Estimating Network Structure via Random Sampling: Cognitive Social Structures and the Adaptive Threshold Method

Michael Siciliano[a*], Deniz Yenigun[b], Gunes Ertan[a]

[a] University of Pittsburgh, Pittsburgh, PA, United States
[b] Bilkent University, Ankara, Turkey

Abstract: This paper introduces and tests a novel methodology for measuring networks. Rather than collecting data to observe a network or several networks in full, which is typically costly or impossible, we randomly sample a portion of individuals in the network and estimate the network based on the sampled individuals' perceptions on all possible ties. We find the methodology produces accurate estimates of social structure and network level indices in five different datasets. In order to illustrate the performance of our approach we compare its results with the traditional roster and ego network methods of data collection. Across all five datasets, our methodology outperforms these standard social network data collection methods. We offer ideas on applications of our methodology, and find it especially promising in cross-network settings.

Keywords: network sampling, cognitive social structures, cross-network research, social networks

*Pre-Publication Version of Paper Appearing in the Journal *Social Networks*. Available at: http://www.sciencedirect.com/science/article/pii/S0378873312000408

*Corresponding author at: University of Pittsburgh, Pittsburgh PA, United States. E-mail address: mds55@pitt.edu; Telephone: 412-427-9621; Mailing address: 1916 W. Dickens Ave, Chicago, IL, 60614

Estimating Network Structure via Random Sampling: Cognitive Social Structures and the Adaptive Threshold Method

"For the last thirty years, empirical social research has been dominated by the sample survey. But as usually practiced, …, the survey is a sociological meat grinder, tearing the individual from his social context and guaranteeing that nobody in the study interacts with anyone else in it."

Allen Barton, 1968 (Quoted in Freeman 2004)

1. Introduction

Most cross-group or cross-organizational research studies rely on random sampling for the collection of data on economic and organizational variables. Such an approach precludes the measurement of the network within each organization as complete or near complete participation rates are needed (Wasserman and Faust, 1994). We offer advances in data collection methods to enable researchers to maintain a random sample framework while also collecting network data on the relationships among the individuals in the organizations under study. Our method begins by randomly sampling a portion of individuals in the network and then estimates the complete network based on the sampled individuals' perceptions of all possible ties, which are referred to as cognitive slices. Thus, rather than collecting data from each actor in the organization to observe the network in full, which is typically costly or impossible in a cross-organizational setting involving multiple networks, we provide a methodology to aggregate sampled individuals' perceptions of the full network.

There are two interrelated areas of methodological research on networks associated with our current agenda: network sampling (Butts 2003; Frank 2005; and Heckathorn 1997) and network measurement under conditions of missing data (see for instance Butts 2003; Costenbader and Valent 2003; and Borgatti et al. 2006). This study speaks to these ongoing research areas but addresses them from a distinctly different angle as our goal is to recreate an accurate network representation from a small sample of network members.

The following section will offer a brief justification of our methods. Section 3 provides an overview of cognitive social structures and combination methods to deal with three-way data. Section 4 will discuss our estimation and aggregation methods for sampling and combining cognitive slices to produce accurate representations of the "true" network. Section 5 will introduce the datasets we analyze and provide the results of our analysis involving a comparison of our methodology's performance against the standard roster approach and the ego network approach. Section 6 offers thoughts on implementation and potential limitations.

2. Rationale

Perhaps the most challenging step for researchers wishing to measure a network is data collection. The data collection phase is especially difficult in a cross-network study as one has to measure a number of networks. Typically network researchers employ one of two approaches to data collection. The first is to attempt to sample every individual in the organization and collect data on the direct network ties. The second is to sample a subset of individuals and collect ego network data (i.e. direct ties as well as the ties among the alters). Each method presents a different set of problems.

The first method, utilizing a standard roster survey, requires the questionnaire to be distributed to each actor in each network. This is a time consuming process for the researcher and one reason why network studies with a comparative or cross-network framework tend to incorporate only a few networks. The traditional roster method also requires high participation rates in order to produce valid network data (Wasserman and Faust 1994). Recent meta-analyses on survey response rates indicate organizational research achieves an average response rate of just over 50% (Baruch and Holtom 2008; Anseel et al. 2010). Issues of missing data potentially create statistical power issues by reducing the total sample size as some organizations with low response rates will be unsuitable for analysis given concerns over the validity of the network structure. For instance, Sparrowe et al. (2001) had to eliminate nearly 20% of the groups in their analysis due to low response rates within those groups. In terms of the accuracy of the network for a single group, Stork and Richards

(1992) note that in network of 60 actors a 75% response rate provides complete data for only 55% of the relationships in the network. The remaining relationships have only partial data or no data at all and therefore determining the existence or inexistence of a social tie becomes problematic. The accuracy of both the whole network and individual level measures in the roster method are completely dependent on the response rate. As the response rate increases the accuracy and validity of the measures increase. Therefore, researchers employing this method must expel additional time and resources in sending reminders, offering participation incentives, and following up with participants to improve response rates.

While the roster method is more suited to organizational research given the bounded network, a second approach, using ego network data, could be employed. For example, one could estimate a network variable, such as density, in each ego-net and average across the values of the egos sampled to produce an estimate of the global density measure in a particular network. The reliance on averaging across individuals alleviates some of the problems associated with time and response rates. If one is randomly sampling ego networks within an organization, then demand for a high response rate is reduced and therefore the need for multiple follow-ups or additional incentives for organizational actors to participate is lessened. However, the capability of ego network data to produce accurate whole network or global measures is uncertain, as a random sample of ego-networks may not result in information concerning all areas of the network under study. Consequently, an ego network approach does not allow for the estimation of the overall network structure as it does not attempt to gather information on all actors. That said, in an organizational research setting, one could employ a personal network strategy to capture global properties and therefore we include the ego network approach as another point of comparison with the novel methodology we present in this paper.

The method we propose requires only a small random sample to produce accurate estimates of both network structure and global network measures. The powerful combination of random sampling and network data collection provide a researcher with the opportunity to fruitfully explore not only economic and organizational factors related to an outcome of interest, but also the social structure of the human associations under study. Because our methodology relies, in part, on perceptions of ties in a network, we begin with a discussion of cognitive social structures.

3. Network Data and Cognitive Social Structures

Our methods utilize cognitive social structures (defined below) as an alternative means of data collection to address issues of time, response rate, and accurate measurement that arise when collecting network data for a large number of groups or organizations. Specifically we ask, can we randomly sample a small subset of individual perceptions of the network to produce an accurate representation of the overall network those individuals are embedded in? If accurate network representations can be produced from only a few individuals, then data collection can be more easily conducted by interviewing only a handful of randomly selected actors within an organization, and, under any research settings, issues of response rate can be minimized.

3.1. Cognitive Social Structures

Cognitive social structures (CSS) are three dimensional network structures, which represent each individual actor's perception of the entire network (Krackhardt 1987). Thus, unlike traditional data collection methods asking an actor to indicate only his/her ties with other actors in the network, when collecting CSS data the actor needs to provide information on all of the possible ties in the network. These structures are represented as $\mathbf{R}_{i,j,m}$, where i is the sender of the relation, j is the receiver of the relation, and m is the perceiver of the relation (Krackhardt 1987). Here $\mathbf{R}_{i,j,m} = 1$ means that person m perceives a relation to exist from actor i to actor j. If $\mathbf{R}_{i,j,\mathrm{m}} = 0$ then person m perceives the relation to not exist. Therefore, a full CSS network of size N, would be an $NxNxN$ array containing 0 or 1 entries representing the existence of possible ties.[1]

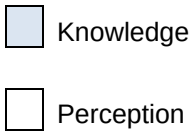
In a directed network consisting of 5 actors, CSS data collection would require each actor to make a judgment about 20 possible ties. So in any CSS matrix, say for actor A, there are two types of data. The first type involves information concerning the ties that involve actor A, which we call *knowledge*. The second type involves actor A's opinion of the ties

[1] It is possible to have valued CSS, but for our current purposes we are only interested in binary networks.

between the other actors in the network, which we call *perception*. For example, Figure 1 below represents a cognitive social structure for actor *A* in a 5 actor network, also referred to as *A*'s cognitive slice. As noted, the knowledge data provided by *A* is in the shaded row and column and the perception data is in the non-shaded areas.

Figure 1: Cognitive Social Structure for Actor *A*, Knowledge versus Perception

A	A	B	C	D	E
A	0	0	1	0	1
B	0	0	1	0	0
C	1	1	0	0	0
D	0	0	0	0	0
E	1	0	0	0	0



To date, the study of CSS has focused primarily on the psychological aspects of perception and cognitive accuracy. Cognitive accuracy has been defined as "the degree of similarity between an individual's perception of the structure of informal relationships in a given social context and the actual structure of those relationships" (Casciaro et al. 1999, p. 286). Research on CSS has explored the connections between cognitive accuracy and power in an organization (Krackhardt 1990) as well as the structural and/or psychological reason for variations in cognitive accuracy (Pattison 1994; Casciaro 1998).

Of importance for the context of this paper, is the fact that research on CSS has shown that individuals have large variations in their ability to accurately perceive the network. Individuals make errors of omission (claiming a tie does not exist when it does) and errors of commission (claiming a tie exists when it does not). Thus, any single individual's reconstruction of the network will be flawed and will generally provide a poor representation of the true network. Our hypothesis guiding the research objective is that by combining a small number of flawed cognitive social structures the errors of omission and commission uniquely associated with a single individual can be washed away through aggregation with other individuals' flawed perceptions.

3.2. CSS Reduction/Aggregation Methods

In order to work with cognitive social structures, because they are three-dimensional datasets, one needs to engage in some form of data reduction or data aggregation. Krackhardt (1987) provided three methods of aggregating cognitive social structures in order to transform three-dimensional data into two-dimensional data. The three methods discussed were slices, locally aggregated structures, and consensus structures. Slices are defined as one individual's perception of the network. Thus, it indicates all ties between *i* and *j*, holding the perceiver constant.

$$R'_{i,j} = R_{i,j,m} ,$$

where *m* is a constant, indicating the perceiver. Locally aggregate structures (LAS) are traditional means of network data collection relying only on information provided by the receiver or sender of a particular tie. The rationale being that the individuals best suited to determine if a tie exists are the members of the dyad under question. Such structures are termed locally aggregated because "the resulting relation between *i* and *j* depends on information provided by the most local members in the network, namely *i* and *j* themselves" (Krackhardt 1987, p.116). Because the CSS data contains information on both *i's* and *j's* knowledge of their individual ties to others in the network, it is possible to combine their knowledge through an intersection rule

$$R'_{i,j} = \{R_{i,j,i} \cap R_{i,j,j}\} ,$$

or a union rule

$$R'_{i,j} = \{R_{i,j,i} \cup R_{i,j,j}\} .$$

Based on the preceding, for a tie from actor *i* to actor *j* to exist in the aggregated network under the LAS intersection rule, both actor *i* and actor *j* must agree on its existence. Alternatively, under the LAS union rule, a tie can exist in the aggregate network if either *i* or *j* claim the tie to exist. Thus, the LAS intersection rule is a more conservative rule requiring mutual agreement by both actors in a dyad. Note, however, that such methods use only knowledge data and ignore all of the perception data provided by the respondent.

Consensus structures rely on information provided by every individual's perception of the tie between *i* and *j* in the network. Thus, as noted by Krackhardt, a practical implementation of a consensus structure is to set a threshold value, where a tie is defined to exist once a certain percentage of network members claim a tie exists between *i* and *j*. The threshold function can be defined as:

$$R'_{i,j} = \begin{cases} 1, & if \sum_m R_{i,j,m} > Threshold, \\ 0, & otherwise. \end{cases}$$

Other methods of aggregation for three-way data have been used, see for instance Batchelder et al. (1997).

4. The Adaptive Threshold Method

As indicated in the brief discussion on cognitive social structures, much of the current research in cognitive network theory focuses on perception as the phenomena to be explained and compares individual perceptions to reality. This paper seeks to combine a small subset of individual perceptions to construct an approximation of reality by estimating the "true" network structure from a sample of cognitive slices. The methods to produce a single network from a random sample of cognitive slices will be discussed in detail below. There are two general steps: (i) sampling individuals to obtain cognitive slices and (ii) aggregating cognitive slices. Once the sampling and aggregation of slices have been discussed, details on the performance of our methods using actual datasets will be provided in Section 5.

4.1. Sampling and Aggregating Cognitive Slices

Given our desire to maintain a random sampling framework, every actor in the network has an equal probability of being selected and upon being selected would be asked to provide his or her cognitive social structure. Once data has been gathered from a random sample of size *n*, aggregation of the cognitive slices relies on a two part procedure utilizing both types of data available in an actor's cognitive social structure: knowledge and perception. As a simple example of aggregation, assume there are 5 actors in a network: *A*,*B*,*C*,*D*,*E*. Each actor's cognitive social structure is given in Figure 2. Note that there is no requirement that the CSS be symmetric.

Figure 2: Sample Cognitive Slices for a 5 Actor Network

A	A	B	C	D	E
A	0	0	1	0	1
B	0	0	1	0	0
C	1	1	0	0	0
D	0	0	0	0	0
E	1	0	0	0	0

B	A	B	C	D	E
A	0	1	0	0	0
B	1	0	1	0	1
C	0	1	0	0	1
D	0	0	0	0	0
E	0	0	0	0	0

C	A	B	C	D	E
A	0	0	0	0	0
B	1	0	1	0	0

D	A	B	C	D	E
A	0	0	1	0	1
B	0	0	1	1	0

C	1	1	0	0	0
D	0	0	0	0	0

C	1	1	0	0	0
D	0	1	0	0	1

Following Krackhardt (1990), we derive the "true' network through LAS intersection.[2] For affective relations, that cannot be observed or verified objectively, the most informed actors concerning the existence of a tie are the members of the dyad under question. Thus, truth is based on an ideal situation, where we have information about each directional tie from both actors in the dyad. When determining whether the $R_{i,j}$ tie exists, we can assess the information provided by *i* and *j*. For example, *i* may claim to hold a friendship tie with *j*, but j may not confirm *i*'s claim. In this case $R_{i,j,i}$=1, $R_{i,j,j}$=0. There is no way to objectively determine who is correct. As Krackhardt (1990; 1996) claims, if both *i* and *j* agree on the *i-j* tie then it is more likely to be true than if they do not agree. This approach of defining a friendship tie only when both parties agree that it exists has obvious face validity (Krackhardt 1990 p. 349). Furthermore, the use of LAS intersection offers another advantage. When comparing our method with the traditional roster method in Section 5, LAS intersection produces a criterion graph that can be reproduced by both methodologies. Because the LAS approach "mimics the typical form in which network data are collected" (Krackhardt 1990, p.349), if the informants in the traditional method are accurate, then the row-dominated roster method and the LAS intersection method will produce similar results. Thus, the LAS intersection approach to the true network provides an achievable criterion state for both methodologies and therefore offers a valid point of comparison.

The goal of our methodology is to produce an accurate representation of this "true" network when only sampling a few of the individuals in the network. In other words, if we were only able to survey 3 actors of this 5 person network, how accurately could we combine the knowledge and perception data contained in each of their cognitive slices to approximate the "true" network as defined by LAS intersection?

When a sample is drawn, the primary issue is how to determine the proper means of aggregating the sampled cognitive slices. Because we are sampling actors we are forced to deal with both knowledge and perception as many of the dyads in the network contain actors who were not sampled and thus have no local knowledge to bear on their relations. Under our sampling conditions, three different scenarios for the determination of a potential tie in the aggregated network exist. The first is when knowledge data is present for both actors involved in the tie (i.e. both actors were sampled). The second is when no knowledge exists for a tie, and thus there is only perception about the tie's existence (i.e. neither of the two actors involved in the tie under question were sampled). The third is when knowledge exists for only one of the actors involved in the dyad (i.e. only one of the two actors involved in the tie were sampled).

In order to illustrate these three scenarios, consider the case where *A*, *D*, and *E* are selected as a sample. Each actor, excluding the diagonal, provides eight pieces of knowledge and twelve pieces of perception data. In the first scenario when both actors in the dyad under question are sampled, we combine the knowledge components of each of the sampled actors using the LAS intersection rule. Thus, for instance: (i) the *D-E* tie was both claimed by *D* and *E*, so the tie exists in the aggregate network; (ii) *A-E* tie was claimed by *A*, but denied by *E*, so the tie does not exist in aggregate network, and (iii) the *A-D* tie was denied by both *A* and *D*, so the tie does not exist in aggregate network. According to our method, no perceptions can change the existence or non-existence of these ties.

In the second scenario, neither of the actors was sampled and so the existence of the tie can only be determined by the perception of others. Thus, in our sample of *A*, *D*, *E*, the existence of a tie between *B-C* in the aggregated network relies on the sampled actors' perception that such a tie exists because neither actor *B* nor actor *C* were sampled. The critical question in these instances is how much perception evidence must be brought to bear on a particular tie before we claim that it exists in the aggregate network. This requires us to set an evidence threshold, *k*, where if *k* or more sampled actors perceive a tie to exist, then the tie will be created in the aggregate network. Discussion about how *k* is determined is detailed in Section 4.2.

[2] There are other ways to construct the "true" network given a set of informant reports. Butts (2003) and Romney et al. (1986) argue that truth can be approximated via consensus methods.

In the third scenario, where only one individual of a possible tie is sampled, we must also rely on perception. For example, *A* claims a tie with *C*, but *C* was not sampled. We cannot treat this as the *A*-*E* case, since *C* was not sampled, so *C* did not have a chance to accept or deny this tie. Therefore, we treat this as perception, and the existence of the *A*-*C* tie has one piece of perception evidence indicating that the tie exists. In such a case, if our perception threshold is *k*, then *k*-1 additional perceptions from the other sampled actors must be present to conclude that *A*-*C* tie exists.

Following this approach, the network estimated from sample *A*, *D*, *E*, and the "true" network obtained by the CSS intersection of all five slices are given in Figure 3. The estimated network in Figure 3 was determined by setting a simple evidence threshold, *k*, to a value of 2.

Figure 3: True Network Derived Via LAS Intersection (TRUE) and the Estimated Network (EST.)

TRUE	A	B	C	D	E
A	0	0	0	0	0
B	0	0	1	0	1
C	1	1	0	0	0
D	0	0	0	0	1
E	0	0	0	1	0

EST.	A	B	C	D	E
A	0	0	1	0	0
B	0	0	1	0	0
C	1	1	0	0	0
D	0	0	0	0	1
E	0	0	0	1	0

In this simple example, the estimated network produced both an error of omission (*B*-*E* tie) and an error of commission (*A*-*C* tie). It is evident that determining the threshold *k* becomes the most important factor for our aggregation methods to produce an accurate representation of the "true" network. In this example, *k* was arbitrarily defined to be 2. This means that if two or more sampled individuals claim a tie existed between any two unsampled individuals, then the tie will be established in the aggregate network. This is a naive way to determine *k*. As sample sizes drawn from different networks will vary, a static level of *k* will tend to perform poorly (see Section 5 below). In addition, each sample will contain actors with varying capacity to accurately perceive the network and thus varying propensities to commit errors. A more sophisticated approach to defining and adapting the threshold level for a particular sample should take into account both the size and error rate.

4.2. Setting an Adaptive Threshold

From previous research on cognitive social structures it is known that individuals vary in the amount of errors of omission and errors of commission they make. Because of this, any sample of cognitive slices may be more or less prone to error and hence more or less trustworthy as a whole. It would be helpful to provide a measure of the accuracy of the sample drawn and use the measure of accuracy to set the most appropriate threshold for *k*. This would allow *k* to be adjusted as a means of controlling the amount of error that occurs when aggregating a sample of cognitive slices. Borrowing terminology from statistics, two error types are possible when determining the existence of ties: Type 1 errors or errors of commission, and Type 2 errors or errors of omission. Then for a given threshold *k*, we have

P(Type 1 Error)= P(Perception says there is a tie | There is no tie) = α_k, (1)

P(Type 2 Error)= P(Perception says there is no tie | There is tie) = β_k, (2)

where *P* denotes probability. We will estimate these probabilities from the observed frequencies. Because each sample of cognitive slices contains both knowledge and perception, we are able to estimate the probability of error from the sample itself without the need for a full dataset. In our adaptive threshold method we focus on Type 1 errors due to the lack of knowledge about the origins of Type 2 errors. A Type 2 error can occur because either an actor does not believe that a tie exists between two individuals or because the actor is simply unaware of whether a tie exists and the default decision when unaware of a tie may be to claim its inexistence. This creates problems when attempting to measure and

reduce Type 2 errors during the aggregation of cognitive slices. Type 1 error is also more important for aggregating cognitive slices as the evidence threshold in the consensus methods for cognitive social structures is based on those who perceive a tie to exist, and thus errors of commission are potentially more costly.

For an example of a Type 1 error in the context of a single cognitive slice, return to Figure 2 and view the sampled slices of *A*, *D*, and *E*. Note that *A* claims to not send a tie to *D* and *D* claims to not receive a tie from *A*. Thus, based only the sampled slices we know that the directional tie *A*-*D* does not exist in the "true" network. An example of a Type 1 error occurs in the sampled cognitive slices because actor *E* perceives that a tie exists from *A* to *D* when in fact we know that it does not. Thus, actor *E* committed a Type 1 error and we can acknowledge this error based solely on the sampled cognitive slices. This is a crucial point. We can locate and count the number of Type 1 errors committed by a cognitive slice based only on information from individuals who were sampled. With larger networks, the opportunities to create Type 1 errors increase and thus it is possible to develop a reasonable estimate of the overall accuracy of the sampled actors' perceptions. This allows *k* to be determined by setting a tolerable level of Type 1 error and then calculating the threshold value of *k* that is necessary to meet the pre-defined tolerance level.

Hence, *k* can adjust based on the measure of the accuracy of each sample drawn. Formally, our estimator of a Type 1 error rate, $\hat{\alpha}_k$, is simply:[3]

$$\hat{\alpha}_k = \frac{\text{Number of Type 1 Errors Comitted by the Sample}}{\text{Number of Possible Type 1 Errors in the Sample}}, \quad (3)$$

where *n* is the sample size. Type 1 errors or the number of perception ties cancelled by knowledge can be increased or decreased based on the value of *k*. For instance, in a sample of 8 actors (say person *A* through person *H*) from a larger network we would directly observe 56 interactions among those sampled individuals. These interactions are pieces of knowledge or ties that are known based on local interaction and thus known to exist or not exist in the "true" network. We can then observe how each of the sampled actors perceives those ties to be distributed among the other sampled actors. For instance, assume there is no tie from actor *A* to actor *B*. Under LAS intersection rules, this would mean that actor *A* denied sending a tie to *B* and actor *B* denied receiving a tie from *A*. We can look to see how many of the remaining six sampled actors (*C* through *H*) perceived a tie to exist. The six other sampled actors would collectively make a Type 1 error if *k* or more of them perceived the tie to exist. By adjusting *k* we can adjust the number of Type 1 errors committed by any particular sample of actors. Because we can determine the accuracy of each sample, we can determine how high or low our threshold level of *k* needs to be set to keep that sample's Type 1 error below some pre-defined level.

Given that the actors are randomly sampled, the error rate calculated from the sample of cognitive slices is assumed to be an accurate representation of the overall error rate the sampled actors would make for the entire network. Thus, for a given sample we can set a tolerable Type 1 error rate and let the accuracy or accountability of the sample determine the threshold level *k* necessary for the sample to not exceed the Type 1 error rate. The algorithm to determine the exact *k* for a given sample operates as follows:

Step 1. Set α, the tolerable error rate. Typical values are 0.05, 0.10, 0.15

Step 2. Draw a random sample of size *n*.

Step 3. Find the smallest *k* such that $\hat{\alpha}_k < \alpha$, and denote this by *k**.

Step 4. Compute the estimated network using the aggregation method with threshold *k**.

In what follows we will refer to this methodology as the *adaptive threshold method*. For a formulized handling of the methodology please see Appendix A.

[3] If we assume that all known ties between sampled actors, in other words the local knowledge of the ties in the sample, are all zero, then in a sample of n actors there are n(n-1)(n-2) opportunities to make a Type 1 error. This is because each of the n sampled actors is not involved in (n-1)(n-2) of the ties and therefore is capable of making a Type 1 error for these dyads only.

5. Performance of the Adaptive Threshold Method

In this section we illustrate the performance of the adaptive threshold method through an extended simulation study. The simulation study consists of repeated sampling from real datasets where the complete CSS is known for each individual. Section 5.1 describes the datasets, Section 5.2 gives a short review of the global network measures we assess, Section 5.3 provides some basic properties of our estimators, and Section 5.4 presents the main findings comparing the adaptive threshold method with the traditional roster and ego network methods.

5.1. Data

To test our method of aggregation using the algorithm to adaptively define *k*, we analyze five datasets. The five datasets are: (i) High Tech Mangers – 21 managers of machinery firm, (ii) Silicon Systems – 36 semiskilled production and service workers from an small entrepreneurial firm, (iii) Pacific Distributors – 33 key personnel from the headquarters of an electronics components distributor, (iv) Government Office – 36 government employees at the federal level, and (v) Italian University – 25 researchers across three interrelated research centers at a university. Each dataset contains a CSS for all or nearly all actors in the network for both friendship and advice relationships. In this study we focus on friendship ties.

Because the datasets we use contain cognitive slices for all individuals in the network, we are able to run multiple trials. Each trail pulls various random samples and creates an approximation of the network from the sampled individuals' cognitive data. Every random sample produces a different estimation of the network and we evaluate the variability from sample to sample. More importantly, because the datasets contain cognitive slices for all actors, we are able to generate the "true" network, or criterion graph. By establishing the "true" network, we can dependably assess the performance of our methodology. Obviously, when employing our methodology in practice, as with any sampling methodology, one will never have data on the full population under study and thus never know the true value of the variable of interest. In our case, this means that a cognitive slice will not be available for every actor as the goal of the methodology is to rely on a small random sample of cognitive slices to estimate the overall network. The primary objective of this paper is to demonstrate the behavior of our methods under various sampling conditions when the "true" structure is known. This allows us to understand how such procedures will behave in the field.

5.2. Global Network Measures

We evaluate our performance through the *correlation* of the estimated network with the true network, and through our ability to estimate several global network measures, namely, *density*, *clustering coefficient*, and *average path length*. We give a short review of all these concepts, but refer the reader to Wasserman and Faust (1994) for a general treatment of network concepts.

Following Krackhardt (1990), we utilize a correspondence measure to assess the accuracy of the aggregated network with the true network. The correspondence measure, labeled as S_{14} by Gower and Legendre (1986), calculates the accuracy of the aggregated network. Given that the matrices of relationships in the five datasets contain only ones and zeros, there are four possible states of relationship between the aggregated network and the true network:

- *a* – matching zeros, meaning the *ij* cell in the true network is zero and the corresponding *ij* cell in the aggregated network is zero
- *b* – omission error, meaning the *ij* cell in the true network is 1 but the corresponding *ij* cell in the aggregated network is zero
- *c* – commission error, meaning the *ij* cell in the true network is zero but the corresponding *ij* cell in the aggregated network is one
- *d* – matching ones, meaning the *ij* cell in the true network is one and the corresponding *ij* cell in the aggregated network is one

Given these definitions, *correlation* is calculated as follows:[4]

[4] As noted by Krackhardt (1990), S_{14} provides a value that is equal to what one would obtain if the matrices were vectorized and a simple Pearson correlation coefficient was calculated.

$$S_{14} = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}.$$

Network density is a measure of the connectedness of the overall network. Density of a network is the ratio of existing ties to the number of possible ties. For directed graphs the formula for density is:

$$l \,/\, n(n\text{-}1),$$

where l is the number of existing ties, and $n(n\text{-}1)$ is the number of possible directed ties among the n actors.

The clustering coefficient, also referred to as transitivity, is the probability that two neighbors of a randomly chosen node are themselves connected. In other words, looking at a particular node in the network, the clustering coefficient measures the interconnectedness among the alters of that node. The formula for the clustering coefficient of a single node *i* is defined as:

$$C_i = \frac{A_i}{m_i(m_i-1)},$$

where A_i is the number of ties between node *i*'s m_i adjacent nodes (Kilduff et al. 2008)[5]. The global clustering coefficient is simply the average of the individual clustering coefficients for all nodes.

The shortest path length between two nodes, known as the geodesic distance, is defined as the smallest number of ties needed to connect two nodes in a network. If two nodes can be connected via existing ties in the network they are said to be reachable. The *average path length* is simply the mean path length of all pairs of reachable nodes, which is defined as:

$$\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} L_{min}(i,j),$$

where $L_{min}(i,j)$ is the geodesic distance between node *i* and node *j* (Kilduff et al. 2008).[6]

5.3. Performance of Adaptive Threshold Method and Effect of α

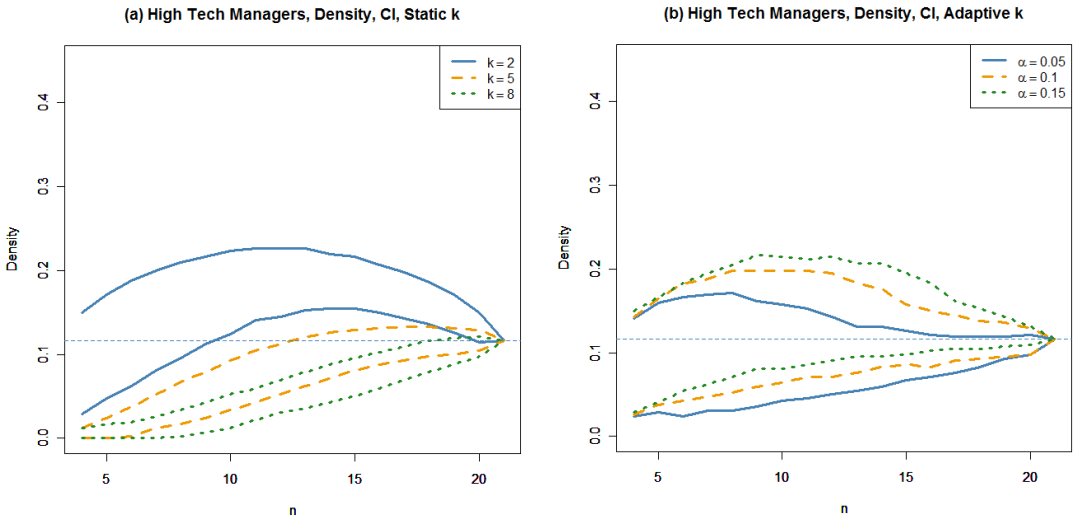
We now illustrate the performance and characteristics of the proposed estimation method through graphs summarizing our simulation results. For each graph in this section and in Section 5.4, the horizontal axis represents the sample size taken from the network under study, and the vertical axis represents the value of an estimated network measure. A horizontal line in each graph displays the true network measure. The colored lines in the graphs correspond to the 95% empirical confidence intervals (CI) for every sample size considered. For a given sample size, say *n*, an empirical 95% confidence interval is obtained by randomly drawing 1,000 samples of size *n*, estimating the network measure of interest using the indicated method for each sample, and reporting the 2.5th and 97.5th percentiles of the estimates.

For an initial assessment, we begin by comparing a static threshold approach with our adaptive threshold method. We use the High Tech Managers data and consider all sample sizes between 4 and 21. In the simulation we draw 1,000 random samples for each sample size, and estimate the network density for every sample. Figure 4-a displays the results of a simulation study for static k=2, static k=5, and static k=8. Figure 4-b displays the results of the adaptive threshold method using 3 different levels of α.

[5] Note the formula for transitivity for a graph is often written differently. It is calculated as $\frac{3\ X\ the\ number\ of\ triangles}{number\ of\ connected\ triples\ of\ nodes}$. The numerator is the number of subgraphs of 3 nodes all of which are connected. The denominator is the number of connected and non-connected subgraphs of 3 nodes.

[6] For two nodes that are unreachable, their values are ignored in the calculation of the shortest path length. Thus, the average path length calculation only considers distances between reachable nodes in the network.

Figure 4: Static versus Adaptive Methods



We can see from Figure 4-a that selection of a static threshold *k* is crucial, as small *k* leads to overestimation and large *k* leads to underestimation of the network measure. The threshold of *k*=2 is too low when sampling, say, 12 individuals. If just 2 of those 12 individuals make the same error of commission then an erroneous tie is placed in the network. The *k*=2 curve begins to decrease after the half waypoint because more and more ties are determined by local knowledge and thus there are fewer chances for errors of commission to occur. We find a very different effect for the *k*=5 and the *k*=8 lines, as these threshold tend to be too high. This sensitivity is not present for the adaptive threshold method as seen in Figure 4-b, as the threshold method specifies α as opposed to k, giving the researcher control of the probability of a Type 1 error. Accordingly, with a low α level of 0.05, the threshold values are the highest and therefore the distribution of density estimates is slightly lower when compared with the other α levels. As the α value moves to 0.10 and to 0.15 the threshold values drop, resulting in density measures with slightly larger predicted values. Regardless of the α level, the true density is consistently covered within the confidence intervals, and as the sample size increases, the variability decreases and the estimated density converges to the true value. These results provide a strong motivation for utilizing the adaptive threshold method.

The choice of α is important in the adaptive threshold method. For each dataset and for each measure, different α levels will produce different results. As with setting a Type 1 error rate in traditional statistics, there are no hard rules to finding the optimal level. While the results are not shown here, our investigation of various α levels across the datasets and across network measures revealed that an α set between 0.08 to 0.12 produced the most reliable results. The final choice of an α rate is ultimately the researcher's and depends on the importance placed on errors of commission and errors of omission.

5.4 Comparing the Adaptive Threshold Method to Standard SNA Methods

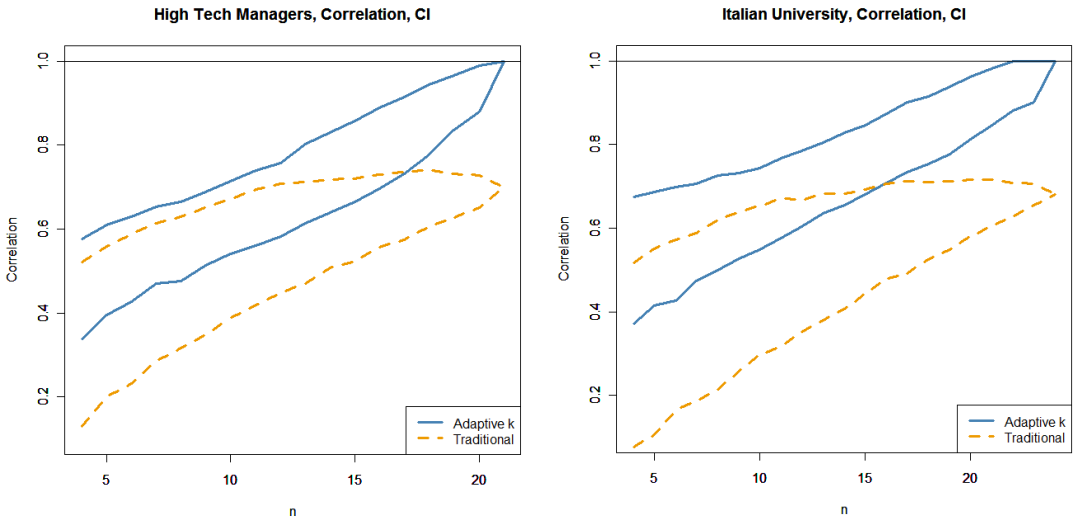
In this section we incorporate all five datasets and analyze how the adaptive threshold method performs at estimating various network level indices. For each of the five networks, simulation studies were carried out in order to produce the 95% empirical confidence intervals for each of the network measures. We report the 95% empirical confidence intervals for the graph correlation with the true network, density, clustering coefficient, and average path length. In order to assess how well the adaptive threshold method performs we compare it's performance with two other standard approaches. These approaches are the traditional roster method, which collects data on each actor in the network about his or her direct ties, and the ego network approach, which collects data on the focal actor's direct ties as well as the ties among his/her alters.

For the traditional roster approach, the network at each sample size was generated by combining the self-reports of the sampled actor's ties. In cases where some individuals were not sampled, missing data was imputed through symmetrization. For example, if actor *i* was sampled but actor *j* was not, then initially the row for actor *j* would be empty. To estimate the structure of *j*'s ties we used a simple strategy commonly employed by network researchers when dealing with (what are believed to be) logically symmetric relations - the missing *j-i* tie is set to the same value as non-missing *i-j* tie.[7] Because we intend to compare our method to the traditional roster method when the roster method achieves moderate to high response rate, imputing missing data only affects a small proportion of ties in these ranges.

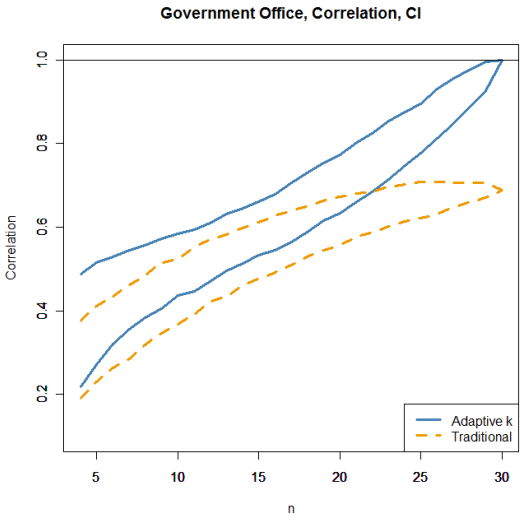
For the ego network approach, global measures were estimated by calculating the value of the measure in each sampled actor's ego-network and then averaging those values across all *n* sampled actors. Note, that because the ego network approach does not produce an approximation of the network structure it is not displayed in the graphics comparing graph correlation. Additionally, because the ego network approach asks individuals to identify their alters and the connections among their alters, by definition, all individuals in the ego network must be a distance of no greater than 2 steps away from each other. Therefore, average path length measures were not calculated for ego network samples.

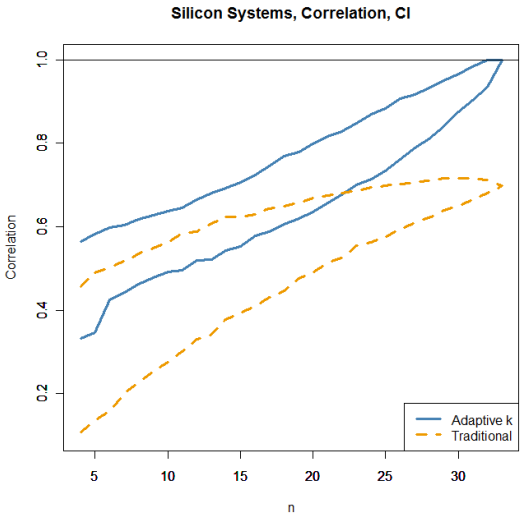
The results of our comparison between the adaptive threshold approach with α=0.1 , the traditional roster approach, and the ego network approach are displayed in Figure 5 and Figure 6. Figure 5 presents the confidence intervals for correlation, and Figure 6 presents the confidence intervals and mean square error (MSE) plots for global network measures. Mean square error is a standard tool in evaluating the performance of an estimator in terms of bias and variability. For a given *n*, mean square error for a network measure is simply computed as $MSE=variance+bias^2$, where *variance* is the observed variance of the estimated network measure, and *bias* is the difference between the true value of the network measure and the average of the estimated network measure, based on 1000 repetitions. As stated before, horizontal line in the confidence interval graphs indicates the true value. For the MSE graphs, the horizontal dashed yellow line is provided as a point of comparison, and indicates the MSE of the roster method when a 70% response rate is achieved.

Figure 5: Graph Correlations



[7] In results not shown here, we aggregated the roster approach data without using symmetrization to impute the missing relations. Such an approach leaves missing actor's rows blank and thus does a very poor job of estimating structure. With no other information to determine non-sampled actors' ties, symmetrizing provides the most informed approach for estimating structure. There are other ways to impute missing data, including the use of exponential random graph models, which are beyond the scope of this paper.


Government Office, Correlation, CI
Correlation
n
Adaptive k
Traditional


Silicon Systems, Correlation, CI
Correlation
n
Adaptive k
Traditional

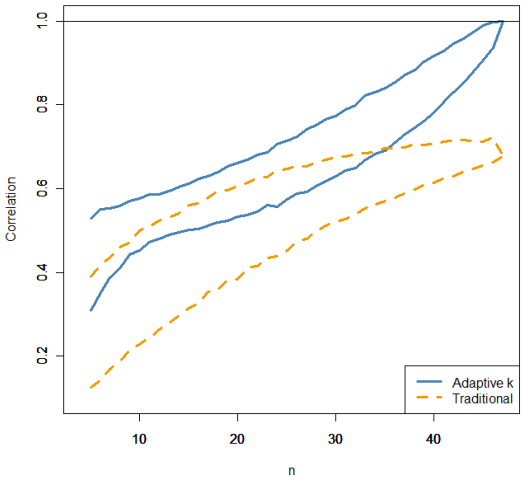
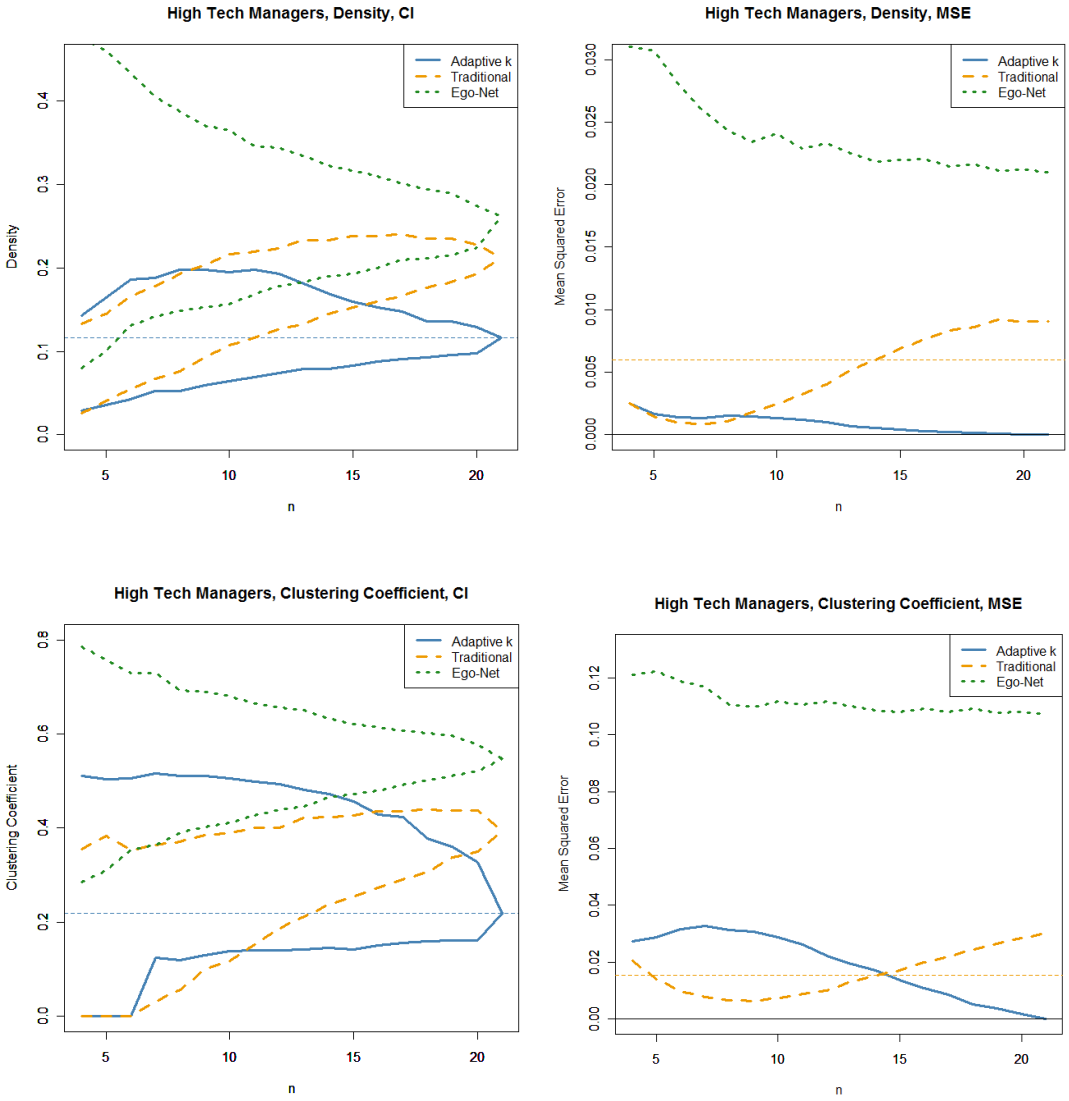
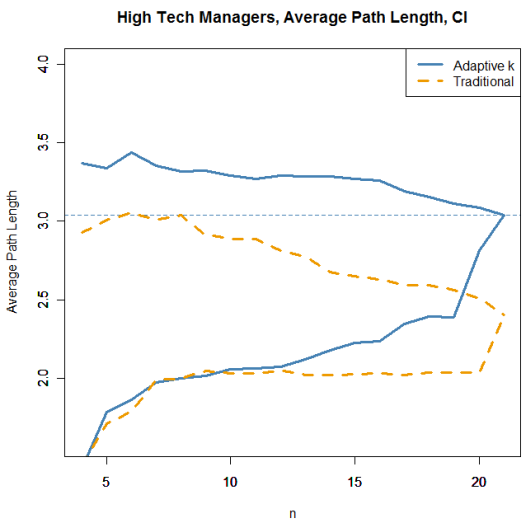

Pacific Distributers, Correlation, CI
Correlation
n
Adaptive k
Traditional

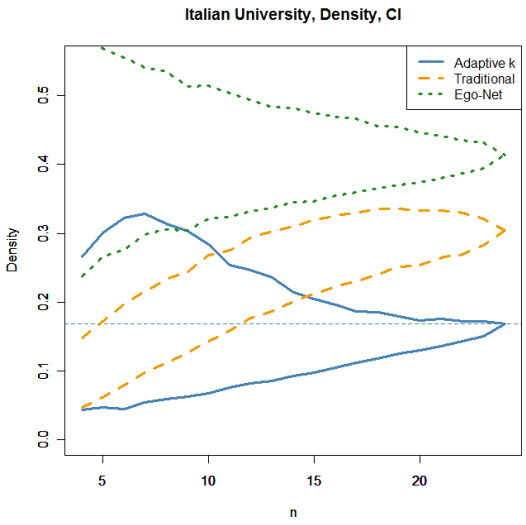
Figure 6: Global Network Measures

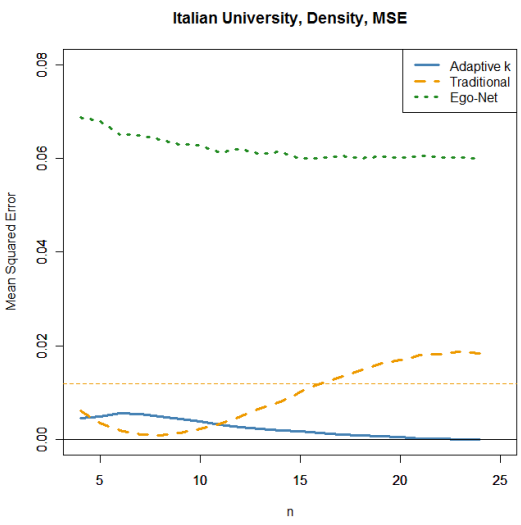
High Tech Managers Data

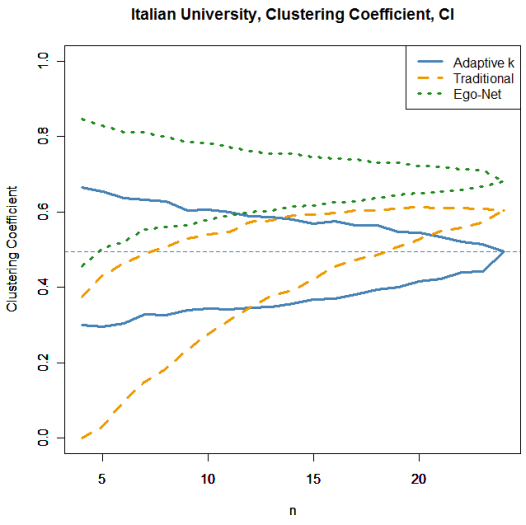


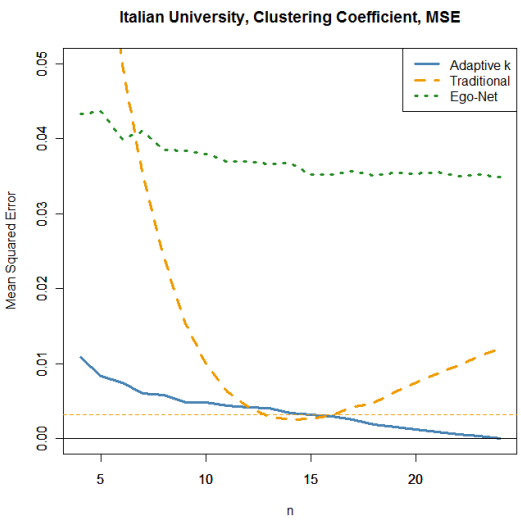


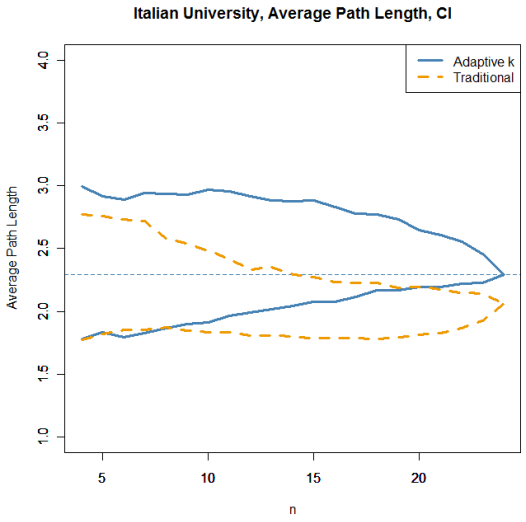
Italian University Data

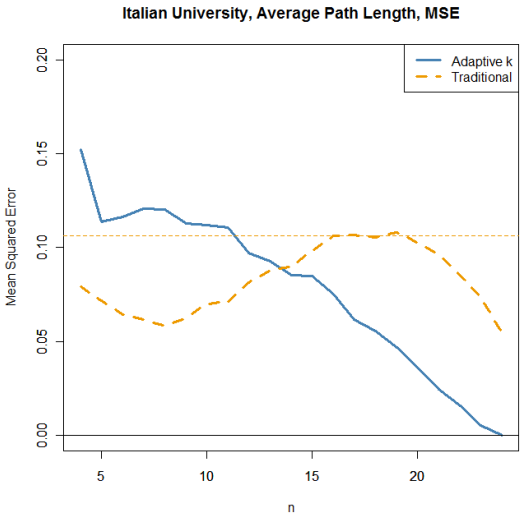




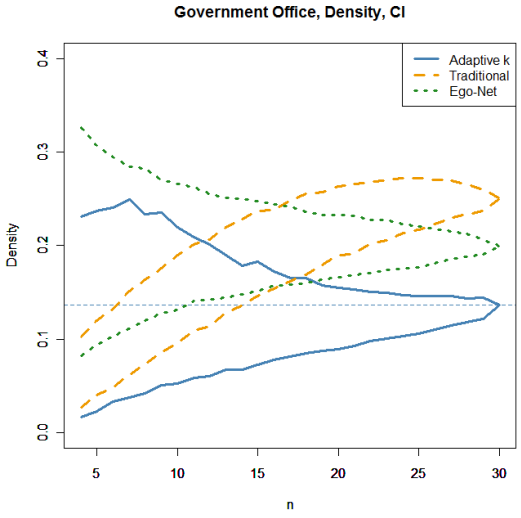


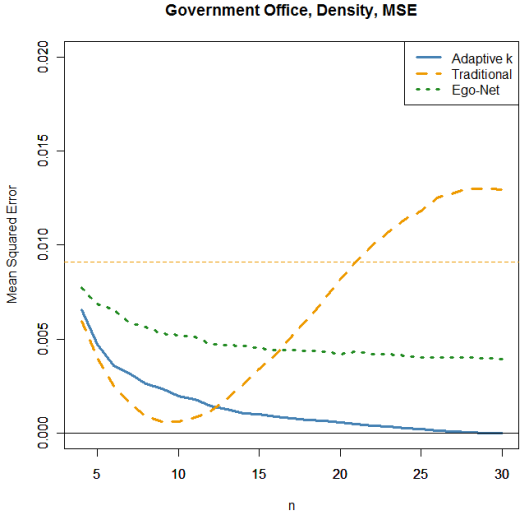


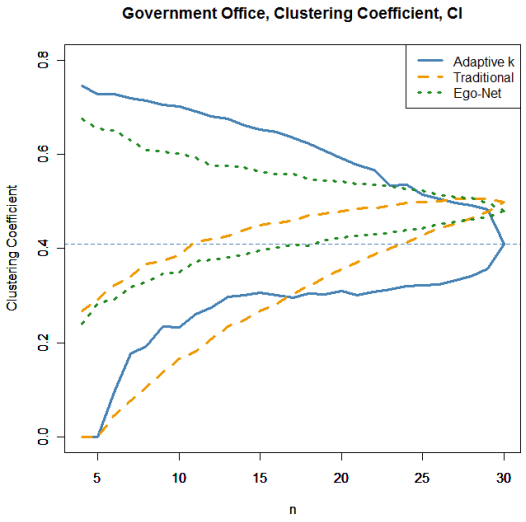


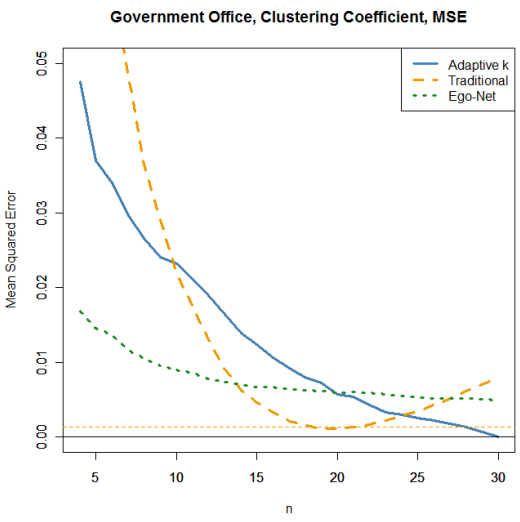


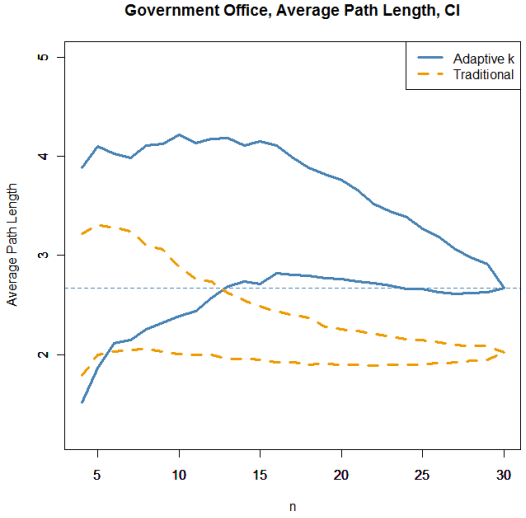
Government Office Data

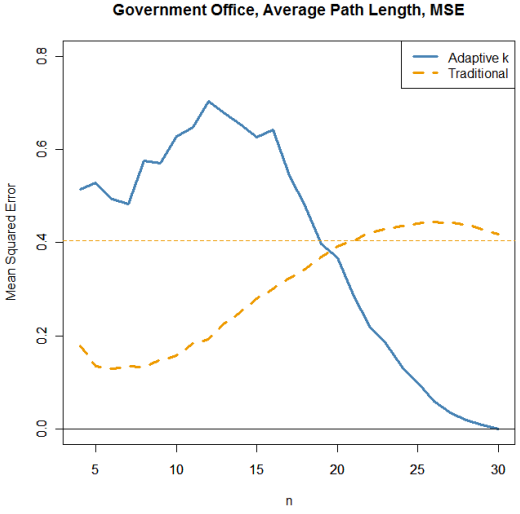




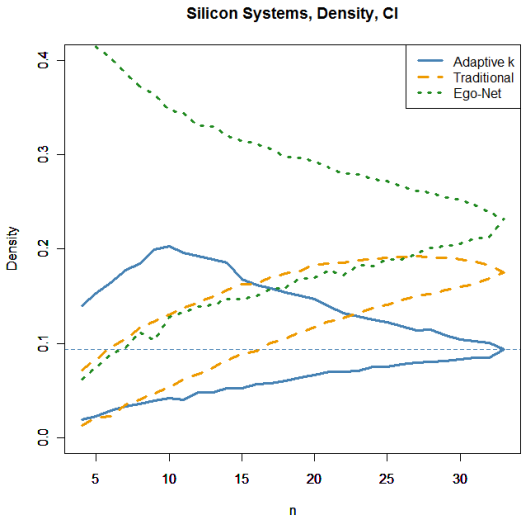

Government Office, Clustering Coefficient, CI
Adaptive k
Traditional
Ego-Net
Clustering Coefficient
n

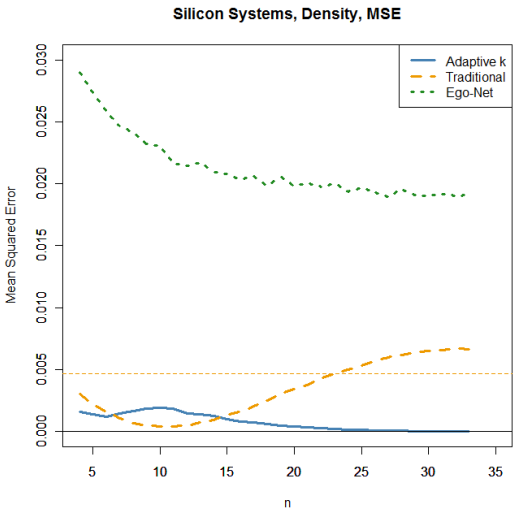

Government Office, Clustering Coefficient, MSE
Adaptive k
Traditional
Ego-Net
Mean Squared Error
n

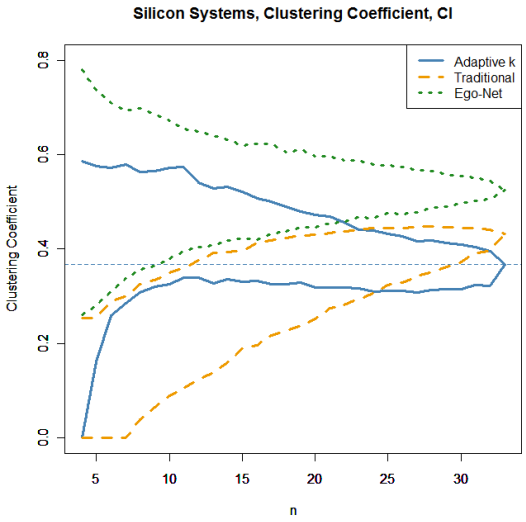

Government Office, Average Path Length, CI
Adaptive k
Traditional
Average Path Length
n

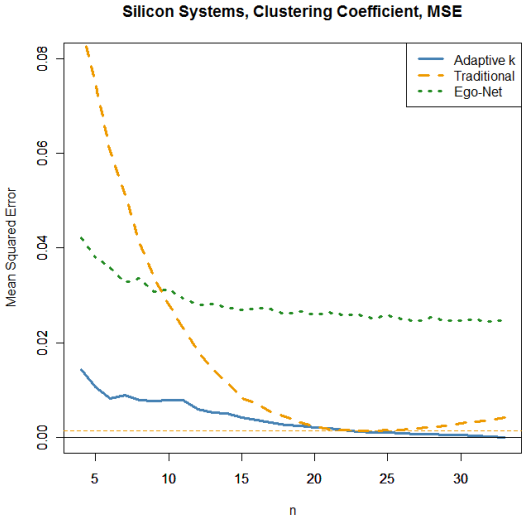

Government Office, Average Path Length, MSE
Adaptive k
Traditional
Mean Squared Error
n

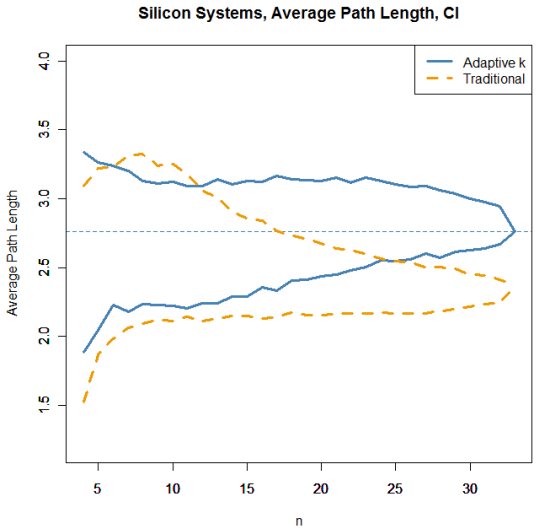
Silicon Systems Data

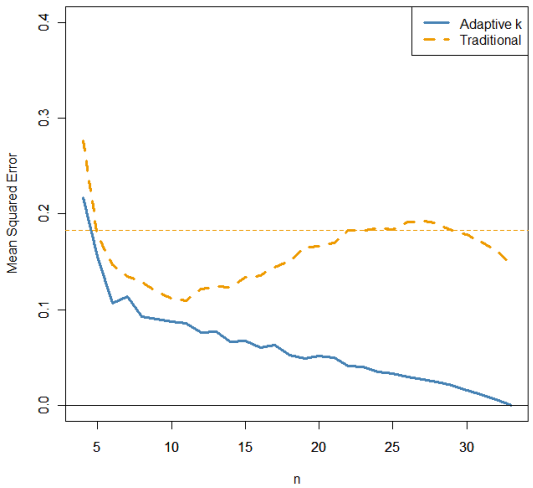




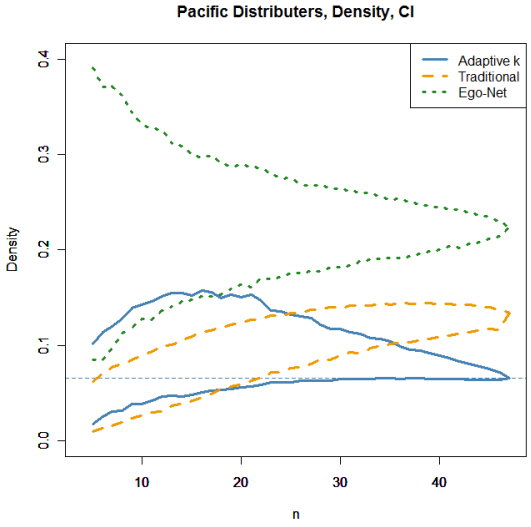


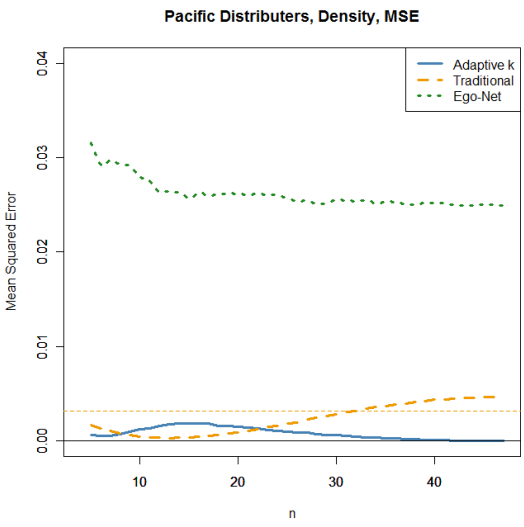


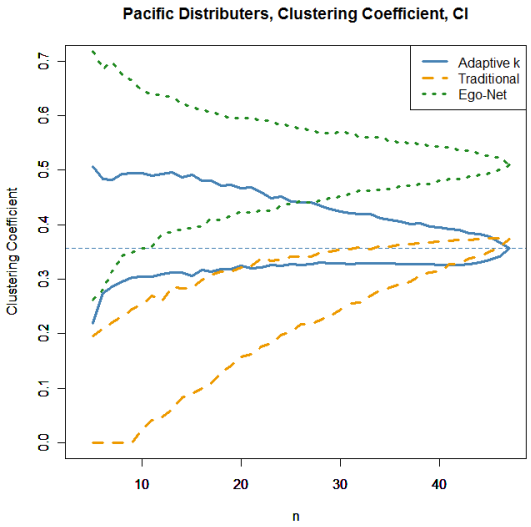


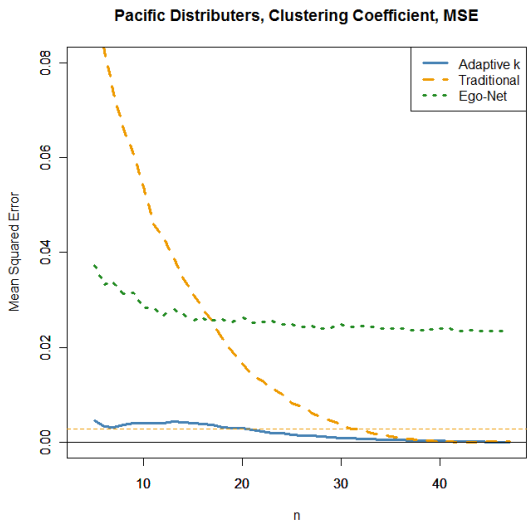


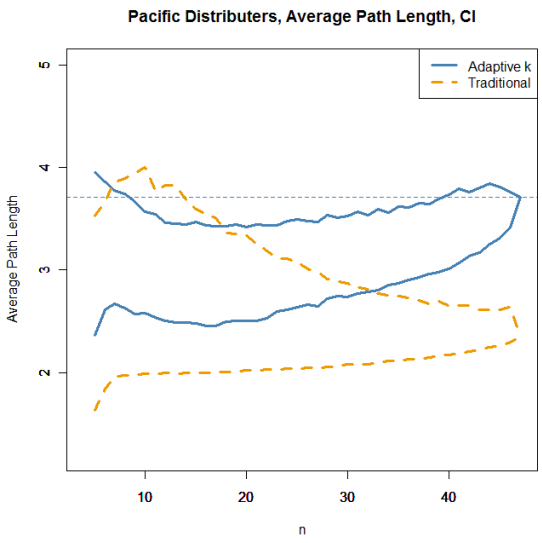
Pacific Distributers Data

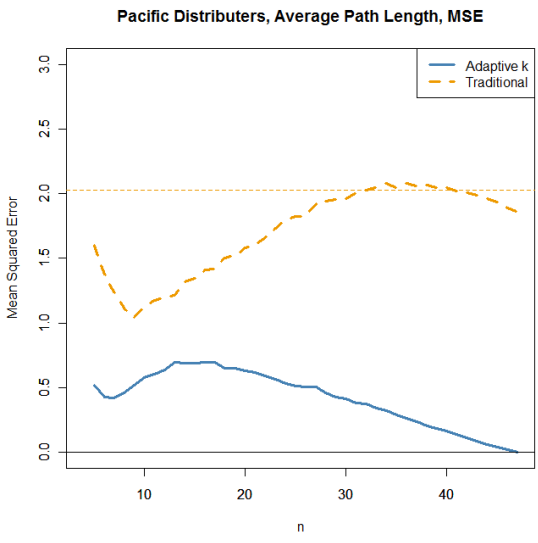












In Figure 5, both methods of estimating network structure have an upwardly sloping trend as sample size increases indicating a closer approximation to the true network. However, the traditional roster approach does not achieve 100% correlation in our trials, as it does not converge to the true network. The main reason for this lack of convergence is that under the traditional roster method, actor *i* is only asked about those individuals with whom he/she is friends. Therefore, if actor *i* claims to be tied to actor *j* then the *i-j* tie is said to exist in the network. This presents a significant opportunity for error as the existence of this tie does not take into account actor *j*'s belief of the *i-j* tie. Hence, the existence of a tie is never verified; its existence solely relies on actor *i*'s individual claim.[8] While it is theoretically possible for the roster method to perfectly reproduce the true network, in practice, due to the rosters methods reliance on row data, the resulting network tends to be biased due to inaccurate self reports.

In Figure 6, one should view *n* as representing sample size for adaptive threshold and ego-net methods, and response rate for traditional roster method. Thus, one can compare the performance of the roster method with an 70% response rate to the adaptive threshold method and ego-net method using a 30% sample size. Figure 6 clearly indicates that the adaptive threshold method provides better coverage of the true network measure than either of the two standard SNA approaches. Even with very small sample sizes, we find that the adaptive threshold method produces a 95% confidence interval which consistently captures the true global measure with a reasonable interval width. In fact, across all datasets and all global measures, except for the path length estimate in the Government Office, the true value is within the confidence bands at all sample sizes. We can also compare the different methods' performance through MSE plots. By comparing the MSE plots, one can determine the point at which the adaptive threshold method outperforms the other methods. For example, the MSE plot for High Tech Managers network density indicates that, no matter how small the sample size used with the adaptive threshold method, it performs better than the traditional method with 70% response rate, in terms of MSE. When we look at the MSE graphs for the five datasets and three network measures, we see that overall the adaptive threshold method outperforms the two traditional methods in terms of MSE, especially for density and average path length.

The graphs reveal that the ego network and the roster method approach tend to overestimate the density and clustering coefficients. One reason for the overestimation of density is that respondents tend to over-report the number of friends they have (Kumasar et al. 1994). While this tendency is still present with the adaptive threshold method when determining ties based on knowledge, its influence is diminished by checking the validity if *i*'s claim of the *i-j* tie with *j*'s claim of the *i-j* tie. In instances where neither *i* nor *j* are sampled, respondents' inclination to claim more ties than are actually present is eliminated due to the reliance on the perception of others.

Increases in density tend to lead to increases in clustering simply due to the larger number of ties. Heider (1958) provides an additional, psychological reason, why the clustering coefficient may be inflated for the ego network method. Based on Heider's (1958) balance theory, individuals tend to view relationships as being transitive. If actor J is friends with actor A and actor B, but unaware of the relationship between A and B, actor J will tend to assume a tie exists between A and B to form a balanced triad. Freeman (1992) discovered that a large number of the errors in a respondent's recall of a previously observed network could be attributed to his/her inclination to correct intransitivity. With regard to calculating path length, the overestimation of density results in an underestimation of the average distance between actors.

Overall, the comparisons strongly indicate that the adaptive threshold method outperforms the traditional SNA approaches to data collection. In addition to the improved accuracy in estimating the network structure, the adaptive threshold method also provides the researcher with the benefits of random sampling and a reduced need for high participation rates. These benefits can greatly enhance comparative research grounded in the network perspective. There are two cases in which the adaptive threshold method does not perform as well. Both the clustering coefficient and the average path length in the Government Office dataset are poorly estimated at small sample sizes relative to the other methods. A primary reason for

[8] The fact that *i* and *j* may differ in their perception of the *i-j* friendship tie may appear unintuitive. We refer the reader to Carly and Krackhardt (1996), which details the processes by which non-symmetry and non-reciprocation occur with interaction based behaviors. Because of this, we would advise researchers interested in employing the traditional roster approach to always ask about directed relationships in both directions. While this often takes place with instrumental ties, such as advice seeking/advice providing, such an approach is also necessary for affective ties that appear logically symmetric.

the poor performance is the large confidence intervals arising from the variation in the actors' perception of the network. The incorporation of a significant number of additional datasets would be necessary to better understand the contextual conditions that may give rise to large variation in actor perception. However, it is also possible that the results are simply idiosyncratic to the Government Office. Regardless of the factors influencing perception, a researcher seeking to a measure a network would, on average, be best served by using the adaptive threshold method.

6. Thoughts on Implementation and Limitations

Research utilizing a network perspective (Brass 1995; Cross et al. 2003) is seen as an improvement upon traditional econometric or statistical models as it investigates social influence, and not solely individual attributes, as explanation of social phenomena (Burt 1992, p. 4). Previous research has demonstrated that informal social structures facilitate communication, collaboration, knowledge transfer, and innovation within an organization (Kilduff and Tsai 2003; Kilduff and Krackhardt 2008). Experimental work dating back to the 1950s has demonstrated the importance of communication patterns for group performance (Bavelas 1950; Guetzkow and Simon 1955; Leavitt 1951; Shaw 1964). The effect of network structure on collective outcomes has been demonstrated for bank profitability (Krackhardt and Hansen 1993), work groups of a telecommunications firm (Cummings and Cross 2003), mental health networks (Provan and Milward 1995; Provan and Sebastian 1998), artistic groups (Uzzi and Spiro 2005), and electronic product development projects (Hansen 1999). Such cross-network research could be enhanced through the adaptive threshold method by facilitating the researcher's ability to gather data on a larger number of networks. Furthermore, there are several current research areas that could benefit by comparing structure and performance across multiple networks.

One example is the role of social networks in teacher and school performance. Daly et al. (2010, p. 363) state that "teachers working in collaboration tend to have a wider skill variety, be more informed about their colleagues' work and student performance, report increased instructional efficacy, and are more likely to express higher levels of satisfaction". Schools with higher levels of social capital have been show to have higher performance (Leana and Pil 2006). However, the relationship between social capital and school performance has not been rigorously tested using network methods. A comparative study could test the relationship between the actual structure of a social network in a school that gives rise to social capital and the collective performance of the teacher's in the school. Gathering data in the 50 to 100 schools necessary for statistical analysis may be prove difficult for the researcher using a standard roster approach. With the adaptive threshold method, a small random sample of teachers in each school could provide all of the necessary structural information. In addition, because adaptive threshold uses a random sample, additional information concerning the school climate, organizational commitment, and other important school level factors can be estimated from the participating teachers.

When researchers look to employ the adaptive threshold method in a study such as the one described above, two key decisions must be made. One is determining the α rate and the other is determining the sample size needed from each organization or network under study. As noted above, α rates in the 0.08 to 0.12 range tended to perform the best for our measures and for our networks. Based on the results in Section 5, it appears that sampling percentages as low as 25-40% can produce accurate results and provides the researcher with a large enough sample to deal with several non-respondents. Clearly, in any given instance, the necessary sample size is dependent upon the accuracy of the actors in the network. If individuals in a particular network have higher quality perceptions about the relations around them, then even smaller sample sizes will produce an accurate picture of the network. Cases like the Government Office, where the clustering coefficient and average path length were poorly estimated at low sample sizes, provide a perfect example indicating that our methods are sensitive to the quality of perception slices. When the correlations between the slices are low, the variances of our estimators may be large for small sample sizes as the few 'knowledge' ties are not able to correct the false 'perception' ties yet. This is illustrated by the large MSE values for small sample sizes in the Government Office dataset. The researcher should be aware of this sensitivity issue when selecting a sample size and a threshold value alpha.

This paper represents only a first attempt at probing the applicability of these methods. Future work on this topic can attempt to improve upon the current performance of our estimator. A potential means of improvement would be to measure the accuracy for each individual in the sample rather than the accuracy of the overall sample. Given that

individuals have varying propensities to commit Type 1 and Type 2 errors, such information could be used to weight the perception of each individual. One could identify the individuals who tend to make errors of commission and those who tend to make errors of omission and use this information to better determine the existence of ties based on perception data. These methods may also be adapted to aid researches with needs that extend beyond a random sample or wish to utilize other sampling techniques to improve estimate accuracy. For example, research has shown that people with higher level positions in companies tend to have poorer cognitive accuracy (Casciaro 1998). Given this, a researcher may choose to over sample individuals in the lower ranks. An alternative approach would be to apply an adaptive sampling framework where once an initial actor's CSS is given, the researcher can determine the next individual in the network to sample by selecting an individual who is not friends with the initial respondent. Such a method would provide greater coverage in all areas of a network that may not be achieved through random sampling alone. The current research could also be broadened by looking beyond network level measures to track the accuracy of network representation on an individual level (i.e. how often is the most central person in the network correctly identified; see for instance the work by Borgatti et al. 2006).

Another promising avenue of potential use for our methodology is in the area of hard to reach networks. Given the inherent difficulty of accessing a significant number of actors in hidden or hard to reach networks, network data collection is often impossible. However, it may be feasible to locate a few individuals in the hard to reach network, work with them to bound the network, and then use our techniques to generate estimates of the actual structure of the network.

There are limitations to our approach as well. The size of the network under study controls the applicability of our sampling methodology as a data collection tool. As noted by Krackhardt (1987), in cases where the network is reasonably large, having a respondent provide his or her perception of every tie in the network would be a difficult task. However, for networks of small to moderate sizes, cognitive structures can and have been used effectively as a data collection method. However, there is opportunity to expand the adaptive threshold method to larger networks. Burt and Ronchi (1994) attempted to measure the structure of a large intra-organizational network using a capture-recapture method by interviewing only a subset of the population understudy. Relationships between individuals not directly interviewed were determined based on the informant's perception of the strength of connection. Burt and Ronchi (1994) found that strong relations tended to be "recaptured", meaning they were perceived by multiple informants in the network as occurring. Therefore, it might be possible to map relationships among individuals in a large network using a partial cognitive social structure approach where not every individual is asked about every tie. In a network of 500 people, a 20% sample size provides 100 pieces of evidence for every tie. This may be more evidence than is necessary to accurately determine a relationship and clearly the demands on the sampled informants would be much too great to actually attempt to gather cognitive social structure data on a network that size.

One potential solution to the demands placed on a respondent to a cognitive social structure questionnaire in a larger network is the link sampling design discussed by Butts (2003). With a link sampling design, individuals are not required to provide information on all possible relationships in his/her network but rather on only a subset of them. Finding the proper balance between the size of the sample and the size of the subset of links an individual in the sample would be required to provide information on is an interesting next step for this approach. Locating the balance necessary to reduce respondent burden and maintain accurate network representations would greatly expand the potential range of application for our method.

7. Conclusion

The adaptive threshold method proposed in this study is well suited for network research where complete data collection is costly or impossible, particularly for cross-network studies. Large scale cross-network studies are rare due to the time and expense required to collect network data on a large number of networks for statistical analysis as well as concerns over the validity of network structure when missing data is present. Our sampling methods can drastically reduce researcher time and effort needed to uncover network structures, and, as demonstrated in the paper, are capable of producing accurate representations of the true network. More importantly, our methods proved to be more reliable than either of the two alternative measures to collect network data.

Acknowledgements:

The authors wish to thank David Krackhardt and Tiziana Casciaro for graciously providing the datasets used in the analysis in Section 5. The authors would also like to thank Clayton Wukich and Annemie Maertens, for their helpful comments on earlier drafts of this paper.

Works Cited:

Anseel, F., Lievens, F., Schollaert, E., Choragwicka, B., 2010. Response Rates in Organizational Science, 1995–2008: A Meta-analytic Review and Guidelines for Survey Researchers. Journal of Business and Psychology 25, 335-349.

Barton, A., 1968. Bringing Society Back In: Survey Research and Macro-Methodology. American Behavioral Scientist 12, 1-9.

Baruch, Y., Holtom, B.C., 2008. Survey response rate levels and trends in organizational research. Human Relations 61, 1139-1160.

Batchelder, W., Kumbasar, E., Boyd, J.P., 1997. Consensus Analysis of Three-Way Social Network Data. The Journal of Mathematical Sociology 22, 29-58.

Bavelas, A., 1950. Communication Patterns in Task-Oriented Groups. The Journal of Acoustical Society of America 22, 271-282.

Borgatti, S.P., Carley, K., Krackhardt, D., 2006. On the Robustness of Centrality Measures Under Conditions of Imperfect Data. Social Networks 28, 124-136.

Brass, D.J., 1995. A Social Network Perspective on Human Resources Management, Research in Personnel and Human Resources Management. JAI Press, Greenwich, CT, pp. 39-79..

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

Burt, R.S., Ronchi, D., 1994. Measuring a Large Network Quickly. Social Networks 16, 91-135.

Butts, C.T., 2003. Network Inference, Error, and Informant (in)accuracy: A Bayesian Approach. Social Networks 25, 103-140.

Carley, K.M., Krackhardt, D., 1996. Cognitive inconsistencies and non-symmetric friendship. Social Networks 18, 1-27.

Casciaro, T., 1998. Seeing Things Clearly: Social Structure, Personality, and Accuracy in Social Network Perception. Social Networks 20, 331-351.

Casciaro, T., Carley, K., Krackhardt, D., 1999. Positive Affectivity and Accuracy in Social Network Perception. Motivation and Emotion 23, 285-306.

Costenbader, E., Valente, T.W., 2003. The Stability of Centrality Measures When Networks are Sampled. Social Networks 25, 283-307.

Cross, R.L., Parker, A., Sasson, L., 2003. Networks in the Knowledge Economy. Oxford University Press, Oxford.

Cummings, J.N., Cross, R.L., 2003. Structural Properties of Work Groups and Their Consequences for Performance. Social Networks 25, 197-210.

Daly, A.J., Moolenaar, N.M., Bolivar, J.M., Burke, P., 2010. Relationships in Reform: The Role of Teachers' Social Networks. Journal of Education Administration 48, 359-391.

Frank, O., 2005. Network Sampling and Model Fitting, in: Carrington, P.J., Scott, J., Wasserman, S. (Eds.), Models and Methods in Social Network Analysis. Cambridge University Press, New York, pp. 31-56.

Freeman, L.C., 1992. Filling in the Blanks: A Theory of Cognitive Categories and the Structure of Social Affiliation. Social Psychology Quarterly 55, 118-127.

Freeman, L.C., 2004. The Development of Social Network Analysis : A Study in the Sociology of Science. Empirical Press, Vancouver, BC.

Gower, J.C., Legendre, P., 1986. Metric and Euclidean Properties of Dissimilarity Coefficients. Journal of Classification 3, 5-48.

Guetzkow, H., Simon, H.A., 1955. The Impact of Certain Communication Nets Upon Organization and Performance in Task-Oriented Groups. Management Science 1, 233-250.

Hansen, M.T., 1999. The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits. Administrative Science Quarterly 44, 82-111.

Heckathorn, D., 1997. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. Social Problems 44, 174-199.

Heider, F., 1958. The Psychology of Interpersonal Relations. Wiley, New York,.

Kilduff, M., Crossland, C., Tsai, W., Krackhardt, D., 2008. Organizational Network Perceptions Versus Reality: A Small World After All? Organizational Behavior and Human Decision Processes 107, 15-28.

Kilduff, M., Krackhardt, D., 2008. Interpersonal Networks in Organizations: Cognition, Personality, Dynamics, and Culture. Cambridge University Press.

Kilduff, M., Tsai, W., 2003. Social Networks and Organizations. Sage, London.

Krackhardt, D., 1987. Cognitive Social Structures. Social Networks 9, 109-134.

Krackhardt, D., 1990. Assessing the Political Landscape: Structure, Cognition, and Power in Organizations. Administrative Science Quarterly 35, 342-369.

Krackhardt, D. (1996). "Comment on Burt and Knez's Third Party Effects on Trust." Rationality and Society 8(1): 111-120.

Krackhardt, D., Hanson, J.R., 1993. Informal Networks: The Company Behind the Chart. Harvard Business Review 71, 104-111.

Kumbasar, E., Rommey, A.K., Batchelder, W.H., 1994. Systematic Biases in Social Perception. American Journal of Sociology 100, 477-505.

Leana, C.R., Pil, F.K., 2006. Social Capital and Organizational Performance: Evidence from Urban Public Schools. Organization Science 17, 353-366.

Leavitt, H.J., 1951. Some effects of certain communication patterns on group performance. Journal of Abnormal and Social Psychology 46, 38-50.

Pattison, P., 1994. Social Cognition in Context, in: Wasserman, S., Galaskiewicz, J. (Eds.), Advances in Social Network Analysis. Sage Publications, Thousand Oaks, pp. 79-109.

Provan, K.G., Milward, B.H., 1995. A Preliminary Theory of Interorganizational Network Effectiveness: A Comparative Study of Four Community Mental Health Systems. Administrative Science Quarterly 40, 1-33.

Provan, K.G., Sebastian, J.G., 1998. Networks Within Networks: Service Link Overlap, Organizational Cliques, and Network Effectiveness. Academy of Management Journal 41, 453-563.

Romney, A.K., Weller, S.C., Batchelder, W.H., 1986. Culture as Consensus: A Theory of Culture and Informant Accuracy. American Anthropologist 88, 313-338.

Shaw, M.E., Leonard, B., 1964. Communication Networks, Advances in Experimental Social Psychology. Academic Press, pp. 111-147.

Sparrowe, R.T., Liden, R.C., Wayne, S.J., Kraimer, M.L., 2001. Social Networks and the Performance of Individuals and Groups. Academy of Management Journal 44, 316-325.

Stork, D., Richards, W.D., 1992. Nonrespondents in Communication Network Studies: Problems and Possibilities. Group and Organization Management 17, 193-209.

Uzzi, B., Spiro, J., 2005. Collaboration and Creativity: The Small World Problem. American Journal of Sociology 111, 447-504.

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

Appendix A: Formulation of the Adaptive Threshold Methodology

In order to formulize the methodology described in the text, we introduce the following notation and summarize the methodology in a three-step procedure. Consider an unknown *NxN* network θ. Assume we randomly selected *n NxN* CSS

slices from this network, say $X_1, X_2, \ldots, X_n$. For a given vector *t*, let 1(*t*) denote an *NxN* matrix whose row and column entries corresponding to *t* are 1's and the rest are 0's. Similarly, let 1(*-t*) denote an *NxN* matrix whose row and column entries corresponding to *t* are 0's and the rest are 1's. Let *s* denote a vector containing the index numbers of the sampled slices. For example, in the 5x5x5 example in Section 4, where *A,D,* and *E* are sampled, *s*={1,4,5}. We denote the sampled and unsampled portions of the network by *S* and *U* respectively, where

$$S = 1(s),$$

$$U = 1(-s).$$

Let $I_t(A)$ denote a function which assigns 1's to all the entries of a matrix *A* that are greater than a given constant *t*, and assigns 0's to all the remaining entries. Let "*" denote the element by element multiplication of two matrices. For a given threshold *k*, we would like to find $\hat{\theta}$, an estimate of θ, based on the CSS slices $X_1, X_2, \ldots, X_n$.

Step 1: Find exact entries of $\hat{\theta}$

For all *i*=1,…,*n*, decompose X_i such that

$X_i = X_{K,i} + X_{K,i}$,

where $X_{K,i}$ denotes the knowledge portion of X_i, and $X_{P,i}$ denotes the perception portion of X_i. We have

$X_{K,i} = 1(i) * X_i$,

$X_{P,i} = 1(-i) * X_i$.

The combined knowledge and combined perception in the observed sample, denoted by *K* and *P* respectively, are given as

$K = \sum_{i=1}^{n} X_{K,i}$,

$P = \sum_{i=1}^{n} X_{P,i}$.

Then the exact entries of $\hat{\theta}$, computed from the knowledge and will not be changed by other perceptions, are contained in the matrix *E* given by

$$E = I_1(S * K).$$

Step 2: Find perception entries of $\hat{\theta}$

As we discussed in Section 4, unverified (or undenied) knowledge ties will be treated as a perception. Then the perception contribution of knowledge from Step 1 is contained in the matrix *C*, where

$$C = U * K.$$

We will decompose perception *P* into two parts, the active part (denoted by P_A) and the inactive part (denoted by P_I). We will use P_A to find the perception entries of $\hat{\theta}$. P_I will be used to find $\hat{\alpha}$. We have

$P_A = U * P$,

$P_I = I_1(S * P)$.

Combining *C* and P_A, we have the final perception matrix P_F, given by

$P_F = I_k(P_A + C)$,

which contains the perception entries of $\hat{\theta}$.

Step 3: Combine Steps 1 and 2 to find $\hat{\theta}$

The estimated network is given by

$\hat{\theta} = E + P_F$.

Recall that in the adaptive threshold method we estimate the Type 1 error probability for a given *k*, and denote this quantity by $\hat{\alpha}_k$. Using the above notation we have

$\hat{\alpha}_k = sum(P_l)/(sum(S) - sum(E)),$

where *sum*(*A*) denotes the sum of all entries of a given matrix *A*. Note that this is equivalent to equation 3 given in Section 4.