Hindawi
Complexity
Volume 2019, Article ID 6871874, 22 pages
https://doi.org/10.1155/2019/6871874

Research Article

Unit Disk Graph-Based Node Similarity Index for Complex Network Analysis

Natarajan Meghanathan

Professor of Computer Science, Jackson State University, Jackson, MS 39217, USA

Correspondence should be addressed to Natarajan Meghanathan; natarajan.meghanathan@jsums.edu

Received 10 December 2018; Revised 16 February 2019; Accepted 21 February 2019; Published 14 March 2019

Academic Editor: Ana Meštrović

We seek to quantify the extent of similarity among nodes in a complex network with respect to two or more node-level metrics (like centrality metrics). In this pursuit, we propose the following unit disk graph-based approach: we first normalize the values for the node-level metrics (using the sum of the squares approach) and construct a unit disk graph of the network in a coordinate system based on the normalized values of the node-level metrics. There exists an edge between two vertices in the unit disk graph if the Euclidean distance between the two vertices in the normalized coordinate system is within a threshold value (ranging from 0 to $\sqrt{k}$, where k is the number of node-level metrics considered). We run a binary search algorithm to determine the minimum value for the threshold distance that would yield a connected unit disk graph of the vertices. We refer to "1 – (minimum threshold distance/ $\sqrt{k}$)" as the *node similarity index* (NSI; ranging from 0 to 1) for the complex network with respect to the k node-level metrics considered. We evaluate the NSI values for a suite of 60 real-world networks with respect to both neighborhood-based centrality metrics (degree centrality and eigenvector centrality) and shortest path-based centrality metrics (betweenness centrality and closeness centrality).

1. Introduction

The weights assigned to nodes (a.k.a. vertices) in a complex network are either topology-based or domain-based or a combination of both. Centrality metrics quantify the topological importance of the nodes in a network [1]. There exist several centrality metrics, each proposed to capture a particular topological aspect; the four commonly studied centrality metrics are degree centrality (DEG), eigenvector centrality (EVC), betweenness centrality (BWC), and closeness centrality (CLC). While DEG and EVC could be categorized as neighborhood-based centrality metrics, BWC and CLC could be categorized as shortest path-based centrality metrics. More detailed information about these four centrality metrics and the procedures to individually compute them is available in [1]. Some of the examples for domain-based metrics are age, height, and weight of a patient (health information networks), number of publications and h-index of an author (citation networks), the processing capacity and the number of ports available for a router (communication networks), etc. Throughout the paper, the terms 'node' and 'vertex', 'link' and 'edge', and 'network' and 'graph' are used interchangeably. They mean the same.

Similarity assessment of nodes in complex networks has been so far conducted only at the node-level (e.g., [2–9]) and not at the network-level. To the best of our knowledge, all the similarity measures available in the literature quantify the extent of similarity between two nodes (like cosine similarity [10], matching index [11], etc.) or a set of nodes (the notion of equivalence classes [1], Rich club coefficient [12], etc.), but not among all the nodes in a network. It would not be appropriate to quantify the similarity among all the nodes in a network as a statistical function (like average or median) of the pair-wise similarity metric values. Also, the currently available similarity measures (like assortative index [11, 13]) use just one node-level metric (typically, the degree centrality metric) to assess the similarity between two vertices or a set of vertices. There is currently no quantitative measure available to rate the extent of similarity among all the vertices in a network with respect to a combination of node-level

metrics (topological metrics and/or domain-based metrics). In this paper, we seek to develop a "network-level" *node similarity index* (NSI) to comprehensively quantify the extent of similarity (in a scale of 0 to 1) among "all" the nodes in a network with respect to a set of node-level metrics.

We propose that two vertices are to be considered "similar" with respect to a set of node-level metrics if the vertices are "closer" on the basis of the Euclidean distance between their coordinates (represented by the normalized values of the node-level metrics for the vertices in the network). For example, let BWC and CLC be the two node-level metrics considered. Let there be four vertices v1, v2, v3 and v4 in the network whose normalized BWC values are 0.49, 0.62, 0.11, and 0.79, respectively, and normalized CLC values are 0.38, 0.42, 0.87, and 0.48, respectively. Then, the coordinates of the vertices v1, v2, v3, and v4 are given by (0.49, 0.38), (0.62, 0.42), (0.11, 0.87), and (0.79, 0.48), respectively, wherein the first entry in the coordinate tuples represents the normalized BWC values of the vertices and the second entry in the coordinate tuples represents the normalized CLC values. The Euclidean distance between vertices v1 and v2 is $\sqrt{(0.49-0.62)^2+(0.38-0.42)^2} = 0.136$ and the Euclidean distance between vertices v3 and v4 is $\sqrt{(0.11-0.79)^2+(0.87-0.48)^2} = 0.784$. According to our notion of similarity, vertices v1 and v2 are relatively more similar to each other, compared to vertices v3 and v4 with respect to BWC and CLC.

Our approach to determine the NSI for a network is briefly summarized below (more details are in Section 2). Given a network of nodes and edges and a set of node-level metrics of interest (let k be the number of node-level metrics considered), we first determine the raw values for the nodes with respect to each of the k node-level metrics and individually normalize them (using the sum of the squares approach). We then distribute the vertices in a k-dimensional coordinate system wherein the coordinate of a vertex is a tuple represented by the normalized values for the k node-level metrics. We seek to construct a unit disk graph of the vertices in the k-dimensional normalized coordinate system (the range of coordinate values for any dimension is 0 to 1) such that two vertices are connected with an edge if the Euclidean distance between them is within a threshold value. We run a binary search algorithm to determine the minimum value for this threshold distance so that the unit disk graph of the vertices in the k-dimensional normalized coordinate system is connected. Our hypothesis is that the closer the vertices in this coordinate system (i.e., more similar the vertices based on the node-level metric values), the smaller the value for the minimum threshold distance to obtain a connected unit disk graph. We hence propose the value for the node similarity index (NSI) to be 1 − (minimum threshold distance/$\sqrt{k}$), where $\sqrt{k}$ is the maximum distance between any two vertices in a coordinate system based on the normalized values of the k node-level metrics considered for similarity assessment.

Some of the applications we envision for the proposed NSI measure and the normalized coordinate system of the node-level metrics used to compute the measure are as follows: a communication network with a smaller NSI value is more likely to have a single point of failure (one or few routers would have more connections and through which more traffic flows compared to the rest) and is also more vulnerable for security attacks. A social network with a larger NSI value could be considered to comprise of users who are more peers/similar to each other. Health professionals may decide on coming up with a single treatment plan or different treatment plans for the patients depending on the NSI value (with respect to a set of node-level metrics) for a health information network; if the values for the health metrics for all the patients are similar (a larger NSI value), then a single treatment plan for all the patients might be a good choice to at least begin with. Further, we could run clustering algorithms on the unit disk graph corresponding to the NSI value for a network and determine clusters of "similar" vertices that need not be directly connected to each other. For example, we could identify the cluster/set of vertices that have similar values for the health parameters physically spread (but need not be connected) over a health information network. Finally, the proposed model of unit disk graph-based node similarity index could be applied for outlier detection: for any unlabeled dataset of features and their normalized values, we could construct a unit disk graph (to represent the dataset) wherein the vertices are the data points (rows) in the dataset with coordinates corresponding to the normalized feature values and two vertices are connected if the Euclidean distance between the two vertices is within a threshold distance. The NSI value for such a dataset would quantify the extent of similarity among the data points with respect to the feature values. Any vertex with a degree of one in the unit disk graph (especially with a larger NSI value) is a potential candidate for being classified as an outlier.

In Sections 3 and 4, we consider a total of 60 real-world networks for similarity assessment and determine their NSI values with respect to a combination of node-level metrics. Since these networks belong to different domains, we do not consider the domain-based metrics as node-level metrics in our assessment calculations. We consider only the topology-based centrality metrics (DEG, EVC, BWC, and CLC) as the node-level metrics for the similarity assessment tests conducted in this paper. The rationale behind the choice of the above four centrality metrics is that they are widely considered as representatives of neighborhood-based (DEG and EVC) and shortest path-based (BWC and CLC) centrality metrics as well as are considered "prototypical" metrics representing three of the four classes of centrality metrics (radial versus medial metrics and volume versus length-based metrics) [14]. The DEG and EVC metrics are radial metrics that capture the volume (number) of walks originating or terminating at a node. The BWC metric is a medial metric capturing the volume of walks passing through a node and the CLC metric is a radial metric capturing the length of the walks originating or terminating at a node. Nevertheless, the NSI measure could be computed for any combination and any number of domain-based and/or topology-based node-level metrics for a complex network.

The rest of the paper is organized as follows: Section 2 describes the proposed procedure to construct the unit disk

graph of the vertices based on a coordinate system comprising of the normalized values for the node-level metrics as well as explains the use of the binary search algorithm to determine the minimum threshold distance value that is required to obtain a connected unit disk graph; the section also analyzes the time complexity and memory space requirements of the binary search algorithm as well as illustrates the whole process using a toy network of eight vertices. Section 3 provides a brief overview of the 60 real-world networks used to evaluate the proposed unit disk graph-based NSI measure. Section 4 tabulates the results obtained for the NSI measure for the 60 real-world networks with respect to neighborhood-based centrality metrics and shortest path-based centrality metrics, considered separately as well as together. Section 4 compares the DEG-EVC NSI values and the BWC-CLC NSI values obtained for the real-world networks with that of the Pearson's correlation coefficient between these centrality metrics. Section 4 also compares the NSI values for the real-world networks based on a coordinate system of all the four centrality metrics with those of the NSI values of random networks with the same number of nodes and edges (generated using the well-known Erdos-Renyi model [15] and the Configuration model [16]); the purpose of this comparison is to highlight that the notion of node similarity captured by the unit disk graph-based NSI values is not a random phenomenon (unless the nodes in the real-world network are connected using randomly generated edges). Finally, we evaluate the correlation between the proposed NSI measure with that of recently proposed network-level measures (such as randomness index and spectral radius ratio for node degree) as well as classical network-level measures (such as assortative index and ratio of standard deviation to average path length) to showcase its uniqueness. Section 5 reviews the related work on similarity assessment in complex networks. Section 6 concludes the paper.

2. Node Similarity Index (NSI)

In this section, we describe the methodology to compute the proposed node similarity index (NSI) for a complex network. The NSI is a quantitative measure of the extent of similarity of the nodes in a complex network with respect to two or more node-level metrics. Let k be the number of node-level metrics considered for the similarity assessment. The sequence of steps to compute the NSI measure is first outlined below and then explained in detail. We use the graph shown in Figure 1 as a running example graph to illustrate the different steps in the procedure to compute the NSI measure.

(i) Compute the raw values of the k node-level metrics.

(ii) Normalize the values for each of the k node-level metrics using the sum of the squares approach.

(iii) Distribute the vertices in a k-dimensional coordinate system based on the normalized values for the node-level metrics.

(iv) Run a binary search algorithm to determine the minimum threshold distance that would be needed to obtain a connected unit disk graph of the vertices.

2.1. Raw Values for the Node-Level Metrics. As mentioned earlier, we use the centrality metrics as the basis to illustrate the procedure to compute the NSI for a network. Depending on the centrality metrics considered, we would need to use the appropriate algorithms to compute the (raw) values for each of these metrics for the vertices. In this paper, we consider the degree centrality (DEG), eigenvector centrality (EVC; [17]), betweenness centrality (BWC; [18, 19]), and closeness centrality (CLC; [20, 21]) for similarity assessment. The procedures to compute these metrics are available in several sources in the literature (e.g., [1]).

Here, we briefly outline the procedures, assuming the networks analyzed are modeled as undirected graphs and the edges are of unit weights:

(i) The DEG value for a vertex is simply the number of edges incident on the vertex.

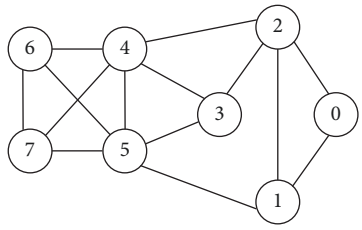
(ii) The EVC of a vertex is computed using the power-iteration method [17] according to which we start with a unit vector (all 1s) as the tentative principal eigenvector (that eventually has all the EVC values) and go through a series of iterations by multiplying (in each iteration) the adjacency matrix of the graph with the tentative principal eigenvector obtained in the previous iteration. At the end of an iteration, we normalize the entries in the resulting product vector (using the sum of the squares approach, see Section 2.2 for an example that illustrates this approach) and use the vector of normalized values as the tentative principal eigenvector for the next iteration. We stop the iterations when the values for the entries in the tentative principal eigenvector between two successive iterations converge to a certain level of precision.

(iii) The BWC of a vertex is obtained by running the Breadth First Search (BFS [21])-based version of the Brandes' algorithm [19]: we run the BFS algorithm at each vertex to determine the number of shortest paths from the vertex to every other vertex in the graph. Using this information, for each vertex, we determine the fractions of the number of shortest paths between any two vertices that go through the vertex and the sum of all these fractions is the BWC of the vertex.

(iv) The CLC of a vertex is basically the inverse of the sum of the shortest paths lengths (number of hops) from the vertex to every other vertex in the graph and is computed using the BFS algorithm.

2.2. Normalization of the Raw Values for the Node-Level Metrics. For each node-level metric, we normalize the raw values for the vertices and transform the values to a scale of 0 to 1. We use the sum of the squares approach for the normalization. As part of this process, we first obtain the square root of the sum of the squares of the raw node-level metric values of the vertices and then divide each of the raw values by this square root value.

For example, to obtain the normalized DEG values of the vertices in Figure 1, we first obtain the square root



Vertex	DEG	EVC	BWC	CLC
0	2	0.1535	0.00	0.0714
1	3	0.2518	2.33	0.0909
2	4	0.3264	4.00	0.1000
3	3	0.3396	0.33	0.0909
4	5	0.4842	4.33	0.1111
5	5	0.4685	5.00	0.1111
6	3	0.3444	0.00	0.0833
7	3	0.3444	0.00	0.0833

Figure 1: Example graph and the "raw" centrality values for the vertices.

of the sum of the squares of the raw DEG values, which is $\sqrt{2^2 + 3^2 + 4^2 + 3^2 + 5^2 + 5^2 + 3^2 + 3^2} = 10.29$. We then divide each of the raw DEG values by 10.29 to obtain the normalized DEG values of the vertices. Figure 2 displays the normalized centrality values of the vertices in the example graph.

2.3. Distribution of the Vertices in a Coordinate System of the Normalized Node-Level Metric Values. We now distribute the vertices in a coordinate system of the normalized values for the node-level metrics. Each node-level metric is considered as a dimension. If the number of node-level metrics considered is k, we basically distribute the vertices in a k-dimensional coordinate system of the normalized values for the node-level metrics. The coordinate for a vertex is represented as a tuple comprising of the normalized values for the k node-level metrics, which are centrality metrics in this paper.

For example, if all the four centrality metrics (DEG, EVC, BWC, CLC) are considered to form the coordinate system, the coordinate for vertex 0 in the example graph of Figures 1 and 2 would be (0.1943, 0.1535, 0.0000, 0.2696). For ease of presentation and visualization, we show the distribution of the vertices in the example graph using two dimensions at a time (see Figure 3): the neighborhood-based DEG and EVC metrics and the shortest path-based BWC and CLC metrics. As we can see, the distribution of the vertices is different in both the coordinate systems. Sometimes, it is possible that two or more vertices may be located at the same coordinate (like V6 and V7 in both the coordinate systems).

Just with a cursory look at the distributions of the vertices in the two coordinate systems of Figure 3, we could conclude that the vertices are more similar to each other with respect to DEG-EVC rather than BWC-CLC. We could also infer that vertices V3, V6, and V7 are more similar with respect to both DEG-EVC as well as BWC-CLC, even though V3 is not directly connected to V6 and V7. We could as well run some clustering algorithm to find clusters of similar vertices with respect to two or more centrality metrics.

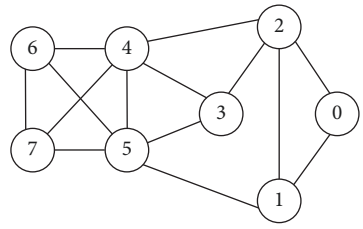
2.4. Binary Search Algorithm to Obtain a Unit Disk Graph with Minimum Threshold Distance. We now seek to construct a unit disk graph that could capture the similarity among the vertices in the coordinate system of the normalized values for the node-level metrics. In a k-dimensional coordinate system of the normalized values (in the range of 0 to 1), the maximum value for the distance between any two vertices is $\sqrt{k}$ (for example, the maximum distance between any two points in a unit square is $\sqrt{2}$) and the minimum value for the distance is of course 0. The binary search algorithm maintains three auxiliary variables: a left index, a right index, and a middle index. For any iteration, the middle index is the average of the left index and right index values at the beginning of the iteration and is more appropriately called the *threshold distance* for that iteration. During each iteration, we construct a unit disk graph of the vertices such that there exists an edge between two vertices if the Euclidean distance between the two vertices is less than or equal to the value of the threshold distance for the particular iteration. During any iteration, we maintain the invariant that the unit disk graph is guaranteed to be connected when the right index is used as the threshold distance and not connected when the left index is used as the threshold distance (unless all the vertices are colocated at the same coordinate). The procedural details of the binary search algorithm (see Algorithm 1 for the pseudo code) are as follows:

(i) To begin with, the left index is 0 and the right index is $\sqrt{k}$. We go through a sequence of iterations (during which the left index and right index approach each other) until the difference between the right index and left index is greater than or equal to ε; in this paper, we use $\varepsilon = 0.001$. In a particular iteration, either the left index moves to the right (i.e., is increased) or the right index moves to the left (i.e., is decreased).

(ii) At the beginning of each iteration, we compute the value for the middle index (threshold distance) as the average of the left index and right index that are updated at the end of the previous iteration.

(iii) As part of the iteration, we construct a unit disk graph of the vertices such that there exists an edge between two vertices if the Euclidean distance between the two vertices is less than or equal to the threshold distance. After constructing such a unit disk graph, we run the Breadth First Search (BFS) algorithm on the graph to check if it is connected or not.

  (a) If the unit disk graph constructed on the basis of the threshold distance for an iteration is connected, the final value for the minimum threshold distance should be greater than the left index, but less than or equal to the current threshold distance (middle index); accordingly,



Vertex	DEG	EVC	BWC	CLC
0	0.1943	0.1535	0.0000	0.2696
1	0.2914	0.2518	0.2887	0.3431
2	0.3885	0.3264	0.4949	0.3774
3	0.2914	0.3396	0.0412	0.3431
4	0.4856	0.4842	0.5361	0.4193
5	0.4856	0.4685	0.6186	0.4193
6	0.2914	0.3444	0.0000	0.3145
7	0.2914	0.3444	0.0000	0.3145

Figure 2: Example graph and the "normalized" centrality values for the vertices.

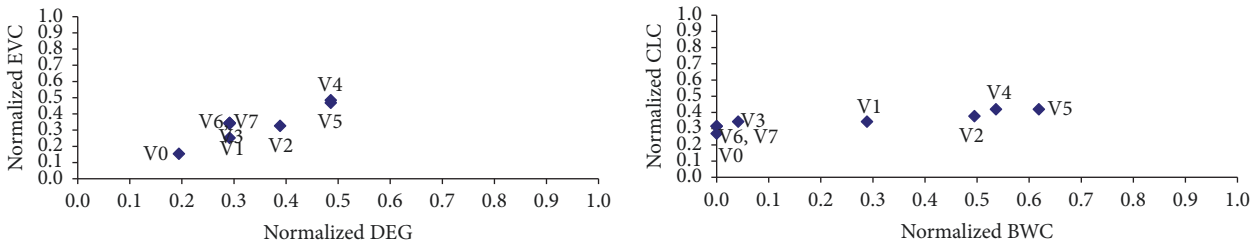


Figure 3: Distribution of the vertices (in the example graph of Figures 1 and 2) in a 2-dimensional coordinate system of the normalized centrality values (DEG-EVC and BWC-CLC).

we update (decrease) the value for the right index to be the current value of the middle index.

(b) If the unit disk graph constructed on the basis of the threshold distance for an iteration is not connected, the final value for the minimum threshold distance should be greater than the current threshold distance (middle index), but less than or equal to the right index; accordingly, we update (increase) the value for the left index to be the current value of the middle index.

(c) When the difference between the right index and left index becomes less than ε, we stop the iterations and consider the value for the right index during the last iteration as the value for the minimum threshold distance (since we always maintain the invariant that the unit disk graph for any iteration is connected when the right index is used as the threshold distance). The NSI value for the network is then simply computed as "1 − (minimum threshold distance/$\sqrt{k}$)".

(1) In a coordinate system based on the normalized values of k node-level metrics, the largest possible value for the minimum threshold distance will be $\sqrt{k}$ (when the vertices are the most dissimilar from each other) and the smallest possible value would be slightly above 0 (unless all the vertices are colocated/exactly similar). Hence, the above formulation of 1 − (minimum threshold distance/$\sqrt{k}$) would restrict the NSI values to a range of 0...1 such that larger the NSI value, the more similar are the vertices with respect to the metrics considered. Also, the division of the minimum threshold distance by $\sqrt{k}$ (where k is the number of dimensions: node-level metrics considered) would negate the impact of the number of node-level metrics considered for similarity assessment and capture the impact of the actual node-level metrics considered in their entirety. For example, with the above formulation, it is possible that the NSI value for a network with respect to (DEG, EVC, BWC, CLC) could end up being larger than the NSI value for a network with respect to (BWC, CLC) and be smaller than the NSI value for a network with respect to (DEG, EVC). That is, the significantly larger similarity among the vertices with respect to DEG and EVC could contribute to increasing the similarity among the vertices with respect to all the four centrality metrics and offset the relatively lower similarity among the vertices with respect to BWC and CLC.

(2) Note that we do not consider the value of the threshold distance (middle index) for the last iteration as the value for the minimum threshold distance because it might be the case that the unit disk graph of the last iteration was not connected for the threshold distance (middle index) of that iteration (see Table 1 for such a scenario).

2.5. Example to Illustrate the Working of the Proposed Binary Search Algorithm. Figure 4 illustrates the sequence of iterations of the binary search algorithm executed on the example

graph of Figures 1 and 2 with the coordinates of the vertices represented by the normalized values of DEG and EVC. As it is a 2-dimensional coordinate system, the initial value for the right index is $\sqrt{2}$ = 1.414. With the initial left index of 0, the initial value for the middle index is (0 + 1.414) / 2 = 0.707. The unit disk graph for the first iteration is constructed with 0.707 as the threshold distance and we notice the graph to be a connected graph (in this example, we actually see a complete graph wherein each vertex is connected to every other vertex). Hence, for the second iteration, we set the right index to be 0.707 and retain the left index as 0, leading to a new middle index value of (0 + 0.707) / 2 = 0.3535. The unit disk graph for this threshold distance (0.3535) value is also connected and we further reduce the search range by setting the right index to 0.3535. We continue the iterations by either increasing the left index or decreasing the right index. During the 12th iteration, we observe the difference between the right index and left index to be less than 0.001 (ε), and we finalize the value for the minimum threshold distance to correspond to the value for the right index during the 12th iteration. We use a precision of at most 6 decimal digits (if needed) for the threshold distance.

In Figure 4, along with the iteration #, we indicate the threshold distance (referred to as TD) used to obtain the unit disk graph for that iteration. Table 1 lists the values for the left index, right index and middle index (threshold distance) for each iteration as well as the difference between the values for the right index and left index and whether the unit disk graph for each iteration is connected or not. At the end of the 11th iteration, we notice that the difference between the right index and left index is less than 0.001 and we stop the iterations and conclude the value of the right index at the beginning of the iteration as the minimum threshold distance (0.172607) for the network under study. The NSI value for the network is then 1 − 0.172607 / $\sqrt{2}$ = 0.877948, where $\sqrt{2}$ corresponds to the number of node-level metrics (DEG, EVC) considered for the analysis. With a cursory look at the unit disk graph for the minimum threshold distance of 0.172607 (see It # 10 in Figure 4), one could conclude that there are three clusters of similar vertices with respect to DEG-EVC: V1, V2, V3, V6, and V7 form the largest cluster (actually a clique); V4 and V5 form another cluster, and V0 is on its own cluster.

2.6. Number of Iterations, Time Complexity, and Space Complexity. An interesting property of the binary search algorithm applied in the search space of $(0, ..., \sqrt{k}]$ is that the number of iterations of the algorithm for any real-world network just depends on the value of k (the number of node-level metrics/coordinates) and the parameter ε (we stop the algorithm if the difference between the right index and left index is less than ε) and does not depend on the actual number of nodes and edges as well as not on the actual values of the centrality/node-level metrics involved. Even if the range of searchable values in each iteration would vary with the real-world network and the centrality/node-level metrics involved, the size of the search space reduces by half in each iteration (a characteristic of the binary search approach). For example if $k = 2$: at the end of the first iteration, the search space is either $(0, ..., 0.707]$ or $(0.707, ..., 1.414]$; in either case, the size of the search space is 0.707. In a similar vein, at the end of the second iteration, the search space is either $(0, ..., 0.3535]$ or $(0.3535, ..., 0.707]$ or $(0.707, ..., 1.0605]$ or $(1.0605, ..., 1.414]$: the size of each of these search spaces is 0.3535. The size of the search space for the third iteration will be half of 0.3535 = 0.17675 and so on. With the size of the search space reducing by half in each iteration, the number of iterations needed for the search space to reduce from $\sqrt{k}$ to a value less than ε would be simply $\log_2^{\sqrt{k}/\epsilon}$ and will be independent of the centrality metrics and their values as well as independent of the actual number of nodes and edges in the real-world network analyzed.

The time complexity of the algorithm is dominated by the time to construct the logical graph G_L (based on the normalized centrality values of the vertices as coordinates) for each iteration, which would be of complexity $O(V^2)$ for a real-world network of V nodes. The possibility of an edge between any two vertices in the real-world network needs to be evaluated, and hence the time complexity to construct the logical graph G_L will be $O(V^2)$. After the logical graph G_L is constructed during an iteration, we would need to check for its connectivity to decide whether to change the left index or right index for the next iteration. The Breadth First Search or Depth First Search algorithms of time complexity $O(V + E)$ could be used for this purpose. Putting together the number of iterations and the time complexity for each iteration, the overall time complexity of the proposed binary search algorithm(run for k node-level metrics with a terminating search space size of ε) for a given graph of V vertices is $O(V^2 * \log_2^{\sqrt{k}/\epsilon})$.

With regard to space complexity, for each iteration, the algorithm constructs a logical graph G_L (a data structure) and checks for its connectivity. As mentioned above, the number of edges in G_L would be $O(V^2)$, where V is the number of vertices in the graph G_L. Note that the logical graph constructed during an iteration could be cleared from memory at the end of the iteration. Also, the number of auxiliary variables used remains the same irrespective of the size of the real-world network graph analyzed. Hence, the memory requirements of the algorithm is $O(V^2)$, where V is the number of vertices in the real-world network graph G_R as well.

3. Overview of the Real-World Networks Used for Analysis

In this section, we provide a brief overview of the 60 real-world networks that are analyzed for the proposed node similarity index (NSI) measure. The real-world networks are spread over several domains, such as (listed below along with the number of networks considered for each domain): acquaintance network (12), friendship network (9), biological network (8), coappearance network (8), citation network (4), employment network (4), collaboration network (3), literature network (3), political network (3), communication

Table 1: Iteration Details for the Execution of the Binary Search Algorithm Illustrated in Figure 4.

Left Index	Right Index	Right Index - Left Index	It #	Middle Index (Threshold Distance)	Connectivity of the Unit Disk Graph
0	1.414	1.414	1	0.707	Connected
0	0.707	0.707	2	0.3535	Connected
0	0.3535	0.3535	3	0.17675	Connected
0	0.17675	0.17675	4	0.088375	Not connected
0.088375	0.17675	0.088375	5	0.132562	Not connected
0.132562	0.17675	0.044188	6	0.154656	Not connected
0.154656	0.17675	0.022094	7	0.165703	Not connected
0.165703	0.17675	0.011047	8	0.171226	Not connected
0.171226	0.17675	0.005524	9	0.173988	Connected
0.171226	0.173988	0.002762	10	0.172607	Connected
0.171226	0.172607	0.001381	11	0.171917	Not connected
0.171917	0.172607	0.00069 < 0.001 STOP!!!			

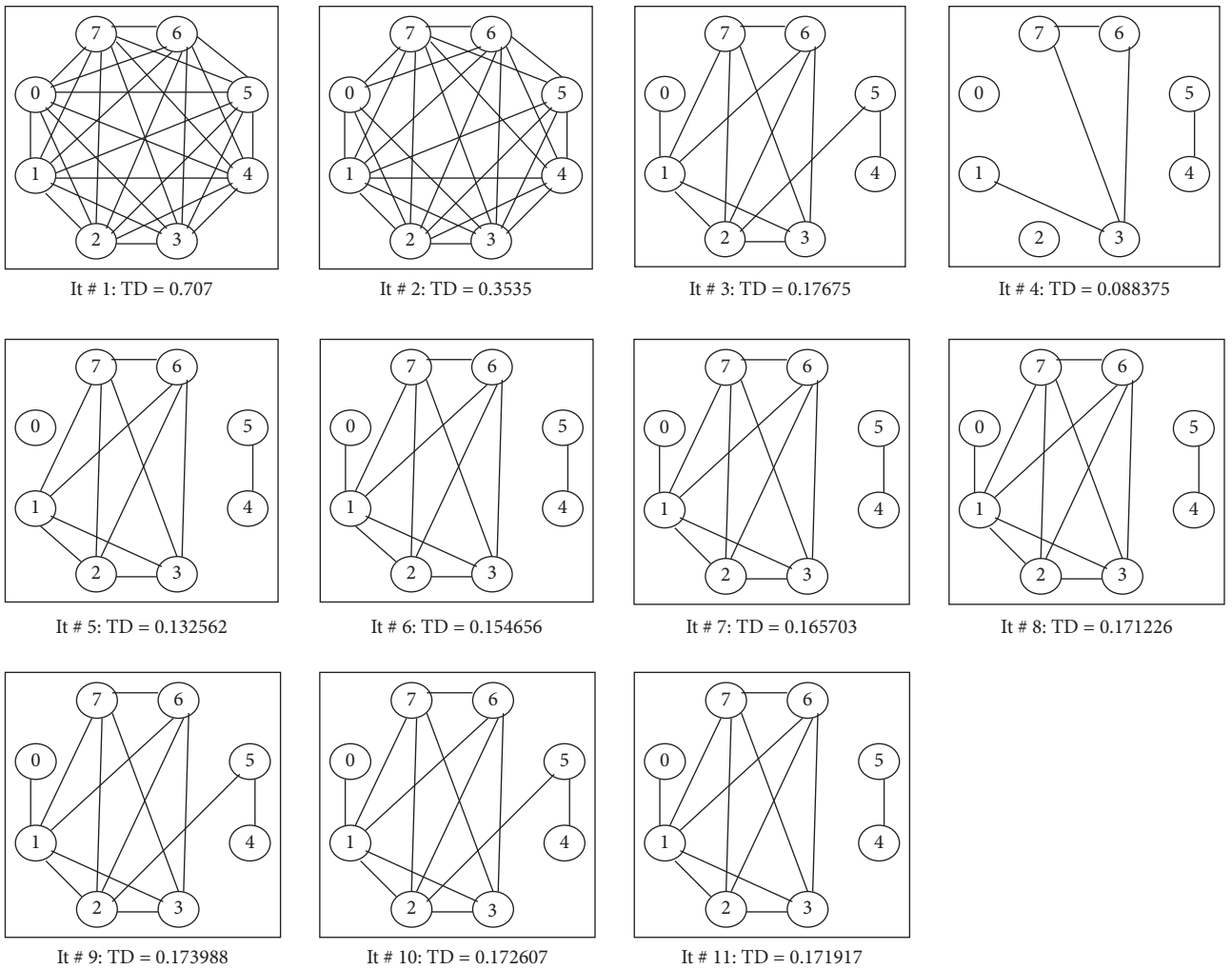


Figure 4: Unit disk graphs constructed during the iterations of the binary search algorithm for the example graph of Figures 1 and 2 and the normalized coordinate system of (DEG, EVC).

network (2), game network (2), and transportation network (2). We now briefly describe these networks: an *acquaintance network* is a social network comprising of people who are not close to each other, but slightly know each other (like an acquaintance) that is typically learnt during an observation period. A *friendship network* is a social network in which the participant nodes closely know each other, and no observation period is typically used to learn about the friendships. A *biological network* is a network that models the interactions involving genes, proteins and the associated transcriptions as well as models the interactions between animals of a species, etc. A *coappearance network* is a network based on the appearance of characters or words (extracted from novels/books/dictionary) alongside each other. A *citation network* is a network in which there exists a link between two nodes (papers) if one of the two papers has cited the other paper as reference. An *employment network* is a network in which the interactions between employees (nodes) are due to the job requirements and not due to any personal liking. A *collaboration network* is a network of authors who are linked if two authors share at least one publication. A *literature network* is a network of books/papers/terminologies/authors (other than citation, collaboration or coauthorship) in a particular area of literature. A *political network* is a network of entities (typically politicians) involved in politics. A *communication network* is a network of entities that communicate in an organizational setting or over a common agenda (e.g., email network, criminal network, trade network, etc.). A *game network* is a network of teams or players playing for different teams and their associations. A *transportation network* is a network of entities (like airports and their flight connections) involved in public transportation. In Table 2, we list the 60 real-world networks, their 3 character-code acronym used in the paper, the domain of the network as well as the number of nodes, edges, average degree and the spectral radius ratio for node degree (a measure of the variation in node degree, with a minimum value of 1.0; [22]). In a recent work [23], we had analyzed these 60 real-world networks for assortativity with respect to the neighborhood and shortest path-based centrality metrics and observed the

real-world networks to be more assortative with respect to EVC and CLC and more disassortative with respect to BWC and DEG.

4. Node Similarity Index of the Real-World Networks

In this section, we present the results obtained for the proposed node similarity index (NSI) measure for the 60 real-world networks with respect to neighborhood-based centrality metrics (DEG, EVC), shortest path-based centrality metrics (BWC, CLC), and both the neighborhood and shortest path-based centrality metrics (DEG, EVC, BWC, CLC) forming the coordinate systems. With a ϵ value of 0.001, the number of iterations incurred (for any complex network) by the binary search algorithm with two and four centrality metrics used for the coordinate systems are respectively $|\log_2^{\sqrt{2}/0.001}| = 11$ and $|\log_2^{\sqrt{4}/0.001}| = 11$. The median of the NSI values for the (DEG, EVC), (BWC, CLC), and (DEG, EVC, BWC, CLC)-based coordinate systems is 0.92, 0.89, and 0.89, respectively.

Table 3 presents the numerical NSI values for the real-world networks with respect to all the three coordinate systems. For domains that have at least 5 real-world networks (there are four such domains), we group the networks together to present the results in Table 3. For each of these four domains (acquaintance networks, friendship networks, biological networks, and coappearance networks), we make the numbers bold for which the NSI value for a particular coordinate system is greater than or equal to the median value for all the real-world networks with respect to the same coordinate system. For example, we make the numbers bold for a (DEG, EVC) coordinate system if the NSI value in the cell is greater than or equal to 0.92. Based on this coloring scheme, we introduce a measure called *relative node similarity score for a network domain* that is computed as the ratio of the bold numbers in the domain divided by the total number of cells in that domain across all the three coordinate systems. For example, in the case of acquaintance networks, there are 21 bold numbers in a total of 36 cells and hence the *relative node similarity score for acquaintance networks* (in comparison to any real-world network; with respect to any coordinate system) is 21/36 = 0.58. Likewise, the relative similarity scores of the Friendship networks, Biological networks and coappearance networks are respectively: 19/27 = 0.70, 11/24 = 0.46 and 5/24 = 0.21. We can thus infer that the nodes in friendship and acquaintance networks are more likely to be similar to each other with respect to the centrality metrics compared to the nodes in the biological and coappearance networks. Nodes in a coappearance network (especially, when it involves the appearance of characters in the same chapter/scene) are less likely to be similar to each other with respect to the centrality metrics.

A visual comparison of the NSI values for the three coordinate systems is presented in Figures 5(a)–5(c). For 43 of the 60 real-world networks (i.e., more than 70% of the networks), the (DEG, EVC)-based NSI values are greater than the (BWC, CLC)-based NSI values (see Figure 5(a)). Hence, nodes in real-world networks are more likely to be similar with respect to the neighborhood-based (DEG, EVC) centrality metrics rather than the shortest path-based (BWC, CLC) centrality metrics. A notable exception to this trend is the Roget Network (#49: ROG) whose (DEG, EVC)-based NSI is 0.57 and (BWC, CLC)-based NSI is 0.88. In Figures 5(b) and 5(c), when the (DEG, EVC)-based NSI values and the (BWC, CLC)-based NSI values are plotted against the (DEG, EVC, BWC, CLC)-based NSI values, we observe the (DEG, EVC, BWC, CLC)-based NSI values are lower than that of the (DEG, EVC)-based NSI values for more than 85% (i.e., for 52/60) of the real-world networks; on the other hand, the (DEG, EVC, BWC, CLC)-based NSI values are greater than that of the (BWC, CLC)-based NSI values for more than 50% (i.e., for 32/60) of the real-world networks. The relatively larger similarity among the vertices with respect to (DEG, EVC) contributes to the larger values for the (DEG, EVC, BWC, CLC)-based NSI measure compared to the (BWC, CLC)-based NSI measure. As a result, nodes in real-world networks tend to be more similar to each other when both the neighborhood-based (DEG, EVC) and shortest path-based (BWC, CLC) centrality metrics are considered together rather than when the shortest path-based (BWC, CLC) centrality metrics are considered alone. This corroborates our earlier assertion in Section 2.4 that our formulation for NSI as "1 − (minimum threshold distance/$\sqrt{k}$)" negates the number of node-level metrics (k) considered and captures the contribution of the node-level metrics in their entirety to quantify the extent of similarity among the vertices.

4.1. Comparison of NSI Values with the Pearson's Correlation Coefficient of the Centrality Metrics. Correlation studies involving centrality metrics have been extensively conducted in the literature (e.g., [70–72]), with the Pearson's correlation coefficient [73], whose values range from -1 to 1, being the most commonly used correlation measure. A larger positive value (or a smaller negative value) for the Pearson's correlation coefficient between two centrality metrics means that the two centrality metrics are strongly and positively (or negatively) related as well as one centrality metric could be predicted using a linear function of the other centrality metric (e.g., [74]). If the Pearson's correlation coefficient between two centrality metrics is closer to 0, it implies the two metrics are not linearly related to each other.

In this subsection, we compare the NSI values obtained for the real-world networks based on the neighborhood (DEG, EVC)-based centrality metrics and the shortest path (BWC, CLC)-based centrality metrics with the Pearson's correlation coefficient values for DEG versus EVC and BWC versus CLC for these networks (see Figure 6). The purpose of this comparison is to showcase that the NSI values based on a coordinate system of a particular combination of centrality metrics are independent of the correlation between the corresponding centrality metrics. Thereby, we claim that the correlation coefficient between two centrality metrics for a real-world network cannot be construed as a network-level measure of the extent of similarity among the nodes in the network.

Table 2: Real-World Networks used for Node Similarity Assessment [Table 2 is reproduced from Meghanathan (2018): [23], (under the Creative Commons Attribution License/public domain)].

#	Net.	Net. Description	Ref.	Network Type	λ_{sp}	#nodes	#edges	k_{avg}
1	TEN	Taro Exchange Network	[24]	Acquaintance Network	1.06	22	39	3.55
2	SSM	Sawmill Strike Comm. Net.	[25]	Acquaintance Network	1.22	24	38	3.17
3	KCN	Karate Club Network	[26]	Acquaintance Network	1.47	34	78	4.59
4	DLN	Dutch Literature 1976 Net.	[27]	Literature Network	1.49	37	81	4.38
5	MPN	Mexican Political Elite Net.	[28]	Political Network	1.23	35	117	6.69
6	SWC	Soccer World Cup 1998 Net	[29]	Game Network	1.45	35	118	6.74
7	FHT	Friendship in Hi-Tech Firm	[30]	Friendship Network	1.57	33	91	5.52
8	MMN	ModMath Network	[29]	Friendship Network	1.59	30	61	4.07
9	KFP	Korea Family Planning Net.	[31]	Acquaintance Network	1.70	37	85	4.59
10	WSB	Windsurfers Beach Network	[32]	Friendship Network	1.22	43	336	15.63
11	FTC	Flying Teams Cade Net.	[33]	Employment Network	1.21	48	170	7.08
12	TWF	Teenage Female Friend Net.	[34]	Friendship Network	1.49	47	77	3.28
13	CDF	College Dorm Fraternity Net	[35]	Acquaintance Network	1.11	58	967	33.35
14	DON	Dolphin Network	[36]	Acquaintance Network	1.40	62	159	5.13
15	MDN	Macaque Dominance Net.	[37]	Biological Network	1.04	62	1167	37.65
16	PFN	Prison Friendship Network	[38]	Friendship Network	1.32	67	142	4.24
17	MTB	Madrid Train Bombing Net.	[39]	Acquaintance Network	1.95	64	295	9.22
18	GLN	Graph Glossary Network	[29]	Literature Network	2.01	67	118	3.52
19	HCN	Huckleberry Coappear. Net.	[40]	Co-appearance Network	1.66	76	302	7.95
20	SJN	San Juan Sur Family Net.	[41]	Acquaintance Network	1.29	75	155	4.13
21	LMN	Les Miserables Network	[40]	Co-appearance Network	1.82	77	254	6.59

Table 2: Continued.

#	Net.	Net. Description	Ref.	Network Type	λ_{sp}	#nodes	#edges	k_{avg}
22	MCE	Manufact. Comp. Empl. Net.	[42]	Employment Network	1.12	77	1549	40.23
23	WTN	World Trade Metal Network	[43]	Communication Network	1.38	80	875	21.88
24	UKF	UK Faculty Friendship Net.	[44]	Friendship Network	1.35	83	578	13.93
25	CFN	Copperfield Network	[40]	Co-appearance Network	1.83	89	407	9.15
26	SPR	Senator Press Release Net.	[45]	Political Network	1.57	92	477	10.37
27	PBN	US Politics Books Network	[46]	Literature Network	1.42	105	441	8.40
28	ADJ	Word Adjacency Network	[47]	Co-appearance Network	1.73	112	425	7.59
29	HTN	Hypertext 2009 Network	[48]	Acquaintance Network	1.21	115	2164	37.64
30	FON	US Football Network	[49]	Game Network	1.01	115	613	10.66
31	CLN	Centrality Literature Net.	[50]	Citation Network	2.03	118	613	10.39
32	AKN	Anna Karnenina Network	[40]	Co-appearance Network	2.48	140	494	7.06
33	MUN	Marvel Universe Network	[51]	Co-appearance Network	2.54	167	301	3.61
34	GD96	Graph Drawing 1996 Net	[29]	Citation Network	2.38	180	228	2.53
35	AFB	Author Facebook Network	-	Friendship Network	2.29	171	940	10.99
36	JBN	Jazz Band Network	[52]	Employment Network	1.45	198	2742	27.69
37	FMH	Faux Mesa High School Net	[53]	Friendship Network	2.81	147	202	2.75
38	RHF	Residence Hall Friend Net.	[54]	Friendship Network	1.27	217	1839	16.95
39	PSN	Primary School Contact Net.	[55]	Acquaintance Network	1.22	238	5539	46.55
40	SDI	Scotland Corp. Interlock Net	[56]	Employment Network	1.94	230	359	3.12

Table 2: Continued.

#	Net.	Net. Description	Ref.	Network Type	λ_{sp}	#nodes	#edges	k_{avg}
41	CEN	C. Elegans Neural Network	[57]	Biological Network	1.68	297	2148	14.47
42	DRN	Drug Network	[58]	Acquaintance Network	2.76	212	284	2.68
43	ISP	Infectious Socio-Patterns Net	[48]	Acquaintance Network	1.69	309	1924	12.45
44	CGD	Citation Graph Drawing Net	[59]	Citation Network	2.24	259	640	4.94
45	APN	US Airports 1997 Network	[29]	Transportation Network	3.22	332	2126	12.81
46	ERD	Erdos Collaboration Net.	[29]	Collaboration Network	3.00	433	1314	6.07
47	MSJ	Soc. Net. Journal Co-authors	[60]	Collaboration Network	3.48	475	625	2.63
48	HIV	Human HIV Gen. Inter. Net.	[61]	Biological Network	6.16	1005	1189	2.37
49	ROG	Roget Network	[40]	Co-appearance Network	1.68	1022	3648	7.14
50	MTN	Mouse Transcription Net.	[62]	Biological Network	4.30	1130	2403	4.03
51	RVU	RVU Email Network	[63]	Communication Network	2.16	1133	10903	9.62
52	ERN	European Union Road Net	[64]	Transportation Network	1.66	1174	1417	2.41
53	PGI	Plasmodium Gen. Inter. Net.	[65]	Biological Network	7.28	1203	1205	2.00
54	YIN	Yeast Interactome Network	[66]	Biological Network	4.51	1278	1809	2.70
55	YPN	Yeast Phosphorylation Net.	[62]	Biological Network	4.90	1407	4083	5.79
56	NDN	Norwegian Director Network	[67]	Co-appearance Network	2.10	1421	7710	5.43
57	PBL	Political Blogs Network	[68]	Political Network	3.30	1490	16715	22.44
58	LCI	Literature Curated Inter. Net.	[69]	Biological Network	4.23	1536	2925	3.76
59	JDN	Java Dependency Network	[29]	Citation Network	4.25	1538	8032	10.17
60	NSC	Network Sci. Co-author Net.	[47]	Collaboration Network	5.51	1589	2742	3.45

Table 3: NSI Values for the Real-World Networks.

#	Net.	(DEG, EVC)	(BWC, CLC)	(DEG, EVC, BWC, CLC)	#	Net.	(DEG, EVC)	(BWC, CLC)	(DEG, EVC, BWC, CLC)
		Acquaintance Network					*Co-appearance Network*		
1	TEN	**0.94**	**0.92**	**0.93**	19	HCN	0.76	0.47	0.59
2	SSM	0.89	0.73	0.80	21	LMN	0.87	0.57	0.68
3	KCN	0.91	0.81	0.84	25	CFN	0.72	0.38	0.52
9	KFP	0.90	**0.89**	**0.90**	28	ADJ	0.86	0.76	0.81
13	CDF	**0.98**	**0.96**	**0.97**	32	AKN	0.89	0.77	0.81
14	DON	**0.95**	**0.90**	**0.90**	33	MUN	0.90	0.72	0.74
17	MTB	**0.93**	**0.92**	0.88	49	ROG	**0.98**	**0.98**	**0.97**
20	SJN	0.88	0.86	0.85	56	NDN	0.89	**0.94**	**0.91**
29	HTN	**0.97**	0.75	0.82			*Other Networks*		
39	PSN	**0.99**	**0.96**	**0.97**	4	DLN	0.93	0.93	0.91
42	DRN	0.87	**0.94**	0.88	5	MPN	0.90	0.85	0.85
43	ISP	**0.97**	**0.90**	**0.92**	6	SWC	0.90	0.89	0.88
		Friendship Network			11	FTC	0.94	0.90	0.92
7	FHT	**0.92**	**0.88**	**0.90**	18	GLN	0.88	0.88	0.88
8	MMN	0.81	0.85	0.80	22	MCE	0.96	0.84	0.88
10	WSB	**0.97**	**0.89**	**0.92**	23	WTN	0.97	0.87	0.91
12	TWF	0.91	**0.94**	**0.92**	26	SPR	0.91	0.92	0.91
16	PFN	**0.94**	**0.91**	**0.92**	27	PBN	0.96	0.93	0.93
24	UKF	**0.95**	0.82	0.84	30	FON	0.99	0.98	0.98
35	AFB	**0.97**	0.88	**0.91**	31	CLN	0.89	0.83	0.86
37	FMH	**0.92**	**0.89**	0.78	34	GD96	0.90	0.88	0.88
38	RHF	**0.97**	**0.90**	**0.93**	36	JBN	0.96	0.73	0.78
		Biological Network			40	SDI	0.86	0.92	0.89
15	MDN	**0.98**	**0.96**	**0.97**	44	CGD	0.97	0.92	0.94
41	CEN	0.88	0.48	0.62	45	APN	0.96	0.89	0.89
48	HIV	0.57	0.68	0.62	46	ERD	0.96	0.97	0.95
50	MTN	0.89	**0.93**	**0.90**	47	MSJ	0.88	0.94	0.91
53	PGI	0.29	0.29	0.29	51	RVU	0.94	0.98	0.94
54	YIN	0.73	0.74	0.73	52	ERN	0.95	0.98	0.93
55	YPN	**0.94**	**0.92**	**0.92**	57	PBL	0.96	0.93	0.92
58	LCI	**0.96**	**0.95**	**0.89**	59	JDN	0.78	0.75	0.77
					60	NSC	0.85	0.95	0.89

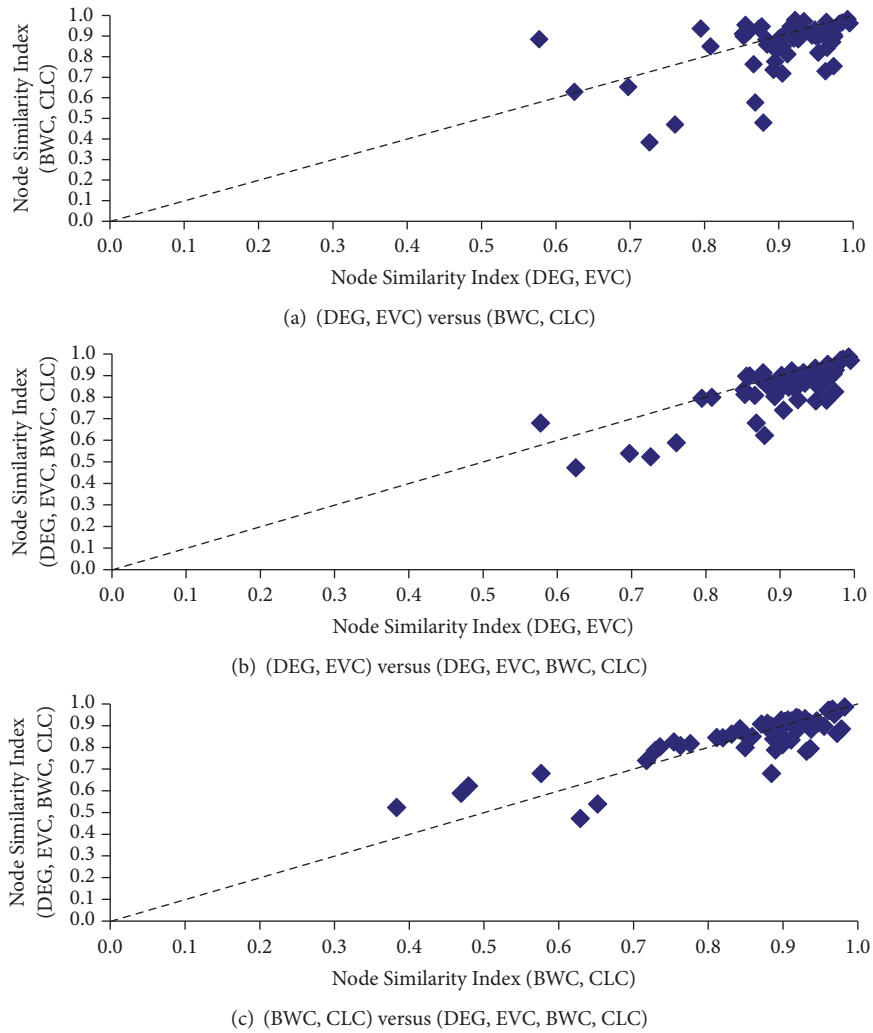


Figure 5: Comparison of the NSI values for the real-world networks based on the coordinate systems.

The plots in Figure 6 for both the neighborhood and shortest path-based centrality metrics indicate that the NSI values for the real-world networks based on the coordinate systems of these centrality metrics are independent of the Pearson's correlation coefficient between the constituent centrality metrics for the real-world networks. Though the Pearson's correlation coefficient values range from -1 to 1 (for DEG, EVC) or from 0 to 1 (for BWC, CLC), the NSI values for most of the real-world networks are 0.85 or above (for DEG, EVC) or 0.80 or above (for BWC, CLC). We could not identify any sort of relationship between the NSI values and the correlation coefficients.

Numerically, the (DEG, EVC)-based NSI values are greater than the Pearson's correlation coefficient between DEG and EVC for about 2/3rds of the real-world networks, with the median of the difference being 0.12; on the other hand, for the other 1/3rd of the real-world networks (for which the Pearson's correlation coefficients between DEG and EVC are relatively larger than the NSI values for the networks based on these two metrics), the median of the difference in the values is only 0.04. Though DEG and EVC are positively correlated for a majority of the real-world networks, the Pearson's correlation values between DEG and EVC are negative (-0.5 or lower) for the following four networks: Marvel Universe Network (#33: MUN), Author Facebook Network (#35: AFB), Yeast Phosphorylation Network (#55: YPN) and Network Science Coauthorship Network (#60: NSC). In the case of (BWC, CLC), the NSI values are larger than the Pearson's correlation coefficient between BWC and CLC for more than 85% of the real-world networks, with the median of the difference being 0.43. Thus, the (DEG, EVC)-based NSI values are relatively more closer to the Pearson's correlation coefficients between DEG and EVC compared to the proximity of the (BWC, CLC)-based NSI values to the Pearson's correlation coefficients between BWC and CLC.

4.2. Comparison of the NSI Values for the Real-World Networks and Random Networks. In this subsection, we compare the NSI values for the real-world networks with the NSI values

Inputs
Real-world network graph, G_R
Number of centrality metrics, k
The normalized k centrality values $(C_1, C_2, \ldots, C_k)$ for each vertex in G_R
// The centrality-based logical coordinates for a vertex i is represented as $(C_1^i, C_2^i, \ldots, C_k^i)$
Auxiliary Variables
Left Index = 0, Right Index = $\sqrt{k}$, Middle Index, $\varepsilon = 0.001$
Begin Binary Search Algorithm
while (| Right Index - Left Index | > ε) **do**
Middle Index = (Left Index + Right Index) / 2
Construct Logical Graph G_L for the vertices using the Middle Index as the threshold distance
/∗ Two vertices i and j in G_R are connected with an edge in G_L if the Euclidean distance $\sqrt{(C_1^i - C_1^j)^2 + (C_2^i - C_2^j)^2 + \cdots + (C_k^i - C_k^j)^2} \leq$ Middle Index ∗/
if (G_L is connected) **then**
Right Index = Middle Index
else
Left Index = Middle Index
end if
end while
return NSI = (1 − Right Index) / $\sqrt{k}$
End Binary Search Algorithm

Algorithm 1: Pseudo Code for the Binary Search Algorithm to Determine the Node Similarity Index (NSI).

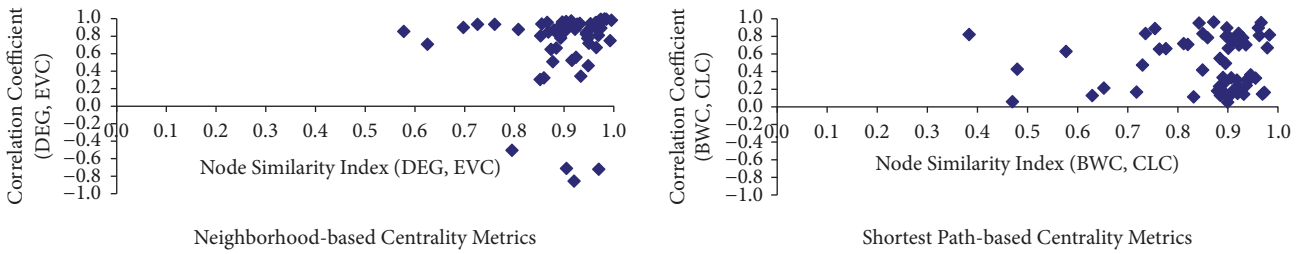


Figure 6: NSI values versus Pearson's correlation coefficient values for the centrality metrics.

obtained for random networks generated using the well-known Erdos-Renyi [15] and Configuration [17] models. For a given real-world network, both the models generate a random network with the same number of vertices and edges, but the edges between the vertices are randomly assigned. The degree distribution of the vertices in the random network generated using the Configuration model will be the same as the degree distribution of the vertices in the corresponding real-world network. On the other hand, the degree distribution of the vertices in the random network generated using the Erdos-Renyi model will always be Poisson in nature, irrespective of the degree distribution of the vertices in the corresponding real-world network. We expect relatively less variation in the centrality values of the nodes in the random network generated using the Erdos-Renyi model compared to those generated using the Configuration model. Nevertheless, our hypothesis is that since the edges are randomly assigned under both these models, the NSI values of the random networks with respect to any combination of centrality metrics should be different from the NSI values of the corresponding real-world networks.

For each of the 60 real-world networks, we generated hundred random networks according to each of the above two models. For a real-world network with N nodes and L links, to generate a random network per the Erdos-Renyi model, we first determine the probability ($p_{link} = L/(N(N-1)/2)$) for a link between any two nodes in the random network; we then consider all possible node pairs of two different vertices and generate a random number for each pair. If the random number generated for a node pair is less than or equal to p_{link}, there is an edge between the two nodes in the random network; otherwise, not. To generate a random network according to the Configuration model, we first determine the degree sequence of the vertices in the corresponding real-world network. We set up a list LD that has the vertex IDs such that the number of times a vertex is included in this list corresponds to the degree of the vertex in the real-world network. We then randomly shuffle the vertices in the list LD ten times (to decrease the chances of the same vertex ID appearing consecutively). Finally, we sequentially parse through the shuffled list and connect the adjacent vertices in the list with an edge. For complex real-world networks with a larger number of nodes, the average

number of self-loops and multilinks in the random networks generated according to the Configuration model is a constant and their density approaches zero as the number of nodes tends to infinity [75].

After rigorous simulations for a coordinate system based on all the four major centrality metrics (DEG, EVC, BWC, CLC), we observe our hypothesis to be indeed true. For 85% and 63% of the real-world networks (i.e., 52 and 38 of the 60 networks), the average of the NSI values for the random networks generated respectively according to the Erdos-Renyi model and the Configuration model are greater than 0.90. Unlike the corresponding random networks, for only 19 of the 60 real-world networks (i.e., less than 1/3rd of the real-world networks), the (DEG, EVC, BWC, CLC)-based NSI values are greater than 0.90. The relatively larger NSI values for the random networks per the Erdos-Renyi (ER) model vis-a-vis the Configuration model could be attributed to the lower variation in the values of the centrality metrics of the vertices in the ER-random networks that exhibit a Poisson degree distribution.

Figure 7 shows the distribution of the (DEG, EVC, BWC, CLC)-based NSI values of the real-world networks and the average of the (DEG, EVC, BWC, CLC)-based NSI values for the corresponding random networks generated according to the Erdos-Renyi model (Figure 7(a)) and the Configuration model (Figure 7(b)). We do not see any relationship between the two NSI values in each of Figures 7(a) and 7(b), indicating that the NSI values measured for a real-world network are not random and they do capture the extent of similarity among the nodes with respect to the centrality metrics considered. The median of the difference in the NSI values for a real-world network and the random network generated per the Configuration model is 0.06 and the random network generated per the Erdos-Renyi model is 0.10.

For only nine of the sixty real-world networks, the NSI value for the real-world network is greater than the average of the NSI values for the corresponding random networks (per the Erdos-Renyi model). These nine real-world networks are as follows: Taro Exchange Network (#1: TEN), Friendship Network in a Hi-Tech Firm (#7: FHT), Windsurfers Beach Network (#10: WSB), College Dorm Fraternity Network (#13: CDF), Macaque Dominance Network (#15: MDN), Manufacturing Company Employee Network (#22: MCE), World Trade Metal Network (#23: WTN), US Football Network (#30: FON), and Primary School Contact Network (#39: PSN). The values for the spectral radius ratio for node degree for these nine real-world networks range from 1.01 to 1.57 with a median of 1.12. Real-world networks with such a low spectral radius ratio for node degree could be indeed considered to be randomly generated.

4.3. Comparison of the NSI Values with the Values for Other Network-Level Measures. In this subsection, we compare the NSI values obtained for the real-world networks with those of other recently proposed and classical network-level measures. These measures are (i) spectral radius ratio for node degree; (ii) randomness index; (iii) assortative index, and (iv) ratio of the standard deviation to the average path length. Below, we provide a brief description of each of these measures and analyze the relationship vis-a-vis the appropriate coordinate system-based NSI values with which we compare them:

(i) The *spectral radius ratio for node degree* [22] quantifies the extent of variation in node degree in a way that is independent of the number of nodes and edges in the network (unlike the classical standard deviation measure that is dependent on the number of nodes). The spectral radius ratio for node degree is computed as the ratio of the principal eigenvalue of the adjacency matrix and the average node degree. The smallest possible value for the measure is 1.0 and it corresponds to a regular network where there is no variation in node degree. For random networks that are characteristic of a smaller variation in the node degree, the spectral radius ratio for node degree is typically closer to 1.0. For scale-free networks that are characteristic of a larger variation in node degree, the spectral radius ratio for node degree is appreciably greater than 1.0. As it is a degree-based measure, we compare the (DEG, EVC)-based NSI values of the real-world networks with their spectral radius ratio for node degree (see Figure 8(a)). We could observe an increasing trend of the (DEG, EVC)-NSI values with decrease in the spectral radius ratio for node degree. However, the R^2 values for all the models that we tried to fit to relate these two measures are at most 0.25.

(ii) The *randomness index* [76] quantifies the extent of randomness in any complex network. It is computed as the Pearson's correlation coefficient between the degree of the vertices and the average local clustering coefficient of the vertices with the particular degree. The local clustering coefficient of a vertex [1] is the probability that any two neighbors of the vertex are directly connected. For a theoretically random network (say, a random network generated according to the ER model [15]), the local clustering coefficient of a vertex is independent of the degree of the vertex, and the expected randomness index is 0. For real-world networks that are not random, the local clustering coefficient of the vertices decreases with increase in the degree of the vertices (as it is less likely that all the neighbors of a high-degree vertex will be directly connected to each other), and there is a negative correlation between the two measures, resulting in negative values for the randomness index. The more negative is the randomness index value (i.e., closer to -1) for a real-world network, the lower the extent of randomness in the network. As we expect the vertices in a theoretically random network to be similar to each other with respect to all the centrality metrics (like we saw in Section 4.2), we compare the (DEG, EVC, BWC, CLC)-based NSI values of the real-world networks with their randomness index (see Figure 8(b)). We do not observe any trend of decrease or increase in the NSI values of the real-world networks vis-a-vis their randomness index values: for example, the randomness index of real-world networks whose NSI values are in the vicinity of 0.90 range from -0.92 to -0.16.

(iii) The *assortative index* measure [13] quantifies the extent of similarity between the end vertices of a network with respect to node degree. It is calculated as the Pearson's correlation coefficient (ranging from -1 to 1) of the remaining

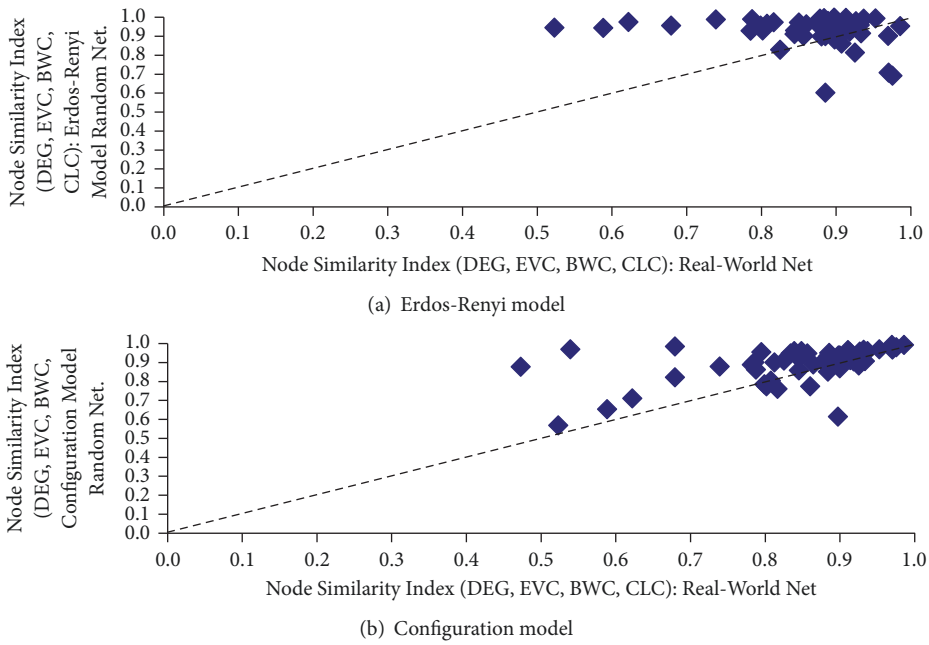


FIGURE 7: Comparison of the (DEG, EVC, BWC, CLC)-based NSI values for the real-world networks and the random networks generated using the Erdos-Renyi model and the Configuration model.

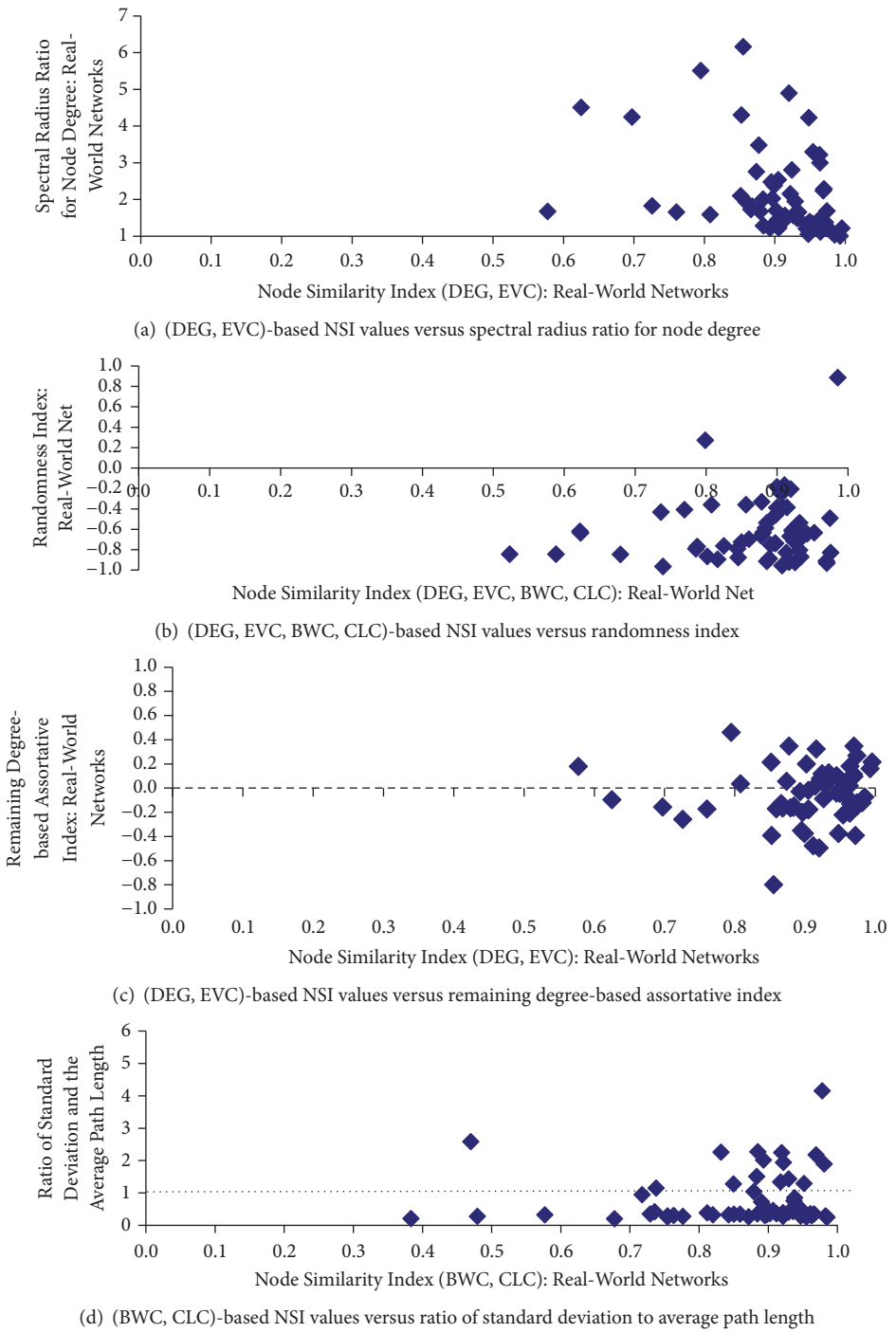
degree of the end vertices of the edges in a network. The remaining degree of a vertex is one less than the degree of the vertex. Networks with larger positive values (closer to 1) for the assortative index are considered to be assortative and networks with smaller negative values (closer to -1) for the assortative index are considered to be disassortative. As it is a degree-based measure, we compare the (DEG, EVC)-based NSI values of the real-world networks with their remaining degree-based assortative index (see Figure 8(c)). We observe larger NSI values for both assortative as well as disassortative networks. For example, the assortative index of real-world networks whose NSI values are in the vicinity of 0.90 range from -0.49 to 0.20.

(iv) The *ratio of the standard deviation to the average path length* has been a classical measure for getting an estimate of the similarity among the shortest path lengths between any two nodes in a network. If there is no significant variation in the shortest path lengths, the ratio is expected to be lower than 1.0 (and more closer to 0.0). The larger the ratio (especially, if greater than 1.0), the larger the variation in the shortest path lengths. As it is a shortest path-based measure, we compare the (BWC, CLC)-based NSI values with the ratio of the standard deviation to the average shortest path length. There is no trend of increase or decrease in the NSI values with the ratio (see Figure 8(d)). The R^2 values for the different models that we tried to fit the data do not exceed 0.10. Hence, like the other three network-level measures compared with, the proposed NSI measure captures the extent of similarity among the nodes with respect to the BWC and CLC metrics, and this is not captured with the classical approach of determining the ratio of the standard deviation to the average path length.

5. Related Work

To the best of our knowledge, similarity assessment in complex networks has been conducted only at the node-level (i.e., between any two nodes or a set of nodes, also referred to as pair-wise node similarity) and not at the network-level (i.e., among all the nodes in the network). The objective of this paper is to develop a measure to comprehensively (i.e., at the network-level) quantify the extent of the similarity among the vertices in a coordinate system based on the normalized values of the node-level metrics. In this section, we review the prominent measures available in the literature for pair-wise node similarity assessment.

One of the classical approaches for pair-wise node similarity assessment is based on the notion of "equivalence classes" [1]; there are three levels of equivalence classes: structural, automorphic and regular. Two nodes are structurally equivalent if they share many of their neighbors [1]. Some of the measures available to quantify structural equivalence are [1]: cosine similarity, Pearson's coefficient and Euclidean distance, all of which are computed based on the rows associated with the corresponding two vertices in the adjacency matrix of the graph. Two vertices u and v are automorphically equivalent if all the vertices can be relabeled to form an isomorphic graph such that the labels of u and v are interchanged [77]. Two vertices u and v are regularly equivalent if they have neighbors who are themselves similar [5, 77]. Similar to structural equivalence, there exist quantitative measures to assess automorphic equivalence and regular equivalence. In [9], the authors proposed four measures (based on maximum common neighborhood, neighborhood patterns, random walks and k-hop neighbors) to assess the



(a) (DEG, EVC)-based NSI values versus spectral radius ratio for node degree

(b) (DEG, EVC, BWC, CLC)-based NSI values versus randomness index

(c) (DEG, EVC)-based NSI values versus remaining degree-based assortative index

(d) (BWC, CLC)-based NSI values versus ratio of standard deviation to average path length

Figure 8: Comparison of the centrality-based node similarity index values for the real-world networks with some of the related network-level measures.

automorphic equivalence of two nodes. SimRank [7] and its variants such as PathSim [8] are examples of well-known measures to assess the similarity of two nodes based on the similarity of their neighbors. However, none of these quantitative measures can be seamlessly extended to quantify the similarity among nodes at the network-level. Also, from the definitions of the three equivalence classes and the measures available to quantify them, we conjecture that it is very unlikely for two distant nodes (i.e., several hops away from each other) to belong to the same equivalence class, especially in the case of structural equivalence, which is the superclass of the three classes [1]. Note that two structurally equivalent nodes are also automorphically and regularly equivalent. Two nodes that are automorphically equivalent

are regularly equivalent too, but need not be structurally equivalent. Two nodes that are regularly equivalent need not be structurally or automorphically equivalent [1].

In addition to the above, quantitative measures to assess pair-wise node similarity based on the neighborhood of the nodes were proposed by Ravasz et al. [78], Burt [79] and Goldberg and Roth [80]. Thiel and Berthold [2] proposed that two nodes (need not be directly connected to each other) are structurally similar if their neighborhoods are structurally similar to each other. In [3], Symeonidis et al. recommended that for two nodes that are not directly connected to each other, their similarity could be quantified as the product of the similarity of the end vertices constituting the edges of the shortest path between the two nodes. For weighted graphs, Chen et al. [4] introduced a measure called *relation strength similarity* (RSS) to assess similarity between two nodes: the RSS of two nodes (u, v) connected to each other is the ratio of the weight of the edge (u, v) to that of the sum of the weights of the edges incident on u and v. The transitive node similarity formulation proposed by Symeonidis et al. [3] for two nodes that are not directly connected to each other could be extended to the RSS measure as well. Though neighborhood-based methods are more common and widely used, there also exist pair-wise node similarity assessment measures that are not neighborhood-based. For example, in [6], the authors applied the notion of "mutual information" from Information Theory to quantify the extent of similarity between two nodes: the similarity score for two nodes is a function of the "information loss" encountered in the network by replacing the two nodes as one node.

While centrality metrics have been traditionally explored for their individual usability to analyze the characteristics of a real-world network, more recent studies [70–72] have focused on analyzing the correlation between any two centrality metrics to explore the usability of one centrality metric (typically, a computationally light metric) in lieu of another centrality metric (typically, a computationally heavy metric) at different levels (i.e., for prediction, network-wide ranking, pair-wise ranking, etc.). However, as seen in Section 4, correlation studies do not reveal or quantify the extent of similarity among the vertices on the basis of their centrality values with respect to two or more metrics. In [81], the authors introduced the notion of "centrality distance" to quantify the similarity of two graphs with respect to a centrality metric and is measured as the sum of the absolute differences of the centrality values (without any normalization) of the individual vertices in the two graphs.

6. Conclusions

The high-level contribution of this paper is the proposal for a unit disk graph-based approach to quantify the similarity among all the nodes in a network with respect to two or more node-level metrics. As part of this approach, we propose the use of a k-dimensional coordinate system wherein the coordinate of a vertex is composed of the normalized values of the k node-level metrics considered for similarity assessment. We propose the use of a binary search algorithm to determine the minimum value for the threshold distance (in a search space ranging from 0 to $\sqrt{k}$) that would be needed to obtain a connected unit disk graph of the vertices in the normalized coordinate system. Our hypothesis is that the larger the similarity among the vertices, the smaller the value for the minimum threshold distance needed to obtain a connected unit disk graph. We propose a measure called the node similarity index (NSI) computed as 1 − (minimum threshold distance/$\sqrt{k}$) to quantify the extent of similarity among the vertices in a scale of 0 to 1. The division by $\sqrt{k}$ in the NSI formulation (where 'k' is the number of node-level metrics considered for similarity assessment) negates the impact of the number of node-level metrics considered and solely captures the impact of the actual node-level metrics considered. With the binary search approach, for a given k and the terminating search space size $\in$, the number of iterations needed for the algorithm is the same for any complex network; the overall time complexity and space complexity of the algorithm are, respectively, $O(V^2 * \log_2^{\sqrt{k}/\in})$ and $O(V^2)$.

We evaluate our proposed model with respect to the four commonly studied centrality metrics (neighborhood-based degree and eigenvector centrality and the shortest path-based betweenness and closeness centrality) on a suite of 60 real-world networks belonging to different domains. Overall, we observe the nodes in real-world networks to be more similar with respect to the neighborhood-based centrality metrics rather than the shortest path-based centrality metrics. For all the combinations of centrality metrics considered, we observe nodes in friendship and acquaintance networks to be relatively more similar among themselves compared to the nodes in biological and coappearance networks. We showcase the uniqueness of the NSI values by comparing them with several quantitative measures such as correlation coefficient, spectral radius ratio of node degree, assortative index, randomness index and ratio of standard deviation to average path length. We do not observe any significant trend of increase or decrease in the NSI values with respect to each of these measures.

We also observed the NSI values of the real-world networks with respect to all the four centrality metrics to be different from the NSI values of the random networks (generated with the ER model) that have the same number of nodes and edges as that of the real-world networks. Thus, the notion of node similarity captured by the unit disk graph-based NSI values is not a random phenomenon and the proposed NSI measure is a unique measure whose values are also not correlated with several of the existing measures for complex network analysis.

Data Availability

The real-world network data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

References

[1] M. E. J. Newman, *Networks: An Introduction*, Oxford University Press, Oxford, UK, 2010.

[2] K. Thiel and M. R. Berthold, "Node similarities from spreading activation," in *Proceedings of the 2010 IEEE 10th International Conference on Data Mining (ICDM)*, pp. 1085–1090, Sydney, Australia, December 2010.

[3] P. Symeonidis, E. Tiakas, and Y. Manolopoulos, "Transitive node similarity for link prediction in social networks with positive and negative links," in *Proceedings of the 4th ACM Recommender Systems Conference (RecSys '10)*, pp. 183–190, September 2010.

[4] H. Chen, L. Gou, X. Zhang, and C. Giles, "Discovering missing links in networks using vertex similarity measures," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 138–143, Trento, Italy, March 2012.

[5] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 73, no. 2, Article ID 026120, 2006.

[6] Y. Li, P. Luo, and C. Wu, "A new network node similarity measure method and its applications," March 2014.

[7] G. Jeh and J. Widom, "SimRank: a measure of structural-context similarity," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 538–543, Edmonton, Canada, July 2002.

[8] W. Yu, X. Lin, W. Zhang, L. Chang, and J. Pei, "More is simpler: effectively and efficiently assessing node-pair similarities based on hyperlinks," *Proceedings of the VLDB Endowment*, vol. 7, no. 1, pp. 13–24, 2013.

[9] Y. Yang, J. Pei, and A. Al-Barakati, "Measuring in-network node similarity based on neighborhoods: a unified parametric approach," *Knowledge and Information Systems*, vol. 53, no. 1, pp. 43–70, 2017.

[10] A. Singhal, "Modern information retrieval: a brief overview," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, vol. 24, no. 4, pp. 35–43, 2001.

[11] N. Meghanathan, "Maximal assortative matching and maximal dissortative matching for complex network graphs," *The Computer Journal*, vol. 59, no. 5, pp. 667–684, 2016.

[12] S. Zhou and R. J. Mondragón, "The rich-club phenomenon in the internet topology," *IEEE Communications Letters*, vol. 8, no. 3, pp. 180–182, 2004.

[13] M. E. Newman, "Mixing patterns in networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 67, no. 2, 2003.

[14] S. P. Borgatti and M. G. Everett, "A graph-theoretic perspective on centrality," *Social Networks*, vol. 28, no. 4, pp. 466–484, 2006.

[15] P. Erdös and A. Rényi, "On random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[16] T. Britton, M. Deijfen, and A. Martin-Lof, "Generating simple random graphs with prescribed degree distribution," *Journal of Statistical Physics*, vol. 124, no. 6, pp. 1377–1397, September 2006.

[17] P. Bonacich, "Power and centrality: a family of measures," *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[18] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[19] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[20] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.

[21] T. H. Cormen, C. E. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*, The MIT Press, 2009.

[22] N. Meghanathan, "Spectral radius as a measure of variation in node degree for complex network graphs," in *Proceedings of the 2014 7th International Conference on u- and e- Service, Science and Technology (UNESST)*, pp. 30–33, Hainan Island, China, December 2014.

[23] N. Meghanathan, "Centrality and partial correlation coefficient-based assortativity analysis of real-world networks," *The Computer Journal*, 2018.

[24] E. Schwimmer, *Exchange in the Social Structure of the Orokaiva: Traditional and Emergent Ideologies in the Northern District of Papua*, C Hurst and Co-Publishers Ltd., London, UK, 1973.

[25] J. H. Michael, "Labor dispute reconciliation in a forest products manufacturing facility," *Forest Products Journal*, vol. 47, no. 11-12, pp. 41–45, 1997.

[26] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[27] W. De Nooy, "A literary playground: Literary criticism and balance theory," *Poetics*, vol. 26, no. 5-6, pp. 385–404, 1999.

[28] J. Gil-Mendieta and S. Schmidt, "The political network in Mexico," *Social Networks*, vol. 18, no. 4, pp. 355–381, 1996.

[29] V. Batagelj and A. Mrvar, "Pajek datasets," http://vlado.fmf.uni-lj.si/pub/networks/data/, 2005.

[30] D. Krackhardt, "The ties that torture: simmelian tie analysis in organizations," *Research in the Sociology of Organizations*, vol. 16, pp. 183–210, 1999.

[31] E. M. Rogers and D. L. Kincaid, *Communication Networks: Toward a New Paradigm for Research*, Free Press, New York City, NY, USA, 1980.

[32] L. C. Freeman, S. C. Freeman, and A. G. Michaelson, "How humans see social groups: a test of the sailer-gaulin models," *Journal of Quantitative Anthropology*, vol. 1, pp. 229–238, 1989.

[33] J. L. Moreno, *The Sociometry Reader*, The Free Press, Glence, IL, USA, 1960.

[34] M. Pearson and L. Michell, "Smoke Rings: Social network analysis of friendship groups, smoking and drug-taking," *Drugs: Education, Prevention and Policy*, vol. 7, no. 1, pp. 21–36, 2000.

[35] H. R. Bernard, P. D. Killworth, and L. Sailer, "Informant accuracy in social network data IV: a comparison of clique-level structure in behavioral and cognitive network data," *Social Networks*, vol. 2, no. 3, pp. 191–218, 1979.

[36] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: can geographic isolation explain this unique trait?" *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003.

[37] Y. Takahata, "Diachronic changes in the dominance relations of adult female japanese monkeys of the arashiyama B group," *The Monkeys of Arashiyama*, pp. 124–139, 1991.

[38] D. MacRae, "Direct factor analysis of sociometric data," *Sociometry*, vol. 23, no. 4, p. 360, 1960.

[39] B. Hayes, "Connecting the dots," *American Scientist*, vol. 94, no. 5, pp. 400–404, 2006.

[40] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*, Addison-Wesley, Reading, MA, USA, 1993.

[41] C. P. Loomis, J. O. Morales, R. A. Clifford, and O. E. Leonard, *Turrialba Social Systems and the Introduction of Change*, The Free Press, 1953.

[42] R. L. Cross and A. Parker, *The Hidden Power of Social Networks : Understanding How Work Really Gets Done in Organizations*, Harvard Business School Press, Boston, Mass, USA, 2004.

[43] D. A. Smith and D. R. White, "Structure and dynamics of the global economy: Network analysis of international trade 1965-1980," *Social Forces*, vol. 70, no. 4, pp. 857–893, 1992.

[44] T. Nepusz, A. Petróczi, L. Négyessy, and F. Bazsó, "Fuzzy communities and the concept of bridgeness in complex networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 77, no. 1, Article ID 016107, 2008.

[45] J. Grimmer, "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases," *Political Analysis*, vol. 18, no. 1, pp. 1–35, 2010.

[46] V. Krebs, "Proxy networks –analyzing one network to reveal another," *Bulletin de Méthodologie Sociologique*, vol. 79, no. 1, pp. 61–70, 2003.

[47] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 74, no. 3, Article ID 036104, 19 pages, 2006.

[48] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. van den Broeck, "What's in a crowd? Analysis of face-to-face behavioral networks," *Journal of Theoretical Biology*, vol. 271, no. 1, pp. 166–180, 2011.

[49] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[50] N. P. Hummon, P. Doreian, and L. C. Freeman, "Analyzing the structure of the centrality-productivity literature created between 1948 and 1979," *Science Communication*, vol. 11, no. 4, pp. 459–480, 1990.

[51] P. M. Gleiser, "How to become a superhero," *Journal of Statistical Mechanics: Theory and Experiment*, no. 9, 2007.

[52] P. M. Gleiser and L. Danon, "Community structure in jazz," *Advances in Complex Systems (ACS)*, vol. 6, no. 4, pp. 565–573, 2003.

[53] M. D. Resnick, P. S. Bearman, R. W. Blum et al., "Protecting adolescents from harm: findings from the national longitudinal study on adolescent health," *The Journal of the American Medical Association*, vol. 278, no. 10, pp. 823–832, 1997.

[54] L. C. Freeman, C. M. Webster, and D. M. Kirke, "Exploring social structure using dynamic three-dimensional color images," *Social Networks*, vol. 20, no. 2, pp. 109–118, 1998.

[55] V. Gemmetto, A. Barrat, and C. Cattuto, "Mitigation of infectious disease at school: Targeted class closure vs school closure," *BMC Infectious Diseases*, vol. 14, no. 1, 2014.

[56] J. P. Scott, *The Anatomy of Scottish Capital: Scottish Companies and Scottish Capital*, Croom Helm, London, UK, 1980.

[57] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, "The structure of the nervous system of the nematode Caenorhabditis elegans," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 314, no. 1165, pp. 1–340, 1986.

[58] J-S. Lee, "Generating networks of illegal drug users using large samples of partial ego-network data," in *Proceedings of the Intelligence and Security Informatics*, vol. 3073 of *Lecture Nodes in Computer Science*, pp. 390–402, 2004.

[59] T. Biedl and F. J. Brandenburg, "Graph-drawing contest report," in *Graph Drawing Software*, vol. 2265 of *Lecture Notes in Computer Science*, pp. 513–521, Springer, Berlin, Germany, 2002.

[60] C. McCarty and L. Freeman, http://moreno.ss.uci.edu/data.html, 2008.

[61] M. De Domenico, M. A. Porter, and A. Arenas, "MuxViz: A tool for multilayer analysis and visualization of networks," *Journal of Complex Networks*, vol. 3, no. 2, pp. 159–176, 2015.

[62] N. Bhardwaj, K.-K. Yan, and M. B. Gerstein, "Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 107, no. 15, pp. 6841–6846, 2010.

[63] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 68, no. 6, Article ID 065103, 2003.

[64] L. Subelj and M. Bajec, "Robust network community detection using balanced propagation," *The European Physical Journal B*, vol. 81, no. 3, pp. 353–362, 2011.

[65] M. De Domenico, V. Nicosia, A. Arenas, and V. Latora, "Structural reducibility of multilayer networks," *Nature Communications*, vol. 6, 2015.

[66] H. Yu, P. Braun, M. A. Yildirim et al., "High-quality binary protein interaction map of the yeast interactome network," *Sciencexpress*, p. 1-11, 2008.

[67] C. Seierstad and T. Opsahl, "For the few not the many? The effects of affirmative action on presence, prominence, and social capital of women directors in Norway," *Scandinavian Journal of Management*, vol. 27, no. 1, pp. 44–54, 2011.

[68] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: divided they blog," in *Proceedings of the 3rd International Workshop on Link Discovery*, pp. 36–43, ACM, Chicago, IL, 2005.

[69] T. Reguly, A. Breitkreutz, L. Boucher et al., "Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae," *Journal of Biology*, vol. 5, no. 4, pp. 1–11, 2006.

[70] N. Meghanathan, "Correlation coefficient analysis of centrality metrics for complex network graphs," in *Proceedings of the 4th Computer Science Online Conference, (CSOC-2015), Intelligent Systems in Cybernetics and Automation Theory: Advances in Intelligent Systems and Computing*, vol. 348, pp. 11–20, April 2015.

[71] C. Li, Q. Li, P. Van Mieghem, H. Stanley, and H. Wang, "Correlation between centrality metrics and their application to the opinion model," *The European Physical Journal B*, vol. 88, no. 3, Art. 65, 13 pages, 2015.

[72] T. W. Valente, K. Coronges, C. Lakon, and E. Costenbader, "How correlated are network centrality measures?" *Connections*, vol. 28, no. 1, pp. 16–26, 2008.

[73] G. Strang, *Linear Algebra and its Applications*, Brooks Cole, Pacific Grove, CA, USA, 2005.

[74] N. Meghanathan and X. He, "Correlation and regression analysis for node betweenness centrality," *International Journal in Foundations of Computer Science & Technology*, vol. 6, no. 6, pp. 01–20, 2016.

[75] A. L. Barabasi, *Network Science*, Cambridge University Press, 1st edition, August 2016.

[76] N. Meghanathan, "Randomness index for complex network analysis," *Social Network Analysis and Mining*, vol. 7, no. 1, 2017.

[77] S. Borgatti, M. Everett, and L. Freeman, *UCINET IV Version 1.0 User's Guide*, Analytic Technologies, Columbia, SC, USA, 1992.

[78] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[79] R. S. Burt, "Positions in networks," *Social Forces*, vol. 55, no. 1, pp. 93–122, 1976.

[80] D. S. Goldberg and F. P. Roth, "Assessing experimentally derived interactions in a small world," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 100, no. 8, pp. 4372–4376, 2003.

[81] M. Roy, S. Schmid, and G. Tredan, "Modeling and measuring graph similarity," in *Proceedings of the 10th ACM International Workshop*, pp. 47–52, Philadelphia, Pennsylvania, USA, August 2014.