Vietnam J Comput Sci
DOI 10.1007/s40595-017-0097-1

REGULAR PAPER

Concordance-based Kendall's correlation for computationally-light vs. computationally-heavy centrality metrics: lower bound for correlation

Natarajan Meghanathan[1]

Received: 30 January 2017 / Accepted: 16 June 2017

Abstract We identify three different levels of correlation (pairwise relative ordering, network-wide ranking, and linear regression) that could be assessed between a computationally-light centrality metric and a computationally-heavy centrality metric for real-world networks. The Kendall's concordance-based correlation measure could be used to quantitatively assess how well we could consider the relative ordering of two vertices v_i and v_j with respect to a computationally-light centrality metric as the relative ordering of the same two vertices with respect to a computationally-heavy centrality metric. We hypothesize that the pairwise relative ordering (concordance)-based assessment of the correlation between centrality metrics is the most strictest of all the three levels of correlation and claim that the Kendall's concordance-based correlation coefficient will be lower than the correlation coefficient observed with the more relaxed levels of correlation measures (linear regression-based Pearson's product–moment correlation coefficient and the network-wide ranking-based Spearman's correlation coefficient). We validate our hypothesis by evaluating the three correlation coefficients between two sets of centrality metrics: the computationally-light degree and local clustering coefficient complement-based degree centrality metrics and the computationally-heavy eigenvector centrality, betweenness centrality, and closeness centrality metrics for a diverse collection of 50 real-world networks.

Keywords Relative ordering · Ranking · Linear regression · Centrality · Correlation

✉ Natarajan Meghanathan
natarajan.meghanathan@jsums.edu

[1] Computer Science, Jackson State University, Jackson, MS 39217, USA

1 Introduction

Network Science deals with analyzing complex networks (e.g., biological networks, social networks, citation networks, web, etc.) from a graph theoretic perspective [1]. We model a complex network as an abstract graph of vertices (nodes) and edges (links). Centrality of a vertex is a quantitative measure of the topological significance of the vertex in a graph [1]. There exists a slew of centrality metrics for complex network analysis. Among these, the commonly studied metrics are the degree-based degree centrality (DegC) [1] and eigenvector centrality (EVC) metrics [2] and the shortest path-based betweenness centrality (BWC) [3] and closeness centrality (CLC) metrics [4]. The degree centrality of a vertex is a measure of the number of neighbors of the vertex. The eigenvector centrality [2] of a vertex is a measure of the degree of the vertex as well as the degree of its neighbors. A vertex has a higher EVC if it has a high-degree and its neighbors also have a high-degree. The betweenness centrality [3] of a vertex is a measure of the number of the shortest paths (between any two vertices in the network) that go through the vertex. The closeness centrality [4] of a vertex is a measure of the hop count of the shortest paths (or the weight of the shortest paths in a weighted graph) from the vertex to the rest of the vertices in a graph. For graphs that are not connected, the centrality metrics are typically computed for the largest connected component of the graph.

Among the above four centrality metrics (see Sect. 5 for a comparison of the computation time), the degree centrality metric is the only computationally-light metric (i.e., could be computed quickly) and the other three metrics are computationally-heavy (i.e., would take more computation time). For a graph of V vertices and E edges, it takes $\Theta(V^3)$ time to compute the EVC of the vertices [2] and $\Theta(V^2+VE)$ time to compute each of the BWC [3] and CLC metrics [4].

Recently, some research articles (e.g., [5,6]) have evaluated the correlation between these four commonly used centrality metrics for real-world network graphs to see if one or more of the computationally-heavy centrality metrics (EVC, BWC, and CLC) exhibit a strong correlation with the degree centrality metric (on the basis of the Pearson's correlation coefficient [7]), so that one could then employ linear regression to predict the strongly correlated computationally-heavy metric(s) using the degree centrality metric. On similar lines, the Pearson's correlation coefficient between each of the above four centrality metrics and the maximal clique size per node (another node-level computationally-heavy metric, the computation of which is a NP-hard problem) was evaluated in [8].

In a recent work [9], a new metric called the localized clustering coefficient complement-based degree centrality (LCC′DC) has been proposed as a computationally-light alternative to the computationally-heavy BWC metric. LCC′DC is computed as the product of 1-LCC and DegC, where LCC (local clustering coefficient) of a vertex [1] is a measure of the probability that any two neighbors of the vertex are connected and are computed as the ratio of the actual number of edges between the neighbors of a vertex to that of the maximum possible number of edges between the neighbors of the vertex. For several real-world networks analyzed in [9], the Pearson's correlation coefficient values observed for LCC′DC–BWC are larger than the correlation coefficient values observed for DegC–BWC. In another recent work [10], it was observed that compared to DegC, LCC′DC could be used to more accurately predict BWC values using linear regression (with the standard error of residual values smaller than those incurred for regression using DegC). In both [9] and [10], the correlation analysis was focused on BWC vs. the other four centrality metrics (DegC, LCC′DC, EVC, and CLC). In this paper, our correlation analysis is more comprehensive: the two computationally-light centrality metrics (DegC and LCC′DC) vs. the three computationally-heavy centrality metrics (EVC, BWC, and CLC).

In this paper, we identify three different levels of correlation that could be evaluated between any two centrality metrics of the vertices (more specifically, between a computationally-light metric and a computationally-heavy metric) in complex network graphs: (i) a pairwise relative ordering-based correlation that would be a quantitative measure of how well the relative ordering of a pair of vertices based on a computationally-light metric could be considered as the relative ordering of the same pair of vertices with respect to a computationally-heavy metric. For example, if $\text{LCC}'\text{DC}(v_i) < \text{LCC}'\text{DC}(v_j)$, how sure are we to say $\text{BWC}(v_i) < \text{BWC}(v_j)$ for some two vertices v_i and v_j? (ii) A network-wide ranking-based correlation that would be a quantitative measure of the extent we could use the ranking of the vertices based on a computationally-light metric as the ranking of the vertices based on a computationally-heavy metric. (iii) A linear regression-based correlation that would be a quantitative measure of the extent we could use the values of the computationally-light metric to predict the values for a computationally-heavy metric.

The Pearson's product–moment correlation coefficient is not the only correlation measure used in statistical analysis. There are at least two other well-known correlation measures such as the Spearman's Rank-based correlation measure [5] and the Kendall's concordance-based correlation measure [5] that are widely used in statistical analysis, but not that commonly used in complex network analysis. We opine that the Kendall's correlation coefficient (rather than the Pearson's correlation measure) could be more apt to do pairwise relative ordering of the vertices with respect to a computationally-heavy metric based on the values incurred for a computationally-light metric. Likewise, the Spearman's rank-based correlation coefficient could be an apt measure to decide whether a computationally-light metric could be used to rank the vertices in a graph in lieu of a computationally-heavy metric. We claim that real-world network graphs are more likely to incur different values for the correlation coefficient with respect to each of the above three correlation measures and the Pearson's correlation coefficient alone cannot be used to infer the nature of correlation between any two centrality metrics with respect to each of the three levels of correlation that are of interest in this paper. For example, for the well-known US Politics Books Network [11], we observed the following values for the Kendall's, Pearson's, and Spearman's correlation coefficients with respect to LCC′DC–BWC: 0.69, 0.78, and 0.86.

Our hypothesis in this paper is that the pairwise relative ordering-based correlation is the most strictest of the three levels of correlation and the Kendall's correlation coefficient is more likely to be the lowest of the three correlation coefficients evaluated for real-world network graphs. This is because the correlation measure is quantified as the ratio of the difference between the number of concordant pairs and the number of discordant pairs to that of the total number of pairs of vertices. A pair of vertices v_i and v_j are said to be concordant with respect to centrality metrics X and Y if $\{X(v_i) < X(v_j)$ and $Y(v_i) < Y(v_j)\}$ or $\{X(v_i) > X(v_j)$ and $Y(v_i) > Y(v_j)\}$ or $\{X(v_i) = X(v_j)$ and $Y(v_i) = Y(v_j)\}$; and discordant if $\{X(v_i) < X(v_j)$ and $Y(v_i) > Y(v_j)\}$ or $\{X(v_i) > X(v_j)$ and $Y(v_i) < Y(v_j)\}$. The Kendall's concordance-based correlation is evaluated at the level of vertex–vertex pairs, and hence, for two centrality metrics to be strongly correlated according to this measure, the number of concordant pairs of vertices should be significantly larger than the number of discordant pairs of vertices. The presence of even few discordant pairs of vertices could significantly reduce the value for the Kendall's correlation coefficient. For two different centrality metrics,

if the number of concordant pairs of vertices is significantly larger than the number of discordant pairs of vertices, the network-wide ranking of the vertices with respect to the two centrality metrics is expected to be more or less the same. Likewise, the larger the number of concordant pairs of vertices with respect to two centrality metrics X and Y, larger the chances of a dependence of the values for the centrality metric Y on the values for centrality metric X and vice-versa. Unless the centrality value of a vertex with respect to metric Y increases (or decreases) with an increase (or decrease) in the centrality value of the vertex with respect to metric X, it would be difficult to find a significant number of concordant pairs of vertices with respect to the two centrality metrics X and Y. Hence, our hypothesis in this paper is that the correlation coefficient between two centrality metrics for a real-world network graph could be bounded below by the Kendall's concordance-based correlation coefficient. In other words, if we could evaluate the Kendall's concordance-based correlation coefficient between two centrality metrics for a real-world network graph, the correlation coefficients expected between the same two centrality metrics with respect to the other two correlation measures (i.e., the Spearman's and Pearson's measures) are more likely to be at least the value obtained for the Kendall's concordance-based correlation coefficient.

We determine the correlation coefficient for DegC and LCC′DC with each of the three computationally heavy centrality metrics (EVC, BWC, and CLC) with respect to the three different measures of correlation for a total of 50 real-world networks. This generates a huge data set of correlation coefficient values (50 networks * 2 computationally-light metrics: DegC and LCC′DC * 3 computationally-heavy metrics: EVC, BWC, and CLC = 300 combinations) for each of the three correlation measures. We determine the fraction of the combinations for which each of the three correlation coefficient measures incur the lowest and largest values. We observe the Kendall's concordance-based correlation coefficient to be the lowest for 75% of the combinations, thus confirming our hypothesis.

Throughout the paper, the terms 'network' and 'graph', 'node' and 'vertex', and 'link' and 'edge' are used interchangeably; they mean the same. All the real-world network graphs and the example graphs analyzed in this paper are modeled as undirected graphs. The adjacency matrix of an undirected graph of V vertices is a $V \times V$ binary matrix, wherein there is an entry of 1 for cells (v_i, v_j) and (v_j, v_i) if and only if there is an edge between the two vertices v_i and v_j; otherwise, the entry is a 0. The rest of the paper is organized as follows: Sect. 2 reviews the five centrality metrics DegC, LCC′DC, EVC, BWC, and CLC, and illustrates their computation with an example graph. Section 3 reviews the three correlation measures (Kendall's, Spearman's, and Pearson's) and illustrates their computation for a computationally-light metric vs. a computationally-heavy metric computed for the example graph in Sect. 2. Section 4 presents the 50 real-world networks analyzed in this paper and tabulates the values for some of the fundamental metrics. We also tabulate the computation time for the five centrality metrics (on the 50 real-world networks) justifying their classification as computationally-light or computationally-heavy. Section 5 presents the results of the correlation analysis conducted on the 50 real-world networks on the basis of computationally-light vs. computationally-heavy centrality metrics with respect to the three correlation measures. Section 6 discusses related work and highlights the unique contributions of the work conducted in this paper. Section 7 concludes the paper.

2 Review of centrality metrics

Centrality metrics quantify the importance of a vertex with respect to their position in a graph. In this paper, we consider centrality metrics on the basis of whether they are computationally-light or computationally-heavy. We identify the degree centrality (DegC) [1] and the recently proposed localized clustering coefficient complement-based degree centrality (LCC′DC) [9] as the two computationally-light centrality metrics (as they could be computed quickly with time; see Sect. 4) and identify the other three well-known centrality metrics: eigenvector centrality (EVC) [2], betweenness centrality (BWC) [3], and closeness centrality (CLC) [4] as the computationally-heavy metrics. In this section, we briefly review each of these five metrics and illustrate their computation with a running example graph.

2.1 Degree centrality

The degree centrality (DegC) of a vertex is the number of neighbors incident on the vertex. Figure 1 illustrates the degree centrality of the vertices (listed above the vertices) in the example graph used in Sects. 2 and 3. A key weakness of the degree centrality metric is that the metric can take only integer values and ties among vertices (with same degree) is quite common and unavoidable in network graphs of any size (in the graph of Fig. 1, we observe five of the nine vertices to have a degree of 3). Due to this inherent weakness, we opine that degree centrality might not be an apt metric for network-wide ranking of the vertices or pairwise relative ordering of the vertices in lieu of the computationally-heavy metrics, even though DegC has been observed [5,6] to be strongly correlated with the computationally-heavy centrality metrics (EVC, BWC, and CLC) with respect to the Pearson's correlation measure for linear dependence.

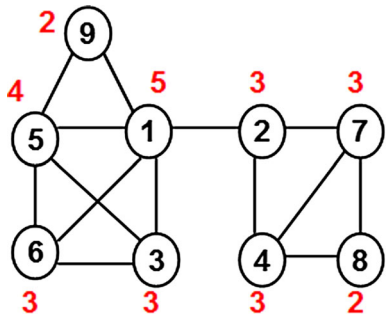


Fig. 1 Degree centrality of the vertices in an example graph

2.2 Eigenvector centrality

The eigenvector centrality (EVC) of a vertex is a measure of the degree of the vertex as well as the degree of its neighbors [2]. The EVC of the vertices is a column vector computed using the power-iteration algorithm [12]. The algorithm takes the adjacency matrix of the graph (say, $A[v_i, v_j]$ for $1 \leq v_i, v_j \leq V$, where V is the number of vertices) as input and processes it through a sequence of iterations. The EVC column vector is initialized to a unit vector (all the entries are 1s). In the first iteration, we multiply the adjacency matrix A with the EVC column vector of all 1s and divide the entries in the product vector P (also a column vector) by the normalized value of its entries. For the subsequent iterations, we set the EVC column vector to be the product vector obtained (after dividing the individual entries with the normalized value) at the end of the previous iteration. We continue the iterations by multiplying the adjacency matrix with the EVC column vector obtained at the end of the previous iteration. The algorithm stops when the entries in the EVC column vector converge (i.e., do not change further to a certain level of precision) and the vector is then called the principal eigenvector.

There is an entry for each vertex in the principal eigenvector and the values of these entries correspond to the eigenvector centrality of the vertices. The normalized value of the entries in the final product vector that is transformed to the principal eigenvector is called the principal eigenvalue (a.k.a. the spectral radius) of the adjacency matrix of the network graph. The power-iteration method is of time-complexity $\Theta(V^3)$ as we do $\Theta(V^2)$ multiplications in each iteration (to compute the product vector) and there could be at most V iterations before the entries in the product vector converge and the product vector becomes the principal eigenvector.

Figure 2 presents an example to illustrate the computation of the principal eigenvector (i.e., the EVC of the vertices) for the example graph. The example aptly illustrates the impact of the DegC and EVC of the neighbors of a vertex on the EVC of the vertex. We notice that though vertices 8 and 9 have the same degree of 2, vertex 9 has a relatively larger EVC (0.290) compared to vertex 8 (0.069): this is because, vertex 9 is attached to two neighboring vertices (vertices 1 and 5) that have a larger DegC as well as a larger EVC, whereas, vertex 8 is attached to two neighboring vertices (vertices 4 and 7) that have a relatively lower DegC and lower EVC values.

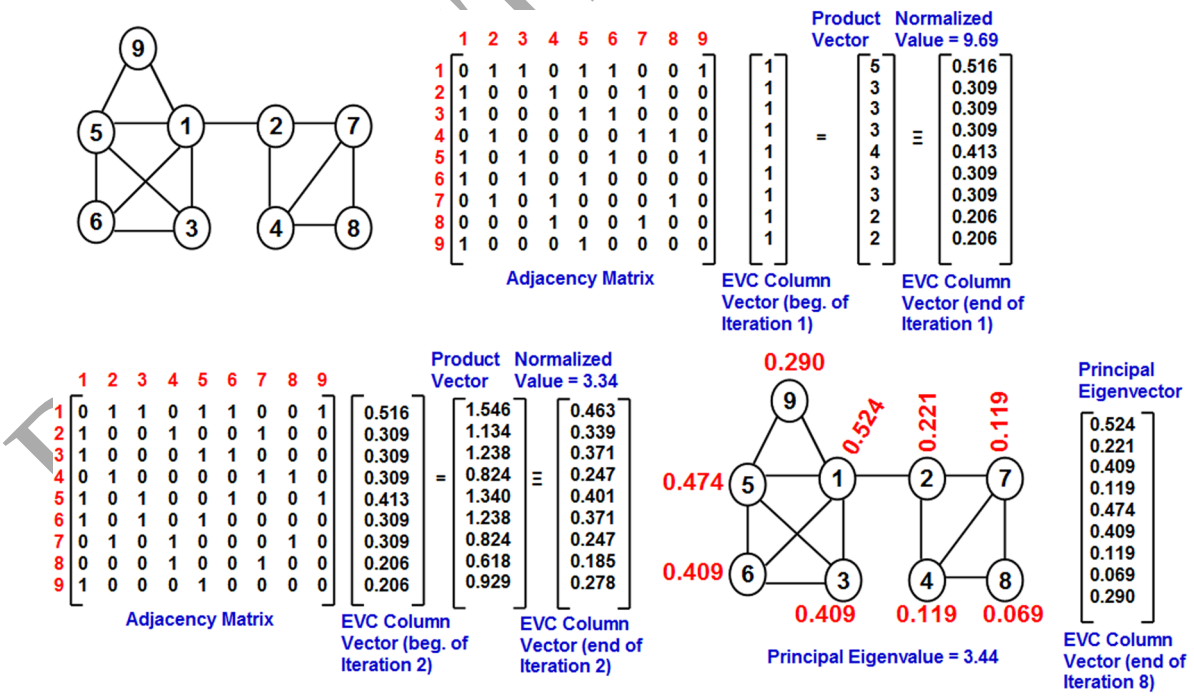


Fig. 2 Eigenvector centrality of the vertices in an example graph

2.3 Betweenness centrality

The betweenness centrality (BWC) of a vertex is a measure of the number of the shortest paths between any two vertices that go through the vertex [3]. The BWC of a vertex v_i is quantitatively computed as follows: $\mathrm{BWC}(v_i) = \sum_{\substack{v_j \neq v_i \\ v_k \neq v_i}} \frac{\#\mathrm{sp}_{v_i}(v_j, v_k)}{\#\mathrm{sp}(v_j, v_k)}$, where $\#\mathrm{sp}(v_j, v_k)$ is the total number of shortest paths between any two vertices v_j and v_k (other than v_i) and $\#\mathrm{sp}_{v_i}(v_j, v_k)$ is the number of such shortest paths between vertices v_j and v_k that go through vertex v_i.

BWC is a computationally-heavy metric and the best algorithm known so far is the classical Brandes' algorithm [13] of time-complexity $\Theta(V^2+VE)$ for undirected graphs. We now briefly describe a breadth first search (BFS)-based implementation [14] of the Brandes' algorithm. We compute a BFS tree rooted at each of the vertices in the graph; we keep track of the level number of every vertex (say, v_i in general) in each of these BFS trees. The level number of a vertex v_i in a BFS tree rooted at vertex v_j corresponds to the number of hops on the shortest path from vertex v_j to v_i. One or more vertices could exist at a particular level in the BFS trees; a vertex v_x is considered to be a predecessor for a vertex v_y in a BFS tree if there exists an edge between v_x and v_y and v_x is at a level one less than the level of v_y (i.e., v_x is relatively closer to the root of the BFS tree). The root of a BFS tree is considered to be at level 0 for the particular tree. The number of the shortest paths from the root of a BFS tree to itself is 1. The number of shortest paths for a vertex v_i from the root v_j of a BFS tree is the sum of the number of shortest paths from the root v_j to each of the predecessors of v_i in the BFS tree rooted at v_j. Using the level numbers and the set of predecessors of a vertex in a BFS tree rooted at a vertex v_j, we could calculate the number of the shortest paths from the root v_j to every other vertex in the graph. To calculate the number of the shortest paths from two vertices v_j to v_k that go through vertex v_i, we would simply take the maximum of the number of shortest paths from v_j to v_i (on the BFS tree rooted at v_j) and the number of the shortest paths from v_k to v_i (on the BFS tree rooted at v_k). We can then calculate the ratio $\mathrm{BWC}(v_i; v_j, v_k) = \frac{\#\mathrm{sp}_{v_i}(v_j, v_k)}{\#\mathrm{sp}(v_j, v_k)}$ for every vertex v_i with respect to the pair of vertices v_j and v_k ($v_j \neq v_i$ and $v_k \neq v_i$) and add these ratios to calculate the BWC of a vertex v_i. Figure 3 illustrates the computation of the BWC of the vertices in the example graph of Figs. 1 and 2. To avoid cluttering in the figure, we only show the non-zero BWC fractions of a vertex with respect to the pairs of vertices.

2.4 Closeness centrality

The closeness centrality (CLC) of a vertex [4] is a measure of the closeness of the vertex to the rest of the vertices in a graph. The CLC of a vertex is computed as the inverse of the sum of the hop counts of the shortest paths from the vertex to the rest of the vertices in the graph. To determine the CLC of a vertex, we could use the $\Theta(V + E)$–BFS algorithm to determine a shortest path tree rooted at the vertex and find the sum of the level numbers of the vertices on this shortest path tree. We want to maintain the convention that larger the centrality value for a vertex, more important is the vertex. Hence, we find the inverse of the final sum of the level numbers of the vertices on the BFS tree of a vertex and use it as the CLC of the vertex (rather than using just the sum of the level numbers as the CLC). Since we need to run the BFS algorithm once for each vertex, the overall time complexity to determine the CLC of the vertices is $\Theta(V(V + E)) = \Theta(V^2 + VE)$. Figure 4 illustrates the distance matrix (hop counts of the shortest paths between any two vertices) for the example graph of Figs. 1, 2 and 3 and also displays the CLC of the vertices. Vertex 1 is the closest vertex to the rest of the vertices (sum of the distances is 12, the minimum) and hence has the largest CLC value of $1/12 = 0.083$.

2.5 Localized clustering coefficient complement-based degree centrality

The localized clustering coefficient (LCC) of a vertex is a measure of the probability for any two neighbors of the vertex to be connected [1]. The LCC of a vertex is computed as 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]. The LCC of a vertex ranges from 0.0 to 1.0. If any two neighbors of a vertex are directly connected to each other, then the LCC of the vertex is 1.0. On the hand, if no two neighbors of a vertex have a link between them, then the LCC of the vertex is 0.0. Note that the LCC of a vertex v_i with just one neighbor is 1.0 as the neighbor is connected to itself and need not go through the vertex v_i to reach itself.

If two neighbors v_j and v_k of a vertex v_i are not directly connected to each other, then it is more likely that the two vertices would use vertex v_i for the shortest path communication. The larger the fraction of the pairs of neighbors of a vertex that are not directly connected to each other (i.e., lower the LCC of a vertex), the larger the chances for several of the neighbors of the vertex to go through the vertex for the shortest path communication. This observation leads to the proposal of a new centrality metric [9] called the local clustering coefficient complement-based degree centrality (LCC′DC). The local clustering coefficient complement (LCC′ = 1-LCC) essentially captures the probability that any two neighbors of a vertex would go through the vertex for shortest path communication. The LCC′DC of a vertex is simply the product of LCC′ and DegC, the degree centrality of the vertex.

The hypothesis behind the proposal for LCC′DC is that larger the number of neighbors for a vertex and larger the

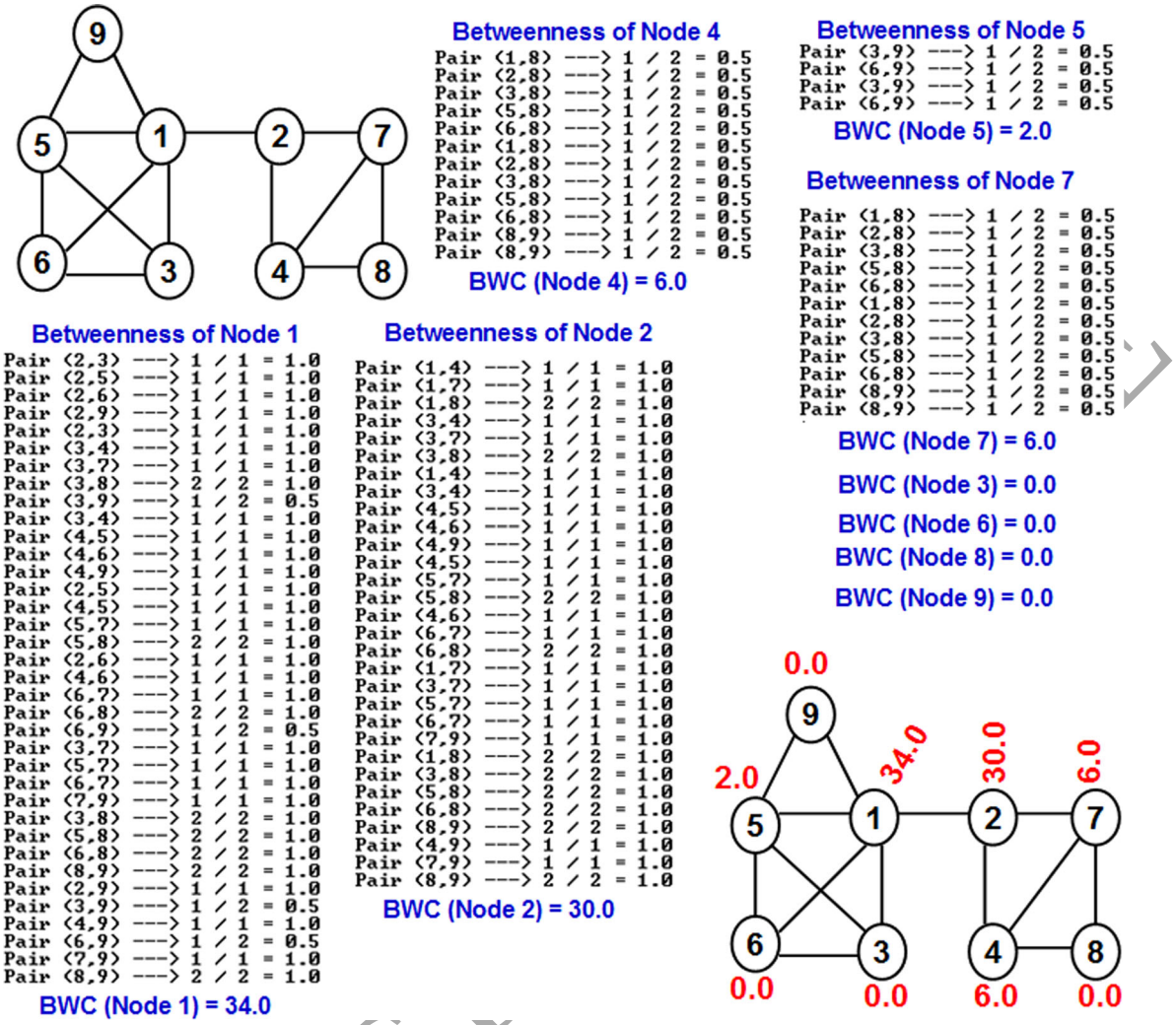


Fig. 3 Betweenness centrality of the vertices in an example graph

fraction of pairs of these neighbors going through the vertex for the shortest path communication, larger the chances of the vertex having a higher BWC. It has been observed in [9] that LCC′DC is strongly correlated to BWC for a suite of 18 real-world networks of different domains (a subset of the networks analyzed in this paper) with wide-ranging variations in degree distribution. Note that to compute the BWC of even a single vertex, we would need to determine the shortest path trees rooted at every vertex. On the other hand, LCC′DC is a computationally-light metric that could be computed simply on the basis of the two-hop neighborhood of a vertex. Figure 5 illustrates the computation of the LCC′DC of the vertices in the example graph of Figs. 1, 2, 3 and 4. A comparison of the BWC and the LCC′DC values incurred for the example graph in Figs. 3 and 5 indicates a strong correlation between the two metrics. Vertices 1 and 2 are the top two vertices to have the largest BWC values of 34 and 30, respectively; vertices 1 and 2 are also the top two vertices to have the largest LCC′DC values of 3.0 and 2.0, respectively. Likewise, the BWC of vertices 3, 6, 8, and 9 are 0.0 each and the LCC′DC values of these vertices are also 0.0 each.

3 Levels of correlation and the correlation coefficient measures

We identify three different levels of correlation that could be explored between a computationally-light centrality metric and computationally-heavy centrality metric. For discussion purposes, let X be a computationally-light centrality metric and Y be a computationally-heavy centrality metric. The three levels of correlation that are of interest in this paper are as follows:

(i) *Pairwise relative ordering of the vertices*: For any two vertices v_i and v_j, we are interested to quantify how well we can use the relative ordering of the two vertices

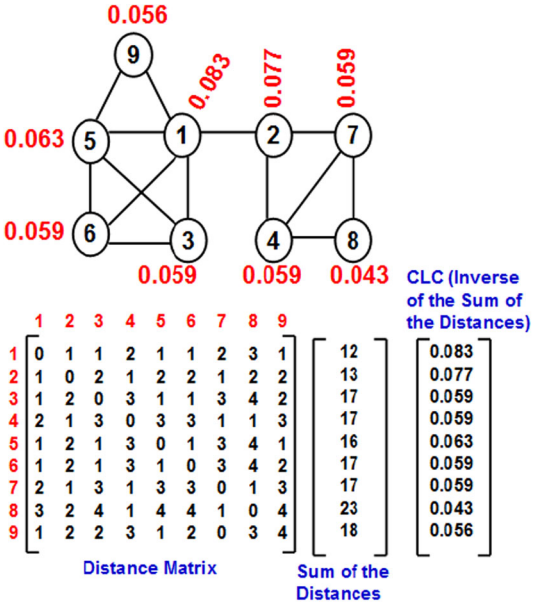


Fig. 4 Closeness centrality of the vertices in an example graph

with respect to the computationally-light metric X (i.e., whether $X(v_i) < X(v_j)$ or $X(v_i) > X(v_j)$ or $X(v_i) = X(v_j)$) as the relative ordering of the same two vertices with respect to the computationally-heavy metric Y.

(ii) *Network-wide ranking of the vertices*: We are interested to quantify how well we can use the network-wide ranking of the vertices with respect to a computationally-light metric Xas the network-wide ranking of the vertices with respect to a computationally-heavy metric Y.

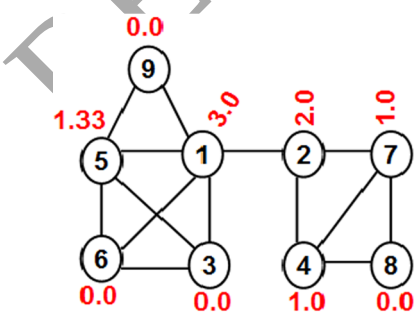
(iii) *Predicting the actual centrality values*: We are interested to quantify how well we can predict the actual centrality values for the vertices with respect to a computationally-heavy metric Y based on the actual centrality values for the vertices with respect to a computationally-light metric X.

The Pearson's product–moment-based correlation measure (r) has been the commonly used measure in the literature (e.g., [5,6]) to assess the correlation between centrality metrics for complex networks. However, the Pearson's correlation measure can accurately capture only one of the three levels of correlation (i.e., predicting the actual centrality values) and not the other two levels of correlation. The Kendall's concordance-based correlation measure (τ) and the Spearman's rank-based correlation measure (ρ) are the correlation measures that can effectively capture the pairwise relative ordering of the vertices and the network-wide ranking of the vertices, respectively. This is quite evident from their formulation itself (as will be seen in this section). The values for all the three correlation coefficient measures range from -1 to 1; the closer is the value to 1 or -1, the more stronger (positive or negative) the correlation between the two centrality metrics in consideration. If the value for the correlation coefficient is closer to 0, the two centrality metrics are considered to be independent of each other.

Our hypothesis in this paper is that the pairwise relative ordering of the vertices is the most restrictive level of correlation one could impose to assess the correlation among vertices with respect to any node-level metric (like centrality metric), and hence, the Kendall's concordance-based correlation coefficient has a higher chance of being the lowest of the correlation coefficient values (compared to Pearson's r and Spearman's ρ) for the three levels of correlation. On the other hand, we conjecture that the Spearman's rank-based correlation is the least restrictive of the three correlation measures as it does not require the two centrality metrics to have a linear dependence (as is required for the Pearson's correlation measure) and minor differences in the rank of a vertex with respect to the two centrality metrics in consideration does not significantly affect the value for the correlation coefficient (see formulation 4 in Sect. 3.2 and the accompanying discussion).

3.1 Kendall's concordance-based correlation

A pair of vertices v_i and v_j are said to be concordant with respect to centrality metrics X and Y if $\{X(v_i) < X(v_j)$ and $Y(v_i) < Y(v_j)\}$ or $\{X(v_i) > X(v_j)$ and $Y(v_i) > Y(v_j)\}$



Vertex	Degree DegC	Max. Edges among neighbs	Actual Edges among neighbs	LCC	LCC' 1-LCC	LCC'DC
1	5	5(5-1)/2 = 10	4	0.4	0.6	3.0
2	3	3(3-1)/2 = 3	1	0.33	0.67	2.0
3	3	3(3-1)/2 = 3	3	1.0	0.0	0.0
4	3	3(3-1)/2 = 3	2	0.67	0.33	1.0
5	4	4(4-1)/2 = 6	4	0.67	0.33	1.33
6	3	3(3-1)/2 = 3	3	1.0	0.0	0.0
7	3	3(3-1)/2 = 3	2	0.67	0.33	1.0
8	2	2(3-1)/2 = 1	1	1.0	0.0	0.0
9	2	2(2-1)/2 = 1	1	1.0	0.0	0.0

Fig. 5 LCC′DC of the vertices in an example graph

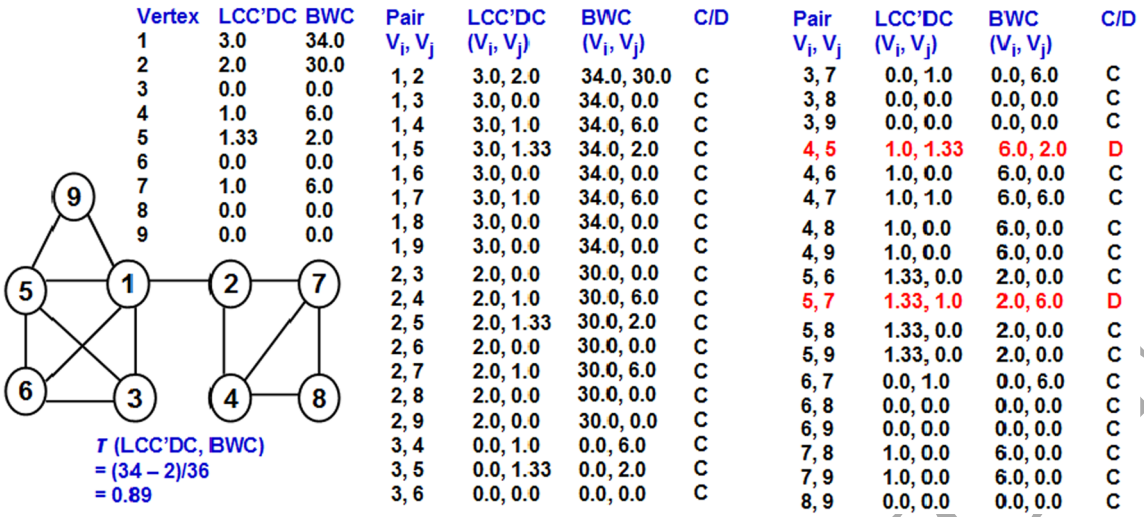


Fig. 6 Example to compute the Kendall's concordance-based correlation coefficient

or $\{X(v_i) = X(v_j)$ and $Y(v_i) = Y(v_j)\}$. A pair of vertices v_i and v_j are said to be discordant with respect to centrality metrics X and Y if $\{X(v_i) < X(v_j)$ and $Y(v_i) > Y(v_j)\}$ or $\{X(v_i) > X(v_j)$ and $Y(v_i) < Y(v_j)\}$. The Kendall's concordance [5]-based correlation coefficient (τ) is computed (see formulation 1) as the ratio of the difference between the number of concordant pairs (#conc.pairs) and the number of discordant pairs (#disc.pairs) to that of the total number of pairs of vertices (which is also the sum of the number of concordant pairs and discordant pairs):

$$\tau(X, Y) = \frac{\#\text{conc.pairs}(X, Y) - \#\text{disc.pairs}(X, Y)}{\#\text{conc.pairs}(X, Y) + \#\text{disc.pairs}(X, Y)}. \quad (1)$$

Figure 6 illustrates an example to calculate the Kendall's concordance-based correlation between the LCC′DC and BWC metrics. We use the alphabets 'C' (for concordance) and 'D' (for discordance) to indicate whether a pair of vertices is concordant or discordant. In this example graph, there are a total of $9 \times (9-1)/2 = 36$ pairs of vertices that could be tested for concordance. Except the two pairs of vertices: pairs 4–5 and 5–7, all the other 34 pairs of vertices are observed to be concordant with respect to LCC′DC and BWC. Hence, the Kendall's concordance-based correlation coefficient τ (LCC′DC and BWC) $= (34-2)/36 = 0.89$.

Note that due to the nature of the formulation (# concordant pairs − # discordant pairs) in the numerator, Kendall's concordance-based correlation coefficient has the tendency to reduce appreciably even in the presence of few discordant pairs. We more formally analyze the relationship between the number of concordant pairs and the number of discordant pairs on the Kendall's correlation coefficient as follows. Let $f_c(X, Y)$ be the fraction of the concordant pairs of vertices with respect to any two metrics X and Y; the formulation for Kendall's concordance-based correlation coefficient could be written as follows.

Fraction of concordant pairs of vertices, $f_c(X, Y) = \frac{\#\text{conc.pairs}(X,Y)}{\#\text{conc.pairs}(X,Y)+\#\text{disc.pairs}(X,Y)}$:

$$1 - f_c(X, Y) = \frac{\#\text{disc.pairs}(X, Y)}{\#\text{conc.pairs}(X, Y) + \#\text{disc.pairs}(X, Y)}$$

$$\tau(X, Y) = \frac{\#\text{conc.pairs}(X, Y)}{\#\text{conc.pairs}(X, Y) + \#\text{disc.pairs}(X, Y)} - \frac{\#\text{disc.pairs}(X, Y)}{\#\text{conc.pairs}(X, Y) + \#\text{disc.pairs}(X, Y)}$$

$$\tau(X, Y) = f_c(X, Y) - (1 - f_c(X, Y)))$$

$$\tau(X, Y) = 2 f_c(X, Y) - 1. \quad (2)$$

Figure 7 illustrates how $\tau(X, Y)$ decreases with decrease in $f_c(X, Y)$. We can notice that for a 0.01 decrease in $f_c(X, Y)$, $\tau(X, Y)$ decreases by 0.02.

3.2 Spearman's rank-based correlation

The rank of a vertex with respect to a centrality metric is a measure of where the vertex stands if the vertices in the network are to be ordered in the decreasing order of the values for the centrality metric (we assume decreasing order for all the centrality metrics). The earlier a vertex appears in the listing with respect to a particular centrality metric, the higher the rank for the vertex with respect to the metric. We use the Spearman's rank [5]-based correlation measure (ρ) to quantify the extent of similarity in the ranking of the vertices with respect to two centrality metrics. We calculate this correlation coefficient measure as follows with respect to any two centrality metrics (say X and Y). For each centrality metric,

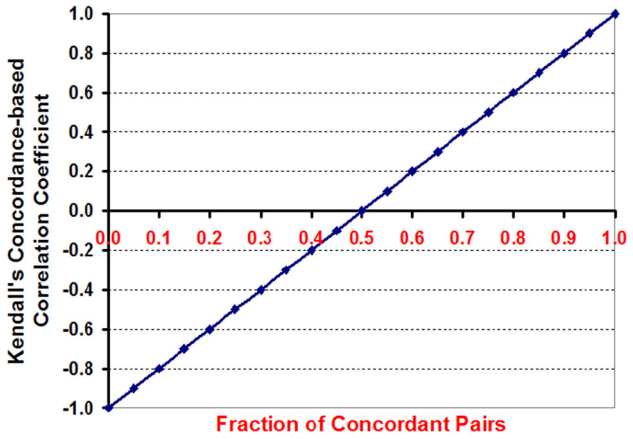


Fig. 7 Relationship between Kendall's concordance-based correlation coefficient and the fraction of concordant pairs

we first obtain a listing of the vertices in the decreasing order of the centrality values. If two or more vertices have the same centrality value, we break the tie in favor of the vertex with the smaller ID. The index at which a vertex appears in this list is the tentative ranking for the vertex. The final ranking for a vertex with respect to a centrality metric is the same as the tentative ranking for the vertex if it has no tie with any other vertex for the centrality metric. If two or more vertices have a tie with respect to a centrality metric, their final ranking with respect to the centrality metric is the average of the tentative rankings for the vertices with respect to the metric. Let d_i be the difference in the final ranking for the vertices with respect to the two centrality metrics X and Y, where $1 \leq i \leq n$ and n is the number of vertices in the graph. The Spearman's rank-based correlation coefficient $\rho(X, Y)$ is computed using formula (3):

$$\rho(X, Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}. \tag{3}$$

Figure 8 illustrates the computation of the Spearman's rank-based correlation coefficient on the example graph of Figs. 1, 2, 3, 4 and 5. With respect to BWC, we observe vertices 4 and 7 to have tie (BWC = 6 for both) and we break the tie on the basis of the vertex ID: vertex 4 with a lower ID gets a tentative rank of 3 and vertex 7 gets a tentative rank of 4; the final ranking for the two vertices is the average of their tentative rankings ($(3 + 4)/2 = 3.5$). A similar tie between the two vertices exists with respect to LCC′DC. We also observe tie between vertices 3, 6, 8, and 9 with respect to both BWC and LCC′DC. We observe a non-zero difference in the ranking of the vertices for only three of the nine vertices and the magnitudes of these differences are not that high to significantly reduce the correlation coefficient value (0.95).

With respect to formulation (3), for larger values of n, the term in the denominator $n(n^2 - 1)$ dominates the summation term $\sum_{i=1}^{n} d_i^2$ in the numerator. Hence, even if the differences in the ranking of the vertices are larger, the Spearman's rank-based correlation coefficient is more likely to relatively stay high (compared to the Kendall's measure) for graphs with larger number of vertices. Using formulation (3), we could also extract an upper bound for the average of the absolute difference (per vertex) in the final ranking for the vertices with respect to two data sets X and Y (in this case, the centrality metrics), so that the Spearman's rank-based correlation coefficient is positive (i.e., >0) and is at least a targeted value (indicated using $\rho_{\text{tgt}}^{(X,Y)}$, such that $0 < \rho_{\text{tgt}}^{(X,Y)} \leq 1$). The upper bound is shown in formulation (4), with $n^2 - 1 \sim n^2$. For a given $\rho_{\text{tgt}}^{(X,Y)}$ value, we observe the upper bound to linearly increase proportional with increase in the number of vertices in the graph (thus confirming our earlier assertion):

$$\text{For } \rho(X, Y) \geq \rho_{\text{tgt}}^{(X,Y)}, 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \geq \rho_{\text{tgt}}^{(X,Y)}.$$

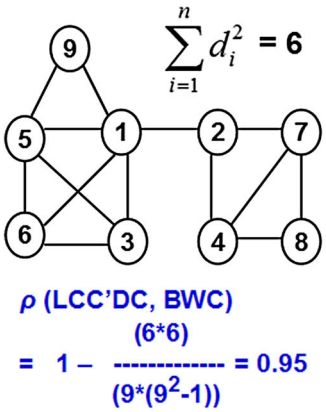
$$\text{That is, } \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n^2}} \leq \sqrt{\frac{\left(1 - \rho_{\text{tgt}}^{(X,Y)}\right) \times n}{6}}. \tag{4}$$

3.3 Pearson's product–moment correlation

The Pearson's product–moment correlation (r) when applied for centrality metrics is a measure of the linear dependence between any two metrics in consideration [5]. It is referred to as the product–moment-based correlation as we calculate the deviation of the data points from their mean value ('mean' is also referred to as 'first moment' in statistics) and use them in the formulation to calculate the correlation coefficient [see formulation (5)]. If X and Y are the data sets for two centrality metrics, let X_i and Y_i indicate the centrality values for the individual vertices v_i ($1 \leq i \leq n$, where n is the number of vertices) and $\overline{X}$ and $\overline{Y}$ are the average of the centrality values; $r(X, Y)$ is calculated as follows:

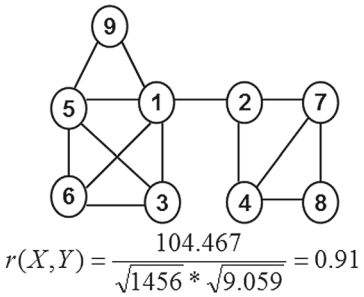
$$r(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}. \tag{5}$$

Figure 9 illustrates the computation of Pearson's product–moment correlation coefficient on the example graph of Figs. 1, 2, 3, 4 and 5. We observe the Pearson's correlation coefficient to be 0.91 and is in between the values of 0.89 and 0.95 observed, respectively, for the Kendall's and Spearman's correlation coefficients. As seen for several real-world networks analyzed in this paper, the Kendall's correlation coefficient measure is the lowest of the three correlation coefficient values.



Vertex, i	BWC	Tentative Rank: BWC	Final Rank: BWC_i	LCC'DC	Tentative Rank: LCC'DC	Final Rank: $LCC'DC_i$	Rank Difference d_i = BWC_i - $LCC'DC_i$	d_i^2
1	34	1	1	3	1	1	0	0
2	30	2	2	2	2	2	0	0
3	0	6	7.5	0	6	7.5	0	0
4	6	3	3.5	1	4	4.5	-1	1
5	2	5	5	1.33	3	3	2	4
6	0	7	7.5	0	7	7.5	0	0
7	6	4	3.5	1	5	4.5	-1	1
8	0	8	7.5	0	8	7.5	0	0
9	0	9	7.5	0	9	7.5	0	0

Fig. 8 Example to compute the Spearman's rank-based correlation coefficient



Vertex, i	BWC	BWC_i - Avg(BWC)	(BWC_i - Avg(BWC))²	LCC'DC	$LCC'DC_i$ - Avg(LCC'DC)	($LCC'DC_i$ - Avg(LCC'DC))²	(BWC_i- Avg(BWC)* ($LCC'DC_i$ - Avg(LCC'DC)
1	34	25.33	641.61	3	2.07	4.285	52.433
2	30	21.33	454.97	2	1.07	1.145	22.823
3	0	-8.67	75.17	0	-0.93	0.865	8.063
4	6	-2.67	7.13	1	0.07	0.005	-0.187
5	2	-6.67	44.49	1.33	0.4	0.160	-2.668
6	0	-8.67	75.17	0	-0.93	0.865	8.063
7	6	-2.67	7.13	1	0.07	0.005	-0.187
8	0	-8.67	75.17	0	-0.93	0.865	8.063
9	0	-8.67	75.17	0	-0.93	0.865	8.063
	Avg = 8.67		Sum = 1456	Avg = 0.93		Sum = 9.059	Sum = 104.467

Fig. 9 Example to compute the Pearson's product–moment correlation coefficient

4 Real-world networks

In this section, we provide a brief description of the 50 real-world networks analyzed in this paper and tabulate the values for some of the fundamental metrics for complex network analysis observed for these networks as well as tabulate the computation time per node of the five centrality metrics (discussed in Sect. 2) for these networks. All the real-world networks are modeled as undirected graphs. Table 1 lists the number of nodes and edges in these graphs as well as the values for fundamental metrics like average node degree (k_{avg}), spectral radius ratio for node degree (λ_{sp}) [15], graph density (G_d), and the number of components (#comps). The spectral radius ratio for node degree [15] is a measure of the variation in node degree and is calculated as the ratio of the principal eigenvalue [2] of the adjacency matrix of the graph to that of the average node degree. The spectral radius ratio for node degree is independent of the number of vertices and the actual degree values for the vertices in the graph. The spectral radius ratio for node degree is always greater than or equal to 1; the farther is the ratio from the value of 1, the larger the variation in node degree. The spectral radius ratio for node degree for the real-world network graphs analyzed in this paper ranges from 1.01 to 5.34 (indicating that the real-world network graphs analyzed range from random networks with smaller variation in node degree to scale-free networks of larger variation in node degree). The individual references to the real-world networks are also listed in Table 1.

The networks considered cover a broad range of categories (as listed below along with the number of networks in each category): (I) acquaintance network (12), (II) friendship network (9), (III) co-appearance network (6), (IV) employment network (4), (V) citation network (3), (VI) collaboration network (3), (VII) literature network (3), (VIII) biological network (3), (IX) political network (2), (X) game network (2), (XI) transportation network (1), (XII) geographical network (1), and (XIII) trade network (1). A brief description about each category of networks is as follows: an *acquaintance network* is a kind of social network in which the participant nodes slightly (not closely) know each other, as observed typically during an observation period. A *friendship network* is a kind of social network in which the participant nodes closely know each other and the relationship is not captured over an observation period. A *co-appearance network* is a network typically extracted from novels/books in such a way that two characters or words (modeled as nodes) are con-

Table 1 Fundamental metric values of the real-world network graphs used for correlation analysis

#	Net.	Network name	Refs.	λ_{sp}	#Nodes	#Edges	k_{avg}	G_d	#Comps
1	ADJ	Word Adjacency Net^{III}	[17]	1.73	112	425	7.589	0.068	1
2	AKN	Anna Karnenina Net^{III}	[18]	2.48	140	494	7.057	0.051	2
3	JBN	Jazz Band Net^{IV}	[19]	1.45	198	2742	27.697	0.141	1
4	CEN	C. Elegans Neural Net^{VIII}	[20]	1.68	297	2148	14.465	0.049	1
5	CLN	Centrality Literature Net^{V}	[21]	2.03	118	613	10.39	0.089	1
6	CGD	Citation Graph Draw. Net^{V}	[22]	2.24	259	640	4.942	0.019	6
7	CFN	Copperfield Net^{III}	[18]	1.83	89	407	9.146	0.104	2
8	DON	Dolphin Net^{I}	[23]	1.40	62	159	5.129	0.084	1
9	DRN	Drug Net^{I}	[24]	2.76	212	284	2.679	0.013	9
10	DLN	Dutch Lit. 1976 Net^{VII}	[25]	1.49	37	81	4.378	0.122	2
11	ERD	Erdos Collaboration Net^{VI}	[26]	3.00	433	1314	6.069	0.014	3
12	FMH	Faux Mesa High Sch. Net^{II}	[27]	2.81	147	202	2.748	0.019	11
13	FHT	Friendship Hi-Tech Firm^{II}	[28]	1.57	33	91	5.515	0.172	1
14	FTC	Flying Teams Cade Net^{IV}	[29]	1.21	48	170	7.083	0.151	1
15	FON	US Football Net^{X}	[30]	1.01	115	613	10.661	0.094	1
16	CDF	College Dorm Fraternity^{I}	[31]	1.11	58	967	33.345	0.585	1
17	GD96	Graph Drawing 1996 Net ^{V}	[26]	2.38	180	228	2.533	0.014	1
18	MUN	Marvel Universe Net^{III}	[32]	2.54	167	301	3.605	0.022	20
19	GLN	Graph Glossary Net^{VII}	[26]	2.01	67	118	3.522	0.053	4
20	HTN	Hypertext 2009 Net^{I}	[33]	1.21	115	2164	37.635	0.33	2
21	HCN	Huckleberry Co. Net^{III}	[18]	1.66	76	302	7.947	0.106	4
22	ISP	Infec. Socio-Patt. Net^{I}	[33]	1.69	309	1924	12.453	0.04	1
23	KCN	Karate Club Net^{I}	[34]	1.47	34	78	4.588	0.139	1
24	KFP	Korea Family Plan. Net^{I}	[35]	1.70	37	85	4.595	0.128	2
25	LMN	Les Miserables Net^{III}	[18]	1.82	77	254	6.597	0.087	1
26	MDN	Macaque Dom. Net^{VIII}	[36]	1.04	62	1167	37.645	0.617	1
27	MTB	Madrid Train Bomb. Net^{I}	[37]	1.95	64	295	9.219	0.146	1
28	MCE	Manufact. Comp. Empl^{IV}	[38]	1.12	77	1549	40.23	0.529	1
29	MSJ	Social Net. Journal Net^{VI}	[39]	3.48	475	625	2.632	0.006	104
30	AFB	Author Facebook Net^{II}	–	2.29	171	940	10.994	0.065	4
31	MPN	Mexican Pol. Elite Net^{IX}	[40]	1.23	35	117	6.686	0.197	1
32	MMN	ModMath Net^{II}	[26]	1.59	30	61	4.067	0.14	1
33	NSC	Net. Science Co-author^{VI}	[17]	5.51	1589	2743	3.45	0.002	269
34	PBN	US Politics Books Net^{VII}	[11]	1.42	105	441	8.4	0.081	1
35	PSN	Primary Sch. Contact Net^{I}	[41]	1.22	238	5539	46.546	0.196	1
36	PFN	Prison Friendship Net^{II}	[42]	1.32	67	142	4.239	0.064	1
37	SJN	San Juan Sur Family Net^{I}	[43]	1.29	75	155	4.133	0.056	1
38	SDI	Scotland Corp. Interlock^{IV}	[44]	1.94	230	359	3.122	0.014	5
39	SPR	Senator Press Release Net^{IX}	[45]	1.57	92	477	10.37	0.114	1
40	SWC	Soccer World Cup Net^{X}	[26]	1.45	35	118	6.743	0.198	1
41	SSM	Sawmill Strike Comm. Net^{I}	[46]	1.22	24	38	3.167	0.138	1
42	TEN	Taro Exchange Net^{I}	[47]	1.06	22	39	3.545	0.169	1

Table 1 continued

#	Net.	Network name	Refs.	λ_{sp}	#Nodes	#Edges	k_{avg}	G_d	#Comps
43	TWF	Teenage Fem. Friend Net[II]	[48]	1.49	47	77	3.277	0.071	4
44	UKF	UK Faculty Friend Net[II]	[49]	1.35	83	578	13.928	0.17	2
45	APN	US Airports 1997 Net[XI]	[26]	3.22	332	2126	12.807	0.039	1
46	USS	US States Net[XII]	[50]	1.25	49	107	4.367	0.091	1
47	RHF	Residence Hall Friend Net[II]	[51]	1.27	217	1839	16.949	0.078	1
48	WSB	Windsurfers Beach Net[II]	[52]	1.22	43	336	15.628	0.372	1
49	WTN	World Trade Metal Net[XIII]	[53]	1.38	80	875	21.875	0.277	1
50	YPI	Yeast PPI Net[VIII]	[54]	3.20	1870	2203	2.387	0.001	149

nected if they appear alongside each other. An *employment network* is a network in which the interaction/relationship between people is primarily due to their employment requirements and not due to any personal liking. A *citation network* is a network in which two papers (nodes) are connected if one paper cites the other paper as reference. A *collaboration network* is a network of researchers/authors who are listed as co-authors in at least one publication. A *literature network* is a network of papers/terminologies/authors (other than collaboration, citation or co-authorship) involved in a particular area of literature. A *biological network* is a network that models the interactions between genes, proteins, animals of a particular species, etc. A *political network* is a network of entities (typically politicians) involved in politics. 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. A *geographical network* is a network of states and their shared borders in a country. A *trade network* is a network of countries/people involved in certain trade. More information about the individual real-world networks can be found in [16].

We measure the computation time per node (total computation time divided by the number of nodes) incurred for each of the five centrality metrics for the 50 real-world network graphs. 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 the five centrality metrics on each of the 50 real-world networks for 25 iterations and averaged the results. Table 2 lists the average computation time per node for the centrality metrics. Though there is no prescribed threshold in the literature, we propose that a centrality metric could be referred to as computationally-heavy if its average computation time per node is 0.01 millisecond or above (highlighted in Table 2) for at least 50% of the real-world networks analyzed, provided the suite of real-world networks analyzed is as diverse as it is in this paper (with respect to the number of nodes and edges and the fundamental metrics listed in Table 1). We observe the degree centrality metric to be computationally-light for all the real-world networks and the LCC′DC metric to be computationally-heavy for only 6% of the real-world networks. Hence, we refer to the DegC and LCC′DC metrics as computationally-light centrality metrics. On the other hand, we observe the CLC, EVC, and BWC metrics to be computationally-heavy for 52, 84, and 100% of the 50 real-world networks studied. Hence, we consider these three centrality metrics as computationally-heavy.

5 Correlation analysis

In this section, we present in detail the results of the correlation analysis conducted for the computationally-light (DegC and LCC′DC) vs. computationally-heavy (CLC, EVC, and BWC) centrality metrics (six combinations of metrics) for the 50 real-world network graphs listed in Sect. 4. To validate our hypothesis, we measure the following: (i) the difference in the correlation coefficient values between any two correlation measures; (ii) the fraction (a total of $50 \times 6 = 300$ combinations) of the 50 real-world networks and the six combinations of computationally-light vs. computationally-heavy centrality metrics for which each of the three correlation measures incur the lowest correlation coefficient values; and (iii) the median of the correlation coefficient values observed for each correlation level between a computationally-light metric and a computationally-heavy metric.

Table 3 lists the values for the correlation coefficient incurred with the three correlation measures (Kendall—τ; Spearman—ρ; Pearson—r) for DegC vs. {CLC, EVC, BWC} and Table 4 lists the values for LCC′DC vs. {CLC, EVC, BWC}. The correlation coefficient values reported in Tables 3 and 4 are the average of 25 runs for each of the 300 combinations of computationally-light vs. computationally-heavy centrality metrics and the real-world networks. In both these tables, we highlight the cells (combinations) for which a correlation measure incurs the lowest value (in bold font) and

Table 2 Computation time per node of the centrality metrics for the real-world network graphs

#	Net.	#Nodes	#Edges	Computation time per node (ms)				
				Computationally-light		Computationally-heavy		
				DegC	LCC′DC	CLC	EVC	BWC
1	ADJ	112	425	0.00043	0.00723	**0.09777**	**0.30250**	**2.40223**
2	AKN	140	494	0.00068	0.00965	**0.05050**	**0.44021**	**3.94329**
3	JBN	198	2742	0.00017	**0.04402**	**0.12066**	**0.22212**	**8.98010**
4	CEN	297	2148	0.00018	**0.01157**	**0.07825**	**0.47899**	**19.16182**
5	CLN	118	613	0.00023	0.00404	**0.05186**	**0.14542**	**1.45644**
6	CGD	259	640	0.00022	0.00083	**0.11286**	**0.48031**	**19.13170**
7	CFN	89	407	0.00017	0.00137	0.00674	**0.02854**	**0.46247**
8	DON	62	159	0.00018	0.00071	0.00419	**0.02097**	**0.31935**
9	DRN	212	284	0.00025	0.00058	**0.10759**	**0.27104**	**17.85425**
10	DLN	37	81	0.00027	0.00089	0.00216	**0.01919**	**0.12676**
11	ERD	433	1314	0.00019	0.00110	**0.20591**	**1.16956**	**48.16531**
12	FMH	147	202	0.00024	0.00050	**0.04871**	**0.12871**	**5.54497**
13	FHT	33	91	0.00024	0.00103	0.00182	**0.01485**	**0.13364**
14	FTC	48	170	0.00017	0.00079	0.00292	**0.01646**	**0.18542**
15	FON	115	613	0.00019	0.00121	**0.01330**	**0.08209**	**1.36739**
16	CDF	58	967	0.00028	0.00997	**0.01810**	**0.03414**	**0.67879**
17	GD96	180	228	0.00017	0.00052	**0.02817**	**0.09189**	**4.26378**
18	MUN	167	301	0.00018	0.00054	**0.02305**	**0.06587**	**1.50102**
19	GLN	67	118	0.00030	0.00046	0.00910	**0.03149**	**0.32149**
20	HTN	115	2164	0.00018	0.00724	**0.01165**	**0.05365**	**1.79522**
21	HCN	76	302	0.00026	0.00074	0.00855	**0.02579**	**0.32276**
22	ISP	309	1924	0.00017	0.00130	**0.10476**	**0.55414**	**21.06320**
23	KCN	34	78	0.00018	0.00047	0.00147	0.00529	0.06882
24	KFP	37	85	0.00030	0.00097	0.00324	**0.01216**	**0.16216**
25	LMN	77	254	0.00016	0.00083	0.00545	**0.01792**	**0.37195**
26	MDN	62	1167	0.00026	0.00560	0.00694	**0.03210**	**0.67774**
27	MTB	64	295	0.00017	0.00063	0.00500	**0.01609**	**0.32844**
28	MCE	77	1549	0.00017	0.00516	0.00558	**0.02377**	**0.74909**
29	MSJ	475	625	0.00020	0.00038	**0.18269**	**0.63120**	**28.86568**
30	AFB	171	940	0.00019	0.00137	**0.03135**	**0.38982**	**3.36468**
31	MPN	35	117	0.00017	0.00086	0.00171	0.00743	**0.10314**
32	MMN	30	61	0.00027	0.00047	0.00233	0.00700	**0.08767**
33	NSC	1589	2743	0.00016	0.00072	**2.52165**	**31.21962**	**457.18801**
34	PBN	105	441	0.00020	0.00092	**0.01848**	**0.07352**	**1.05924**
35	PSN	238	5539	0.00016	**0.01128**	**0.04836**	**0.31601**	**13.87235**
36	PFN	67	142	0.00016	0.00048	0.00567	**0.01925**	**0.33701**
37	SJN	75	155	0.00017	0.00055	0.00573	**0.02573**	**0.42813**
38	SDI	230	359	0.00018	0.00049	**0.05117**	**0.22422**	**10.97583**
39	SPR	92	477	0.00058	0.00133	**0.03793**	**0.14533**	**0.79196**
40	SWC	35	118	0.00017	0.00054	0.00143	0.00743	**0.08714**
41	SSM	24	38	0.00021	0.00033	0.00125	0.00417	**0.03292**
42	TEN	22	39	0.00018	0.00032	0.00091	0.00364	**0.03045**
43	TWF	47	77	0.00017	0.00032	0.00255	0.00979	**0.07106**

Table 2 continued

#	Net.	#Nodes	#Edges	Computation time per node (ms)				
				Computationally-light		Computationally-heavy		
				DegC	LCC′DC	CLC	EVC	BWC
44	UKF	83	578	0.00016	0.00149	0.00675	**0.02675**	**0.62578**
45	APN	332	2126	0.00016	0.00323	**0.09518**	**0.49545**	**18.50593**
46	USS	49	107	0.00041	0.00045	0.00265	**0.01224**	**0.17469**
47	RHF	217	1839	0.00016	0.00212	**0.04083**	**0.24429**	**9.31433**
48	WSB	43	336	0.00019	0.00128	0.00209	0.00977	**0.17558**
49	WTN	80	875	0.00058	0.00283	**0.02513**	**0.10938**	**0.67938**
50	YPI	1870	2203	0.00018	0.00086	**3.45965**	**77.53588**	**834.37062**
Fraction of networks for which average computation time per node ≥ 0.01 ms				0/50 = 0.0	3/50 = 0.06	26/50 = 0.52	42/50 = 0.84	50/50 = 1.00

Table 3 Degree centrality vs. computationally-heavy metrics: results of correlation analysis

#	Net.	Degree centrality (DegC) vs. closeness centrality (CLC)			Degree centrality (DegC) vs. eigenvector centrality (EVC)			Degree centrality (DegC) vs. betweenness centrality (BWC)		
		Kendall	Spearman	Pearson	Kendall	Spearman	Pearson	Kendall	Spearman	Pearson
1	ADJ	**0.764**	*0.901*	0.841	**0.801**	0.929	*0.957*	**0.773**	0.901	*0.915*
2	AKN	**0.626**	0.767	*0.846*	**0.763**	0.897	*0.936*	**0.657**	0.759	*0.892*
3	JBN	**0.736**	*0.890*	0.859	**0.750**	0.890	*0.901*	**0.579**	*0.744*	0.610
4	CEN	**0.553**	*0.738*	0.700	**0.629**	0.811	*0.871*	**0.736**	*0.889*	0.780
5	CLN	0.847	*0.956*	**0.282**	**0.892**	*0.976*	0.961	**0.750**	*0.903*	0.825
6	CGD	0.754	*0.893*	**0.497**	**0.722**	*0.876*	0.810	**0.745**	*0.890*	0.797
7	CFN	**0.882**	*0.945*	0.908	**0.870**	*0.965*	0.935	**0.697**	*0.818*	0.808
8	DON	**0.548**	*0.718*	0.713	**0.512**	0.627	*0.720*	0.667	*0.814*	**0.598**
9	DRN	0.718	*0.856*	**0.608**	**0.603**	*0.758*	0.650	0.758	*0.875*	**0.649**
10	DLN	**0.856**	*0.953*	0.908	**0.768**	0.904	*0.947*	**0.672**	*0.804*	0.791
11	ERD	0.709	*0.858*	**0.261**	**0.675**	0.827	*0.916*	**0.708**	*0.860*	0.782
12	FMH	0.739	*0.871*	**0.624**	**0.541**	*0.704*	0.558	0.711	*0.832*	**0.630**
13	FHT	0.866	*0.956*	**0.409**	**0.812**	0.920	*0.937*	**0.755**	*0.902*	0.816
14	FTC	**0.650**	0.802	*0.837*	**0.596**	0.730	*0.822*	**0.582**	0.723	*0.783*
15	FON	**0.272**	*0.344*	0.291	**0.606**	0.722	*0.750*	**0.260**	*0.336*	0.282
16	CDF	0.998	*1.000*	**0.990**	**0.972**	0.991	*0.997*	**0.809**	*0.940*	0.857
17	GD96	0.552	*0.659*	**0.513**	**0.568**	0.684	*0.844*	**0.759**	0.859	*0.951*
18	MUN	0.395	*0.486*	**0.303**	*−0.356*	−0.479	**−0.712**	**0.603**	0.699	*0.704*
19	GLN	0.664	*0.806*	**0.366**	**0.578**	0.718	*0.853*	**0.773**	0.888	*0.932*
20	HTN	**0.990**	*0.999*	0.993	**0.954**	*0.995*	0.994	0.899	*0.983*	**0.829**
21	HCN	0.743	*0.874*	**0.241**	**0.791**	0.922	*0.936*	**0.552**	0.656	*0.829*
22	ISP	**0.602**	*0.786*	0.722	**0.644**	0.813	*0.893*	0.566	*0.737*	**0.469**
23	KCN	0.786	*0.895*	**0.772**	**0.647**	0.775	*0.917*	**0.811**	0.905	*0.918*
24	KFP	0.766	*0.877*	**0.470**	**0.843**	*0.945*	0.931	**0.370**	*0.500*	0.467
25	LMN	**0.551**	0.675	*0.800*	**0.738**	*0.868*	0.847	**0.612**	0.745	*0.747*
26	MDN	0.997	*1.000*	**0.992**	**0.940**	0.990	*0.994*	**0.807**	*0.936*	0.935
27	MTB	0.737	*0.872*	**0.341**	**0.682**	0.835	*0.924*	**0.622**	*0.746*	0.729

Table 3 continued

#	Net.	Degree centrality (DegC) vs. closeness centrality (CLC)			Degree centrality (DegC) vs. eigenvector centrality (EVC)			Degree centrality (DegC) vs. betweenness centrality (BWC)		
		Kendall	Spearman	Pearson	Kendall	Spearman	Pearson	Kendall	Spearman	Pearson
28	MCE	0.990	*1.000*	**0.982**	**0.874**	0.957	*0.977*	**0.701**	0.834	*0.885*
29	MSJ	0.488	*0.580*	**0.217**	**0.090**	0.120	*0.508*	0.453	*0.520*	**0.392**
30	AFB	0.272	*0.303*	**−0.183**	*−0.267*	−0.361	**−0.720**	0.424	*0.576*	**0.259**
31	MPN	**0.643**	0.780	*0.881*	**0.692**	0.838	0.907	**0.780**	*0.905*	0.892
32	MMN	0.865	*0.943*	**0.733**	**0.734**	0.851	0.877	**0.781**	*0.903*	0.842
33	NSC	0.595	*0.711*	**0.240**	*−0.092*	−0.107	**−0.511**	**0.416**	*0.485*	0.431
34	PBN	**0.418**	*0.585*	0.582	**0.515**	0.663	*0.670*	**0.515**	0.677	*0.712*
35	PSN	**0.869**	*0.974*	0.952	**0.895**	*0.983*	0.982	**0.749**	*0.913*	0.838
36	PFN	**0.761**	*0.884*	0.875	**0.733**	*0.863*	0.843	**0.659**	0.804	*0.849*
37	SJN	**0.486**	0.618	*0.672*	**0.413**	0.536	*0.664*	**0.577**	0.722	*0.812*
38	SDI	0.416	*0.520*	**0.379**	0.398	*0.512*	**0.324**	**0.660**	*0.792*	0.737
39	SPR	**0.870**	*0.968*	0.930	**0.866**	0.968	*0.976*	0.723	*0.872*	0.835
40	SWC	**0.864**	*0.954*	0.941	**0.874**	0.964	*0.968*	0.742	0.863	*0.905*
41	SSM	**0.610**	0.696	*0.782*	**0.585**	0.714	*0.780*	0.584	0.708	*0.851*
42	TEN	**0.524**	*0.629*	0.612	**0.650**	0.774	*0.776*	0.624	0.750	*0.859*
43	TWF	**0.279**	*0.344*	0.326	**0.235**	0.294	*0.523*	0.338	*0.433*	**0.218**
44	UKF	**0.759**	0.904	*0.918*	**0.799**	0.928	*0.944*	**0.624**	*0.794*	0.782
45	APN	**0.670**	*0.823*	0.803	**0.725**	0.864	*0.956*	0.719	*0.863*	**0.705**
46	USS	**0.582**	0.746	*0.755*	**0.667**	0.799	*0.832*	**0.730**	*0.864*	0.744
47	RHF	**0.724**	0.881	*0.891*	**0.715**	0.876	*0.892*	**0.669**	*0.843*	0.841
48	WSB	**0.904**	0.971	*0.975*	**0.909**	*0.983*	0.982	**0.866**	*0.964*	0.895
49	WTN	0.993	*0.999*	**0.987**	**0.851**	0.954	*0.983*	**0.845**	*0.949*	0.908
50	YPI	0.398	*0.506*	**0.191**	**0.330**	*0.422*	0.349	**0.834**	*0.917*	0.847
# Lowest		**27/50**	**0/50**	**23/50**	**46/50**	**0/50**	**4/50**	**41/50**	**0/50**	**9/50**
# Largest		*0/50*	*40/50*	*10/50*	*3/50*	*13/50*	*34/50*	*0/50*	*34/50*	*16/50*

the largest value (in italics font) for the correlation coefficient. We also plot the coefficient values (Kendall's vs. Spearman's and Pearson's correlation measures) in Figs. 10, 11 and 12 as well as the difference in the correlation coefficient values between any two correlation measures (in Figs. 13, 14) to facilitate visual analytics of the results.

5.1 On the sufficiency of a single correlation measure

The results presented in Tables 3 and 4 and Figs. 10, 11 and 12 confirm our claim that a single correlation measure (like the most commonly used Pearson's correlation measure) is not sufficient to assess all the three levels of correlation. There are a total of 600 data points in Figs. 10, 11 and 12: if a single correlation measure is sufficient to assess all the three levels of correlation, we would need a majority of these data points to fall on the diagonal line, implying that the correlation coefficient values determined using the three correlation measures should be equal or close enough to each other. However, we do not observe such a trend in Figs. 10, 11 and 12 as well as in Tables 3 and 4. There are several instances in Figs. 10, 11 and 12 for which the values for the Kendall's concordance-based correlation coefficient is significantly different from that of the Spearman's and Pearson's correlation coefficients, implying the latter is not an appropriate measure to assess the predictability of the pairwise relative ordering of the vertices on the basis of a computationally-light metric in lieu of a computationally-heavy metric.

For each combination of computationally-light vs. computationally-heavy centrality metrics, we also determine the difference in the correlation coefficient values between any two correlation measures and plot a sorted list of this difference in the correlation coefficient values (a total of 150 data points for each of the six combinations of the centrality metrics), as shown in Fig. 13. If all the three correlation measures were to yield the same or close enough values for

Table 4 LCC′DC vs. computationally-heavy metrics: results of correlation analysis

#	Net.	LCC′DC vs. closeness centrality (CLC)			LCC′DC vs. eigenvector centrality (EVC)			LCC′DC vs. betweenness centrality (BWC)		
		Kendall	Spearman	Pearson	Kendall	Spearman	Pearson	Kendall	Spearman	Pearson
1	ADJ	**0.655**	*0.824*	0.799	**0.676**	0.850	*0.920*	**0.789**	0.916	*0.930*
2	AKN	**0.507**	0.621	*0.769*	**0.540**	0.664	*0.855*	0.951	*0.994*	**0.948**
3	JBN	**0.726**	*0.891*	0.782	**0.608**	0.788	*0.793*	**0.717**	*0.860*	0.757
4	CEN	**0.499**	*0.685*	0.661	**0.535**	0.719	*0.825*	**0.774**	*0.923*	0.816
5	CLN	0.720	*0.874*	**0.221**	**0.759**	*0.908*	0.907	**0.837**	*0.954*	0.887
6	CGD	0.697	*0.852*	**0.432**	**0.633**	*0.797*	0.744	**0.846**	*0.956*	0.860
7	CFN	**0.649**	0.767	*0.903*	**0.622**	0.766	*0.823*	0.954	*0.993*	**0.897**
8	DON	**0.604**	*0.780*	0.765	**0.513**	0.663	*0.703*	0.711	*0.861*	**0.709**
9	DRN	0.573	*0.700*	**0.490**	**0.495**	0.610	*0.613*	0.894	*0.975*	**0.696**
10	DLN	**0.768**	*0.903*	0.882	**0.654**	0.817	*0.845*	**0.755**	*0.872*	0.846
11	ERD	0.639	*0.798*	**0.221**	**0.581**	0.741	*0.870*	**0.810**	*0.936*	0.831
12	FMH	0.560	*0.684*	**0.464**	**0.463**	*0.586*	0.511	0.888	*0.973*	**0.718**
13	FHT	0.787	*0.923*	**0.303**	**0.646**	0.827	*0.829*	**0.863**	*0.959*	0.900
14	FTC	**0.652**	0.821	*0.845*	**0.432**	0.579	*0.700*	**0.784**	*0.918*	0.913
15	FON	**0.366**	0.506	*0.552*	**−0.009**	−0.007	*0.011*	**0.447**	0.608	*0.673*
16	CDF	**0.895**	0.981	*0.982*	**0.850**	*0.967*	0.946	**0.869**	*0.968*	0.935
17	GD96	**0.552**	*0.659*	0.562	**0.568**	0.684	*0.860*	**0.759**	0.859	*0.942*
18	MUN	0.379	*0.472*	**0.222**	*−0.344*	−0.440	**−0.548**	0.955	*0.995*	**0.861**
19	GLN	0.498	*0.642*	**0.307**	**0.411**	0.530	*0.753*	**0.856**	*0.952*	0.944
20	HTN	**0.914**	0.987	*0.990*	**0.864**	*0.972*	0.963	0.939	*0.994*	**0.884**
21	HCN	0.539	*0.645*	**0.100**	**0.486**	0.603	*0.784*	0.948	*0.993*	**0.938**
22	ISP	**0.559**	*0.756*	0.692	**0.583**	0.771	*0.848*	0.611	*0.787*	**0.509**
23	KCN	**0.759**	*0.874*	0.766	**0.549**	0.680	*0.867*	**0.886**	*0.960*	0.930
24	KFP	0.600	*0.749*	**0.408**	**0.521**	0.674	*0.736*	**0.663**	*0.807*	0.705
25	LMN	**0.516**	0.639	*0.757*	**0.525**	*0.683*	0.585	**0.923**	*0.987*	0.931
26	MDN	**0.792**	0.925	*0.950*	**0.711**	0.871	*0.913*	**0.950**	*0.995*	0.982
27	MTB	0.528	*0.662*	**0.186**	**0.431**	0.548	*0.701*	0.896	*0.981*	**0.874**
28	MCE	**0.679**	0.802	*0.946*	**0.546**	0.638	*0.790*	0.955	*0.996*	**0.942**
29	MSJ	0.331	*0.401*	**0.277**	**0.061**	0.076	*0.082*	0.955	*0.996*	**0.610**
30	AFB	0.429	*0.549*	**0.123**	−0.044	−0.045	**−0.224**	0.726	*0.871*	**0.543**
31	MPN	**0.618**	0.780	*0.862*	**0.597**	0.778	*0.838*	**0.830**	0.938	*0.941*
32	MMN	0.732	*0.856*	**0.705**	**0.573**	0.701	*0.761*	**0.868**	*0.962*	0.888
33	NSC	0.312	*0.383*	**0.281**	**0.003**	0.005	*0.020*	0.963	*0.997*	**0.703**
34	PBN	**0.479**	*0.674*	0.627	**0.381**	0.516	*0.591*	**0.691**	*0.864*	0.779
35	PSN	**0.881**	*0.981*	0.954	**0.786**	0.941	*0.943*	**0.824**	*0.955*	0.883
36	PFN	**0.628**	*0.788*	0.777	**0.485**	0.632	*0.677*	**0.811**	*0.929*	0.882
37	SJN	**0.462**	0.615	*0.670*	**0.333**	0.432	*0.579*	**0.708**	0.851	*0.861*
38	SDI	0.407	*0.514*	**0.363**	0.391	0.502	**0.318**	**0.665**	*0.793*	0.730
39	SPR	**0.747**	*0.904*	0.882	**0.713**	0.886	*0.914*	**0.763**	*0.905*	0.880

Table 4 continued

#	Net.	LCC′DC vs. closeness centrality (CLC)			LCC′DC vs. eigenvector centrality (EVC)			LCC′DC vs. betweenness centrality (BWC)		
		Kendall	Spearman	Pearson	Kendall	Spearman	Pearson	Kendall	Spearman	Pearson
40	SWC	**0.621**	0.768	*0.841*	**0.597**	0.767	*0.848*	**0.771**	0.883	*0.927*
41	SSM	**0.570**	0.686	*0.804*	**0.390**	0.494	*0.613*	**0.795**	*0.906*	0.847
42	TEN	**0.562**	0.703	*0.717*	**0.401**	0.520	*0.636*	**0.850**	0.939	*0.942*
43	TWF	0.382	*0.478*	**0.344**	**0.177**	0.241	*0.388*	0.795	*0.904*	**0.696**
44	UKF	**0.637**	0.806	*0.848*	**0.554**	0.719	*0.801*	**0.818**	*0.949*	0.908
45	APN	**0.583**	*0.733*	0.687	**0.579**	0.735	*0.827*	0.882	*0.973*	**0.825**
46	USS	**0.528**	*0.701*	0.693	**0.579**	0.733	*0.766*	**0.751**	*0.889*	0.770
47	RHF	**0.748**	*0.907*	0.902	**0.596**	0.777	*0.808*	**0.787**	*0.934*	0.903
48	WSB	**0.850**	0.962	*0.967*	**0.810**	*0.947*	0.940	**0.912**	*0.986*	0.948
49	WTN	**0.820**	0.929	*0.992*	**0.672**	0.827	*0.948*	0.956	*0.995*	**0.944**
50	YPI	0.390	*0.496*	**0.199**	**0.324**	*0.414*	0.333	0.910	*0.980*	**0.849**
# Lowest		**32/50**	**0/50**	**18/50**	**47/50**	**0/50**	**3/50**	**32/50**	**0/50**	**18/50**
# Largest		*0/50*	*33/50*	*17/50*	*1/50*	*8/50*	*41/50*	*0/50*	*43/50*	*7/50*

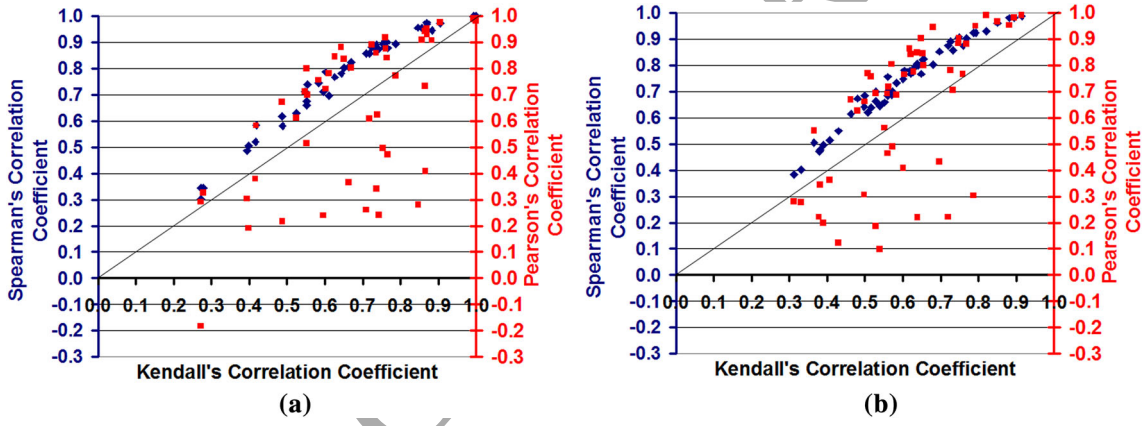


Fig. 10 Kendall's vs. Spearman's and Pearson's correlation coefficients: {DegC, LCC′DC} vs. CLC. **a** DegC − CLC correlation analysis, **b** LCC′DC − CLC correlation analysis

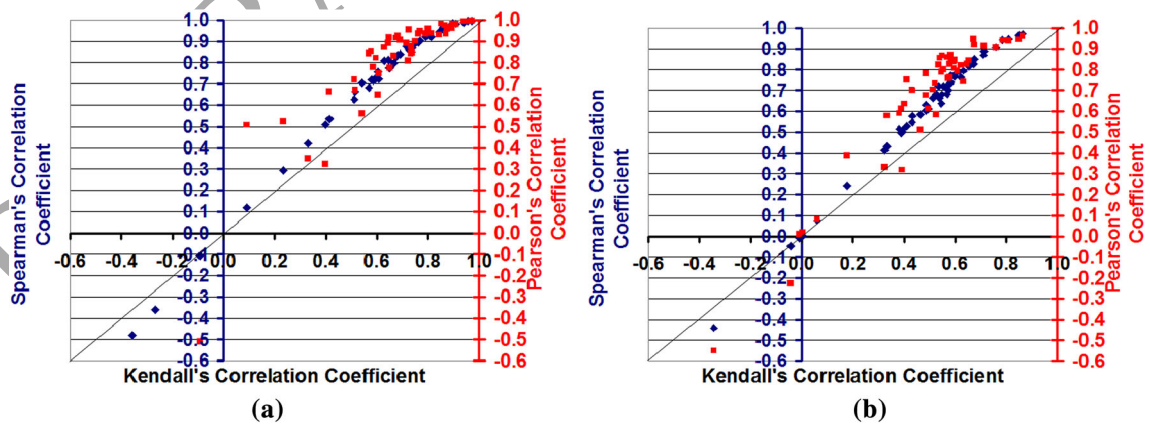


Fig. 11 Kendall's vs. Spearman's and Pearson's correlation coefficients: {DegC, LCC′DC} vs. EVC. **a** DegC − EVC correlation analysis, **b** LCC′DC − EVC correlation analysis

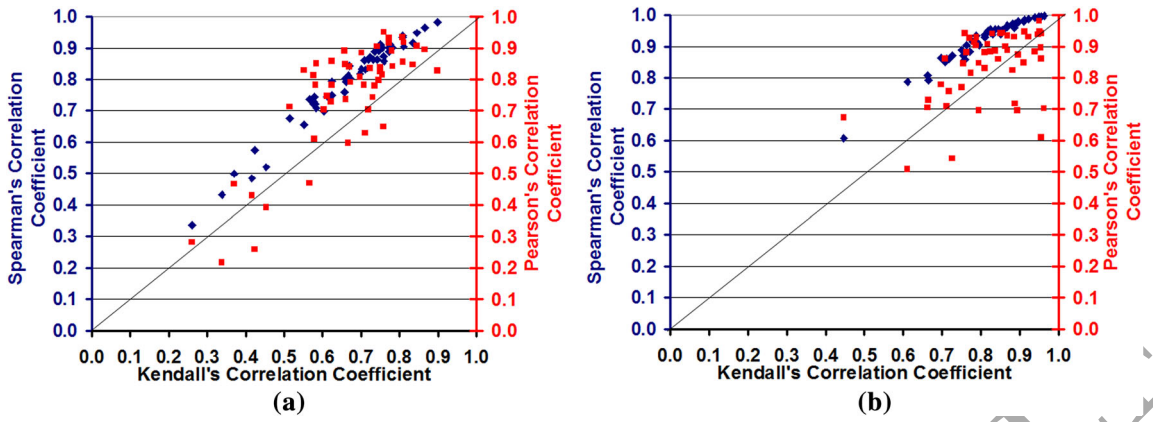


Fig. 12 Kendall's vs. Spearman's and Pearson's correlation coefficients: {DegC, LCC′DC} vs. BWC. **a** DegC − BWC correlation analysis, **b** LCC′DC − BWC correlation analysis

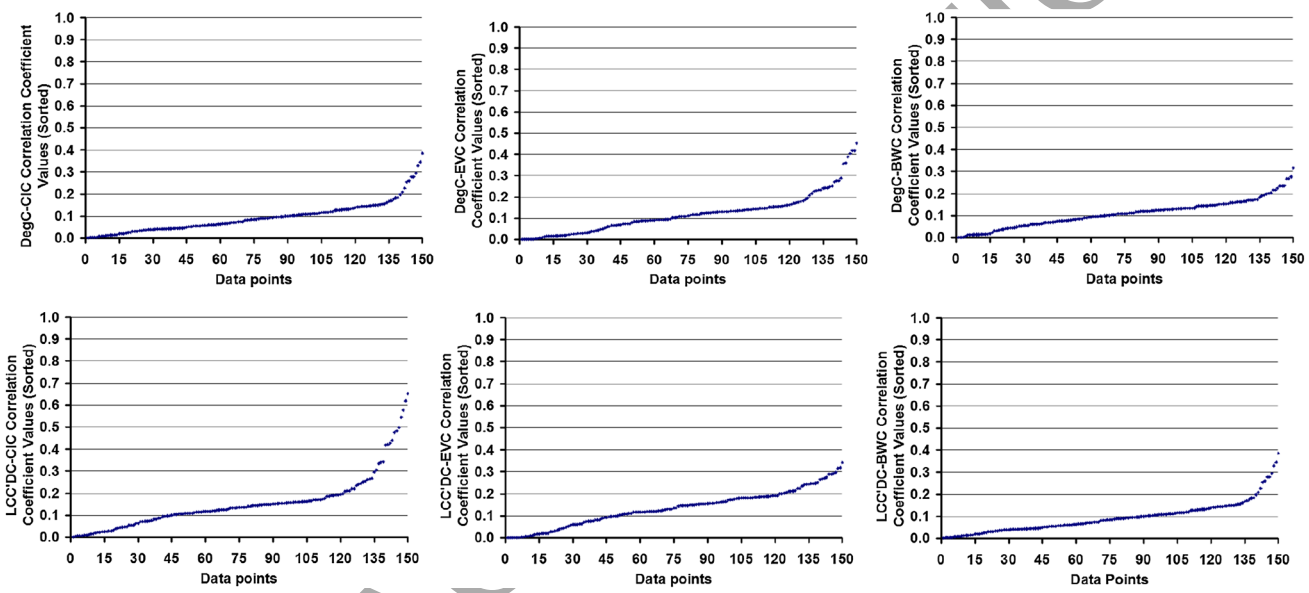


Fig. 13 Distribution of the difference in the correlation coefficient values (sorted order) between any two correlation measures computed for {DegC, LCC′DC} vs. {CLC, EVC, BWC} metrics

the correlation coefficient, we should have only obtained a flat line for each of the plots in Fig. 13. However, we see that the difference in the correlation coefficient values could be as large as 0.3–0.7. We used threshold values of 0.05 and 0.10 for the difference in the correlation coefficient values and determined the fraction of the 150 data points (for each of the six combinations) for which the difference exceeds the threshold (see Fig. 14). We observe that at least 40% of the data points had a difference in the correlation coefficient values of 0.10 or above for each of the six combinations of the centrality metrics evaluated using any two of the three correlation measures. All of the above confirm our claim that a single correlation measure would not be sufficient to assess all the three levels of correlation that are of interest in this paper.

5.2 Kendall's correlation measure for lower bound of the correlation coefficient

We observe the Kendall's correlation coefficient measure to incur the lowest of the correlation coefficient values for 114 of the 150 combinations in the case of DegC vs. the computationally-heavy centrality metrics {CLC, EVC, BWC} and for 111 of the 150 combinations in the case of BWC vs. the three computationally-heavy centrality metrics. Hence, we observe the Kendall's concordance-based correlation measure to be the lowest of the three correlation coefficient values for a total of $(114 + 111) = 225$ of the 300 combinations. As we analyze real-world networks, whose degree distribution ranges from Poisson to Power law (with spectral radius ratio for node degree [15]

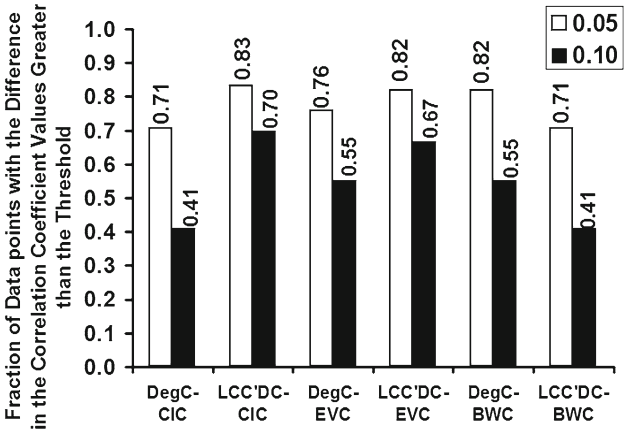


Fig. 14 Fraction of the data points with the difference in the correlation coefficient values between any two correlation measures greater than the threshold values of 0.05 and 0.10

ranging from 1.01 to 5.5) and covering different domains (such as social networks, citation networks, geographical networks, co-appearance networks, biological networks, etc.), we claim that 75% (or the equivalent decimal value of $225/300 = 0.75$) could be considered as the probability with which the Kendall's concordance-based correlation coefficient observed for a computationally-light metric vs. a computationally-heavy metric would serve as a lower bound for the correlation coefficient expected between the same two centrality metrics under the Spearman's and Pearson's measures for any real-world network. The Spearman's rank-based correlation measure did not incur the lowest among the three correlation coefficient values for even one of the 300 combinations. The Pearson's correlation measure incurred the lowest correlation coefficient values for the remaining 25% of the 300 combinations of the computationally-light vs. computationally-heavy centrality metrics and the real-world network graphs.

Figures 10, 11 and 12 present a visual analysis of the Kendall's correlation coefficient values vs. the Spearman's and Pearson's correlation coefficient values obtained for the computationally-light {DegC, LCC′DC} vs. the computationally-heavy {CLC, EVC, BWC} centrality metrics. If a data point lies above the diagonal line, then the Kendall's correlation coefficient for that combination is lower compared to the other correlation measure (either Spearman's or Pearson's depending on the case). Hence, larger the number of data points that are above the diagonal line, the larger the number of combinations of centrality metrics and real-world network graphs for which the Kendall's correlation coefficient is the lowest. We observe more than 95% of the blue data points (corresponding to the Spearman's correlation measure) to be above the diagonal line in both the sub-figures (a) and (b) of Figs. 10, 11 and 12. It is only the 25% of the red data points (corresponding to the Pearson's correlation measure) that are below the diagonal line, especially in the case of the computationally-light metrics vs. the CLC centrality metric. The Kendall's correlation coefficient is the lowest of the three correlation measures for more than 90% of the data points corresponding to the case of the computationally-light metrics vs. the EVC centrality metric.

The results thus convince us that the Kendall's concordance-based correlation measure should ideally be the first correlation measure one should compute between two centrality metrics (especially for correlation studies involving computationally-light vs. computationally-heavy centrality metrics) for a chosen real-world network and decide to proceed further based on the correlation coefficient value obtained. If we observe a strong correlation between a computationally-light centrality metric and a computationally-heavy centrality metric for a real-world network with respect to the Kendall's measure, there would not be even a need to compute the correlation coefficient with respect to the other two correlation measures (Spearman's and Pearson's) as there is a 0.75 chance that these correlation coefficient values will be at least the value observed for the Kendall's concordance-based correlation coefficient. From Tables 3 and 4, we also observe that the Kendall's correlation measure incurs the largest correlation coefficient value for only 8 of the 300 combinations (i.e., less than 3% of the 300 combinations). Hence, we could also conclude that the Kendall's correlation coefficient is more likely not the largest of the three correlation values with a probability of $1 - 8/300 \sim 0.97$.

5.3 Analysis of the median values for the correlation coefficient

Figure 15a–c display the median values for the correlation coefficient observed for the three levels of correlation between a computationally-light metric {DegC, LCC′DC} with each of the computationally-heavy metrics {CLC, EVC, BWC} for the 50 real-world network graphs. Figure 15d displays the median when the correlation coefficient values for all the three levels of correlation are considered together (here after referred simply as *overall*) for a particular combination of the computationally-light and computationally-heavy metrics. Similar to the trend observed in Figs. 3, 4, 10, 11 and 12, we also notice that irrespective of the computationally-light vs. computationally-heavy centrality metric combination, the median of the correlation coefficient observed for pairwise relative ordering of the vertices is the lowest among the correlation coefficient values for all the three levels of correlation.

With respect to the individual combination of centrality metrics, we consistently observe the degree centrality metric (DegC) to exhibit higher levels of correlation with the closeness and eigenvector centrality metrics for each of the three levels of correlation as well as overall, whereas we observe

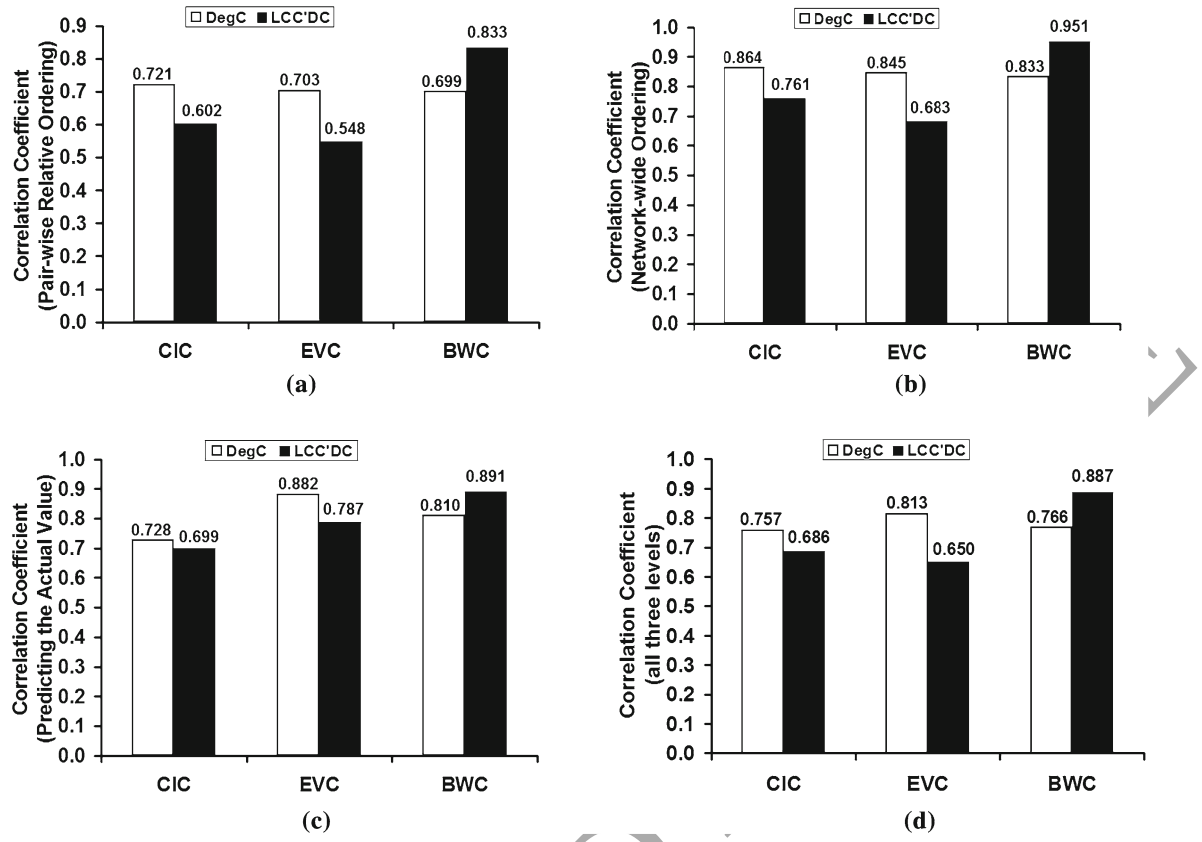


Fig. 15 Median values for the correlation coefficient for each level of correlation and all the three levels: {DegC, LCC′DC} vs. {CLC, EVC, BWC}. **a** Pair-wise relative ordering (Kendall's measure), **b** network-wide ranking (Spearman's measure), **c** predicting the actual values (Pearson's measure), **d** all three levels (Kendall's, Spearman's, Pearson's measures)

the local clustering coefficient complement-based degree centrality (LCC′DC) metric to exhibit relatively stronger correlation with the betweenness centrality (BWC) metric for each of the three levels of correlation as well as overall. We thus conclude that for each of the three levels of correlation, the DegC metric could serve as the computationally-light alternative for the CLC and EVC metrics, whereas, the LCC′DC metric could serve as the computationally-light alterative for the BWC metric.

5.4 LCC′DC–BWC correlation vs. DegC–BWC correlation

We observe the LCC′DC metric to exhibit a very strong correlation with the BWC metric (the most time-consuming metric of the three computationally-heavy centrality metrics) and the data points (in Fig. 12b) are located relatively closer to 1 and also closer to each other, indicating that the correlation coefficient values with respect to the three correlation measures converge towards the largest possible value of 1 for a majority of the real-world network graphs. Considering a total of 150 LCC′DC–BWC correlation coefficient values obtained with respect to the three correlation measures for the 50 real-world network graphs, we observe (see Fig. 15d) the median to be 0.887 (the largest median value for each of the six combinations of computationally-light vs. computationally-heavy centrality metrics: see Fig. 15a–c) and only 12 of the 150 correlation coefficient values (i.e., less than 10%) are below 0.7. On the other hand, the median of the 150 DegC–BWC correlation coefficient values for the three correlation measures analyzed for the 50 real-world network graphs is 0.766 (see Fig. 15d), appreciably lower than the median value of 0.887 for the LCC′DC–BWC correlation.

5.5 Applications

The results of our correlation study confirm our hypothesis that the Kendall's concordance-based correlation coefficient is more likely to be the smallest of the three correlation coefficients computed between a computationally-light vs. computationally-heavy metric for a real-world network. An equally interesting observation is that the Spearman's correlation coefficient is not the lowest for any real-world network.

Thus, if one were to observe a higher value for the Kendall's correlation coefficient between a computationally-light metric (say, LCC′DC) and a computationally-heavy metric (say, BWC), then it is more likely that the ranking of the vertices with respect to the computationally-light metric is more likely to be similar to the ranking of the vertices with respect to the computationally-heavy metric.

6 Related work and our contributions

The idea of studying correlation between computationally-light centrality metrics and the computationally-heavy centrality metrics was recently mooted by Li et al. [5] in which the Pearson's correlation coefficient was used as the correlation measure to evaluate the extent to which one could rank the vertices using a computationally-light centrality metric in lieu of a computationally-heavy centrality metric. However, as seen in this paper, the Pearson's correlation coefficient values are different from those of the Spearman's and Kendall's rank-based correlation measures for at least the computationally-light centrality metrics vs. computationally-heavy shortest path-based centrality metrics. In another recent work [55], it has been observed that the Kendall's concordance-based correlation measure is more suitable to evaluate pairwise correlation, especially among the top-k ranked vertices, whereas the Spearman's correlation measure is more suitable to evaluate rank-based correlation involving all the vertices, especially if several of them have equal ranks. The three correlation measures were also recently used [56] to analyze the extent to which one can predict flux changes using the functional centrality metric [57] proposed to quantify the functional relevance of individual biochemical reactions in metabolic networks. In [8], the computationally-light degree centrality metric and the computationally-heavy eigenvector centrality metric were observed to be strongly correlated with the computationally-heavy maximal clique size for a node (the maximal clique size for a node is the clique of the largest size that a node is part of) under all the three levels of correlation. The BWC metric has been observed to be weakly correlated (correlation coefficient values in the range 0...0.5) with the maximal clique size. Unlike the work in [8], wherein the correlation between centrality metrics and maximal clique size was studied, in this paper, we investigate the correlation among the centrality metrics themselves on the lines of computationally-light {DegC, LCC′DC} vs. computationally-heavy {CLC, EVC, BWC} centrality metrics. The LCC′DC metric was also not considered in [8].

In [16], the author analyzed the assortativity of 50 real-world network graphs (also considered in this paper) based on each of the four centrality metrics DegC, BWC, EVC, and CLC. The assortativity index of a network with respect to a centrality metric is a quantitative measure of how similar are the values for the end vertices of the edges with respect to the centrality metric, and is measured as the Pearson's correlation coefficient of the centrality values of the end vertices of the edges. Networks with larger positive (negative) values for the assortative index are considered to be assortative (dissortative) with respect to the centrality metric under consideration. Networks whose assortative index values are closer to zero are considered to be neutral with respect to the centrality metric considered. It was observed that real-world networks are more likely to be neutral with respect to DegC and BWC, and more likely to be assortative with respect to EVC and CLC. Our work in this paper primarily differs from [16] on these lines: we do not measure the correlation coefficient between the centrality values of the end vertices of the edges with respect to a metric. We rather measure the correlation coefficient between the vertices with respect to two different centrality metrics (computationally-light vs. computationally-heavy). In addition, the correlation coefficient in [16] was measured only using the Pearson's correlation measure; in this paper, we use three different measures of correlation. Likewise, LCC′DC was not considered in the assortativity analysis study of [16], whereas, LCC′DC is considered in this paper.

In [58], the author developed a new centrality metric called CIRank (that keeps track of the changes propagating among classes in a software dependency network) and observed it to be significantly correlated with the degree and PageRank centrality metrics on the basis of the Spearman's rank-based correlation coefficient. In [59], the Kendall's concordance-based correlation measure was used to assess the correlation between eight different centrality metrics that are suitable for gene regulatory networks in *E. Coli*. It was observed that the ranking of the genes with respect to the centrality metrics is significantly different (leading to a low correlation coefficient between any two centrality metrics), especially when vertices (genes) with non-zero out degree are only considered. In another related study [60], the Kendall's measure was used to study the correlation between DegC, CLC, BWC, and EVC metrics for the *M. musculus* protein–protein interaction network. In [61], the authors studied the impact of removing the top-k ranked vertices (with respect to a centrality metric) on the traffic-carrying capacity of the remaining nodes and the connectivity of ISP (Internet Service Provider) networks: removal of the top-k vertices with respect to the locally computable degree centrality metric had a similar impact on the traffic-carrying capacity of the remaining nodes vis-a-vis removal of the top-k vertices with respect to the globally computable centrality metrics.

Though some of the recent works in the literature (as mentioned above) have also used the three correlation measures (Kendall's, Spearman's and Pearson's) for analyzing the correlation between centrality metrics with respect to the three

levels of correlation (pairwise relative ordering, network-wide ranking, and prediction of the actual values), ours is the first work to evaluate the three levels of correlation from the point of view of computationally-light vs. computationally-heavy centrality metrics and demonstrate that the pairwise relative ordering of the vertices could be the most restrictive and that the corresponding Kendall's concordance-based correlation measure could serve (with a probability as large as 0.75) as the lower bound for correlation coefficient among the three levels of correlation. We could also conclude that the Kendall's correlation coefficient is not the largest among the correlation coefficients of the three measures with a probability of 0.97.

7 Conclusions

We observe the pairwise relative ordering of vertices based on a computationally-light metric in lieu of a computationally-heavy centrality metric to be the most restrictive of all the three levels of correlation and the Kendall's concordance-based correlation coefficient (that is a measure of the level of correlation to assess the pairwise relative ordering of vertices) could be considered (with a probability as large as 0.75) to serve as the lower bound for correlation coefficient between a computationally-light and computationally-heavy centrality metric. Likewise, we could also conclude that the Kendall's correlation coefficient is more likely not the largest of the three correlation measures (between a computationally-light and computationally-heavy centrality metric) with a probability of 0.97. Such significant observations on the nature of the correlation coefficient values obtained for the centrality metrics (especially for computationally-light vs. computationally-heavy metrics) with respect to the Kendall's concordance-based correlation measure have been hitherto not reported in the literature. To the best of our knowledge, it is not known in the literature which of these three commonly studied correlation measures is likely to incur the lowest value for the correlation coefficient when a computationally-light vs. computationally heavy metric are correlated. Through this research, we have established that the pairwise comparison-based correlation is the most strongest form of correlation analysis that one could conduct between a computationally-light vs. computationally-heavy centrality metric. As an application of this result, if one were to observe a higher (lower) value for the Kendall's concordance-based correlation coefficient between a computationally-light metric (say, LCC'DC) vs. a computationally-heavy metric (BWC), then one is more likely to observe the ranking of the vertices with respect to the computationally-light metric to be more (less) similar to the ranking of the vertices with respect to the computationally-heavy metric.

References

1. Newman, M.E.J.: Networks: An Introduction, 1st edn. Oxford University Press, Oxford (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. Li, C., Li, Q., Van Mieghem, P., Stamey, H.E., Wang, H.: Correlation between centrality metrics and their application to the opinion model. Eur. Phys. J. B **88**(65), 1–13 (2015)
6. Meghanathan, N.: 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, 27–30 April 2015 (2015)
7. Triola, M.F.: Elementary Statistics, 12th edn. Pearson, New York (2012)
8. Meghanathan, N.: Maximal clique size vs. centrality: a correlation analysis for complex real-world network graphs. In: Proceedings of the 3rd International Conference on Advanced Computing, Networking, and Informatics. Springer Smart Innovation, Systems and Technologies Series, vol. 44, pp. 95–101, 23–25 June 2015, Orissa, India (2015)
9. Meghanathan, N.: A computationally lightweight and localized centrality metric in lieu of betweenness centrality for complex network analysis. Vietnam J. Comput. Sci. **4**(1), 23–38 (2017)
10. Meghanathan, N., He, X.: Correlation and regression analysis for node betweenness centrality. Int. J. Found. Comput. Sci. Technol. **6**(6), 1–20 (2016)
11. Krebs, V.: Proxy networks: analyzing one network to reveal another. Bulletin de Méthodologie Sociologique **79**, 61–70 (2003)
12. Lay, D.C.: Linear Algebra and its Applications, 4th edn. Pearson, New York (2011)
13. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25**(2), 163–177 (2001)
14. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd edn. MIT Press, Cambridge (2009)
15. 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, December 2014 (2014)
16. Meghanathan, N.: Assortativity analysis of real-world network graphs based on centrality metrics. Comput. Inf. Sci. **9**(3), 7–25 (2016)
17. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**(3), 036104 (2006)
18. Knuth, D.E.: The Stanford GraphBase: A Platform for Combinatorial Computing, 1st edn. Addison-Wesley, Reading (1993)

19. Geiser, P., Danon, L.: Community structure in jazz. Adv. Complex Syst. **6**(4), 563–573 (2003)
20. White, J.G., Southgate, E., Thomson, J.N., Brenner, S.: The structure of the nervous system of the nematode *Caenorhabditis elegans*. Philos. Trans. B **314**(1165), 1–340 (1986)
21. Hummon, N.P., Doreian, P., Freeman, L.C.: Analyzing the structure of the centrality-productivity literature created between 1948 and 1979. Sci. Commun. **11**(4), 459–480 (1990). doi:10.1177/107554709001100405
22. Biedl, T., Franz, B.J.: Graph-drawing contest report. In: Proceedings of the 9th International Symposium on Graph Drawing, pp. 513–521, September 2001 (2001)
23. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behav. Ecol. Sociobiol. **54**(3), 396–405 (2003)
24. Lee, J.-S.: Generating networks of illegal drug users using large samples of partial ego-network data. In: Intelligence and Security Informatics. Lecture Notes in Computer Science, vol. 3073, pp. 390–402 (2004)
25. de Nooy, W.: A literary playground: literary criticism and balance theory. Poetics **26**(5–6), 385–404 (1999)
26. Batagelj, V., Mrvar, A.: Pajek datasets (2006). http://vlado.fmf.uni-lj.si/pub/networks/data/
27. Resnick, M.D., Bearman, P.S., Blum, R.W., Bauman, K.E., Harris, K.M., Jones, J., Tabor, J., Beuhring, T., Sieving, R.E., Shew, M., Ireland, M., Bearinger, L.H., Udry, J.R.: Protecting adolescents from harm. Findings from the national longitudinal study on adolescent health. J. Am. Med. Assoc. **278**(10), 823–832 (1997)
28. Krackhardt, D.: The ties that torture: Simmelian tie analysis in organizations. Res. Sociol. Organ. **16**, 183–210 (1999)
29. Moreno, J.L.: The Sociometry Reader, pp. 534–547. The Free Press, Glencoe (1960)
30. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**(12), 7821–7826 (2002)
31. Bernard, H.R., Killworth, P.D., Sailer, L.: Informant accuracy in social network data IV: a comparison of clique-level structure in behavioral and cognitive network data. Soc. Netw. **2**(3), 191–218 (1980)
32. Gleiser, P.M.: How to become a superhero. J. Stat. Mech. Theory Exp. **2007**(9), P09020 (2007)
33. Isella, L., Stehle, J., Barrat, A., Cattuto, C., Pinton, J.F., Van den Broeck, W.: What's in a crowd? Analysis of face-to-face behavioral networks. J. Theor. Biol. **271**(1), 166–180 (2011). doi:10.1016/j.jtbi.2010.11.033
34. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**(4), 452–473 (1977)
35. Rogers, E.M., Kincaid, D.L.: Communication Networks: Toward a New Paradigm for Research. Free Press, New York (1980)
36. Takahata, Y.: Diachronic changes in the dominance relations of adult female Japanese monkeys of the Arashiyama B group. In: The Monkeys of Arashiyama, pp. 124–139. State University of New York Press, Albany (1991)
37. Hayes, B.: Connecting the dots. Am. Sci. **94**(5), 400–404 (2006)
38. 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, Brighton (2004)
39. McCarty, C., Freeman, L.: (2008). http://moreno.ss.uci.edu/data.html
40. Gil-Mendieta, J., Schmidt, S.: The political network in Mexico. Soc. Netw. **18**(4), 355–381 (1996)
41. Gemmetto, V., Barrat, A., Cattuto, C.: Mitigation of infectious disease at school: targeted class closure vs. school closure. BMC Infect. Dis. **14**(695), 1–10 (2014)
42. MacRae, D.: Direct factor analysis of sociometric data. Sociometry **23**(4), 360–371 (1960)
43. 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)
44. Scott, J.P.: The Anatomy of Scottish Capital: Scottish Companies and Scottish Capital, 1900–1979, 1st edn. Croom Helm, London (1980)
45. Grimmer, J.: A Bayesian hierarchical topic mode for political texts: measuring expressed agendas in senate press releases. Polit. Anal. **18**(1), 1–35 (2010)
46. Michael, J.H.: Labor dispute reconciliation in a forest products manufacturing facility. For. Prod. J. **47**(11–12), 41–45 (1997)
47. Schwimmer, E.: 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 (1973)
48. 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)
49. 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)
50. Meghanathan, N., Lawrence, R.: Centrality analysis of the United States network graph. In: Proceedings of the 3rd International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences, pp. 23–28, Tadepalligudem, India, June 1–2, 2016 (2016)
51. 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)
52. Freeman, L.C., Freeman, S.C., Michaelson, A.G.: How humans see social groups: a test of the Sailer–Gaulin models. J. Quant. Anthropol. **1**, 229–238 (1989)
53. Smith, D.A., White, D.R.: Structure and dynamics of the global economy: network analysis of international trade 1965–1980. Soc. Forces **70**(4), 857–893 (1992)
54. Jeong, H., Mason, S.P., Barabasi, A.L., Oltvai, Z.N.: Lethality and centrality in protein networks. Nature **411**, 41–42 (2001)
55. Aprahamian, M., Higham, D.J., Higham, N.J.: Matching exponential-based and resolvent-based centrality measures. J. Complex Netw. **4**(2), 157–176 (2016)
56. Sajitz-Hermstein, M., Nikoloski, Z.: Functional centrality as a predictor of shifts in metabolic flux states. BMC Res. Notes **9**(317), 1–4 (2016)
57. Sajitz-Hermstein, M., Nikoloski, Z.: Restricted cooperative games on metabolic networks reveal functionally important reactions. J. Theor. Biol. **314**, 192–203 (2012)
58. Wang, R., Huang, R., Qu, B.: Network-based analysis of software change propagation. Sci. World J. **2014**, 1–10 (2014) (Article ID 237243)
59. Koschutzki, D., Schreiber, F.: Centrality analysis methods for biological networks and their application to gene regulatory networks. Gene Regul. Syst. Biol. **2008**(2), 193–201 (2008)
60. Dawyer, T., Hong, S.-H., Koschutzki, D., Schreiber, F., Xu, K.: Visual analytics of network centralities. In: Proceedings of the Asia-Pacific Symposium on Information Visualization, vol. 60, pp. 189–197, Tokyo, Japan (2006)
61. Nomikos, G., Pantazopoulos, P., Karaliopoulos, M., Stavrakakis, I.: Comparative assessment of centrality indices and implications on the vulnerability of ISP networks. In: Proceedings of the 26th International Teletraffic Congress, pp. 1–9, Karlskrona, Sweden, September 2014 (2014)