Mathematical Modeling and Inference for Degree-capped Ego-centric Network Sampling

Babak Fotouhi\textsuperscript{1,2} and Steven Rytina\textsuperscript{3}

\textsuperscript{1} Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, USA
\textsuperscript{2} Institute for Quantitative Social Sciences, Harvard University, Cambridge, MA, USA
\textsuperscript{3} Department of Sociology, McGill University, Montreal, QC, Canada

Abstract

The structure of social networks is usually inferred from limited sets of observations via suitable network sampling designs. In offline social network sampling, for practical considerations, researchers sometimes build in a cap on the number of social ties any respondent may claim. It is commonly known in the literature that using a cap on the degrees begets methodologically-undesirable features because it discards information about the network connections. In this paper, we consider a mathematical model of this sampling procedure and seek analytical solutions to recover some of the lost information about the underlying network. We obtain closed-form expressions for several network statistics, including the first and second moments of the degree distribution, network density, number of triangles, and clustering. We corroborate the accuracy of these estimators via simulated and empirical network data. Our contribution highlights notable room for improvement in the analysis of some existing social network data sets.

I. Introduction

In studies of social networks, explanations and predictions are conducted and expressed in terms of the structural properties of the underlying webs of connections and interactions. In order for the theoretical results to be applicable and the empirical studies to be practical, these structural properties must be observed and measured. This highlights the need for methods to efficiently and reliably estimate these structural properties from available observations. This is the focal task of network sampling.

The network sampling literature embodies several data collection methods (sampling designs), suitable for different contexts, such as the World Wide Web (Leskovec & Faloutsos, 2006),
the Internet (Achlioptas et al., 2005), and biological networks (Dempsey et al., 2012; Jordano, 2015)—also see Kolaczyk (2014) for a review. These methods differ on how they select subsets of the networks to be observed. Their corresponding mathematical and statistical methods for estimation also differ. In this paper, we focus on offline social network sampling.

Sampling offline social networks has particular practical challenges. Different practical constraints limit the possibilities of observation and measurement of social connections—as shall be discussed below. The simplest procedure would be to select a sample of the population as respondents (ego), and ask them to nominate their social connections (alters) according to given instructions. The contents of these ‘name generator’ questions depend on the context of the study. Ideally, one would like to find the alters that egos nominate, and to interview those alters, and continue to expand until a certain limit is reached. This would be called a sociocentric design (Perkins et al., 2015), which provides richer data, but is sometimes prohibitively costly. Thus some social network studies employ the ego-centric design, which is the focus of the present paper. Both ego-centric and socio-centric designs face sources of noise such as imperfect recollections, memory distortions, question order effects, and satisficing (Van Tilburg, 1998; Marsden, 2003, 2011; Pustejovsky & Spillane, 2009). Another major problem (which is not the focus of our paper) arises in the sampling designs that involve observing marginalized and/or stigmatized hidden populations. See Gile et al. (2018) and Goel and Salganik (2010), and the references therein, for ‘respondent-driven sampling’, which addresses this problem.

Due to time and economic constraints, and due to respondent fatigue (which can intensify the noise-inducing factors discussed above), some ego-centric designs ask respondents to nominate a fixed number of alters (which we refer to as the cap and denote by \(c\)). This fixed-choice approach dates back to the pioneering empirical studies in social network research: Alexander Jr and Campbell (1964) used \(c = 5\), Coleman et al. (1966) used \(c = 3\), Wellman (1979) and Wellman et al. (1973) used \(c = 6\), Laumann (1973) and Fischer (1977) used \(c = 3\), and Fischer (1982) used \(4 \leq c \leq 10\) for different name generators. The network component of the General Social Survey employs \(c = 5\) (Burt, 1984; McPherson et al., 2006). National Longitudinal Study of Adolescent Health (Add Health) uses \(c = 10\) (Harris et al., 2009). The National Social Life, Health, and Aging Project (NSHAP) employs \(c = 5\) (Cornwell et al., 2009). The network part of the American National Election Study (ANES) used \(c = 4\) (Burns et al., 2001). The network part of the Gallup Panel reported by Omalley et al. (2012) employs \(c = 4\). In the supplementary material, we provide over 60 other examples of social network studies in various domains which
use fixed-choice data, along with their \( c \) values.

There are contexts where the introduction of the degree cap might not be of major concern. For example, if the type of tie that the name generator corresponds to typically returns few alters, then the sampling results might be reliable (and one can use methods for inferring network parameters from full ego networks, some of which are discussed in Section II). This would be the case, for instance, when studying intimate social relations, such as secret sharing, money borrowing, and advice seeking. In this paper, we assume that the typical number of ties is larger than \( c \). In such cases, the effect of applying the degree cap is more pronounced and discards information about the network structure.

It is known in the literature that the network acquired via applying the degree cap (and thereby truncating the degrees) is markedly different from the original social network. This “can dramatically alter estimates of network-level statistics” (Kossinets, 2006), and will “… underestimate the total linkages present in the system” (Shrum & Cheek Jr, 1987). Also “The restriction of the nomination process also introduces error into the measurement of other network properties, such as properties of triads…” (Wasserman & Faust, 1994). Hence fixed-choice methods have been deemed “for most purposes undesirable” (Newman, 2010). See also Holland and Leinhardt (1973), Prell (2012), and Robins (2015) for other critique of this method and warnings against using it. Moreover, the network obtained after degree truncation can have markedly different functional properties than the original network, such as those relevant in the spread of infectious disease over social networks (Harling & Onnela, 2016). This paper takes a step towards remedying these problems.

In this paper, we provide a mathematical model of the fixed-choice sampling procedure. The task we focus on is finding closed-form analytical expressions for several parameters of the original network, using the sampled data. After introducing the sampling setup and model, we propound solutions for the network size, first and second moments of the degree distribution, density, the number of triads and triangles, and the clustering coefficient. We also use our results to obtain approximations for the epidemic threshold. As a byproduct, our results suggest that using the sampled network as proxy for the true underlying social network can underestimate the strength and probability of contagion/diffusion processes, which is consistent with the previous simulation studies in the literature (Harling & Onnela, 2016).
II. RELATED WORK

Our task is to develop analytical expressions to estimate global network parameters from fixed-choice data. There is a rich literature devoted to the statistical inference of the network structure from various sampling designs. We are not aware of previous studies that consider a problem statement identical to ours. Below, we highlight some studies which have considered setups which share similarities with our setup.

Smith (2012) used a novel simulation study that juxtaposes Exponential Random Graph Models (ERGM) and case control logistic regression to infer global network structure from samples of ego networks. The setup of Smith (2012) has two differences with ours: ego networks are fully observed (i.e., the true degree of egos are known), and ties between the alters are elicited. In our paper, we assume that no information about the true number of alters is known, and no information about the ties between alters is available. An improvement to the latter results is obtained by Smith (2015), where the limit on the number of elicited alters is also incorporated into the simulations. The advantage of this improved analysis is robustness against degree skew, i.e., existence of individuals with outstandingly large degrees that skew the degree distribution. There is a difference with the setup that we consider: ties between the alters are assumed to be available, that is, the motifs that are used in the ERGM analysis include the inter-alter ties. In our paper, we assume only the links between the ego and alters have a chance of being observed.

We distinguish between our setup and that of snowball sampling. Snowball sampling involves choosing a set of initial seed nodes, then observing all their links and their neighbors, then surveying every neighbor and observing all links and neighbors of them, and so on. The \(k\)-waves snowball sampling involves iterating this procedure \(k\) times. There is a rich literature on snowball sampling. See Koskinen et al. (2013) and Stivala et al. (2016) and the references therein, for example. See Stivala et al. (2016) for the use of ERGMS to analyze snowball sampling. In this paper, we consider a different setup from snowball sampling. We assume only one wave, and in that one wave, only a fixed number of links for each ego have a chance of being observed, whereas in snowball sampling, all of them are observed. The ERGM approach has also been employed to handle randomly-missing data (Robins et al., 2004; Koskinen et al., 2013).

In Ott et al. (2017), simulation studies are used to obtain improved estimates of associations between behaviors and social relationships by maximizing the likelihood of observed fixed-choice data. Ott et al. (2017) also consider an augmented to fixed-choice sampling, where in addition to
the list of the nominated alters, each respondent also reports its true degree. This is different from our problem because we assume the true degrees are unknown. We also consider no covariates, and our task is to obtain analytical expressions to estimate the global network statistics. Moreover, in our setup, we do not assume that every node in the network is surveyed. We seek to solve for the structural properties of the underlying network from fixed-choice observations of only a small fraction of the nodes.

In the next section, we first develop a model for the respondent selection procedure and the alter nomination process, describe the observed network under this sampling design, and then focus on several structural properties of the underlying network.

III. METHODS

We employ a variant of the method of Frank and Snijders (1994) which is a network version of the method of moments that they applied to snowball sampling. We delineate the method in depth below, but the steps can be summed up as follows: defining a number of observables in terms of original network parameters, calculating their expected value, and equating these expected values with empirically-observed values to get a system of equations, and solving for the parameters by inverting the system.

The sampling procedure involves selecting an initial set of $n_e$ respondents (where $e$ stands for ego), and asking each of them to introduce $c$ alters. We assume that $c$ is greater than the minimum degree of the network.

We can construct a network in which respondents and named alters are the nodes and all the claimed ties are the links. Let $\tilde{G}$ denote this crude network, and let $G$ denote the original social network. We consider the basic setup of an unweighted undirected network. Note that some mentioned alters might themselves be among the respondents, that is, a respondent can name another respondent as an alter. So within $\tilde{G}$, some links connect a respondent to a non-respondent, and some links connect two respondents. Let $\ell_{ee}$ and $\ell_{ea}$ denote the number of these two types of links, respectively. Figure 1 presents the sampling setup in a schematic nutshell.

We proceed by modeling the selection process of the respondents as an independent Bernoulli process with unknown probability $q$. This is a conventional step in network sampling literature (Kolaczyk, 2014; Frank & Snijders, 1994). We seek to infer some structural parameters of $G$ with high practical import. We focus on $N$ (the size of the population from which samples are taken), $\mu_1$ (average degree), $\mu_2$ (second moment of the degree distribution), $\rho$ (network density,
Fig. 1: An illustrative example of the sampling scheme with example cap value $c = 2$. The red nodes are the respondents. The available ties to each node is depicted in bold lines. The arrows signify mentioned ties, and they point from the respondents to the alters. The dashed lines represent the ties of the respondents that were not mentioned in the interview. As visible from the figure, the crude (observed) network can be disconnected. Its structure clearly differs from that of the original network. The values of $n_e$, $\ell_{ee}$, and $\ell_{ea}$, are given for explanatory purposes. These values are used to estimate $\mu_1$, $N$, and $q$, as discussed in the text. The node size in the depicted figure is a linear function of the node degree.

which measures how close-knit the members of a society are, i.e., how close the network is to a complete graph), clustering coefficient (triadic closure ratio, which is a measure of social cohesion due to transitivity, i.e., the degree to which the friends of one’s friends are themselves friends), the number of triangles and open triads (where by ‘open triad’ we mean local structures in which two friends of an ego are not friends with each other, as opposed to triangles), and the epidemic threshold and basic reproduction number for SIS and SIR models (which characterizes the degree to which a network is susceptible to potential epidemic outbreaks under these models). We also infer $q$, which is a parameter of the model. Table I presents a list of variables used throughout the analysis.
To employ the method of moments (Kolaczyk, 2014; Frank & Snijders, 1994), we first calculate the expected value of $n_e$, $\ell_{ee}$ and $\ell_{ea}$ and construct a system of equations which leads us to estimators for $N$, $q$, and $\mu_1$. Using these estimators, we provide an estimator for network density. We then move to the clustering coefficient. To that aim, we focus on the number of triads and triangles. We calculate the expected number of observed triangles and open triads in $\tilde{G}$ as functions of the number of triangles and open triads in $G$, and through the resulting system of equations, we obtain estimators for the number of open triads and triangles, and combine them to estimate the clustering coefficient, as well as the second moment of the degrees. We then use these results to obtain the epidemic threshold and basic reproduction number under the mean-field approximation for SIS and SIR models (Pastor-Satorras et al., 2015; Lindquist et al., 2011). We corroborate the accuracy of these estimators for several synthetic and real network data sets. Finally, we point out stark contrasts between the inferred quantities and the crude ones, highlighting the importance of performing inference and the substantial room for improvement for the accuracy of the existing results in the social networks literature.

### IV. RESULTS AND DISCUSSION

We provide the contours of the calculations for the estimators, leaving the minutia (and the accompanying calculations pertaining to the variances) to be expounded in the supplementary material. We denote the element $ij$ of the adjacency matrix of $G$ by $A_{ij}$. Accommodating the assumption of unweightedness and considering $N$ to be large, the expected value of $\ell_{ee}$ is given

<table>
<thead>
<tr>
<th>observable</th>
<th>symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of respondents (egos)</td>
<td>$n_e$</td>
</tr>
<tr>
<td>number of ego-ego links</td>
<td>$\ell_{ee}$</td>
</tr>
<tr>
<td>number of ego-alter links</td>
<td>$\ell_{ea}$</td>
</tr>
<tr>
<td>number of open triads in $\tilde{G}$</td>
<td>$\tilde{\Lambda}$</td>
</tr>
<tr>
<td>number of triangles in $\tilde{G}$</td>
<td>$\tilde{\Delta}$</td>
</tr>
<tr>
<td>inclusion probability</td>
<td>$q$</td>
</tr>
<tr>
<td>population size</td>
<td>$N$</td>
</tr>
</tbody>
</table>

**TABLE I:** observables used throughout the analysis and their corresponding notation (we use these observables to estimate unknown network statistics of the underlying network.)
by

\[ \mathbb{E}(\ell_{ee}) = \frac{q^2}{2} \sum_{ij} \left( \frac{c}{k_i} + \frac{c}{k_j} - \frac{c^2}{k_i k_j} \right) = \frac{q^2}{2} \left( 2Nc - c^2 \sum_{ij} A_{ij} \right) \simeq \frac{q^2Nc}{2} \left( 2 - \frac{c}{\mu_1} \right), \]

(1)

where the sum runs over all nodes in \( G \). See Equations 1 to 8 in the SI for calculation details.

Note that as an approximation, we have made a crucial assumption in modeling the interview process, which is not universally applicable. We have assumed the network is unweighted, that is, no information of tie weights are available. If such information is at hand, a reasonable approximation would be to assume that all the ties of each respondent have equal probability of being mentioned (any heterogeneity will inevitably need some measure of tie weight). As we discuss later, incorporating tie strength would be an important line of future work. We have also assumed that no node with degree smaller than \( c \) exists, that is, the interview process truncates the degree of every respondent. This also is reasonable for weak ties, because cap values are typically below 10 and individuals have at least an order of magnitude more weak ties than that. Finally, we have assumed that triadic closure tendencies for recollection are negligible. That is, respondent \( Z \) has the same probability of mentioning its neighbors \( W, X, Y \), where \( X \) and \( Y \) are connected to one another but \( W \) is not connected to any other neighbor of \( X \).

For the expected value of \( \ell_{ea} \), the \( q^2 \) factor changes to \( q(1 - q) \) (see Equations 9 to 11 in the SI). For the expected values of \( \ell_{ea} \) we have:

\[ \mathbb{E}(\ell_{ea}) = \frac{q(1 - q)}{2} \sum_{ij} \left( \frac{c}{k_i} + \frac{c}{k_j} \right) A_{ij} = Ncq(1 - q). \]

(2)

Furthermore, the expected value of \( n_e \) is simply \( qN \). The method of moments proceeds by equating the observed values of the said quantities to their expected values simultaneously. We need the solution to this system of three equations. The solution is obtained after straightforward algebraic steps:

\[
\begin{bmatrix}
\hat{N} \\
\hat{q} \\
\hat{\mu}_1
\end{bmatrix} =
\begin{bmatrix}
\frac{n_e}{cN} \\
1 - \frac{\ell_{ea}}{cn_e} \\
1 - \frac{\ell_{ea}}{cn_e}
\end{bmatrix}
\]

(3)

Note that these results also provide information about the minimum sample size. The denominator of the factor in \( \mu_1 \) evinces requirements that the collected sample must fulfill. The
term \( n_c \) is all the names mentioned by all the respondents. So \( n_c - \ell_{ea} \) is the number of times respondents have mentioned other respondents and \( n_c - \ell_{ea} - \ell_{ee} \) is the number of times respondents have mutually named one another. Sampling must continue until there is at least one mutual mention of respondents—so that the positivity of both of these terms is ensured. We can use the delta method to estimate the variance of these estimators (see Equations 15 to 28 in the SI). For \( N \) and \( q \), we have:

\[
\begin{bmatrix}
\hat{\sigma}^2_N \\
\hat{\sigma}^2_q
\end{bmatrix} \approx \begin{bmatrix}
n_c^2 \ell_{ea} \left( \frac{4\ell_{ee}\ell_{ea} + (1 + c)(cn_e - 2\ell_{ea})^2}{(cn_e - \ell_{ea})^4} \right) \\
\ell_{ea} \left( \frac{4\ell_{ee}\ell_{ea} + c\ell_{ea}^2 + (cn_e - 2\ell_{ea})^2}{c^4n_e^4} \right)
\end{bmatrix},
\]

(4)

and for the average degree, we have:

\[
\hat{\sigma}_\mu^2 \approx \frac{1}{4n_e^2(cn_e - \ell_{ea} - \ell_{ee})^4} \left[ c^4n_e^4 - 2c^3n_e^3(2\ell_{ee} + \ell_{ea}) + 8\ell_{ee}\ell_{ea}(\ell_{ee} + \ell_{ea}) + c^2n_e^2[4\ell_{ee}^2 + (c + 13)\ell_{ee}\ell_{ea} + \ell_{ea}^2 - 8cn_e\ell_{ea}\ell_{ee}(\ell_{ee} + 2\ell_{ea})] \right].
\]

(5)

The network density is the ratio of the number of existing links to the number of all possible links. The latter is simply \( \binom{N}{2} \), so we have \( \rho = \frac{2L}{N(N-1)} \), where \( L \) is the total number of existing links, and equals half the sum of all degrees. Thus the estimator for the network density becomes \( \hat{\rho} = \frac{n_c - \ell_{ea}}{N - 1} \), which can be expressed as follows

\[
\hat{\rho} = \frac{c}{2} \frac{(cn_e - \ell_{ea})^2}{(cn_e - \ell_{ee} - \ell_{ea})(cn_e(n_e - 1) + \ell_{ea})}.
\]

(6)

The variance of the density is too long to present in the main text and is presented in the SI.

We verify the accuracy of the estimators with simulations on synthetic networks and empirical data. The first setup network generation model we use is the small-world model (Newman & Watts, 1999) which emulates both the short average path length and the high clustering coefficient, which are evident properties of real-world social networks. The clustering is also tunable, and we diversified it in the generation process of test networks to have assessed performance for different possible levels of clustering. The results are depicted in Figure 2 for different network sizes and different values of \( q \). The performance of the estimators are remarkable: despite the fact that fixed-choice sampling basically truncates and thereby equalsizes
degrees and discards information about the heterogeneity of the degrees, the estimators are able to recover this lost information.

To assess the robustness of performance, we also employed other topologies: Erdős-Renyi (Erdős & Rényi, 1960) (ER), which is the classical random graph model, Random Geometric Graphs (Penrose, 2003) (RGG), which are used to model populations with tie formation based on geographical proximity, the scale-free model of Holme and Kim (Holme & Kim, 2002) (HK), which produces less heavy-tailed networks with high clustering, the scale-free model of Klemm and Eguiluz (Klemm & Eguiluz, 2002) (KE), which generates heavy-tailed networks with high clustering and small-world behavior, and the forest-fire model (Leskovec et al., 2005), which emulates densification and transitivity in tie formation. We also used the Barabási-Albert (Barabási & Albert, 1999) model, although the networks it generates are too heavy-tailed and exhibit too little clustering to be realistic for offline social networks (in fact, their clustering coefficient goes to zero in the large-\(N\) limit, which is certainly not the case for real-world offline social networks in which high transitivity is a universal and central feature). BA networks are more suitable for online social networks and scholarly citation networks. Yet, we consider BA as a worst-case setting. The results for these topologies are presented in Figure 3a. It can be seen that even for BA networks, whose properties deviate substantially from real networks, the accuracy of the estimator is acceptable. The details of the simulations are provided in the SI (Section IV).

We also use examples of real data sets: the collaboration network of physicists in high energy physics (Leskovec et al., 2007), denoted by ‘collaboration1’ in Figure 3b, the collaboration network of computer scientists (Yang & Leskovec, 2015), denoted by ‘collaboration2’, the Gnutella peer-to-peer file sharing network (Leskovec et al., 2007; Ripeanu & Foster, 2002), denoted by P2P, Wikipedia-vote networks (Leskovec et al., 2010b, 2010a), and the email-exchange network of Rovira i Virgili University (Guimera et al., 2003). Implementation details are provided in the SI (Section IV).

To estimate the variance of the degrees, epidemic thresholds, and the clustering coefficient, we first infer the number of open triads and triangles in \(G\). It must be noted that a certain observed triad in \(\tilde{G}\) could have been either a triangle or an open triad in \(G\), between which we must distinguish. In other words, an open triad in \(G\) cannot be observed as a triangle in \(\tilde{G}\), but a triangle in \(G\) can be observed either as an open triad or a triangle in \(\tilde{G}\). We delineate through a few examples for expository purposes.

Consider a triangle in \(G\) consisting of nodes \(\{x, y, z\}\). There are multiple mutually-exclusive
events under which this triangle of $G$ will be observed as a triangle in $\tilde{G}$. One such event is that $x$, $y$ and $z$ are all chosen as respondents, and then all three of them mention the other two as alters. Another example event would be that they are all chosen as respondents, $x$ mentions $y$ but not $z$, $y$ mentions $z$ but not $x$, and $z$ mentions $x$ but not $y$. Another event would be that $x$ and $y$ are chosen as respondents, but $z$ is not, then $x$ mentions both $y$ and $z$, and $y$ mentions $z$ but not $x$. On an unweighted structure, these example events occur with approximate expected probabilities $q^3 \frac{c^3(c-1)^3}{\mu^3(\mu_1-1)^3}$, $q^3 \frac{c^3(\mu_1-c)^3}{\mu^3(\mu_1-1)^3}$, and $q^2(1-q)\frac{c^2(c-1)(\mu_1-c)}{(\mu_1-1)^2}$, respectively. There are 36 distinct mutually-exclusive events under which this triangle in $G$ will be observed as a triangle in $\tilde{G}$. Furthermore, as mentioned above, this triangle can also be observed as an open triad in $\tilde{G}$ under certain events. One example event would be that $x$ is chosen as respondent but $y$ and $z$ are not, and then $x$ names both of them as alters. Another example event would be that both $y$ and $z$ are chosen as respondents and $x$ is not, and both of them mention $x$ among their alters. These two events happen with approximate expected probabilities $q(1-q)\frac{c(c-1)}{\mu_1(\mu_1-1)}$ and $q^2(1-q)\frac{c^2(\mu_1-c)^2}{\mu_1^2(\mu_1-1)^2}$, respectively. There are in total 17 mutually-exclusive events under which the said triangle will be observed as an open triad in $\tilde{G}$. These events are depicted along with their corresponding probabilities and multiplicities (due to symmetry) in Figure 4. We have defined the following auxiliary variables for brevity:

$$\begin{align*}
a_1 & \overset{\text{def}}{=} \frac{\hat{\mu}_1 - 1}{\mu_1} = \frac{c}{\hat{\mu}_1} \\
a_0 & \overset{\text{def}}{=} \frac{\hat{\mu}_1 - 1}{\mu_1} = \frac{c_1(c-1)\hat{\mu}_1(\hat{\mu}_1 - 1)}{\hat{\mu}_1(\hat{\mu}_1 - 1)} \\
b_2 & \overset{\text{def}}{=} \frac{\hat{\mu}_1 - 2}{\mu_1 c - 2} = \frac{c(\mu - 1)}{\mu_1(\mu_1 - 1)} \\
b_1 & \overset{\text{def}}{=} \frac{\hat{\mu}_1 - 2}{\mu_1 c - 2} = \frac{c_1(c - 1)\hat{\mu}_1(\hat{\mu}_1 - 1)}{\mu_1(\hat{\mu}_1 - 1)} \\
b_0 & \overset{\text{def}}{=} \frac{\hat{\mu}_1 - 2}{\mu_1 c - 2} = \frac{(\hat{\mu}_1 - c)(\hat{\mu}_1 - 1)}{\hat{\mu}_1(\hat{\mu}_1 - 1)}. 
\end{align*}$$

(7)

It is straightforward to check what events these probabilities pertain to: $a_1$ is the probability that a non-central node in an open triad who is chosen as respondent mentions the central node, and $a_0$ is the probability that it does not mention the central node. Similarly, $b_2$ is the probability that a central node in an open triad, or any node in a triangle, who is selected as respondent, mentions both other nodes, $b_1$ is the probability that it mentions only one of them, and $b_0$ is the probability that it mentions neither of them. These sets of probabilities satisfy $a_1 + a_0 = 1$ and $b_2 + 2b_1 + b_0 = 1$, respectively.
Now consider an open triad in \( G \), consisting of a central node \( x \) that is connected to nodes \( y \) and \( z \). There are multiple events under which this triad will be observed in \( \tilde{G} \). One example event is that \( x \) is chosen as a respondent and \( y \) and \( z \) are not chosen, and \( x \) mentions both \( y \) and \( z \) as alters. Another event is that both \( y \) and \( z \) are chosen as respondents and \( x \) is not, and both of them mention \( x \) among their alters. Another example event would be that all three of them are chosen as respondents, \( y \) and \( z \) both mention \( x \) as their alters, and \( x \) mentions only \( y \). For an unweighted structure, these example events happen with approximate expected probabilities 

\[
q(1 - q)^2 \frac{c(c-1)}{\mu_1(\mu_1-1)} , q^2(1 - q) \frac{c^2}{\mu_1}, \text{ and } q^3 \frac{c^2(\mu_1-c)(c-1)}{\mu_1^2(\mu_1-1)},
\]

respectively. There are a total of 17 distinct mutually-exclusive events under which the said open triad can appear in \( \tilde{G} \). Figure 4 depicts the 17 events under which a triad will be observed as a triad (also note the multiplicities under each given configuration which are due to symmetry).

We denote the number of open triads in \( G \) by \( \Lambda \) and the observed number of open triads in \( \tilde{G} \) by \( \tilde{\Lambda} \), the number of triangles in \( G \) by \( \Delta \) and the observed number of triangles in \( \tilde{G} \) by \( \tilde{\Delta} \), the sum of the first 11 probabilities in Figure 4 by \( P(\Delta \rightarrow \Lambda) \), the sum of the rest of the probabilities in Figure 4 by \( P(\Delta \rightarrow \Delta) \), and the sum of probabilities in Figure 4 by \( P(\Lambda \rightarrow \Lambda) \).

The moment condition for the number of triangles is 

\[
\tilde{\Delta} = \Delta \frac{P(\Lambda \rightarrow \Lambda)}{P(\Delta \rightarrow \Delta)}
\]

and the moment condition for the number of open triads is 

\[
\tilde{\Lambda} = \Delta \frac{P(\Lambda \rightarrow \Lambda)}{P(\Delta \rightarrow \Delta)} + \Lambda \frac{P(\Lambda \rightarrow \Lambda)}{P(\Delta \rightarrow \Delta)}
\]

Solving these equations for \( \Delta, \Lambda \) yields the estimators for the number of triangles and open triads:

\[
\begin{align*}
\hat{\Delta} & = \frac{\tilde{\Delta}}{P(\Delta \rightarrow \Delta)} \\
\hat{\Lambda} & = \frac{\tilde{\Lambda}}{\tilde{\Delta}P(\Delta \rightarrow \Delta) - \tilde{\Delta}P(\Delta \rightarrow \Lambda)}
\end{align*}
\]

(8)

The clustering coefficient is defined as 

\[
C = \frac{3\Delta}{A+3\Delta}.
\]

Plugging in the estimated values for \( \Delta \) and \( \Lambda \), we obtain

\[
\hat{C} = \frac{3\Delta P(\Lambda \rightarrow \Lambda)}{\tilde{\Delta}P(\Delta \rightarrow \Delta) - \tilde{\Delta}P(\Delta \rightarrow \Lambda) + 3\tilde{\Delta}P(\Lambda \rightarrow \Lambda)}
\]

(9)

Note that this is different from the crude clustering coefficient that would be obtained from \( \tilde{G} \), which would be equal to \( \frac{3\Delta}{3A+3\Lambda} \). Figure 5 presents a comparison between the accuracy of the estimator given by (9) and that of the crude estimate which uses \( \tilde{G} \). It is visible that the crude method consistently underestimates the clustering coefficient. This means that the crude social network exhibits lower levels of cohesion and transitivity (i.e., the degree to which friends of a person are friends with each other) as compared to the original social network. There is experimental evidence that higher clustering renders social networks more conducive to social
contagion and diffusion (Centola, 2010). This means that if we rely on the crude version of the network, the contagion of social phenomena will be underestimated.

Now we focus on \( \mu_2 \), the second moment of the degree distribution. As we discuss in the supplementary material, one approach would be to pursue with the method of moments by calculating the expected number of mentioned alters who are not themselves chosen as respondents, and subsequently equating it with the observed value. In the calculations, \( \mu_2 \) would be invoked, and solving the resulting equation would provide an estimator. However, this approach would lead to an estimator with large variance, because \( \mu_2 \) appears in the exponent, which is a rapidly-varying function. Here we proceed with a more robust alternative. First note that for each node with degree \( k \), there are \( \frac{k(k-1)}{2} \) triads centered at that node. Summing this quantity over all nodes, we have \( \sum_i \frac{k_i(k_i-1)}{2} = \frac{N}{2} (\mu_2 - \mu_1) \). On the other hand, the total number of triads (closed plus open) is estimated by \( \hat{\Lambda} + 3\hat{\Delta} \). Equating these two and solving for \( \mu_2 \), we obtain the following estimator for the second moment of the degrees:

\[
\hat{\mu}_2 = \hat{\mu}_1 + \frac{2}{N} \left[ \frac{\hat{\Lambda} \hat{P}(\Delta \rightarrow \Delta) + 3\hat{\Delta} \hat{P}(\Lambda \rightarrow \Lambda) - \hat{\Delta} \hat{P}(\Delta \rightarrow \Lambda)}{\hat{P}(\Delta \rightarrow \Delta) \hat{P}(\Lambda \rightarrow \Lambda)} \right].
\] (10)

The second moment of the degrees can be utilized to approximate the epidemic thresholds for the SIS and SIR models under the heterogeneous mean-field approximation (Pastor-Satorras et al., 2015), we have \( \lambda_{cr}^{SIS} = \frac{\mu_1}{\mu_2} \) and \( \lambda_{cr}^{SIR} = \frac{\mu_1}{\mu_2 - \mu_1} \). We point out that this result is approximate. The expressions we use for the epidemic threshold are obtained under the approximation that degree correlations are negligible, and they have the advantage of only using the moments of the degree distribution. More accurate results exist in the literature, but they use additional information about degree correlations or the adjacency matrix (Pastor-Satorras et al., 2015). Another quantity in studying epidemics on networks is the Basic Reproduction number (denoted by \( R_0 \)). It is defined as the expected number of susceptible individuals that a single infected individual—in a fully-susceptible population—transmits the disease to. Under the same approximation as above, for the basic reproduction number we have \( R_0^{SIS} = \frac{\beta}{\gamma} \frac{\mu_2}{\mu_1} \) and \( R_0^{SIR} = \frac{\beta}{\beta + \gamma} \left( \frac{\mu_2}{\mu_1} - 1 \right) \), where \( \beta \) is the transmission rate and \( \gamma \) is the recovery rate. Note that \( R_0 \) can be obtained once thresholds are given, so it suffices to analyze one of these values. Here we consider the thresholds. We can estimate \( \lambda_{cr}^{SIS} \) and \( \lambda_{cr}^{SIR} \) by inserting \( \hat{\mu}_1 \) and \( \hat{\mu}_2 \) into the above expressions. Figure 6 presents the comparison between the estimated value and the crude version of the epidemic thresholds of several networks. The accuracy of the presented estimators is readily observable, but perhaps more importantly, it can be seen that the crude version of the epidemic threshold consistently
overestimates the true value both for SIS and SIR dynamics (equivalently, we can say that they consistently underestimate the basic reproduction number (Lindquist et al., 2011). This means that networks might be more susceptible to potential epidemic outbreaks than the crude data suggest.

We close by ascertaining that the estimators yield reasonable values for data sets that have employed fixed-choice sampling. We apply our method to the only fixed-choice offline social network data set that we could acquire, namely, the Add Health data (Moody, 2001). The cap value for this data set is $c = 5$. Table II presents the estimated values, which are reasonable.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>$\hat{\mu}_1$</th>
<th>$\hat{\mu}_2$</th>
<th>$\hat{C}$</th>
<th>$\hat{\rho}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated value</td>
<td>24</td>
<td>814</td>
<td>0.35</td>
<td>0.009</td>
</tr>
</tbody>
</table>

**TABLE II:** Estimated values for the Add Health data set

It is of note that the computational complexity of the posited estimators is equal to the complexity required to count the triangles (the other parts have lower complexity). The triangles can be counted in $O((\ell_{ee} + \ell_{ea})^{1.5})$ (Alon et al., 1997). Hence the computational complexity is remarkably feasible. Note that triangles are being counted in the sampled network, which is small in practice (no existing offline social network data set has more than $O(10^4)$ nodes), and for such sizes, estimation takes a couple of seconds on a regular laptop.

V. CONCLUSION, LIMITATIONS, AND FUTURE EXTENSIONS

Focusing on fixed-choice sampling as a widely-used method in social network studies, we provided a mathematical model for the sampling procedure and obtained analytical solutions to estimate several standard network statistics, and corroborated the results with simulations on synthetic and empirical networks. We highlighted the disparity between the implications of crude and inferred values regarding epidemic processes. An illuminating line of future work would be to investigate the effects of truncation, or other properties of the sampling setups in social network studies, on the estimates of contagion and diffusion processes on networks. It would be informative to know if different dynamics (e.g., simple contagions, complex contagions, other threshold-based contagion models, epidemic-type models, etc.) are affected differently by the properties of the sampling procedure.

The presented results highlight the substantial room for methodological improvement in treating data obtained from fixed-choice egocentric social network designs and takes a step towards
providing estimators for the true structural parameters of the underlying social network. There remains several practically-important problems in this direction. Here we highlight a few important ones. Foremost is to accommodate heterogeneity in the strength of social ties. This heterogeneity undeniably influences the recollections (e.g., the chance of different alters being remembered) in the interviews, thereby affecting the gathered data. The first-order correction to the present conceptualization would be to augment it with tie strengths. For this implementation, the simplest way forward would be assuming dichotomous strengths; strong and weak (akin to the Gemeinschaft/Gesellschaft duo). For example, in the context of high school friendship networks, the assumption of tie uniformity is less reasonable, because ties within a class or a school are likely to be stronger than ties across classes or schools, respectively. A more realistic sampling model would incorporate tie strength in the response process, that is, the probability that a tie is mentioned increases with its strength. To incorporate tie strength into the analysis, one would have to inevitably introduce parameters. One can assume a parametric distribution for the tie strengths. See Saramäki et al. (2014) and Sutcliffe et al. (2012) for examples of empirical parametric distributions of tie strengths. This is equivalent to a weighted graph representation. The next step is to incorporate the tie strengths into the recollection process (e.g., the chance of recalling being proportional to the tie strength), and then to infer the parameters—from which the moments of the degree distribution would follow. This approach would also be effective in improving the performance for networks with highly-skewed degree distribution. As discussed in the text, the estimators for the BA network had the most variability. Parametric assumptions, for example for the family the degree distribution belongs to, can help improve the results.

Another line of improvement upon the present results is to focus on the complete degree distribution. In the present paper we estimated the first two moments of the degree distribution. A stronger approach able to recover the full degree distribution would be desirable because the degree distribution is highly informative of various other structural properties of the social network. Another important line of future research is to focus on small-population corrections, which might be particularly important for network studies in small populations, such as high school friendship networks. The proposed estimators in the present paper perform better as population grows (see Section III of the SI for small-network examples). Devising methods tailor-made for small-scale networks is a worthy direction for future research.

An additional source of heterogeneity in social network ties is the types of relations: Individuals possess distinct interlocking roles in society, thus acquiring different ‘types of ties’—for classics
on how these heterogeneities can be utilized to study structural equivalence via role algebra, see Lorrain and White (1971), White et al. (1976), and Boorman and White (1976). If a social network design is to capture a multilayer picture, the present inference scheme must be extended to multiple layers of relations, incorporating the underlying inter-layer correlations into the interviewee recollection process.

Another improvement in the realism of the sampling setup is to incorporate triadic closure tendencies for recollection. Suppose respondent Z has a pair of neighbors X and Y that are connected to one another, and a neighbor W that is not connected to any other neighbor of X. Future research can explore once Z remembers X, how much is the likelihood of remembering Y increases (since they might be closely associated in the memory and recalling one might elevate the chances of recalling the other), and the results can be used to improve the modeling of the recollection procedure. It might also be the case that the probability of an alter being remembered depends on the degree of that alter (e.g., more ‘popular’ friends are more salient in the memory). Future research in this direction can shed light on the recollection process in interviews.

We hope that this work will encourage further development of methodologically-rich data collection and inference methods that would serve towards more accurate studies of social networks in various disciplines.

REFERENCES


cities in central Canada: Selected papers, 152–167.


Fig. 2: The distribution of the ratio of the estimated values to true values of (a) $N$, (b) $\mu_1$, and (c) $\rho$, for networks constructed via the small-world network described in text. The performance is already stable with network sizes as low as $O(10^3)$ (most real social networks are much larger than this). The bias of the estimator is small across different values of $q$, but the variance is sensitive and higher values of $q$ (larger sample proportions) yield smaller variances. In the side plots, the estimators are compared to the crude estimates. The marked difference in the accuracy of the two approaches are visible. For these scatter plots, we have generated many synthetic networks, sweeping across network sizes between 5000 and 10000, as well as sample proportions in the range $0.05 < q < 0.2$. 

$q$ inclusion probability
$N$ network size
Fig. 3: assessing the performance of the density estimator for synthetic and empirical network data.

Fig. 4: Possible ways that triangles in $G$ can be observed as open triads or triangles in $\tilde{G}$ (top panel), and possible ways that open triads in $G$ can be observed as open triads in $\tilde{G}$ (bottom panel). The configurations are accompanied by their probabilities of occurrence. The prefactors shown in tags pertain to the multiplicities due to symmetry. The coefficients used in the probabilities are defined in Equation 7.
Fig. 5: Comparison of the performance of the provided estimator for the clustering coefficient with that of the crude network. We used 300 synthetic networks of different size (uniformly between 2000 and 5000), different values of $q$ (uniform between 0.1 and 0.2), and different values of $c$ (uniform between 5 and 8).

Fig. 6: The comparison of the performance of the presented estimators with those of the crude network. The simulation setup is similar to that of Figure 5. In both cases, the epidemic thresholds are underestimated by the crude estimates. This means that in the crude networks, the average number of secondary infections each new infected individual causes is smaller than the true network. In other words, using the crude estimates underestimates the risk and prevalence of epidemics.