

## Measurement error in network data: A re-classification<sup>☆</sup>

Dan J. Wang<sup>a,\*</sup>, Xiaolin Shi<sup>b</sup>, Daniel A. McFarland<sup>a</sup>, Jure Leskovec<sup>a</sup>

<sup>a</sup> Stanford University, Stanford, CA 94305, United States

<sup>b</sup> Microsoft Corporation, Redmond, WA 98052, United States

### ARTICLE INFO

#### Keywords:

Measurement error  
Missing data  
Simulation

### ABSTRACT

Research on measurement error in network data has typically focused on missing data. We embed missing data, which we term *false negative nodes and edges*, in a broader classification of error scenarios. This includes *false positive nodes and edges* and *falsely aggregated and disaggregated nodes*. We simulate these six measurement errors using an online social network and a publication citation network, reporting their effects on four node-level measures – degree centrality, clustering coefficient, network constraint, and eigenvector centrality. Our results suggest that in networks with more positively-skewed degree distributions and higher average clustering, these measures tend to be less resistant to most forms of measurement error. In addition, we argue that the sensitivity of a given measure to an error scenario depends on the idiosyncracies of the measure's calculation, thus revising the general claim from past research that the more 'global' a measure, the less resistant it is to measurement error. Finally, we anchor our discussion to commonly-used networks in past research that suffer from these different forms of measurement error and make recommendations for correction strategies.

© 2012 Elsevier B.V. All rights reserved.

### 1. Introduction

Network analysis has long been plagued by issues of measurement error, usually in the form of missing data. For instance, survey data used in early sociometric research often contained misrepresentations of ego-networks due to the limits of respondent memory and survey design (Marsden, 1990).

While missing data remains a major issue, much network research currently faces the opposite problem. The growing availability of large, complex network datasets has transformed a research environment that lacked sufficient data into one with an overabundance (De Choudhury et al., 2010). Network researchers regularly analyze networks with millions of nodes and edges with multiplex relations. However, while much research still

focuses on missing data, it has overlooked other major classes of measurement error that have emerged (for an exception, see Borgatti et al. (2006)).

Here, *measurement error* refers to mistakes in collecting or coding a network dataset. For example, in a sociometric survey, if a respondent misspells the name of a contact, then the contact might erroneously be treated as two different individuals. However, measurement error can also refer to the extent to which a network dataset represents the reality of the relationships within a group under study. For instance, even if all respondents report the correct spellings of their friends' names, the understanding of what qualifies as a friendship tie can vary by respondent.

Thus, in network research, there exist three levels of empirical interpretation (see Table 1) – (1) the *ideal network*: the true set of relations among entities in a network, (2) the *clean network*: the set of relations among entities as coded in a network dataset without data entry mistakes, and (3) the *observed network*: a network dataset, often suffering from coding errors, that is actually available to a researcher. Resolving the differences between the ideal network and clean network is difficult because it entails applying an objective understanding to what is an inherently subjective relationship (see De Choudhury et al. (2010)). Because most measurement errors are a result of inconsistencies in data collection and coding, we focus our analysis on the discrepancies between the clean and observed network.

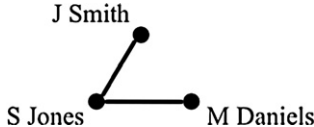
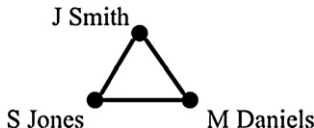
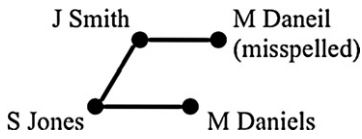
First, we classify network measurement error into six types – missing nodes, spurious nodes, missing edges, spurious edges, falsely aggregated nodes, and falsely disaggregated nodes. In effect, we build on the important work of Borgatti et al. (2006), who

<sup>☆</sup> This material is based upon work supported by the Office of the President at Stanford University and the National Science Foundation under Grant No. 0835614. In addition, this research was in part supported by NSF Grants CNS-101092, IIS-1016909, IIS-1149837, the Albert Yu & Mary Bechmann Foundation, IBM, Lightspeed, Samsung, Yahoo, an Alfred P. Sloan Fellowship, and a Microsoft Faculty Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Stanford University or the National Science Foundation. We are grateful to the comments from members of the Mimir Project workshop at Stanford University on earlier iterations of this work. We also are indebted to the editors of Social Networks and the two anonymous reviewers, who provided helpful suggestions, many of which we incorporated into the final version of this paper.

\* Corresponding author at: Department of Sociology, Stanford University, 450 Serra Mall, Stanford, CA 94305-2047, United States.

E-mail address: [djwang@stanford.edu](mailto:djwang@stanford.edu) (D.J. Wang).

**Table 1**  
Three levels of empirical network interpretation.

	Description	Example 1 Friendship network gathered through sociometric surveys	Example 2 Collaboration network gathered by scraping publication database	Illustration
Ideal network	Network of true relations between entities	Ties represent actual, mutual friendships between individuals	Ties represent active collaborative relationships between individuals	
Clean network	Network of relations encoded in data without measurement error	Ties represent each respondent's own perception of friendship with others	Ties represent co-author relationships, but not necessarily active collaboration between individuals	
Observed network	Network of relations encoded in data with measurement error	Ties represent reported friendships, but some nodes and ties are erroneously coded	Ties represent co-author relations, but some nodes and ties are erroneous coded	

compared the effects of spurious and missing nodes and edges on centrality measures in Erdős-Rényi random graphs. By contrast, we compare the effects of our error scenarios by simulating them in two real-world networks and one random graph, each of which vary in key structural characteristics, such as average clustering and degree distribution. We then observe the effects of these error scenarios on four different node-level network measures – degree centrality, clustering coefficient, network constraint, and eigenvector centrality.

In addition, we make recommendations for error correction strategies, which depend on the type of measurement error scenario, the network measure affected, and the structural attributes of the network under study. We also ground our discussion in common examples of network datasets from past research. By bringing attention to these understudied classes of measurement error, we alert network researchers to important sources of bias that extend beyond missing data.

### 1.1. Related work

Early examinations of the measurement error in networks focused on missing data in sociometric surveys. Common sources of error include respondent bias (Pool and Kochen, 1978), non-response (Stork and Richards, 1992), and the design of questionnaires (Burt, 1984; Holland and Leinhardt, 1973). In addition, this work identified factors in research design that could give rise to measurement error. Specifically, the boundary specification problem refers to the issue of deciding which units qualify as legitimate members of a connected network (Laumann et al., 1983). Barnes (1979), for instance, noted that specifying a restrictive boundary can underestimate centrality measures. As a solution, some used objective criteria for gathering network data, such as only including individuals who are, by definition, members of a group, such as students of a given school (Coleman, 1961; Kapferer, 1969).<sup>1</sup>

<sup>1</sup> See also work on network sampling and correction strategies for missing data (Granovetter, 1976; Marsden, 1990; Leskovec and Faloutsos, 2006; Handcock and Gile, 2010; Sadiqov et al., 2011).

Our paper builds on the more recent work of Kossinets (2006) and Borgatti et al. (2006), who simulate of measurement errors on random Erdős-Rényi networks. Kossinets (2006) focuses on missing network data, finding that clustering coefficients are overestimated when the boundary specification is too restricted, and centrality measures are underestimated when non-response is pervasive. Borgatti et al. (2006) begin an important conversation about the typology of measurement errors. They find that spurious nodes and edges diminish the similarity between a network dataset and its error-afflicted counterpart but not as much as the removal of nodes and edges.

We argue that key structural features of real-world social networks can also influence the robustness of certain network measures to error scenarios. Erdős-Rényi graphs, like those in Kossinets (2006) and Borgatti et al. (2006), tend to have little clustering and more uniform degree distributions, making their comparison to empirical networks unrealistic (Newman and Park, 2003). We therefore use empirical network datasets for simulating measurement error to issue more relevant cautions to empirical researchers about different forms of network measurement error.

## 2. Error scenarios

First, network data can suffer from missing or spurious nodes, which we term *false negative nodes* and *false positive nodes*, missing or spurious edges, which are termed here as *false negative edges* and *false positive edges*, or the erroneous merging or splitting of nodes, which we call *false aggregation* and *false disaggregation*. Since we have already reviewed work on missing nodes and edges, their discussion below will be brief. Table 2 summarizes the measurement error scenarios we consider in this paper.

### 2.1. Nodes

#### 2.1.1. False negative nodes

*False negative nodes* refer to the absence of nodes that should be present in a network. Examples include network data gathered via snowball sampling, which tend to underestimate the total membership of the groups under study (Erickson, 1978). Other sources of false negative nodes include non-response from surveys

**Table 2**  
Description of types of measurement error in network data.

Error scenario	Example	Empirical references
False negative nodes	Non-response in sociometric surveys, boundary misspecification	Stork and Richards (1992)
False positive nodes	Fake user accounts in online communities	Lewis et al. (2008) and Narayanan and Shmatikov (2009)
False negative edges	Imperfect respondent recall, limiting contact list in sociometric surveys	Sudman (1985) and Brewer (2000)
False positive edges	False ties in online communities, tie-decay windows that are too wide	Lewis et al. (2008))
False aggregation	In entity resolution on coauthorship networks, mistakenly treating different authors as the same author	Newman (2002) and Fleming et al. (2007)
False disaggregation	In entity resolution on coauthorship networks, mistakenly treating different spellings of the same author's name as different authors	Newman (2002) and Azoulay and Zivin (2005)

(Stork and Richards, 1992). Also, recent work on citation networks indicate that large portions of a publication's references can be missing from a citation database due to coding errors (Shi et al., 2010).

### 2.1.2. False positive nodes

In contrast to false negatives, *false positive nodes* refer to nodes that are erroneously present in a network. Few network researchers have systematically examined this measurement error scenario (Borgatti et al., 2006). However, just as respondent bias can result in the underreporting of nodes and relations, it can also lead to the overrepresentation of certain elements of a network (Feld and Carter, 2002).

False positive nodes are pervasive in networks when the data being gathered are not subject to close scrutiny. Consider, for example, an online community where spamming scripts generate false user activity. By design, these 'spam-users' mimic human online activity, which is often impossible to filter completely given the amount of data in large online communities.

Nevertheless, growing interest in online communities has resulted in studies of Twitter (Narayanan and Shmatikov, 2009), Facebook (Ackland, 2009; Lewis et al., 2008), and the World of Warcraft (Nardi and Harris, 2006), which contain a wealth of relational information. It would be naive, though, to take the data at face-value because among other issues, it is often difficult to distinguish between real and fake users. For example, independent assessments suggest that 27% of all Facebook accounts are fake (Nyberg, 2010; Richmond, 2010). While data from online communities can provide insight into complex network dynamics, filtering false data stands as a major challenge for network researchers.

## 2.2. Edges

### 2.2.1. False negative edges

*False negative edges* occur when relationships between nodes that should be reported are not observed in a network. In sociometric surveys, the risk of false negative edges comes from respondents' imperfect recall of their ego-networks (Bernard et al., 1984; Brewer, 2000).<sup>2</sup> In addition, the survey itself might restrict the number of contacts a respondent may list, enforcing an artificial limit on measures like node degree (Burt, 1984).

### 2.2.2. False positive edges

*False positive edges* occur when relationships between nodes are erroneously present in a network. In surveys, respondents sometimes report relations that are not actually present. In online community data, many contacts listed by users by no means represent real world relationships. Attempts to discern "real" ties

from "false" or "virtual" ties have generally found that features like temporal or spatial proximity can be used to discern actual relationships (De Choudhury et al., 2010; Wuchty, 2009).

Other settings like co-authorship networks, which represent collaborative relationships, can also contain false positive edges (Newman, 2001; Wuchty et al., 2007). Here, spurious edges come from failing to account for tie decay. Most researchers have dealt with this issue by setting an arbitrary time window to signal the period in which an established tie is meaningful (Fleming and Frenken, 2007). Using windows that are unrealistically long, however, can introduce false positive edges.<sup>3</sup>

For example, in communication networks, such as the Enron email corpus (Carley and Skillicorn, 2005), using every email to represent ties between individuals would be inappropriate because emails from years ago might be irrelevant to signaling more current relationships. Conversely, using a tie-decay window that is too narrow might overlook important ties from the past (De Choudhury et al., 2010).

### 2.3. False aggregation and disaggregation

The final set of measurement errors is less common than false positives or false negatives, but they nonetheless deserve consideration. *False aggregation* refers to the error scenario in which two nodes, *A* and *B*, are mistakenly treated as one node. *False disaggregation* is the opposite problem, in which one node *A*, is erroneously treated as two separate nodes, *A* and *B*.

The false aggregation and disaggregation of nodes typically occur during data cleaning. The problem is closely related to *entity resolution* – the disambiguation of distinct 'entities' in a dataset, such as similar author names in citation databases for bibliometric analysis. Citation databases used to construct co-authorship networks often contain author names under multiple spellings (Ahuja, 2000; Fleming et al., 2007; Azoulay and Zivin, 2005). Sometimes, different but close spellings can represent the same author while in other cases, different spellings can actually refer to different authors (or even worse, the same exact spelling can refer to two different authors).

Some algorithms for entity resolution treat too many actual authors erroneously as multiple authors (false disaggregation), while others can mistakenly group different too many authors as the same author (false aggregation). While it is almost impossible to know the true list of disambiguated authors (given that publication databases can record millions of author names), it is conceivable that one problem is worse than the other.

At the very least, the results of our research can better inform readings of previous studies, which have noted the entity resolution

<sup>2</sup> Brewer (2000) reviews 17 studies, in which the recall of social contacts among surveyed respondents varied between 16% and 99%.

<sup>3</sup> Fleming and Frenken (2007, p. 952) note that varying the tie window in their study does not greatly affect the structural features of their co-patenting networks; we suspect, however, in other contexts, misspecifying a tie-decay window can have more detrimental effects.

**Table 3**  
Descriptive statistics for networks under study.

	Nodes	Edges	Density	Avg. clustering coefficient
<i>Slashdot.com network</i>				
Empirical network	70,416	353,595	0.00014	0.0561
Degree sequence random network	70,416	350,073	0.00014	0.0112
Erdős-Rényi random network	70,416	353,595	0.00014	0.0001
<i>ArXiv citation network</i>				
Empirical network	27,770	352,324	0.00091	0.3121
Degree sequence random network	27,770	349,926	0.00091	0.0123
Erdős-Rényi random network	27,770	352,324	0.00091	0.0009

issue, but have failed to completely account for it (for an exception, see Newman (2002)).

### 3. Data and methods

#### 3.1. Datasets

Following Costenbader and Valente (2003), we compare different types of measurement error by simulating them in two empirical datasets. Furthermore, because we are interested in how key structural features of a network can moderate the impact of error scenarios, we also simulate our error scenarios for rewired networks with the same degree sequences as their empirical counterparts and Erdős-Rényi random graphs that have the same size and density as the real networks.<sup>4</sup>

We use two empirical datasets that have been studied in past research. First, we analyze the friendship network of Slashdot.com users, which contains 77,357 nodes and 353,595 edges (Leskovec et al., 2010). Slashdot.com is an online technology forum where users share and comment on technology news. Slashdot users frequently engage in online discussion and can nominate one another as either ‘friends’ or ‘foes’.<sup>5</sup> Second, we consider a citation network of publications in the field of High-Energy Physics Theory, available in the e-print repository, ArXiv (Gehrke et al., 2003). All articles were published between January 1993 and April 2003. This graph contains 27,770 nodes and 352,324 edges.<sup>6</sup>

Aside from their structural differences, we chose these two datasets because they represent different types of commonly-analyzed networks (Newman and Park, 2003). In addition, both networks are similar in size, but vastly different in other structural features. The average clustering in the citation network is over five times greater than in the Slashdot network (Table 3), and whereas the degree distribution of the Slashdot network follows a power law, the citation network’s degree distribution does not (Fig. 1). In addition, we chose two large networks because manual data cleaning is often unrealistic for such data. Our later discussion of error correction strategies focuses on automated methods, which would be unnecessary and less applicable to small networks. For smaller networks, manual data cleaning is more tractable and likely more accurate.<sup>7</sup>

<sup>4</sup> We used a configuration model to generate a rewired network with a specified degree sequence using the ‘configuration\_model’ function from the Python package *networkx* (Hagberg et al., 2008). The function allows for the creation of loops which is why the number of edges in our rewired networks is slightly lower than that in our empirical networks. Since the difference in edge count is trivial, we are confident that this does not affect our results.

<sup>5</sup> These data were gathered in November 2008. In our analysis, we remove ‘foe’ links, which allow Slashdot users to express negative feelings toward certain other users.

<sup>6</sup> See <http://snap.stanford.edu> for more information about these two networks and download links for the data.

<sup>7</sup> We caution readers, though, that the effects of measurement error are also size-dependent (Borgatti et al., 2006), but this issue is beyond the scope of this paper.

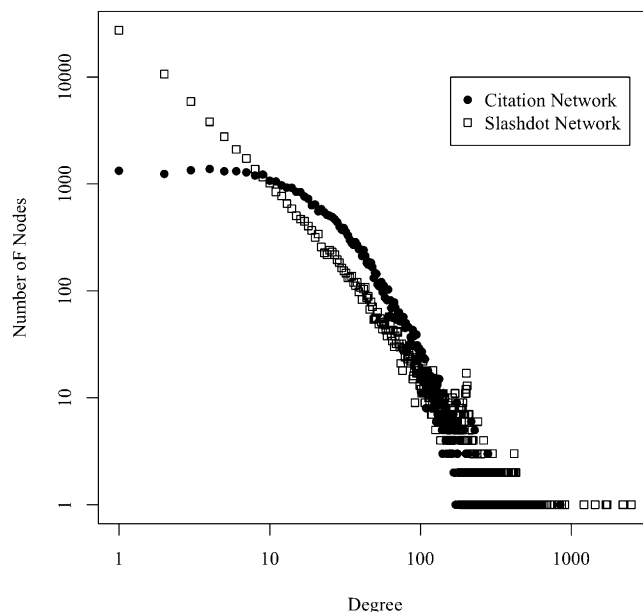
While we are aware that these two networks themselves might suffer from measurement error, we treat them as if they were clean to observe the *effects* of measurement error on networks with real-world features (Table 1). However, we are reasonably confident that the Slashdot.com dataset is free from spam activity, and the ArXiv database is routinely monitored for errors by human administrators (Gehrke et al., 2003). In addition, we use versions of these datasets that have been cleaned thoroughly for analysis in past research.

Although these two networks are directed networks, we treat all ties as undirected to simplify our analysis. Thus, the degree distributions in Fig. 1 use the sum of in- and out-degrees.

#### 3.2. Approach

Our goal is to compare the robustness of four node-level network measures – degree centrality (Freeman, 1979), clustering coefficient (Watts and Strogatz, 1998), network constraint (Burt, 1992), and eigenvector centrality (Bonacich, 1987) – to different error scenarios (Table 4 summarizes these measures for our networks). Our intuition is that if, for instance, the overall ranking of nodes by degree centrality does not change dramatically under an error scenario, then the degree centrality measure is robust. The process of simulating an error scenario follows:

1. Consider a clean network,  $G(V, E)$ ;
2. Apply an error scenario to  $G$ ; call the new perturbed network,  $G'(V', E')$ ;
3. For each node  $u$  that is in both  $G$  and  $G'$  ( $u \in V \cap V'$ ) calculate a given node-level network measure and store the measure for  $u$



**Fig. 1.** Degree distribution of Slashdot and ArXiv citation networks.

**Table 4**  
Summary statistics for network measures used in analysis.

	Slashdot friendship network			ArXiv citation network			Formula
	Mean	SD	Skewness	Mean	SD	Skewness	
Degree centrality	10.0430	34.7102	22.9295	24.3687	30.8759	5.2178	Count of the number of ties for a node $i$ ; usually scaled so that the maximum value is 1 (Freeman, 1978:221).
Clustering coefficient	0.0560	0.1830	4.2856	0.2848	0.2071	1.2612	$CC_i = \frac{2 e_{jk} }{k_i(k_i - 1)}$ $e_{jk}$ is the set of edges between node $i$ 's neighbors; and $k_i$ is the number of $i$ 's neighbors (Watts and Strogatz 1998:441).
Network constraint	0.5450	0.3937	0.0749	0.1623	0.2095	2.7682	$C_i = \sum_{j \neq i} \left( p_{ij} + \sum_{k \neq i, k \neq j} p_{ik} p_{kj} \right)^2$ $p_{ij}$ is strength of the tie between a node $i$ and its neighbor $j$ (in our case, $p = 1$ for all ties); $k$ and $j$ represent $i$ 's neighbors. Higher values of $C_i$ indicate that $i$ acts as less of a structural hole (Burt, 1992:54).
Eigenvector centrality	0.0071	0.0290	13.9200	0.0098	0.0256	8.9063	A node's eigenvector centrality is the unit-normalized sum of its ties to its neighbors, wherein each tie to a neighbor is weighted by the neighbor's ties, and each of the neighbors ties are weighted, and so forth. To facilitate calculation, given a graph $G$ 's representation as an adjacency matrix $A$ , the eigenvector centrality of node $i$ is given by the $i$ th element of $A$ 's unit-normalized principal eigenvector (Bonacich, 1987:1172).

Note: We calculate skewness as the third moment about the mean.

from  $G$  in the vector  $M$  and the measure of  $u$  from  $G'$  in the vector  $M'$ .

#### 4. Calculate the rank correlation (Spearman's rho) of $M$ and $M'$ .

We use Spearman's rho as an evaluation criterion because Pearson's correlation would add noise as a result of its sensitivity to linearity. We prioritize the node ranking by a given network measure because the distribution of our node-centric measures in empirical social networks tends to be highly skewed. Consider a network with a positively skewed degree distribution. If the network's top-ranked node by degree centrality loses half of its edges but remains top-ranked, this would diminish the *linear* correlation between  $M$  and  $M'$  (containing node degree centralities), but would have no effect on the *rank* correlation between  $M$  and  $M'$ . In other words, if we used Pearson's correlation, we would risk exaggerating the effect of measurement error. Spearman's rho is simply a more practical measure.<sup>8</sup>

We compare degree centrality, clustering coefficient, network constraint, and eigenvector centrality because they represent commonly used network measures, the calculations of which range from *local* (degree) to *semi-local* (clustering coefficient and network constraint) to *global* (eigenvector centrality).<sup>9</sup>

## 4. Initial simulations: empirical networks

### 4.1. Simulation procedure

Our approach involves simulating an error scenario on what we take to be a clean network,  $G(V, E)$ . We describe one simulation run for each error scenario below. For each error scenario, we executed

<sup>8</sup> Spearman's rho, which ranges from  $-1$  to  $1$ , gives the Pearson's correlation of the ranks of two variables.

<sup>9</sup> Network measures for a given node are considered local if their calculation only involves features of the focal node itself. For example, degree centrality is a count of a focal node's edges (Table 4). The calculation of semi-local measures, like clustering coefficient, involves the activity of a node's neighbors. Finally, a global network measure is calculated using properties of the entire network. For example, betweenness centrality requires collecting all shortest paths in a network.

10 runs. Because we observed very little variation in our results between runs, we omit error bars and confidence intervals from our results plots because they would be almost unobservable.<sup>10</sup>

**False negative nodes.** Given  $G(V, E)$ , at each step, we remove  $0.05|V|$  randomly chosen nodes from the network, yielding  $G'$ . For each node in both  $G$  and  $G'$ , we calculate the value of some node-level measure and store them in the vectors  $M$  and  $M'$ . We then calculate Spearman's rho for  $M$  and  $M'$ . We continue removing  $0.05|V|$  nodes until only  $0.05|V|$  are left in the network, recalculating and recording Spearman's rho at each step (removing all nodes would result in an empty graph).<sup>11</sup>

**False positive nodes.** We introduce spurious nodes to the initial network one by one. For each spurious node, we extend  $n$  edges ( $n$  is equal to the degree of a randomly chosen node already in  $G$ ) from itself to  $n$  randomly chosen nodes that are already in the network, preserving the network's average degree (Borgatti et al., 2006). Given  $G(V, E)$ , at each step, we add  $0.05|V|$  spurious nodes to the network. We calculate the rank correlation of  $M$  and  $M'$  at the end of each step. We continue attaching spurious nodes until  $|V|$  nodes have been added. After one run,  $G'$  will contain  $2|V|$  nodes.

**False negative edges.** Given  $G(V, E)$ , at each step, we remove  $0.05|E|$  randomly chosen edges from the network, yielding  $G'$  and calculating Spearman's rho for  $M$  and  $M'$ . We remove edges until  $0.05|E|$  remain. Removing edges has the effect of artificially diminishing a network's average degree.

<sup>10</sup> We attribute the small amount of variation between runs to the relatively large size of our empirical and randomly-generated networks.

<sup>11</sup> Another strategy for this simulation would be to first remove  $0.05|V|$  nodes from  $G$  to generate  $G'$  and calculate rank correlations, and then remove  $0.10|V|$  from our original  $G$  to generate a perturbed graph, and then  $0.15|V|$  nodes, and so forth. Our strategy to instead remove an additional  $0.05|V|$  nodes from our perturbed graphs each step is more computationally efficient. There is no reason that these two methods of simulation would yield different results. For example, the  $G'$  generated by removing  $0.10|V|$  nodes from  $G$  has the same probability of occurring as the  $G'$  generated by first removing  $0.05|V|$  nodes from  $G$  to produce  $G'$  and then another  $0.05|V|$  nodes from  $G'$ . More formally, while the probability of  $G'$  occurring given  $G$  and  $G'$  is higher than the probability of  $G'$  occurring just given  $G$ , we are only concerned with comparing  $G$  and  $G'$ . Thus, these two simulation methods would only generate different results if the probability of  $G'$  occurring given  $G$  varies between methods, which it does not.

**Table 5**  
Amount of error tolerated by network measures before rank correlation < 0.95.

Error scenario	Degree centrality	Clustering coefficient	Network constraint	Eigenvector centrality
<i>Slashdot network</i>				
False negative nodes	0.36	0.24	0.04	0.20
False positive nodes	0.04	0.76	0.04	0.96
False negative edges	0.16	0.08	0.03	0.09
False positive edges	0.04	0.76	0.04	0.44
False aggregation	0.16	0.24	0.12	0.16
False disaggregation	0.16	0.24	0.04	0.12
<i>Citation network</i>				
False negative nodes	0.96	0.16	0.24	0.96
False positive nodes	0.64	0.03	0.32	0.08
False negative edges	0.60	0.06	0.12	0.56
False positive edges	0.62	0.03	0.32	0.05
False aggregation	0.13	0.13	0.12	0.10
False disaggregation	0.24	0.10	0.10	0.52

**False positive edges.** Given  $G(V, E)$ , at each step, we add  $0.05|E|$  edges between randomly chosen unconnected pairs of nodes in the network, yielding  $G'$ , artificially increasing the network's average degree. We add edges until  $|E|$  edges have been added. At the end,  $G'$  will contain  $2|E|$  edges.

**False aggregation.** Given  $G(V, E)$ , it is possible to execute  $|V| - 1$  aggregations between all possible pairs of nodes in  $G$ . Aggregating, a pair of nodes entails randomly selecting two nodes from a network,  $A$  and  $B$ , removing  $B$ , and reattaching all of  $B$ 's neighbors to  $A$  (which node is removed is determined randomly). After  $|V| - 1$  merges, only one node would remain in the network. At each step,  $0.05(|V| - 1)$  merges are executed, yielding  $G'$ ; Spearman's rho is likewise calculated at the end of each step, and merges are executed until  $0.05(|V| - 1)$  possible merges remain.

**False disaggregation.** Disaggregating a node entails splitting a node,  $A$ , into two nodes  $A$  and  $B$ . We also randomly remove some of  $A$ 's neighbors and reattach them to the newly added isolate,  $B$  (each of  $A$ 's neighbors has a 0.50 probability of being reattached to  $B$ ). At each step,  $0.05|V|$  nodes are split and Spearman's rho is calculated. We continue splitting nodes until  $|V|$  splits have been executed, doubling the size of  $G$ .

## 4.2. Results

Fig. 2 reports the results of our error scenarios for three different networks: the empirical network of Slashdot users, a random graph with the same degree sequence as the empirical network, and an Erdős-Rényi random graph with the same number of nodes and edges as the empirical network. Fig. 3 contains the corresponding results for the ArXiv citation network.

Each 'cell' in Figs. 2 and 3 plots the rank correlation results for a given network measure under each of the six error scenarios in a given network. The x-axis is the proportion of the graph perturbed, which is specific to each error scenario. If proportion = 0.50 for the false positive nodes error scenario, this means  $0.50|V|$  nodes have been added, while under false negative nodes,  $0.50|V|$  nodes have been removed. For false positive edges, proportion = 0.50 signals that  $0.50|E|$  edges have been added whereas for false negative edges,  $0.50|E|$  edges have been removed. Under false aggregation, if proportion = 0.50,  $0.50(|V| - 1)$  pairs of nodes have been merged, and for false disaggregation,  $0.50|V|$  nodes have been split.

### 4.2.1. Comparing network measures

Comparing the rows in Figs. 2 and 3, degree centrality and eigenvector centrality (the first and last rows in both figures) appear to be the most robust measures under our error scenarios. Their similar behaviors are unsurprising because the two measures tend to be highly correlated. This supports the findings of Borgatti et al.

(2006), who show the reliability of their centrality measures suffered almost identically in their error simulations.

In our findings, eigenvector centrality is slightly less robust than degree centrality under false negative nodes and false negative edges. Eigenvector centrality is calculated based on the recursively weighted degrees of a focal node's neighbors, whereas degree centrality is simply the count of a focal node's neighbors (Table 4). Thus, an error scenario must target a node's immediate ties to affect its degree centrality. However, to affect the node's eigenvector centrality, an error scenario can perturb a node's immediate ties, or the ties of its neighbors, the ties of its neighbors' neighbors, and so forth.

In both empirical networks, clustering coefficient and network constraint are the less robust than degree centrality (Figs. 2 and 3). Again, the calculation of degree centrality is less dependent on the activity of a node's neighbors than the calculation of clustering coefficient or network constraint. We may generalize this result to argue that more global node-level network measures tend to be more robust to measurement error.

Given this intuition, though, why do clustering coefficient and network constraint appear to be less robust than eigenvector centrality, which is our most global measure (compare row 4 to rows 2 and 3 in Fig. 3)? We suggest that the formulas used to calculate some node-level measures are more sensitive to graph perturbations than others regardless of how local or global the measure is.

Consider the formula for clustering coefficient. A focal node that has two neighbors that also share a tie has a clustering coefficient = 1 (i.e. the maximum value of clustering coefficient). If the tie between the focal node's neighbors is removed, then the focal node's clustering coefficient = 0, which is the minimum value a clustering coefficient can take.<sup>12</sup> Thus, removing one edge has the potential to transform a top-ranked node (by clustering coefficient) into a bottom-ranked node. Because of the ways in which they are calculated, the rank change from removing one edge does not perturb eigenvector centrality and degree centrality as much, even though eigenvector centrality may require a more 'global' calculation.

### 4.2.2. Comparing error scenarios

False negative edges pose the biggest problem in almost every plot, whereas false negative nodes are not nearly as detrimental. This is because by removing edges at random, we are more likely

<sup>12</sup> A node  $A$  has two neighbors.  $A$ 's clustering coefficient is calculated by dividing the number of ties that exist between its neighbors by the number of ties that could exist between them. If  $A$ 's two neighbors share a tie, their clustering coefficient is  $1/1 = 1$ . If they do not share a tie, their clustering coefficient is  $0/1 = 0$ .

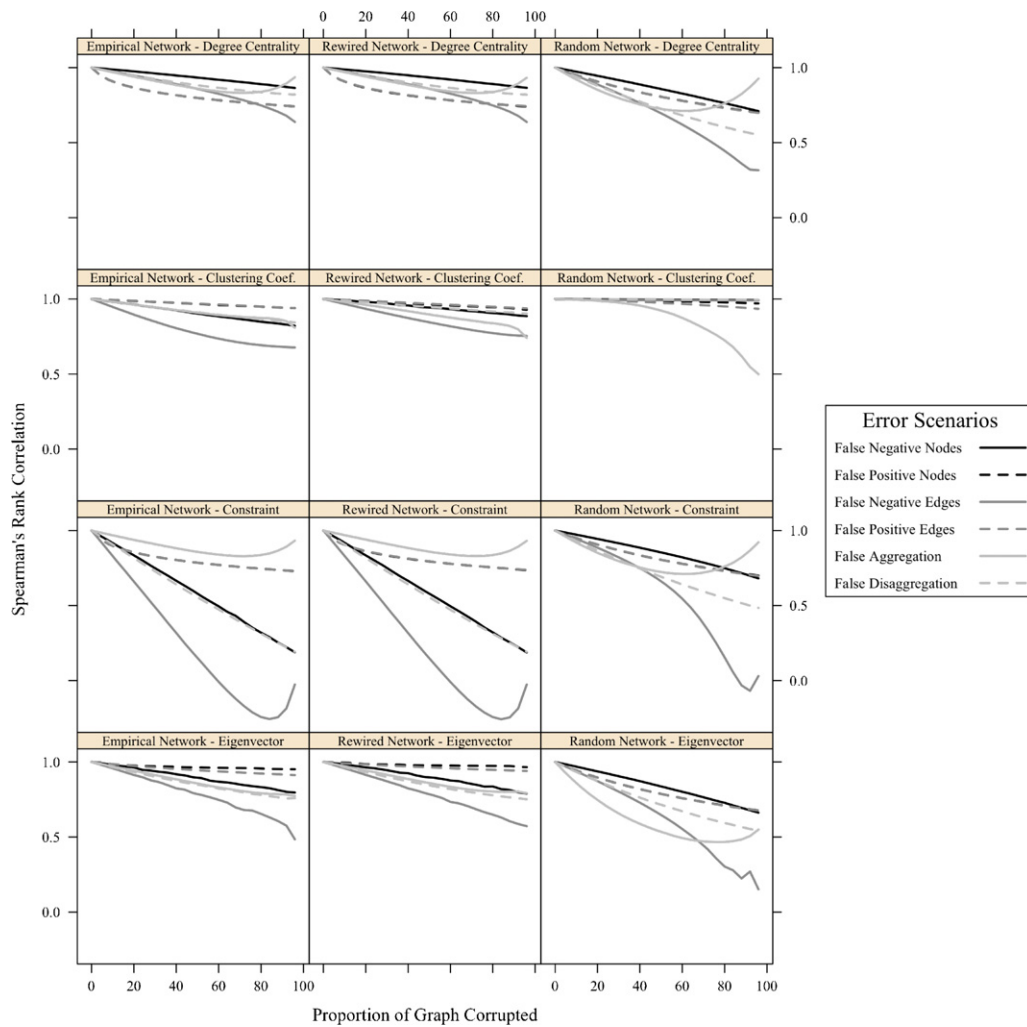


Fig. 2. Measurement error simulation results for Slashdot network, rank correlations.

to remove edges that are attached to high-degree nodes. The false negative nodes scenario, on the other hand, removes nodes at random, most of which are low-degree nodes because of the graphs' positively-skewed degree distributions. Thus, whereas every node has an equal probability of being removed under false negative nodes, the edges of a high-degree nodes are targeted under false negative edges. Because these high-degree nodes are often critical structural features of an empirical network, an error scenario like false negative edges can affect network measures far more than an error scenario like false negative nodes, which targets each node with the same probability.

On the other hand, most node-level network measures are relatively robust to false positive edges, their rank correlations for the different networks in Figs. 2 and 3 hovering at 0.70 or higher even when half the network is composed of either false positive nodes or edges. This suggests that during data collection, for a better approximation of the node-level measures, using a lower tie-strength threshold for defining edges is more desirable, i.e. it is better to have false positive edges than false negative edges.

One exception is the effect of false positive edges on clustering coefficients in the citation network, which appears to be just as damaging as false negative edges (Fig. 3). Again, this can be explained by the sensitivity of the calculation of clustering coefficient to missing and spurious edges (adding an edge between

the two unconnected neighbors of a node can increase the node's clustering coefficient from 0 to 1).

Finally, the effects of false aggregation and false disaggregation lie between the effects of false negatives and false positives. This is likely because false aggregation introduces the combined effects of false negative nodes and false positive edges, whereas false disaggregation reflects the effects of false negative edges and false positive nodes.<sup>13</sup>

While their effects are similar, in some cases, false aggregation renders network measures slightly less robust than false disaggregation. However, unless over 80% of nodes in the empirical networks are falsely aggregated or disaggregated, their effects on the reliability of most network measures is moderate – at their worst, false aggregation and false disaggregation reduces the rank correlation of most of our network measures to 0.50.

#### 4.2.3. Comparing networks with varying structural features

Our results also indicate that error scenarios have similar effects on network measures for our empirical and rewired networks. That the empirical networks and their rewired versions react to error

<sup>13</sup> False aggregation involves removing a node, and erroneously attaching the removed node's edges to another node. False disaggregation involves removing edges from an existing node to attach to the newly-added spurious node.

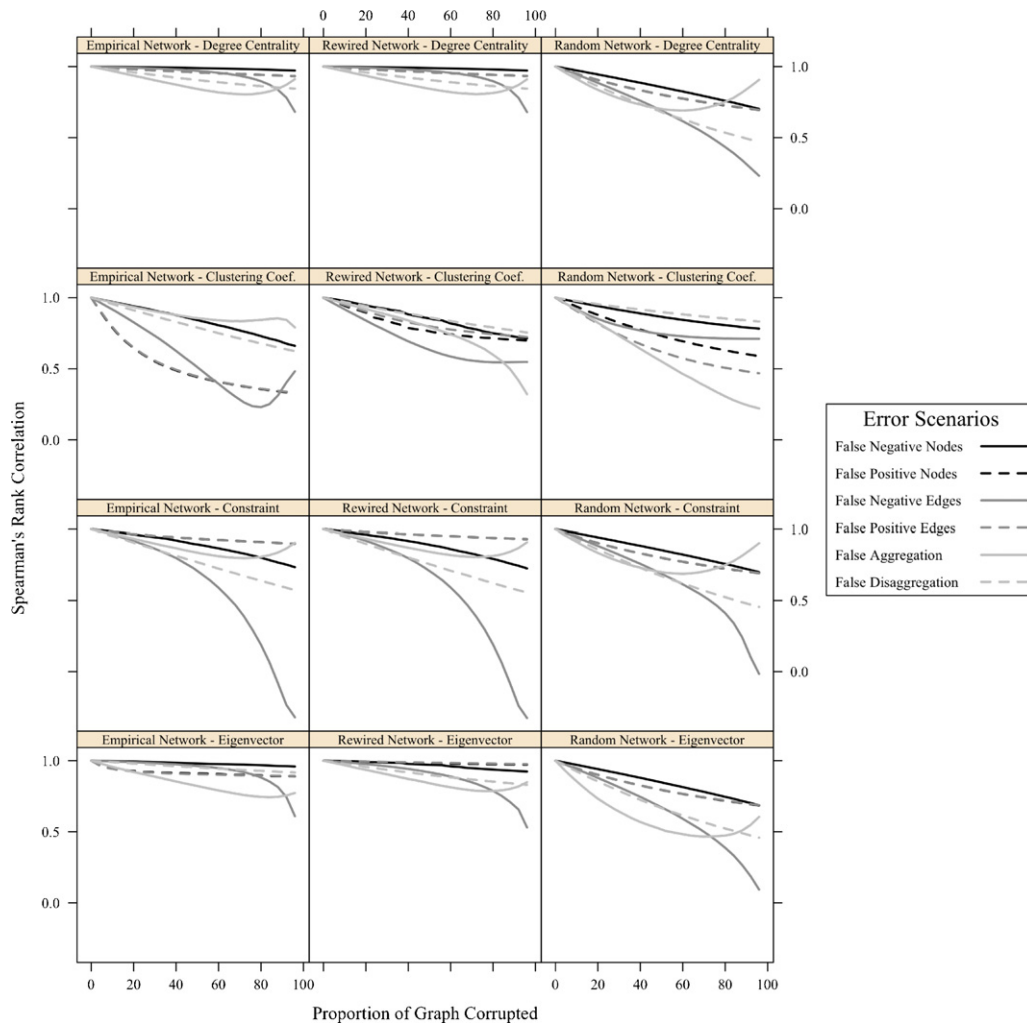


Fig. 3. Measurement error simulation results for ArXiv citation network, rank correlations.

scenarios similarly suggests that the degree distribution of a network might account for its robustness to certain measurement error scenarios.<sup>14</sup>

In one exception, the effects of our measurement errors on clustering coefficient in the citation network are similar for the rewired network and the Erdős–Rényi network (see row 2 of Fig. 3), but different for the empirical citation network. This result can be explained by the high average clustering associated with the empirical citation network compared with the low clustering in the rewired and Erdős–Rényi networks.

Ultimately, these results suggest that the measurement error simulations on Erdős–Rényi random graphs in Borgatti et al. (2006) would benefit from comparison to error simulations conducted

with empirical networks, or at least random networks that have some real-world structural features.

## 5. Further simulations: node subsets and graph structure

In this section, we further investigate the relationship between a network's structural features and its robustness to measurement error. First, we examine differences in how error scenarios affect different subsets of nodes. Second, we simulate the effects of error scenarios on random networks with different degree distributions and average clustering.

### 5.1. Highly-ranked node subsets

In many empirical networks, the distribution of nodes by a given network metric tends to be positively skewed. This especially true in both the Slashdot and citation networks (Fig. 1). Similarly, the distributions of clustering coefficient and network constraint measures among nodes also exhibit positive skew (see Table 4).

By virtue of these skewed distributions, an error scenario would not be able to alter greatly, for example, the degree centrality rank of the many low-ranked nodes in a network. In other words, highly-ranked nodes have more to lose. As such, we would expect the rank

<sup>14</sup> The rank correlations for centrality measures stay close to 0.9 even when almost the entire ArXiv citation network is corrupted (Fig. 3). This is intuitive because in this network, there are many one-degree nodes, a few nodes with hundreds of edges, and even fewer nodes in between. Thus, for a scenario like false positive edges to greatly affect rank correlation in degree centrality, the addition of edges would have to concentrate on a few low-ranked nodes, rather than being randomly distributed. Most low-ranked nodes tend to stay low-ranked. As a result, because the calculation of rank correlation uses the vectors of the measures from all nodes in network, most of which low-ranked (i.e. most have one or zero degrees), it is not surprising that the rank correlation here is so high.



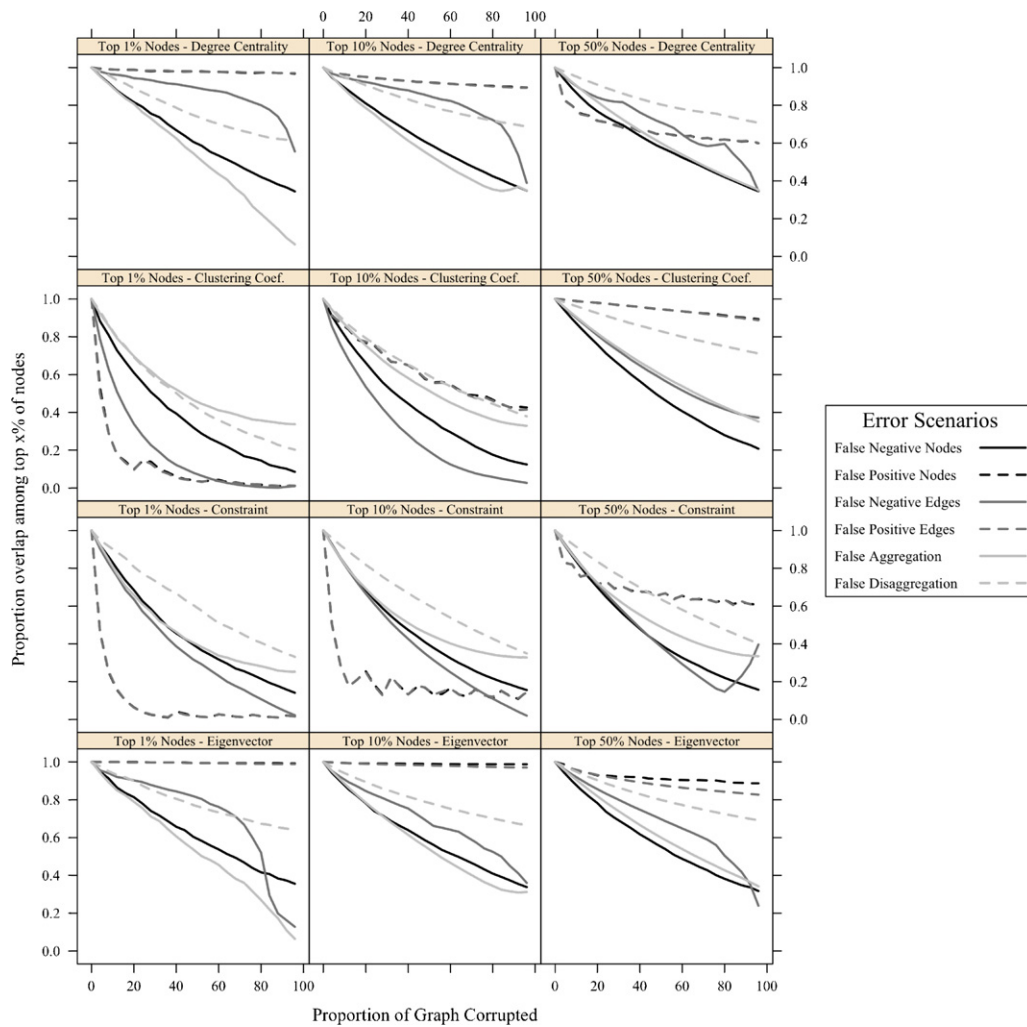


Fig. 4. Measurement error simulation results for Slashdot network, top-ranking node subset overlap.

correlation of a measure for only high-ranking nodes in  $G$  and those same nodes in  $G'$  to be much lower than the rank correlation of a measure for *all* common nodes in  $G$  and  $G'$ .

We conduct further simulations with the Slashdot and citation networks to investigate this question. For each network measure, we calculate the overlap for the top 1% of nodes (by some node-level measure) in  $G$  and  $G'$ . For instance, we obtain the top 1% of nodes by degree centrality in  $G$  and the top 1% of nodes in  $G'$ . We then compute the proportion of these top 1% nodes in  $G$  that remain among the top 1% of nodes in  $G'$ . In the error scenarios in which nodes are removed – false negative nodes and false aggregation – we compute the proportion using those top 1% of nodes in  $G$  that were not removed.<sup>15</sup> We run the same measurement error simulations as described above and report these proportions for our three node-level network measures among the top 1%, 10%, and 50% of nodes in Figs. 4 and 5.

Comparing across columns in both Figs. 4 and 5, overlaps appear to be lower when using the top 1% of nodes as compared with using the top 50% of nodes. As we suggested, this implies that error scenarios affect top ranking nodes far more than middle- or

low-ranking nodes.<sup>16</sup> Thus, in our baseline results reported in Figs. 4 and 5, much of the variation in rank correlation can be attributed to perturbations among top-ranking nodes.

Our results also display one exception, which is that false positive edges appear to affect the degree centrality of top-ranking nodes less than lower-ranking nodes (see row 1 in both Figs. 4 and 5). As mentioned, the positively-skewed degree distributions of the Slashdot and citation networks implies that only a few nodes have more than one neighbor. Thus, when examining the top 50% of nodes, many nodes that previously had only one neighbor have a great deal more to gain in terms of ranking by degree centrality than the already top-ranked nodes.

## 5.2. Degree distribution

Our results also indicate that a network's degree distribution might account for the robustness of certain network metrics. Namely, networks with more skewed degree distributions have

<sup>15</sup>  $D$  is a set containing the top  $x\%$  of nodes in  $G$  by some node-level network measure.  $D'$  contains the top  $|D|$  nodes in  $G'$  by the same network measure. We compute our overlap score by taking  $|D \cap D'|/|D|$ .

<sup>16</sup> We do not compare measures for the top 10% of nodes, middle 10%, and bottom 10%, for instance, because the extreme positive skew for many of our node-level measures would indicate that the middle and bottom 1% of nodes have the same values in  $G$  and  $G'$ . Using the top 1%, top 10%, and top 50%, we show that the rank changes of larger node subsets that include more 'less elite' nodes are less volatile than the rank changes for smaller subsets of 'more elite' nodes.

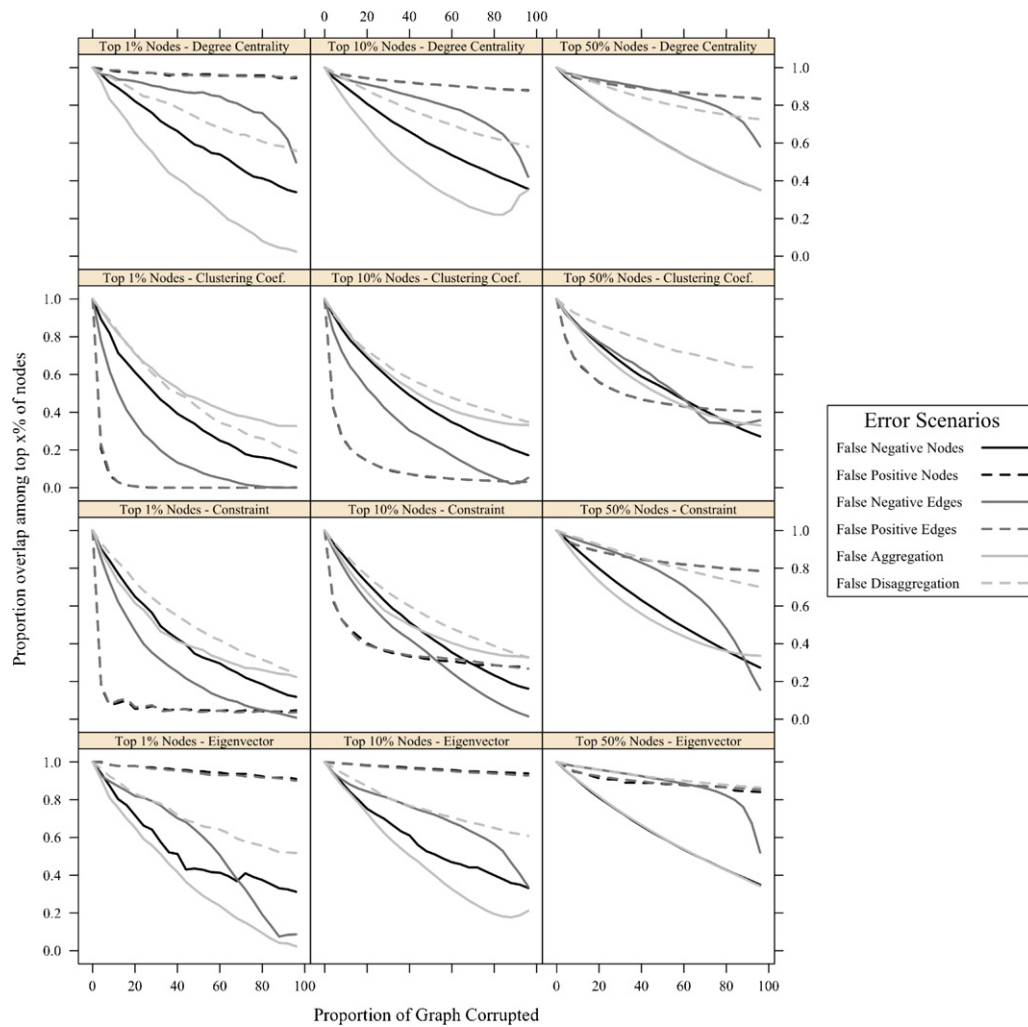


Fig. 5. Measurement error simulation results for ArXiv citation network, top-ranking node subset overlap.

smaller subsets of highly-ranked nodes and larger subsets of nodes with degrees equal to 1.

We examined this question by generating three random preferential attachment graphs, in order of increasing positive skewness in their degree distributions. Specifically, we increased the power law scaling parameter (otherwise known as  $\alpha$ ) (Albert and Barabasi, 2002), which resulted in one graph where  $\alpha=2.5$ , a second where  $\alpha=3.0$ , and finally, a third graph, in which  $\alpha=3.5$ .<sup>17</sup>

Our results in Fig. 6, show evidence that graphs with less skewed degree distributions are more resistant to measurement error. Specifically, the robustness of degree centrality to false positive nodes and edges diminishes as the degree distribution of a network becomes more positively skewed (see row 1 of Fig. 6). Here, adding an edge to a low-degree node affects its degree centrality ranking more than adding an edge to a high-degree node. Given that there are more low-degree nodes in graphs with more positively-skewed degree distributions, the probability of a spurious edge being randomly attached to a low-degree node is also higher, which results in greater changes in degree centrality rank.

The reliability of clustering coefficient suffers more under every error scenario as the positive skew of a graph's degree distribution increases. Random edge removal tends to target the edges of those nodes that have many neighbors. In graphs with more positively-skewed degree distributions, these high-degree nodes tend to have a higher proportion of the graph's edges and thus, are more important to the network's overall connectedness. As a result, edge removal tends to affect these high-degree nodes' ego-networks the most. In networks with less-skewed degree distributions, false negative edges would affect the edges of the network's nodes more uniformly.

### 5.3. Average clustering

The Slashdot and citation networks also differ in average clustering (Slashdot = 0.056; citation = 0.285, see Table 3). We generated three 10,000-node random networks with the same power law degree distribution ( $\alpha=2.5$ ), same densities (density = 0.001), and average clustering coefficients of 0.01, 0.20, and 0.40.<sup>18</sup> As expected, altering the average clustering of a graph did not

<sup>17</sup> We generated these random power law networks using the 'ba.game' function from the Python package *iGraph* (Csardi and Nepusz, 2006). The 'ba.game' function allows for the creation of networks equal in the number of nodes and edges but varying in the scaling parameter for preferential attachment.

<sup>18</sup> We generated these random networks with different average clustering using the 'powerlaw\_cluster\_graph' function from the Python package *networkx* (Hagberg et al., 2008). The 'powerlaw\_cluster\_graph' function allows for the generation of networks with fixed size, level of preferential attachment, and average clustering.

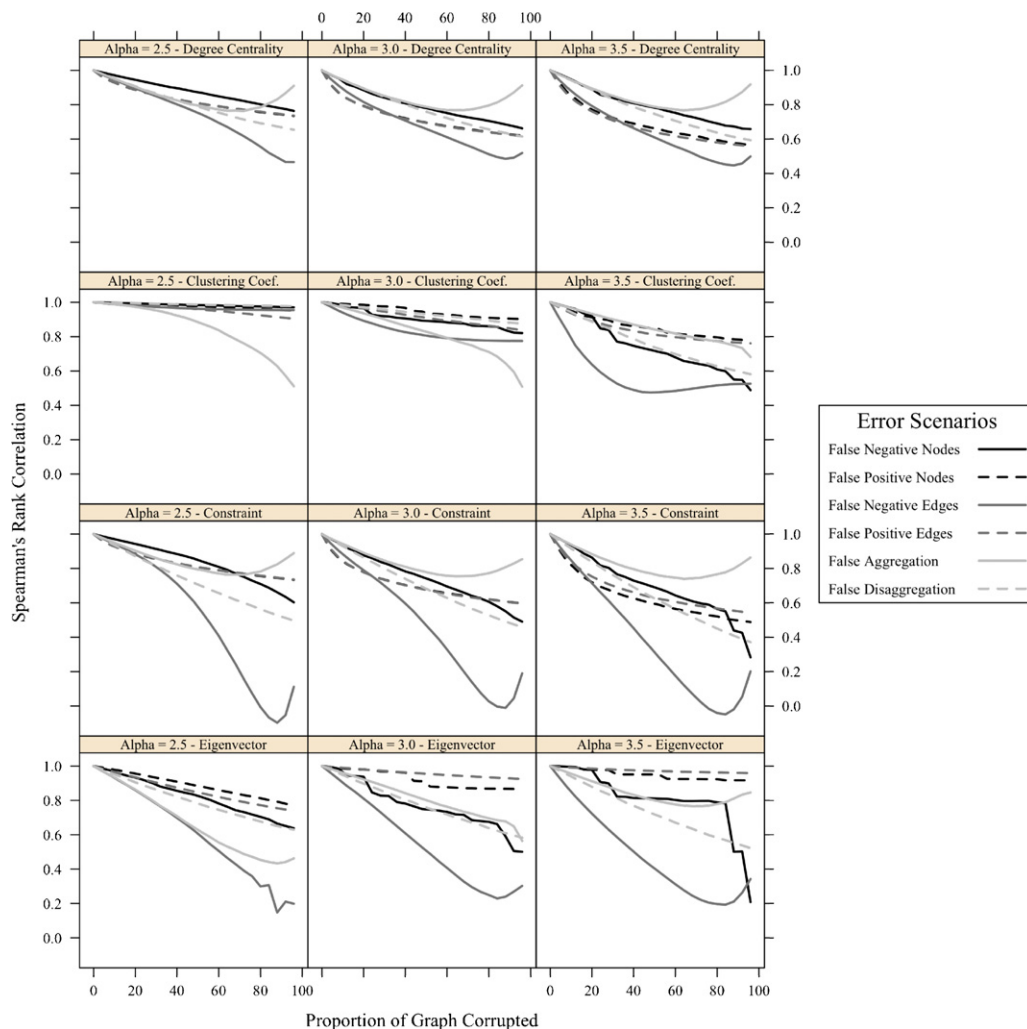


Fig. 6. Measurement error simulation results for random preferential attachment graphs, varying degree distributions, rank correlations.

yield dramatic differences in the robustness of degree centrality, eigenvector centrality, and network constraint to our error scenarios. Variation in these three measures depend more on the number of neighbors attached to a focal node whereas the calculation of clustering coefficient depends on the configuration of ties between a focal node's neighbors.

The clustering coefficients in our simulated networks with higher average clustering were less robust to false positive nodes and edges than in networks with less clustering (see Fig. 7). Consider a focal node that has two neighbors that share a tie. Removing the edge between its two neighbors reduces its clustering coefficient from 1 to 0, while then adding a neighbor that has no other ties diminishes the focal node's clustering coefficient from 1 to 1/3. Because these scenarios are more likely in highly clustered networks, such networks are also more sensitive to measurement error.

According to these results, graphs with low clustering and less positively skewed degree distributions tend to be more resistant to our error scenarios. This underscores the importance of how error scenarios affect empirical networks and Erdős-Rényi random graphs differently. Because Erdős-Rényi networks have little

clustering and more uniform degree distributions, simulating measurement error using such networks can underestimate their actual effects.

## 6. Discussions and conclusion

### 6.1. Summary

In this article, we expanded on prior work by examining a variety of measurement errors often overlooked in network research (e.g. false positive nodes and edges and the false (dis)aggregation of nodes). In addition, we compared a wider assortment of node-level network measures (degree centrality, clustering coefficient, network constraint, and eigenvector centrality), testing their robustness to our different forms of measurement error. We also investigated network-structural properties (average clustering, degree distributions) as explanations for the varying effects of measurement error. Below, we summarize our main results, recommend error correction strategies, and anchor the discussion to examples of commonly used network datasets.

#### 6.1.1. The contingent impact of error scenarios

Table 5 summarizes the robustness of our four network measures to measurement error by reporting the amount of error that

See Holme and Kim (2002) for a description of the algorithm implemented in this function.

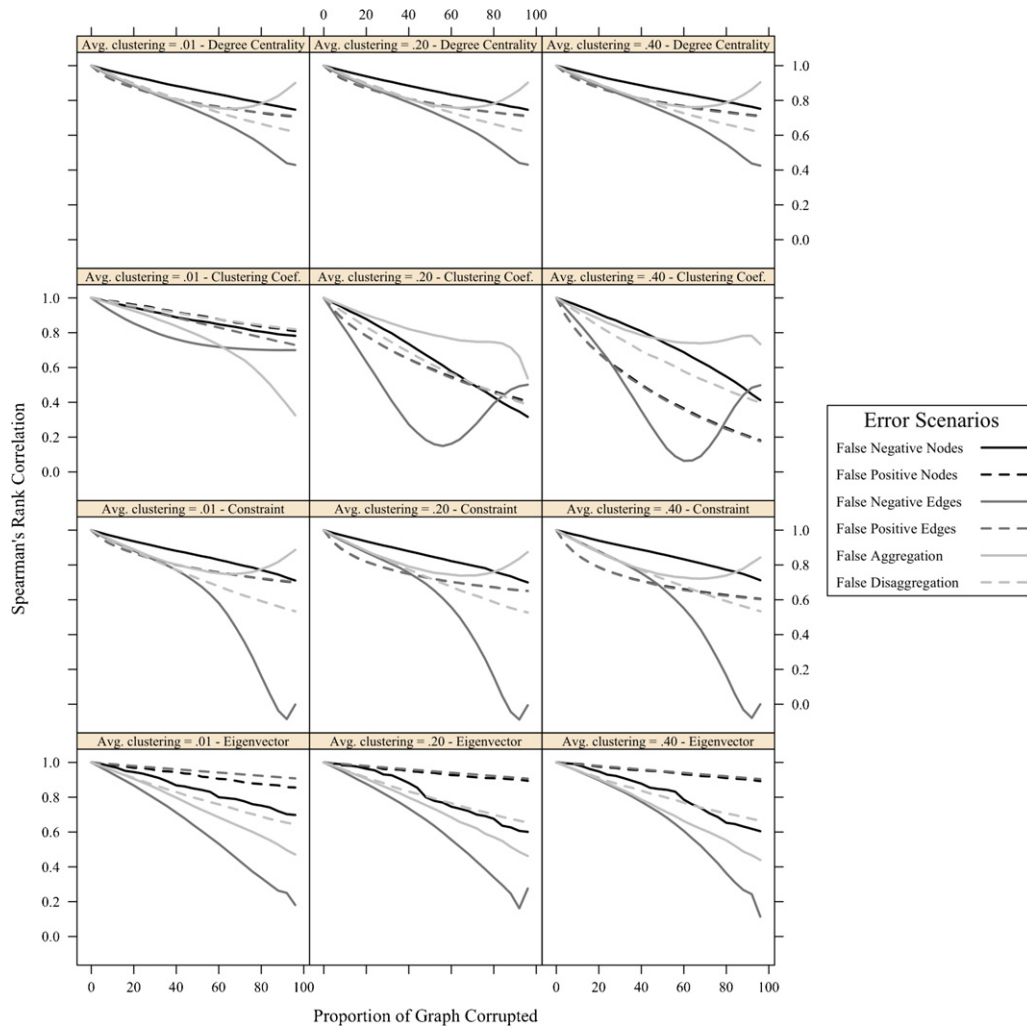


Fig. 7. Measurement error simulation results for random preferential attachment graphs, varying average clustering, rank correlations.

one of our empirical networks can sustain before the rank correlation of a node-level measure (in the perturbed and unperturbed networks) falls below 0.95.<sup>19</sup> We chose 0.95 as a cutoff because it is arguably high enough that any bias introduced by measurement error in a measure would be trivial. In addition, Table 6 shows that there is great variation in the amount that graph must be corrupted before the rank correlation for a measure reaches 0.95.

While generally, we find that networks with low average clustering and less positively-skewed degree distributions are most resistant to measurement error, our results contain important nuances. Unlike Borgatti et al. (2006), we do not observe that missing nodes and edges are consistently more damaging than spurious nodes and edges. For instance, the Slashdot network requires more error in the form of false positive edges than false negative edges to diminish the reliability of eigenvector centrality, while the opposite is true in the citation network (Table 5).

We explain this contradiction by looking to differences in network-structural features. In networks with less positively-skewed degree distributions, false positive and negative edges tend to have similar effects on the reliability of network measures. In contrast, in graphs with more positively-skewed degree distributions, false negative edges cause greater harm than false positive

edges (Fig. 6). In general, these results suggest that erring on the side of representing too many weak ties as real ties makes for more reliable network measures than including only stronger ties.

The effects of false aggregation and false disaggregation also vary with the graph structural features. For example, in graphs with less positively-skewed degree distributions, false aggregation poses a bigger problem than false disaggregation to clustering coefficient and eigenvector centrality (Fig. 6). However, in graphs with more positively skewed degree distributions, this pattern is reversed. Similarly, in graphs with low average clustering, false aggregation diminishes the reliability of clustering coefficient more than false disaggregation, while the opposite holds in graphs with high average clustering.

Table 6  
Forms of measurement error present in common network datasets.

Error scenario	Facebook friendship network	ISI Web of Science citation network	Add Health friendship network
False negative nodes	+	+	+
False positive nodes	+	+	+
False negative edges	+	+	+
False positive edges	+		+
False aggregation		+	
False disaggregation	+	+	

Note: '+' signifies that network dataset suffers from this type of measurement error.

<sup>19</sup> Recall that these proportions reported in Table 6 have different meanings specific to different measurement error scenarios (see Section 4.1).

Finally, we find support for the assertion by Borgatti et al. (2006) that centrality measures are similarly robust to measurement error. Curiously, though, the reliability of less ‘global’ measures like clustering coefficient suffers more compared with more ‘global’ measures like eigenvector centrality. We suspect that the differences in the robustness of our four network metrics has to do with the steps involved in calculation. For instance, the sensitivity of a node’s clustering coefficient to the removal (or addition) of a single edge (or neighbor) is far greater than that of a node’s degree or eigenvector centrality.

## 6.2. Correction strategies

### 6.2.1. Data collection and cleaning

Despite our conclusions above, the network researcher is often not afforded the luxury of choice between network measures or types of networks to use for an empirical analysis. However, based on our results, we can make several recommendations with regard to error correction strategies.

Rather than gathering additional data or cleaning and entire network dataset, cleaning certain node subsets would improve measurement reliability more than focusing on other subsets. Our results in Figs. 4 and 5 indicate that the reliability of network measures for ‘elite’ nodes generally drops more as a result of measurement error than for other node subsets. This is especially useful when data collection or cleaning is labor- and time-intensive and resources are scarce.

Suppose, for instance, that a network dataset suffers from random missing edges. If the degree distribution of the network is positively skewed, then most of the missing edges likely belong to high-degree nodes. Gathering more complete data for highly active nodes is a far better strategy than attempting to collect complete data for all nodes. As shown in Figs. 4 and 5, even collecting data for the top 1% of nodes for a 10,000 node network would constitute a major improvement to the reliability of network measures.

In addition, our results suggest that although the false aggregation and disaggregation of nodes impose the same degree of measurement error, false disaggregation appears to pose less of a problem under some scenarios. Thus, as mentioned, when conducting entity resolution on nodes, stricter matching rules should be employed. Also, when setting a threshold for tie strength, using a lower threshold, which yields more false positive edges, results in more reliable measures than higher thresholds, which yield more false negative edges.

### 6.2.2. Network imputation

While the imputation of network data remains an important error correction strategy for missing nodes or edges (Guimera and Sales-Pardo, 2009; Huisman, 2009; Kim and Leskovec, 2011), our results suggest that it should only be used in scenarios where false negatives are more detrimental than false positives. For example, according to Fig. 2, in the Slashdot network, false negative nodes diminish the reliability of clustering coefficient, network constraint, and eigenvector centrality more than false positive nodes. The imputation of missing nodes would then stand as viable correction strategy for these measures. However, false positive nodes are worse for degree centrality than false negative nodes. Thus, imputation could introduce even greater measurement error with the presence of spurious nodes.

## 6.3. Implications for empirical network research

### 6.3.1. Implications and examples

Table 6 summarizes three examples of common network datasets, identifying the types of measurement error that they

might face. Before citation networks can be generated from publication datasets like the ISI Web of Science, they often require large-scale entity resolution which automates the identification of unique authors (Table 6). According to our results in Table 6, if an entity resolution algorithm leaves 13% of the nodes in a network as improperly matched (false aggregation), then the reliability of semi-local node-level measures like clustering coefficient or network constraint diminishes only marginally (Spearman’s  $\rho = 0.95$ , Table 6).

A second example concerns a social network that suffers from respondent recall or survey design bias. The National Longitudinal Study of Adolescent Health (Add Health) dataset contains friendship network data that likely suffers from false negatives. Artificially limiting the size of an individual’s reported ego-network introduces false negative edges, which can severely affect semi-local measures like network constraint even if just 20% of a network’s edges are missing (Fig. 2). We caution the reader though, that compared to our networks under study, the Add Health data contain a much smaller network, which can make it even more sensitive to measurement error (Borgatti et al., 2006).

Finally, given the abundance of online community network data available, researchers must also be sensitive to false positive nodes and edges in their dataset. As mentioned, an estimated 27% of all accounts on Facebook are fake, which can make measures of clustering or brokerage wholly unreliable (see results in Fig. 4).

### 6.3.2. Future directions

While we have ventured a systematic comparison of six different measurement error scenarios, there is much that we have not covered. We have little intuition about how the size of a network influences the robustness of its measurements to error scenarios. Borgatti et al. (2006) identify density as an important graph-level feature, but we suspect that their results may in part be driven by the sizes of the networks they consider. To a great extent, paying attention to the interaction between a network’s size and structural features would add greater insight to any analysis of network measurement error.

Also, our results are based on random perturbations of our networks (i.e. the random removal or addition of edges), and should only be taken as a baseline. In many cases, measurement error is distributed non-randomly throughout networks. Limiting the contact list of a sociometric survey response, for example, would affect individuals with more ties than they can list, but not those with few contacts. While we compare across measurement error scenarios, we encourage researchers to investigate different variations of the same measurement scenario.

In addition, we have not analyzed the direction of the bias engendered by our measurement error scenarios, nor have we touched on changes in the empirical distribution of a given network measure as a result of measurement error. Both these features can affect the perceived relationship of network measures to non-network outcomes. As such, we view our work as a springboard for further research on the statistical implications of using error-afflicted network measures.

## References

- Ackland, R., 2009. Social network services as data sources and platforms for e-researching social networks. *Social Science Computer Review* 24 (August (4)), 481–492.
- Ahuja, G., 2000. Collaboration networks, structural holes, and innovation: a longitudinal study. *Administrative Science Quarterly* 45 (January (3)), 425–465.
- Albert, R., Barabasi, A.-L., 2002. Statistical mechanics of complex networks. *Review of Modern Physics* 74 (1), 47–97.
- Azoulay, P., Zivin, J.G., 2005. Peer effects in the workplace: evidence from professional transitions for the superstars of medicine. Working Paper.
- Barnes, J.A., 1979. Network analysis: orienting notion, rigorous technique, or substantive field of study? *Perspectives on Social Network Analysis*, 402–423.

- Bernard, H.R., Killworth, P., Kronenfeld, D., Sailer, L., 1984. The problem of informant accuracy: the validity of retrospective data. *Annual Review of Anthropology* 13, 495–517.
- Bonacich, P., 1987. Power and centrality: a family of measures. *American Journal of Sociology* 92, 1170–1182.
- Borgatti, S.P., Carley, K.M., Krackhardt, D., 2006. On the robustness of centrality measures under conditions of imperfect data. *Social Networks* 28, 124–136.
- Brewer, D.D., 2000. Forgetting in the recall-based elicitation of personal and social networks. *Social Network* 22, 29–43.
- Burt, R.S., 1984. Network items and the general social survey. *Social Networks* 6 (4), 293–339.
- Burt, R.S., 1992. *Structural Holes. The Social Structure of Competition*. Harvard University Press, Cambridge, MA.
- Carley, K.M., Skillicorn, D., 2005. Special issue on analyzing large scale networks: the Enron corpus. *Computational & Mathematical Organization Theory* 11, 179–181.
- Coleman, J., 1961. *Social Climates in High Schools*. U.S. Dept. of Health, Education, and Welfare, Office of Education.
- Costenbader, E., Valente, T.W., 2003. The stability of centrality measures when networks are sampled. *Social Networks* 25, 283–307.
- Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695, gute Alternative zu UCI-Net und Pajek. Liest die Standard-Dateiformate aus anderen SNA-Tools. Visualisierungen mit der gewohnten Qualität der url R-Grafik-Ausgabe. <http://igraph.sf.net>.
- De Choudhury, M., Mason, W.A., Hofman, J.M., Watts, D.J., 2010. Inferring relevant social networks from interpersonal communication. In: *WWW'10: Proceedings of the 19th International Conference on World Wide Web*. ACM, New York, NY, USA, pp. 301–310.
- Erickson, B., 1978. Some problems of inference in chain data. *Sociological Methodology*, 276–302.
- Feld, S.L., Carter, W.C., 2002. Detecting measurement bias in respondent reports of personal networks. *Social Networks* 24, 365–383.
- Fleming, L., Frenken, K., 2007. The evolution of inventor networks in the silicon valley and boston regions. *Advances in Complex Systems* 10 (1), 53–71.
- Fleming, L., Mingo, S., Chen, D., 2007. Collaborative brokerage, generative creativity, and creative success. *Administrative Science Quarterly* 52, 443–475.
- Freeman, L.C., 1979. Centrality in social networks: conceptual clarification. *Social Networks* 1 (3), 215–239.
- Gehrke, J., Ginsparg, P., Kleinberg, J., 2003. Overview of the 2003 kdd cup. *SIGKDD Explorations Newsletter* 5 (2), 149–151.
- Granovetter, M., 1976. Network sampling: some first steps. *The American Journal of Sociology* 81 (6), 1287–1303.
- Guimera, R., Sales-Pardo, M., 2009. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences* 106 (52), 22073–22078.
- Hagberg, A.A., Schult, D.A., Swart, P.J., 2008. Exploring network structure, dynamics, and function using NetworkX. In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA USA, pp. 11–15.
- Handcock, M.S., Gile, K.J., 2010. Modeling social networks from sampled data. *Annals of Applied Statistics* 4 (1), 5–25.
- Holland, P., Leinhardt, S., 1973. The structural implications of measurement error in sociometry. *Journal of Mathematical Sociology* 3, 85–111.
- Holme, P., Kim, B.J., 2002. Growing scale-free networks with tunable clustering. *Physical Review E* 65 (2), 026107.
- Huisman, M., 2009. Imputation of missing network data: some simple procedures. *Journal of Social Structure* 10 (1), 1–29.
- Kapferer, B., 1969. Norms and the manipulation of relationships in a work context. *Social Networks in Urban Situations*, 181–244.
- Kim, M., Leskovec, J., 2011. The network completion problem: inferring missing nodes and edges in networks. In: *SDM*, pp. 47–58.
- Kossinets, G., 2006. Effects of missing data in social networks. *Social Networks* 28, 247–268.
- Laumann, E., Marsden, P., Prensky, D., 1983. The boundary specification problem in network analysis. *Applied Network Analysis: A Methodological Introduction*, 18–34.
- Leskovec, J., Faloutsos, C., 2006. Sampling from large graphs. In: *KDD'06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, pp. 631–636.
- Leskovec, J., Huttenlocher, D., Kleinberg, J., 2010. Signed networks in social media. In: *In Proc. 28th CHI*.
- Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., Christakis, N., 2008. Tastes, ties, and time: a new social network dataset using facebook.com. *Social Networks* 30, 330–342.
- Marsden, P.V., 1990. Network data and measurement. *Annual Review of Sociology* 16, 435–463.
- Narayanan, A., Shmatikov, V., 2009. De-anonymizing social networks. *IEEE Symposium on Security and Privacy* 0, 173–187.
- Nardi, B., Harris, J., 2006. Strangers and friends: collaborative play in world of warcraft. In: *CSCW'06: Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*. ACM, New York, NY, USA, pp. 149–158.
- Newman, M.E.J., 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* 98 (2), 405.
- Newman, M.E.J., 2002. Who is the best connected scientist? a study of scientific coauthorship networks. Working Paper, pp. 1–32.
- Newman, M.E.J., Park, J., 2003. Why social networks are different from other types of networks. *Physical Review E* 68 (September (3)), 036122.
- Nyberg, S., February 5, 2010. Fake accounts in facebook – how to counter it. <http://ezinearticles.com/?Fake-Accounts-in-Facebook-How-to-Counter-It&id=3703889>.
- Pool, I., Kochen, M., 1978. Contacts and influence. *Social Networks* 1, 5–52.
- Richmond, R., May 2, 2010. Stolen facebook accounts for sale. [http://www.nytimes.com/2010/05/03/technology/internet/03facebook.html?\\_r=1](http://www.nytimes.com/2010/05/03/technology/internet/03facebook.html?_r=1).
- Sadikov, E., Medina, M., Leskovec, J., Garcia-Molina, H., 2011. Correcting for missing data in information cascades. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. WSDM'11*. ACM, New York, NY, USA, pp. 55–64. <http://doi.acm.org/10.1145/1935826.1935844>.
- Shi, X., Leskovec, J., McFarland, D.A., 2010. Citing for high impact. In: *JCDL'10: Proceedings of the 10th Annual Joint Conference on Digital Libraries*. ACM, New York, NY, USA, pp. 49–58.
- Stork, D., Richards, W.D., 1992. Nonrespondents in communication network studies. *Group & Organization Management* 17, 193–209.
- Sudman, S., 1985. Experiments in the measurement of the size of social networks. *Social Networks* 7, 127–151.
- Watts, D., Strogatz, S., 1998. Collective dynamics of 'small-world' networks. *Nature* 393 (6684), 440–442.
- Wuchty, S., 2009. What is a social tie? *Proceedings of the National Academy of Sciences* 106 (36), 15099–15100.
- Wuchty, S., Jones, B.F., Uzzi, B., 2007. The increasing dominance of teams in production of knowledge. *Science* 316, 1036–1039.