, we test our hypothesis on a suite of 25 biological network graphs and 25 social network graphs of diverse degree distributions and present the results of the Silhouette Index measure evaluated for the physical clusters and logical clusters obtained by running the Louvain algorithm (per the approaches described in the earlier sections). The biological networks analyzed are either gene–gene interaction networks, protein–protein interaction networks, interactions between different animal species in a particular area, etc. The social networks analyzed comprise acquaintance networks that capture the association between two users in a social network over a certain time period and friendship networks for which no such time period is used to capture association between two users.

In Fig. 3, we visually present some of the fundamental statistical information for the biological networks and social networks (for more details, see [12]). The number of nodes in the biological networks ranges from 62 to 2,640 with a median of 813. The number of nodes in the social networks ranges from 22 to 1,882 with a median of 75. The spectral radius ratio for node degree ($\lambda_{\text{sp}} \geq 1$) [13] for a network graph is the ratio of the principal eigenvalue [5] of the adjacency matrix for the graph and the average node degree; the larger the λ_{sp} value, the larger the variation in node degree. The edge density ($0 \leq \rho_{\text{edge}} \leq 1$) is calculated as the ratio of the actual number of edges in the network and the maximum possible number of edges in the network (which is $N(N-1)/2$ for a network of N nodes). The biological networks are characteristic of having larger λ_{sp} values and lower ρ_{edge} values; on the other hand, social networks are characteristic of having smaller λ_{sp} values and larger ρ_{edge} values. For more details on the individual real-world networks analyzed in each of these domains, the interested reader is referred to [12].

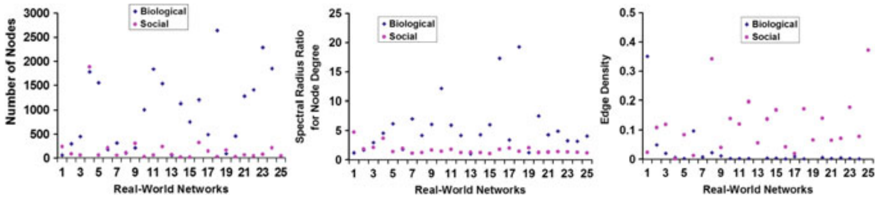
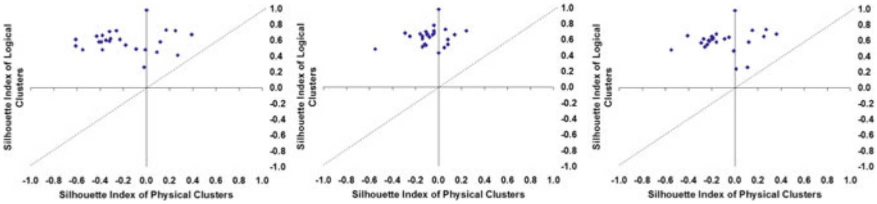


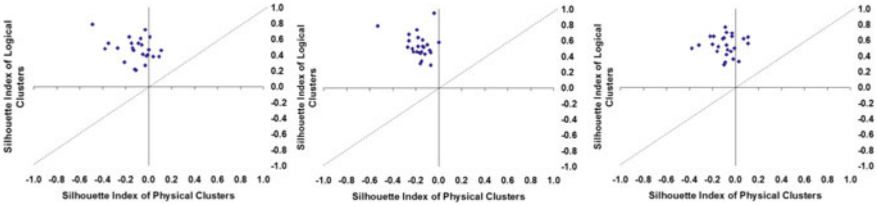
Fig. 3 Statistics for the biological and social networks

In Figs. 4 and 5, we present and visually compare the Silhouette Index values for the physical vs. logical clusters on the basis of the neighborhood-based (DEG, EVC); the shortest path-based (BWC, CLC), and all the four centrality metrics together (DEG, EVC, BWC, CLC). We observe all the data points to be above the dotted diagonal line, indicating that the Silhouette Index values for the logical clusters for all the real-world networks are appreciably larger than the Silhouette Index values for the physical clusters. For each coordinate system and for each network category, we measure (see Fig. 6a for a comparative bar chart) the average of the difference in the Silhouette Index values for the logical clusters versus physical clusters. We observe the biological networks to incur larger average difference in the Silhouette Index values for all the three coordinate systems. The (DEG, EVC) coordinate system incurs, respectively, the lowest (for social networks) and largest (for biological networks) values for the average difference in the Silhouette Indexes for the logical versus physical clusters. We also measure the median (see Fig. 6b for a comparative bar chart) of the Silhouette Index values for the physical clusters versus logical clusters. Through Figs. 4,5, and 6, we observe the Silhouette Index values for the physical (logical) clusters in the biological networks to be relatively lower (larger) than those in the social networks.



(DEG, EVC) (BWC, CLC) (DEG, EVC, BWC, CLC)

Fig. 4 Biological networks: Silhouette Index values for the physical versus logical clusters



(DEG, EVC) (BWC, CLC) (DEG, EVC, BWC, CLC)

Fig. 5 Social networks: Silhouette Index values for the physical versus logical clusters

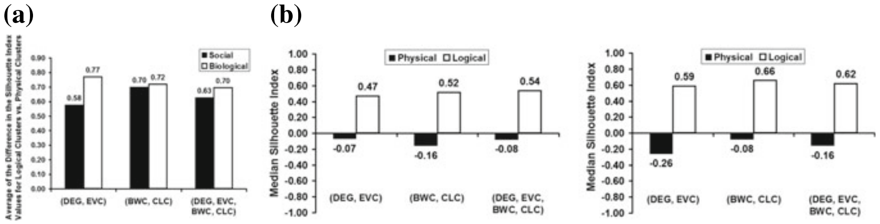


Fig. 6 Statistical comparison of the Silhouette Index values. **a** Avg. difference in the social networks biological networks Silhouette Index values. **b** Median of the Silhouette Index values

5 Conclusions

We show that logical clusters of vertices could be more cohesive with respect to node-level centrality metrics compared to physical clusters of vertices in complex real-world network graphs. In this pursuit, we adapt a recently proposed unit-disk graph approach (for node similarity assessment) [12] to determine such logical clusters of similar vertices. We applied the approach on a suite of 25 biological networks and 25 social networks and evaluated the extent of similarity of the vertices in the logical clusters versus physical clusters using the Silhouette Index measure with respect to three combinations of centrality metrics (DEG, EVC), (BWC, CLC), and (DEG, EVC, BWC, CLC). For each combination, we observe all the 50 real-world network graphs to incur significantly larger Silhouette Index values for the logical clusters compared to the physical clusters. We observe the biological networks to show a relatively larger difference in the Silhouette Index values between the logical clusters and physical clusters. Likewise, the (BWC, CLC) coordinate system has been observed to incur relatively larger Silhouette Index values for several real-world networks. In future, we plan to extend the unit-disk graph approach for outlier detection in complex networks and large datasets.

References

1. Newman MEJ (2010) Networks: an introduction, 1st edn. Oxford University Press, Oxford, UK
2. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99(12):7821–7826
3. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theor Exp P10008:1–11
4. Meghanathan N (2016) A greedy algorithm for neighborhood overlap-based community detection. Algorithms 9(1, 8):1–26
5. Bonacich P (1987) Power and centrality: a Family of measures. Am J Sociol 92(5):1170–1182
6. Freeman L (1977) A set of measures of centrality based on betweenness. Sociometry 40(1):35–41
7. Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25(2):163–177

8. Freeman L (1979) Centrality in social networks: conceptual clarification. Soc Netw 1(3):215–239
9. Cormen TH, Leiserson CE, Rivest RL, Stein C (2009) Introduction to algorithms. MIT Press, Cambridge
10. Meghanathan N (2016) Assortativity analysis of real-world network graphs based on centrality metrics. Comput Inform Sci 9(3):7–25
11. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput Appl Math 20:53–65
12. Meghanathan N (2019) Unit disk graph-based node similarity index for complex network analysis. Complexity. Article ID 6871874, p 22
13. Meghanathan N (2014) Spectral radius as a measure of variation in node degree for complex network graphs. In: The 3rd international conference on digital contents and applications, Hainan, pp 30–33