CORRELATION AND REGRESSION ANALYSIS FOR NODE BETWEENNESS CENTRALITY

Natarajan Meghanathan[1] and Xiaojia He[2]

[1]Jackson State University, MS, USA, [2]University of Georgia, GA, USA,

ABSTRACT

In this paper, we seek to find a computationally light centrality metric that could serve as an alternate for the computationally heavy betweenness centrality (BWC) metric. In this pursuit, in the first half of the paper, we evaluate the correlation coefficient between BWC and the other commonly used centrality metrics such as Degree Centrality (DEG), Closeness Centrality (CLC), Farness Centrality (FRC), Clustering Coefficient Centrality (CCC) and Eigenvector Centrality (EVC). We observe BWC to be highly correlated with DEG for synthetic networks generated based on the Erdos-Renyi model (for random networks) and Watts-Strogatz model (for small-world networks). In the second half of the paper, we conduct a regression analysis for BWC with that of a recently proposed centrality metric called the localized clustering coefficient complement-based degree centrality (LCC'DC) for a suite of 47 real-world networks. The R-Squared metric and Correlation coefficient for the LCC'DC-BWC regression has been observed to be appreciably greater than those observed for the DEG-BWC regression. We also observe the LCC'DC-BWC regression to incur relatively a lower value for the standard error of residuals for a majority of the real-world networks.

KEYWORDS

Betweenness Centrality, Random Networks, Small-World Networks, Real-World Networks, Correlation Coefficient, Regression, Standard Error for Residuals, R-Squared, Localized Clustering Coefficient

1. INTRODUCTION

Over the past decades, a large number of centrality measures have been introduced and developed to quantify the significance and importance of the nodes in various networks. Betweenness centrality (BWC) is one of the most widely used measures, first developed in the 1970s by Freeman [1] and Anthonisse [2], independently. BWC is a measure of the degree to which a node functions as the mediation node by calculating the fraction score of all shortest paths (geodesic) between other pairs of nodes that go through it. It is expected that the network would be disconnected if one or two nodes with high BWC were removed. Thus one can expect that a node with high BWC does not belong to one of the dense groups, but connects them. For the rest of the paper, the terms 'node' and 'vertex', 'link' and 'edge', 'network' and 'graph' are used interchangeably. They mean the same.

BWC has been widely applied to a large number of complex network analyses. For instance, it has been proposed as an indicator of the "interdisciplinary" nature of scientific journals [3]. In general, BWC of the nodes in a network increases with connectivity as a power law with an exponent η [4]. Thus, it is known to be computationally time consuming to obtain exact BWC:

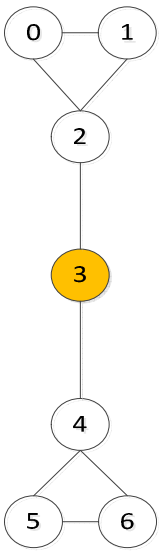
DOI:10.5121/ijfcst.2016.6601

O(nm) time for unweighted graphs and O($nm + n^2 \log n$) time for weighted graphs, where n is the number of vertices and m is the number of edges in the network [5][6][14]. In the first half of this paper, we focus on analyzing the correlation between BWC and five well-known centrality measures, including eigenvector centrality (EVC), degree centrality (DEG), clustering coefficient centrality (CCC), farness centrality (FRC), and closeness centrality (CLC) for synthetic networks generated from theoretical models for random network and small-world networks. In the second half of this paper, we conduct a comprehensive regression analysis to demonstrate that the recently proposed localized clustering coefficient complement-based degree centrality (LCC'DC) [28] could serve as a viable alternative for BWC and be used to predict the BWC of the vertices in real-world networks.

2. COMPUTATION OF BETWEENNESS CENTRALITY

The computation of BWC in this paper follows the algorithm by Brandes (2001) [5]. If the number of shortest paths between two nodes i and j that pass through node k as the intermediate node is denoted as g_{ij}^{k} and the total number of geodesic between the two nodes i and j is denoted as g_{ij}, then the BWC for node k is defined as

$$BWC(k) = \sum_{i} \sum_{j} \frac{g_{ijk}}{g_{ij}} (i \neq j \neq k)$$



Betweenness of node 2:
Node 3→ 0: 1
Node 3→ 1: 1
Node 4→ 0:1
Node 4→1:1
Node 5→0:1
Node 5→1:1
Node 6→0:1
Node 6→1:1
Betweenness Centrality of node 2:8

Betweenness of node 3:
Node 4→ 0: 1
Node 4→ 1: 1
Node 4→ 2:1
Node 5→0:1
Node 5→1:1
Node 5→2:1
Node 6→0:1
Node 6→1:1
Node 6→2:1
Betweenness Centrality of node 3:9

Betweenness of node 4:
Node 5→0:1
Node 5→1:1
Node 5→2:1
Node 5→3:1
Node 6→0:1
Node 6→1:1
Node 6→2:1
Node 6→3:1
Betweenness Centrality of node 4:8

Betweenness Centrality of node 0: 0
Betweenness Centrality of node 1: 0
Betweenness Centrality of node 5: 0
Betweenness Centrality of node 6: 0

Figure 1: Representative Example to Compute the Betweenness Centrality of the Vertices in a Network

The representative BWC calculation is illustrated in Figure 1. On the basis of the algorithm proposed by Brandes (2001) [5], breadth-first search is involved in the computation. It is clear that BWC is different to degree-based ranking as shown in Figure 1. Nodes 3 and 4 have highest degree in this present network; however, node 3 has highest BWC. Nodes 0, 1, 5, and 6 each has a degree of 2, but with a BWC of 0.

3. CORRELATION ANALYSIS FOR THEORETICAL NETWORKS

3.1. Analysis on Random Networks

Firstly, random networks were simulated to investigate all the six centrality measures including BWC, EVC, DEG, CCC, FRC, and CLC. In this section, networks with 100 nodes were simulated. Particularly, the probability of linkage between nodes is varied from 0.05 to 0.9 to evaluate above mentioned centrality measures. The probability of linkage is increased from 0.05 to 0.1 by 0.01; from 0.1 to 0.9 by 0.1. Representative random networks are shown in Figure 2 with a ranking factor of BWC. Correlation between BWC and other five measures, including DEG, EVC, CCC, FRC, and CLC, was then determined. Average correlation coefficient value was calculated based on 100 trials.

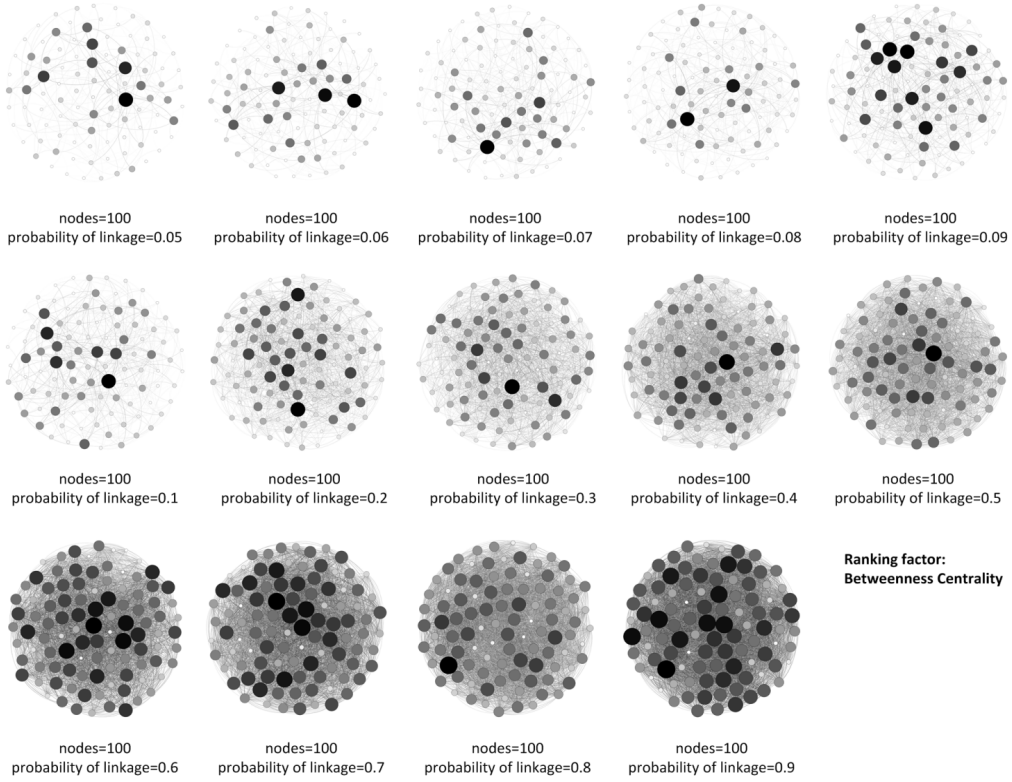


Figure 2: Simulation of Random Networks with Various Probability of Linkage Values
[Ranking Factor is Betweenness Centrality]

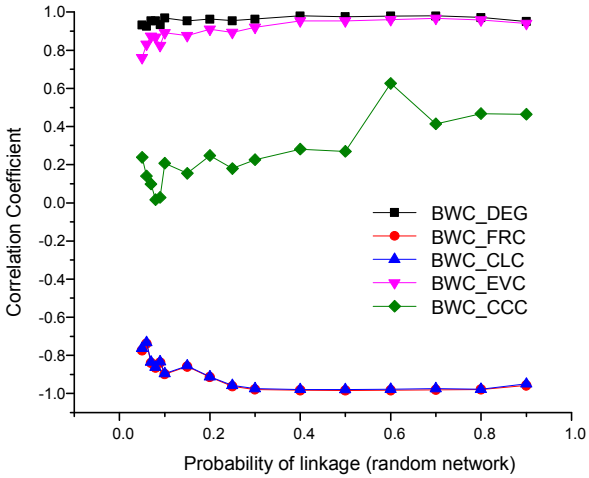


Figure 3: Correlation Coefficient between BWC and the other Five Centrality Measures: DEG, EVC, CCC, FRC and CLC on Random Networks with Various Probability of Linkage Values

As shown in Figure 3, BWC is highly correlated to all measures except CCC. Our data suggests a strong correlation between BWC and DEG, ranging from 0.9316 to 0.9513. The highest correlation of BWC to FRC, CLC, and EVC reaches -0.9576, -0.9495, and 0.94, respectively. The negative correlation indicates that an increase in one variable reliably predicts a decrease in the other one. A high value in negative correlation still suggests high correlation. It is pretty sure that we can select DEG, FRC, CLC, EVC as alternatives to BWC in random networks.

3.2. Analysis on Small-World Networks

We investigated on small-world networks evolved from regular network. Similar to random network simulation, 100 nodes with a *k*-regular value (initial number of links per node) of 10 are set for small-world network simulation. In this section, the probability of rewiring was varied from 0.01 to 0.09 with increment of 0.01; and from 0.1 to 0.9 with increment of 0.1. Representative small-world networks are shown in Figure 4 with a ranking factor of BWC. Correlation between BWC and the other five measures, including DEG, EVC, CCC, FRC, and CLC, was then calculated. Average correlation coefficient value was calculated based on 100 trials.

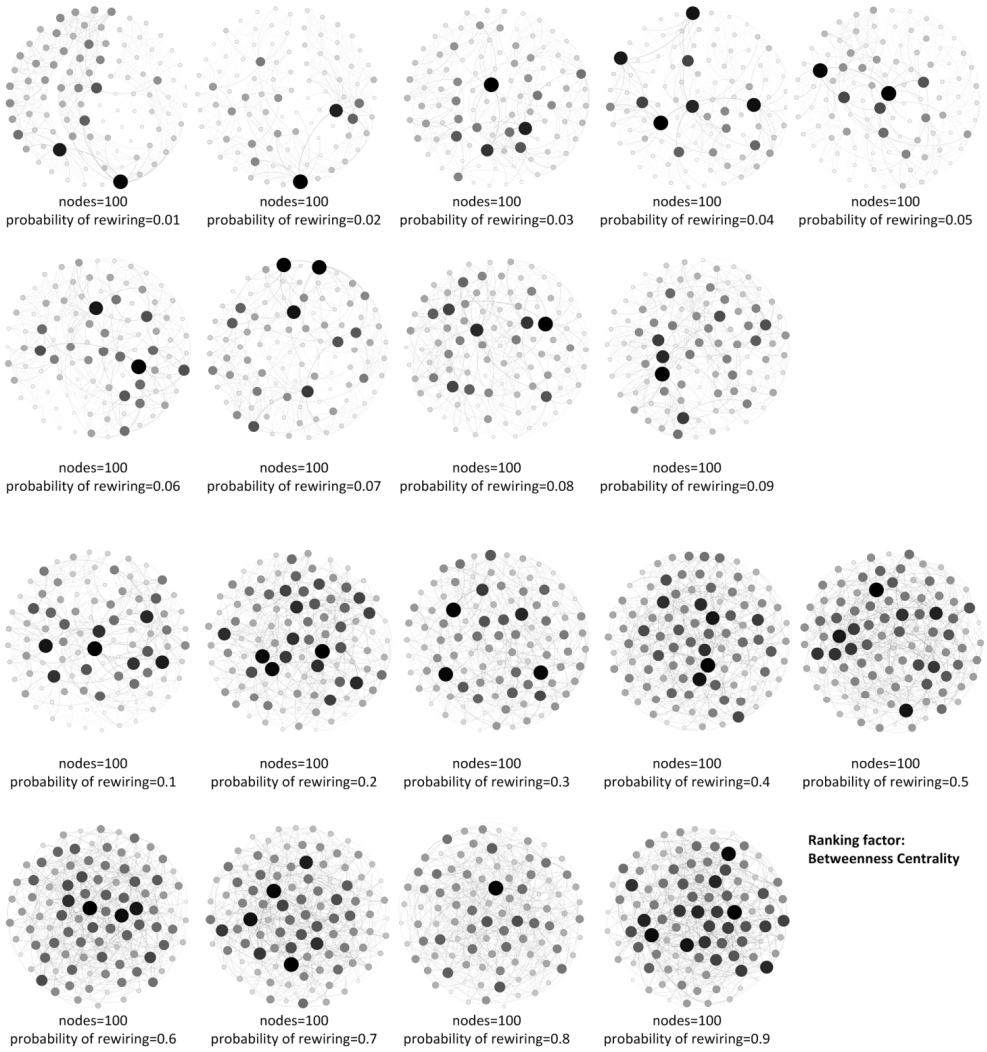


Figure 4: Simulation of Small-World Networks with Various Probability of Rewiring Values [Ranking Factor is Betweenness Centrality]

For small-world networks, there is a strong correlation between BWC and the other centrality metrics, except EVC, at a probability of rewiring lower than 0.2. The correlation coefficient was larger than 0.51 when the probability of rewiring reaches 0.2 for DEG, FRC, CLC, and CCC. The highest correlation coefficient of BWC to DEG, FRC, and CLC reaches to 0.5325, -0.7499, and -0.7348 at probability of rewiring of 0.08. The correlation between BWC and CCC decreases from 0.8131 to 0.0683 along with the increase of probability of rewiring.

In a previous work, a transformation between small-world network and random network was revealed [15]. It was found that simulated network from a regular network would be small-world network when the probability of rewiring is from 0.01 to 0.1; however, it changes to random network when the probability of rewiring is between 0.1 and 1.0. In this study, we also observed a clear turning point at probability of rewiring of 0.1 as shown in Figure 5. Overall, we could

preferably use CCC as alternative to BWC at probability of rewiring lower than 0.07. At a critical probability of rewiring lower than 0.2, we still could use DEG, FRC, CLC, and CCC as alternatives.

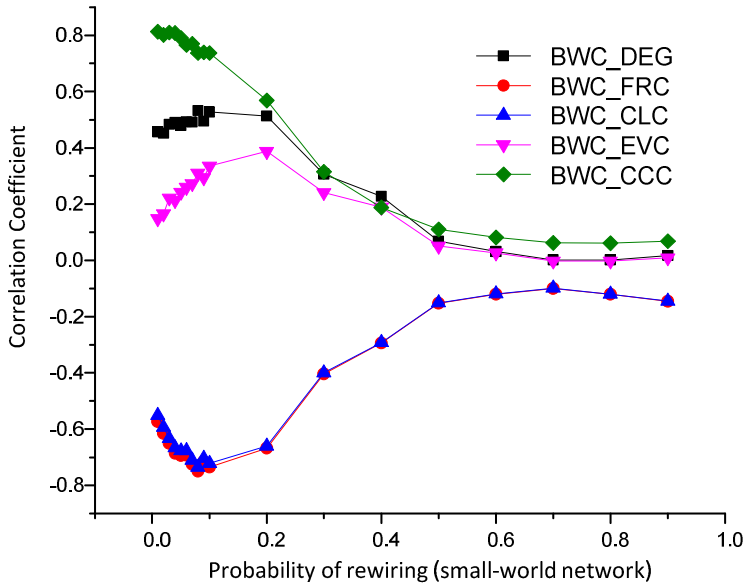
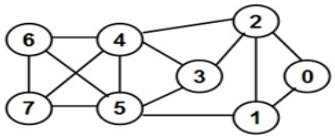


Figure 5: Correlation Coefficient between BWC and the other Five Centrality Measures, including DEG, EVC, CCC, FRC and CLC, on Small-World Networks with Various Probability of Rewiring Values

4. REGRESSION ANALYSIS FOR REAL-WORLD NETWORKS

In a recent work [28], the first author of this paper had proposed a novel centrality metric called localized clustering coefficient complement-based degree centrality (denoted LCC'DC) that has been demonstrated to exhibit the strongest correlation with node betweenness centrality (compared to the other centrality metrics discussed in this paper so far). The localized clustering coefficient (LCC) for a node is a measure of the probability that any two neighbors of the node are connected. The LCC of a node is a measured as the ratio of the actual number of links between any two neighbors of the node to that of the maximum possible number of links between any two neighbors of the node. The localized clustering coefficient complement (denoted LCC') is 1-LCC. Hence, LCC'DC for a node is the product of LCC' and DEG, the degree centrality of the node. Figure 6 shows an example to compute the LCC'DC values of the vertices in a sample graph. Note that LCC of a vertex with degree 1 is 1.0.



Vertex	LCC'DC	BWC
0	0.0	0.0
1	2.0	2.33
2	2.67	4.0
3	1.0	0.33
4	2.5	4.33
5	3.0	5.0
6	0.0	0.0
7	0.0	0.0

Vertex	Degree DegC	Max. Edges among neighbs	Actual Edges among neighbs	LCC	LCC' 1-LCC	LCC'DC
0	2	2(2-1)/2 = 1	1	1.0	0.0	0.0
1	3	3(3-1)/2 = 3	1	0.33	0.67	2.0
2	4	4(4-1)/2 = 6	2	0.33	0.67	2.67
3	3	3(3-1)/2 = 3	2	0.67	0.33	1.0
4	5	5(5-1)/2 = 10	5	0.5	0.5	2.5
5	5	5(5-1)/2 = 10	4	0.4	0.6	3.0
6	3	3(3-1)/2 = 3	3	1.0	0.0	0.0
7	3	3(3-1)/2 = 3	3	1.0	0.0	0.0

Figure 6: Example to Illustrate the Computation of the LCC'DC Metric

In this section, we conduct linear regression analysis between the BWC and LCC'DC values vis-a-vis the DEG values for a suite of 47 real-world networks. In other words, we fit a straight line to predict the BWC values for the nodes using the LCC'DC values and compare the goodness of the fitted straight line with that of straight line fitted using the DEG values. We normalize the values for the BWC, LCC'DC and DEG metrics (i.e., the values for all the three metrics are normalized to a range of 0...1). We use the Ordinary Least Squares (OLS) method for the linear regression [30]; the objective of this regression model is to minimize the sum of the squares of the residuals (in the context of this paper, the residual for a node is the difference between the actual BWC value and the BWC value predicted based on the best fitting straight line). The goodness of the regression is evaluated with respect to the Estimated Standard Error of the Residuals (Standard deviation of the residuals), Coefficient of Determination (R-Squared metric), Pearson's correlation coefficient (square root of the R-Squared value) and the goodness fraction (fraction of the nodes for which the predicted BWC value is within the threshold error of the actual value). The R-Squared (ranges from 0...1) metric indicates the proportion of variation in the dependent variable that is predictable from the independent variable. As stated earlier, the primary objective of the OLS method of linear regression is to fit a straight line that would lower the Estimated Standard Error of the Residuals to as low as possible. The strength of the Standard Error of the Residuals metric is that it could be computed using the same unit as that of the dependent variable and it would be a relatively an apt metric to interpret the goodness of the fit. Moreover, as we use normalized values of the dependent and independent variables for the centrality metrics to fit the regression line, the Standard Error of the Residuals (SER) could be directly construe the range of the predicted values vis-a-vis the actual values. For example, if the Standard Error of the Residuals is 0.05, we could say that in a scale of 0....1 for the actual values of the BWC metric, the predicted values for the BWC metric are more likely to be ± 0.05 of the actual values. We also calculate the fraction of nodes (called the goodness fraction) for which the predicted BWC values are indeed ± SER of the actual BWC values. All the 47 real-world network graphs analyzed in this section are modeled as undirected graphs. Table 1 lists the networks (and the acronym code) along with the number of nodes and edges, average degree (k_{avg}) and spectral radius ratio for node degree (λ_{sp}) [10]: a measure of variation in the degree of the nodes; the farther is this value from 1.0, the larger is the variation in node degree. A more detailed description of the networks can be found in [22]. The regression analysis is conducted using the Java Apache Commons Math package [31].

Table 1: Properties of the Real-World Network Graphs used for Regression Analysis

#	Network Name	Code	# Nodes	# Edges	k_{avg}	λ_{sp}
1	Word Adjacency Net.	ADJ	112	425	7.589	1.73
2	Anna Karnenina Net.	AKN	140	494	7.057	2.48
3	Band Jazz Net.	JBN	198	2742	27.697	1.45
4	C. Elegans Neural Net.	CEN	297	2148	14.465	1.68
5	Centrality Literature Net.	CLN	118	613	10.39	2.03
6	Citation Graph Drawing Net.	CGD	259	640	4.942	2.24
7	Copperfield Net.	CFN	89	407	9.146	1.83
8	Dolphin Net.	DON	62	159	5.129	1.40
9	Drug Net.	DRN	212	284	2.679	2.76
10	Dutch Literature 1976 Net.	DLN	37	81	4.378	1.49
11	Erdos Collaboration Net.	ERD	433	1314	6.069	3.00
12	High School Friendship Net.	FMH	147	202	2.748	2.81
13	Hi-Tech Firm Friendship Net.	FHT	33	91	5.515	1.57
14	Flying Teams Cadet Net.	FTC	48	170	7.083	1.21
15	US Football Net.	FON	115	613	10.661	1.01
16	College Dorm Fraternity Net.	CDF	58	967	33.345	1.11
17	Graph Drawing 1996 Net.	GD96	180	228	2.533	2.38
18	Gleiser Marvel Universe Net.	MUN	167	301	3.605	2.54
19	Graph and Digraph Glossary Net.	GD01	101	190	3.762	1.80
20	Hypertext 2009 Net.	HTN	115	2164	37.635	1.21
21	Huckleberry Coappearance Net.	HCN	76	302	7.947	1.66
22	Infectious Socio-patterns Net.	ISP	309	1924	12.453	1.69
23	Karate Club Net.	KCN	34	78	4.588	1.47
24	Korea Family Planning Net.	KFP	37	85	4.595	1.70
25	Les Miserables Net.	LMN	77	254	6.597	1.82
26	Macaque Dominance Net.	MDN	62	1167	37.645	1.04
27	Madrid Train Bombing Net.	MTB	64	295	9.219	1.95
28	Manufac. Company Employee Net.	MCE	77	1549	40.23	1.12
29	Social Networks Journal Co-Authors	MSJ	475	625	2.632	3.48
30	Author Facebook Net.	AFB	171	940	10.994	2.29
31	Mexican Political Elite Net.	MPN	35	117	6.686	1.23
32	Modern Math Method Net.	MMN	30	61	4.067	1.59
33	US Politics Books Net.	PBN	105	441	8.4	1.42
34	Primary School Contact Net.	PSN	238	5539	46.546	1.22
35	Prison Friendship Net.	PFN	67	142	4.239	1.32
36	San Juan Sur Family Net.	SJN	75	155	4.133	1.29
37	Scotland Corporate Interlocks Net.	SDI	230	359	3.122	1.94
38	Senator Press Release Net.	SPR	92	477	10.37	1.57
39	Soccer World Cup 1998 Net.	SWC	35	118	6.743	1.45
40	Sawmill Strike Communication Net.	SSM	24	38	3.167	1.22
41	Taro Exchange Net.	TEN	22	39	3.545	1.06
42	Teenage Female Friendship Net.	TWF	47	77	3.277	1.49
43	UK Faculty Friendship Net.	UKF	83	578	13.928	1.35
44	US Airports 1997 Net.	APN	332	2126	12.807	3.22
45	Residence Hall Friendship Net.	RHF	217	1839	16.949	1.27
46	Windsurfers Beach Net.	WSB	43	336	15.628	1.22
47	World Trade Metal Net.	WTN	80	875	21.875	1.38

Table 2: Regression Analysis Results for {LCC'DC vs. DEG} and BWC: R-Squared and estimated Standard Error of Residuals (SER) as well as Computation Time of Centrality Metrics

#	Net. Code	R-Squared Metric		Std. Error Residuals (SER)		Computation Time (milli sec.)		
		LCC'DC	DEG	LCC'DC	DEG	BWC	LCC'DC	DEG
1	ADJ	0.865	0.838	0.032	0.035	261.67	0.840	0.016
2	AKN	0.899	0.796	0.026	0.038	378.56	1.287	0.021
3	JBN	0.574	0.372	0.043	0.052	2006.85	9.336	0.039
4	CEN	0.666	0.609	0.033	0.035	5837.20	3.576	0.047
5	CLN	0.787	0.681	0.038	0.047	194.83	0.183	0.019
6	CGD	0.739	0.636	0.025	0.029	5123.47	0.258	0.049
7	CFN	0.805	0.652	0.047	0.063	42.96	0.113	0.017
8	DON	0.503	0.357	0.071	0.081	19.28	0.038	0.010
9	DRN	0.485	0.422	0.038	0.041	3944.56	0.171	0.046
10	DLN	0.716	0.626	0.066	0.076	3.16	0.022	0.005
11	ERD	0.690	0.611	0.023	0.026	21956.46	0.507	0.075
12	FMH	0.515	0.397	0.045	0.051	940.77	0.132	0.032
13	FHT	0.810	0.666	0.059	0.079	2.80	0.021	0.005
14	FTC	0.834	0.613	0.043	0.066	8.67	0.034	0.007
15	FON	0.453	0.079	0.028	0.037	148.68	0.165	0.016
16	CDF	0.874	0.734	0.029	0.042	24.68	0.276	0.018
17	GD96	0.887	0.905	0.024	0.022	870.43	0.131	0.025
18	MUN	0.741	0.495	0.039	0.054	274.62	0.122	0.025
19	GD01	0.891	0.868	0.035	0.038	21.18	0.039	0.011
20	HTN	0.781	0.687	0.036	0.043	211.62	0.745	0.017
21	HCN	0.881	0.687	0.040	0.064	23.23	0.073	0.011
22	ISP	0.259	0.220	0.044	0.045	6413.78	0.458	0.047
23	KCN	0.866	0.843	0.058	0.063	2.35	0.020	0.005
24	KFP	0.498	0.218	0.090	0.113	3.41	0.024	0.006
25	LMN	0.867	0.558	0.040	0.073	30.03	0.060	0.011
26	MDN	0.963	0.875	0.011	0.020	33.81	0.379	0.009
27	MTB	0.764	0.531	0.053	0.074	22.31	0.051	0.010
28	MCE	0.887	0.784	0.034	0.047	55.96	0.449	0.012
29	MSJ	0.372	0.154	0.036	0.041	13454.64	0.345	0.068
30	AFB	0.295	0.067	0.059	0.068	705.27	0.219	0.038
31	MPN	0.885	0.795	0.049	0.065	2.93	0.021	0.005
32	MMN	0.788	0.709	0.060	0.070	2.47	0.019	0.014
33	PBN	0.606	0.507	0.050	0.056	107.30	0.117	0.016
34	PSN	0.780	0.702	0.021	0.025	3426.78	2.560	0.032
35	PFN	0.778	0.721	0.044	0.049	21.72	0.034	0.010
36	SJN	0.742	0.660	0.047	0.054	32.46	0.049	0.017
37	SDI	0.532	0.543	0.037	0.036	2489.91	0.162	0.036
38	SPR	0.775	0.697	0.041	0.048	69.01	0.117	0.014
39	SWC	0.859	0.820	0.054	0.061	3.05	0.022	0.005
40	SSM	0.717	0.723	0.099	0.098	0.78	0.012	0.005
41	TEN	0.888	0.738	0.048	0.073	0.67	0.010	0.003
42	TWF	0.484	0.047	0.081	0.111	3.58	0.021	0.007
43	UKF	0.824	0.611	0.042	0.062	51.65	0.161	0.011
44	APN	0.681	0.496	0.030	0.038	7075.86	0.942	0.047
45	RHF	0.816	0.707	0.023	0.029	2025.47	0.455	0.037
46	WSB	0.899	0.802	0.038	0.054	7.44	0.057	0.006
47	WTN	0.890	0.824	0.035	0.044	54.09	0.251	0.012

Table 2 presents the R-Squared values and the estimated Standard Error of the Residuals (SER) as well as the running time (in milliseconds) of the methods to determine the BWC, LCC'DC and DEG. We have highlighted the cells for which the R-Squared values are high as well as the cells for which we incur a lower SER value (both of which indicating the corresponding centrality metric has a better goodness of fit). The execution time of the LCC'DC method is significantly smaller than that of the BWC method; nevertheless the DEG metric incurs the smallest of the execution time.

We will follow the convention proposed by Evans [29] to assess the strength of the correlation between two datasets based on the values for the correlation coefficient. As the R-Squared metric is simply the square of the Pearson's correlation coefficient (in the case of the OLS linear regression with intercept being considered as part of the model), we propose to adopt the square of the boundary values for each of the ranges proposed for the correlation coefficient as the ranges for the level of association with respect to the R-Squared metric. Accordingly, the range of values for the strength of the correlation (with respect to the Pearson's correlation coefficient) and the strength of the association (with respect to the R-Squared metric) are shown in Table 3. Note that the R-Squared metric for OLS linear regression with intercept model always takes the values between 0...1 (as it is simply the square of the Pearson's correlation coefficient whose values range from -1....1).

Table 3: Range of Correlation Coefficient Values for the Level of Correlation and the Range of R-Squared Values for the Level of Association

Range of Correlation Coefficient	**Level of Correlation**		**Range of Correlation Coefficient**	**Level of Correlation**		**Range of R-Squared Values**	**Level of Association**
0.80 to 1.00	Very Strong Positive		-1.00 to -0.80	Very Strong Negative		0.64 to 1.00	Very Strong
0.60 to 0.79	Strong Positive		-0.79 to -0.60	Strong Negative		0.36 to 0.63	Strong
0.40 to 0.59	Moderate Positive		-0.59 to -0.40	Moderate Negative		0.16 to 0.35	Moderate
0.20 to 0.39	Weak Positive		-0.39 to -0.20	Weak Negative		0.04 to 0.15	Weak
0.00 to 0.19	Very Weak Positive		-0.19 to -0.01	Very Weak Negative		0.00 to 0.03	Very Weak

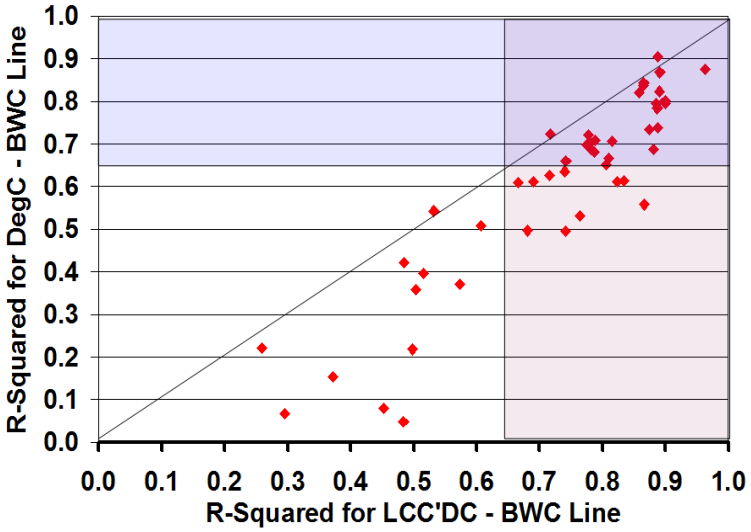


Figure 7: Distribution of the R-Squared Values for the LCC'DC - BWC OLS Regression Line vs. the DEG - BWC OLS Regression Line

Figure 7 illustrates the comparison of the R-Squared values for the LCC'DC-BWC OLS regression line vs. the DEG-BWC OLS regression line. For 44 of the 47 real-world networks, we observe the R-Squared values for the LCC'DC-BWC OLS regression to be larger than that of the DEG-BWC OLS regression. For the LCC'DC-BWC OLS regression, we observe about 35 of the 47 real-world networks to exhibit a R-Squared value of 0.64 or larger. In other words, about 35/47 ~ 75% of the real-world networks have been observed to exhibit a very strong association between LCC'DC and BWC. On the other hand, for the DEG-BWC OLS regression: we observe only about 25 of the 47 real-world networks exhibit a R-Squared value of 0.64 or larger. That is, about 25/47 ~ only 53% of the real-world networks exhibit a very strong association between DEG and BWC. With respect to R-Squared values less than or equal to 0.15: we observe 4/47 ~ 9% of the real-world networks to exhibit a weak to very weak association between DEG and BWC, whereas none of the real-world networks exhibit a weak to very weak association between LCC'DC and BWC. The minimum, maximum and median of the R-Squared values observed for the LCC'DC-BWC regression line fit are respectively 0.26, 0.96 and 0.78 (indicating about half of the real-world networks analyzed exhibit an R-Squared value of 0.78 or above). On the other hand, the minimum, maximum and median of the R-Squared values observed for the DEG-BWC regression are respectively 0.05, 0.90 and 0.66.

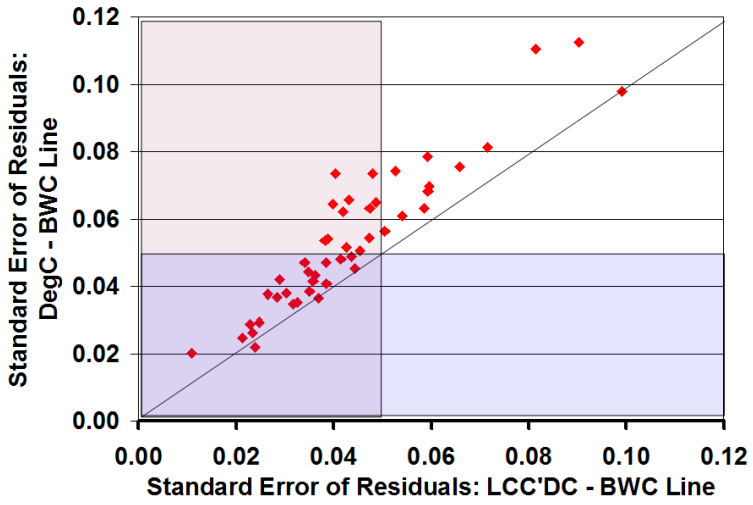


Figure 8: Distribution of the Standard Error of the Residuals (SER) Values for the LCC'DC - BWC OLS Regression Line vs. the DEG - BWC OLS Regression Line

Figure 8 illustrates the comparison of the estimated Standard Error of the Residuals (SER) values for the LCC'DC-BWC OLS regression line vs. the DEG-BWC OLS regression line. Again, for 44 of the 47 real-world networks (the same networks that we observed the R-Squared values to be larger), we observe the SER values for the LCC'DC-BWC OLS regression to be lower than that of the DEG-BWC OLS regression. Like the case of the R-Squared metric, for the LCC'DC-BWC OLS regression, we are more likely to be able to predict a BWC value that is ± 0.05 of the actual BWC values of the nodes (in the range of 0...1) for about 35/47 ~ 75% of the real-world networks. On the other hand, for the DEG-BWC OLS regression: we are more likely to be able to predict a BWC value that is ± 0.05 of the actual BWC values of the nodes for only about 23/47 ~ less than 50% of the real-world networks. The minimum, maximum and median of the SER values observed for the LCC'DC-BWC regression line fit are respectively 0.011, 0.099 and 0.04 (indicating about half of the real-world networks analyzed exhibit an SER value of 0.04 or lower). On the other hand, the minimum, maximum and median of the SER values observed for the DEG-BWC regression line fit are respectively 0.02, 0.113 and 0.051.

Table 4: Regression Analysis Results for {LCC'DC vs. DEG} and BWC: Pearson's Correlation Coefficient, Goodness Fraction of Nodes as well as Parameters for LCC'DC-BWC Regression

#	Net. Code	Pearson's Correlation Coefficient		Goodness Fraction of Nodes		Parameters for LCC'DC - BWC Regression	
		LCC'DC	DEG	LCC'DC	DEG	Intercept: b_0	Slope: b_1
1	ADJ	0.930	0.915	0.705	0.670	-0.040	1.196
2	AKN	0.948	0.892	0.848	0.826	-0.007	0.977
3	JBN	0.757	0.610	0.874	0.904	-0.020	0.980
4	CEN	0.816	0.780	0.936	0.949	-0.027	1.067
5	CLN	0.887	0.825	0.829	0.837	-0.022	1.025
6	CGD	0.860	0.797	0.794	0.791	-0.006	0.970
7	CFN	0.897	0.808	0.851	0.862	-0.022	0.957
8	DON	0.709	0.598	0.871	0.823	-0.024	0.984
9	DRN	0.696	0.649	0.913	0.899	-0.003	0.781

10	DLN	0.846	0.791	0.714	0.743	-0.025	1.041
11	ERD	0.831	0.782	0.869	0.860	-0.005	0.919
12	FMH	0.718	0.630	0.815	0.859	-0.004	0.789
13	FHT	0.900	0.816	0.694	0.778	-0.017	1.009
14	FTC	0.913	0.783	0.583	0.625	-0.054	1.255
15	FON	0.673	0.282	0.687	0.713	-0.066	1.640
16	CDF	0.935	0.857	0.724	0.741	-0.047	1.285
17	GD96	0.942	0.951	0.900	0.900	-0.013	1.020
18	MUN	0.861	0.704	0.879	0.867	-0.006	0.893
19	GD01	0.944	0.932	0.750	0.736	-0.019	1.042
20	HTN	0.884	0.829	0.947	0.947	-0.059	1.409
21	HCN	0.938	0.829	0.878	0.824	-0.015	0.977
22	ISP	0.509	0.469	0.913	0.922	-0.009	0.745
23	KCN	0.930	0.918	0.824	0.853	-0.028	1.030
24	KFP	0.705	0.467	0.795	0.692	-0.004	0.854
25	LMN	0.931	0.747	0.857	0.857	-0.016	0.990
26	MDN	0.982	0.935	0.774	0.710	-0.094	1.693
27	MTB	0.874	0.729	0.871	0.729	-0.003	0.912
28	MCE	0.942	0.885	0.727	0.740	-0.045	1.207
29	MSJ	0.610	0.392	0.931	0.926	-0.001	0.636
30	AFB	0.543	0.259	0.914	0.957	-0.013	0.695
31	MPN	0.941	0.892	0.714	0.771	-0.086	1.327
32	MMN	0.888	0.842	0.737	0.684	-0.050	1.167
33	PBN	0.779	0.712	0.800	0.771	-0.011	0.935
34	PSN	0.883	0.838	0.798	0.782	-0.016	1.141
35	PFN	0.882	0.849	0.806	0.731	-0.027	1.103
36	SJN	0.861	0.812	0.813	0.787	-0.045	1.206
37	SDI	0.730	0.737	0.852	0.848	-0.018	1.023
38	SPR	0.880	0.835	0.783	0.793	-0.038	1.179
39	SWC	0.927	0.905	0.686	0.800	-0.058	1.186
40	SSM	0.847	0.851	0.667	0.792	-0.070	1.112
41	TEN	0.942	0.859	0.682	0.727	-0.097	1.364
42	TWF	0.696	0.218	0.680	0.800	-0.001	0.830
43	UKF	0.908	0.782	0.778	0.815	-0.052	1.229
44	APN	0.825	0.705	0.901	0.883	-0.004	0.861
45	RHF	0.903	0.841	0.783	0.760	-0.037	1.373
46	WSB	0.948	0.895	0.767	0.767	-0.059	1.258
47	WTN	0.944	0.908	0.800	0.763	-0.022	1.044

Table 4 displays values for the Pearson's correlation coefficient (square root of the R-Squared values) and the Goodness Fraction (the fraction of nodes for which the predicted BWC value is within ± SER of the actual value). We also display the values for the regression parameters b_0 and b_1 obtained for each of the real-world networks with respect to the LCC'DC-BWC regression. We decided to use the positive values for the Pearson's correlation coefficient (square root of R-Squared) because the slope (parameter b_1) of the regression line is positive. We use the range of values proposed by Evans [29] for the levels of correlation (shown in Table 3) to determine the percentage of networks for which we observe a very strong correlation between the BWC and LCC'DC metrics vis-a-vis DEG.

Figure 9 displays the distribution of the correlation coefficient values for LCC'DC vs. DEG. As in the case of the R-Squared metric, we observe about 75% (35 out of 47) of the real-world networks

exhibited a very strong correlation between LCC'DC and BWC; whereas, only about 53% (25 out of 47) of the real-world networks exhibited a very strong correlation between DEG and BWC. For 25 of the 47 of the real-world networks, both the LCC'DC and DEG metrics exhibited a very strong correlation with BWC (basically for all the 25 real-world networks for which the DEG exhibited a very strong correlation). For 12 of the 47 real-world networks, both the LCC'DC and DEG do not exhibit a very strong correlation (basically for all the 12 real-world networks for which the LCC'DC did not exhibit a very strong correlation). Overall, there is no real-world network for which the DEG metric exhibited a very strong correlation and the LCC'DC metric did not exhibit a very strong correlation. Thus, the LCC'DC metric is a very promising metric to be considered an alternative measure of the BWC metric for complex real-world networks.

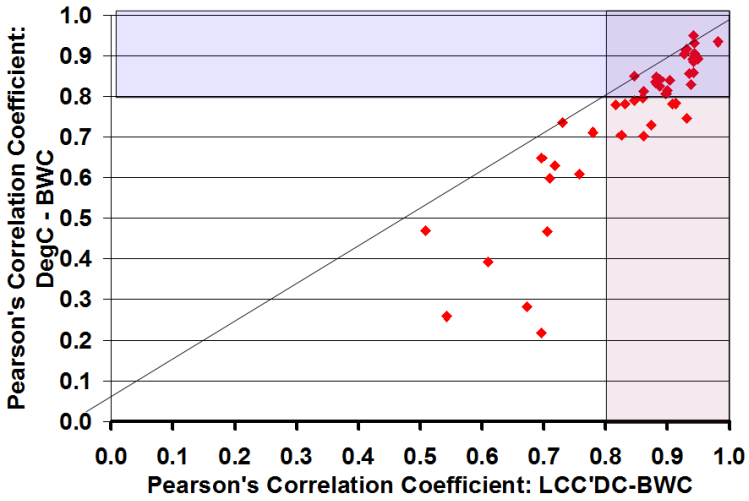


Figure 9: Distribution of the Pearson's Correlation Coefficient Values for the LCC'DC - BWC OLS Regression Line vs. the DEG - BWC OLS Regression Line

Figure 10 displays the distribution of values for the goodness fraction of the nodes for which the predicted BWC (based on the regression line) differs from that of the actual BWC by at most the SER value for the regression. The goodness fraction (ranges from 0...1) for both the regressions (LCC'DC and DEG-based) were observed to be very close to each other (as observed from most of the data points in Figure 10 lying close to the diagonal line), with each regression yielding a relatively slightly larger value for the goodness fraction for about 50% of the real-world networks. The median value for the goodness fraction of the nodes with both the regressions is close to 0.80; thus, for at least 50% of the real-world networks: for about 80% of the nodes, we are able to predict a BWC value that differs from the actual BWC value by at most the SER value.

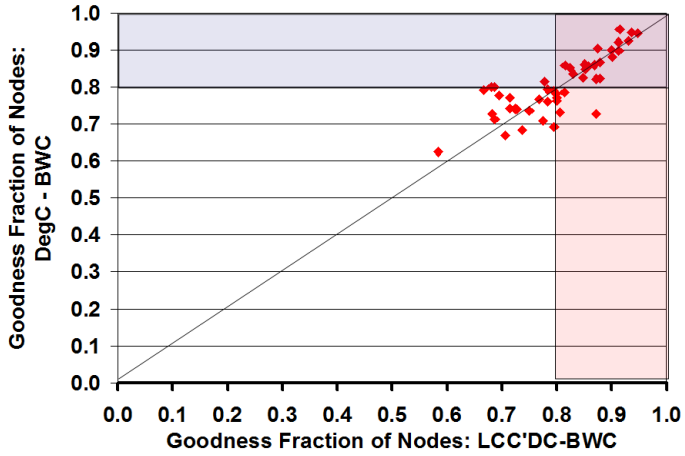


Figure 10: Distribution of the Goodness Fraction of Nodes for the LCC'DC - BWC OLS Regression Line vs. the DEG - BWC OLS Regression Line

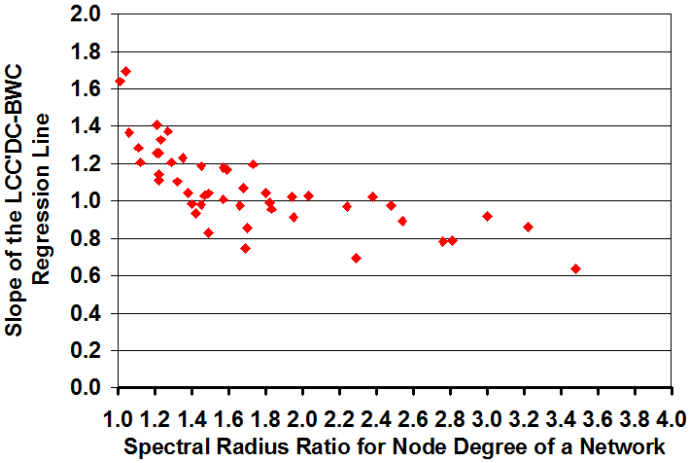


Figure 11: Spectral Radius Ratio for Node Degree for a Network vs. Slope of the OLS LCC'DC-BWC Regression Line for the Network

Figures 11 and 12 respectively illustrate the distribution of the spectral radius ratio for node degree (λ_{sp}) vs. the slope and intercept of the LCC'DC-BWC OLS regression line. We observe the slope and intercept of the regression line to respectively decrease and increase with increase in λ_{sp}. We confirm this through a more detailed look at the distribution of the actual normalized LCC'DC and BWC values for two sample networks: the MSJ network (McCarty Social Network Journal Authors) with λ_{sp} = 3.48 has a slope of 0.6362 and intercept of -0.0009 (see Figure 12) and the FON network (US Football Network) with λ_{sp} = 1.01 (see Figure 13) has a slope of 1.6397 and intercept of -0.0656. The intercept of the regression lines for all the 47 real-world networks is negative and the intercept approaches zero for networks with larger spectral radius ratio for node degree. A larger positive slope of the regression lines for networks with lower λ_{sp} values indicates that the BWC values are predicted to quickly increase with increase in the LCC'DC values of the vertices in such networks, compared to networks with higher λ_{sp} values.

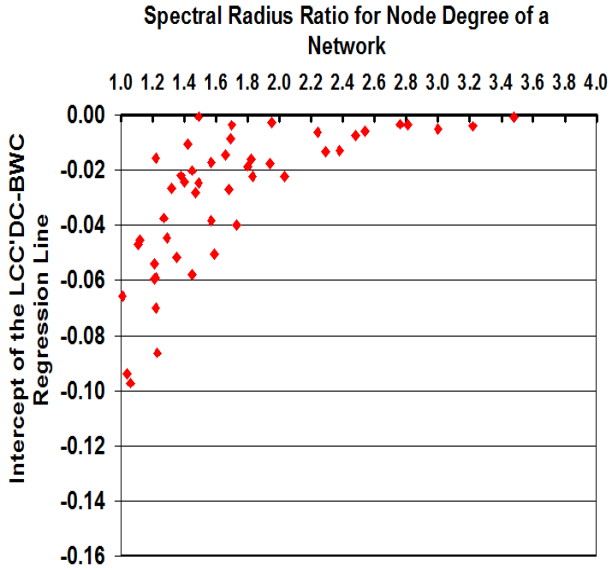


Figure 12: Spectral Radius Ratio for Node Degree for a Network vs. Intercept of the OLS LCC'DC-BWC Regression Line for the Network

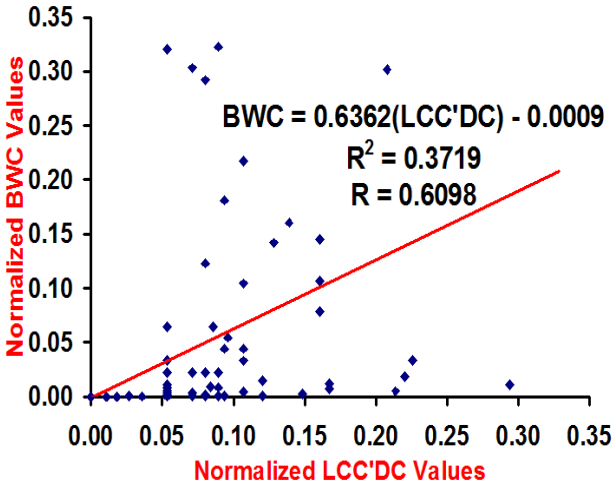


Figure 13: Distribution of the Normalized LCC'DC and BWC Values as well as the OLS LCC'DC-BWC Regression Line for the McCarty Social Journal Authors Network (λ_{sp} = 3.48)

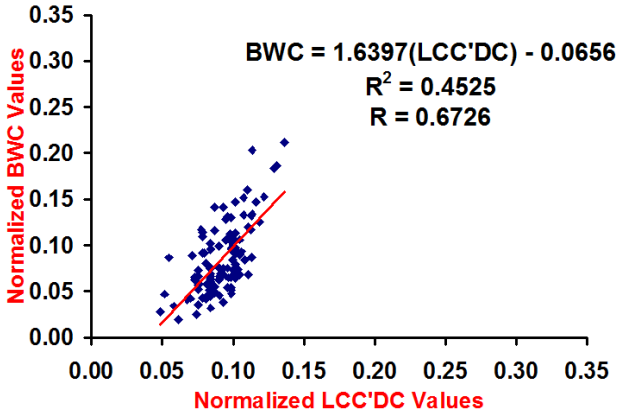


Figure 13: Distribution of the Normalized LCC'DC and BWC Values as well as the OLS LCC'DC-BWC Regression Line for the US Football Network (λ_{sp} = 1.01)

4. RELATED WORK

In [25], the authors evaluated the range-limited centrality measures of vertices and edges in complex networks: i.e., the centrality of the vertices based on their one-hop neighborhood, two-hop neighborhood and etc. It has been observed for both unweighted and weighted network graphs: within smaller shortest path lengths, the ranking of the vertices based on the range-limited centrality measures becomes similar to the ranking of the vertices computed for the entire network. This results corroborates our finding in this paper regarding the correlation observed between LCC'DC and BWC. The LCC'DC metric studied in this paper is essentially computed based on the one-hop neighborhood of the vertices and it captures the fraction of the pairs of neighbors of the vertex that need to go through the vertex for shortest path communication and scaled to the degree of the vertex.

In [26], the authors proposed the use of virtual nodes between weighted edges to transform a weighted graph of V vertices and E edges to an unweighted graph of unit edge weights so that one could run the O(V+E) breadth first search (BFS) algorithm [27] instead of the O(Elog$V$) Dijkstra algorithm [27] to determine the fraction of shortest paths (between any two vertices) going through a vertex. For example, if an edge x-y has weight 3, we introduce two virtual nodes v_1, v_2 such that we replace the edge x-y of weight 3 with a sequence of edges x- v_1- v_2-y each of weight 1. However, this transformation is scalable and time-saving only for light-weighted sparse network graphs (in [26]: the threshold edge weight has been shown to be log$N$/D + 1 where N and D are respectively the number of nodes and average degree of the nodes) and the number of virtual nodes introduced for weighted edges would become an overkill even for sparse graphs with larger edge weights or for dense graphs with even smaller edge weights.

There are a few approaches/algorithms proposed to further develop the application of BWC. For instance, the random-walk betweenness measure calculated for all vertices in a network in worst-case time O($(m+n)n^2$) using matrix methods [8]. Others such as bounded-distance betweenness [9], distance-scaled betweenness [9], edge betweenness [11] and group betweenness [12] are also introduced. In [18], the authors used a metric called the "complex degree centrality" [20] for a co-author network and observed this centrality to exhibit a correlation coefficient of 0.848 with the node betweenness metric. The complex degree centrality of a vertex (author) is computed as an entropy-based measure of the number of co-authors and the number of co-authored papers for the

author. Nevertheless, the computation cost of these betweenness measures is still high. It is more feasible if we could find another centrality measure with low computation cost that is highly correlated to BWC. It was shown that the BWC is related to the degree in social networks [13] and scale-free network [14]. However, there still lacks substantial support on the alternatives to BWC.

BWC measures the interrelationships among vertices. The results of our simulation studies suggest that BWC is highly correlated to LCC'DC on most tested real-world networks with a correlation coefficient of 0.8 or above. Leydesdorff (2007) [3] also observed high correlation between BWC and DEG with a correlation coefficient value of 0.724 on Journal Citing Social Networks. Recently, Pozzi et al (2013) [16] observed a strong correlation of the centrality indices between unweighted BWC and DEG calculated on Planar Maximally Filtered Graphs (PMFG) with a value of 0.97. There is also a moderate correlation between BWC and CLC papered with a value of 0.54 [3]. CLC refers to the relatedness among a set of vertices, providing a global measure of relationships among all vertices. A good correlation between BWC and CLC is valuable when it comes to a connection between global and local view.

There has been some works that also highlight independence of betweenness centrality from node degree. In [21], the authors observed that the scale-free social networks that have been observed to be assortative with respect to node degree (i.e., high-degree vertices are connected to high-degree vertices and vice-versa) are not similarly observed to be assortative with respect to node betweenness (i.e., a node with high betweenness centrality is more likely to be connected with nodes of arbitrary betweenness centrality and need not be neighbors that also have high betweenness centrality). A similar observation has also been made in a recent study of the lead author in [22] to assess the assortativity of real-world networks with respect to different centrality metrics. The assortative index (correlation coefficient between vertices with respect to a particular centrality metric) with respect to betweenness centrality has been observed to be less than 0.5 for a majority of the real-world networks. In addition, in [23]: the correlation between degree centrality and betweenness centrality for fractal scale-free networks (both synthetic as well as real-world networks) has been observed to be weaker compared to the correlation observed for non-fractal scale-free networks. Fractal networks have self-similar network motifs at different scales (i.e., motifs of different sizes) [24].

In [17], the authors evaluated the impact of percentage of samples (like the fraction of nodes among the nodes in a network) that need to be used to construct a regression model to predict each of the commonly used centrality metrics. The regression model for a centrality metric is constructed based on the sample nodes randomly chosen and the values predicted from this model for the non-sampled nodes were compared with the actual value. Though several centrality metrics showed stable behavior (i.e., the metric values were predicted with appreciable accuracy with sampling), the Betweenness Centrality metric was observed to be unstable. That is, it was not possible to predict the BWC of the non-sampled vertices using the BWC of the sampled vertices. In another sampling related work, Bader et al [7] propose an adaptive sampling technique to pick a sequence of source-destination pairs (the subsequent pair in the sequence is decided based on the information gathered by running the single source shortest path problem for the pairs chosen until then) rather than considering all pairs of vertices or random pairs of vertices [19]. An adaptive sampling of the source-destination pairs of vertices has been observed [7] to produce a relatively better approximation to betweenness centrality compared to random sampling.

5. CONCLUSIONS

The high-level contributions of this paper are that we have shown the computationally heavy betweenness centrality metric to be highly correlated with the degree centrality metric for synthetic networks generated from the theoretical random network and small-world network models and with that of the localized clustering coefficient complement-based degree centrality metric (LCC'DC) for real-world networks. As part of the regression analysis conducted for real-world networks, we observe the LCC'DC metric to incur relatively larger R-Squared values and at the same time relatively smaller SER values (Standard error for residuals) compared to those for the degree centrality metric. We also propose an evaluation metric called the *goodness fraction* to estimate the fraction of nodes for which we are able to predict the BWC values that are within ± the SER values observed for the real-world networks. We find both the LCC'DC and DEG-based regression models to incur a larger value for the goodness fraction of nodes. For future work, we plan to identify computationally light metrics for other computationally heavy metrics like eigenvector centrality, closeness centrality and maximal clique size using correlation analysis and corroborate the identified correlations using regression analysis.

REFERENCES

[1] L. C. Freeman, "A Set of Measures of Centrality based on Betweenness," Sociometry, vol. 40, no. 1, pp. 35-41, March 1977.

[2] J. M. Anthonisse, "The Rush in a Directed Graph," Stichting Mathematisch Centrum. Mathemaatische Besliskunde (BN 9/71), pp. 1-10, 1971.

[3] L. Leydesdorff, "Betweenness Centrality as an Indicator of the Interdisciplinarity of Scientific Journals," Journal of the American Society for Information Science and Technology, vol. 58, no. 9, pp. 1303-1319, 2007.

[4] M. Barthelemy, "Betweenness Centrality in Large Complex Networks," European Physical Journal B, vol. 38, pp. 163-168, 2004.

[5] U. Brandes, "A Faster Algorithm for Betweenness Centrality," The Journal of Mathematical Sociology, vol. 25, no. 2, pp. 163-177, 2001.

[6] M. E. J. Newman, "Scientific Collaboration Networks. II. Shortest Paths, Weighted Networks, and Centrality," Physical Review E, vol. 64, 016132, June 2001.

[7] D. Bader, S. Kintali, K. Madduri and M. Mihail, "Approximating Betweenness Centrality," Algorithms and Models for the Web-Graph, Lecture Notes in Computer Science, vol. 4863, pp. 124-137, 2007.

[8] M. E. J. Newman, "A Measure of Betweenness Centrality based on Random Walks," Social Networks, vol. 27, pp. 39-54, 2005.

[9] S. P. Borgatti and M. G. Everett, "A Graph-Theoretic Perspective on Centrality," Social Networks, vol. 28, no. 4, pp. 466-484, October 2006.

[10] N. Meghanathan, "Spectral Radius as a Measure of Variation in Node Degree for Complex Network Graphs," Proceedings of the 3rd International Conference on Digital Contents and Applications, (DCA 2014), pp. 30-33, Hainan, China, December 20-23, 2014.

[11] M. E. J. Newman and M. Girvan, "Finding and Evaluating Community Structure in Networks," Physical Review E, vol. 69, 026113, February 2004.

[12] M. G. Everett and S. P. Borgatti, "The Centrality of Groups and Classes," The Journal of Mathematical Sociology, vol. 23, no. 9, pp. 181-201, 1999.

[13] K-I. Goh, E. Oh, B. Kahng and D. Kim, "Betweenness Centrality Correlation in Social Networks," Physical Review E, vol. 67, 017101, January 2003.

[14] K-I. Goh, B. Kahng and D. Kim, "Universal Behavior of Load Distribution in Scale-Free Networks," Physical Review Letters, vol. 87, 278701, December 2001.

[15] N. Meghanathan, "Distribution of Maximal Clique Size under the Watts-Strogatz Model of Evolution of Complex Networks," International Journal in Foundations of Computer Science and Technology, vol. 5, no. 3, pp. 1-12, May 2015.
[16] F. Pozzi, T. D. Matteo and T. Aste, "Spread of Risk across Financial Markets: Better to Invest in te Peripheries," Scientific Reports, vol. 3, no. 1665, 2013.
[17] E. Costenbader and T. W. Valente, "The Stability of Centrality Measures when Networks are Sampled," Social Networks, vol. 25, pp. 283-307, 2003.
[18] Y. Kwon, D. Jeong, Y. Moon and J. Yoo, "Comparing Analysis Study of Centrality Indices using Paper Information on Secondary Battery," Indian Journal of Science and Technology, vol. 8 (S1), pp. 333-339, January 2015.
[19] U. Brandes and C. Pich, "Centrality Estimation in Large Networks," International Journal of Bifurcation and Chaos, vol. 17, no. 7, pp. 2303-2318, July 2007.
[20] J-Y. Lee, "A Study on Document Citation Indicators based on Citation Network Analysis," Journal of the Korean Society for Library and Information Science, vol. 45, no. 2, pp. 119-143, 2011.
[21] K. I. Goh, E. Oh, B. Kahng and D. Kim, "Betweenness Centrality Correlation in Social Networks," Physical Review E Stat. Nonlin. Soft Matter Phys., vol. 67 (1 Pt 2): 017101, January 2003.
[22] N. Meghanathan, "Assortativity Analysis of Real-World Network Graphs based on Centrality Metrics," Computer and Information Science, vol. 9, no. 3, pp. 7-25, August 2016.
[23] M. Kitsak, S. Havlin, G. Paul, M. Riccaboni, F. Pammolli and H. E. Stanley, "Betweenness Centrality of Fractal and Nonfractal Scale-Free Model Networks and Tests on Real Networks," Physical Review E Stat Nonlin. Soft Matter Phys., vol. 75 (5 Pt 2): 056115, Epub, May 31, 2007.
[24] C. Song, S. Havlin and H. A. Makse, "Self-Similarity of Complex Networks," Nature, vol. 433, no. 7024, pp. 392-395, January 2005.
[25] M. Ercsey-Ravasz, R. N. Lictenwalter, N. V. Chawla and Z. Toroczkai, "Range-limited Centrality Measures in Complex Networks," Phys. Rev. E. Stat Nonlin. Soft Matter Phys., vol. 85 (6 Pt 2): 066103, Epub, Jun 6, 2012.
[26] J. Yang and Y. Chen, "Fast Computing Betweenness Centrality with Virtual Nodes on Large Sparse Networks," PLoS One, vol. 6, no. 7, e22557, 2011.
[27] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, Introduction to Algorithms, 3rd Edition, MIT Press, 2009.
[28] N. Meghanathan, "A Computationally Lightweight and Localized Centrality Metric in lieu of Betweenness Centrality for Complex Network Analysis," Springer Vietnam Journal of Computer science, pp. 1-16, First Online: June 2016. DOI: 10.1007/s40595-016-0073-1.
[29] J. D. Evans, Straightforward Statistics for the Behavioral Sciences, 1st Edition, Brooks Cole Publishing Company, August 1995.
[30] F. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd Edition, Springer Series in Statistics, August 2015.
[31] Java Apache Commons Math API: http://commons.apache.org/proper/commons-math/, last accessed: November 26, 2016.