Social Networks 34 (2012) 493–505

Contents lists available at SciVerse ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Random errors in egocentric networks

Zack W. Almquist*

Department of Sociology, University of California, Irvine, 3151 Social Science Plaza A, Irvine, CA 92697-5100, United States

ARTICLE INFO

Keywords:
Egonet
Egocentric network
Error
Network error
Facebook

ABSTRACT

The systematic errors that are induced by a combination of human memory limitations and common survey design and implementation have long been studied in the context of egocentric networks. Despite this, little if any work exists in the area of random error analysis on these same networks; this paper offers a perspective on the effects of random errors on egonet analysis, as well as the effects of using egonet measures as independent predictors in linear models. We explore the effects of false-positive and false-negative error in egocentric networks on both standard network measures and on linear models through simulation analysis on a ground truth egocentric network sample based on facebook-friendships. Results show that 5–20% error rates, which are consistent with error rates known to occur in ego network data, can cause serious misestimation of network properties and regression parameters.

1. Introduction

The specter of *error* appears within all measurements of the natural world, including the physical world (heat, geology, etc.), and the social world (friends, coworkers, etc.); however, the issue of error is never more present than in the research of social networks (Butts, 2003). In the elicitation of social network information, one can induce error in both the edges used to represent social ties and in the vertices that represent individual people or organizations. One of the most common forms of elicitation for local networks is that of the so-called *egocentric network* or *egonet*. An egonet is produced when a researcher acquires the *neighborhood* of a *focal actor* by characterizing the *alters* of an *ego*. A typical extension of this method is one which attains *all* the relations between ego's alters.

Research on error in the context of egocentric networks has placed a great deal of emphasis on the systematic mismeasurements induced by the interaction between aspects of human memory and the design and instrumentation of surveys (Marsden, 2002, 2003, 1990; Vehovar et al., 2008; Burt, 1984; Brewer, 2000; Brewer and Garrett, 2001). Several studies have demonstrated that a substantial amount of systematic error, induced in the collection of egocentric data, occurs as a result of a variety of cognitive mechanisms (e.g., forgetting) and suggest that there is no systematic method for predicting the severity of these errors in any given context (Brewer, 2000). Brewer and colleagues have demonstrated a number of different cognitive mechanisms which inhibit and/or bias both the elicitation of alters and alter–alter ties, and explore how this affects a variety of network measures such as degree, density, and centrality (Brewer and Webster, 1999; Brewer, 2000).

Paradoxically, given these results, there has been little, if any, work characterizing the effects of simple random error on egocentric networks or on the effects of random error on standard Graph Level Indices (GLI) such as mean degree, triad census, centralization (Anderson et al., 1999), or Node Level Indices (NLI) including degree or other centrality measures (Wasserman and Faust, 1994). A similar paucity exists in studies of the effect of this error on the parameterization of linear models.

A number of disciplines and subfields employ egocentric measures as predictors within a linear model framework. Examples from the migration and urbanization literature include the use of ego's degree and the ratio of group ties within an egonet (one group's ties compared with another's). For example, one might use the number of ties within a city and outside a city to predict measures of immigration and segregation (Guarnizo and Haller, 2002; Brown, 2006; Aguilera and Massey, 2003; Fischer, 1982). In the epidemiological literature the use of egonet density and the egonet degree of different relations (e.g., friendship, kinship, etc.) can be used to predict such things as drug use (Schroeder et al., 2001). And, in the field of criminology, researchers have used egonet mean degree – the sum of egonet degree divided by the number of egos sampled (e.g., spatially stratified sampling schemes) and egonet mean degree for different relations (e.g., number delinquent friends) – to predict different crime metrics (Sampson, 1988; Warner and Rountree, 1997; Browning et al., 2004).

To provide a typology for the effect of random error on egocentric networks this work employs the standard statistical/medical terminology of *false positive* and *false negative* to characterize this random error. In the context of egocentric networks, this notion of false positive and false negative is broken down into two distinct

* Tel.: +1 9494366213.
E-mail address: almquist@uci.edu

0378-8733/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.socnet.2012.03.002

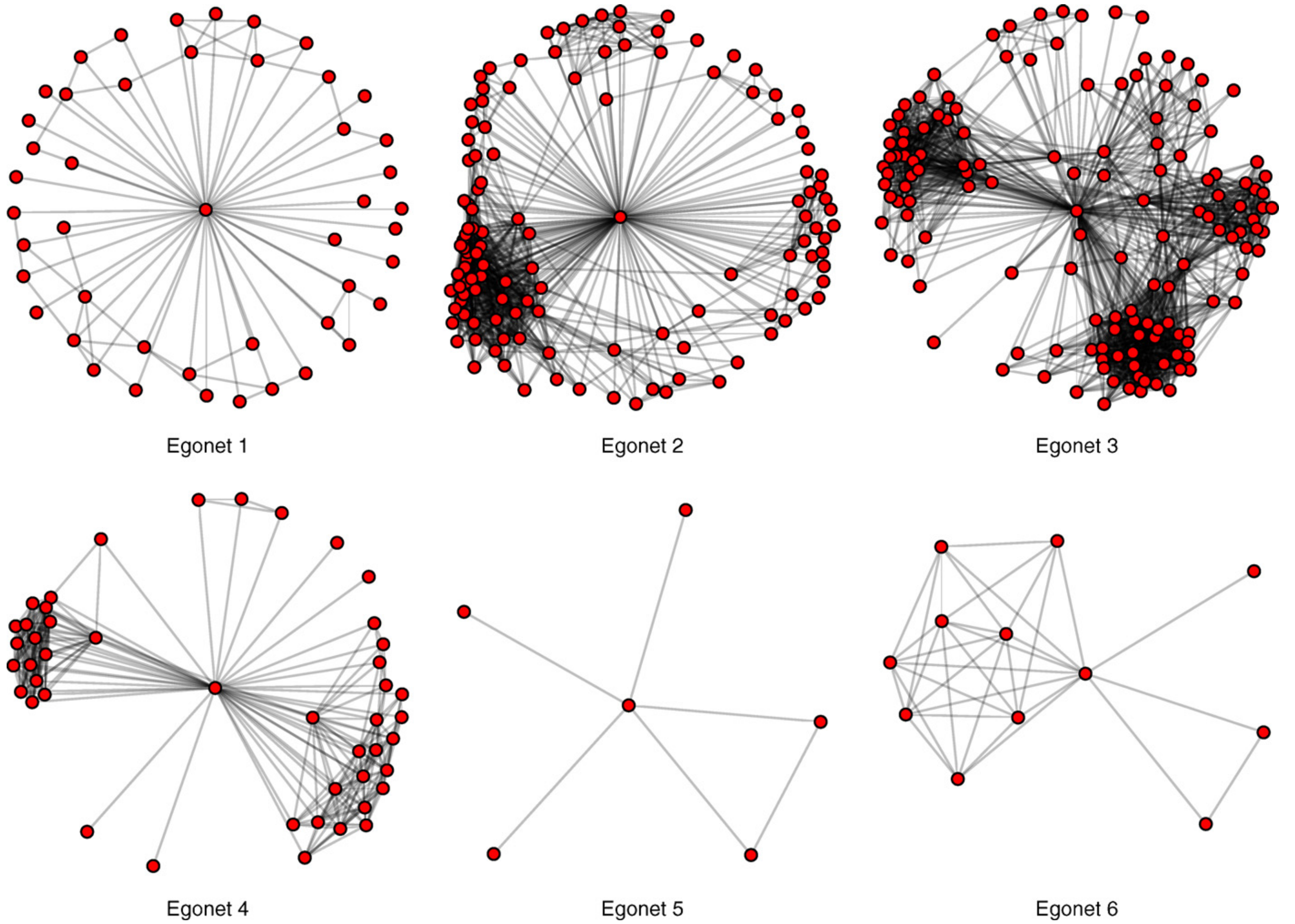


Fig. 1. Six examples of egocentric networks in the facebook sample selected by randomly choosing an egocentric network within 0.15 quantiles based on graph density.

categories: errors in the alters reported (i.e., errors in the vertex set) and errors in the relation or tie between alters (i.e., errors in the edge set).

To inform this discussion of error, we employ the modern technique of simulation analysis to inject different levels of error into a sample of "perfectly" measured egocentric networks. The egocentric sample comes from a Metropolis-Hastings Random Walk (MHRW) sample of *facebook*[1] (Gjoka et al., 2010) that is shown to be uniform asymptotically (Gjoka et al., 2010). For this analysis it is not necessary that the sample be uniform, only that egocentric networks are measured precisely. Because these egocentric networks derive from an online source, which is collected through automated algorithms, it is argued that these egonets represent a so-called "ground truth" of the *public facebook friendship egocentric networks*. By providing an example of a true population sample for which social ties are known exactly, the facebook data serves as a useful case for examining the potential impact of measurement error on empirically encountered social networks. While no one case can be representative of all real-world networks, the use of empirical data provides a source of realistic heterogeneity that can be lacking in simplified, simulated networks.

To further demonstrate potential effects of random errors on egocentric networks we have chosen to analyze a series of standard social network metrics and to carefully dissect a common use case of egocentric data. Specifically, we consider Latkin et al.'s (1995) research on needle sharing where the author uses logistic regression on the binary (yes/no) outcome of needle sharing on several egocentric network metrics.

This paper is laid out in the following manner: (1) background and necessary mathematical notation, (2) error structure, (3) methodology, (4) results, (5) limitations and other considerations, and (6) a discussion of the results and implications.

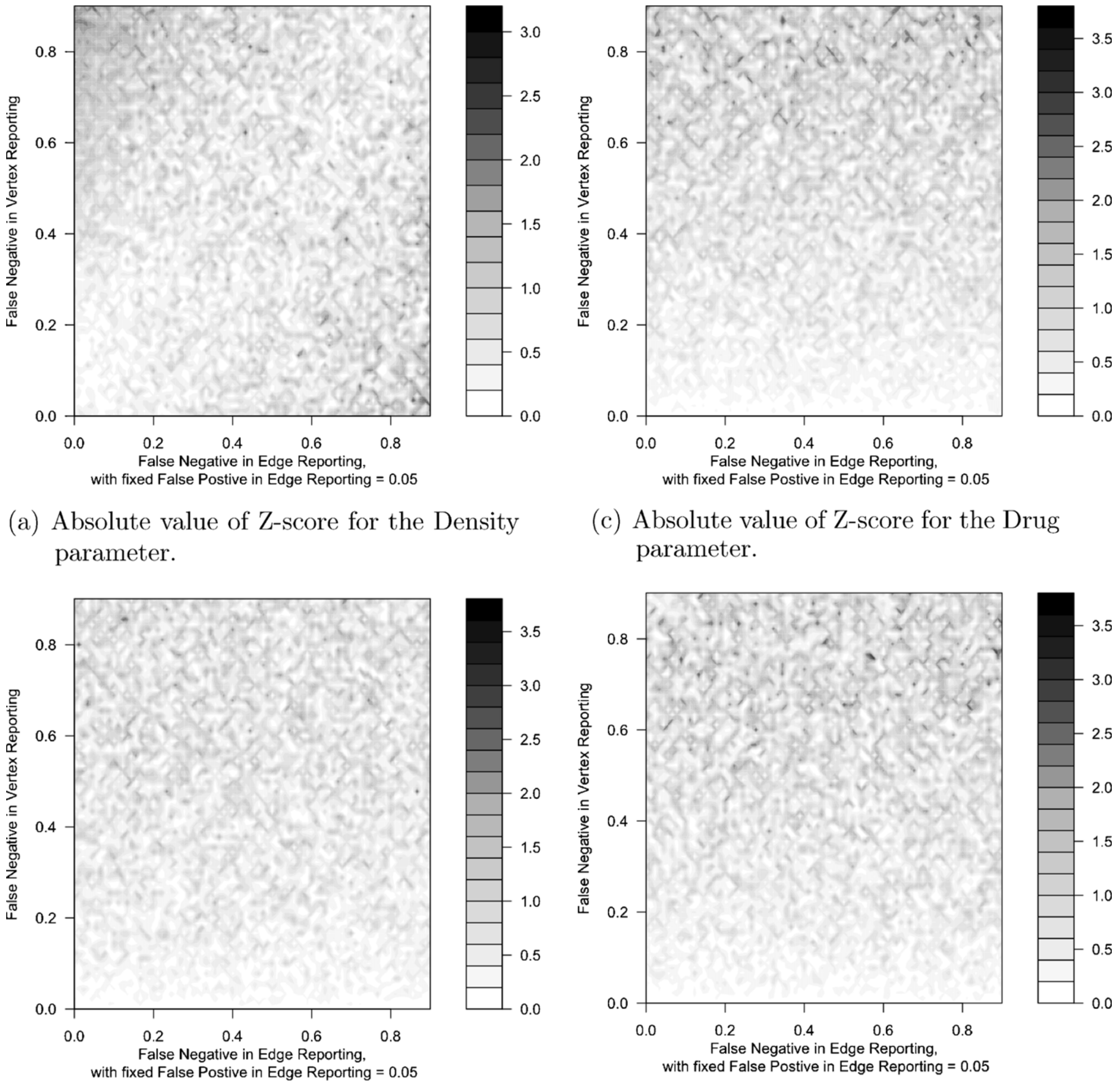
2. Background and notation

This section will first give a brief overview of the literature on error in the area of egocentric networks, and will include some of the larger literature of error on social networks. Following the review of the literature, this work will cover the basic details of egocentric networks and graph theoretic notation employed in this paper.

2.1. Literature review

The collection of social network data, *especially egocentric network data*, is fraught with potential for both random and systematic error. There is a long history of worrying about error in the social network literature (see, Bernard et al., 1984; Butts, 2003; Freeman et al., 1987). In the context of egocentric networks, research has tended to focus on the different effects of alter elicitation such as free recall, card sort, and roster (Brewer et al., 2005). There has been extended study of the effects on recall of friendship egonets, where it was found that individuals "forgot" as much as 20% of their

[1] www.facebook.com.



(a) Absolute value of Z-score for the Density parameter.

(c) Absolute value of Z-score for the Drug parameter.

(b) Absolute value of Z-score for the Multiplexity parameter.

(d) Absolute value of Z-score for the Sex parameter.

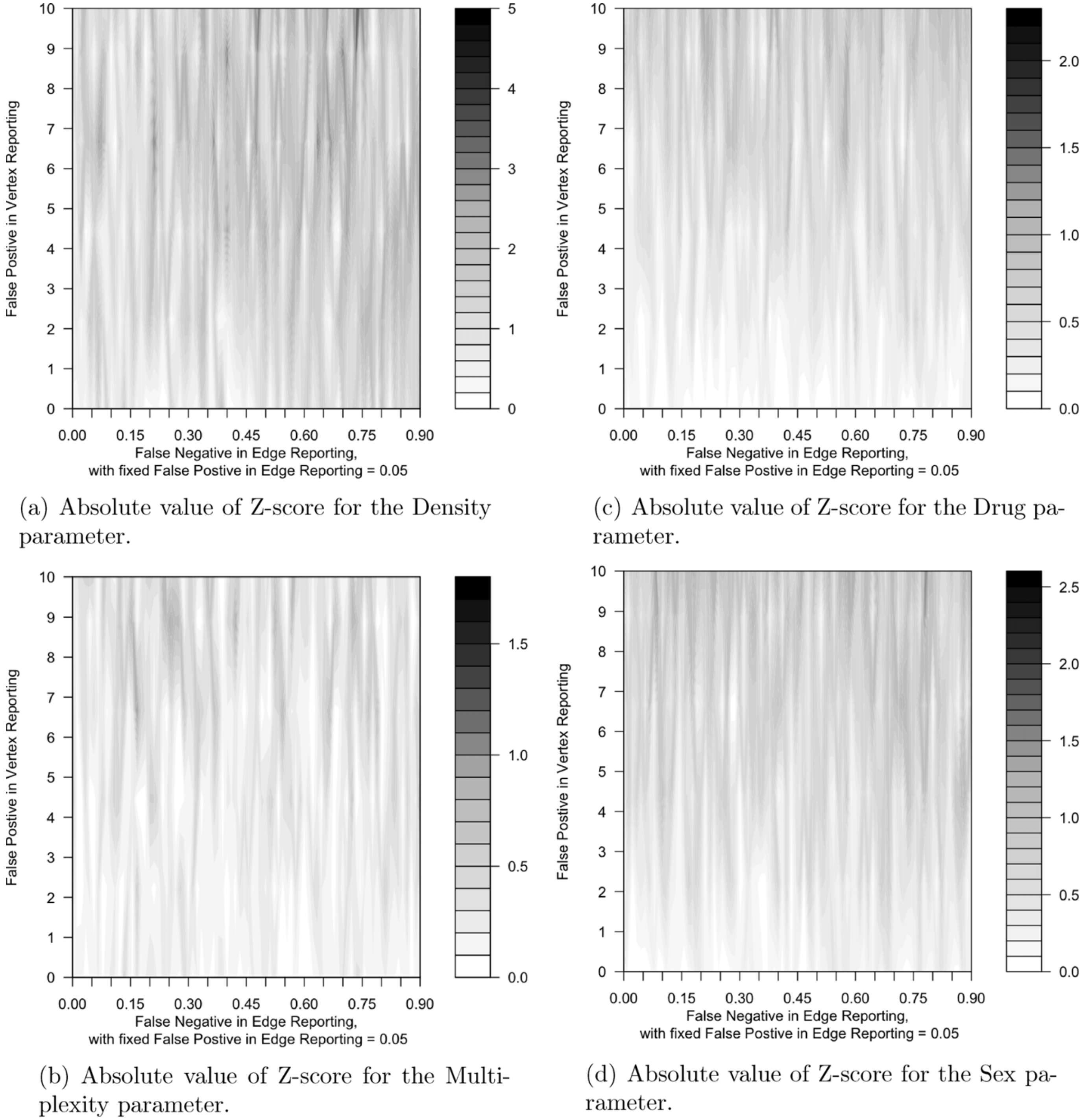
Fig. 2. False negative vertex error versus false negative edge error, with edge false positive set to 0.05, for parameters density, multiplexity, drug and sex.

friends (and even 3% of their best friends) (Brewer and Webster, 1999). There exists extensive research into the bias and error created by the survey instrument itself—such as interviewer, recency, and order effects (Marsden, 2003; Marin, 2004).

Generally speaking, it has been shown that the issue of memory and the form of elicitation of networks can have considerable influence on the size and composition of an egocentric network acquired by a researcher (Brewer, 2000). Along with friendship, careful attention has been paid to such effects in the context of intravenous drug users and individuals involved in high-risk sexual activity and their impact on the ability to acquire accurate egocentric networks in such cases (Brewer and Garrett, 2001; Marsden et al., 2006). There exist attempts to characterize systematic bias caused by these various elicitation schemes (Feld and Carter, 2002).

Recently, work has sought to characterize general principles of the effects of random error on network structure (e.g., Butts, 2003; Borgatti et al., 2006). Butts (2003) found that (i) individuals are more likely to make *false negative* than *false positive* errors in their reports of alter–alter ties and (ii) these types of errors can have significant impact on network structure.

Borgatti et al. (2006) found that centrality measures on random networks were rather robust to random error; however, it is not so obvious that this will be true in the context of non-random networks such as egocentric networks.



(a) Absolute value of Z-score for the Density parameter.

(c) Absolute value of Z-score for the Drug parameter.

(b) Absolute value of Z-score for the Multiplexity parameter.

(d) Absolute value of Z-score for the Sex parameter.

Fig. 3. False positive vertex error versus false negative edge error, with edge false positive set to 0.05, for density, multiplexity, drug and sex parameters.

2.2. Egocentric networks

Egocentric networks are generally composed of a focal actor (*ego*) and a set of component actors (*alters*) who are connected to ego through some predefined relation (e.g., friendship) (see Wasserman and Faust, 1994). This form of data is most often elicited in survey context (e.g., GSS, AddHealth, Personal Networks in Town and City; Burt, 1984; Moody, 2002; Fischer, 1982; Marsden, 1990). Egonets may also be sampled from online populations due to either data limitations (e.g., the amount of memory required to hold the whole network is simply impractical) or to enforced limitation to access of the data (e.g., facebook, friendster, twitter; Gjoka et al., 2010).

When speaking about an egocentric network the two most important levels are (1) a *star* egocentric network and (2) the *first-order* egocentric network (Wasserman and Faust, 1994; Butts, 2008). A star egonet is composed of only ego and his or her alters, and is therefore always a *star*, topologically speaking (Butts, 2008, p. 18). A *first-order* egonet is composed of ego and his or her alters and the connections between the alters (e.g., ego is friends with Bill and Jill, and Jill is also friends with Bill). In this paper, an *egocentric network* (unless specified) will always refer to a *first-order* egonet.

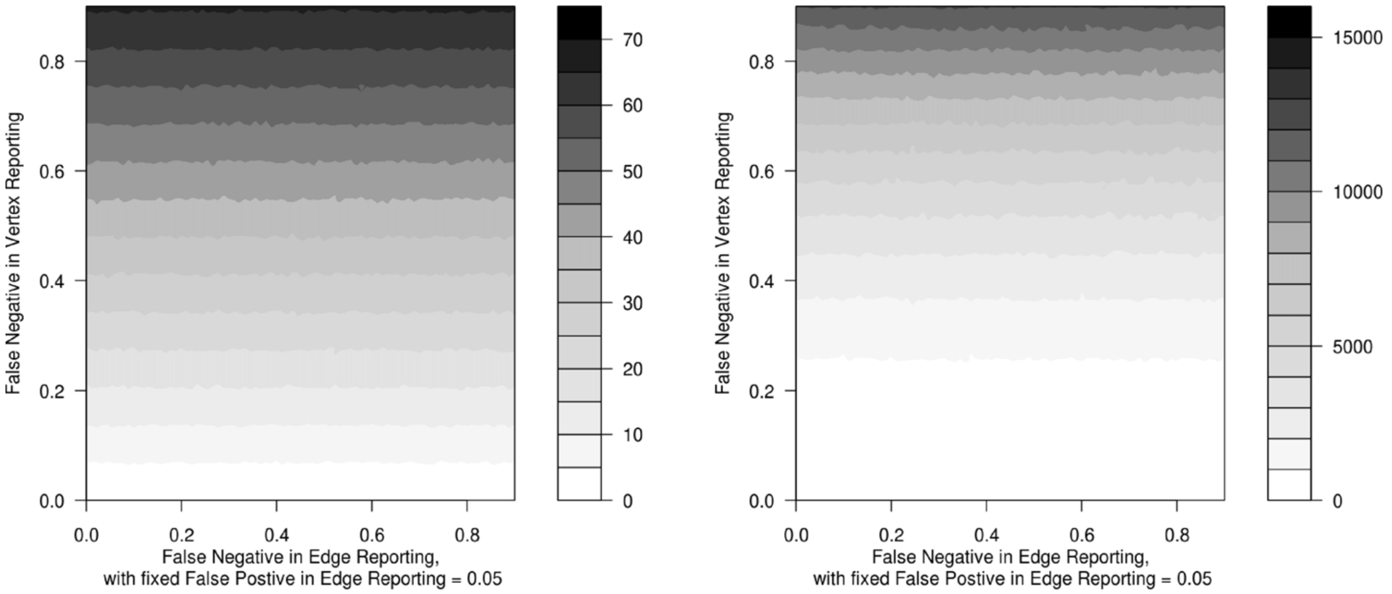


Fig. 4. Left is average absolute error and right is MSE (with edge false positive fixed to 0.05). False negative vertex error versus false negative edge error in network size.

This definition may be expanded until the entire graph is covered (e.g., *second-order*, *third-order*, etc.).

2.3. Graph notation

The prevailing practice in the field of social networks is to represent relational structures in a combination of mathematical and statistical notation. In this paper it is necessary to define a series of basic concepts, the first of which is a *graph* that is comprised of an *edge set* (E) and *vertex set* (V), i.e., $G = (V, E)$, where V represents a set of actors and E represents a set of relations (e.g., friendship, kinship, neighbors, coworkers, etc.). It is often useful to be able to write a graph in its matrix algebra representation, also known as an *adjacency matrix*. An adjacency matrix is a n by n matrix, where n is the size of the vertex set, composed of 1s and 0s (diagonal usually nulled out, i.e., no self-ties). A graph may be either *directed* or *undirected*. An undirected graph is by definition *symmetric* (i.e., if A is friends with B then B is friends with A), while a directed graph may be *non-symmetric* (i.e., if A is friends with B, then B does not have to be friends with A). Undirected relations are representative of relations such as coworkers, friends, or kin, whereas directed networks might be composed of relations like communication, dominance acts, or gift-giving. Finally, $|\cdot|$ is the cardinality operator and when applied to an adjacency matrix provides the *size* of the network or the number or vertices in the network.

In the context of egocentric networks it is necessary to define the concept of a *subgraph*. A graph H is said to be a subgraph of G if $H \subset G$. It is worth pointing out that under this definition an egocentric network is always a subgraph of the complete network. The important take-away from this is that each subgraph can itself be represented as a graph and hence every egocentric network, while being a piece of a larger graph, is itself a graph.

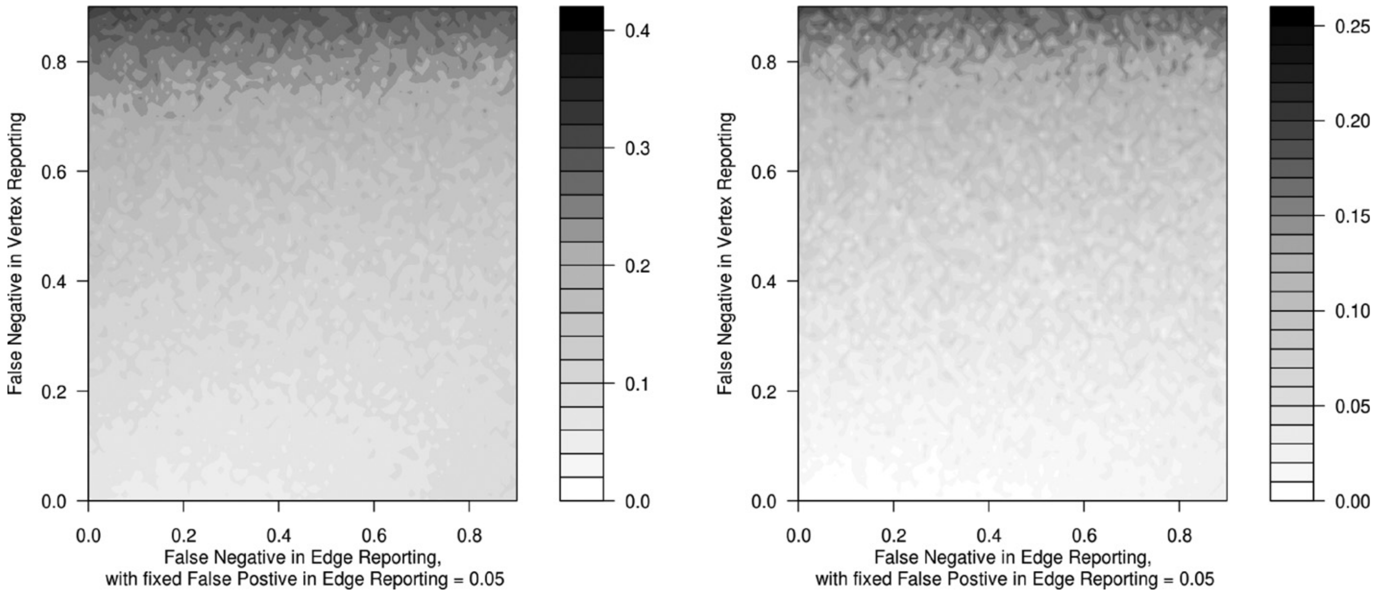


Fig. 5. Left is average absolute error and right is MSE (with edge false positive fixed to 0.05). False negative vertex error versus false negative edge error in network density.

Table 1
Descriptive statistics of the 330 subsample of egocentric networks in the facebook 37+thousand sample of egocentric networks.

	Mean	Median	SD
Size	93.99	38.50	148.45
Density	0.37	0.24	0.30
Deg cent	0.78	0.86	0.23
Triad 0	2913239.73	11636.00	25415499.80
Triad 1	476567.00	4395.00	1872516.45
Triad 2	79016.47	1509.00	283261.99
Triad 3	32341.83	442.00	151420.37

3. Error structure

A number of different ways exist to conceptualize the notion of error. One way to approach this problem is to propose that there is the observed data (in this case $G^o = (V^o, E^o)$, which is a measurement of the *true* data ($G^t = (V^t, E^t)$). Thus we can define a notion of *false positive* and *false negative* in the context of egocentric networks (borrowing from the statistics and medical literature language on errors). There are two *core* types of false positives and false negatives in the context of egonets. The first type is of the *edge*, e.g., when an edge is falsely included or falsely excluded and the second type is of the *vertex*, e.g., when a vertex is falsely included or falsely excluded. Both of these errors occur naturally in most egonet elicitation schemes. For example, Brewer and Webster (1999) show that an individual can forget as much as 20% of their friends.

3.1. Edge error

Edge error in egocentric networks occurs when a tie between alter *i* and alter *j* is either mislabeled as present (false positive) or mislabeled as not present (false negative). One distinctive feature of an egocentric network – which would not be present in a more general discussion of network error, such as that seen in Butts (2003) – is that ego is definitionally connected to every alter. This is represented notationally by conditioning on *ego*.

Before defining false positive and false negative in the context of egocentric networks, a little bit of notation will need to be introduced. *Ego* will be indexed at $i = 1$, without loss of generality, so that $e^k_{1\cdot}$ is the set of edges from ego to all alters $j = 2, \ldots, |G^k|$, where $k = o$ or t.

The probability of a false positive or false negative in the context of egocentric networks may be very naturally written in probabilistic notation where the probability of the observed alter–alter tie (i to j; $i, j \neq 1$) is misclassified as 1 or 0 depending on what the "true" edge value is:

$$\text{False positive}: \ Pr(e^o_{ij} = 1 | e^t_{ij} = 0, e^t_{1\cdot}) = \alpha_e \qquad (1)$$

$$\text{False negative}: \ Pr(e^o_{ij} = 0 | e^t_{ij} = 1, e^t_{1\cdot}) = \beta_e \qquad (2)$$

where the false positive rate for any given alter–alter edge is α_e and the false negative rate for any given alter–alter edge is β_e.

Table 2
Multiple logistic regression for *sharing needles* in previous 6 months at follow-up interview for 330 injection drug users in the SAFE study, Baltimore, MD, 1991–1992.

Variables	*B*	SE *B*	Signif.	Odds ratio
Network size				
Drug	0.11	0.06	0.04	1.12
Sex	−0.04	0.09	0.68	0.96
Intimate interaction	0.08	0.09	0.38	1.08
Physical assistance	−0.07	0.08	0.43	0.94
Material assistance	−0.15	0.08	0.06	0.86
Positive feedback	−0.05	0.07	0.48	0.95
Health information	0.02	0.08	0.79	1.02
Social participation	0.06	0.06	0.31	1.07
Network characteristics				
Network density	1.05	0.43	0.02	2.85
Multiplexity	0.11	0.09	0.21	1.11
Individual characteristics				
Gender	0.12	0.31	0.69	1.13
Education	−0.02	0.25	0.94	0.98
Age	0.00	0.02	0.92	1.00

Table 3 from Latkin et al. (1995).

Table 3
Mean and SD of alter subgroups in Latkin et al. (1995).

	Mean	SD
No needle sharing (121)		
Drug	4.39	2.22
Sex	1.45	1.44
Intimate interaction	2.71	1.98
Physical assistance	3.40	1.95
Material assistance	3.15	1.81
Positive feedback	4.25	2.43
Health information	2.30	1.97
Social participation	4.02	2.54
Needle sharing (209)		
Drug	5.40	3.28
Sex	1.54	1.73
Intimate interaction	3.08	1.83
Physical assistance	3.48	1.97
Material assistance	3.00	1.87
Positive feedback	4.50	2.89
Health information	2.58	1.99
Social participation	4.82	3.05

3.2. Vertex error

Similarly to the case of edge error, the error in the vertex set may be broken down into false positives (i.e., incorrectly, including an alter), and false negatives (i.e., incorrectly, excluding an alter). This statement contains a number of non-trivial assumptions which are derived from more general assumptions on egocentric networks and graphs. Removing a vertex has the effect of removing all ties that are connected to that vertex, which means that removing a vertex also affects the edge structure of an egocentric network. The inclusion or exclusion of alter v_i may again be written into probabilistic notation similar to that of the edge error case, noting that in this case the inclusion or exclusion of a vertex includes whether all pairwise alter$_i$ to all other alters is possible (this is suppressed in the following notation). Again, without loss of generality, *ego* is

Table 4
Simulated logistic regression table based on Table 3 in Latkin et al. (1995).

Variable	*B*	SE *B*	Signif.	Odds ratio
Intercept	−0.03	0.72	0.96	0.97
Network size				
Drug	0.10	0.03	0.00	1.11
Sex	−0.15	0.06	0.01	0.86
Intimate interaction	0.06	0.04	0.17	1.06
Physical assistance	−0.06	0.04	0.08	0.94
Material assistance	−0.19	0.05	0.00	0.82
Positive feedback	−0.09	0.04	0.02	0.91
Health information	0.08	0.04	0.05	1.09
Social participation	0.10	0.04	0.01	1.10
Network characteristics				
Network density	1.17	0.52	0.02	3.22
Multiplexity	0.02	0.08	0.80	1.02
Individual characteristics				
Gender	0.19	0.36	0.60	1.21
Education	−0.01	0.04	0.90	0.99

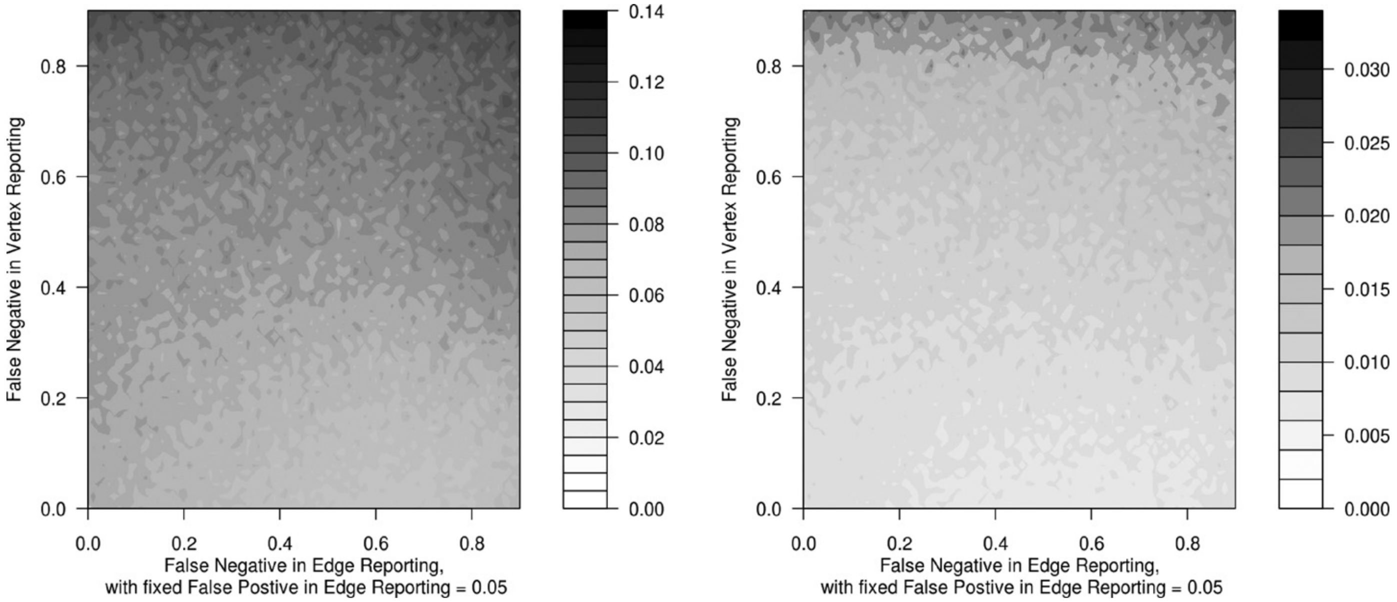


Fig. 6. Left is average absolute error and right is MSE (with edge false positive fixed to 0.05). False negative vertex error versus false negative edge error in dyad and isolates.

index at $i = 1$.

$$\text{False positive}: \; Pr(v_i \in V^o | v_i \notin V^t, v_1 \in V^t) = \alpha_v \quad (3)$$

$$\text{False negative}: \; Pr(v_i \notin V^o | v_i \in V^t, v_1 \in V^t) = \beta_v \quad (4)$$

where the false positive rate for any the false inclusion of an alter i is α_v and the false negative rate for the false exclusion of an alter i is β_v (where $i = 2, \ldots$).

4. Methodology

In this paper we employ simulation to induce different rates of error on both the vertex set and edge set of the *facebook egonet sample*, hereafter referred to as the "facebook sample." First, there will be a discussion of the data followed by a description of the algorithms used to induce random error and, finally, a discussion of the simulation techniques employed to test the effects of random error on egocentric networks and the effects of this error on the parameters of general linear models.

4.1. Data: facebook friends egonets

4.1.1. Facebook

Facebook.com, in 2004, where it was initially introduced at Harvard University. At that time, facebook.com was primarily used to share information about personal interests and activities. Shortly after its introduction to Harvard University, facebook spread to other Ivy League Schools and subsequently to hundreds of institutions around the United States. Currently, facebook membership is open to the general public. Facebook allows users to create a profile where they may choose to share a variety of personal characteristics (e.g., sex, relationship status, school, work, etc.) and general interest information such as favorite artists or political views. Another core aspect of facebook is that it allows for the *joining* together of

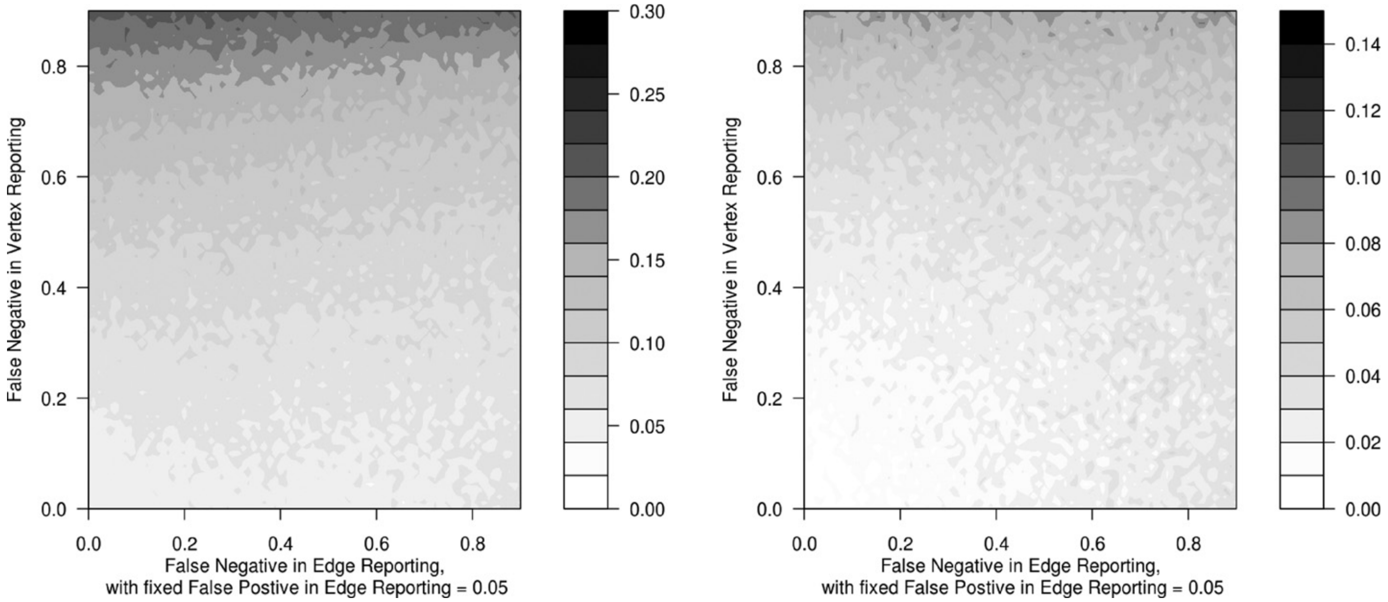


Fig. 7. Left is average absolute error and right is MSE (with edge false positive fixed to 0.05). False negative vertex error versus false negative edge error in two stars.

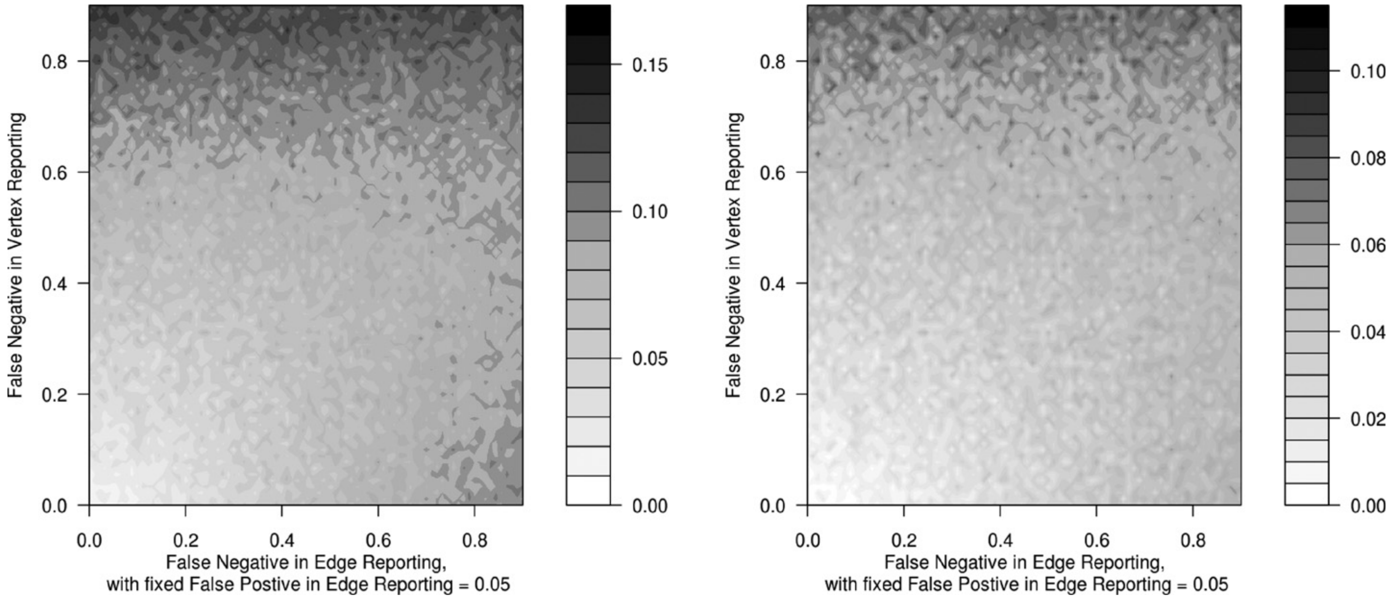


Fig. 8. Left is average absolute error and right is MSE (with edge false positive fixed to 0.05). False negative vertex error versus false negative edge error in triangles.

people from different regions (e.g., New York City or Oregon), schools (e.g., High School, University), or places of work (e.g., Microsoft) into groups. And, most importantly, there is a social network component to facebook made *public* by the declaration of "friends." *Facebook friendship* is a mutual tie where both parties must agree that they are *facebook friends*, and that friendship is made public providing they choose not to hide their information by changing the default settings on their facebook profile.

4.1.2. Facebook sample

The data used for this study is a sample of full egocentric networks (egonets) as derived from the Metropolis Hastings Random Walk sample of *facebook friends* (Gjoka et al., 2010). This sample includes around 37+thousand full egonets (and second-order neighbors). The relation of facebook friendship is publicly displayed, i.e., the facebook user has not modified their default privacy setting in such so as to hide their "friends." In this paper we employ a subsample of 330 egocentric networks uniformly sampled from the full sample of egonets.[2] In this sample, the average degree of ego is 94 with an average egonet density of 0.37 (for more details see Table 1). In Fig. 1, one may visualize the diversity of these egocentric networks.

4.2. Simulation analysis

While there are a number of ways to explore the nature of random error on egocentric networks, one very natural approach is to employ simulation. We treat the facebook sample as "ground truth" and induce various levels of false positive and false negative error into the edge and vertex set of the egocentric networks. To induce error on the edges is quite natural, we induce a Bernoulli change in the "true" network at a specified rate (α_e, β_e).

To induce error in the vertex set is straight forward in the false negative case (vertex deletion), but slightly more complicated in the false positive case (vertex addition). Deletion may be done by removing a vertex (and all edges connected to the removed vertex) at a given rate (β_v).

As a result of the "open boundary problem," (where should the missing vertex come *from*, and, more importantly whom should the new vertex connect to) vertex addition is more complicated. Instead of tackling this open boundary problem head on, we choose to add one vertex at a time up to *n* false positive vertices where we estimate the likelihood of these new vertices being connected to any given node (including the other added vertices) as Bernoulli process at a given density δ.[3]

4.3. Computation

All computation is implemented in the R Statistical Programming Environment (R Development Core Team, 2010).

5. Case study: the effects of random errors in parameter estimation of dependent variables

As social network analysis and relational analysis has grown in popularity within the social science community the practice of employing different GLIs and NLIs from egocentric network data as independent variables in linear models has grown in popularity, e.g., degree regressed on the outcome of segregation or egonet density on the binary outcome of needle sharing (yes/no).

5.1. Personal network characteristics and their influence on needle-sharing

Latkin et al.'s (1995) research is a particularly salient example of the use of egocentric network measures. These researchers performed logistic regression on the binary (yes/no) outcome of needle sharing on several egocentric based GLIs. Latkin et al.'s (1995) paper was published in the Journal *Social Networks*, "Personal Network Characteristics as Antecedents to Needle-sharing and Shooting Gallery Attendance." Latkin et al. (1995) researched

[2] The case study and simulation analysis of GLIs are performed on a 330 subsample for computational reasons. The 37+thousand egonets range from 1 to 4563 making the multiple computations of various network statistics need for this paper prohibitively expensive. However, because we employ a uniform random sample of a uniform random sample this subsample of 330 will be sufficiently representative.

[3] The δ selected in this analysis represents the (homogenous) probability of any two alters being tied together as estimated from the data, and is (we argue) the most reasonable proxy for selecting if two alters should be connected given an absence of observed data for false positive alter–alter ties.

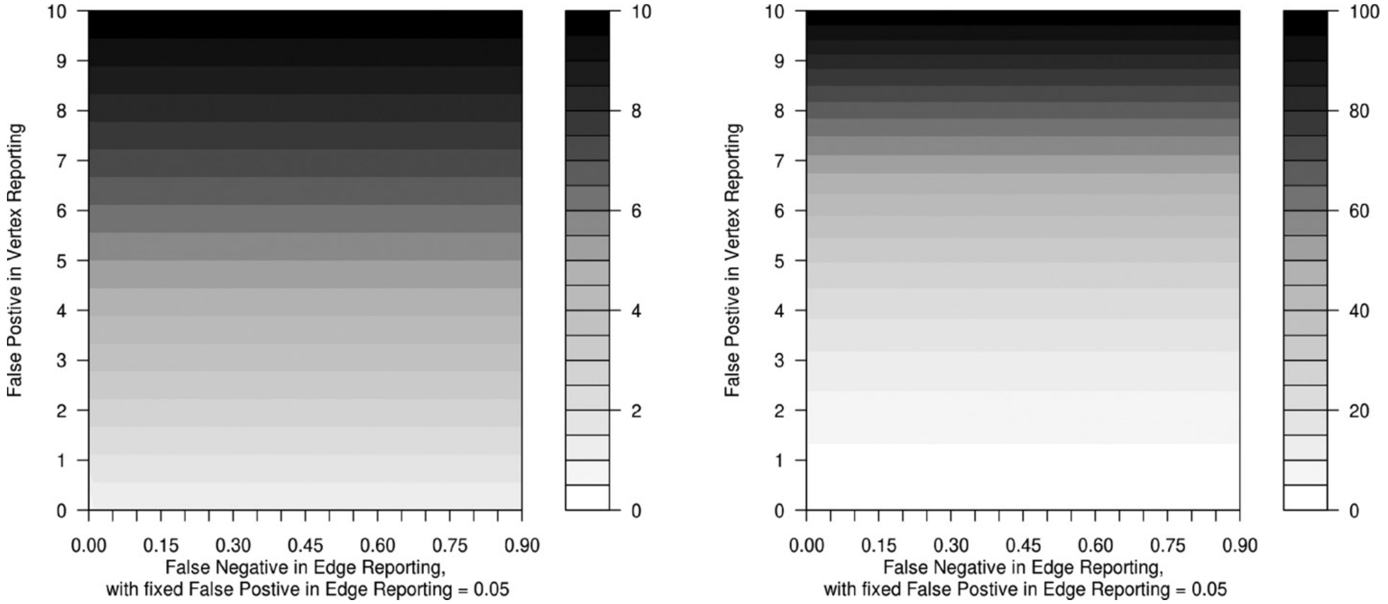


Fig. 9. Left is average absolute error and right is MSE (with edge false positive fixed to 0.05). False positive vertex error versus false negative edge error in network size.

the impact of local network characteristics on the likelihood of needle-sharing within a sample of 330 individuals participating in an HIV-prevention study. Latkin et al. (1995) discovered that needle-sharing is influenced by both ecological and resource factors, specifically that there is potential for network-based strategies to reduce needle-sharing among injecting drug users.

This research employed three different network measures: *density* (egocentric network density), *multiplexity* (number of individuals in ego's network who share 2 or more attributes), and *subgraph degree*. Latkin et al. (1995) measured eight subnetworks in each of their respondent's egocentric network: two "negative" networks (Drug and Sex), and six "positive" networks (Intimate interaction, Physical assistance, Material assistance, Positive feedback, Health information, and Social participation). Latkin et al. (1995) employ logistic regression to predict *sharing needles* in previous six months at follow-up interval for 330 injection drug users in the SAFE study, Baltimore, MD, 1991–1992 (see Table 2).

5.2. Simulating a network of needle-sharing

To emulate the aforementioned case study we employ the method of simulation to generate subgraph groups, and the binary outcome variable (0 or 1). In the original study there are 209 individuals who shared needles out of the sample of 330. We select the 209 out of 330 egocentric networks to be "needle sharers" by a weighted sampling routine to provide similar weighting results to those in Table 2.[4]

To simulate the subgroups we employ draws from a Poisson distribution to provide the number of ego's alters to be in a given subgroup and then uniformly sample individuals from the set of alters to be labeled as part of said subgroup.[5] Our utilization of a Poisson distribution allows us to stay within the spirit of this research, which is concentrating on the effects of random processes on egocentric networks. The Poisson distribution used differs between the "needle users" and non-needle users as it does in the original study. We use a standardized mean times the egonet size for each group because (1) the two groups differed and (2) the egocentric networks in the facebook sample are on average larger than those in the study by Latkin et al. (1995). The covariates Gender and Education are also simulated from Poisson distributions.[6]

STEP 1 Generate subgroups from truncated Poisson distribution.We start by performing a single draw from a Poisson distribution with mean $(n_i \cdot d_{gi}/\alpha)$, where n_i is the size of the egonet i, d_{gi} is the degree of group g and egonet i, and α is a tuning parameter[7]. We redraw from the Poisson distribution if the draw is larger than n_i. The means used for this analysis may be found in Table 3.

STEP 2 Random assignment for groups.We then assign group labels based on the Poisson draws from STEP 1 through a simple random assignment procedure.

STEP 3 Perform logistic regression; Table 4.

5.3. Inducing error on the network's of needle-sharing

We begin by inducing error on the egocentric networks in the way described in Section 4.2 and then recomputing the logistic regression model on the flawed network. A slight addition to the false positive vertex algorithm is implemented to handle the assignment of groups in the "addition" vertices being added to the network. This is handled through a simple binary random variable with the probability of being included in a group chosen by the percentage of ingroup members divided by the total number of alters in a given egocentric network.

5.4. The effects of random errors in egocentric networks as predictors in linear models

In this paper we report the findings of false positive and false negative errors on egocentric networks as predictors of an outcome

[4] We employ the following weighting scheme to select the needle sharing individuals $((\alpha \cdot \ln(d) \cdot \delta)/c)$, where α is tuning constant, d is the degree of the egonet, δ is the density of the egonet and c is a normalizing constant. To acquire Table 4 we use an $\alpha = 0.3$.

[5] In many social contexts we expect these different groups to be tightly clustered (e.g., my friends tend to be friends), so this should be a more conservative test as alters are simply placed into their grouping by random assignment.

[6] Age is left out because it had a zero effect in the original logistic regression.

[7] $\alpha = 10$ for this analysis, because that is the average degree of ego.

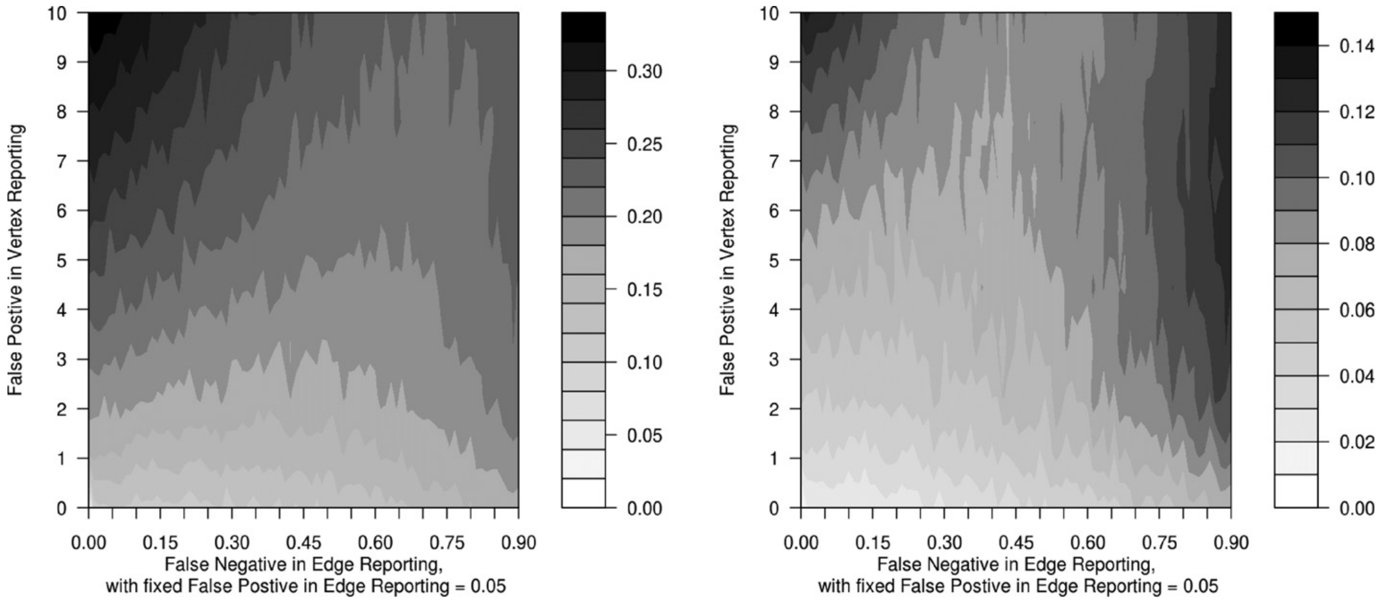


Fig. 10. Left is average absolute error and right is MSE (with edge false positive fixed to 0.05). False positive vertex error versus false negative edge error in network density.

of interest (i.e., needle sharing) through a simulation analysis as discussed in the earlier section. We focus on three egonet measures: density, multiplexity, and two subgroup measures: *sex* and *drug*. When interpreting the effects of the random error on the logistic regression parameters we have chosen to exam the "*z*-score" or normalized effect of the parameter minus the true value divided by the observed standard deviation for the parameter of interest (Figs. 2(a)–3(d)). In the false negative reporting in the vertex set and edge set we begin to see large effects in the density parameter at around 20% error and in the subgroups as early as 5 or 10% error. Notice that since we are looking at *z*-scores this has implications for directionality of the effect and the effect's significance level (Fig. 2 (a)–(d)).

For false positive reporting (Fig. 3(a)–(d)) in the vertex set and false negative reporting in the edge set we begin to see serious changes to the parameter value (and the resulting *p*-value) at as few as one or two individuals. Notice also that the landscape is quite non-linear so that the effect goes in and out of "safe space."

6. Errors in standard network measures

Again we report the false positive and false negative effects on outcomes of interest, in this case standard network summary measures. To reduce the number of combinations, we fix the edge false positive rate to 0.05 (since Butts, 2003, showed that false positive rates are typically very low) and the average mean square error and average absolute error for false negative vertex error versus false negative edge error (Figs. 4–8 and we also look at the false positive vertex error versus false negative edge error, adding 0–10 vertices (Figs. 9–13). This study examines random error in five different network measures: size, density, and the undirected triad census – minus the null triads because egocentric networks are connected by definition – (Dyad and Isolates, Two Stars, and

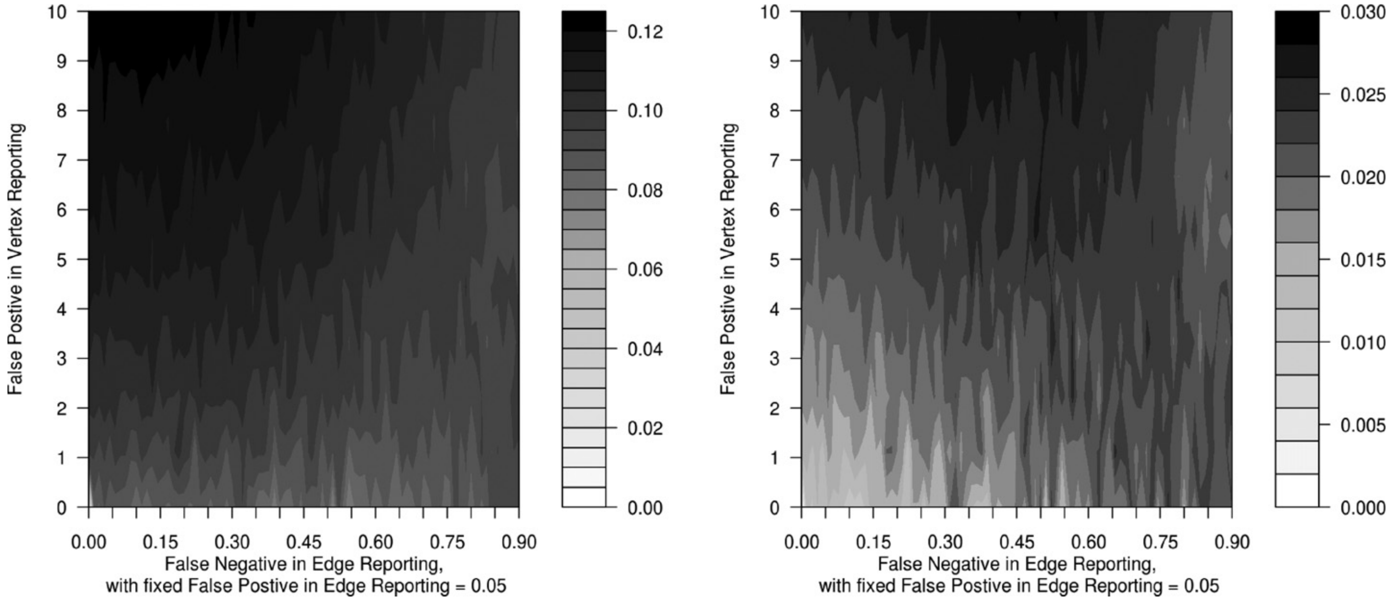


Fig. 11. Left is average absolute error and right is MSE (with edge false positive fixed to 0.05). False positive vertex error versus false negative edge error in dyad and isolates.

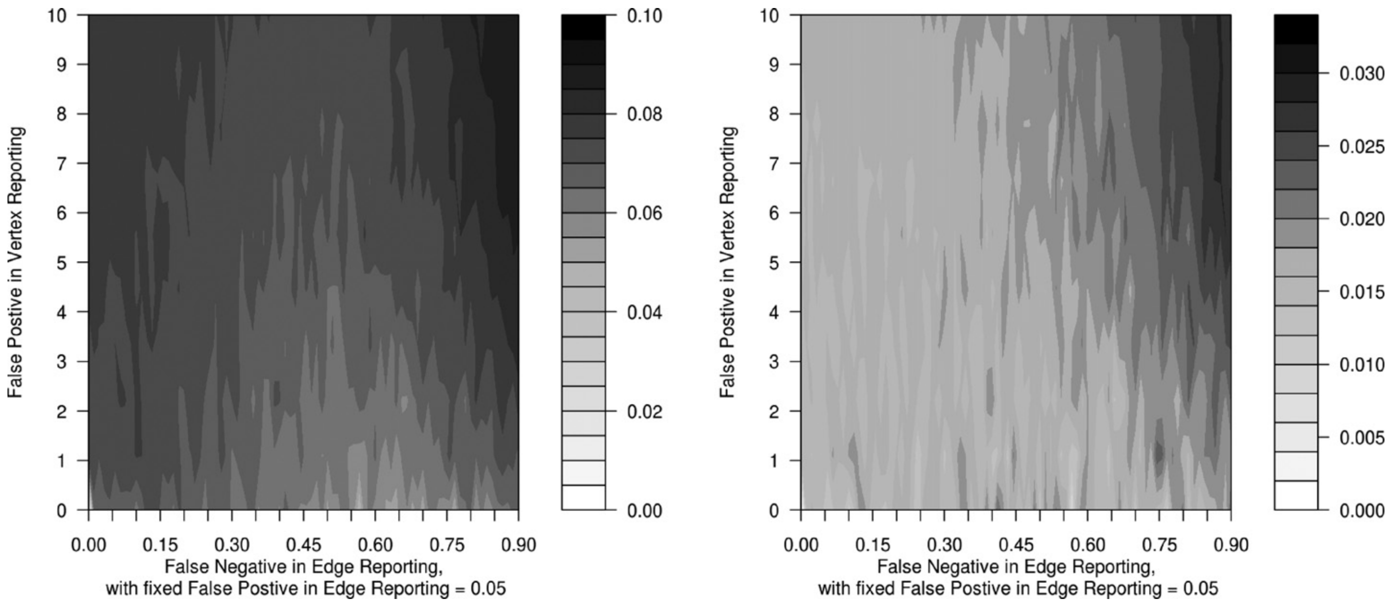


Fig. 12. Left is average absolute error and right is MSE (with edge false positive fixed to 0.05). False positive vertex error versus false negative edge error in two stars.

Triangles; Wasserman and Faust, 1994). All five are classic network measures (see for details Wasserman and Faust, 1994). Each plot is composed of two components, a *mean square error* (MSE) and *average absolute error* (AAE). Notice that in Figs. 4–13 the two different error measures result in largely the same finding, with a tendency for AAE to be more robust to the percent of error induced on the egocentric network.

6.1. Results of simulated errors in standard network measures

The results section will again be broken down into distinct classes that of *false negatives* in the vertex set and *false positives* in the vertex set because of the issues discussed in Section 4.2.

6.1.1. False negatives in vertex reporting

False negatives in vertex reporting effects on size (or egocentric degree) and edges is rather straightforward and affects the egonet's size/egonet degree in linear fashion as expected (Fig. 4). Density effects are large and may become quickly non-trivial (even a shift of 5% in the density can radically change the nature of a graph or egonet). If we view the diagonal in Fig. 5 – which is the region most egocentric network data is likely to be collected (i.e., both error in the vertex set and edge set) – there is largely a linear trend, however, it can be seen that at around 20% error metric like the density of an egonet may be heavily biased. This bias could affect the interpretation of the egonets influence on outcomes such as needle sharing (notice that 20% error is not unheard of, see Brewer and Webster, 1999). Errors in the (standardized) triad census (Figs. 6–8) are non-trivial. Again, 20% seems to be the start of the "serious" amount of error in this network measure.

6.1.2. False positives in vertex reporting

False positives in vertex reporting effects on size (or egocentric degree) and edges is rather straightforward and affects the

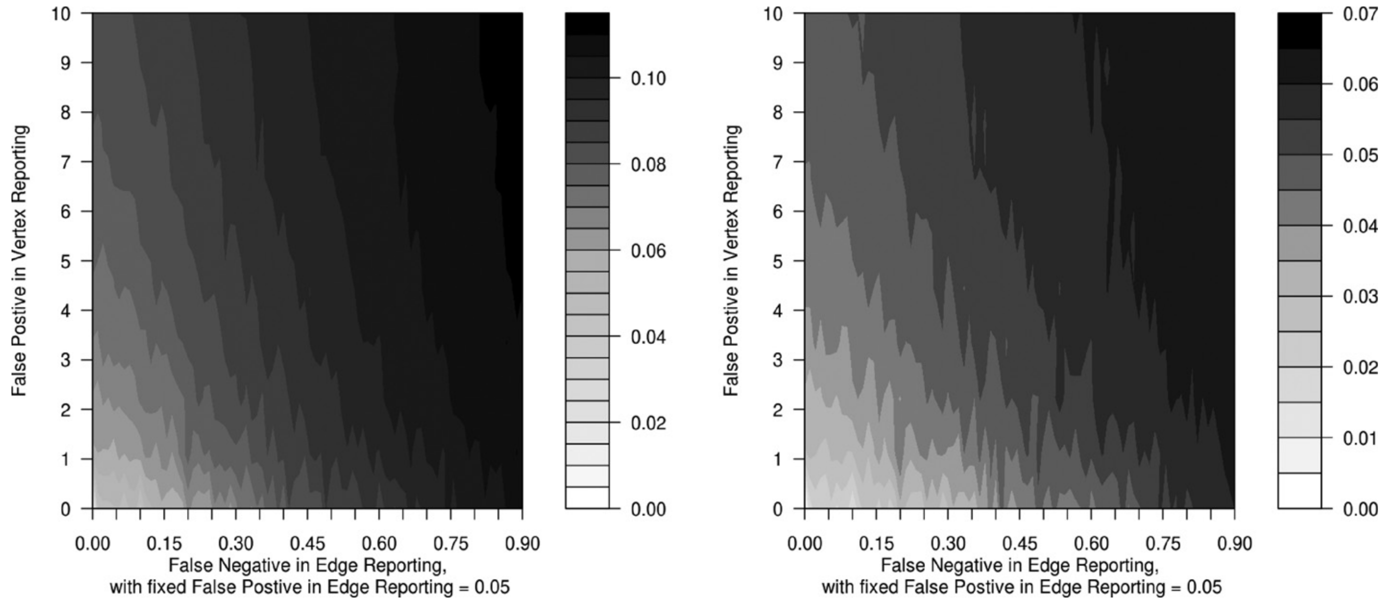


Fig. 13. Left is average absolute error and right is MSE (with edge false positive fixed to 0.05). False positive vertex error versus false negative edge error in triangles.

egonet's size/egonet degree in a linear fashion as expected (Fig. 9). Error in the vertex reporting, where ego is adding one or more vertices to their egonet, quickly become problematic for density as few as three additional alters may move the difference in density up as much 20% (Fig. 10). Errors in the (standardized) triad census (Figs. 11–13) are again non-trivial.

7. Limitations and other considerations

This research is focused on a very particular form of error within the analysis of egocentric networks, namely that of measurement error. While it is important to always take into consideration issues of measurement and the error that ensues, it is also paramount to not give up on a method due to measurement issues and limitations. Egocentric elicitation is a core network metric and a growing tool within the survey literature to allow for local-level network data collection. The characterization of the error which might occur within egocentric measurement is only one of many considerations and should not detract from the actual collection of the data. It should, however, inform and temper the conclusions and interpretation of significant findings based off of egocentric network measures. However, it is worth pointing out at this juncture, not all error within network data collection is random it may in fact be very systematic and clustered (e.g., not only does one forget one friend, but they forget all friends associated with that individual, e.g., all work friends). This clustered error maybe very problematic or much less of an issue depending on one's interest and the research goals. It is critical to mention that this paper and research has very limited implications for systematic error or clustered error which might have very different characteristics or issues.

Other limitations include the data source, i.e., facebook. While the egocentric networks attained from facebook are measured perfectly they may differ in significant ways from many egocentric networks of interest. Facebook is an online friendship network which has limited maintenance issues, and thus the local networks may become much larger (on average) than some face-to-face relationships. Another issue for the facebook friendships is that online users differ from the general population in some important ways, though these differences are rapidly disappearing.[8]

8. Discussion and implications

Research on error in the context of egocentric networks has placed a great deal of emphasis on the systematic mismeasurements induced by the interaction between aspects of human memory and the design and instrumentation of surveys; however, little if any work has been done in the area of random error on egocentric networks. This paper attempts to provide a starting point to understanding what simple errors might do to an egocentric network and what effects this might have on a researcher's analysis of the egonets themselves and the resulting effects of using egonet measures as independent predictors in linear models.

We began by describing the notation and terminology (false positives/false negatives), and then introduced a ground truth data set (a collection of uniformly sampled egocentric networks composed of facebook friendships) that contain a rich and realistic baseline on which to perform our simulation analysis of random errors. Next, we built on the work of Butts (2003) to hone in on the false negative reporting of edge ties (rather than false positive of edge ties).

Our simulation analysis is built around a real-world public health research paper published in the Journal of Social Networks (Latkin et al., 1995). This paper explored the effects of egocentric networks on the binary outcome of needle sharing. This analysis demonstrates that significant change in the network measures and parameter estimates can occur at as little as 5% error and becomes troubling at 20% (which as Brewer and Webster, 1999, demonstrates this may happen in many egonet samples). This error is particularly striking when considering the fluctuations which can occur in the weight values estimated by standard linear models, especially when considering such standard measures as density or subgroup degree—noting that this is likely to be more problematic in the case where the subgroups are clustered as one might expect in many social contexts.

In summary, we have examined the impact of random error on egocentric networks through inducing error on a Facebook sample of egonets and exploring the effects of this error on standard social network measures and their use as predictors in linear models. Here, we discovered that if there is as much as 20% error in egonet reporting, as Brewer and Webster (1999) suggests is not uncommon, then simple random error within egonets could be an important issue to take into consideration when interpreting results—this is especially true if there is even a minor amount of false positive reporting in the vertex set. It becomes obvious, then, that random error is potentially a problem for network researches and deserves continued attention and research in the future growth of survey methodology and network analysis.

Acknowledgments

The author would like to thank both the anonymous reviewers, and Carter T. Butts and Mark S. Handcock for their kind suggestions and comments. Last, the author would like thank his dad for all his help and support over the years. This work was supported in part by ONR award N00014-08-1-1015, NSF Grant OIA-1028394 and NIH/NICHD grant 1R01HD068395-01.

References

Aguilera, M.B., Massey, D.S., 2003. Social capital and the wages of Mexican migrants: new hypotheses tests. Social Forces 82 (2), 671–701.

Anderson, B.S., Butts, C.T., Carley, K., 1999. The interaction of size and density with graph-level indices. Social Networks 21, 239–267.

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.

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 Networks 22, 29–43.

Brewer, D.D., Garrett, S.B., 2001. Evaluation of interviewing techniques to enhance recall of sexual and drug injection partners. Sexually Transmitted Disease 28 (11), 666–677.

Brewer, D.D., Rinaldi, G., Mogoutov, A., Valente, T.W., 2005. A quantitative review of associative patterns in the recall of persons. Journal of Social Structure 6 (1).

Brewer, D.D., Webster, C.M., 1999. Forgetting of friends and its effects on measuring friendship networks. Social Networks 21, 361–373.

Brown, S.K., 2006. Structural assimilation revisited: Mexican-origin nativity and cross-ethnic primary ties. Social Forces 85, 75–92.

Browning, C.R., Feinberg, S.L., Dietz, R.D., 2004. The paradox of social organization: networks, collective efficacy, and violent crime in urban neighborhoods. Social Forces 83 (2), 503–534.

Burt, R.S., 1984. Network items and the general social survey. Social Networks 6, 293–339.

Butts, C.T., 2003. Network inference, error, and informant (in)accuracy: a Bayesian approach. Social Networks 25, 103–140.

Butts, C.T., 2008. Social network analysis: a methodological introduction. Asian Journal of Social Psychology 11, 13–41.

Feld, S.L., Carter, W.C., 2002. Detecting measurement bias in respondent reports of personal networks. Social Networks 24, 365–383.

Fischer, C.S., 1982. To Dwell Among Friends: Personal Networks in Town and City. Chigaco University Press, Chicago, IL.

Freeman, L.C., Romney, A.K., Freeman, S.C., 1987. Cognitive structure and informant accuracy. American Anthropologist 89, 310–325.

Gjoka, M., Kurant, M., Butts, C.T., Markopoulou, A., March 2010. Walking in facebook: a case study of unbiased sampling of OSNs. In: Proceedings of IEEE INFOCOM'10, San Diego, CA.

[8] For current polling on the differences between online and offline individuals see the most recent reports put out by the PEW Research Center (http://www.pewresearch.org).

Guarnizo, L.E., Haller, W.J., 2002. Transnational entrepreneurs: an alternative form of immigrant economic adaptation. American Sociological Review 67 (2), 278–298.

Latkin, C., Mandell, W., Vlahov, D., Knowlton, A., Oziemkowsk, M., Celentano, D., 1995. Personal network characteristics as antecedents to needle-sharing and shooting gallery attendance. Social Networks 17, 219–228.

Marin, A., 2004. Are respondents more likely to list alters with certain characteristics?: implications for name generator data. Social Networks 26, 289–307.

Marsden, P.V., 1990. Network data and measurement. Annual Review of Sociology 16, 435–463.

Marsden, P.V., 2002. Egocentric and sociocentric measures of network centrality. Social Networks 24, 407–422.

Marsden, P.V., 2003. Interview effects in measuring network size using a single name generator. Social Networks 25, 1–16.

Marsden, P.V., Landon, B.E., Wilson, I.B., McInnes, K., Hirschhorn, L.R., Ding, L., Cleary, P.D., 2006. The reliability of survey assessments of characteristics of medical clinics. Health Services Research 41 (1), 265–283.

Moody, J., 2002. The importance of relationship timing for diffusion. Social Forces 81 (1), 25–56.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org.

Sampson, R.J., 1988. Local friendship ties and community attachment in mass society: a multilevel systemic model. American Sociological Review 53 (5), 766–779.

Schroeder, J.R., Latkin, C.A., Hoover, D.R., Curry, A.D., Knowlton, A.R., Celetano, D.D., 2001. Illicit drug use in one's social network and in one's neighborhood predicts individual heroin and cocaine use. Annals of Epidemiology 11 (6), 389–394.

Vehovar, V., Manfreda, K.L., Koren, G., Hlebec, V., 2008. Measuring ego-centered social networks on the web: questionnaire design issues. Social Networks 30, 213–222.

Warner, B.D., Rountree, P.W., 1997. Local social ties in a community and crime model: questioning the systemic nature of informal social control. Social Problems 44 (4), 520–536.

Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, New York, NY.