Comput Math Organ Theory (2009) 15: 303–328
DOI 10.1007/s10588-009-9063-5

Robustness of centrality measures under uncertainty: Examining the role of network topology

Terrill L. Frantz · Marcelo Cataldo · Kathleen M. Carley

Published online: 15 December 2009

Abstract This study investigates the topological form of a network and its impact on the uncertainty entrenched in descriptive measures computed from observed social network data, given ubiquitous data-error. We investigate what influence a network's topology, in conjunction with the type and amount of error, has on the ability of a measure, derived from observed data, to correctly approximate the same of the ground-truth network. By way of a controlled experiment, we reveal the differing effect that observation error has on measures of centrality and local clustering across several network topologies: uniform random, small-world, core-periphery, scale-free, and cellular. Beyond what is already known about the impact of data uncertainty, we found that the topology of a social network is, indeed, germane to the accuracy of these measures. In particular, our experiments show that the accuracy of identifying the prestigious, or key, actors in a network—according observed data—is considerably predisposed by the topology of the ground-truth network.

Keywords Network topology · Data error · Measure robustness · Centrality · Observation error

T.L. Frantz (✉) · K.M. Carley
Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA
e-mail: terrill@org-sim.com

K.M. Carley
e-mail: kathleen.carley@cs.cmu.edu

M. Cataldo
Two North Shore Center, Suite 320, Pittsburgh, PA 15212, USA
e-mail: Marcelo.Cataldo@us.bosch.com

1 Introduction

Social network datasets are often incomplete and prone to observation error due to the intricacy of collection-instrument design and the inherent vagueness of human-informant reliability and bias (Stork and Richards 1992; Feld and Carter 2002). The error in the observed data may be unintentional or intentional (Albert et al. 2000; Carley 2002; Calloway and Morrissey 1993; Freeman et al. 1987; Killworth and Bernard 1976). No matter its nature, the presence of this error is a nontrivial issue (Marsden 1990; McKnight et al. 2007) and it raises the question of the impact of the uncertainty, relative to the accuracy of network measures computed from this data. In extensive efforts to ameliorate this problem, researchers have been examining and estimating the impact of observation error and exploring the reliability of descriptive network measures given the confines of errant data (e.g., Borgatti et al. 2006; Costenbader and Valente 2003; Kossinets 2006; Marsden 1993; Zemljic and Hlebec 2005).

Past research taking the functional approach[1] to the problem of uncertain data has primarily consisted of two complementary strategies; although, there are certainly several other strategies (see Butts 2003; Robins et al. 2004). In the first strategy, researchers have used a variety of sampling techniques on the observed data (Erickson and Nosanchuk 1983; Frank 1978, 1981; Galaskiewicz 1991; Gile and Handcock 2006; Granovetter 1976; Handcock and Gile 2007). This strategy has been found to provide "reasonable, if not excellent" (Galaskiewicz 1991, p. 347) estimates of the bona fide centrality measures for the true network. In the Galaskiewicz study, it was shown that the lower the density of the network, the better the estimates of centrality were, while network size did not appear to be an important factor. In another study, Costenbader and Valente (2003) showed that most of the centrality measures were sensitive to sampling size, network size and network density; however, the measures remained robust for sampling levels between 50% and 80%. Further, they found that eigenvector centrality was considerably more robust relative to the other measures.

A second strategy used to estimate robustness of network measures, involves controlled experiments that statistically analyze computer-generated network data, i.e., virtual experiments. Bolland (1998) uncovered considerable redundancy in centrality measures under the condition of data uncertainty and found that such redundancy actually increases with the level of error. In a study that explored both generated and real-world networks, Kossinets (2006) found that observation error involving missing nodes can significantly alter network-level statistics such as average degree centrality, clustering, and a variety of other descriptive measures. In another controlled experiment, Borgatti et al. (2006) explored measure robustness specific to case of uniform-random networks. They found that the common measures of centrality have a similar pattern of robustness across network size and density, and that the different type of errors (missing or superfluous; nodes and edges) have surprisingly similar robustness profiles. Recently, Kim and Jeong (2007) found that in certain networks, closeness was generally the most robust of the basic centrality measures.

[1] Much research investigating missing or over sampled data has taken the structural approach, which focuses on *what* the incorrect data might be, whereas the functional approach focuses on the *impact* of the incorrect data and *what to do* about it (McKnight et al. 2007).

Concurrent to the mounting research on the robustness of network measures, there has also been growing interest in understanding the characteristics of the more holistic aspects of networks, and in particular, the topological aspects of a social network (e.g., Newman et al. 2002). The mathematically elegant Erdös and Rényi (1959) network topology, which we refer herein as the *uniform random* network, has been well studied by social scientists for many years. Recently, attention has shifted toward understanding the more complex, stylized networks—networks with a more intricately defined topology—to the point that some related terms have even entered the populous vernacular. One topology, the *small-world* (Milgram 1967; Newman and Watts 1999), has been found in numerous social settings (Davis et al. 2003; Watts 1999a), such as in the networks of film actors (Watts and Strogatz 1998) and computer discussion groups (Ravid and Rafaeli 2004), and is a notion that has even been the basis for a popular play (Guare 1990). In other cases, such as business networks (Powell et al. 2005), the Internet and other communication networks, it has been found that they have a *scale-free* topology (Albert et al. 1999; Faloutsos et al. 1999), likely as a result of the social phenomenon of preferential attachment (Simon 1955). As Borgatti and Everett (1999) reported, the *core-periphery* topology has been found in areas such as systems, economics, collective actions, interlocking directorates, and other organizational areas, e.g., inter-organizational alliances (Stuart and Robinson 2000). Lastly, the *cellular* topology (Frantz and Carley 2005; Krebs 2002; Mayntz 2004; Tsvetovat and Carley 2005) is a socially-constructed network often associated with covert organizations such a terrorist groups and is a topology that is frequently associated with the "dark side" (Ronfeldt and Arquilla 2001, p. 2). Since the events of 9/11, the cellular form has garnered a great deal of academic and government interest.

In this study, we conjoin the interest in the robustness of network measures with that of network topology, and conjecture that the topology of the network has critical relevance to the robustness profiles of descriptive social-network measures. In the formal sense, we scrutinize the following conjecture: *in the ubiquitous circumstance of data uncertainty, the topological form of the true network has a measurable effect on the robustness of node-level measures when computed from the observed network data, relative to the analogous values computed from the ground-truth network.*

By undertaking a combined perspective and endeavoring to conduct an experiment involving both the robustness of network measures and network topology, social scientists and practitioners will improve their understanding of the impact that observation error has on network measures and, thus, will move closer to conceiving a remedy to the inherent data-uncertainty problem. We explore the robustness of network measures via a virtual experiment in a manner akin to that of Borgatti et al. (2006), but with a poignant focus on the topology of the network; specifically, we handle network topology of the true social-network as an independent variable in our experiments. We concentrate on the five topologies mentioned above, as we examine the robustness of four traditional node-level, centrality measures: *degree* (Freeman 1979), *betweenness* (Freeman 1977), *closeness* (Freeman 1979), *eigenvec-*

tor (Bonacich 1987). We also examine the *local clustering*[2] (Watts and Strogatz 1998; Watts 1999b) measure because we suspect that it may be relevant to future social network studies that are oriented toward topology-related classification.

The remainder of the paper is organized as follows: we detail the methodology of the experiment, present the quantitative results, qualitatively discuss the findings, and present our conclusions. Finally, the limitations of the study and suggested future research directions are presented.

2 Method

The aim of this experiment is to examine the relationship of a network's ground-truth topology to the robustness of network measures that are derived from observed data. In order to accomplish this, we designate the topology of a *true-network* as an independent control variable of primary focus; other independent control variables include the size and density of the true network as well as specific characteristics of error approximating the uncertainty, resulting in an *observed-network*. Based on their involvement in the process, the control variables are organized into two groups: the network class and the uncertainty class. The network class variables specify the characteristics of the true network and are used in the generation of the true-network data. The uncertainty class variables are those that specify the characteristics of the data error, i.e., the error in the observed network data and are used in transforming the ground-truth data to an observed dataset. For response outcome variables, we focus on observed-network-based metrics determined from rank-ordered lists of five widely-used node-level measures, namely: degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and local clustering. These measures will be further expressed using two types of metrics: top-node and top-group; detail about these outcome variables is provided later in this section (see Sect. 2.3).

This experiment has an amalgamated factorial block and randomized design: four of the control variables are non-probabilistic, tiered samples constructed in a factorial and purposeful fashion, and one control variable has a combined quota and purposeful sampling methodology. Considering the five control variables, as shown in Table 1, some topological forms are infeasible at some network densities, so some experiment cell expectations are not fully satisfied. For example, a cellular network generated with a density of 0.70 is not possible to achieve, thus there are no samples of cellular topologies at this density level; while for other topologies, such as uniform random, density at this level is realizable.

Procedurally, for each replication, we start by generating an undirected graph of the prearranged topology, size (number of nodes) and of a specific density (number of edges), and then label it the *true* network, representing what the absolute truth network scenario is to be. Next, an exact copy of the true network is systematically perturbed according to the prearranged error class variables (type and level), resulting

[2]Local clustering is a measure of how well connected the neighbors of a specific node are. The exact definition is: the count of the number of existing ties among an ego's alters, divided by the count of the number of ties possible among the same alter nodes; its resulting value will be between 0 and 1.

Table 1 Parameter space of control variables

Variable	Number/nature of values	Values
Network class: (used in generation of true network)		
Network size[a]	3	25, 50, 100
Network density	Range	0.01–0.90
Topology	5	Uniform-random, small-world, scale-free, core-periphery, cellular
Uncertainty class: (used in transformation to observed network)		
Error type	4	Edge-remove, edge-add, node-remove, node-add
Error level[b]	5	1%, 5%, 10%, 25%, 50%
Replications[c]	Range	10–250

[a]Number of nodes

[b]Percentage of nodes or edges to be perturbed, according to the error type

[c]Actual number of replication per cell varies due to stochastic nature of the network generation algorithms

in a transformation to a separate, undirected graph, which is labeled, the *observed* network. This observed network represents what uncertain data the researcher in point of fact collects; therefore, in the real-world, it is this data that is actually used to compute measures that are reported. It follows that then each experimental replication is represented by a pair of networks composed of a *true* network and its corresponding *observed* network.

For each network in the pair, we compute the five node-level measures for each individual node in the network, and then rank order the nodes according to their measure value in decreasing order. This results in five node-lists for each network and therefore, five paired node-lists; each pair consists of one rank-ordered list for the true network and one for the observed network. Using the paired node-list for each measure, several metrics are computed using the lists that indicate the congruence of the two lists, resulting in the ability to quantify and evaluate the discrepancy between the ground-truth values and the observed values.

For the remainder of this section, we describe how the true networks are generated specific to their topology, how uncertainty is introduced in the corresponding observed network, and how the robustness metrics are determined.

2.1 Generating the true networks

In order to independently generate each true network, we applied an algorithm designed to produce a plausible approximation of the particular topology; accordingly, we have five generating-algorithms, each unique to a specific topology. Each algorithm takes as one of its input parameters the size of the desired network (number of nodes), whose value is set as specified by the network size variable. The properties of each algorithm restricts the implementation of a consistent network density specification across all the algorithms. In some cases density is a direct input parameter and results in networks of that exact density, e.g., uniform random and core-periphery,

while in other cases, e.g., small-world, scale-free and cellular, the density of the actual network generated is approximately the value of the desired density. The generative algorithms have been reviewed and the density of the networks produced is sufficiently near the desired density value. The details and the parameter settings for each of these topologies are described below.

To generate a *uniform random* network, we used the method to generate an Erdös Random Graph (Bollobás 2001) as implemented in the Organizational Risk Analyzer (ORA) software. We selected this generator over equivalent others both because it was used in the Borgatti et al. (2006) experiment—therefore our experimental procedures can be straightforwardly authenticated—and because its parameter requirements fit precisely with the experiment's variables. The algorithm requires two input parameters: number-of-nodes, and number-of-edges. We simply compute the value for the number-of-edges parameter based on the density and network size variables. To generate the graph, the algorithm randomly selects pairs of vertices (without replacement) from the set of all possible and joins the pair, until the number-of-edges parameter has been satisfied. Each pair has an equal probability of being selected. The probability value is the inverse of the number of nodes, squared. Any resulting self-loops are discarded. This algorithm produces a graph with the network density value matching exactly the density parameter and variable.

To generate a *small-world* network, we used an edge-add (as opposed re-wire) approach described as TOPOLOGY 3 (Small-world) in Airoldi and Carley (2005). However, we modify the original procedure slightly by converting the graph's directed edges to be undirected and by disabling some of the algorithm's advanced features by effectively nullifying several inessential parameters. Subsequently, the reconfigured algorithm calls for three parameters: number-of-nodes, number-of-edges, and distance-of-close-neighbors. The algorithm first constructs a ring lattice with the specified number-of-nodes set to the variable network size value, and the distance-of-close-neighbors parameter set to one value in {1, 2, 3, 4, 5, 8, 10, 20}; this value is bounded by what is possible given the number of nodes in the graph. This will generate a multitude of fixed ring-lattices with a pre-determined number of edges;[3] thus, this is also a graph with a pre-determined density. Next, edges are added randomly until the desired density, according to the density variable, is reached. Generated networks are then assigned into the appropriate experiment cell according to value of the actual density of the true network.

To generate a *scale-free* network, we used the algorithm, SCALE FREE 2, proposed by Airoldi and Carley (2005), which applies a power law with preferential attachment technique, as proposed by Barabasi and Albert (1999), but allows for generating graphs with a specified and fixed number of nodes and edges. The algorithm requires three parameters: number-of-nodes, number-of-edges, and power-law-exponent. Like other algorithms, we simply compute the value for the number-of-edges parameter based on the density and network size variables. For all graphs of

[3]For example, a minimal lattice of 10 nodes, connected to 1 neighbor, has a fixed density of 0.11, therefore it is impossible to have a small-world network of this node-size with density of 0.01, 0.02, etc. The experiment cells for these impossibilities are left unfilled. The 100-node lattice provides a full range of densities to fill the range of experiment cells for the small-world topology, accordingly.

this topology, we arbitrarily fix the power-law-exponent parameter to 2.0 because this rounded value approximates the exponent of scale-free networks empirically found in the real-world, e.g. the Internet. The algorithm constructs an ordered degree distribution vector of normalized probabilities based on the sequential order of the node, the total number of nodes, and power-law-exponent parameter. The normalized probability is then computed by dividing the number of ties for the node (the current vector value) by the total number of ties in the desired graph (based on the density desired). Then the ties-creation processes based on the vector of probabilities is repeated until all edges have been created as proscribed; that is, we randomly select a node pair and add an edge or not between the pair with probability according to the distribution probabilities in the vector.

To generate a *core-periphery* network, we used an algorithm, conceived and operationalized by Borgatti and Reminga (2005), that is part of the Organizational Risk Analyzer (ORA) software, which reproduces the formalization described by Borgatti and Everett (1999). The algorithm requires three parameters: number-of-nodes, network-density, and power-law-exponent. As was done with the scale-free topology generation process, we fixed the power-law-exponent to 2.0. The algorithm first creates a node-degree distribution vector in the same manner as the scale-free algorithm, i.e., constructing an ordered degree distribution vector of normalized probabilities based on the sequential order of the node, the total number of nodes, and power-law-exponent parameter. From this vector, a two-dimension matrix of all possible edge combinations is then formed, where each cell in the matrix is then filled with a score value that is the product of the values from degree distribution vector for the two paired nodes. Potential edges for the graph are then randomly selected from cells in this matrix. The value of the selected cell is then used as an-exists probability until the number of edges called for in the network construction is reached.

To generate a *cellular* network, we used the algorithm, CELLULAR 2, proposed by Airoldi and Carley (2005). The algorithm requires five parameters: number-of-nodes, number-of-cells, probability-of-edges-within-a-cell, probability-of edges-between-cells, and a power-law-exponent. The number-of-cells parameter is determined by rotation though a value in {0.2 0.4, 0.6, 0.8} that is multiplied by the number-of-nodes and rounded, to establish the value for the number-of-cells parameter, resulting in {2, 4, 6, 8} cells in the case of a 100-vertex graph, and {1, 2} cells in the case of a 10-vertex graph. Both the probability-of-edges-within-a-cell and the probability-of edges-between-cells parameters are each separately determined by rotation though the values in {0.01, 0.02, 0.05, 0.10, 0.30, 0.50, 0.70, 0.90}. As with the scale-free and core-periphery topology generation processes, we fix the power-law-exponent to 2.0. Each generated network is then assigned into the appropriate experiment cell according to value of the actual density of the true network.

2.2 Introducing uncertainty

The data for the observed network is generated by perturbing a copy of a previously-generated true network. The uncertainty in the observed data is modeled using a random process that perturbs the true network according to the error class variables (error type and error level) of the experiment cell in which the network pair is a member of.

Table 2 Description of error types

Error type	Process	Uncertainty model
Edge		
Remove	From set of all existing, edges are uniformly, randomly selected for removal until level of error is achieved.	Models the uncertainty of ties being erroneously unreported in the observed network.
Add	From the set of all non-adjacent node pairs, a pair is selected and an edge added between them until the level of error is achieved.	Models the uncertainty of non-existing ties being erroneously reported in the observed network.
Node		
Remove	From set of all existing, nodes are uniformly, randomly selected for removal until level of error is achieved. All edges incident to the removed nodes are also removed.	Models the uncertainty of nodes and their incident edges being erroneously unreported in the observed network.
Add	Nodes are added until the level of error is achieved. In addition, edges are added incident to each new node to join the node to any other non-adjacent node. The number of edges added to each new node is determined by randomly selecting a node from the original set and matching its degree value.	Models the uncertainty of non-existing nodes and their non-existing incident edges being erroneously reported in the observed network.

The error type is one of {edge-remove, edge-add, node-remove, node-add}. The error level is a specific percentage value from {1%, 5%, 10%, 25%, 50%}, that is relative to the number of edges or nodes in the original true network according to the error type variable. For example edge-remove at the 10% error level for a 10-node, 10% density network, would result in 5 edges being removed; in the same scenario for edge-add, 5 edges would be inserted between 5 sets of two randomly selected nonadjacent nodes. Essentially, the change in edges is equivalent to the absolute-value of the hamming distance between the paired, true and the observed networks.

For error types involving nodes, the error level is related to the number of nodes in the true network and also necessarily involves those edges directly related to the affected nodes. With node-remove, all edges incident to each removed node are also removed. In the case of node-add, each edge is added connecting the new node and a random other node. The number of edges added to each new node is matched to the degree of a randomly selected other, pre-existing node in the same true network.

The four error types each model a specific type of uncertainty found in observational data. Edge-remove, which may be the most common type of uncertainty experienced in real-world studies, situates the observed data for missing edge information. Edge-add models extraneous edges being reported in the observed network data that do not exist in truth. Node-remove replicates the missing node, and thus the additional adjacent edge situation. Node-add models the situation whereas false nodes and adjacent ties are included in the observed network data, while in truth they do not exist. Table 2 summarizes the description of the error type and their application and implication to the uncertainty characteristic in the observed network data.

2.3 Computing measure accuracy

Each distinct network-pair has several statistics associated with it that indicate measure agreement from a variety of standpoints, i.e., its exact accuracy from various perspectives at several levels. There are 20 separate statistics computed for each distinct network-pair. The assemblage of these statistics can be envisaged as a two-dimension table: across one dimension are the five network measures, e.g., degree centrality; across the other dimension are four specially-crafted metrics.[4] We conceptualize these metrics into two groups, labeled: top-node and top-group. Each metric is delineated according to how it is derived; this, therefore reflects the level of its expressiveness towards a measure's accuracy. The top-node metrics are named: top-1, top-3, and top-10%. The value assigned to each metric is the result of a comparison between a rank-ordered list of nodes obtained from the true network with a similarly constructed list obtained from the observed network. The metrics are so named according to the size of a rank-ordered list of nodes obtained from the observed network. The resulting value indicates a quantity reflecting the level of agreement between the paired true and observed networks; thus, the value for each statistic indicates a given measure's congruence, thus accuracy, at a different level of strictness. The top-node metrics are indicative of how uncertainty affects the reliability of a measure to accurately identify the individual node that truly has the highest value for that measure, e.g., the top-ranked node according to degree centrality is the one note with the highest value of all nodes in the network. Essentially these indicate the agreement, or not, of The top-node metrics are particularly useful with respect to the key player question (Borgatti 2006), identifying nodes of relative importance (White and Smyth 2003), and identifying network elites (Burt 1978; Masuda and Konno 2006), but can also be quantitatively indicative of measure robustness as established by Borgatti et al. (2006).

To determine the value of the various metrics in the top-node group, we dichotomize (a binary value, either 1 or 0, for true and false, respectively) a comparison of the membership between the two rank-based lists: *Top 1* is an indication that the top-rank node in the true network is also the top-ranked node in the observed; *Top 3* is an indication that the top-rank node in the true network is one of the top 3 nodes in the observed network; and *Top 10%* is an indication that the top-rank node in the true network is also ranked in the top 10% (relative to number of nodes in the true network) in the observed network.

The metric in the top-group is overlap, which is a value indicative of how uncertainty affects the reliability of a measure to accurately identify the top set of nodes in

[4]Herein, we adopted some of the same accuracy metrics as those crafted by Borgatti et al. (2006), so as to ease the feasibility of conjoining our result and because of inherent limitations of the traditional statistical approaches. Indeed, other more traditional statistical approaches for evaluating rank data are available, e.g., Spearman *rho* and Kendall's *tau*, but are not utilized here due to their constraint of requiring squared data. In this study, there are situations when the two rank-ordered lists may not be complete, i.e., in the case of node-remove error type. It should be noted that some work has been done on handling non-square data in this realm (see Papaioannou and Loukas 1984; Alvo and Cabilio 1995), but we purposefully opt to be in harmony with Borgatti et al. For an experiment that assessed the reliability of complete rank-ordered lists under data uncertainty, we suggest reading Kim and Jeong (2007).

Table 3 Metrics computation step 1: Determine true and observed network ordered rankings

Ranking	True network		Observed network	
	Node identifier = *Set A*	Node-level measure value	Node identifier = *Set B*	Node-level measure value
1	node15	0.561	node13	0.485
2	node13	0.511	node15	0.478
3	node22	0.482	node25	0.451
4	node25	0.482	node28	0.410
5	node28	0.455	node56	0.410
6	node31	0.332	node02	0.400
7	node56	0.173	node14	0.396
8	node02	0.152	node17	0.373
9	node14	0.149	node19	0.104
10	node17	0.113	node58	0.098
etc.	etc.	etc.	etc.	etc.

a network according to the measure; whereas, the number of nodes in this top-ranked set is determined relatively by taking 10% according to the number of nodes in the true network. The sole top-group metric, *overlap*, is a value that reflects the broader extent to which the measures are robust; it evaluates the how well the set of top-10 percent nodes in the true network match the set of top-10 percent nodes in the observed network. The overlap metric is a similarity metric, in the form of the Jaccard's coefficient between the two node sets, which provides an index between 0 and 1, such that 0 indicates the extreme circumstance that there is no overlap and 1 indicates strict agreement.

A synopsis of the sequence of steps for determining the values for these metrics is illustrated via Tables 3 through 6; these tables show the details of computing the metrics for one example network-pair case. For clarity, we will use standard set notation throughout this process description by referring to specific sets of individual nodes using capitalized alphabetic characters. Step 1 (Table 3) requires separately rank-ordering the individual nodes from each network in the specific network-pair according to their value based on the particular measure, e.g., degree centrality. To determine the rank-order of the nodes, from highest-to-lowest value for the particular measure, we use the ordinal ranking approach; the node internal-identification number arbitrarily breaking value-based ties. The resulting rank-ordered list from the true network we label, A, and the list from the observed network is labeled B.

Table 4 (step 2) is an illustration of how subsets of A and B are constructed. From A, we identify the top node from the true network and label the unity-member C. C is an instrumental aspect of the majority of metrics as it represents the most prestigious, or key player that we want to be able to locate in the observed network at the same most prestigious position. A set of varied length across the different samples of network sizes is constructed and labeled D. This set has its number of members set to 10% of the number of nodes in the true network. For example, a 100 node network would have 10 elements in D. With respect to the observed network, three subsets

Table 4 Metrics computation step 2: Construct the sets for top-node metrics

Count	True network (Set A)		Observed network (Set B)		
	Top 1 = *Set C*	Top 10% = *Set D*	Top 1 = *Set E*	Top 3 = *Set F*	Top 10% = *Set G*
1	node15	node15	node13	node13	node13
2		node13		node15	node15
3		node22		node25	node25
4		node25			node28
5		node28			node56
6		node31			node02
7		node56			node14
8		node02			node17
9		node14			node19
10		node17			node58

Table 5 Metrics computation step 3: Construct the sets for top-group metric

Count	Intersection $A \cap B$ = *Set H*	Union $A \cup B$ = *Set I*
1	node02	node02
2	node13	node13
3	node14	node14
4	node15	node15
5	node17	node17
6	node25	node19
7	node28	node22
8	node56	node25
9		node28
10		node31
11		node56
12		node58

are constructed from B. The top node makes up E, and the top 3 nodes make up F. G is constructed in the same manner as D, but using B as its source.

Table 5 (step 3) is an illustration of how traditional intersection and a union sets are constructed from both the true (A) and observed (B) rank-ordered lists using the traditional set-algebraic techniques. This results in two distinct subsets, H and I, from the true and observed networks, respectively.

Table 6 (step 4) is an illustration of the final step in the process which results in the top-node and the top-group metrics. The value assigned to Top 1, a binary value, is a truth value that specifies the subset agreement between C and E, and indicates whether or not the top 1 node in the true network is also the top 1 node in the observed network. The value assigned to Top 3, a binary value, is a truth value that specifies the subset agreement between C and F, and indicates whether or not the top 1 node in the true network is found within the top 3 nodes in the observed network. The

Table 6 Metrics computation step 4. Evaluate set algebras and computation

Metric	Determination	Value
Top 1	$\begin{cases} 1 & \text{if } C \in E \\ 0 & \text{otherwise} \end{cases}$	0
Top 3	$\begin{cases} 1 & \text{if } C \in F \\ 0 & \text{otherwise} \end{cases}$	1
Top 10%	$\begin{cases} 1 & \text{if } C \in G \\ 0 & \text{otherwise} \end{cases}$	1
Overlap	$\lvert H \rvert / \lvert I \rvert$	$8/12 = 0.750$

Table 7 Example of independent control variables for a single network-pair

Name	Value
Topology	Small-world
Node size	50
Density	0.10
Error level	20%
Error type	Edge-remove

Table 8 Example of dependent outcome variables for a single network-pair

Name	Top 1	Top 3	Top 10%	Overlap
Degree centrality	0	1	1	0.468
Betweenness centrality	1	1	1	0.451
Closeness centrality	0	0	1	0.334
Eigenvector centrality	1	1	1	0.486
Local clustering	0	1	1	0.400

value assigned to Top 10%, a binary value, is a truth value that specifies the subset agreement between C and G, and indicates whether or not the top 1 node in the true network is found within the top 10% nodes in the observed network. To compute the overlap value, the count of the number of members of H is divided by I. Recall, this produces a ratio between zero and unity that reflects the amount of overlap between the true and the observed rank-ordered lists. A value of 0 indicated no overlap and a value of 1 indicates complete overlap, i.e., perfect accuracy between the true and observed measurements.

To recapitulate the outcome of the above steps in a complete summarization for a single network-pair, Tables 7 and 8 show an all-inclusive set of the variables and their values for an arbitrary, example; Table 7 shows the complete independent control variables and Table 8 shows the complete set of dependent outcome variables.

3 Results

We organize the discussion of the results in three sub-sections. First, we report on our preliminary analysis of the sample data that were produced by the generative

processes described in the Method section. Next, we report on our statistical analysis of the effect of the network topology on the accuracy of top-node outcome metrics. Lastly, we report on the statistical effects of network topology on the accuracy of top-group outcome metrics.

3.1 Preliminary analysis of sample data

The sample-data generation process yielded 622,719 independent network-pairs; each whose true-network was constructed by one of the topology-specific generative algorithms, and a copy then perturbed according to specific error characteristics. The true-networks were constructed of fixed network size of 25, 50 or 100 nodes and were effectively drawn from a range of possible densities according to the precincts of the stylized topology. The uncertainty in the observed network was introduced in the form of a specific error type and level, e.g., edge-remove, 10%.

The distribution of the resulting data is reported in Table 9 in the form of a contingency table, which presents the count of independent samples according to the topology and the error type; this provides a general sense of the actual outcome of the generation process. The average number of replications for each distinct experimental cell (as per all control variables, i.e., both the network and uncertainty classes) is 177 independently constructed network-pairs. Since the high-level accuracy measures are either a probability or an average, experimental cell combinations with less than 10 replications were excluded from the analysis and are therefore excluded from this 622,719 count: this is specific to the network density as the actual value of density is topology-dependent according to the properties of the network generation algorithm.

In order to provide a general sense of the robustness across the various topologies, we present three graphs (Figs. 1–3) that illustrate a characteristic profile according to control variables: network size, network density, and error level. Each graph shows a very small cross-section of the sample data, which is a representation of the variable's profile more generally; however, in actuality, the details of the profiles can vary somewhat across the various cross-sections of the data. To keep the presentation simple, we report only on the degree centrality, top 3 metric in these figures.

Figure 1 shows an accuracy profile of network size. The graphic indicates that network size (number of nodes) can be described, by and large, as having a near-zero slope and that the topologies are parallel, but are somewhat offset from one another. This suggests, from a visual perspective, that network size may have relatively minor influence on the robustness of the measures.

Table 9 Sample sizes by topology and error type (number of independent network-pairs)

Topology	Edge		Node		Total
	Remove	Add	Remove	Add	
Uniform	30,000	30,000	30,000	30,000	120,000
Small world	29,649	29,645	29,658	29,667	118,619
Core periphery	30,000	30,000	30,000	30,000	120,000
Cellular	36,103	36,071	36,061	35,998	144,233
Scale free	29,958	29,979	29,996	29,964	119,867
Totals	155,710	155,695	155,685	155,629	622,719

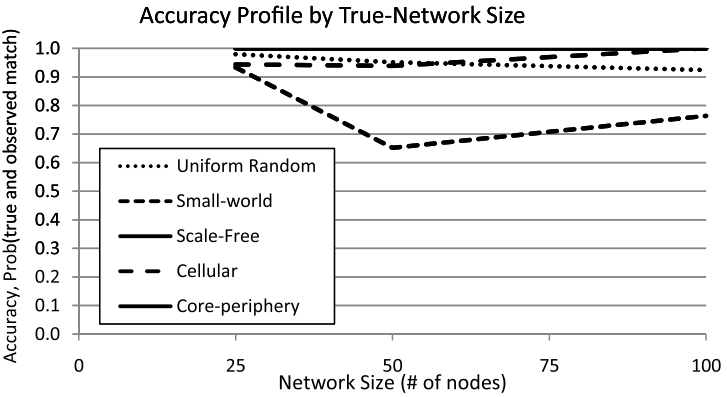

Fig. 1 Accuracy profile for topologies by network size (degree centrality, top 3 metric; density 10%, edge remove, error level 10%)

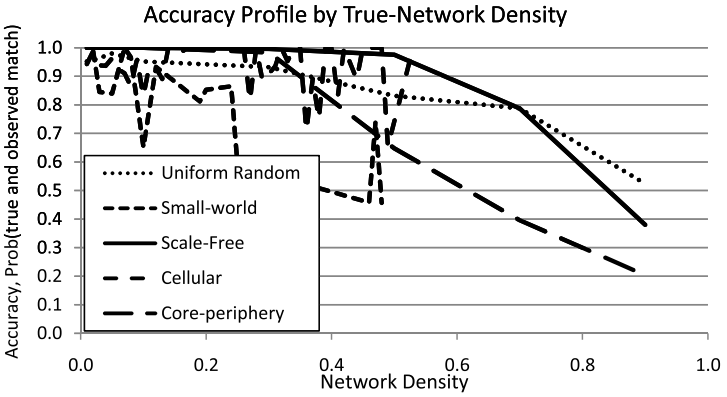

Fig. 2 Accuracy profile for topologies by network density (degree centrality, top 3 metric; size 50, edge remove, error level 10%)

Figure 2 shows an accuracy profile of network density. The graphic indicates that network density can be described, by and large, as having mixed slopes at the lower density values, but becomes more consistent with even greater prominence as density increases. The topologies profiles are quite different at lower densities. This suggests, from a visual perspective, that network density may have a material effect on the robustness and that the topology has even more influence than the density per se. In this graphic, the limitations of the network generation algorithms and topological characteristics become quite apparent, as discussed in the Method section. Specifically, the accuracy for small-world has more variability due to data points being available at each percentage unit of density, e.g., data for 20%, 21%, 22% density, etc., and with differing sample sizes at each density; whereas, the sample-set for the small-world has densities of no greater than 50% and the cellular topology has densities of no greater than 20%.

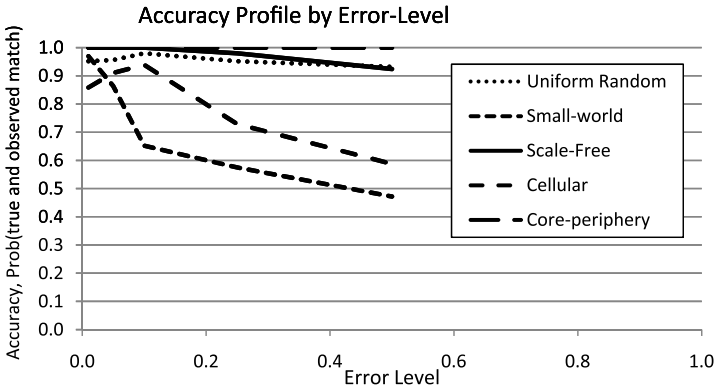

Fig. 3 Accuracy profile for topologies by error level (degree centrality, top 3 metric; size 50, density 10%, error remove)

Figure 3 shows an accuracy profile of error level. The graphic indicates that error level can be described, as having very different slopes for each of the topologies. The measure robustness in the Small-world network is substantially less that of the Core-periphery and the others. This suggests, from a visual perspective, that the error level may have a varying effect on the robustness and that the topology has even more influence than the error level per se.

As a concluding facet of this preliminary analysis, we conducted an informal comparison of our results with those of the Borgatti et al. (2006) study; thus, we considered only the results pertaining to the uniform random topology. We concluded that our results are abundantly consistent with those analogously presented by the earlier study. This confirmation provides an unsophisticated, but imperative, affirmation of the consistency of the operational aspects of our experiment across the two independent studies.

To summarize, via this preliminary analysis, we conclude that the operational aspect of the experimental process is sound and foreshadowing the more formal statistical analysis of the data, we can expect that there will likely be a noticeable difference in the measure accuracy across topologies, although the statistical significance is not yet presented. The next two sub-sections report the results of the more definitive, statistical analysis of the data.

3.2 Effect of topology on top-node metrics

The top-node metrics are binary measures; hence, we constructed logistic regression models to examine the impact of topology, and other factors, on the robustness of the metrics. Since the independent variables for all of the models are the same, the transformations of variables in both the top-node and the top-group models are likewise the same. Recognizing that the topology variable is an unordered categorical variable with five possible values, we transformed the topology variable into four distinct, binary dummy variables; the uniform random topology was designated as the baseline value. Further, since the error type variable is also an unordered group of categorical

values, it too was transformed into set of dummy variables; in the this case, we captured their values within two distinct, binary variables. The error type edge remove was fashioned as the baseline value for this group, i.e., both dummy variables are simultaneously set with a value of zero. Moreover, we examined each of the remaining three non-categorical, main effect variables (size, density and error level) for linearity and determined that error level is the only variable that necessitated a transformation to meet the assumptions of the regression technique. We found that the logarithmic function provided the best linear transformation for the error level variable and used this transformation for all models. Furthermore, it was not necessary to transform either the size or the density variables. Moreover, we ruled out the utility of using interaction terms in the regression equations as there was little expectation of a material statistical interaction between the independent variables; our initial examination of various regression mixture models corroborated this. An analysis of pairwise correlation showed no colinearity problems. The resulting full, logit based, log-odds regression model is shown as (1)

$$\begin{aligned}\text{logit}(p_i) = \beta_0 + \beta_1 \times \text{size} + \beta_2 \times \ln(\text{density}) + \beta_3 \times \text{errorLevel} \\ + \sum_{i=4}^{5}(\beta_i \times \text{errorType}) + \sum_{i=6}^{9}(\beta_i \times \text{topology})\end{aligned} \tag{1}$$

In order to appraise the extent to which a specific logit model is better than its base model (the model with only an intercept), the deviance difference is regularly computed between the two models of interest. These and all other combinations of possible models were constructed and deviance computed. Table 10 reports the coefficients of the logistic regression for a selected set of these models, specific to the degree centrality measure, top-node metrics. We describe three complete models for Top 1, constructed step-wise by increasing the number of factors, then we present only the selected, best full-models for the Top 3 and Top 10% metrics. Since the models are constructed using logistic regression, thus the regression equation is evaluated using maximum likelihood, the deviance is used as a criterion for quantifying the lack of fit for a model to the observed data; therefore, the lower the value of the deviance statistic, the better the model. When looking at only the first three Top 1 models reported in the table, the deviance is smallest for the model with the topology factors and is an improvement over the simpler models. As would be expected, the deviance increases as more terms are added to at the model, but since the Akaike (1973) information criterion (AIC), which introduces a penalty for the number of parameters in a model, confirms that the additional parameters, thus increased complexity, of the model is justified. Focusing on the three finalized models Top 1, Top 3, and Top 10%, reported in Table 10, error level has high negative-effect on accuracy and both size and density have relatively little effect on the accuracy. However, collectively, the topology variables have a larger effect on accuracy that approaches and even surpasses the effect of log (error level). The negative coefficient of a logit model implies that the predicted probability curve goes down to the right, which is the case in all of the factors except the Core Periphery and Scale Free topologies. From this we can conclude there is a relationship between topology and the measures that differs across differing topologies.

Table 10 Logistic regression models for degree centrality measure (across all top-node metrics)

Factor	Degree centrality				
	Top 1	Top 1	Top 1	Top 3	Top 10%
(Intercept)	2.896***	2.907***	3.144***	4.181***	3.941***
Network size	−0.006***	−0.006***	−0.007***	−0.008***	0.002***
Density	−0.002***	−0.002***	−0.007***	−0.008***	−0.005***
Log(error level)	−0.758***	−0.765***	−0.817***	−0.868***	−0.904***
Error type dummy 1		0.324***	0.347***	0.310***	0.230***
Error type dummy 2		−0.293***	−0.313***	−0.001	0.252***
Cellular topology			−0.070***	−0.024*	0.006
Core periphery topology			0.838***	0.533***	0.481***
Scale free topology			0.584***	0.617***	0.505***
Small world topology			−0.920***	−1.049***	−1.003***
Deviance–null model	296310.9	296310.9	296310.9	261686	228823.7
Deviance	184367.0	178839.3	141597.2	141779.6	133232.2
Deviance ratio	0.3777921	0.396447	0.5221328	0.458207	0.4177518
AIC	198699.3	193175.7	155941.6	153182.1	143035.4

*$z < 0.05$

***$z < 0.001$

In can be useful to provide a more intuitive interpretation of the logistic regression model. For instance, we can convert the coefficients reported in Table 10 for the degree centrality, top 1 model by changing the topology coefficients to odds values (take the anti-log of the odd-based coefficient); the model indicates that Core-periphery has a 2.31 times greater positive effect on accuracy than the Uniform Random topology. Likewise, Scale-Free has 1.79 times greater effect on accuracy, relative to uniform random. However, Small-world and Cellular affect accuracy at 0.39 and 0.93 times, respectively, thus less than that of Uniform Random topology.

Table 11 reports the selected best models for the remaining centrality measures (betweenness, closeness and eigenvector) and their respective top-node metrics (top 1, top 3, and top 10%). Across all of these statistical models, the sign of the coefficient for each topology is uniformly either positive or negative, which indicates consistency of the nature of the effect of topology in these metrics; a negative value means that the factor has less effect than the baseline factor value. The Core Periphery and Scale Free topologies both have a greater effect on the measure accuracy, relative to uniform random topology, while Cellular and Small World have a less so effect on measure accuracy. Moreover, it is unambiguous that topology has a far greater relative effect on measure accuracy than does the size or density of the network according to the size of the coefficients.

Table 12 reports the selected best models for the three top-node metrics computed from the local clustering measure. For each of the models featured in the table, all of the topologies have a relatively lesser effect on accuracy relative to the Uniform Random topology; although, the statistical strength of the coefficients do vary in these

Table 11 Logistic regression models for all top-node metrics for centrality measures: betweenness, closeness and eigenvector

Factor	Betweenness centrality			Closeness centrality			Eigenvector centrality		
	Top 1	Top 3	Top 10%	Top 1	Top 3	Top 10%	Top 1	Top-3	Top 10%
(Intercept)	2.527***	3.573***	3.219***	2.698***	3.504***	3.070***	3.022***	3.951***	3.563***
Network size	−0.007***	−0.008***	0.004***	−0.008***	−0.009***	0.003***	−0.008***	−0.009***	0.003***
Density	−0.005***	−0.005***	−0.003***	−0.000*	0.000	0.003***	−0.009***	−0.008***	−0.005***
Log(error level)	−0.755***	−0.795***	−0.816***	−0.792***	−0.794***	−0.793***	−0.765***	−0.791***	−0.811***
Error type dummy 1	0.282***	0.291***	0.324***	0.407***	0.360***	0.364***	0.097***	0.045***	0.060***
Error type dummy 2	−0.510***	−0.384***	−0.173***	−0.483***	−0.319***	−0.104***	−0.788***	−0.695***	−0.481***
Cellular topology	−0.071***	−0.099***	−0.062***	−0.698***	−0.699***	−0.590***	−0.345***	−0.605***	−0.574***
Core periphery topology	1.188***	0.857***	0.746***	1.061***	0.802***	0.754***	1.112***	0.749***	0.676***
Scale free topology	0.820***	0.952***	0.903***	0.656***	0.784***	0.721***	0.592***	0.676***	0.660***
Small world topology	−0.926***	−0.979***	−0.974***	−0.934***	−0.937***	−0.872***	−1.010***	−1.087***	−1.074***
Deviance–null model	320525.2	289773.2	255132.4	343826.2	301210.4	257180.9	352980.3	317320.7	275334.2
Deviance	150936.7	150066.0	141961.6	158907.7	151283.0	139090.3	174430.7	163917.6	149058.8
Deviance ratio	0.5290957	0.4821259	0.4435766	0.5378256	0.4977498	0.4591731	0.5058344	0.4834323	0.4586259
AIC	165645.8	162823.8	153230.0	173389.2	164115.1	150628.3	188628.7	177020.6	161212.4

*$z < 0.05$

***$z < 0.001$

Table 12 Logistic regression models for local clustering measure, across all top-node metrics

Factor	Local clustering		
	Top 1	Top 3	Top 10%
(Intercept)	3.443***	3.757***	3.334***
Network size	−0.011***	−0.011***	−0.002***
Density	−0.011***	0.000***	0.003***
Log(error level)	−0.848***	−0.852***	−0.813***
Error type dummy 1	0.894***	0.824***	0.821***
Error type dummy 2	−1.393***	−1.297***	−1.048***
Cellular topology	−0.007	−0.017	−0.016
Core periphery topology	−0.024*	−0.294***	−0.352***
Scale free topology	−0.070***	−0.327***	−0.431***
Small world topology	−0.718***	−0.484***	−0.430***
Deviance–null model	379814.4	342406.7	320473.0
Deviance	168120.7	158070.2	174891.6
Deviance ratio	0.557361	0.5383554	0.4542704
AIC	182324.4	172015.1	188064.4

*$z < 0.05$

***$z < 0.001$

models, particularly in the case of the Cellular topology, which suggests that accuracy for Cellular differs little from the uniform random topology. As with the previous models, relative to the effect that network size and density have on accuracy, the network's topology has substantially more effect and varies between the different stylized topology types.

In summary, the analysis of the regression models for the top-node metrics convincingly demonstrates, in nearly all circumstances, that the type of topology does have various amount of effects on the accuracy of the measure metric. For example, as detailed earlier, in the case of the Degree Centrality for the top 1 metric model, the Core-periphery topology has over twice as much effect on the accuracy than does the Uniform Random topology.

3.3 Effect of topology on top-group metrics

In order to investigate the effect of topology on the top-group metric, e.g., overlap, we constructed several independent statistical models for each of the five measures. Since the overlap metric is characteristically a continuous value, the models are constructed using a linear regression. We followed the same procedure as described above to transform the categorical independent variables and examined each of the main effect variables for linearity. Recall, we determined that error level is the only variable that necessitated a transformation to meet the linear assumptions of the regression technique; hence, we applied the logarithm function to the error level variable. Furthermore, we ruled out the utility of using interaction terms in the regressions as there was little expectation of a statistical interaction between the independent variables; our exploratory analysis corroborated the prudence of this. An analysis of pairwise correlation showed no colinearity problems. The resulting full, linear regression

Table 13 Linear regression models for all measures (top-group metrics: overlap)

Factor	Degree centrality overlap	Betweenness centrality overlap	Closeness centrality overlap	Eigenvector centrality overlap	Local clustering overlap
(Intercept)	0.869^{***}	0.851^{***}	0.853^{***}	0.905^{***}	1.040^{***}
Network size	-0.000^{***}	-0.001^{***}	-0.000^{***}	-0.001^{***}	-0.001^{***}
Density	0.000^{***}	0.001^{***}	0.001^{***}	-0.000^{***}	-0.002^{***}
Log(error level)	-0.117^{***}	-0.124^{***}	-0.129^{***}	-0.126^{***}	-0.134^{***}
Error type dummy 1	0.135^{***}	0.134^{***}	0.137^{***}	0.105^{***}	0.187^{***}
Error type dummy 2	-0.088^{***}	-0.132^{***}	-0.113^{***}	-0.146^{***}	-0.222^{***}
Cellular topology	0.068^{***}	0.048^{***}	-0.039^{***}	0.016^{***}	0.019^{***}
Core periphery topology	0.204^{***}	0.195^{***}	0.215^{***}	0.227^{***}	0.000
Scale free topology	0.097^{***}	0.126^{***}	0.103^{***}	0.093^{***}	-0.044^{***}
Small world topology	-0.155^{***}	-0.128^{***}	-0.135^{***}	-0.141^{***}	-0.103^{***}
Adjusted R^2	0.4760208	0.4156237	0.4674364	0.3922854	0.4290301

$^{***}z < 0.001$

model is shown as (2)

$$p_i = \beta_0 + \beta_1 \times \text{size} + \beta_2 \times \ln(\text{density}) + \beta_3 \times \text{errorLevel} + \sum_{i=4}^{5} \beta_i \text{errorType} + \sum_{i=6}^{9} \beta_i \text{topology} + \varepsilon. \quad (2)$$

Table 13 reports the OLS coefficients of the best regression models for the overlap metric for each of the five measures. These models are not nearly as consistent as those reported in the prior section for the top-node metrics. For example, the size of the coefficients is much smaller than in the prior models and the coefficients of the stylized topology factors are generally positive, which implies that they affect greater metric accuracy, relative to the uniform random network. The adjusted R^2 values indicate that there is much of the final accuracy explained by these models. The various R^2 values in these reported models reflect that roughly 43% of the data can be explained via each model.

Since the models reported in Table 13 are constructed using linear regression, transformation of the coefficients is not necessary prior to interpreting them. Each model indicates that the effect on accuracy of the Small-world topology is statistically different than the effect of the baseline Uniform Random topology and that Scale-free differs as well, albeit positively.

In summary, the analysis of the statistical models for the top-group metric, presented in this section, strongly shows, in all circumstances, that the type of topology does have statistically different levels of effect on the accuracy of the measures' overlap metric.

4 Discussion

This paper reports a systematic examination of the robustness of several network measures under different conditions of uncertainty, when controlling across several stylized network topologies. The results supported the conjecture: in the circumstance of uncertainty, the topological form of the true network has a measurable effect on the robustness of the measures when taken from the observed network data, relative to value taken from the ground truth. The chief finding this study is that the topology of the in-truth network does have a statically strong effect on the level of reliability of the centrality and the local clustering measures as computed from the observed data. Our findings suggest that making a priori classification of the topology of a network provides important additional information about the probabilistic reliability of the network measures that are computed over the observed data. Moreover, our results highlight the specific quantitative vulnerabilities of different topologies to the various combinations of error type, level and specific network measures.

Relative to a uniformly random network—one exclusive of being characterized as being of a stylized topology—topology can have either a greater positive or negative effect on the accuracy of measure, but often a material effect nonetheless. In other words, we found that the probability of the most prestigious actor or group of actors identified by an observed network dataset being accurate, in-actuality, is impacted by the topology of the in-truth network.

In the situation of the uniform random topology, our results are consistent with those reported by Borgatti et al. (2006). In terms of the other four topologies, the magnitudes of the effects on accuracy differ significantly across error type and the topology. In particular, the results show that the level of the affect on accuracy differs according to topology, more so than it differs across network size, density, error type and error level. For instance, node-remove errors tend to impact scale-free networks severely, particularly in the case of betweenness centrality. On the other hand, node-add error types have significantly higher impact on small-world networks, when considering the local clustering measure, than to the rest of the topologies.

Our findings have implications to researchers and practitioners alike. In short, the amount of confidence we can have for estimating the truly prestige actors in a network, according to their network position, is affected, either hindered or improved, by the topology of the network. When estimating the level of confidence for a measure taken from the data, we must first estimate the topology classification of the true network. The identification of these weaknesses in a particular network topologies provide important guidance for future empirical work, particularly in the area of identifying "key players" (Borgatti 2006), in terms of the reliability of results under conditions of measurement error. This level of confidence may even be applicable to individuals during the selection process of them seeking new relationships (see Fortunato et al. 2006).

Finally, our results also have important implications from a data collection point of view. In the inherent trade-off between resource cost and completeness of data collected, our findings indicate that the researchers should include an estimate of the true topology of the subject network in making an assessment of this important trade-off. For instance, if we are interested in a key node based on degree centrality, in

order to achieve an 85% level of confidence that we accurately identified the most prestigious actor in a 100-node, 10% density network, within a group of three, we would require a 50% or greater level of accuracy in the data collection process if the network is a Small World, but only a 40% level of accuracy for a Cellular network. Indeed, at this juncture, how to determine this accuracy level itself is an unknown task. In other words, our findings provide guidelines on how accurate the data must be when collected in order to reach particular levels of accuracy in the network measures under consideration.

4.1 Limitations

We draw attention to three limitations of the methodology applied in this study. Each of these is a matter that warrants future research. The most relevant limitation is the unproven nature of the network creation algorithms used in the true-network generation procedure. It remains an open question whether the algorithms statistically and accurately simulate the drawing of a random sample from the complete distribution of *all* the possible networks of a specific topology. There may be an unequal bias towards particular instances of isomorphic networks, within the topology; we are working with generative models rather than utilizing a much preferred random sampling technique that draws uniformly from all possible configurations of a topology with the given parameters. Verifying these algorithms is well known as a computationally prohibitive task. Secondly, we acknowledge that the various characteristics particular to a given topology may alter the results when the entire range of possible parameters for a topology is studied. That is to say, for example, the findings for a scale-free network may differ for alpha being set to 3, rather than a value of 2 as set in this experiment. Moreover, we operate on the supposition that errors in social data are random in nature, when they may in fact have a variety of non-independent biases, sources and patterns (see Rubin 1976), which in turn, may present influences in the data actually observed and, therefore the character of the observed error. Finally, an significant limitation of this study is that we only investigated errors of a single type rather than the more real-world scenario of a mixture of errors, i.e., observed network data having a blend of over and underreported, ties and nodes, in the same dataset.

4.2 Future research

We present five recommendations for future research that we believe would provide further advancement towards a remedy to the inherent data-uncertainty problem. First, while we have shown that topology has an affect on the robustness of centrality measures, there is the next question about the precise extent to which each of the many different topologies and their variants distinctively affect the robustness profiles. In the guise of network topology labels, subtle differences in the methodology for generating a given network may possibly result in diverse robustness levels. Perhaps it is a characteristic of a topology (thus a family of topologies) that matter, not a specific topology itself. For example, there are many ways to generate a core periphery network; each variant needs to be explored and individually related to a specific robustness profile. Further, technically, characteristically, a small-world network is

also a scale-free network (Amaral et al. 2000), so there needs to be more attention to the precise topological characteristics, beyond the approach of merely using the label-names of a topology, as is used herein.

Second, as we and others have openly acknowledged, errors in observed social network data most likely are not truly random in nature. Early, as well as this, research specific toward investigating robustness has been limited to being based on random error as opposed to more realistic, systemic or non-randomly influenced errors in the data. One notable exception to this is the Marsden (1990) study which examined both random and non-random errors. Certainly, this makes the research much more complicated, but the community will be rewarded with theories based on richer scenarios.

Third, it should prove invaluable to analysts when they have statistically valid confidence levels and error bounds of descriptive network measures applicable to their specific observed network. Such quantities may possibly be based on the known parameters and characteristics of the observed network combined with the a priori true network information and error characteristics. To date, analysts are constrained by using measures determined only from the observed network, thus are being limited to working with descriptive statistic only. The analysis of networks will take a huge leap forward when confidence levels can be assigned to collected data that will ultimately lead to including p-vales with the statistics we calculate from observed data.

Fourth, the issue of alias nodes is a matter of consequence. Particularly in studies with data collection automated by computer, the problem of entity disambiguation, i.e., a single node being recorded as separate multiple nodes, is type of error that likely has very real implications to analysis. The resulting network measures based on the observed data would most likely lead to meaningless information without some mathematical adjustments being applied; foremost, the impact of this type of error must be understood.

Finally, the fifth area that can contribute to addressing the data-error scenario is to begin to verify the results drawn from virtual experiments, such as this, by testing the results against real-world, empirical networks; although, at this point we only briefly ponder how such a study might be designed, since the ground-truth in the real world is never really known and thus cannot be used to verify the accuracy of the observed.

Acknowledgements The authors acknowledge and thank several indispensable people for their priceless contribution to this study. Edo Airoldi and Jeff Reminga provided critical algorithms and software components for the study; Eunice J. Kim provided supportive feedback on the statistics and Brian Hirshman and Stephen Borgatti for scholarly advice as well. This work was supported by the Department of Defense and National Science Foundation under MKIDS program (IIS0218466) and the Office of Naval Research under Dynamic Network Analysis program (N00014-02-1-0973) and WVHTCFTAVI071375 and N00014-06-1-0104, with additional support from the Air Force Office of Sponsored Research (MURI: Computational Modeling of Cultural Dimensions in Adversary Organizations, FA9550-05-1-0388), TAVI, HEINZ-IGERT TRAINING PROGRAM (NSF, DGE-9972762), VIBES (ARMY DAAD19-01C0065), Dynamic Network Analysis Applications to Counter-Narcotic Investigations Related to Marijuana (ONR, N00014-06-1-0104 (Mod#2)), MOAT Phase II, C2 Insight (DARPA, N0000610921), Network Analysis and Computational Modeling for Combating Terrorist Threats (DYNET:MMVOIA Quick Reaction Fund) (ONR, N000140610921), and Determine Ability to Model Terrorist Networks Unit (SPAWAR). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense, the National Science Foundation, the Office of Naval Research, or the US government.

References

Airoldi EM, Carley KM (2005) Sampling algorithms for pure network topologies: Stability and separability of metric embeddings. SIGKDD Explor 7:13–22. Special Issue on Link Mining

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csàki F (eds) Second international symposium on inference theory. Akadémiaia Kaidó, Budapest, pp 267–281

Albert R, Jeong H, Barabasi AL (1999) Diameter of the World Wide Web. Nature 401:130–131

Albert E, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406:378–382

Alvo M, Cabilio P (1995) Rank correlation methods for missing data. Can J Stat (La Revue Canadienne de Statistique) 23:345–358

Amaral LAN, Scala A, Barthelemy M, Stanley HE (2000) Classes of small-world networks. Proc Nat Acad Sci USA 97:11149–11152

Barabasi AL, Albert E (1999) Emergence of scaling in random networks. Science 286:509–512

Bollobás B (2001) Random graphs, 2nd edn. Cambridge University Press, Cambridge

Bonacich P (1987) Power and centrality: A family of measures. Am J Soc 92:1170–1182

Borgatti SP (2006) Identifying sets of key players in a network. Comput Math Organ Theory 12:21–34

Borgatti SP, Everett MG (1999) Models of core/periphery structures. Soc Netw 21:375–395

Borgatti S, Reminga J (2005) Personal communication

Borgatti SP, Carley KM, Krackhardt D (2006) On the robustness of centrality measures under conditions of imperfect data. Soc Netw 28:124–136

Burt R (1978) Stratification and prestige among elite experts in methodological and mathematical sociology circa 1975. Soc Netw 1(1978):105–158

Butts CT (2003) Network inference, error, and informant (in)accuracy: a Bayesian approach. Soc Netw 25:103–140

Carley KM (2002) Dynamic network analysis. In: Breiger R, Carley KM, Pattison P (eds) Dynamic social network modeling and analysis: 2002 workshop summary and papers. National Academies Press, Washington

Calloway M, Morrissey JP (1993) Accuracy and reliability of self-reported data in interorganizational networks. Soc Netw 15:377–398

Costenbader E, Valente TW (2003) The stability of centrality measures when networks are sampled. Soc Netw 25:283–307

Davis GF, Yoo M, Baker WE (2003) The small world of the American corporate elite, 1982–2001. Strateg Organ 1:301–326

Erdös P, Rényi A (1959) On random graphs. Publ Math 6:290–297

Erickson BH, Nosanchuk TA (1983) Applied network sampling. Soc Netw 5(4):367–382

Feld SL, Carter WC (2002) Detecting measurement bias in respondent reports of personal networks. Soc Netw 24:365–383

Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the Internet topology, ACM SIGCOMM'99. Comput Commun Rev 29:251–262

Fortunato S, Flammini A, Menczer F (2006) Scale-free network growth by ranking. Phys Rev Lett 96:218701

Frank O (1978) Sampling and estimation in large social networks. Soc Netw 1:91–101

Frank O (1981) A survey of statistical methods for graph analysis. In: Leinhardt S (ed) Sociological methodology. Jossey-Bass, San Francisco

Frantz TL, Carley KM (2005) A formal characterization of cellular networks. Technical report CMU-ISRI-05-109, School of Computer Science, Carnegie Mellon University

Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40:35–41

Freeman LC (1979) Centrality in social networks: I conceptual clarification. Soc Netw 1:215–239

Freeman LC, Rommey Kimball A, Freeman SC (1987) Cognitive structure and informant accuracy. Am Anthropol 89:310–325

Galaskiewicz J (1991) Estimating point centrality using different network sampling techniques. Soc Netw 13:347–386

Gile K, Handcock MA (2006) Model-based assessment of the impact of missing data on inference for networks. Working paper, no. 66, Center for Statistics and Social Sciences, University of Washington

Granovetter M (1976) Network sampling: Some first steps. Am J Soc 81:1287–1303

Guare J (1990) Six degrees of separation: A play. Vintage, New York

Handcock MS, Gile K (2007) Modeling social networks with sampled or missing data. Working paper, no. 75, Center for Statistics and Social Sciences, University of Washington
Killworth PD, Bernard HR (1976) Informant accuracy in social network data. Human Organ 35:269–286
Kim PJ, Jeong H (2007) Reliability of rank order in sampled networks. Eur Phys J B 55:109–114
Kossinets G (2006) Effect of missing data in social networks. Soc Netw 28:247–268
Krebs VE (2002) Mapping networks of terrorist cells. Connections 24:43–52
Marsden PV (1990) Network data and measurement. Ann Rev Soc 16:435–463
Marsden PV (1993) The reliability of network density and composition measures. Soc Netw 15:300–421
Masuda N, Konno N (2006) VIP-club phenomenon: Emergence of elites and masterminds in social networks. Soc Netw 28:297–309
Mayntz R (2004) Organizational forms of terrorism: Hierarchy, network, or a type sui generis? MpIfG discussion paper 04.4. Max Planck Institute for the Study of Societies
McKnight PE, McKnight KM, Sindani S, Figueredo AJ (2007) Missing data. Guilford, New York
Milgram S (1967) The small world problem. Psychol Today 2:60–67
Newman ME, Watts DJ (1999) Scaling and percolation in the small-world network model. Phys Rev E 60:7332–7342
Newman ME, Watts DJ, Strogatz SH (2002) Random graphs models of social networks. Proc Nat Acad Sci 99:2566–2572
Papaioannou T, Loukas S (1984) Inequalities on rank correlation with missing data. J R Stat Soc, Ser B 46:68–71
Powell WW, White DR, Koput KW, Owen-Smith J (2005) Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. Am J Soc 110:1132–1205
Ravid G, Rafaeli S (2004) Asynchronous discussion groups as small world and scale free networks. First Monday 9(9), September
Robins G, Pattison P, Woolcock J (2004) Missing data in networks: Exponential random graph ($p*$) models for networks with non-respondents. Soc Netw 26:257–283
Ronfeldt D, Arquilla J (2001) Networks, netwars, and the fight for the future. First Monday 6
Rubin DB (1976) Inference and missing data. Biometrika 63:581–592
Simon HA (1955) On a class of skew distribution functions. Biometrika 42:425–440
Stork D, Richards WD (1992) Nonrespondents in communication network studies: Problems and possibilities. Group Organ Manag 17:193–209
Stuart T, Robinson D (2000) The origins of inter-organizational networks. Working paper, Columbia Business School
Tsvetovat M, Carley KM (2005) Structural knowledge and success of anti-terrorist activity: The downside of structural equivalence. J Soc Struct 6 (2005)
Watts DJ (1999a) Small worlds: The dynamics of networks between order and randomness. Princeton University Press, Princeton
Watts DJ (1999b) Networks, dynamics, and the small-world phenomenon. Am J Sociol 105:493–527
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
White S, Smyth P (2003) Algorithms for estimating relative importance in networks. ACM SIGKDD'03, August 24–27, Washington
Zemljic A, Hlebec V (2005) Reliability of measures of centrality and prominence. Soc Netw 27:73–88

Terrill L. Frantz received an Ed.D. in Organization Change from Pepperdine University and is currently a Ph.D. student in Computation, Organizations and Society at Carnegie Mellon University. He holds an M.S. from the School of Computer Science at Carnegie Mellon and an MBA from the Stern School of Business at New York University; He earned his B.S. in Business Administration from Drexel University in 1984. He has extensive experience in Information Technology as a Vice President at Morgan Stanley and at JP Morgan and other global and local institutions. Terrill's research interests include Organizational Post-Merger Integration, Organizational Network Analysis, and Computational Organization Theory.

Marcelo Cataldo received M.S. and Ph.D. degrees in Computation, Organizations and Society from Carnegie Mellon University in 2007. He also received a B.S. in Information Systems from Universidad Tecnologica Nacional (Argentina) in 1996 and a M.S. in Information Networking from Carnegie Mellon University in 2000. His research interests are in the area of distributed product development with special focus on the relationship between the product architecture and the structure of the technical organization. Marcelo Cataldo is a Senior Research Engineer at Robert Bosch's Research and Technology Center.

Kathleen M. Carley is the director of the Center for Computational Analysis of Social and Organizational Systems (CASOS), a university wide interdisciplinary center that brings together network analysis, computer science and organization science (www.casos.cmu.edu). Her research combines cognitive science, social networks and computer science to address complex social and organizational problems; Specific research areas are dynamic network analysis, computational social and organization theory, adaptation and evolution, text mining, and the impact of telecommunication technologies and policy on communication, information diffusion, disease contagion and response within and among groups particularly in disaster or crisis situations. Dr. Carley is the director of the center for Computational Analysis of Social and Organizational Systems (CASOS) which has over 35 members, both students and research staff. She is the founding co-editor of CMOT and has co-edited several books in the computational organizations and dynamic network area.