Vietnam J Comput Sci (2017) 4:23–38
DOI 10.1007/s40595-016-0073-1

REGULAR PAPER

A computationally lightweight and localized centrality metric in lieu of betweenness centrality for complex network analysis

Natarajan Meghanathan[1]

Received: 6 April 2016 / Accepted: 2 June 2016 / Published online: 17 June 2016

Abstract The betweenness centrality (BWC) of a vertex is a measure of the fraction of shortest paths between any two vertices going through the vertex and is one of the widely used shortest path-based centrality metrics for the complex network analysis. However, it takes $O(|V|^2 + |V||E|)$ time (where **V** and **E** are, respectively, the sets of nodes and edges of a network graph) to compute the BWC of just a single node. Our hypothesis is that nodes with a high degree, but low local clustering coefficient, are more likely to be on the shortest paths of several node pairs and are likely to incur a larger BWC value. Accordingly, we define the local clustering coefficient-based degree centrality (LCCDC) for a node as the product of the degree centrality of the node and one minus the local clustering coefficient of the node. The LCCDC of a node can be computed based on just the knowledge of the two-hop neighborhood of a node and would take significantly lower time. We conduct an exhaustive correlation analysis and observe the LCCDC to incur the largest correlation coefficient values with BWC (compared to other centrality metrics under three different correlation measures) and to hold very strong levels of positive correlation with BWC for at least 14 of the 18 real-world networks analyzed. Hence, we claim the LCCDC to be an apt metric to rank the nodes or compare any two nodes of a real-world network graph in lieu of BWC.

Keywords Betweenness centrality · Degree centrality · Local clustering coefficient · Correlation coefficient · Complex network graphs

✉ Natarajan Meghanathan
natarajan.meghanathan@jsums.edu

[1] Department of Computer Science, Jackson State University, Jackson, USA

1 Introduction

Network science (a.k.a. complex network analysis) is an emerging area of interest in the data science discipline and corresponds to analyzing complex real-world networks from a graph theory point of view. Among the various metrics used for complex network analysis, node centrality is a prominently used metric of immense theoretical interest and practical value. The centrality of a node is a link statistics-based quantitative measure of the topological importance of the node with respect to the other nodes in the network [1]. Applications for node centrality metrics could be, for example, to identify the most influential persons in a social network, the key infrastructure nodes in an internet, the super-spreaders of a disease, etc. The existing centrality metrics could be broadly classified into two categories [1]: neighbor-based and shortest path-based. Degree centrality (DegC) and eigenvector centrality (EVC) [2] are well-known metrics for neighbor-based centrality, while Betweenness centrality (BWC) [3] and closeness centrality (ClC) [4] are well-known metrics for shortest path-based centrality. Throughout the paper, the terms 'node' and 'vertex', 'link' and 'edge', and 'network' and 'graph' are used interchangeably. They mean the same.

The degree centrality of a vertex is the number of neighbors connected to the vertex and can be determined just based on the one-hop neighborhood knowledge. The eigenvector centrality of a vertex is a measure of the degree of the vertex as well as the degree of its neighbors. The betweenness centrality of a vertex is a measure of the fraction of the shortest paths between any two vertices that go through the vertex; whereas the closeness centrality of a vertex is a measure of the shortest path distances to every other vertex in the network. Other than degree centrality, all the above three centrality metrics require the global knowledge of the network for their computation.

With respect to the running time of the algorithms to compute the centrality metrics, for an arbitrary network graph of $|V|$ vertices and $|E|$ edges: the EVC of all the vertices en masse can be computed in $O(|V|^3)$ time, whereas it would take $O(|V|+|E|)$ and $O(|V|^2+|V||E|)$ time, respectively, to compute the closeness centrality and betweenness centrality of an individual vertex. The BWC, thus, incurs the longest running time to be computed for just a single node. As the BWC for a node u is defined as the sum of the fraction of shortest paths between any two nodes i and j ($i \neq j \neq u$) that go through node u, one would have to run the shortest path algorithm on every node in the graph to compute the BWC of even a single node. Even though the BWC of all the vertices could be determined once the shortest path algorithm is run on every node in a network graph, it is still too much of a computation overhead on network graphs with a larger number of nodes and/or edges (especially, if one is interested in just knowing the relative importance of a selected few vertices with regards to their location on the shortest paths among any two vertices in the network graph). Thus, the motivation of this research is to explore the possibility of using a computationally lightweight localized centrality metric that is highly correlated to the BWC and could be used to rank the vertices or compare selected vertices in a network graph in lieu of the BWC.

Our high-level contribution in this paper is the proposal of a local clustering coefficient-based degree centrality (LCCDC) metric as a computationally lightweight centrality alternative for the betweenness centrality (BWC). The local clustering coefficient of a node in a graph is the fraction of the pairs of its neighbors that are directly connected to each other. The underlying theoretical basis for the proposed LCCDC metric is that if none of the neighbors of a vertex go through the vertex for shortest path communication, and then none of the other vertices in the graph go through the vertex for shortest path communication. Accordingly, we define the LCCDC of a vertex as the product of the degree of the vertex and one minus the local clustering coefficient of the vertex. The LCCDC metric, thus, quantifies the extent, to which the degree centrality of a vertex facilitates shortest path communication through the vertex and could be at most the degree centrality of the vertex. If a vertex has a high degree, but a low local clustering coefficient, it implies that though the vertex has several neighbors—a very few of these neighbors are directly connected to each other. Hence, a high-degree vertex with a low local clustering coefficient is likely to be on the shortest path for several pairs of vertices in the network (at least for the neighbors of the node). On the other hand, a vertex with a higher clustering coefficient (even if it has a higher degree) is not likely to be on the shortest paths connecting its neighbors and thereby not likely to be on the shortest paths between any two vertices in the graph. All of the above arguments form the basis of our hypothesis that a high-degree vertex with a low local clustering coefficient is more likely to exhibit a larger value for the betweenness centrality.

We explore the level of correlation between LCCDC and BWC through extensive experimental studies involving a suite of 18 real-world networks, whose degree distribution ranges from Poisson to Power-law [5] under three different correlation measures [5]. We observe the LCCDC to exhibit highest values for the correlation coefficient with BWC (compared to DegC, EVC, and ClC under all the three correlation measures). In addition to the quantitative values, we also qualitatively classify the level of correlation for BWC with the other centrality metrics studied in this paper, and observe the newly proposed LCCDC metric to exhibit strong-very strong levels of positive correlation with BWC for at least 16 of the 18 real-world networks analyzed. High levels of positive correlation between time-efficient LCCDC and time-consuming BWC are an indicator that if two vertices are to be compared based on their BWC values, it would be more likely sufficient to just compare their LCCDC values. Similarly, the ranking of the vertices in a real-world network graph based on their BWC values is more likely to be the same as the ranking of the vertices based on the LCCDC metric. Thus, we claim that the LCCDC could be used to compare vertices in lieu of their BWC.

The rest of the paper is organized as follows: Sect. 2 reviews the classical centrality metrics (DegC, EVC, BWC, and ClC) and the calculation of the BWC metric with an example. Section 3 introduces the local clustering coefficient-based degree centrality (LCCDC) metric and justifies its proposal as an alternate for BWC with a motivating example. Section 4 introduces the three measures of correlation used in the experimental studies on real-world networks. Section 5 presents the 18 real-world network graphs and discusses the results of correlation coefficient analysis for BWC with each of LCCDC, DegC, EVC, and ClC as well as ranks the five centrality metrics on the basis of the execution time incurred to compute them on these graphs. Section 6 reviews related work on correlation studies involving the centrality metrics. Section 7 concludes the paper and explores directions for future research.

2 Node centrality metrics

We now review the centrality metrics that are used for the correlation coefficient analysis studies in this paper. These are the neighbor-based degree centrality (DegC) and eigenvector centrality (EVC) metrics and the shortest path-based betweenness centrality (BWC) and closeness centrality (ClC) metrics.

The degree centrality (DegC) of a vertex is the number of neighbors for the vertex in the graph and can be easily computed by counting the number of edges incident on the vertex. If $\mathbf{A}$ is the $n \times n$ adjacency matrix for a graph, such that $\mathbf{A}[i, j] = 1$ if there is an edge connecting v_i to v_j (for undirected graphs) and $\mathbf{A}[i, j] = 0$ if there is no edge connecting v_i and v_j. The degree centrality of a vertex v_i is quantitatively defined as follows: $\text{DegC}(v_i) = \sum_{j=1}^{n} A[i, j]$. It would take $O(|V|)$ time to determine the degree centrality of a vertex, as there would be $n = |V|$ entries in the row corresponding to each vertex in the adjacency matrix.

The eigenvector centrality (EVC) of a vertex is a quantitative measure of the degree of the vertex as well as the degree of its neighbors. A vertex that has a high degree for itself as well as located in the neighborhood of high-degree vertices is likely to have a larger EVC. The EVC values of the vertices in a graph correspond to the entries for the vertices in the principal eigenvector of the adjacency matrix of the graph. An $n \times n$ adjacency matrix has n eigenvalues and the corresponding eigenvectors. The principal eigenvector is the eigenvector corresponding to the largest eigenvalue (principal eigenvalue) of the adjacency matrix, $\mathbf{A}$. Moreover, if all the entries in a square matrix are positive (i.e., greater than or equal to zero), the principal eigenvalue as well as the entries in the principal eigenvector are also positive [6]. We determine the EVC of the vertices using the power-iteration method [6] of complexity $O(|V|^3)$ in a graph of $|V|$ vertices, as there are $O(|V|^2)$ multiplications in each iteration of the power-iteration method, and there could be at most $|V|$ iterations before the normalized value of the eigenvector converges to the principal eigenvalue (typically, the number of iterations needed for the convergence to happen would be far less than the number of vertices in the graph).

The betweenness centrality (BWC) of a vertex is the sum of the fraction of shortest paths going through the vertex between any two vertices, considered over all pairs of vertices. In this paper, we determine the BWC of the vertices using the breadth first search (BFS)-variant of the well-known Brandes algorithm [7]. We run the BFS algorithm [8] on each vertex in the graph and determine the level of each vertex (the number of hops/edges from the root) in each of these BFS trees. The root of a BFS tree is said to be at level 0 and the number of shortest paths from the root to itself is 1. On a BFS tree rooted at vertex r, the number of shortest paths for a vertex i at level l ($l > 0$) from the root r is the sum of the number of shortest paths from the root r to each the neighbors of vertex i (in the original graph) that are at level $l-1$ in the BFS tree. Since we are working on undirected graphs, the total number of shortest paths from vertex i to vertex j (denoted sp_{ij}) is simply the number of shortest paths from vertex i to vertex j in the shortest path tree rooted at vertex i or vice-versa. The number of shortest paths from a vertex i to a vertex j that go through a vertex k (denoted $\text{sp}_{ij}(k)$) is the maximum of the number of shortest paths from vertex i to vertex k in the shortest path tree rooted at i and the number of shortest paths from vertex j to vertex k in the shortest path tree rooted at vertex j. Thus, $\text{BWC}(k) = \sum_{\substack{k \neq i \\ k \neq j}} \frac{\text{sp}_{ij}(k)}{\text{sp}_{ij}}$. With regard to the run-time complexity of the Brandes algorithm, it would take $O(|V| + |E|)$ time to run the BFS shortest path algorithm on a particular vertex and a total of $O(|V|*(|V|+|E|))$ time on the $|V|$ vertices of a network graph. In addition, for each vertex: one has to trace through the $|V|$ shortest path trees to determine the number of shortest paths from the root vertices of these shortest path trees to the particular vertex for which we want to find the BWC. This could take another $|V||E|$ time for all the vertices in the graph. Thus, the computation time incurred to determine the BWC values of all the vertices in a graph would be: $O(|V|^2 + |V||E| + |V||E|)$, which for all theoretical purposes is written simply as: $O(|V|^2 + |V||E|)$.

Figure 1 illustrates an example to calculate the BWC of the vertices on a sample graph that is used as a running example in Figs. 1, 2, 3, 4, 5, and 6. We can observe the betweenness values for vertices 0, 6, and 7 are zero each, because no shortest path between any two vertices go through them. We observe that even though vertices 4 and 5 have the same larger degree, the average degree of the neighbors of vertex 5 is slightly lower than the average degree of the neighbors of vertex. As a result, vertex 5 is more likely to occupy a relatively larger fraction of the shortest path between any two vertices and incur a relatively larger BWC value compared to vertex 4 (even though vertex 4 has a larger EVC value). In addition, even though vertex 3 has a larger degree than vertex 1, the BWC of vertex 1 is significantly larger than that of vertex 3. This could be attributed to vertex 1 lying on the shortest path from vertices 0 and 2 to vertices 4, 5, 6, and 7; on the other hand, vertex 3 lies only on the shortest path between 2 and 5.

The closeness centrality (ClC) of a vertex is the inverse of the sum of the number of shortest paths from the vertex to every other vertex in the graph. We determine the ClC of the vertices by running the BFS algorithm on each vertex and summing the number of shortest paths from the root vertex to every other vertex in these BFS trees. It would take $O(|V| + |E|)$ time to run the BFS algorithm once and determine the shortest path tree rooted at a particular vertex. To determine the closeness centrality of all the vertices in a graph, one would have to run the BFS algorithm on each of the vertices: thus, incurring an overall time complexity of $O(|V|*(|V| + |E|)) = O(|V|^2 + |V||E|)$. However, unlike the BWC metric, there is no additional computation overhead incurred to determine the ClC values of the vertices.

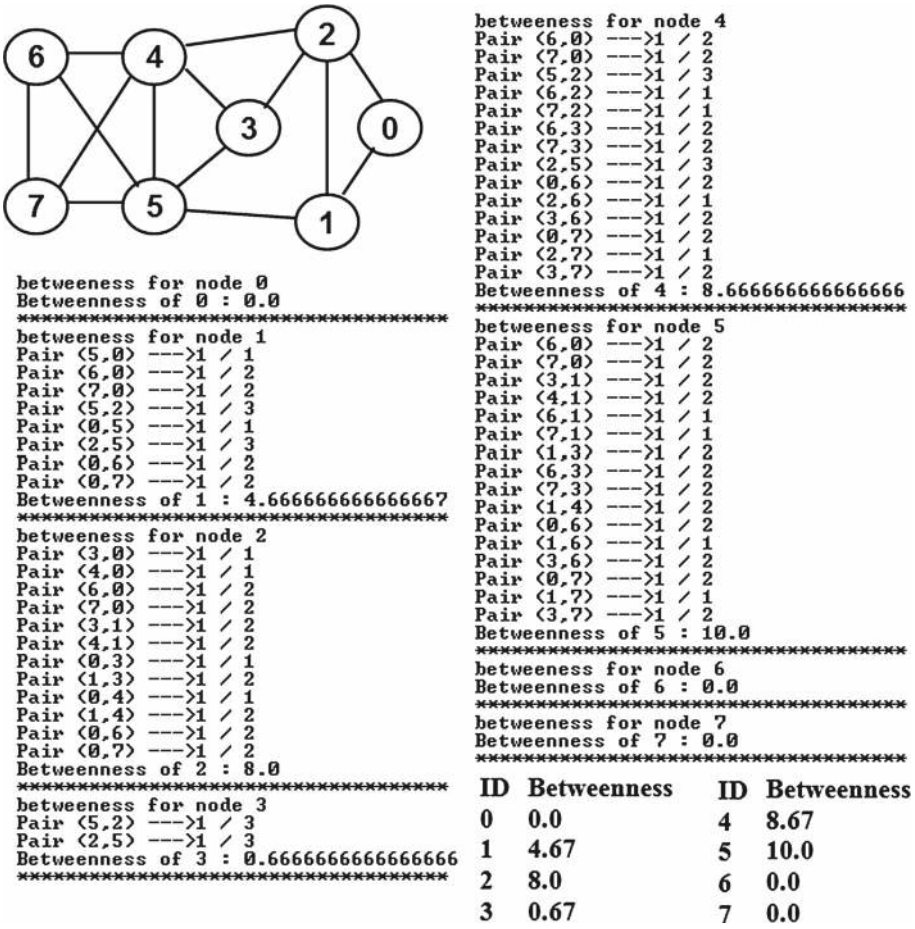
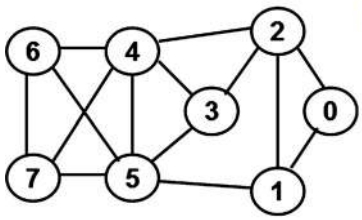


Fig. 1 Example to illustrate the calculation of betweenness centrality



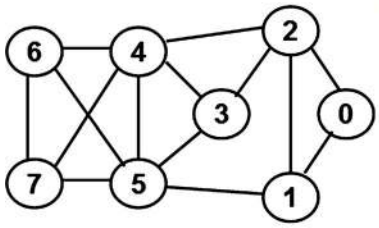
Vertex ID, v_i	Degree (k_i) Centrality	Actual # Links among Neighbors	Max. Possible # Links among Neighbors	Local Clustering Coefficient, LCC
0	2	1	2(2-1)/2 = 1	1/1 = 1.0
1	3	1	3(3-1)/2 = 3	1/3 = 0.33
2	4	2	4(4-1)/2 = 6	2/6 = 0.33
3	3	2	3(3-1)/2 = 3	2/3 = 0.67
4	5	5	5(5-1)/2 = 10	5/10 = 0.5
5	5	4	5(5-1)/2 = 10	4/10 = 0.4
6	3	3	3(3-1)/2 = 3	3/3 = 1.0
7	3	3	3(3-1)/2 = 3	3/3 = 1.0

Fig. 2 Example to illustrate the calculation of local clustering coefficient

3 Local clustering coefficient-based degree centrality

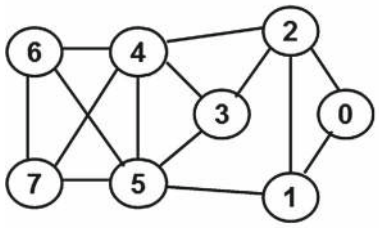
The local clustering coefficient (LCC) of a vertex is the ratio of the actual number of links between the neighbors of the vertex to that of the maximum possible number of links between the neighbors of the vertex [1]. For a vertex v_i with degree k_i (i.e., k_i neighbors), the maximum possible number of links between the neighbors of the node is $k_i(k_i-1)/2$. Figure 2 illustrates the computation of the LCC values of the vertices on the example graph used in Fig. 1. We see that a vertex having high degree need not necessarily have a higher LCC, as it would be difficult to expect direct links between any two neighbors of the vertex. In Fig. 2, we observe that both vertices 4 and 5 that have a degree of 5 each incur LCC values that are lower than the LCC of vertices 6 and 7 that have a degree of 3 each. In addition, vertices with the same degree need not have the same LCC, as the connectivity among the neighbors of each vertex could be different from that of the others. We notice that though vertices 3, 6, and 7 have a degree of 3 each, the LCC of vertex 3 is only 0.33, whereas vertices 6 and 7 have an LCC of 1.0 each.

Our hypothesis behind the proposed local clustering coefficient-based degree centrality (LCCDC) metric is as follows: a high-degree vertex with a lower clustering coefficient is essential to at least connect the neighbors (that are not



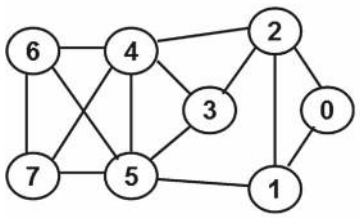
Vertex ID, v_i	Degree (k_i) Centrality	Local Clustering Coefficient, LCC	(1-LCC)	LCCDC k_i * (1-LCC)	BWC
0	2	1.0	0.0	0.0	0.0
1	3	0.33	0.67	2.0	4.67
2	4	0.33	0.67	2.67	8.0
3	3	0.67	0.33	1.0	0.67
4	5	0.5	0.5	2.5	8.67
5	5	0.4	0.6	3.0	10.0
6	3	1.0	0.0	0.0	0.0
7	3	1.0	0.0	0.0	0.0

Fig. 3 Example to illustrate the calculation of local clustering coefficient-based degree centrality



Vertex	BWC (B)	LCCDC (C)	$B - B_{avg}$	$C - C_{avg}$	$(B - B_{avg}) * (C - C_{avg})$	$(B - B_{avg})^2$	$(C - C_{avg})^2$
0	0	0	-4	-1.39	5.56	16	1.93
1	4.67	2	0.67	0.61	0.41	0.45	0.37
2	8	2.67	4	1.28	5.12	16	1.64
3	0.67	1	-3.33	-0.39	1.30	11.09	0.15
4	8.67	2.5	4.67	1.11	5.18	21.81	1.23
5	10	3	6	1.61	9.66	36	2.59
6	0	0	-4	-1.39	5.56	16	1.93
7	0	0	-4	-1.39	5.56	16	1.93
Avg	4	1.39		Sum	38.35	133.35	11.78
	Correlation Coefficient = 38.35 / √(133.35*11.78) = 0.97						

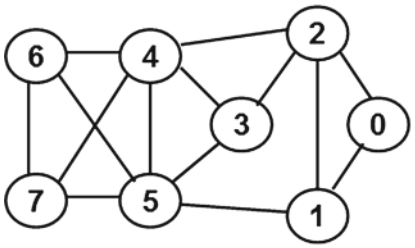
Fig. 4 Example to illustrate the computation of Pearson's correlation coefficient (betweenness centrality: B and local clustering coefficient-based degree centrality: C)



Vertex	BWC (B)	Tentative Rank: B	Final Rank: b_i	LCCDC (C)	Tentative Rank: C	Final Rank: c_i	Rank Difference (d_i): $b_i - c_i$	d_i^2
0	0	1	2	0	1	2	0	0
1	4.67	5	5	2	5	5	0	0
2	8	6	6	2.67	7	7	-1	1
3	0.67	4	4	1	4	4	0	0
4	8.67	7	7	2.5	6	6	1	1
5	10	8	8	3	8	8	0	0
6	0	2	2	0	2	2	0	0
7	0	3	2	0	3	2	0	0
							Sum	2
		Correlation Coefficient = 1 - (6*2)/(8*(8^2-1)) = 0.98						

Fig. 5 Example to illustrate the computation of Spearman's correlation coefficient (betweenness centrality: B and local clustering coefficient-based degree centrality: C)

directly connected to each other) of the vertex on a shortest path. In addition, such a high-degree vertex with a lower LCC might be on the shortest path of several other pairs of vertices (especially, for those vertices that are in the 2-hop and 3-hop neighborhood), eventually contributing to a higher BWC for the vertex. On the other hand, a vertex in a connected graph incurs a BWC of zero if none of the neighbors of the vertex go through it for their shortest path(s) to any other vertex in the graph. In other words, a vertex sustains a BWC value of zero if it is either a stub vertex (has a degree of 1: that is connected to only one other vertex) or there exists a link between any two neighbors of the vertex. In both the cases, the LCC of the vertex is 1 and the BWC value for the vertex will be zero. Considering all of the above, we propose to calculate the LCCDC metric for a vertex as the product of the degree centrality of the vertex and one minus the local clustering coefficient of the vertex. That is, $\text{LCCDC}(v_i) = k_i * (1 - \text{LCC}(v_i))$. The proposed formulation also sets up meaningful upper bound and lower bound for the LCCDC metric. With the above formulation, the maximum possible value for the local clustering coefficient-based degree centrality of a vertex is the degree centrality of the vertex itself (if the LCC of the vertex is 0) and the minimum possible value for the LCCDC of a vertex is 0 (if the LCC of the vertex is 1). Thus, the proposed formulation for LCCDC of a vertex captures the extent to which the degree centrality of a vertex is useful



Vertex	BWC (B)	LCCDC (C)
0	0	0
1	4.67	2
2	8	2.67
3	0.67	1
4	8.67	2.5
5	10	3
6	0	0
7	0	0

#conc.pairs = 25
#disc.pairs = 1
Total # pairs = 8(8-1)/2 = 28

Correlation Coefficient
(25-1)/28 = 0.86

Vertex Pairs (v_i, v_j)	B_i, C_i	B_j, C_j	Type of Pairs
(0, 1)	0, 0	4.67, 2	Concordant
(0, 2)	0, 0	8, 2.67	Concordant
(0, 3)	0, 0	0.67, 1	Concordant
(0, 4)	0, 0	8.67, 2.5	Concordant
(0, 5)	0, 0	10, 3	Concordant
(0, 6)	0, 0	0, 0	N/A
(0, 7)	0, 0	0, 0	N/A
(1, 2)	4.67, 2	8, 2.67	Concordant
(1, 3)	4.67, 2	0.67, 1	Concordant
(1, 4)	4.67, 2	8.67, 2.5	Concordant
(1, 5)	4.67, 2	10, 3	Concordant
(1, 6)	4.67, 2	0, 0	Concordant
(1, 7)	4.67, 2	0, 0	Concordant
(2, 3)	8, 2.67	0.67, 1	Concordant
(2, 4)	8, 2.67	8.67, 2.5	Discordant
(2, 5)	8, 2.67	10, 3	Concordant
(2, 6)	8, 2.67	0, 0	Concordant
(2, 7)	8, 2.67	0, 0	Concordant
(3, 4)	0.67, 1	8.67, 2.5	Concordant
(3, 5)	0.67, 1	10, 3	Concordant
(3, 6)	0.67, 1	0, 0	Concordant
(3, 7)	0.67, 1	0, 0	Concordant
(4, 5)	8.67, 2.5	10, 3	Concordant
(4, 6)	8.67, 2.5	0, 0	Concordant
(4, 7)	8.67, 2.5	0, 0	Concordant
(5, 6)	10, 3	0, 0	Concordant
(5, 7)	10, 3	0, 0	Concordant
(6, 7)	0, 0	0, 0	Concordant

Fig. 6 Example to illustrate the computation of Kendall's correlation coefficient (betweenness centrality: B and local clustering coefficient-based degree centrality: C)

in facilitating shortest path communication through the vertex, and we claim it to be lightweight alternative to the BWC metric (as verified in Sect. 4).

Figure 3 illustrates the computation of the LCCDC values of the vertices of the example graph used in Figs. 1 and 2. We observe that larger the LCCDC value for a vertex, the larger the BWC value for the vertex and vice-versa. We observe that vertices 0, 6, and 7 that do not lie on the shortest path for any two vertices in the graph have a BWC of zero each and also have LCCDC value of zero each. Notice that for each of these 3 vertices 0, 6, and 7: the neighbors of the vertex have direct links to each other and are not required to go through the vertex (this is one of the two scenarios for which the BWC value of a vertex will be zero, as explained above). We also notice that though both vertices 4 and 5 have a degree of 5 each, vertex 5 has relatively larger values for both the LCCDC and BWC metrics owing to relatively fewer fraction of direct links among its neighbors. Likewise, though both vertices 1 and 3 have a degree of 3 each, vertex 1 has relatively larger BWC and LCCDC values due to a relatively fewer fraction of direct links among its neighbors.

The local clustering coefficient of a vertex can be computed by checking whether the neighbors of the vertex are directly connected to each other. For a vertex i with k_i neighbors, there is a possibility of $k_i(k_i-1)/2$ edges among the neighbors of vertex i. This could be efficiently done in O(1) time for each pair of neighbors by checking their corresponding entry in the adjacency matrix, leading to a time complexity of $O(k_i^2)$ for a vertex i of degree k_i. Thus, the time complexity incurred to compute the local clustering coefficient of the vertices in a graph narrows down to the problem of determining an upper bound for the sum of the squares of the degrees of the vertices in a graph. This has been derived to be $O(|E| * (\frac{2*|E|}{|V|-1} + |V| - 2))$ for a graph of $|V|$ vertices and $|E|$ edges [36]. It would take $O(|V|^2)$ time to compute the degree centrality of the vertices in a graph. Hence, the time complexity incurred to compute the LCCDC of the vertices in a network graph of $|V|$ vertices and $|E|$ edges can be written as: $O(|V|^2 + |E| * (\frac{2*|E|}{|V|-1} + |V| - 2))$.

4 Correlation coefficient measures

We now discuss the three well-known correlation coefficient measures that are used to evaluate the correlation between BWC and LCCDC as well as the correlations

between BWC and each of the other three centrality metrics (DegC, EVC and ClC) presented in Sect. 2. These are the product moment-based Pearson's correlation coefficient, Rank-based Spearman's correlation coefficient, and Concordance-based Kendall's correlation coefficient. The Spearman's and Kendall's correlation measures are rank-based and the Pearson's correlation measure is a measure of the linear relationship between two variables (in our case, the LCCDC and BWC metrics) [6]. The Pearson's measure captures the correlation between the two metrics as follows: If we were to list the vertices in the monotonically increasing order of their BWC values, are the LCCDC values of these vertices are also in the monotonically increasing order or decreasing order or neither. The Spearman's measure captures the correlation as follows: How close is the ranking of the vertices based on the increasing order of their BWC values and in the increasing order of their LCCDC values? Kendall's measure captures the correlation between the two metrics as follows: Consider any two vertices v_i and v_j. If BWC(v_i) $>$ BWC(v_j), is the LCCDC(v_i) $>$ LCCDC(v_j) or LCCDC(v_i) $<$ LCCDC(v_j) or LCCDC(v_i) = LCCDC(v_j)? All the three correlation measures are independent of each other. We use three different and independent correlation measures to more rigorously validate our hypothesis that the time-efficient LCCDC metric can be used to rank the nodes or compare any two nodes in a real-world network graph in lieu of the time-consuming BWC metric.

The correlation coefficient values obtained for all the three measures range from -1 to 1. Correlation coefficient values closer to 1 indicate a stronger positive correlation between the two metrics considered (i.e., a vertex having a larger value for one of the two metrics is more likely to have a larger value for the other metric too), while values closer to -1 indicate a stronger negative correlation (i.e., a vertex having a larger value for one of the two metrics is more likely to have a smaller value for the other metric). Correlation coefficient values closer to 0 indicate no correlation (i.e., the values incurred by a vertex for the two metrics are independent of each other). We will adopt the ranges (rounded to two decimals) proposed by Evans [9] to indicate the various levels of correlation, shown in Table 1. The color code to be used for the various levels of correlation are also shown in this table.

For simplicity, we refer to the two data sets as B and C, respectively, corresponding to the betweenness centrality and each of the other four centrality metrics (including the LCCDC). We will use the results from Fig. 3 to illustrate examples for the computation of the correlation coefficient under each of the three correlation measures.

4.1 Pearson's product moment-based correlation coefficient

The Pearson's product moment-based correlation coefficient for two data sets is defined as the covariance of the two data sets divided by the product of their standard deviation [5]. Let B_{avg} and C_{avg} denote the average values for the BWC and the LCCDC centrality metric for a graph of n vertices and let B_i and C_i denote, respectively, the values for the BWC and LCCDC incurred for vertex v_i. The Pearson's correlation coefficient (indicated PCC) is quantitatively defined as shown in Eq. (1). The term product moment is associated with the product of the mean (first moment) adjusted values for the two metrics in the numerator of the formulation. Figure 4 presents the calculation of the PCC for the betweenness centrality (B) and local clustering coefficient-based degree centrality (C) values obtained for the example graph used in Figs. 1, 2, 3. We obtain a correlation coefficient value of 0.97 (see Fig. 4) indicating a very strong positive correlation between the two metrics for the example graph.

$$\text{PCC}(B, C) = \frac{\sum_{i=1}^{n} (B_i - B_{\text{avg}})(C_i - C_{\text{avg}})}{\sqrt{\sum_{i=1}^{n} (B_i - B_{\text{avg}})^2 \sum_{i=1}^{n} (C_i - C_{\text{avg}})^2}} \ldots \quad (1)$$

4.2 Spearman's rank-based correlation coefficient

Spearman's rank correlation coefficient (SCC) is a measure of how well the relationship between two data sets (variables) can be assessed using a monotonic function [5]. To compute the SCC of two data sets $B$and C, we convert the raw scores B_i and C_i for a vertex $i$to ranks b_i and c_i and use formula (2)

Table 1 Range of correlation coefficient values and the corresponding levels of correlation

Range of Correlation Coefficient Values	Level of Correlation		Range of Correlation Coefficient Values	Level of Correlation
0.80 to 1.00	Very Strong Positive		-1.00 to -0.80	Very Strong Negative
0.60 to 0.79	Strong Positive		-0.79 to -0.60	Strong Negative
0.40 to 0.59	Moderate Positive		-0.59 to -0.40	Moderate Negative
0.20 to 0.39	Weak Positive		-0.39 to -0.20	Weak Negative
0.00 to 0.19	Very Weak Positive		-0.19 to -0.01	Very Weak Negative

shown below, where $d_i = b_i - c_i$ is the difference between the ranks of vertex i in the two data sets. We follow the convention of assigning the rank values from 1 to n for a graph of n vertices, even though the vertex IDs range from 0 to $n-1$. To obtain the rank for a vertex based on the list of values for a centrality metric, we first sort the values (in ascending order). If there is any tie, we break the tie in favor of the vertex with a lower ID; we will thus be able to arrive at a tentative, but unique, rank value for each vertex with respect to the centrality metric. We determine a final ranking of the vertices as follows: For vertices with unique value of the centrality metric, the final ranking is the same as the tentative ranking. For vertices with an identical value for the centrality metric, the final ranking is assigned to be the average of their tentative rankings. Figure 5 illustrates the computation of the tentative and final ranking of the vertices based on their betweenness centrality and local clustering coefficient-based degree centrality values in the example graph used in Figs. 1, 2, 3, 4 as well as illustrates the computation of the Spearman's rank-based correlation coefficient.

$$\mathrm{SCC}(B, C) = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \ldots. \tag{2}$$

In Fig. 5, we observe ties among vertices with respect to both BWC and LCCDC. The tentative ranking is obtained by breaking the ties in favor of vertices with lower IDs. In the case of BWC (B), we observe the 3 vertices 0, 6, and 7 to have an identical BWC value of 0 each and their tentative rankings are, respectively, 1, 2, and 3 (ties for tentative rankings are broken in favor of vertices with lower IDs); the final ranking (2) of each of these 3 vertices is thus the average of 1, 2, and 3. A similar scenario could be observed for LCCDC: vertices 0, 6, and 7 have an identical LCCDC value of 0 each and the final ranking of each of these three vertices is 2, based on their tentative rankings of 1, 2, and 3. The Spearman's rank-based correlation coefficient (SCC) computed for maximal clique size and degree centrality for the example graph used from Figs. 1, 2, 3, 4 is 0.98. We observe the SCC value to be slightly larger than the PCC value obtained in Fig. 4 for the same graph and the level of correlation for both the measures falls in the range of very strong positive correlation.

4.3 Kendall's concordance-based correlation coefficient

The Kendall's concordance-based correlation coefficient (KCC) for any two centrality metrics (say, B and C) is a measure of the similarity (a.k.a. concordance) in the ordering of the values for the metrics incurred by the vertices in the graph [5]. We define a pair of distinct vertices v_i and v_j as concordant if $\{B_i > B_j$ and $C_i > C_j\}$ or $\{B_i < B_j$ and $C_i < C_j\}$. In other words, a pair of vertices v_i and v_j are concordant if either one of these two vertices strictly have a larger value for the two metrics B and C compared to the other vertex. We define a pair of distinct vertices v_i and v_j as discordant if $\{B_i > B_j$ and $C_i < C_j\}$ or $\{B_i < B_j$ and $C_i > C_j\}$. In other words, a pair of vertices v_i and v_j are discordant if a vertex has a larger value for only one of the two centrality metrics. A pair of distinct vertices v_i and v_j are neither concordant nor discordant if either $\{B_i = B_j\}$ or $\{C_i = C_j\}$ or $\{B_i = B_j$ and $C_i = C_j\}$. The Kendall's concordance-based correlation coefficient is simply the difference between the number of concordant pairs (denoted #*conc.pairs*) and the number of discordant pairs (#*disc.pairs*) divided by the total number of pairs considered. For a graph of $n$vertices, KCC is calculated as shown in formulation (3).

$$\mathrm{KCC}(B, C) = \frac{\#conc.pairs - \#disc.pairs}{\frac{1}{2}n(n-1)} \ldots. \tag{3}$$

Figure 6 illustrates the calculation of the Kendall's correlation coefficient between BWC and LCCDC for the example graph used in Figs. 1, 2, 3, 4, 5. For a graph of 8 vertices, the total number of distinct pairs that could be considered is $8(8-1)/2 = 28$, and out of these, 25 pairs are classified to be concordant and just 1 pair as discordant (this itself is a direct indication of the very strong positive correlation between BWC and LCCDC). The remaining 2 pairs are neither concordant nor discordant (denoted as N/A) in the figure. We get a correlation coefficient of 0.86: still falling in the range of very strong positive correlation, though the absolute value of the correlation coefficient is lower than the correlation coefficient values obtained with the Pearson's and Spearman's measures. The KCC is also observed to return the lowest correlation coefficient values for all our experiments with the real-world networks (Sect. 5). Thus, the KCC could be construed to provide a lower bound for the correlation coefficient values and the level of correlation between BWC and the centrality metrics considered.

5 Real-world network graphs

We consider a suite of 18 real-world network graphs for our correlation analysis. We list below and identify these graphs in the increasing order of their variation in node degree, captured in the form of a metric called the spectral radius ratio for node degree (denoted λ_{sp}) [10]. The spectral radius ratio for node degree for a graph is the ratio of the principal eigenvalue of the adjacency matrix of the graph to that of the average node degree. The λ_{sp} values are always greater than or equal to 1.0. The larger the value, the larger the variation in node degree. The λ_{sp} values of the real-world networks considered in this paper range from 1.01 to 3.48 (i.e., from random networks to scale-free networks). Random networks exhibit a Poisson-style degree distribution and have a lower

Table 2 Fundamental properties of the real-world network graphs used in the correlation studies

#	Net.	λ_{sp}	#nodes	#edges	k_{avg}	G_c	D	PL_{avg}	G_a	G_m	CC_{avg}	#*comps*
1	FON	1.01	115	613	10.7	1.46	4	2.51	0.191	0.604	0.403	1
2	EAN	1.12	77	1549	40.2	10.6	2	1.47	−0.040	0.211	0.770	1
3	FTC	1.21	48	170	7.1	0.68	5	2.40	−0.014	0.455	0.438	1
4	RFN	1.27	217	1839	16.9	1.71	4	2.40	0.097	0.431	0.363	1
5	SJF	1.29	75	155	4.1	0.29	7	3.49	0.030	0.595	0.322	1
6	UKF	1.35	81	577	14.2	1.33	4	2.10	0.039	0.449	0.574	1
7	PBN	1.42	105	441	8.4	0.32	7	3.08	−0.023	0.521	0.488	1
8	BJN	1.45	198	2742	27.7	0.57	6	2.24	0.031	0.444	0.633	1
9	TFF	1.49	50	122	3.3	0.10	8	2.65	0.363	0.741	0.599	4
10	HCN	1.66	74	302	7.9	0.67	4	2.14	0.030	0.546	0.854	4
11	KFP	1.70	39	85	4.3	0.10	10	3.23	0.241	0.448	0.361	5
12	LMN	1.82	77	254	6.6	0.21	5	2.64	−0.077	0.553	0.736	1
13	CFN	1.83	87	407	9.1	0.98	3	1.95	−0.166	0.372	0.777	2
14	MTB	1.95	70	295	9.2	0.33	2	1.85	0.029	0.380	0.794	1
15	FBN	2.29	187	939	10.0	0.10	7	3.07	0.349	0.687	0.631	21
16	AKN	2.48	138	494	7.1	0.33	5	2.45	−0.081	0.371	0.798	2
17	ERN	3.00	472	1314	6.1	0.05	11	4.02	0.182	0.534	0.347	3
18	SJC	3.48	475	625	2.6	0.03	17	6.49	0.350	0.945	0.818	104

variation in node degree; their λ_{sp} values are typically closer to 1.0. Scale-free networks have a larger variation in node degree (especially those like the airline networks that have a few hubs—high degree nodes, and the rest of the nodes are of relatively much lower degree)—incurring a larger λ_{sp} value.

The real-world network graphs are briefly introduced below, in the increasing order of their λ_{sp} value. We also identify these networks with their ID (ranging from 1 to 18 as listed below) as well as with a three-character abbreviation—listed along with the λ_{sp} value. Table 2 lists the values for the following fundamental properties for each of these networks: average degree (k_{avg}), algebraic connectivity (G_c) [11], diameter (D), average path length (PL_{avg}), assortativity (G_a) [12], modularity (G_m) [13], average clustering coefficient (CC_{avg}) [1], and number of components (#*comps*). The values for each of the above properties for the real-world network graphs were obtained using our own implementation of the algorithms to determine these properties and their validity is verified using the Gephi [14] tool. We restrict ourselves to networks of moderate size due to the excessive computation time involved in computing the betweenness centrality for larger networks. In addition, we restrict ourselves to undirected network graphs (i.e., those that have a symmetric adjacency matrix) for the analysis conducted in this paper. Note that betweenness centrality is a symmetric centrality metric (i.e., unlike in-degree and out-degree, there do not exist in and out versions of BWC).

1. US Football Network (FON; $\lambda_{sp} = 1.01$) [15]: this is a network of 115 football teams (nodes) of US universities that played in the Fall 2000 season; there is an edge between two nodes if the corresponding teams have played against each other in the league games.
2. Employee Awareness Network (EAN; $\lambda_{sp} = 1.12$) [16]: this is a network of 77 employees (nodes) from a research team in a manufacturing company; there exists an edge between two nodes if the two employees are aware of each other's knowledge and skills.
3. Flying Teams Cadet Network (FTC; $\lambda_{sp} = 1.21$) [17]: this is a network of 48 cadet pilots (vertices) at an US Army Air Forces flying school in 1943, and the cadets were trained in a two-seated aircraft; there exists an edge between two vertices if at least one of the two corresponding cadet pilots have identified the other pilot among his/her preferred partners with whom she/he likes to fly during the training schedules.
4. Residence Hall Friendship Network (RFN; $\lambda_{sp} = 1.27$) [18]: this is a network of 217 residents (vertices) living at a residence hall located on the Australian National University campus. There exists an edge between two vertices if the corresponding residents are friends of each other.
5. San Juan Sur Family Network (SJF; $\lambda_{sp} = 1.29$) [19]: this is a network of 75 families (vertices) in San Juan Sur, Costa Rica, 1948. There exists an edge between two vertices if at least one of the two corresponding

families have visited the other family's household at least once.

6. UK Faculty Friendship Network (UKF; $\lambda_{sp} = 1.35$) [20]: this is a network of 81 faculty (vertices) at a UK university. There exists an edge between two vertices if the corresponding faculty are friends of each other.
7. US Politics Books Network (PBN; $\lambda_{sp} = 1.42$) [21]: this is a network of books (vertices) about US politics sold by Amazon.com around the time of the 2004 US presidential election. There exists an edge between two vertices if the corresponding two books were co-purchased by the same buyer (at least one buyer).
8. Jazz Band Network (JBN; $\lambda_{sp} = 1.45$) [22]: this is a network of 198 Jazz bands (vertices) that recorded between the years 1912 and 1940; there exists an edge between two bands if they shared at least one musician in any of their recordings during this period.
9. Teenage Female Friendship Network (TFF; $\lambda_{sp} = 1.49$) [23]: this is a network of 50 female teenage students (vertices) who studied as a cohort in a school in the West of Scotland from 1995 to 1997. There exists an edge between two vertices if the corresponding students reported (in a survey) that they were best friends of each other.
10. Huckleberry Coappearance Network (HCN; $\lambda_{sp} = 1.66$) [24]: this is a network of 74 characters (vertices) that appeared in the novel Huckleberry Finn by Mark Twain; there is an edge between two vertices if the corresponding characters had a common appearance in at least one scene.
11. Korea Family Planning Network (KFP; $\lambda_{sp} = 1.69$) [25]: this is a network of 39 women (vertices) at a Mothers' Club in Korea; there existed an edge between two vertices if the corresponding women were seen discussing family planning methods during an observation period.
12. Les Miserables Network (LMN; $\lambda_{sp} = 1.81$) [24]: this is a network of 77 characters (nodes) in the novel Les Miserables; there exists an edge between two nodes if the corresponding characters appeared together in at least one of the chapters in the novel.
13. Copperfield Network (CFN; $\lambda_{sp} = 1.83$) [26]: this is a network of 87 characters in the novel David Copperfield by Charles Dickens; there exists an edge between two vertices if the corresponding characters appeared together in at least one scene in the novel.
14. Madrid Train Bombing Network (MTB; $\lambda_{sp} = 1.95$) [27]: this is a network of suspected individuals and their relatives (vertices) reconstructed by Rodriguez using press accounts in the two major Spanish daily newspapers (El Pais and El Mundo), regarding the bombing of commuter trains in Madrid on March 11, 2004. There existed an edge between two vertices if the corresponding individuals were observed to have a link in the form of friendship, ties to any terrorist organization, co-participation in training camps and/or wars, or co-participation in any previous terrorist attacks.
15. Facebook Network (FBN; $\lambda_{sp} = 2.29$): this is a network of the 187 friends (vertices) of the author in the well-known social media network, Facebook [28]. There exists an edge between two nodes if the corresponding people are also friends of each other.
16. Anna Karnenina Network (AKN; $\lambda_{sp} = 2.47$) [24]: this a network of 138 characters (vertices) in the novel Anna Karnenina; there exists an edge between two vertices if the corresponding characters have appeared together in at least one scene in the novel.
17. Erdos Collaboration Network (ECN; $\lambda_{sp} = 3.00$) [29]: this is a network of 472 authors (nodes) who have either directly published an article with Paul Erdos or through a chain of collaborators leading to Paul Erdos. There is an edge between two nodes if the corresponding authors have co-authored at least one publication.
18. Social Journal Network (SJN; $\lambda_{sp} = 3.48$) [30]: this is a network of 475 authors (vertices) involved in the production of 295 articles for the Social Networks Journal, since its inception until 2008; there is an edge between two vertices if the corresponding authors co-authored at least one paper published in the journal.

We measured the execution time incurred (measured in milliseconds) to compute each of the 5 centrality metrics: LCCDC, DegC, BWC, EVC, and ClC for the above 18 real-world networks. The executions were conducted on a computer with Intel Core i7-2620M CPU @ 2.70 GHz and an installed main memory (RAM) of 8 GB. We ran the procedures for each of these 5 centrality metrics on each of the real-world networks for 20 iterations and averaged the results. Table 3 lists the raw values for the average execution time (in milliseconds) for each of the 5 centrality metrics on the 18 real-world networks. Figure 7 plots the natural logarithm of the average execution time (for the values to be plotted on a comparable scale) incurred for the centrality metrics on each of the real-world networks. While the networks are listed in Table 3 and Fig. 7 in the increasing order of their spectral radius ratio for node degree (the same order as in Table 2); for each network, the centrality metrics are shown in the decreasing order of the execution times. Overall, we observe that networks with a larger number of nodes incur a larger execution time; for networks with comparable number of nodes, the execution time for the centrality metrics increases with increase in the edge-node ratio (ratio of the number of nodes to the number of edges), especially to compute the time-consuming centrality metrics, such as the

Table 3 Average execution time to compute the centrality metrics for the real-world network graphs

#	Net.	# nodes	Edge-node ratio	Average execution time to compute the centrality metrics (ms)				
				BWC	EVC	ClC	LCCDC	DegC
1	FON	115	5.33	166149.5	6229.7	1403.8	136.5	26.2
2	EAN	77	20.12	61915.4	3203.2	582.8	459.2	17.5
3	FTC	48	3.54	9694.8	921.4	136.8	25.8	10.6
4	RFN	217	8.47	2,198,077.4	54,264.1	8925.4	472.1	50.6
5	SJF	75	2.07	33,514.1	1924.1	407.1	36.9	17.1
6	UKF	81	7.12	56,355.1	2321.3	507.5	133.0	18.2
7	PBN	105	4.20	1,16,321.7	4802.3	992.1	94.3	24.3
8	BJN	198	13.85	1,970,503.6	74,771.2	17,774.7	12,137.8	56.3
9	TFF	50	2.44	4527.5	548.1	109.5	13.7	9.5
10	HCN	74	4.08	25,299.1	1520.1	347.8	62.3	17.1
11	KFP	39	2.18	3782.5	318.9	76.8	13.3	7.2
12	LMN	77	3.30	35,168.4	1361.2	470.2	42.2	17.1
13	CFN	87	4.68	56,355.1	2321.3	507.5	133.0	18.2
14	MTB	70	4.21	23,998.0	1170.4	308.7	52.8	16.7
15	FBN	187	5.02	8,17,865.3	24,435.9	5166.5	184.4	40.3
16	AKN	138	3.58	3,96,377.5	1,54,722.5	27,190.2	1270.7	33.8
17	ERN	472	2.78	23,106,718.9	5,38,524.0	81,444.4	1238.4	100.7
18	SJC	475	1.32	14,564,978.3	3,49,242.1	82,584.8	181.3	89.7

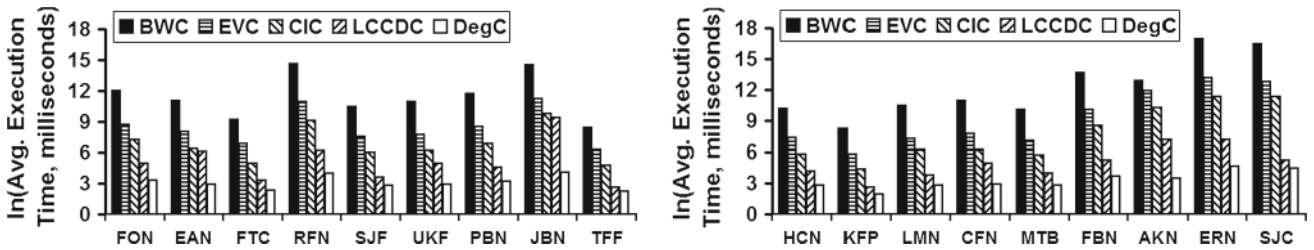


Fig. 7 Average execution time to compute the centrality metrics for the real-world network graphs (natural logarithm scale)

BWC and EVC. Table 3 and Fig. 7 display a clear ranking of the centrality metrics with respect to the execution time: BWC and DegC incur, respectively, the largest and smallest values for the average execution time for each real-world network analyzed. As the LCCDC values are computed by making use of the DegC values, it is natural to expect the execution time of the procedure to compute the LCCDC values to be larger than that of the DegC values. The execution time of the degree centrality metric appears to be anywhere from 0.4–69 % of the execution time of the LCCDC metric.

From Table 3 and Fig. 7, we could clearly observe the LCCDC metric to consistently incur a lower execution time compared to the BWC, EVC, and ClC metrics for each of the real-world networks analyzed. We observe the execution time incurred to compute the LCCDC metric to be significantly smaller than that of the BWC metric. The ratio of the average execution time for computing the BWC and LCCDC values for the real-world networks ranges from 117 to 80,330. The ClC metric incurs an execution time that is at least 25 % larger than the execution time of the LCCDC metric and appears to be even significantly larger for several real-world networks evaluated. The EVC metric incurs an execution time that is 6 to 1926 times larger than the execution time of the LCCDC metric. Considering all of the above, our claim that LCCDC is a computationally lightweight metric is well justified.

Table 4 presents the raw values for the correlation coefficient obtained for the Betweenness centrality metric and each of the four centrality metrics: LCCDC, DegC, EVC, and ClC based on the PCC, SCC, and KCC measures. We color code the levels of correlation in Table 4 according to the color codes listed in Table 1. Under all the three correlation measures, we observe the proposed LCCDC metric to demonstrate significantly larger correlation coefficient values with BWC vis-a-vis the correlation coefficient values incurred by the other centrality metrics. Among the three correlation measures, the Spearman's rank-based correla-

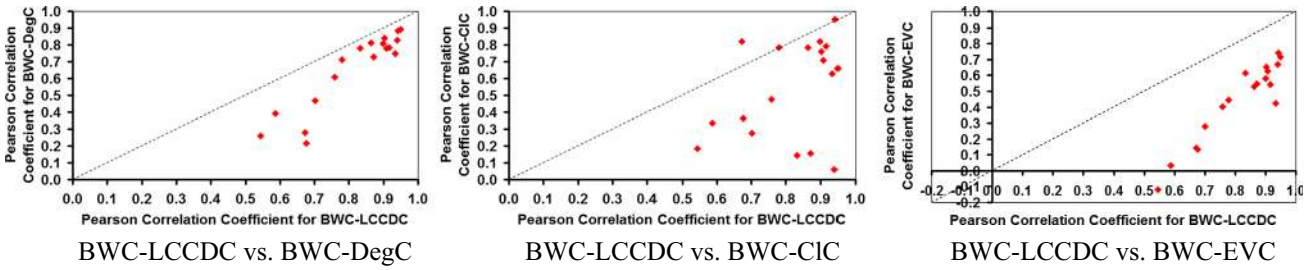
Table 4 Correlation coefficient values between betweenness centrality and the other centrality metrics for real-world network graphs

		Pearson Correlation Coeff.				Spearman Correlation Coeff.				Kendall's Correlation Coeff.			
#	Net.	LCC DC	Deg C	ClC	EVC	LCC DC	Deg C	ClC	EVC	LCC DC	Deg C	ClC	EVC
1	FON	0.67	0.28	0.82	0.15	0.61	0.40	0.84	0.17	0.44	0.20	0.65	0.12
2	EAN	0.94	0.89	0.95	0.74	1.00	0.83	0.83	0.68	0.95	0.69	0.69	0.57
3	FTC	0.92	0.78	0.79	0.54	0.92	0.73	0.80	0.41	0.77	0.55	0.61	0.30
4	RFN	0.90	0.84	0.76	0.65	0.93	0.84	0.86	0.62	0.79	0.66	0.67	0.45
5	SJF	0.86	0.81	0.79	0.53	0.85	0.73	0.77	0.41	0.66	0.52	0.57	0.29
6	UKF	0.91	0.78	0.71	0.63	0.95	0.79	0.75	0.60	0.82	0.61	0.57	0.45
7	PBN	0.78	0.71	0.78	0.44	0.86	0.68	0.81	0.37	0.69	0.49	0.61	0.26
8	BJN	0.76	0.61	0.48	0.40	0.86	0.74	0.73	0.57	0.71	0.57	0.56	0.42
9	TFF	0.68	0.22	0.36	0.14	0.88	0.46	0.47	-0.19	0.61	0.29	0.34	-0.11
10	HCN	0.94	0.83	0.06	0.67	0.92	0.70	0.69	0.65	0.55	0.41	0.41	0.37
11	KFP	0.70	0.47	0.28	0.28	0.80	0.51	0.61	0.40	0.62	0.35	0.46	0.26
12	LMN	0.93	0.75	0.63	0.42	0.88	0.77	0.68	0.72	0.60	0.48	0.43	0.43
13	CFN	0.90	0.81	0.82	0.58	0.95	0.83	0.77	0.77	0.73	0.60	0.55	0.53
14	MTB	0.87	0.73	0.15	0.55	0.91	0.76	0.68	0.56	0.64	0.53	0.46	0.35
15	FBN	0.54	0.26	0.18	-0.12	0.86	0.58	0.70	-0.22	0.67	0.40	0.52	-0.14
16	AKN	0.95	0.89	0.66	0.72	0.88	0.78	0.66	0.69	0.54	0.49	0.39	0.41
17	ERN	0.83	0.78	0.15	0.62	0.92	0.86	0.72	0.64	0.69	0.63	0.51	0.44
18	SJC	0.59	0.39	0.34	0.03	0.78	0.65	0.56	0.16	0.29	0.22	0.19	-0.08

tion measure yields the largest values for the correlation coefficient between LCCDC and BWC, such that the level of correlation is very strongly positive for 16 of the 18 networks analyzed and strongly positive for the remaining two networks. Similarly, with respect to the Pearson's product moment-based correlation measure, we observe the LCCDC metric to exhibit correlation levels of strongly to very strongly positive for 16 of the 18 networks (11 networks exhibit very strongly positive correlation and 5 networks exhibit strongly positive correlation). The Kendall's concordance-based correlation measure yields the lowest values for the correlation coefficient between BWC and the other centrality metrics. Nevertheless, even under the Kendall's correlation measure: we observe the LCCDC metric to exhibit strong to very strong positive correlation with BWC for 14 of the 18 real-world networks analyzed. Overall, considering all the three correlation measures, we could say that the LCCDC metric exhibits strong to very strong levels of positive correlation for at least 14 of the 18 real-world networks analyzed. Such a high level of correlation with BWC is not observed for the other three centrality metrics analyzed in this paper, as well as for any other network analysis metric in the literature.

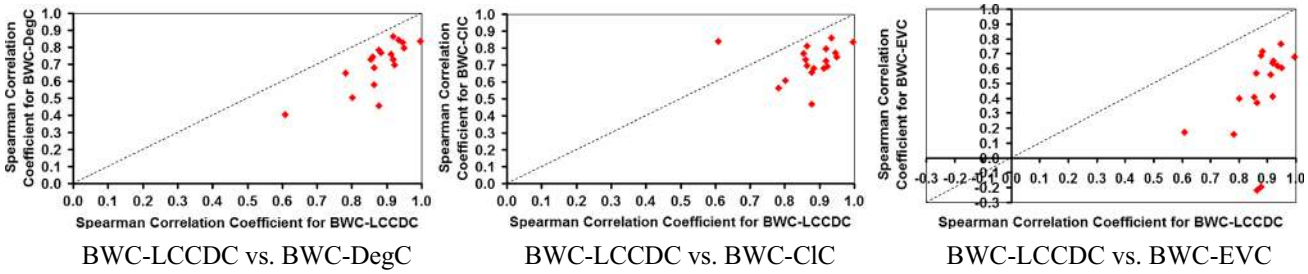
Figures 8, 9 and 10 compare the relative magnitude of the values for the correlation coefficient (based on the proximity of the data points to the diagonal line in these figures) obtained for BWC-LCCDC with each of the other three combinations of centrality metrics: BWC-DegC, BWC-ClC, and BWC-EVC under each of the three correlation measures. Each data point in these figures corresponds to a particular real-world network. If a data point is below the diagonal line, it implies the correlation coefficient incurred for BWC-LCCDC is larger than the correlation coefficient incurred for the BWC-centrality metric combination for the real-world network that the data point represents. If a data point lies above the diagonal line, it implies the BWC-LCCDC correlation coefficient is lower than the BWC-centrality metric combination for the corresponding real-world network. If a data point lies on the diagonal line, it implies the correlation coefficient values are almost equal. Among the other three centrality metrics analyzed (see Figs. 8, 9, 10 for a comparison), the degree centrality metric exhibits relatively higher levels of correlation with BWC. Nevertheless, when compared to the correlation coefficient values incurred for BWC-LCCDC, the BWC-DegC correlation coefficient values are at least lower by 0.05 (in a scale of -1 to 1) for all the 18 real-world networks and lower by at least 0.10 for at least 10 of the 18 real-world networks under each of the three correlation measures.

The only centrality metric which exhibits correlation coefficient values (with BWC) matching or exceeding to that incurred for LCCDC-BWC for at least one of the real-world networks under at least one of the three correlation measures is the closeness centrality (ClC) metric. The best case scenario for ClC is that there exists just one real-world network (among the 18 networks analyzed) for which the BWC-ClC correlation coefficient is larger than the BWC-LCCDC correlation coefficient under all the three correlation measures;



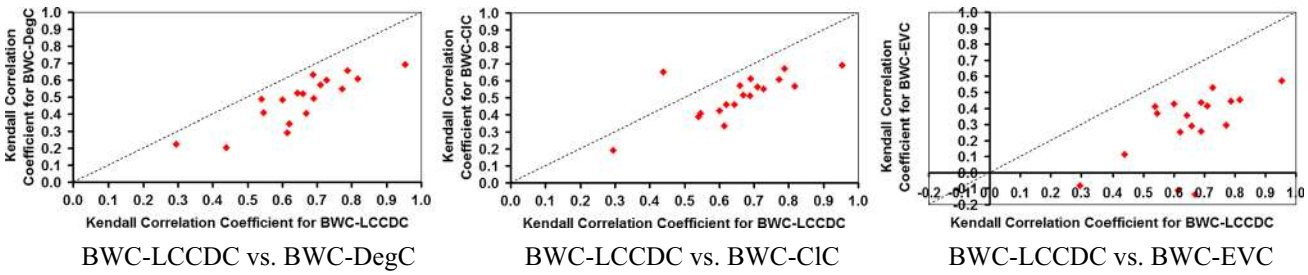
BWC-LCCDC vs. BWC-DegC BWC-LCCDC vs. BWC-ClC BWC-LCCDC vs. BWC-EVC

Fig. 8 Distribution of the correlation coefficient values for real-world networks under the Pearson's product moment-based correlation measure (from the centrality metrics viewpoint)



BWC-LCCDC vs. BWC-DegC BWC-LCCDC vs. BWC-ClC BWC-LCCDC vs. BWC-EVC

Fig. 9 Distribution of the correlation coefficient values for real-world networks under the Spearman's Rank-based correlation measure (from the centrality metrics viewpoint)



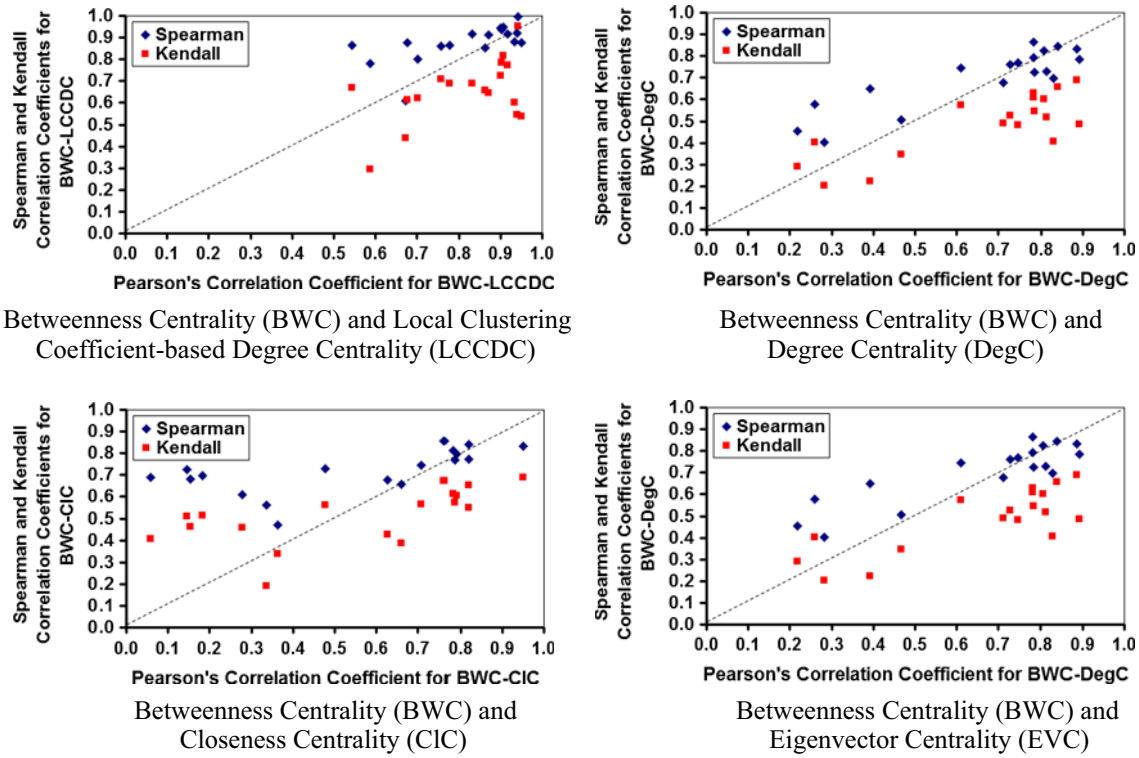
BWC-LCCDC vs. BWC-DegC BWC-LCCDC vs. BWC-ClC BWC-LCCDC vs. BWC-EVC

Fig. 10 Distribution of the correlation coefficient values for real-world networks under the Kendall's concordance-based correlation measure (from the centrality metrics viewpoint)

in addition, under the Pearson's and Spearman's correlation measures: the correlation coefficient values incurred for ClC with BWC equal to those incurred for LCCDC with BWC for two of the 18 real-world networks. Note that the closeness centrality metric is relatively more computation-intensive (a shortest path algorithm needs to be run at every vertex), as is also vindicated by the results in Table 3 and Fig. 7. The Eigenvector centrality (EVC) metric exhibits relatively lower levels of correlation with BWC among all the centrality metrics analyzed and under all the three correlation measures. This could be attributed to the relatively larger clustering coefficient values incurred for vertices with higher EVC. A node i with a higher EVC is more likely surrounded by nodes having higher degree: a majority of these nodes could be directly connected to each other and there would be no need to go through node i. As a result, vertices with higher EVC are very less likely to lie on the shortest path for their neighbor nodes.

Among the three correlation measures used to evaluate the correlation of BWC with LCCDC and the other centrality metrics, we observe the Spearman's measure to yield correlation coefficient values that are relatively more closer to that of the Pearson's measure. This could be deduced by observing the relative proximity of the data points to the diagonal line in Fig. 11: the data points corresponding to the Spearman's and Pearson's correlation measures are relatively more closer to the diagonal line when compared to the data points corresponding to the Kendall's and Pearson's correlation measures. Overall, for a majority of the real-world networks analyzed, the Spearman's and Kendall's correlation measures appear to, respectively, provide the upper bound and lower bound for the values of the correlation coefficient (and the correlation levels) incurred between BWC and each of the other four centrality metrics.

With respect to the impact of the variation in node degree on the correlation levels, overall: we observe the level of cor-



Betweenness Centrality (BWC) and Local Clustering Coefficient-based Degree Centrality (LCCDC)

Betweenness Centrality (BWC) and Degree Centrality (DegC)

Betweenness Centrality (BWC) and Closeness Centrality (ClC)

Betweenness Centrality (BWC) and Eigenvector Centrality (EVC)

Fig. 11 Distribution of the correlation coefficient values for real-world network graphs (from the correlation measures viewpoint)

relation between BWC and each of the four centrality metrics to decrease with increase in the spectral radius ratio for node degree (more predominantly observed with the Kendall correlation measure and to a certain extent with the Pearson's and Spearman's correlation measures). A high-level view of the results in Table 4 indicates that the correlation level tends to reduce from a higher positive level to a relatively lower level as the spectral radius ratio for node degree of the real-world network graphs increases. As the networks become increasingly scale-free (i.e., the variation in node degree in the network increases), the trend we could deduce is a decrease in the correlation coefficient values between BWC and each of the four centrality metrics (especially in the case of Eigenvector centrality under all the three correlation measures).

6 Related work

Several centrality metrics have been proposed for the complex network analysis. UCINET 6 [31] employs the following eight of these centrality metrics: degree, betweenness, closeness, eigenvector, power, information, flow, and reach. As mentioned earlier, the most frequently used centrality metrics are: degree, closeness, betweenness, and eigenvector. In one of the first studies on correlations among centrality metrics, Bolland [32] observed that degree centrality and closeness centrality are highly correlated, while the betweenness centrality is relatively uncorrelated with degree, and closeness and eigenvector centralities. Rothenberg et al. [33] observed the information centrality and distance metrics (eccentricity, mean, and median of the path length between any two vertices) to be not so strongly correlated with the degree and betweenness centrality metrics. Rotherberg et al. [33] observed the degree centrality to be the most strongly correlated metric with betweenness centrality: we also observe that next to LCCDC, the degree centrality could be claimed as the centrality metric that exhibits stronger correlation with BWC. With respect to the impact of symmetry in the adjacency matrix on the correlation levels observed, Valente et al. [34] observed that the disparity between symmetric centrality metrics (like betweenness) and asymmetric centrality metrics (like degree) increases when computed on the undirected instances of directed network graphs.

For scale-free networks [35], the distribution of the betweenness centrality of the vertices has been observed to follow a power-law pattern (similar to that of the degree centrality) [37]. It was also observed in [38] that for scale-free networks that are either dissortative [12] or neutral with respect to node degree, the average of the betweenness centralities of the neighbors of a vertex is proportional to the

betweenness centrality of the vertex considered; whereas, for assortative scale-free networks, the betweenness centralities of the neighbors of a vertex is independent of the betweenness centrality of the vertex considered.

Among the various localized centrality metrics proposed in the literature, the "leverage" centrality metric proposed by Joyce et al. [39] for brain networks has gained prominence. Leverage centrality of a node is a measure of the extent of connectivity of the node relative to the connectivity of its neighbors. For a node i with degree k_i and set of neighbors N_i, the leverage centrality of node i, $\mathrm{LVC}(i) = \frac{1}{k_i} \sum_{j \in N_i} \frac{k_i - k_j}{k_i + k_j}$ [39]. Leverage centrality is based on the notion that a node with degree higher than the degree of its neighbors is likely to be more influential on its neighbors and vice-versa. The above formulation for LVC restricts its use only for vertices with degree 1 or above and not applicable for isolated vertices. On the other hand, our proposed LCCDC metric (also a localized centrality metric) could be computed for any vertex and the entire network graph need not be just one single connected component. Moreover, the above formulation for leverage centrality metric compares the degree of a node with the degree of an individual neighbor node, and fails to take into consideration the connectivity among the neighbor nodes themselves (without involving the node in consideration). Hence, the leverage centrality metric cannot be a suitable alternate for the betweenness centrality (BWC) metric, as is also evidenced in the correlation studies of [39]: the correlation between leverage centrality and BWC is lower than the correlation between degree centrality and BWC. On the other hand, we observe that the correlation between LCCDC and BWC is even stronger than the correlation between degree centrality and BWC that has been observed in the literature until now. Thus, our proposed LCCDC metric is significantly different from that of the leverage centrality, closeness centrality, and the other centrality metrics.

Li et al. [40] conducted an extensive correlation study for the centrality metrics on 34 real-world network graphs as well as the theoretical graphs generated from the Erdos-Renyi (ER; for random networks) [41] and Barabasi-Albert (BA; for scale-free networks) [36] models. It has been observed in [40] that the degree centrality metric exhibits the strongest levels of correlation with the betweenness centrality metric for both the ER and BA networks. Likewise, for about two-thirds of the 34 real-world network graphs, the BWC-DegC correlation coefficient values were observed to be the largest incurred compared to the correlation coefficient values incurred for BWC-ClC, BWC-LVC, and BWC-EVC. Unlike our paper, the correlation study in Li et al. [40] has been conducted only with the Pearson's product moment-based correlation measure. We observe from the results of this paper that the Kendall's concordance-based correlation measure gives a lower estimate for the levels of correlation between any two centrality metrics. The LCCDC metric withstands the test with respect to all the three correlation measures and consistently incurs larger values for the correlation coefficient with BWC compared to the correlation coefficient values incurred for any other centrality metric with BWC.

7 Conclusions

The high-level contribution of this paper is the proposal of a localized, computationally lightweight alternate centrality metric for the computation-intensive betweenness centrality (BWC) metric that is widely used for the complex network analysis. We effectively magnify the importance of a node to connect its neighbors on the shortest path (evaluated through the local clustering coefficient) with the node's degree to assess its importance to connect any two nodes in the network on a shortest path. Our hypothesis is that nodes with higher degree, but lower local clustering coefficient, are more likely to be part of several shortest paths between any two node pairs in the network. Accordingly, we propose the local clustering coefficient-based degree centrality (LCCDC) for a vertex as the product of the degree of the vertex and one minus the local clustering coefficient. We observe the LCCDC to exhibit a strong-very strong positive correlation with BWC (under all the three correlation measures used) for a majority of the real-world network graphs analyzed. Even with the Kendall's concordance-based correlation measure (that is observed to return lower values for the correlation coefficient among the three correlation measures considered), we observe the LCCDC metric to exhibit strong-very strong levels of correlation with BWC for 14 of the 18 real-world networks analyzed (whereas the degree centrality and closeness centrality metrics could at most exhibit strong correlation with BWC for at most 4–5 of the 18 real-world networks analyzed). Under the Spearman's rank-based correlation measure, we observe the LCCDC to be very strongly correlated to BWC (correlation coefficient values of 0.80 or above) for 16 of the 18 real-world networks. Thus, we confidently claim that the LCCDC could effectively serve as an alternate metric for ranking the vertices of a graph in lieu of the BWC. To the best of our knowledge, we have not come across such a computationally lightweight centrality metric that is highly correlated with betweenness centrality. As part of future work, we will explore extending the application of the LCCDC metric (with appropriate modifications) for directed real-world network graphs as well as conduct a correlation study between LCCDC and BWC for network graphs generated from theoretical models (like the ER and BA models).

References

1. Newman, M.: Networks: an introduction, 1st edn. Oxford University Press, Oxfrod (2010)
2. Bonacich, P.: Power and centrality: a family of measures. Am. J. Sociol. **92**(5), 1170–1182 (1987)
3. Freeman, L.: A set of measures of centrality based on betweenness. Sociometry **40**(1), 35–41 (1977)
4. Freeman, L.: Centrality in social networks conceptual clarification. Soc. Netw. **1**(3), 215–239 (1979)
5. Triola, M.F.: Elementary statistics, 12th edn. Pearson, NY (2012)
6. Lay, D.C.: Linear algebra and its applications, 4th edn. Pearson, NY (2011)
7. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25**(2), 163–177 (2001)
8. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to algorithms, 3rd edn. MIT Press, Cambridge (2009)
9. Evans, J.D.: Straightforward Statistics for the Behavioral Sciences, 1st edn, Brooks Cole Publishing Company (1995)
10. Meghanathan, N.: Spectral radius as a measure of variation in node degree for complex network graphs,. In: Proceedings of the 7th international conference on u- and e- service, science and technology, pp. 30–33, Haikou, China (2014)
11. Maia de Abreu, N.M.: Old and new results on algebraic connectivity of graphs. Linear Algebra Appl. **423**(1), 53–73 (2007)
12. Newman, M.E.J.: Assortative mixing in networks. Phys. Rev. Lett. **89**(2), 208–701 (2002)
13. Newman, M.E.J.: Modularity and community structure in networks. J. Natl. Acad. Sci. USA **103**(23), 8557–8582 (2006)
14. Cherven, K.: Mastering Gephi network visualization. Packt Publishing, UK (2015)
15. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**(12), 7821–7826 (2002)
16. Cross, R.L., Parker, A., Cross, R.: The hidden power of social networks: understanding how work really gets done in organizations. 1st edn. Harvard Business Review Press, NY (2004)
17. Moreno, J.L.: The sociometry Reader, pp. 534–547, The Free Press, Glencoe (1960)
18. Freeman, L.C., Webster, C.M., Kirke, D.M.: Exploring social structure using dynamic three-dimensional color images. Soc. Netw. **20**(2), 109–118 (1998)
19. Loomis, C.P., Morales, J.O., Clifford, R.A., Leonard, O.E.: Turrialba social systems and the introduction of change, pp. 45–78, The Free Press, Glencoe (1953)
20. Nepusz, T., Petroczi, A., Negyessy, L., Bazso, F.: Fuzzy communities and the concept of bridgeness in complex networks. Phys. Rev. E **77**(1), 016107 (2008)
21. Krebs, V.: Proxy networks: analyzing one network to reveal another. Bulletin de Méthodologie Sociologique **79**, 40–61 (2003)
22. Geiser, P., Danon, L.: Community structure in Jazz. Adv. Complex Syst. **6**(4), 563–573 (2003)
23. Pearson, M., Michell, L.: Smoke rings: social network analysis of friendship groups, smoking and drug-taking. Drugs Educ. Prev. Policy **7**(1), 21–37 (2000)
24. Knuth, D.E.: The Stanford GraphBase: a platform for combinatorial computing, 1st edn. Addison-Wesley, Reading (1993)
25. Rogers, E.M., Kincaid, D.L.: Communication networks: toward a new paradigm for research, Free Press, USA (1980)
26. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**(3), 036104 (2006)
27. Hayes, B.: Connecting the dots. Am. Sci. **94**(5), 400–404 (2006)
28. Facebook Netvizz Application. https://apps.facebook.com/netvizz/
29. Pajek Datasets. http://vlado.fmf.uni-lj.si/pub/networks/data/
30. Freeman, L.: Datasets. http://moreno.ss.uci.edu/data.html
31. Borgatti, S.P., Everett, M.G., Johnson, J.C.: Analyzing social networks. 1st edn. SAGE Publications, UK (2013)
32. Bolland, J.M.: Sorting out centrality: an analysis of the performance of four centrality models in real and simulated networks. Soc. Netw. **10**(3), 233–253 (1988)
33. Rothenberg, R.B., Potterat, J.J., Woodhouse, D.E., Darrow, W.W., Muth, S.Q., Klovdahl, A.S.: Choosing a centrality measure: epidemiologic correlates in the colorado springs study of social networks. Soc. Netw. **17**(3-4), 273–297 (1995)
34. Valente, T.W., Coronges, K., Lakon, C., Costenbader, E.: How correlated are network centrality measures? Connections **28**(1), 16–26 (2008)
35. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
36. de Caen, D.: An upper bound on the sum of squares of degrees in a graph. Discret. Math. **185**(1–3), 245–248 (1998)
37. Goh, K., Oh, E., Jeong, H., Kahng, B., Kim, D.: Classification of scale-free networks. J. Natl. Acad. Sci. USA **99**(20), 12583–12588 (2002)
38. Goh, K., Oh, E., Kahng, B., Kim, D.: Betweenness centrality correlation in social networks. Phys. Rev. E **67**(1), 017101 (2003)
39. Joyce, K.E., Laurienti, P.J., Burdette, J.H., Hayasaka, S.: A new measure of centrality for brain networks. PLoS One **5**(8), e12200, 1–13 (2010)
40. Li, C., Li, Q., Van Mieghem, P., Stanley, H.E., Wang, H.: Correlation between centrality metrics and their application to the opinion model. Eur. Phys. J. B **88**(65), 1–13 (2015)
41. Erdos, P., Renyi, A.: On random graphs I. Publ. Math. **6**, 290–297 (1959)