Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Social Networks 25 (2003) 283–307

SOCIAL NETWORKS

www.elsevier.com/locate/socnet

The stability of centrality measures when networks are sampled

Elizabeth Costenbader[a,∗], Thomas W. Valente[b,1]

[a] *Department of Population and Family Health Sciences, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD 21205, USA*

[b] *Department of Preventive Medicine, University of Southern California School of Medicine, 1000 Fremont Avenue, Building A Room 5133, Alhambra, CA 91803, USA*

Abstract

The ability to measure centrality in social networks has been a particularly useful development in social network analysis. For researchers trying to decide which centrality measure is most meaningful and valid for their research purposes, various papers have explored the conceptual foundations of centrality measures. Less well documented is the empirical performance of centrality measures under different research scenarios or constraints. This study uses bootstrap sampling procedures to determine how sampling affects the stability of 11 different network centrality measures. Results indicate that some measures are more stable than others, and that stability is also a function of network and study properties.

Keywords: Social networks; Sampling; Centrality; Simulation; Surveys

1. Introduction

The ability to measure centrality in social networks has been a particularly useful development in social network analysis. Measures of centrality describe actors' positions in a network relative to others and in relation to the complete network. Several centrality measures have been created to measure which individuals in a network possess influential and prestigious roles (Freeman, 1979; Bonacich, 1972, 1987; Scott, 2000; Wasserman and Faust, 1994). Additional measures have been developed to indicate the social influence of an individual on the other individuals in his or her network (Friedkin, 1991). Most recently,

[∗] Corresponding author. Tel.: +1-410-262-1789.

E-mail addresses: ecostenb@jhsph.edu (E. Costenbader), tvalente@usc.edu (T.W. Valente).

[1] Tel.: +1-626-457-6678; fax: +1-626-457-6699.

0378-8733/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0378-8733(03)00012-1

measures of network centrality have been proposed to assess an individual's integration into his or her network (Valente and Foreman, 1998).

Two approaches have been used to help identify which centrality measure is most meaningful or useful for a particular research project. First, some researchers have explored the conceptual foundations of centrality measures (Freeman, 1979; Friedkin, 1991). Second, others have studied the empirical performance of centrality measures under different research scenarios or constraints (Bolland, 1988; Galaskiewicz, 1991). For example, Bolland (1988) evaluated the performance of measures of centrality degree, betweenness, closeness and continuing flow (Bonacich, 1972) in real and simulated networks. Using data from a network of influence relationships among 40 political participants in an education program in Chillicothe, Ohio, Bolland simulated the addition of random links between a target node and a randomly selected set of other nodes in the network. Each measure was then evaluated on its robustness to random error, and sensitivity to systematic variation in the network.

Galaskiewicz (1991) used data from a study of interorganizational linkages in two communities and sampled from the original networks at 75, 50 and 25%, 10 times at each sampling percentage. He then compared the resulting individual centrality scores to the individual centrality scores calculated on the original matrices. He found that as the sampling percentage increased or the number of trials increased the size of the errors of estimation shrank as did the variance of the errors and that he was able to derive more accurate estimates of actor's point centrality scores in networks that were sparser and for less popular actors (i.e. actors with fewer ties to others in the network). There was no evidence that network size affected the accuracy of the estimates.

The present study follows logic similar to the Galaskiewicz study but we include 11 centrality measures, expand the sampling levels and expand the empirical analysis. We simulate non-response in order to determine how stable different centrality measures are when respondents are not interviewed or do not respond. Network studies, particularly network-based interventions, have been hampered by the perceived cost of interviewing a large portion of the network. Complete population interviewing is motivated by concern that network measures become increasingly unreliable as sample proportion decreases. Burt (1983) argued that less than 100% participation seriously affects network data. Further, since every participant contributes $N - 1$ pieces of information (who they did and did not nominate) each missing interview constitutes a lot of missing data.

Even in settings that contain enumerated populations with delineated boundaries and thus appear ideally suited for network studies, for example, schools, organizations, and small rural communities, some network members are missed or refuse to participate resulting in less than 100% participation. Given that data are always missing on some segment of the population, conclusions that can be drawn about the entire population are subject to a degree of uncertainty. Frank has provided statistical approaches to quantify the degree of uncertainty in these conclusions (Frank, 2002).

This study compares centrality measures on their performance in different networks at decreasing sampling levels. Specifically, we compare the stability of 11 centrality measures at eight different sampling proportions, in eight different study settings containing 59 networks. We investigate the factors that may improve or retard their stability. Our results underline the importance of understanding how the network measures are calculated when selecting the most appropriate measure.

2. Methods

2.1. Datasets

Eleven centrality measures are considered; in-degree, out-degree, degree symmetrized, betweenness directed and betweenness symmetrized, closeness directed and closeness symmetrized, the first eigenvector, which we refer to as simple eigenvector, Bonacich's 1972 eigenvector centrality,[2] radiality and integration.[3] We also attempted to calculate Friedkin's measures of social influence centrality but were unsuccessful due to the fact that our networks were nonergodic. Each measure is calculated then correlated with the measure obtained when samples of the original network are taken. Data were originally collected in eight studies, which included 63 sociometric networks in a variety of settings. All of these studies interviewed or attempted to interview all of the members of bounded communities.

Table 1 presents characteristics of the datasets. The oldest study is the 1955 classic Medical Innovation study (Coleman et al., 1966; Burt, 1987). Physicians in four Illinois communities: Peoria, Bloomington, Quincy, and Galesburg, were asked to name three general practitioners who lived in their communities with whom they discussed medical practices, from whom they sought advice, and whom they considered friends.

Data for study two were collected in 1973 in a study of the diffusion of family planning practices in Korea (Rogers and Kincaid, 1981). Women in rural villages were asked to nominate five other village residents from whom they sought advice about family planning. Data from the third study were collected in rural villages in 1966 in a study of the spread of farming practices in Brazil (Rogers et al., 1970). Farmers were asked to name their three best friends, the three most influential people in their community, and the three most influential farmers in their community.

Data for studies four and five were collected in 1993 from women's voluntary associations, tontines, in urban Cameroon using both nominations and roster data collection techniques (Valente et al., 1997). Study participants initially were asked to nominate five friends who were members of their voluntary organization. In a separate question, study participants were asked to circle the names of friends on a roster, which listed the names of all members of the voluntary organization. These two questions may generate different networks and therefore were considered as two distinct datasets and centrality measures are calculated for each separately.

In these first five studies, network data were collected to study the spread of a new idea, opinion or practice (Valente, 1995; Rogers, 1995). In the last three studies, network data were collected in order to assist executives in organizations to better understand the information flows within and between organizations (Burt, 1992). Data for study six were collected in 1991 from all the attorneys, partners and associates, employed in a law firm (Lazega and

[2] Use of the first eigenvector and another eigenvector measure of centrality were proposed by Bonacich in his 1972 article. In 1987, Bonacich developed an additional measure of centrality using the eigenvector, which he referred to as "power". We were not able to include Bonacich's "power" measure in our analyses as by symmetrizing the missing data we created linear dependencies in the matrix and thus could not calculate "power".

[3] We were able to calculate both the directed and symmetrized versions of betweenness, closeness and degree because all of the ties reported were initially directed.

Table 1
Description of datasets

Dataset	Year of data collection	Setting	Make up of networks	No. of network questions	Question(s) asked
1	1955	Illinois communities	Physicians	3	Name three physicians who you consider friends, with whom you discuss medical practices, and from whom you seek advice
2	1973	Rural villages in Korea	Married women of childbearing age	1	Name five people in the village from whom you seek advice about family planning
3	1966	Rural villages in Brazil	Farmers	3	Name three best friends, three most influential people in the community, and three most influential farmers
4	1993	Urban Cameroon	Women belonging to a voluntary organization	1	Name five friends belonging to the voluntary organization
5	1993	Urban Cameroon	Women belonging to a voluntary organization	1	Circle names of all organization members considered friends
6	1991	Corporate law firm in the US	All attorneys	3	Circle names of all other attorneys considered strong coworkers, friends and individuals to whom you would go for advice
7	1996	IT department in a company in Latin America	All information technology (IT) employees	7	Seven separate questions regarding information exchange at work
8	1996	IT department in a company in the US	All information technology (IT) employees	7	Seven separate questions regarding information exchange at work

van Duijn, 1997; Lazega, 2000; Lazega and Pattison, 1999). A second distinction is that the boundary for this network was functional rather than geographic. The law firm had multiple offices throughout the US and as such the network data were collected among employees working in offices located in three different US cities. Data for studies seven and eight were collected in 1996 from the information technology (IT) personnel within a Latin American company and within a US company, respectively (Krebs, 1996).

In the law firm, attorneys were asked in three separate questions to nominate other lawyers within the firm whom they would consider to be close coworkers, friends, and individuals to whom they went for advice. Attorneys were given a roster of names and were allowed to nominate as many other attorneys from the roster as they chose for each question. In the high tech firms, IT employees were asked seven separate questions regarding the exchange of specific types of work information. For each question, they were allowed to select an unlimited number of names from a roster, which listed the names of all other IT personnel employed by their firm.

2.2. Network characteristics

These sociometric networks differ in their size, the number of questions asked of respondents, the type of questions asked, and the number of nominations allowed. Table 2 summarizes these differences and shows that most of these studies collected data from more than one network. For example, the Brazilian farmer's study interviewed farmers living in 11 different villages. The total number of networks in these eight studies was 63.

Given that our aim was to determine how well centrality measures calculated from sampled nodes approximate those calculated from the entire network, we felt it would be more difficult to make this comparison if information from a large portion of the network was not collected. Therefore, we excluded from our study any network in which less than 50% of the enumerated population initially responded to the network questions. Using this criterion we excluded one of the Illinois communities, one Korean village, and one of the Cameroonian women's voluntary organizations, leaving a final sample of 59 networks.[4] Since it would be cumbersome to present the characteristics of all 59 of these networks and since networks in the same study often shared similar attributes, Table 2 presents the average properties of the networks in the eight studies.

In these data response rates ranged from 51 to 100%. The medical innovation study had the lowest response rates while the organizational studies had the highest. Undoubtedly, the response rates are related to how well the network boundaries were defined and to study settings. Within a clearly bounded organization, response rates are likely to be higher as it is easier to locate all network members initially as well as to track down non-respondents. Response rates are also affected by the wording of the questions and the method in which surveys are administered.

Network size ranged from 34 to 169 people. The women's voluntary organizations in Cameroon were the largest while the physicians in Illinois were the smallest. However,

[4] Since the roster data and the nominations data for the Cameroonian women's voluntary organizations were considered as two distinct datasets, exclusion of data from one of the women's voluntary organizations resulted in the loss of two networks.

Table 2
Characteristics of networks

Dataset	Number of networks analyzed[a]	Average network size	Average response rate (%)	Average network density	Total number of nominations possible	Average number of nominations	Range of out-degree nominations sent	Average network centralization (symmetrized) (%)	Average network centralization (in-degree) (%)	Average network centralization (out-degree) (%)
1	3	64	56	0.06	9	2.61	0–8	24.11	20.04	12.26
2	24	68	64	0.03	5	1.64	0–5	20.02	21.06	5.12
3	11	76	82	0.03	9	1.94	0–7	27.35	30.04	5.77
4	9	83	76	0.04	5	3.13	0–5	22.08	28.65	2.03
5	9	83	76	0.49	Unlimited	39.06	0–152	28.82	16.77	49.77
6	1	71	100	0.32	Unlimited	22.15	2–49	33.23	30.64	39.46
7	1	72	82	0.20	Unlimited	14.19	0–34	24.39	24.34	28.69
8	1	45	96	0.38	Unlimited	16.62	0–40	43.45	35.74	54.34

[a] Networks in which the response rate was less than 50% were excluded from our analysis.

presentation of the average network size masks the fact that the Cameroon networks ranged in size from 55 to 169 nodes and the Illinois physician networks ranged in size from 34 to 199 nodes. There was a smaller range of network sizes in the other datasets.

Average network density ranged from 3 to 49% in the studies (Table 2) and from 1 to 47% in the networks. Network density is a result of four factors: network size, the number of nominations permitted, number of network questions asked, and the type of questions asked. All other things being equal, as network size decreases network density will increase (Scott, 2000) whereas as number of nominations recorded decreases network density decreases. Studies that limited the number of nominations to five or nine nominations were less dense, while those that did not limit the number were denser. A roster increases the number of nominations measured by cueing participants to think of every other individual in the network. Indeed, in the roster data some respondents circled everyone.

In addition, some networks included responses from multiple questions. For all of the datasets that asked study participants multiple network questions, centrality measures were calculated on the network created from the responses to all of the questions; thereby, allowing for a maximum of nine nominations in the case of the physicians, Brazilian farmers and corporate lawyers. However, it should be noted that asking people to name X other individuals to whom they are connected in Z different ways is different than allowing people to name as many other individuals as they choose for one question. In the former case, it is likely that the same individuals will be named for more than one question and therefore will not be double counted. Comparing the upper bound on the range of out-degree nominations with the total number of nominations possible confirms that this is the case for both the physicians' and the farmers' datasets (Table 2). Despite the fact that respondents were allowed to name up to nine other individuals in these networks, the greatest number of unique individuals that anyone nominated was eight and seven, respectively.

Network centralization describes the extent to which nodes in a network are connected to one or a group of individuals. Table 2 presents symmetric and asymmetric degree centralization scores and shows that that the IT employees in the US firm were the most centralized network, network centralization (symmetrized)—43.45%. A graph of this network reveals a central core of individuals (not shown). For centralized networks, we expect the centrality measures to be stable across sampling levels (e.g. a star of 20 nodes would be perfectly correlated with one of 10 nodes except when the central star is omitted).

2.3. Simulation procedures

Eleven centrality measures were calculated for each network. Symmetric measures were calculated on adjacency matrices symmetrized on the maximum (i.e. a nomination sent or received treated as a link). We then took repeated random samples of the network (bootstrapping) at each of eight different sampling proportions; starting at 80%, decrementing by 10%, down to 10%. We sampled by rows rather than by columns as we assumed that most researchers in the field would choose to use as much data as is available to them even if it means that some ties are only partially described.

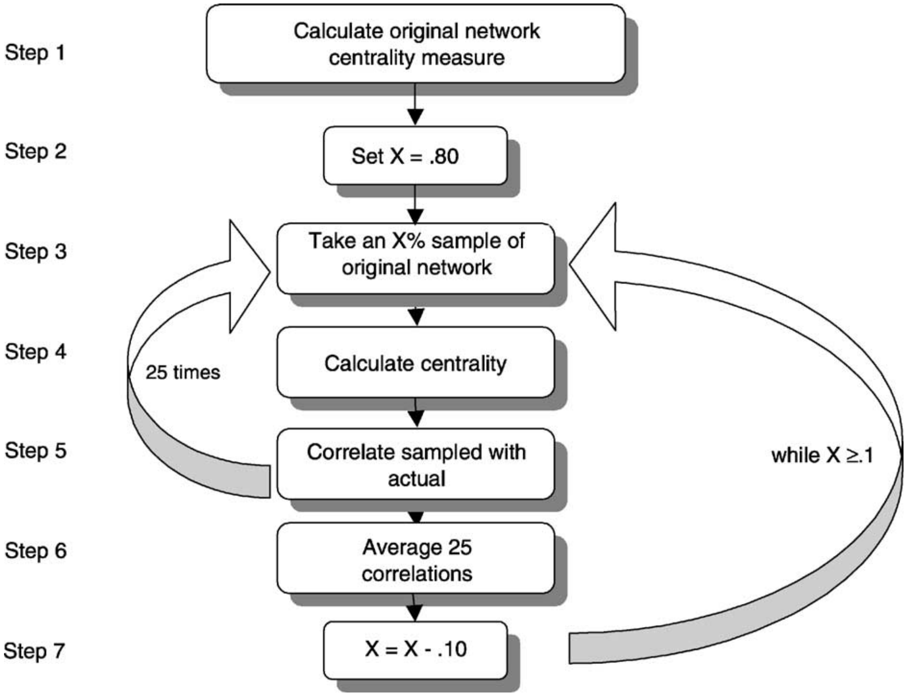

Fig. 1. Flowchart of simulation procedures.

At each sampling level, we correlated the original and sampled centrality measure 25 times, and then calculated the average correlation. We also calculated the difference and standard deviation of the differences.[5] Thus, the data consist of 5192 correlations (59 networks × 11 measures × 8 sampling levels) and the same number of differences and standard deviations. These correlations (and differences and variances) were then collapsed into 472 cases (59 networks × 8 sampling levels). Fig. 1 provides a flowchart of the simulation procedures. Multivariate linear regression analyses were used to determine which network characteristics best predicted the correlations between actual and sampled centrality measures.

Note that in some of the sampled networks (samples), centrality measures could not be calculated. This happened at the lower sampling proportions (10 and 20%) and most frequently for the betweenness and closeness measures since the networks became too sparse and disconnected. When this occurred, the correlation between the original centrality measure and the new centrality measure obviously could not be calculated. Nevertheless since this routine was run 25 times at each sampling level, it was always the case that some of the 25 samples provided enough ties for the new centrality measures and subsequently the correlations to be calculated. In these instances, the average of the 25 correlations between

[5] Our thanks to Kevin Reynolds for suggesting we look at the differences as well as the correlations.

the old and the new centrality measures was calculated dividing still by 25 despite the fact that there were not 25 correlations in the numerator.

3. Results

3.1. In-degree centrality

In-degree centrality measures the volume of ties directed to a node, calculated by summing the number of nominations received. Even at low sampling rates, in-degree had higher correlations between the actual and the sampled network measures than all of the other centrality measures with the exception of simple eigenvector centrality. Fig. 2 shows the results for in-degree centrality correlation among the 11 networks of farmers in Brazil. These results suggest that had the interviewers only been able to interview a random selection of 50% of the farmers in these networks, the calculation of in-degree centrality nevertheless would have shown a greater than 0.90 correlation with the measure of in-degree calculated using data from all of the interviews.

Fig. 3 provides further evidence of the stability of in-degree centrality at low sampling levels by showing the average correlation across all networks for each dataset. The average correlation coefficient when only 50% of the network responded was greater than or equal to 0.86 in all but one dataset.

3.2. Additional centrality measures

Fig. 4 displays the decline in average correlations across sampling levels for eight of the other 10 measures. Taking samples from the data for seven of these measures affects them to a greater degree than for in-degree centrality.

Out-degree centrality, also a measure of volume of ties but self-reported, is calculated by summing the number of nominations sent. A 50% sample of the original network had average correlations ranging from 0.39 to 0.62. The correlations decline more rapidly than for in-degree centrality. When we sampled the data, nodes dropped from the data were treated as if they were absent on the day of the survey. Consequently, these nodes had their out-degree centrality measure drop to zero. In contrast, the in-degree centrality measure was much less affected by the sampling because although individuals dropped from the sample were no longer able to send any nominations, they were still able to receive them. Sampling also had a greater affect on symmetrized degree centrality than on asymmetric in-degree (lower average correlations) because it includes nominations sent.

Betweenness centrality, the frequency a node lies on the shortest path between other nodes in the network, measures gatekeeping and control of information in a network (Freeman, 1979). The betweenness of a node is assessed by determining the extent to which a node lies on the geodesics connecting all other nodes in the network. The average correlation coefficient based on a simulated 50% sample of the network ranged from 0.38 to 0.54. Fig. 4 does not include the results obtained for betweenness centrality calculated on a symmetrized matrix (network). Similar to the results obtained for degree directed and symmetrized, the symmetrized measure had lower correlations than the directed one.

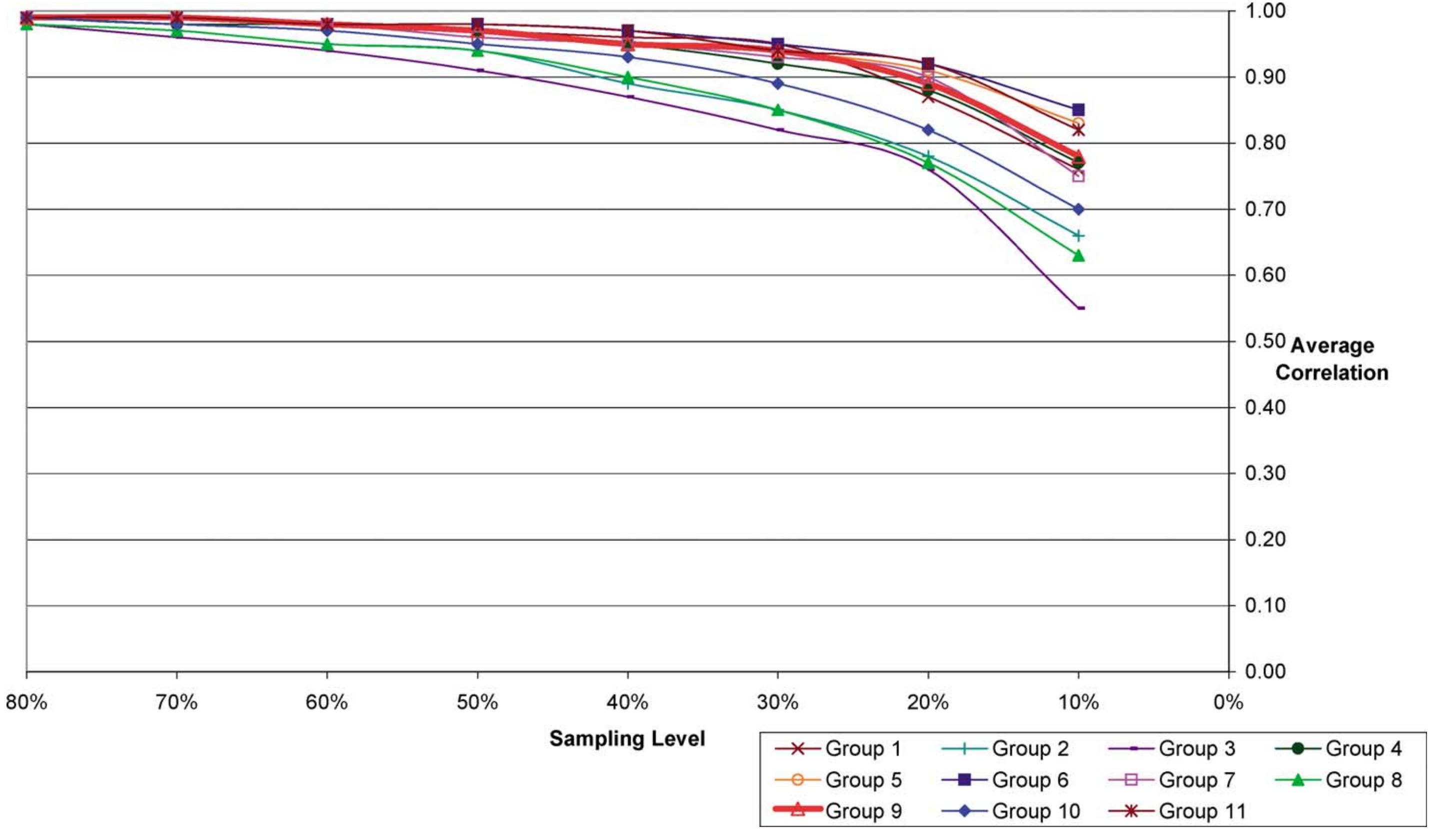

Fig. 2. Average correlation between actual and sampled in-degree centrality measures for the networks in 11 groups of farmers.

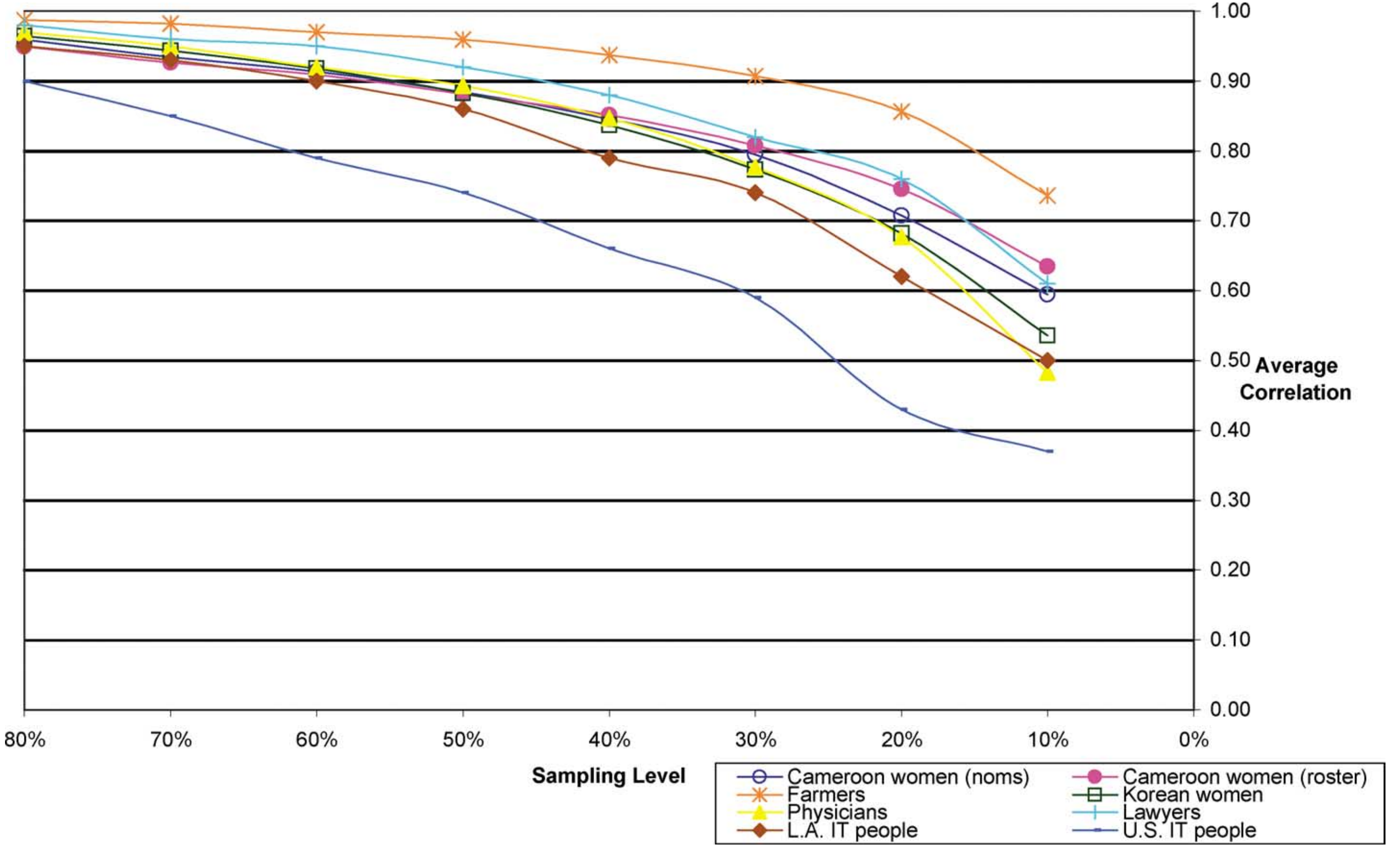

Fig. 3. Average correlation between actual and sampled in-degree centrality measure for the networks in eight datasets.

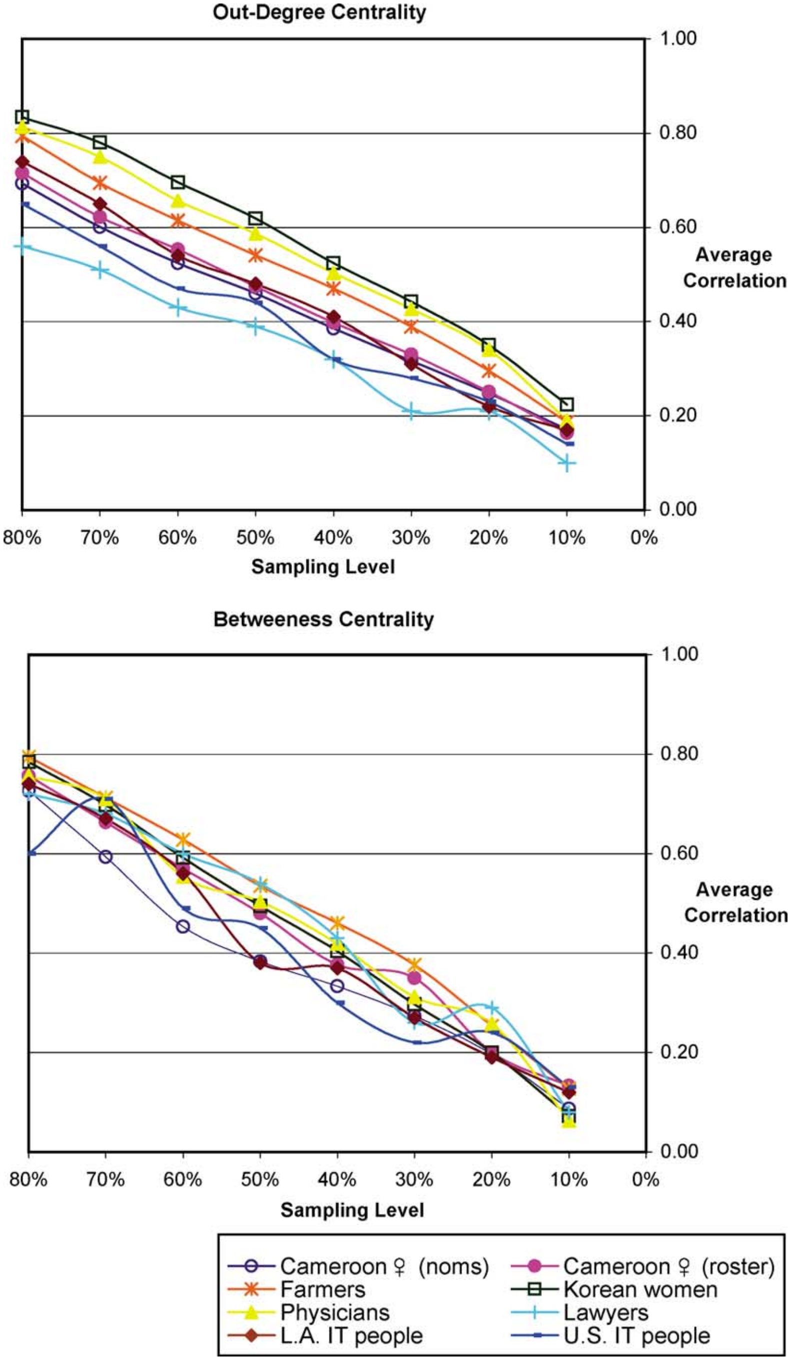

Fig. 4. Average correlation between actual and sampled data for eight different centrality measures in eight datasets.

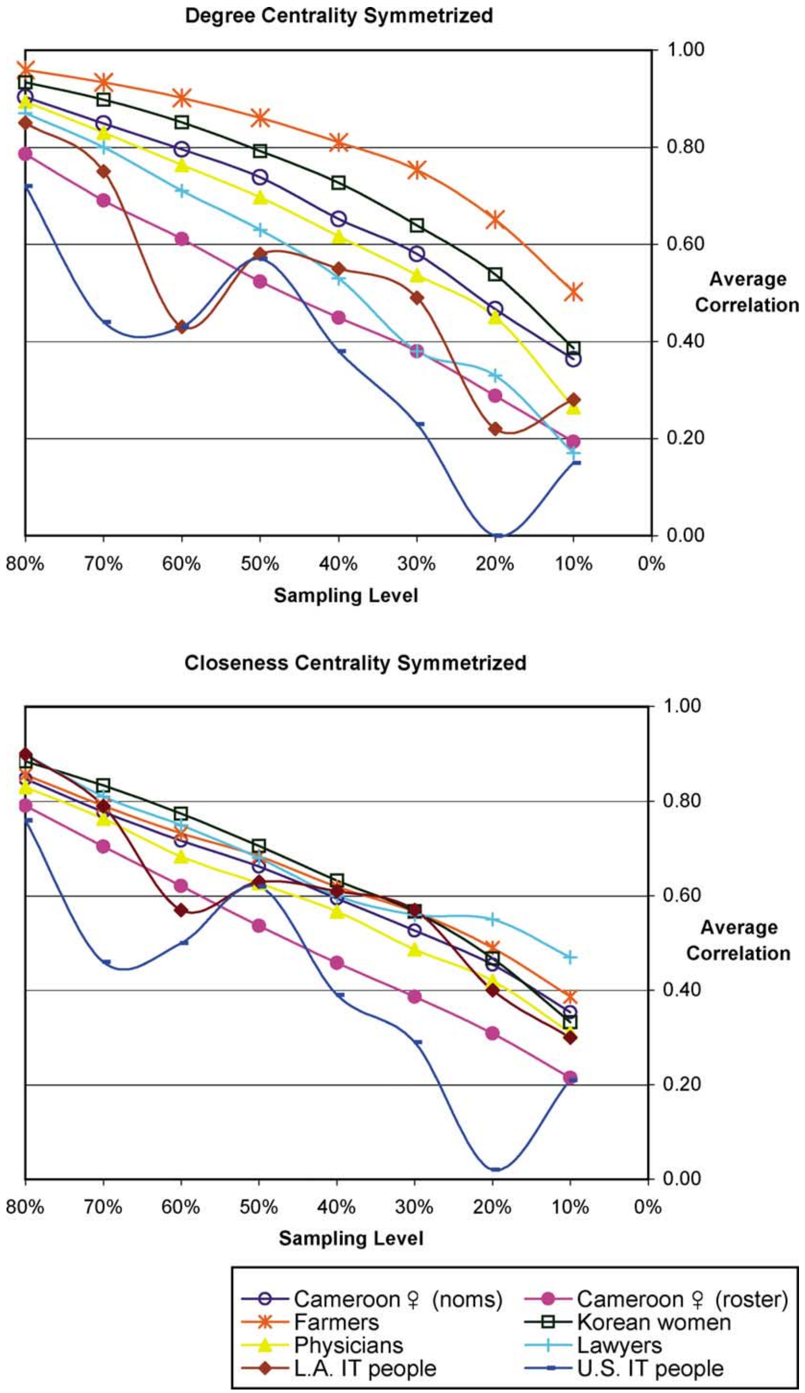

Fig. 4. (*Continued*)

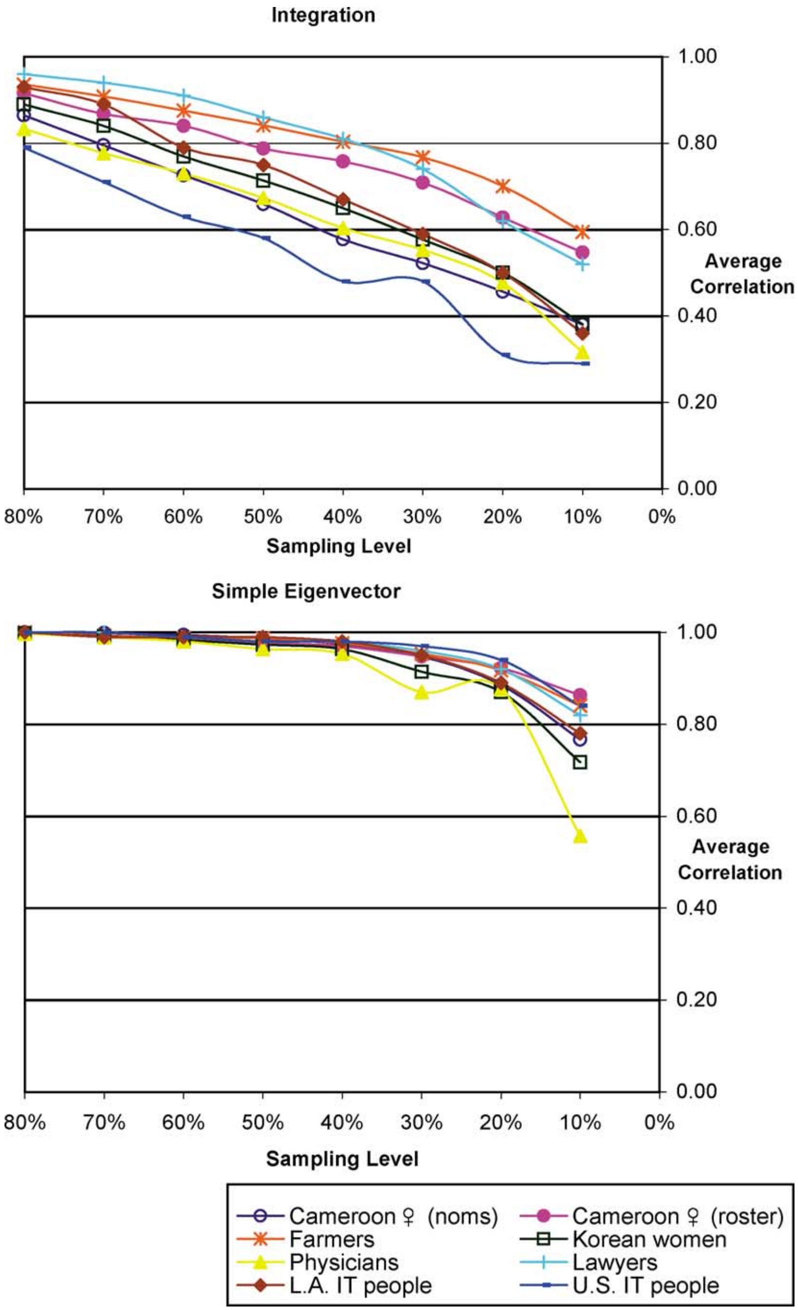

Fig. 4. (*Continued*)

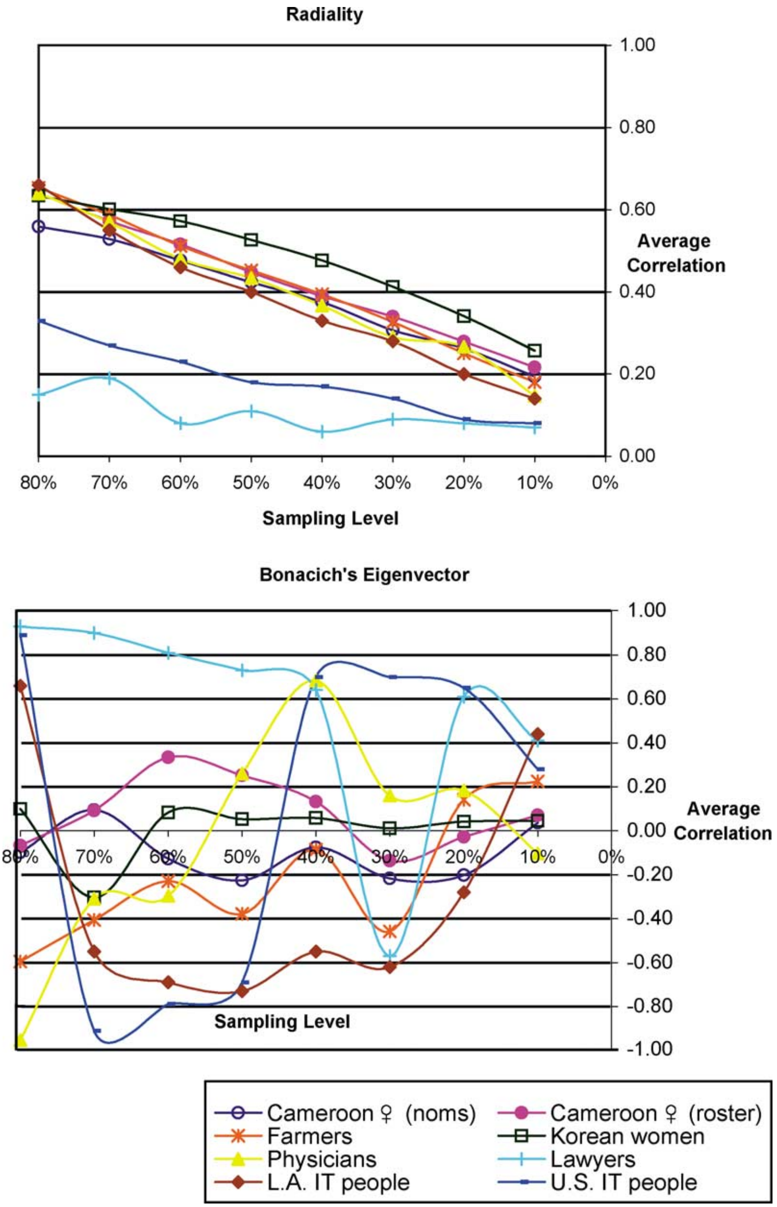

Fig. 4. (*Continued*).

Closeness centrality measures how many steps on average it takes for an individual to reach everyone else in the network. Individuals who have high closeness centrality measures can most efficiently make contact with others in the network (Freeman, 1979). Closeness centrality was measured by taking the reciprocal of the sum of the distances between individuals. For isolated individuals who have no ties to anyone else in the network, the sum of the distances between themselves and everyone else in the network would actually be zero. However, the reciprocal of 0 is an infinite number. Therefore, for these individuals the maximum geodesic or the largest path in the network was used in place of zero.[6] With a 50% sample of the original network nodes, average correlations for the closeness measure ranged from 0.54 to 0.71.

Typically closeness centrality is measured on a symmetrized matrix as it is assumed that if Kathy knows Sam that Sam can reach Kathy. However, since this may not always be the case we calculated closeness centrality on a directed matrix as well. Although these results are not shown, they indicate that in these 59 networks closeness centrality performs better as a directed measure than as a symmetric measure. Directed closeness performs better than symmetrized closeness due to the fact that these are relatively sparse networks, especially when only counting directed ties. As such, a large number of individuals in the asymmetrized network have the same closeness centrality in the original dataset as they do in each of the samples (i.e. the sum of the maximum geodesics).

Integration measures the degree nominations received integrate a node into the network (Valente and Foreman, 1998). Integration is similar to closeness centrality both conceptually and computationally. However, while closeness centrality totals the geodesic values and takes the reciprocal, integration totals the reverse distances between nodes. This difference, between reversing and reciprocating the sum of the geodesics, allows integration to be a directed measure. Integration shows correlations nearly as high as in-degree. With a 50% sample of the original network nodes, average correlations for the integration measure are 0.70 or greater in 41 of the 59 datasets.

Radiality measures the degree nominations sent reach out into the network (Valente and Foreman, 1998). An individual with a radial network has direct contact with individuals who do not have direct contact with each other. An individual with high radiality is able to reach everyone else in the network in fewer steps, on average, than an individual who has contact with individuals who are connected to each other. Radiality is calculated by computing integration on the transpose of the adjacency matrix. There is greater variation in the correlations between radiality calculated on the original networks and the sampled network than for any of the previous measures.

Of note in several of these graphs, most notably, degree centrality and closeness centrality, are the wave-like results for the networks of US and Latin American IT employees, and the network of lawyers. We were unable to determine what combination of factors may be causing this irregular pattern in the correlations for these datasets; although, the fact that these datasets did not limit the number of nominations an individual could send is likely

[6] There is debate in the social network community regarding path lengths for unreachable nodes. Mathematically, unreachable nodes have infinite distance. Substantively, a node in a network is part of the network and although not connected in the data is in some manner connected to the other nodes. Therefore, the maximum geodesic captures this distance.

to have contributed. If an individual originally sent 30 nominations and subsequently was not included into the sample this individual would experience a bigger change in his/her centrality score than would an individual who originally was only allowed to send five nominations and subsequently was dropped from the sample.

Eigenvector centrality when calculated using the first eigenvector as a simple raw score measure was the most stable measure being more highly correlated with its original calculation than were any of the other measures at low sampling levels. When only 10% of the network responded the average correlation coefficient for this simple eigenvector centrality measure was 0.70 or greater in 53 of the 59 networks. Eigenvector centrality when calculated in this manner is not a graph-theoretic measure, but instead is based on correlations (or more exactly shared variation). Consequently, sampling from the network seems to have had less effect on the rank ordering of the nodes than for the graph-theoretic measures. Simple eigenvector centrality's stability may indicate that it is the preferred centrality measure when the network data are incomplete. It should also be noted that the simple eigenvector centrality computation was done on symmetrized matrices so its stability should be compared to the other symmetrized measures.

Bonacich's 1972 eigenvector centrality measure quantifies the extent nodes are connected to other central nodes. Nodes connected to highly central nodes have greater centrality than those connected to less central ones. Because this measure relies not only on ego's centrality but also on the centrality of ego's alters this measure experiences the greatest fluctuations when nodes are sampled. It is interesting to note that while this measure is highly unstable across all sampling levels, its stability actually increases at smaller sampling levels. For example, when 80% of the network is sampled the average correlation coefficients range from −0.89 to 0.95 whereas when only 10% of the network is sampled the average correlation coefficients range from −0.10 to 0.44. This finding indicates that failure to interview even a few people in the network could lead to extremely different results for Bonacich's eigenvector measure of centrality. It would seem that this measure is best suited to situations where the interviewer clearly knows the boundaries of the network and is able to interview 100% of the network members.

Fig. 5 compares the performance of each of the 11 network centrality measures as sample size decreases. Eigenvector centrality as a simple raw score appears to be the most stable when sampled, followed by the in-degree and integration measures; both calculated primarily with nominations received. Out-degree and betweenness centrality, which rely more heavily on nominations sent, were among the least stable measures. By far, the most unstable measure across sampling levels was Bonacich's eigenvector centrality measure. Also among the least stable of centrality measures was radiality. Radiality and Bonacich's eigenvector measure are similar in as much as they both reach out into the network to determine ego's centrality.

3.3. Regression analysis

To understand the network characteristics that may account for variation in these correlations, we conducted multivariate linear regression analyses on the correlation of actual and sampled centrality measures, on the average difference between actual and sampled centrality measures and on the standard deviation of the difference between actual and sampled

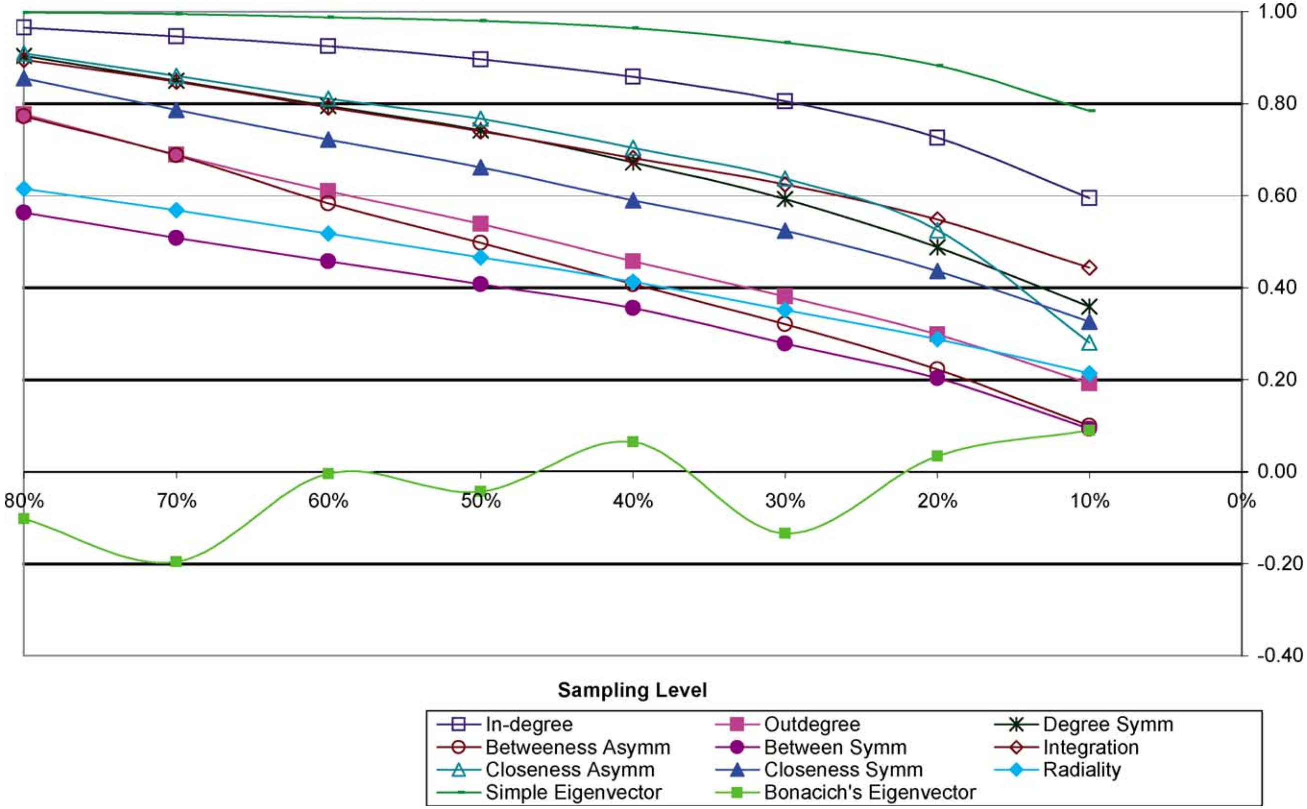

Fig. 5. Average correlation for 11 actual and sampled centrality measures computed for 59 networks.

centrality measures. Tables 3–5 present standardized coefficients for the correlations, average differences and standard deviations, respectively. In each table, the dependent variables were regressed on the following covariates: sampling level, response rate, network size, network density, and network centralization as well as indicator variables for each dataset (Cameroonian nomination dataset treated as reference category).

With the exception of Bonacich's eigenvector centrality measure, all models for the correlations had at least 65% variance explained (R^2). As expected, sampling level was strongly and significantly positively associated with the correlation between actual and sampled centrality measures for all of the measures other than Bonacich's eigenvector. The correlation between the actual and sampled Bonacich eigenvector measure was not significant but was negative as was shown in Figs. 4 and 5. Note that network centralization was significantly associated with the correlation for seven of the 11 measures and response rate, network size and network density were significantly associated with the correlation for five of the 11 measures.

Interestingly unlike sampling level, which had a positive association with correlation for 10 of the 11 measures, response rate had a negative association for seven of the 11 measures. The negative association most likely reflects the nature of missing data due to non-response. It is likely that the individuals who typically are missed or refuse to participate in a study are individuals who are on the periphery and have fewer connections to the rest of the social network. It seems plausible that individuals with fewer connections to the group are more likely to be absent at the time of the survey or to refuse to participate. Network density was negatively associated with correlations for out-degree, degree symmetrized, betweenness directed and symmetrized, closeness symmetrized and simple eigenvector centrality indicating that centrality measures for sparse networks are relatively stable.

The results in Table 3 indicate that different factors affect the stability of different measures. For example, density is positively associated with integration and radiality, and negatively associated with symmetrized degree and betweenness, indicating that the accuracy of integration and radiality measures may be more stable at higher density than other centrality measures. The measure most affected by sampling seems to be directed betweenness, which has the strongest coefficient for sampling level. Since different factors affect different measures, the choice of which measure to use depends on its theoretical formulation, not its stability.

Since the correlations are similarity measures, in some sense we anticipated that a regression of the differences would provide us with results that were in large part the inverse of the results for the correlations. Contrary to our expectations, this was not the case. Most notably, network characteristics rather than sampling level were more important covariates of the differences between actual and sampled data. As seen in Tables 4 and 5, the explained variance for the average and standard deviation differences was lower than for correlations indicating that other network or study characteristics may account for the differences. The models for the average difference explained very little of the variance in the directed closeness and Bonacich eigenvector measures and only from 45 to 76% of the variance (R^2) in the other measures. The models for the standard deviation of the difference explained as little as 8 and 37% of the variance in the Bonacich eigenvector and in-degree centrality measures, respectively but from 47 to 89% of the variance in the other nine measures.

Table 3
Standardized mulitvariate regression coefficients for the correlation between actual and sampled centrality measures on sampling level, study characteristics, and studies ($N = 472$)

Covariate	Measure										
	Degree			Betweenness		Closeness		Integration	Radiality	Simple eigen-vector	Bonacich's eigenvector
	In-degree	Out-degree	Symmetrized	Directed	Symmetrized	Directed	Symmetrized				
Sampling level	0.68*	0.92*	0.81*	0.95*	0.83*	0.82*	0.87*	0.74*	0.72*	0.79*	−0.84
Response rate	0.07	−0.17*	−0.19*	−0.07*	0.03	0.01	−0.12*	0.07	−0.23*	−0.02	−0.01
Network size	0.04	0.01	−0.10*	0.003	−0.06*	0.003	−0.11*	−0.06*	−0.20*	−0.02	0.02
Network density	0.01	−0.10*	−0.25*	−0.03	−0.19*	0.14	−0.12	0.21*	0.21*	−0.07	0.02
Network centralization (in-degree)	0.38*	−0.02	0.16*	0.08*	0.01	0.09*	0.03	0.13*	0.07*	0.08*	−0.01
Tontine nominations data (reference)											
Tontine roster data	0.15	0.10*	−0.01	0.11*	0.19*	−0.14	−0.10	0.11	−0.10*	0.15	0.07
Latin American IT department data	−0.02	0.03*	0.05*	0.06*	0.11*	−0.01	0.001	0.02	−0.03*	0.02	−0.03
Corporate law firm data	0.001	0.01	0.02	0.04*	0.12*	0.03	0.08*	0.05	−0.21*	0.04	0.12*
Brazilian farmer's data	0.19*	0.17*	−0.24*	0.13*	0.21*	0.14*	0.03	0.34*	0.05	0.04	−0.06
Medical innovation data	0.07*	0.06*	−0.09*	0.02	0.02	−0.04	−0.09*	0.03	−0.10	−0.07*	0.02
Korean family planning data	0.14*	0.20*	0.03	0.06*	0.05	−0.07	−0.005	0.18*	0.10*	−0.12*	0.08
US IT department data	0.08*	0.002	−0.08*	0.000	0.03	0.01	−0.09*	0.01	−0.20*	0.02	0.04
R^2	0.66	0.94	0.89	0.93	0.75	0.73	0.85	0.71	0.70	0.66	0.04

* $P < 0.05$.

Table 4
Standardized mulitvariate regression coefficients for the average difference between actual and sampled centrality measures on sampling level, study characteristics, and studies ($N = 472$)

Covariate	Measure										
	Degree			Betweenness		Closeness		Integration	Radiality	Simple eigen-vector	Bonacich's eigenvector
	In-degree	Out-degree	Symmetrized	Directed	Symmetrized	Directed	Symmetrized				
Sampling level	−0.19*	−0.19*	−0.27*	−0.28*	−0.36*	0.14*	−0.12*	−0.39*	−0.39*	−0.16*	−0.01
Response rate	0.09*	0.09*	0.08*	0.55*	−0.01	0.08	0.14*	0.35*	0.35*	0.02	0.33*
Network size	0.23*	0.23*	0.21*	0.21*	0.22*	0.08	0.01	−0.03	−0.03	0.09*	0.08
Network density	−0.35*	−0.36*	0.21*	−0.79*	−0.03	−0.11	0.33*	−0.31*	−0.30*	0.45*	−0.60*
Network centralization (in-degree)	0.07*	0.07*	0.08*	−0.11*	−0.26*	0.02	0.08	−0.02	−0.02	0.06	−0.06
Tontine nominations data (reference)											
Tontine roster data	1.09*	1.10*	0.60*	0.09	−0.32*	0.11	−0.12	0.51*	0.50*	0.24*	0.43*
Latin American IT department data	0.11*	0.11*	0.06*	−0.22*	−0.10*	0.01	−0.06	0.06*	0.05*	−0.05	−0.22*
Corporate law firm data	0.16*	0.16*	0.05	−0.16*	−0.10*	0.008	−0.14*	0.10*	0.10*	−0.09*	0.05
Brazilian farmer's data	0.01	0.01	0.01	−0.81*	0.38*	0.004	−0.19*	−0.58*	−0.58*	0.16*	−0.34*
Medical innovation data	0.08*	0.08*	0.06*	−0.07	0.09*	−0.12*	0.04	−0.02	−0.02	0.02	0.05
Korean family planning data	0.08	0.08	0.07	−0.52*	0.28*	0.001	−0.48*	−0.38*	−0.39*	0.07	−0.07
US IT department data	0.29*	0.29*	0.09*	−0.08	−0.03	0.02	−0.04	0.09*	0.09*	0.31*	−0.1
R^2	0.71	0.71	0.75	0.51	0.61	0.05	0.45	0.76	0.76	0.56	0.18

* $P < 0.05$.

Table 5
Standardized mulitvariate regression coefficients for the standard deviation of the difference between actual and sampled centrality measures on sampling level, study characteristics, and studies ($N = 472$)

Covariate	Measure										
	Degree			Betweenness		Closeness		Integration	Radiality	Simple eigen-vector	Bonacich's eigenvector
	In-degree	Out-degree	Symmetrized	Directed	Symmetrized	Directed	Symmetrized				
Sampling level	−0.07	−0.04*	−0.17*	−0.14*	0.01	−0.30*	−0.30*	−0.38*	−0.07*	−0.57*	0.02
Response rate	0.16*	0.08*	0.03	0.52*	0.31*	0.53*	−0.000	0.31*	0.09*	0.08*	0.18*
Network size	0.39*	0.26*	0.26*	0.33*	0.60*	−0.07*	−0.04*	−0.06*	−0.05*	0.12*	−0.15*
Network density	−1.04*	−0.42*	0.11	−0.85*	−0.77*	−0.44*	0.61*	−0.15	0.21*	0.21*	−0.29
Network centralization (in-degree)	0.14*	0.09*	0.13*	−0.03	0.09*	0.24*	0.07*	0.15*	−0.04	0.05*	−0.08
Tontine nominations data (reference)											
Tontine roster data	1.28*	1.25*	0.80*	0.11	0.14	0.03	0.02*	−0.61*	0.33*	0.44*	0.13
Latin American IT department data	0.13*	0.13*	0.08*	−0.05	−0.04	−0.12*	−0.01*	−0.27*	0.11	0.06*	0.03
Corporate law firm data	0.19*	0.17*	0.09*	−0.16*	−0.09*	−0.07*	−0.03*	−0.34*	0.07*	0.04	−0.07
Brazilian farmer's data	0.03	0.01	0.05	−0.70*	0.06	−0.43*	−0.001	−0.29*	−0.48*	−0.05	0.04
Medical innovation data	0.15*	0.09*	0.06*	0.00	0.10*	−0.06	0.00	0.02	0.00	0.05	0.1
Korean family planning data	0.18*	0.09*	0.06*	−0.43*	0.00	−0.43*	−0.14*	−0.21*	−0.35*	−0.02	0.12
US IT department data	0.42*	0.31*	0.10*	−0.05	0.04	0.19*	−0.01	−0.28*	0.07*	0.08*	0.01
R^2	0.37	0.84	0.86	0.47	0.73	0.69	0.89	0.72	0.86	0.82	0.08

* $P < 0.05$.

4. Discussion

Our results indicate relatively high correlation, albeit in some instances substantial absolute differences, between actual network properties and those calculated on randomly selected sub-samples for some network measures. This indicates that under some circumstances researchers may be still be able to use network data for which some data are missing to study network properties or create network-based interventions. In other words, researchers who do not interview all members of a community or network may still be able to take advantage of some aspects of network theory and techniques.

As noted previously, the stability of eigenvector centrality when calculated as a simple raw score may indicate that it is the preferred centrality measure when the network data are incomplete. However, the fact that sampling has less effect on this centrality measure may be due to the fact that in comparison to the other centrality measures, which measure the ones (i.e. the actual nominations), this measure is able to effectively capture the similarity of zeros. Since many of the studies restrict nominations to five people, there are a lot of zeros in the original networks. Consequently, eigenvector centrality as a simple raw score is less affected by sampling from the networks as the zeros are preserved. Conversely, radiality and Bonacich's eigenvector centrality measure, which reach out into the network and rely on information about ego's alters, appear to be less preferred measures when network data are incomplete. Yet since these measures rely on more global network information, they may do a better job of measuring centrality relative to the whole network.

The results of this study should be interpreted with caution as there are a number of limitations to the generalizability of these findings. Although an effort was made to include sociometric network datasets that were collected for networks of different sizes, for different reasons, in different locations and using different questions, these eight studies are by no means a comprehensive subset of the many varieties of sociometric network studies that have been conducted to date. For instance, this subset of datasets does not include either a study of disease transmission or a longitudinal study of network properties.

Furthermore, given our limited information about several of these aforementioned aspects of the network data (i.e. motivation for conducting the study, study setting and choice of network generating questions) we were not able to distinguish in our analysis between the influences of these different factors on the resulting network properties. In addition it should be noted that another network factor that may be influencing our results is the content of the tie since we include studies of friendship ties, coworker ties, discussion partners and advice and information sources. In fact, the regression results shown in Table 3 reveal that for nine of the 11 measures the Brazilian farmers' dataset showed higher correlations across sampling levels than did the Cameroon nominations dataset. Nevertheless, why the Brazilian farmers study had higher correlations is a matter of conjecture given that in our regression analysis we controlled for these various influences simultaneously by simply including an indicator variable for each dataset.

Our results are premised on the notion that sampling provides data missing at random, rather than missing due to some attribute of the respondents or study setting. Clearly, non-random factors influence respondent participation that could be accounted for in future studies. In addition, it is possible that there are other factors unknown to us that may have influenced how study participants responded to the network questions or the way in which the network centrality measures correlated across sampling levels. For instance, the rapport the researcher had with the study participants may have influenced how much attention study participants gave to answering each question carefully. Alternatively, other ties may have existed between network members that may have correlated better across sampling levels, but about which no data were collected.

Rather than providing a definitive answer to the question of whether or not sociometric network data collection should entail complete population interviewing, these simulations provide a tentative benchmark for future studies. In other words, network researchers may find it helpful to refer to our findings when contemplating the appropriateness of sampling under the circumstances and interests of their particular studies. We do not mean to imply however that concerns about sampling should be a primary motivating factor. Finally, we would encourage future network studies to employ our methodology on their own data in order to contribute to the existing body of knowledge regarding sampling from sociometric networks.

The next phase of this research agenda will be to determine appropriate snowball sample sizes, essentially reframing the sampling problem from a snowball perspective: how many initial nodes (index cases) and their alters are needed to confidently measure a network. This problem is at an order of magnitude more complex than the one presented here since we have to vary the sampled proportion of index cases, alters, and steps (direct nominees, nominees of nominees and so on). The results, however, would be useful for contact tracing studies, and alternative network study designs.

Our tentative conclusion is that some centrality measures such as in-degree, integration, and the simple eigenvector are relatively stable provided the network boundary can be specified. In-degree centrality is relatively stable even at a low sampling level (i.e. 50% missing at random) and so can be used for network-based interventions (Valente and Davis, 1999) and as an indicator of network position. Eigenvector centrality as a simple raw score was also stable due in large part to its ability to capture network structure based on nominations not made as well as those made. Study characteristics mattered, measures in some studies were more stable than others, and so we recommend that future studies with less than 100% response rates compare their results to these in order to estimate the potential bias in their centrality measures.

Acknowledgements

Our thanks to two anonymous reviewers, who provided helpful comments on this manuscript, and Robert Foreman for assistance with some computer programming. We also thank Noah Friedkin for making SNAPS available. Partial support for this research was provided by NIH grant DA 10172.

References

Bolland, J.M., 1988. Sorting out centrality: an analysis of the performance of four centrality models in real and simulated networks. Social Networks 10 (3), 233–253.

Bonacich, P., 1972. Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology 2 (1), 113–120.

Bonacich, P., 1987. Power and centrality—a family of measures. American Journal of Sociology 92 (5), 1170–1182.

Burt, R.S., 1983. Studying status/role-sets using mass surveys. In: Applied Network Analysis: A Methodological Introduction. Sage, Beverly Hills, CA (Chapter 5).

Burt, R.S., 1987. Social contagion and innovation: cohesion versus structural equivalence. American Journal of Sociology 92, 1287–1335.

Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA.

Coleman, James S. Katz Elihu, Menzel Herbert, 1966. Medical Innovation: A Diffusion Study. Bobbs Merrill, New York.

Frank, O., 2002. Using centrality modeling in network surveys. Social Networks 24 (4), 385–394.

Freeman, L.C., 1979. Centrality in social networks: conceptual clarification. Social Networks 1, 215–239.

Friedkin, N.E., 1991. Theoretical foundations for centrality measures. American Journal of Sociology 96 (6), 1478–1504.

Galaskiewicz, J., 1991. Estimating point centrality using different network sampling techniques. Social Networks 13 (4), 347–386.

Krebs, V., http://www.orgnet.com/IHRIM.html. 1996. Ref Type: Electronic Citation.

Lazega, E., 2000. Teaming up and out: getting durable cooperation in a collegial organization. European Sociological Review 16 (3), 245–266.

Lazega, E., van Duijn, M., 1997. Position in formal structure, personal characteristics and choices of advisors in a law firm: a logistic regression model for dyadic network data. Social Networks 19 (4), 375–397.

Lazega, E., Pattison, P.E., 1999. Multiplexity, generalized exchange and cooperation in organizations: a case study. Social Networks 21 (1), 67–90.

Rogers, E.M., 1995. Diffusion of Innovations, 4th Edition. Free Press, New York.

Rogers, E.M., Kincaid, D.L., 1981. Communication Networks: A New Paradigm for Research. Free Press, New York.

Rogers, E.M., Ascroft, J.R., Röling, N.G., 1970. Diffusion of Innovations in Brazil, Nigeria, and India, Michigan State University. Ref Type: Unpublished Work.

Scott, J., 2000. Social Network Analysis: A Handbook, 2nd Edition. Sage, Thousand Oaks, CA.

Valente, T.W., 1995. Network Models of the Diffusion of Innovations. Hampton Press, Cresskill, NJ.

Valente, T.W., Foreman, R.K., 1998. Integration and radiality: measuring the extent of an individual's connectedness and reachability in a network. Social Networks 20 (1), 89–105.

Valente, T.W., Davis, R.L., 1999. Accelerating the diffusion of innovations using opinion leaders. The Annals of the American Academy of the Political and Social Sciences 566, 55–67.

Valente, T.W., 1997. Social network associations with contraceptive use among Cameroonian women in voluntary associations. Social Science and Medicine 45 (5), 677–687.

Wasserman, S., Katherine, F., 1994. Social Network Analysis Methods and Applications. Cambridge University Press, Cambridge.