Toward A Confidence Estimate For The Most-Central-Actor Finding

Terrill L. Frantz & Kathleen M. Carley

ABSTRACT

We introduce a innovative practice of providing a statistical confidence estimate to accompany the reporting of the most-central actor according to social network data; often times, collected relationship-data has some amount of error, which may affect the accuracy of the top-actor finding. This confidence estimate is the frequency-based likelihood that the *data-based* finding accurately identifies the most-central actor in the *ground-truth* social network. Among other possibilities, this confidence estimate is immediately applicable to: degree, betweenness, closeness and eigenvector centrality. This paper describes and illustrates a practical, non-parametric resampling approach for determining, applying and evaluating this confidence estimate.

Keywords:

social network analysis, statistical confidence, network centrality

INTRODUCTION

Social network analysts commonly seek to identify the most-central actor, according to any one of several measures of centrality, because of the significance of this distinctive network position (Ibarra, 1993; Ibarra & Andrews, 1993) and its association with social power (Brass & Burkhardt, 1993; Klein, Lim, Saltz, & Mayer, 2004) and individual performance (Cross & Cummings, 2003; Sparrowe, Liden, Wayne, Kraimer, 2001). For example, identifying the actor with the most relationships in a friendship network: the actor with the highest degree centrality is indicative of the individual with the highest potential-access to gossip (Burt, 2001; Wittek & Wielers, 1998). However, given the regularity that data collected on a social network has some amount of missing data--arising from various systemic and other factors (Kossinets, 2006) —, how confident should the analyst be in this data correctly indicating the in-truth top-actor?

Depending on the question being considered, analysts typically identify the most-central actor according to one of the four traditional measures of centrality: degree, betweenness, closeness and eigenvector (Bonacich, 1987, Freeman, 1979). This undertaking is accomplished by computing the value of the particular centrality measure for each unique actor in the network, rank-ordering the actors according to the computed value, and referring to the actor with the highest value as the being the top-actor (Carley, Columbus, DeReno, Reminga, & Moon, 2008). Unfortunately, the data collected from which the measures are computed often contains error, such as the under-reporting of network links . Despite the well-known problems with accuracy of real-world data, the analyst nevertheless reports the central-actor according to what data is available. Historically, social network analysts have blindly treated the top-actor finding as if it were an unbiased estimator of the ground-truth network; we posit that this may indeed be an inaccuracy as the top-actor finding may have an undetermined amount of bias. While a small

amount of data error may be unimportant to the correctness of the resulting central-actor claim, an extraordinarily large amount of error will undoubtedly result in a false claim (Frantz & Carley, 2006; Frantz, Cataldo & Carley, 2008).

In this paper, we somewhat marginalize the ubiquitous data error problem by providing a manner in which the analyst can provide an quantitative estimate of the confidence they have in their claim: a confidence estimate to accompany the most-central actor pronouncement. We introduce a process to identify this boundary condition and facilitate its practical application in research and in the field. Including this confidence estimate is invaluable to users of the analysis; essentially this provides much more complete information from which to make later inferences and draw substantiated conclusions about the characteristics of the true social network according to the data collected.

Concerning the organization of this paper, we first provide a simplistic example of the notion of providing an algebraically-based confidence estimate, which also illustrates an unintuitive dynamic of such a confidence estimate--at least as evident in the case of this simple example. We then present a more complex example to illustrate the limits of the straightforward algebraic approach by presenting a more versatile methodology that utilizes computer-based, experimental trialing. Finally, we present a realistic application of the complete process according to how an analyst would actually apply this approach in their analytic work.

INSTRUCTIVE EXAMPLE

We introduce the practice of estimating a confidence statistic for the central-actor identification task through the presentation of a very simple and tractable example. We begin by looking at a conventional, organization-reporting structure--a relationship strucutre consisting of only five actors. This network is constructed in the stylized form of a hierarchical network, also called a

star network. This simple organization consists of a single manager and four subordinates, also identified as an *ego* and four *alters* (see Fig. 1). We approach this example by classifying the network links as being undirected in the relation of "associated with", as opposed to the directed-link relationship "reports to", or "is the manager of." Contrary to the data that analysts typically have to work with, i.e., the collected sample, this example network is the actual in-truth data; it follows then, that in-truth, this undirected network has density of 40%, e.g. 5 actors and 4 links, out of a possible 10 links. The most-central actor according to degree centrality, in-truth, is the Actor A (the ego) as it the node with the most adjacent links.

Insert Figure 1 about here

Before we explore this example further, we should make clear the operational procedure of identifying the most-central actor, hereafter also referred to as the *top-actor*. We apply the standard competition ranking strategy, whereas ties are grouped and rank-numbered with the number gap following the group rank value, e.g., "1224". Moreover, there are two idiosyncratic viewpoints of what the unary top-actor is when a tie-group is formed. Sometimes more than one actor has the value equal to the highest centrality value and therefore more than one actor can correctly be identified as the top-actor; in this circumstance, there is a multi-member *top-group*. For the first view, the analyst selects one member of the top-group and identifies that actor as the top-actor; we consider this view is the *strict* approach. The second view considers that the top-actor is "one of the members of the top-group"- this a *relaxed* approach. Certainly, the strict approach will have lower confidence values than that of the relaxed approach, ceteris paribus.

To better appreciate the nuance between the two views, for example, imagine a hypothetical network dataset, where two actors (Actor A, and Actor B) have exactly the same degree

centrality value, say 0.15, which is the highest value for all actors in an observed network dataset. Assume the analyst arbitrarily labels Actor A as the top-actor, as opposed to selecting and labeling Actor B as the top-actor. With the relaxed approach, the analyst will offer a 100% probability that Actor A is "a" top-actor. However, using the strict approach, where the analyst must chose either one of Actor A *or* Actor B as the top-actor, there is a 50% probability that the correct actor will be selected (a 1 of 2 chance). Since herein, we are introducing a new idea, we prefer parsimony and will therefore apply only the relaxed approach throughout this paper.

Insert Table 1 about here

Returning to the simple, 5-actor hierarchy example we began to describe above and shown in Fig. 1: if in fact, the analyst has this complete information, without any data error, the top-actor according to degree centrality is Actor A (the ego). However, if the dataset was actually missing any one of the relational links, thus the data error is 25% (1 missing link of 4 in-truth links), Actor A will still be correctly identified as the top-actor, according to the data. We thereof assign a confidence estimate of 100% (see Table 1), according to this probability. Moreover, if two links are missing (50% error), Actor A will still be correctly identified as the top-actor, with 100% confidence.

Insert Table 2 about here

As shown in Table 2, perhaps astonishingly, the analyst will *always* correctly identify the true top-actor according to the collected data, even if all of the in-truth links are missing; the top-actor confidence estimate is 100% across all possible error levels of a hierarchical network. Such

precision is due to the particular features of the *relaxed* approach for handling the top-group situation-- as discussed above. The frequency probabilities and their distribution across data error level is anticipated to differ under the strict approach.

COMPUTING THE CONFIDENCE ESTIMATE ALGEBRAICALLY

In this section we delve deeper into determination of the confidence estimate by introducing a algebraic approach to the process of computing the probability value and therefore the confidence estimate. Staying with a small network, albeit a bit more complex than the hierarchy network above, we set up a five-actor kite network (See Fig. 2). The kite network is an illustrative topology that is often used in the classroom to teach the centrality measures.

Insert Figure 2 about here

To compute the confidence estimate, we follow a process similar to that described in the prior section, but since individual links may have a assorted amount of impact on the centrality measure values (note that the prior hierarchy network has links that are isomorphic, so they are all equivalent in the centrality computation), we need to compute a summarized metric for the specific combination of independent variables. What is actually a probabilistic distribution can be simplified to being computed as a summation of the number of accurate results divided by the total number combinations possible; basically, a relative frequency of successful outcomes. Table 3 shows the impact of removing one link (20% error) according to the specific tie being removed, which is similar in result to the previous, hierarchical-network example.

Insert Table 3 about here

However, Table 4 shows the attention-grabbing dynamics of the outcomes when two links are removed (40% error). We present a table of the five links in the kite network and the outcome when any two of them are missing from the data being analyzed. For example, if link DE and AB are missing, then Actor D is still correctly labeled as the top-actor; in this case D is one in the top-group of two actors. The converse is true in the case of links DE and CD being missing. In this case, actors A and B make up the top-group and actor D is not a member of the top-group; therefore, the actor identified will be incorrect, resulting in a value of 0 for this cell.

Summarizing this same outcome matrix into a probability matrix that is based on the method for breaking top-group ties (See Table 5), makes this a bit clearer to understand the impact on the final computation of the resulting confidence estimate. To compute the final metric, the values in Table 5 are combined and the mean will represent the confidence value; in this case the final value is 70% (7/10).

Insert Table 4 about here

Insert Table 5 about here

While this procedure is tractable for up to two missing links, going beyond two links makes the frequency tables unwieldy as the number of dimensions in the table equals the number of missing links as does the number of cells increase with the total number of links possible for the given network. An approach to working with these higher dimensional circumstances is provided in the following section.

DETERMINING THE CONFIDENCE ESTIMATE EXPERIMENTALLY

The simplistic examples presented in the prior sections are merely illustrative and not directly relevant to the complexity of real-world, social network analysis. Nevertheless, as the number of actors in the network increases, we quickly depart the problem space of having algebraically-tractable circumstances as the necessary combinatorics quickly become computationally unwieldy, at best, and practically impossible, as worst. In order to continue, we must instead attempt to approximate the confidence estimate subject to several relevant network parameters, such as size, density, topology, and error level using a different, less exacting approach.

Instead, we will conduct a computer-based experiment to approximate the frequency probability-, or likelihood-based, confidence estimate for a particular scenario using a nonparametric, bootstrap procedure similar to the generalized Efron (1987) strategy for broader statistical applications. Similar non-parametric resampling techniques have been shown to be effective in developing confidence estimates in other scholarly domains, e.g., medicine (Campbell & Torgerson, 1999; Walling, Visscher, & Haley, 1998). We also borrow from and modified the experimental design developed for earlier research into the broader aim of studying the quantitative robustness of the centrality measures (cf. Frantz, Cataldo, & Carley, under review; Frantz & Carley, 2005; Borgatti, Carley & Krackhardt, 2006).

Herein, we execute independent replications of a Monte Carlo experiment that involves randomly generating a social network. This network replicates the relevant characteristics of the observed network, according to size, density and topology. We then make a copy of that network data exactly, then perturb the exact copy according to the error level parameter. To perturb the copy, we add the number of links according to the error level, relative to the number of links that would produce the number in the dataset. Links are repetitively added randomly between pairs

of actors that do not presently have a link until the number of previously-missing links is reached. The centrality measures of interest for each actor in the two networks are then computed.

The actors in each network are then rank-ordered according to their corresponding centrality value. A comparison of the top-actor finding is made between the pair of networks. If there is a match then a binary count variable is set to 1, indicating that the observed top-actor is a match to the in-truth top-actor, else the indicator is set to 0. To compute the confidence estimate, we compute the relative frequency of correct identifications (there is a match, thus the binary variable is set to 1) relative to the number of trials, resulting in the standard frequency probability of being correct; this is therefore the confidence estimate for the network specific to: size, observed topology, density and estimated error level.

Conceptually, we are repetitively drawing a sample from the set of all possible universes, or ground-truth networks, that are bounded by the analyst a priori approximation of the network topology and data error level, and the a posteriori variables evident from the network data being investigated, i.e. network size and density. From each sample (with replacement) from the possible universe of ground-truth networks, we determine the most-central actor according to the specified centrality measure. We then compare this finding with the finding from the original network dataset, resulting in a binary outcome. This process repeats for each replication of the experiment. The confidence estimate we, therefore, compute is the proportion of times that the possible-universe sampled finding matches the dataset finding, relative to the total number of replications. This value is the likelihood estimate for social network of the same characteristics as the original, that the top-actor finding from the dataset is actually the top-actor in the actual social network.

To assess the soundness of this experiment-based approach being used as a proxy for the full combinatory computation, we conducted this procedure on the identical kite network example discussed in the prior section. The results based on 1,000 replications each cell is reported in Table 6. We point out that the experimental approach provides an estimate of 70.1% for a confidence estimate for degree centrality with 40% error. This corresponds with the 70% determined for the same scenarios using the algebraic approach. Table 6 takes the example further by also reporting the confidence estimates for the three other traditional centrality measures as well. Keep in mind, this is an extremely simplistic example and the reader should not make too much out of the correlation of the values across the measures. We expect that this is a byproduct of the characteristically small network being investigated.

Insert Table 6 about here

The confidence estimates reported in the above table can be used as accessories to the top-actor claim being reported by the analyst. For example, for this network and a estimated error of 20%, the analyst can state the claim that the actor is the *true* top-actor according to the betweenness measure, but with estimate of only 60.8% confidence. Further, if the same was under the pretense of 100% error, the same claim would be guaranteed to be correct. Of course, this broader condition has little real-world practical value.

APPLYING THE CONFIDENCE ESTIMATE

We now present a hypothetical, yet realistic, situation that illustrates the complete process and results of this technique. Figure 3 shows a node-link diagram of an undirected social network with 32 actors and 86 links, thus the network is characterized as having a density of 17.3%. We

loosely conjecture that the social network has a scale-free topology, as it is network assumed to be constructed under the pretense of preferential attachment. We also estimate from past experiences in similar-style surveys, that the data reported will be underreported by 22 unique links. Therefore, an a priori error level of 20% error (22/[86+22]=80%) is assigned.

Insert Figure 3 about here

According to the dataset, Actor A-31 is identified as the top-actor according to degree centrality. Moreover, Actor A-31 is also the top-actor for closeness and betweenness, while Actor A-1 is the top-actor in eigenvector centrality. To estimate the separate confidence estimates for these four claims, we used the experiment methodology described above, tailored for this specific scenario, i.e., 32 actors, 17.3% density, scale-free topology, and an error level of 20%. We set the software to run through 1,000 independent trials. Though not part of the actual process, but for illustrative purposes we also ran the process for several other error-level settings to provide a sense of the implications of the a priori error-level estimate, The detailed results of all these runs, for the degree centrality, is reported in the *Scale-Free Topology* column in Table 7. In addition, to develop greater understanding of the pre-analysis assumption of the network topology, Table 7 also reports results for a comparable network of *uniformly random* topology.

Figures 4, 5, 6, and 7, show the results of the entire in graphic form. Using these results, the analyst can claim that using data with an estimated 10% level of error, that: (a) Actor A-31 is the top-actor according to degree centrality with 86% confidence, (b) Actor A-31 is the top-actor according to betweenness with 79% confidence, (c) Actor A-31 is the top-actor according to closeness with 80% confidence, and (d) Actor A-1 is the top-actor according to eigenvector centrality with 68% confidence.

Insert Table 7 about here

Insert Figure 4 about here

Insert Figure 5 about here

Insert Figure 6 about here

Insert Figure 7 about here

IMPLICATIONS

We put forth that this technique can be useful in any social network analysis situation. Broadly, the capability to assign a confidence estimate to top-actor identification has immense applicability and value in the realm of social and organizational analysis: business and military, in research and practice, and for improvement (think management) and demolition (think terrorism). Specifically, this scheme will be beneficial to the top-actor identification in an email-based analysis of a military unit (e.g., McCulloh, Ring, Frantz, & Carley, 2008); or likewise, for strengthening the conclusions reached in a study of university faculty (e.g., Frantz & Carley, 2005). This metric may also prove to be especially valuable to perfecting automated decision-making tools and other social network related practices that are computationally-based.

SUMMARY

The purpose of this paper is to establish a method by which an analyst can provide a confidence estimate to the identification of the top-actor in a network according to one of the traditional centrality measures. We provided an example of the process using algebraic techniques for a simplistic social network. We then presented a bootstrap-like process using computer-based trialing to estimate the statistical confidence for a more complex network example. Finally, we demonstrated precisely how the likelihood-based estimate can be used to provide an indication of confidence to the identification of the top-actor in a social network, which is an important task in a network analysis in any social network analytic setting. This approach to working with social network data provides greater transparency to the consumers of analyses derived from relationship data collected in the real-world. While the nonparametric approach we take to determine the confidence estimate is an established method in other realms, it has yet to be applied to social network data. We believe that this approach, as we provide, has immediate benefit to analysts and consumers of social network data as is, though there is more work to be done in this realm.

LIMITATIONS AND FUTURE WORK

While it is readily apparent that social network analysis can benefit from the use of confidence estimates in reporting analytic results, there is much more research needed in this area in order to develop the statistical soundness of the estimates and operationalize the process of computing a precise estimate. The most pressing issue to this end, is the present inability to accurately approximate the a priori error-level and the topology of the data, which is a requirement of the technique we present, and possibly any future, alternative techniques as well. While much is known about types of missing data and other variants of error in social network data collection,

little is known about estimating the actual amount of error in specific circumstances. Moreover, the topological aspects of real-world social network data is only now beginning to be explored, but there is much more to do in this area as well. The forward advance of utilizing confidence estimates, we believe, is highly dependent on scholarly progress in these two areas. This void will remain a significant limitation to estimating confidence in this and numerous other statistical aspects of social network analysis.

Herein, we considered only the missing-link data error, which is only one type of error that analysts must be equipped for – namely, missing nodes, extra nodes, and extra links. Moreover, the random-network topologies generated in this experiment are based on just one variant of the same-classified topology, e.g., it is likely that a full set of computer runs need to be conducted on the full parameter space of a stylized scale-free, or cellular network, for example. As in any socially constructed real-world dataset, one must be cautious of non-random, systemic, or intentional error; herein, we are parsimonious and thus assume an entirely random error characteristic (see Rubin, 1976), which may impact the accuracy of the confidence estimate in real-world applications. We leave these complications and many others not yet discovered for our future work in this important area of social network analysis.

REFERENCES

Bonacich, P. 1987. Power and centrality: A family of measures. **American Journal of Sociology**, 92: 1170-1182.

Borgatti S., Carley, K., & Krackhardt, D. 2006. On the robustness of centrality measures under conditions of imperfect data. **Social Networks**, 28: 124-136.

Brass, Daniel J., & Burkhardt, Marlene E. (1993). Potential power and power use: An investigation of structure and behavior. **Academy of Management Journal**, 36: 441-470.

Burt, R. S. 2001. Bandwidth and echo: Trust, information, and gossip in social networks. In James E. Rauch & Alessandra Casella (Eds.), **Networks and markets**: 30-74. NY: Russell Sage.

Campbell, M. K., & Torgerson, D. J. 1999. Bootstrapping: Estimating confidence intervals for cost-effectiveness ratios**. Quarterly Journal of Medicine,** 92:177-182.

Carley, K.. Columbus, D., DeReno, M., Reminga, J. & Moon, I-L. 2008. **ORA user's guide 2008.** Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISR-08-125.

Cross, R., & Cummings, J. N. (2004). Time and network correlates of individual performance in knowledge-intensive work. **Academy of Management Journal**, 47: 928-937.

Efron, B. 1987. Better bootstrap confidence intervals. **Journal of the American Statistical Association**, 82(397), 171-185.

Efron, B. 1994. Missing data, imputation, and the bootstrap. **Journal of the American Statistical Association**, 89(426), 463-475.

Frantz, T., & Carley, K. 2005. **An automated methodology for conducting a social network study of a university faculty.** Carnegie Mellon University, School of Computer Science (SCS), Institute for Software Research International (ISRI), Center for Computational Analysis of Social and Organizational Systems (CASOS) - Technical Report CMU-ISRI-05-106.

Frantz, T., & Carley, K. 2005. **Relating network topology to the robustness of centrality measures.** Carnegie Mellon University, School of Computer Science (SCS), Institute for Software Research International (ISRI), Center for Computational Analysis of Social and Organizational Systems (CASOS) - Technical Report CMU-ISRI-05-117.

Frantz, T., Cataldo, M., & Carley, K. 2006. **Social network data, given error: Evidence for the construction of confidence intervals around network measures.** Presented at the International Network for Social Network Analysis, Sunbelt XXVI., Vancouver, British Columbia, Canada, April 25-30.

Frantz, T., Cataldo, M., & Carley, K. 2008**. Measure robustness under uncertainty: Topology matters too.** (Under review).

Freeman, L. C. 1979. Centrality in social networks: Conceptual clarification. **Social Networks,** 1: 215-239.

Ibarra, H., 1993. Network centrality, power, and innovation involvement: Determinants of technical and administrative roles. **Academy of Management Journal,** 36: 471-501.

Ibarra, H., & Andrews, S. B. 1993. Power, Social Influence, and Sense Making: Effects of Network Centrality and Proximity on Employee Perceptions. **Administrative Science Quarterly,** 38: 277-303.

Klein, K. J., Lim, B-C., Saltz, J. L, & Mayer, D. M. 2004. How did they get there? An examination of the antecedents of centrality in team networks. **Academy of Management Journal,** 47: 952-963.

Kossinets, G. 2006. Effects of missing data in social networks. **Social Networks**, 28: 247-268.

McCulloh, I., Ring, B., Frantz, T., & Carley, K. 2008. **Unobtrusive social network data from email.** Presented at the 26th Army Science Conference, Orlando, FL, USA, 1-4 December.

Rubin, Donald B. (1976). Inference and missing data. **Biometrica**, 63: 581-592.

Sparrowe, R. T., Liden, R. C., Wayne, S. J., & Kraimer, M. L. 2001. Social networks and the performance of individuals and groups. **Academy of Management Journal**, 44: 316-325.

Walling, G. A., Visscher, P. M., & Haley, C. S. 1998. A comparison of bootstrap methods to construct confidence intervals in QTL mapping. **Genetic Research**, 71: 171-180.

Wittek, R., & Wielers, R.. 1998. Gossip in Organizations. **Computational & Mathematical Organization Theory**, 4:189–204.

TABLE 1

Confidence if one link missing for hierarchy example

Missing Link:	AB	AC	AD	AE
Confidence Estimate (Frequency probability %)	100%	100%	100%	100%

TABLE 2

Confidence estimate across error levels for hierarchy example

Number of Missing Links	Data Error Level	Derived Top-Actor (s)	Confidence Estimate (Frequency Probability %)
0	0%	A	100%
1	25%	A	100%
2	50%	A	100%
3	75%	A or the sole alter	100%
4	100%	A or any other	100%

TABLE 3

Remove one link – Frequency probability

Removed link:	AB	AC	BD	CD	DE
Fraction of time correct	1/1	1/1	1/3	1/3	1/4
Confidence Estimate: (Frequency probability %)	100%	100%	100%	100%	100%

TABLE 4

Remove two links – Fraction of time correct

Removed links	AB	AC	BD	CD	DE
AB					
AC	1/1				
BD	1/1	1/2			
CD	1/2	1/1	0		
DE	1/2	1/1	0	0	

TABLE 5

Remove two links – Frequency probability %

Removed links	AB	AC	BD	CD	DE
AB					
AC	100%				
BD	100%	100%			
CD	100%	100%	0%		
DE	100%	100%	0%	0%	

TABLE 6

Confidence estimates for 5-actor kite network

Centrality Measure	Estimated Error Level (%)				
	20	40	60	80	100
Degree	1.00	0.70	0.72	0.60	1.00
Betweenness	0.61	0.70	0.72	1.00	1.00
Closeness	0.61	0.70	0.72	0.60	1.00
Eigenvector	0.61	0.70	0.29	0.60	0.00

TABLE 7

Confidence estimate for hypothetical network- Degree centrality; scale-free versus uniform random

Error Level (%)	Scale-Free Topology	Uniform Random Topology
5	0.91	0.85
10	0.86	0.77
15	0.79	0.69
20	0.75	0.60
25	0.71	0.53
30	0.68	0.49
35	0.64	0.48
40	0.60	0.39
45	0.58	0.35
50	0.54	0.33

FIGURE 1

Hierarchical (star) network

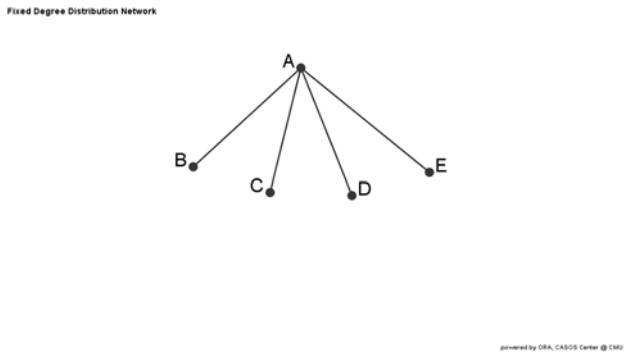

FIGURE 2

Kite network

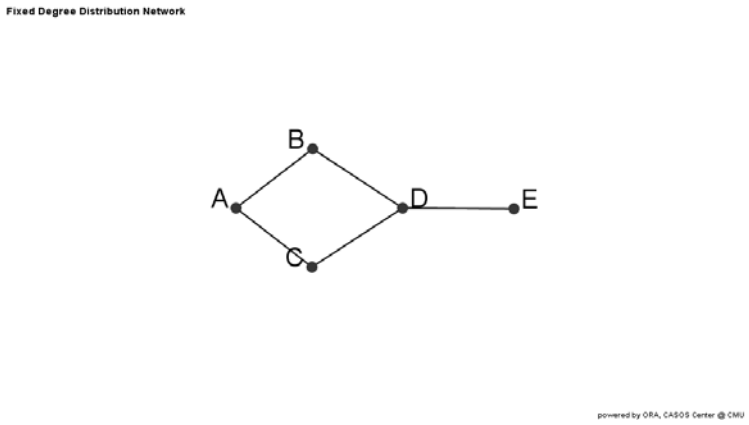

FIGURE 3

Node-Link diagram of the hypothetical social network

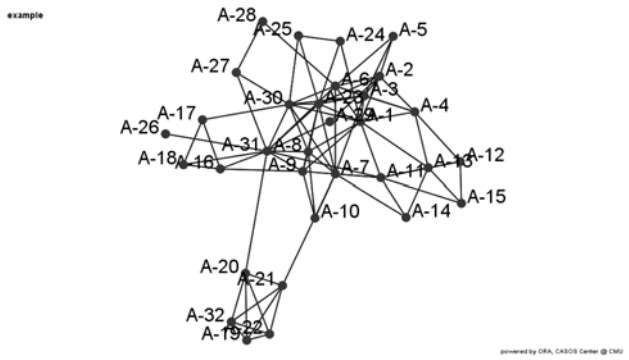

FIGURE 4

Confidence estimate for hypothetical social network – Degree centrality

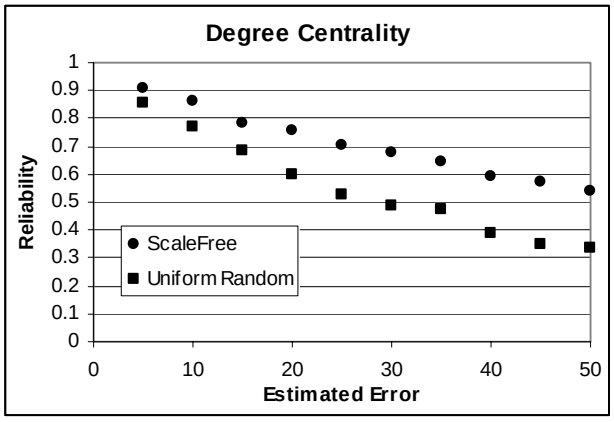

FIGURE 5

Confidence estimate for hypothetical social network – Betweenness centrality

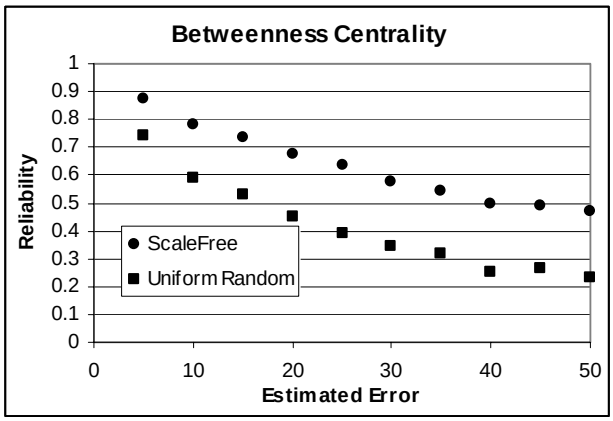

FIGURE 6

Confidence estimate for hypothetical social network – Closeness centrality

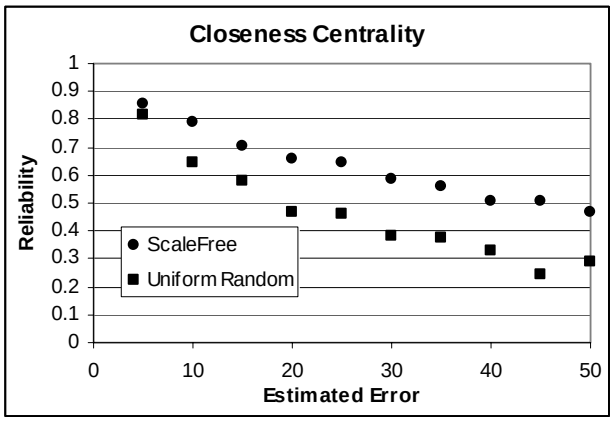

FIGURE 7

Confidence estimate for hypothetical social network – Eigenvector centrality