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

Babak Fotouhi^{1,2} and Steven Rytina³

¹ *Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, USA*

² *Institute for Quantitative Social Sciences, Harvard University, Cambridge, MA, USA*

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 (*egos*), 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 ℓ_{ee} and ℓ_{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), μ_1 (average degree), μ_2 (second moment of the degree distribution), ρ (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 , ℓ_{ee} , and ℓ_{ea} , are given for explanatory purposes. These values are used to estimate μ_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.

observable	symbol
number of respondents (egos)	n_e
number of ego-ego links	ℓ_{ee}
number of ego-alter links	ℓ_{ea}
number of open triads in $\tilde{G}$	$\tilde{\Lambda}$
number of triangles in $\tilde{G}$	$\tilde{\Delta}$
inclusion probability	q
population size	N

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

To employ the method of moments (Kolaczyk, 2014; Frank & Snijders, 1994), we first calculate the expected value of n_e , ℓ_{ee} and ℓ_{ea} and construct a system of equations which leads us to estimators for N , q , and μ_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 ℓ_{ee} is given

by

$$\mathbb{E}(\ell_{ee}) = \frac{q^2}{2} \sum_{ij} \left(\frac{c}{k_i} + \frac{c}{k_j} - \frac{c^2}{k_i k_j} \right) = \frac{q^2}{2} (2Nc - c^2 \sum_{ij} \frac{A_{ij}}{k_i k_j}) \simeq \frac{q^2 Nc}{2} (2 - \frac{c}{\mu_1}), \quad (1)$$

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 no 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 ℓ_{ea} , the q^2 factor changes to $q(1 - q)$ (see Equations 9 to 11 in the SI). For the expected values of ℓ_{ea} we have:

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

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:

$$\begin{bmatrix} \hat{N} \\ \hat{q} \\ \hat{\mu}_1 \end{bmatrix} = \begin{bmatrix} \frac{n_e}{1 - \frac{\ell_{ea}}{cn_e}} \\ 1 - \frac{\ell_{ea}}{cn_e} \\ \frac{c}{2} \frac{n_e c - \ell_{ea}}{n_e c - \ell_{ea} - \ell_{ee}} \end{bmatrix} \quad (3)$$

Note that these results also provide information about the minimum sample size. The denominator of the factor in μ_1 evinces requirements that the collected sample must fulfill. The

term $n_e c$ is all the names mentioned by all the respondents. So $n_e c - \ell_{ea}$ is the number of times respondents have mentioned other respondents and $n_e c - \ell_{ea} - \ell_{ee}$ 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:

$$\begin{bmatrix} \hat{\sigma}_{\hat{N}}^2 \\ \hat{\sigma}_{\hat{q}}^2 \end{bmatrix} \simeq \begin{bmatrix} n_e^2 \ell_{ea} \frac{4\ell_{ee}\ell_{ea} + (1+c)(cn_e - 2\ell_{ea})^2}{(cn_e - \ell_{ea})^4} \\ \ell_{ea} \frac{4\ell_{ee}\ell_{ea} + c\ell_{ea}^2 + (cn_e - 2\ell_{ea})^2}{c^4 n_e^4} \end{bmatrix}, \quad (4)$$

and for the average degree, we have:

$$\begin{aligned} \hat{\sigma}_{\hat{\mu}_1}^2 \simeq \frac{1}{4n_e^2(cn_e - \ell_{ea} - \ell_{ee})^4} & \left[c^4 n_e^4 - 2c^3 n_e^3(2\ell_{ee} + \ell_{ea}) + 8\ell_{ee}\ell_{ea}^2(\ell_{ee} + \ell_{ea}) \right. \\ & \left. + c^2 n_e^2[4\ell_{ee}^2 + (c+13)\ell_{ee}\ell_{ea} + \ell_{ea}^2 - 8cn_e\ell_{ea}\ell_{ee}(\ell_{ee} + 2\ell_{ea})] \right]. \end{aligned} \quad (5)$$

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 $\frac{\hat{\mu}_1}{N-1}$, which can be expressed as follows

$$\hat{\rho} = \frac{c}{2} \frac{(cn_e - \ell_{ea})^2}{(cn_e - \ell_{ee} - \ell_{ea})[cn_e(n_e - 1) + \ell_{ea}]}. \quad (6)$$

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 equalizes

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-Rényi (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 (Barabasi & 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 $q^3 \frac{c^3(c-1)^3}{\mu_1^3(\mu_1-1)^3}$, $q^3 \frac{c^3(\mu_1-c)^3}{\mu_1^3(\mu_1-1)^3}$, 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 $q(1-q)^2 \frac{c(c-1)}{\mu_1(\mu_1-1)}$ 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:

$$\left\{ \begin{array}{l} a_1 \stackrel{\text{def}}{=} \frac{\binom{\hat{\mu}_1-1}{c-1}}{\binom{\hat{\mu}_1}{c}} = \frac{c}{\hat{\mu}_1} \\ a_0 \stackrel{\text{def}}{=} \frac{\binom{\hat{\mu}_1-1}{c}}{\binom{\hat{\mu}_1}{c}} = \frac{(\hat{\mu}_1 - c)}{\hat{\mu}_1}, \end{array} \right. \quad \left\{ \begin{array}{l} b_2 \stackrel{\text{def}}{=} \frac{\binom{\hat{\mu}_1-2}{c-2}}{\binom{\hat{\mu}_1}{c}} = \frac{c(c-1)}{\hat{\mu}_1(\hat{\mu}_1-1)} \\ b_1 \stackrel{\text{def}}{=} \frac{\binom{\hat{\mu}_1-2}{c-1}}{\binom{\hat{\mu}_1}{c}} = \frac{c(\hat{\mu}_1-c)}{\hat{\mu}_1(\hat{\mu}_1-1)} \\ b_0 \stackrel{\text{def}}{=} \frac{\binom{\hat{\mu}_1-2}{c}}{\binom{\hat{\mu}_1}{c}} = \frac{(\hat{\mu}_1-c)(\hat{\mu}_1-c-1)}{\hat{\mu}_1(\hat{\mu}_1-1)}. \end{array} \right. \quad (7)$$

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 $q(1-q)^2 \frac{c(c-1)}{\mu_1(\mu_1-1)}$, $q^2(1-q) \frac{c^2}{\mu_1^2}$, and $q^3 \frac{c^2(\mu_1-c)(c-1)}{\mu_1^3(\mu_1-1)}$, 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 Λ and the observed number of open triads in $\tilde{G}$ by $\tilde{\Lambda}$, the number of triangles in G by Δ 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 $\tilde{\Delta} = \Delta P(\Delta \rightarrow \Delta)$, and the moment condition for the number of open triads is $\tilde{\Lambda} = \Delta P(\Delta \rightarrow \Lambda) + \Lambda P(\Lambda \rightarrow \Lambda)$. Solving these equations for Δ, Λ yields the estimators for the number of triangles and open triads:

$$\begin{cases} \hat{\Delta} = \frac{\tilde{\Delta}}{P(\Delta \rightarrow \Delta)} \\ \hat{\Lambda} = \frac{\tilde{\Lambda}P(\Delta \rightarrow \Delta) - \tilde{\Delta}P(\Delta \rightarrow \Lambda)}{P(\Delta \rightarrow \Delta)P(\Lambda \rightarrow \Lambda)} \end{cases} \quad (8)$$

The clustering coefficient is defined as $C = \frac{3\Delta}{\Lambda+3\Delta}$. Plugging in the estimated values for Δ and Λ , we obtain

$$\hat{C} = \frac{3\tilde{\Delta}P(\Lambda \rightarrow \Lambda)}{\tilde{\Lambda}P(\Delta \rightarrow \Delta) - \tilde{\Delta}P(\Delta \rightarrow \Lambda) + 3\tilde{\Delta}P(\Lambda \rightarrow \Lambda)} \quad (9)$$

Note that this is different from the crude clustering coefficient that would be obtained from $\tilde{G}$, which would be equal to $\frac{3\tilde{\Delta}}{3\tilde{\Delta}+\tilde{\Lambda}}$. 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 μ_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, μ_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 μ_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 μ_2 , we obtain the following estimator for the second moment of the degrees:

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

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 $\mathcal{R}_0$). 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 $\mathcal{R}_0^{SIS} = \frac{\beta}{\gamma} \frac{\mu_2}{\mu_1}$ and $\mathcal{R}_0^{SIR} = \frac{\beta}{\beta + \gamma} (\frac{\mu_2}{\mu_1} - 1)$, where β is the transmission rate and γ is the recovery rate. Note that $\mathcal{R}_0$ can be obtained once thresholds are given, so it suffices to analyze one of these values. Here we consider the thresholds. We can estimate λ_{cr}^{SIS} and λ_{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.

Quantity	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{C}$	$\hat{\rho}$
Estimated value	24	814	0.35	0.009

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

- Achlioptas, D., Clauset, A., Kempe, D., & Moore, C. (2005). On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. In *Proceedings of the thirty-seventh annual acm symposium on theory of computing* (pp. 694–703).
- Alexander Jr, C. N., & Campbell, E. Q. (1964). Peer influences on adolescent educational aspirations and attainments. *American Sociological Review*, 568–575.
- Alon, N., Yuster, R., & Zwick, U. (1997). Finding and counting given length cycles. *Algorithmica*, 17(3), 209–223.
- Barabasi, A., & Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439), 509–512.
- Boorman, S. A., & White, H. C. (1976). Social structure from multiple networks. ii. role structures. *American journal of sociology*, 1384–1446.

- Burns, N., Kinder, D. R., Rosenstone, S. J., & Sapiro, V. (2001). the national election studies. american national election studies 2000: Pre-and post-election survey [computer file]. icpsr version. *Ann Arbor, MI: University of Michigan, Center for Political Studies.*
- Burt, R. S. (1984). Network items and the general social survey. *Social networks*, 6(4), 293–339.
- Centola, D. (2010). The spread of behavior in an online social network experiment. *science*, 329(5996), 1194–1197.
- Coleman, J. S., Katz, E., Menzel, H., et al. (1966). *Medical innovation: A diffusion study.* Bobbs-Merrill Indianapolis.
- Cornwell, B., Schumm, L. P., Laumann, E. O., & Graber, J. (2009). Social networks in the nshap study: rationale, measurement, and preliminary findings. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 64(suppl 1), i47–i55.
- Dempsey, K., Duraisamy, K., Bhowmick, S. S., & Ali, H. (2012). The development of parallel adaptive sampling algorithms for analyzing biological networks. In *Parallel and distributed processing symposium workshops & phd forum (ipdpsw), 2012 ieee 26th international* (pp. 725–734).
- Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 17–61.
- Fischer, C. S. (1977). *Networks and places: Social relations in the urban setting.* Free Press.
- Fischer, C. S. (1982). *To dwell among friends: Personal networks in town and city.* University of chicago Press.
- Frank, O., & Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, 10, 53–53.
- Gile, K. J., Beaudry, I. S., Handcock, M. S., & Ott, M. Q. (2018). Methods for inference from respondent-driven sampling data. *Annual Review of Statistics and Its Application*, 5(1).
- Goel, S., & Salganik, M. J. (2010). Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences*, 107(15), 6743–6747.
- Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., & Arenas, A. (2003). Self-similar community structure in a network of human interactions. *Physical review E*, 68(6), 065103.
- Harling, G., & Onnela, J.-P. (2016). Impact of degree truncation on the spread of a contagious process on networks. *arXiv preprint arXiv:1602.03434*.
- Harris, K., Halpern, E., C.T.and Whitsel, Hussey, J., Tabor, J., Entzel, P., & Udry, J. (2009). The national longitudinal study of adolescent health: Research design [Computer software man-

- ual]. (<http://www.cpc.unc.edu/projects/addhealth/design> [accessed 19-Feb.-2016])
- Holland, P. W., & Leinhardt, S. (1973). The structural implications of measurement error in sociometry. *Journal of Mathematical Sociology*, 3(1), 85–111.
- Holme, P., & Kim, B. J. (2002). Growing scale-free networks with tunable clustering. *Physical review E*, 65(2), 026107.
- Jordano, P. (2015). Sampling networks of ecological interactions. *bioRxiv*, 025734.
- Klemm, K., & Eguiluz, V. M. (2002). Growing scale-free networks with small-world behavior. *Physical Review E*, 65(5), 057102.
- Kolaczyk, E. D. (2014). *Statistical analysis of network data (volume 69 of springer series in statistics)* (Vol. 65). Springer-Verlag New York.
- Koskinen, J. H., Robins, G. L., Wang, P., & Pattison, P. E. (2013). Bayesian analysis for partially observed network data, missing ties, attributes and actors. *Social Networks*, 35(4), 514–527.
- Kossinets, G. (2006). Effects of missing data in social networks. *Social networks*, 28(3), 247–268.
- Laumann, E. O. (1973). *Bonds of pluralism: The form and substance of urban social networks*. Wiley-Interscience.
- Leskovec, J., & Faloutsos, C. (2006). Sampling from large graphs. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 631–636).
- Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010a). Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on world wide web* (pp. 641–650).
- Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010b). Signed networks in social media. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1361–1370).
- Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh acm sigkdd international conference on knowledge discovery in data mining* (pp. 177–187).
- Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 2.

- Lindquist, J., Ma, J., Van den Driessche, P., & Willeboordse, F. H. (2011). Effective degree network disease models. *Journal of mathematical biology*, 62(2), 143–164.
- Lorrain, F., & White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1), 49–80.
- Marsden, P. V. (2003). Interviewer effects in measuring network size using a single name generator. *Social Networks*, 25(1), 1–16.
- Marsden, P. V. (2011). Survey methods for network data. *The SAGE handbook of social network analysis*, 370–388.
- McPherson, M., Smith-Lovin, L., & Brashears, M. E. (2006). Social isolation in america: Changes in core discussion networks over two decades. *American sociological review*, 71(3), 353–375.
- Moody, J. (2001). Peer influence groups: identifying dense clusters in large networks. *Social Networks*, 23(4), 261–283.
- Newman, M. E. (2003). Ego-centered networks and the ripple effect. *Social Networks*(1), 83–95.
- Newman, M. E. (2010). *Networks: an introduction*. Oxford, UK: Oxford University Press.
- Newman, M. E., & Watts, D. J. (1999). Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4), 341–346.
- Ott, M. Q., Harrison, M. T., Gile, K. J., Barnett, N. P., & Hogan, J. W. (2017). Fixed choice design and augmented fixed choice design for network data with missing observations. *Biostatistics*.
- Omalley, A. J., Arbesman, S., Steiger, D. M., Fowler, J. H., & Christakis, N. A. (2012). Egocentric social network structure, health, and pro-social behaviors in a national panel study of americans. *PLoS One*, 7(5), e36250.
- Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of modern physics*, 87(3), 925.
- Penrose, M. (2003). *Random geometric graphs* (No. 5). Oxford University Press.
- Perkins, J. M., Subramanian, S., & Christakis, N. A. (2015). Social networks and health: A systematic review of sociocentric network studies in low-and middle-income countries. *Social Science and Medicine*, 125, 60–78.
- Prell, C. (2012). *Social network analysis: History, theory and methodology*. Thousand Oaks, CA: Sage.

- Pustejovsky, J. E., & Spillane, J. P. (2009). Question-order effects in social network name generators. *Social networks*, 31(4), 221–229.
- Ripeanu, M., & Foster, I. (2002). Mapping the gnutella network: Macroscopic properties of large-scale peer-to-peer systems. In *international workshop on peer-to-peer systems* (pp. 85–93).
- Robins, G. (2015). *Doing social network research: Network-based research design for social scientists*. London, UK: Sage.
- Robins, G., Pattison, P., & Woolcock, J. (2004). Missing data in networks: exponential random graph (p) models for networks with non-respondents. *Social networks*, 26(3), 257–283.
- Saramäki, J., Leicht, E., López, E., Roberts, S. G., Reed-Tsochas, F., & Dunbar, R. I. (2014). Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3), 942–947.
- Shrum, W., & Cheek Jr, N. H. (1987). Social structure during the school years: Onset of the degrouping process. *American Sociological Review*, 218–223.
- Smith, J. A. (2012). Macrostructure from microstructure: Generating whole systems from ego networks. *Sociological methodology*, 42(1), 155–205.
- Smith, J. A. (2015). Global network inference from ego network samples: testing a simulation approach. *The Journal of Mathematical Sociology*, 39(2), 125–162.
- Stivala, A. D., Koskinen, J. H., Rolls, D. A., Wang, P., & Robins, G. L. (2016). Snowball sampling for estimating exponential random graph models for large networks. *Social Networks*, 47, 167–188.
- Sutcliffe, A., Dunbar, R., Binder, J., & Arrow, H. (2012). Relationships and the social brain: integrating psychological and evolutionary perspectives. *British journal of psychology*, 103(2), 149–168.
- Van Tilburg, T. (1998). Interviewer effects in the measurement of personal network size a nonexperimental study. *Sociological Methods and Research*, 26(3), 300–328.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge university press.
- Wellman, B. (1979). The community question: The intimate networks of east yorkers. *American journal of Sociology*, 1201–1231.
- Wellman, B., Craven, P., Whitaker, M., Stevens, H., Shorter, A., DuToit, S., & Bakker, H. (1973). Community ties and support systems: From intimacy to support. *The form of*

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

White, H. C., Boorman, S. A., & Breiger, R. L. (1976). Social structure from multiple networks.

i. blockmodels of roles and positions. *American journal of sociology*, 730–780.

Yang, J., & Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1), 181–213.

Fig. 2: The distribution of the ratio of the estimated values to true values of (a) N , (b) μ_1 , and (c) ρ , 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 $0.05 < q < 0.2$.

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.

Supplementary Information

CONTENTS

I	Notation and Sampling Scheme	2
II	Network Inference	2
II-A	Moment Estimators for q , N and average degree	2
II-B	Network Density	9
II-C	Second moment of the degree distribution: A Rudimentary Approach	10
II-D	Number of Triangles and Triads	12
II-E	Second moment of the degree distribution	17
III	Discussion About Settings with Small Network Size	18
IV	Details of Simulations for Synthetic Network Models and Network Data Sets	18
V	Studies that Use Degree Caps in the Name Generator	19
	References	21

Abstract

Here we present details on the calculation of estimators as well as calculations for their variances. We then provide a list of example studies who adopt fixed-choice sampling designs.

I. NOTATION AND SAMPLING SCHEME

Every network in this paper is undirected. There is an underlying network that we select the respondents from, and there is an *observed* network which is the final result of the study, constructed from the responses. The underlying network is denoted by G , with set of nodes V and set of links E . The quantities and distributions pertaining to the observed graph will be marked by asterisks; the observed graph is $\tilde{G}$, whose set of nodes and links are V^* and E^* , respectively. The degree distribution of the underlying network is $p(k)$, that is, $p(k)$ is the fraction of the nodes in the population who have degree k . Similarly, $p^*(k)$ denotes the degree distribution of the observed network. The size of the population is denoted by N , and the number of links in the network is denoted by L . The first and second moments of the degree distribution are denoted by μ_1 and μ_2 , respectively. The first moment of the degree distribution is simply the average degree of the network.

An initial subset of the population is first selected as respondents. We assume the selection process is a simple Bernoulli selection, in which each node is selected with probability q , independent of every other node. The set of selected respondents is denoted by V_0 . Each respondent is asked to claim c alters. There will be $G \times |V_0|$ alters mentioned overall, but the number of unique alters might differ. Some alters might have been mentioned more than once. Also note that some alters might themselves be in V_0 . We denote by V_1 the set of mentioned alters who are not themselves in V_0 . Consequently, in the constructed networks $\tilde{G}$, there cannot be a link between nodes in V_1 (by definition), but there can be links between nodes that belong to V_0 .

The number of nodes in V_0 is denoted by n_e , and the number of nodes in V_1 is denoted by n_1 . The number of links that connect a node from V_0 to another node in V_0 is denoted by ℓ_{ee} . The number of links who connect a node from V_0 to a node in V_1 is denoted by ℓ_{ea} .

II. NETWORK INFERENCE

In this section we present moment estimators for several network measures. For some measures, we present more than one estimator. The performance of the estimators are investigated via Monte Carlo simulations.

A. Moment Estimators for q , N and average degree

We begin by taking the expected value of ℓ_{ee} , the number of links between nodes in V_0 . Let us denote the element ij of the adjacency matrix of G by A_{ij} , which is 0 if i and j are not connected, and is 1 otherwise. Let α_{ij} denote the conditional probability of i mentioning j as an alter. That is, if node i is in V_0 , then it would

mention j among its c alters with probability α_{ij} . Also, let I_{ij} denote the indicator function which is 1 if i has mentioned j (conditional upon i having been interviewed), and 0 otherwise. This means that I_{ij} is one with probability α_{ij} , and is 0 with probability $1 - \alpha_{ij}$. Finally, let us denote the probability that node i is selected into V_0 by s_i . If a link exists in G between nodes i and j , it will be observed in $\tilde{G}$ if both nodes are selected into V_0 and at least one of them mentions the other one. The expected value of ℓ_{ee} is given by

$$\mathbb{E}(\ell_{ee}) = \frac{1}{2} \sum_{ij} (s_i s_j I_{ij} + s_i s_j I_{ji} - s_i s_j I_{ij} I_{ji}) A_{ij}. \quad (1)$$

The factor $\frac{1}{2}$ is used due to the double counting of links in the sum. Also note that we have assumed the independence of the two events that either of the nodes mentions the other one. This is an approximation and might not hold in certain practical scenarios.

For a simple Bernouli selection process, we have

$$\mathbb{E}(\ell_{ee}) = \frac{q^2}{2} \sum_{ij} (\alpha_{ij} + \alpha_{ji} - \alpha_{ij} \alpha_{ji}) A_{ij}. \quad (2)$$

Each respondent is assumed to select c neighbors uniformly at random. Note that this is an approximation. As discussed in the main text, considering a parametric model can allow wights into the network which can be used to model the nomination process more realistically (i.e., one where respondent mentions stronger ties with higher probabilities). See the discussions in the main text.

For a node with degree k , a specific neighbor is selected with probability $\frac{c}{k}$. So we have

$$\mathbb{E}(\ell_{ee}) = \frac{q^2}{2} \sum_{ij} \left(\frac{c}{k_i} + \frac{c}{k_j} - \frac{c}{k_i} \frac{c}{k_j} \right) A_{ij}. \quad (3)$$

First, we need to perform the following summation that appears on the right hand side:

$$\sum_{ij} \frac{1}{k_i} A_{ij} = \sum_i \frac{1}{k_i} \sum_j A_{ij} = \sum_i \frac{1}{k_i} k_i = N. \quad (4)$$

Using this result, we can rewrite (3) as follows:

$$\mathbb{E}(\ell_{ee}) = \frac{q^2}{2} \left[Nc + Nc - c^2 \sum_{ij} \left(\frac{1}{k_i k_j} \right) A_{ij} \right]. \quad (5)$$

Now we need to compute $\sum_{ij} \frac{A_{ij}}{k_i k_j}$. We can alternatively express this sum as follows:

$$\sum_{ij} \frac{A_{ij}}{k_i k_j} = \sum_k [Np(k)] k \sum_{\ell} \frac{p(\ell|k)}{k\ell}. \quad (6)$$

We now use the approximation $p(\ell|k) = \frac{\ell p(\ell)}{\mu_1}$, which performs well in practice and has been invoked in a wide range of analytical problems in network science. Examples include percolation on networks [1], robustness against random and targeted node or link breakdown [2]–[4], diffusion of information on networks [5], [6], epidemics over networks [7]–[9], network synchronization [10], navigability and search in large networks [11]–[13], distribution

of loops and cycles in large networks [14], studying random walks [13], contact processes [15], [16], and opinion dynamics [17], [18]. This approximation is exact for regular networks, as well as for uncorrelated networks such as Erdős-Rényi (ER) networks and networks—with any degree distribution—generated under the configuration model [19], and any model for which the connection probability of two nodes conditional on their degrees k and ℓ is proportional to $k\ell$). We shall also verify, via simulations, the accuracy of the results this approximation yields. Plugging this approximation into (6), we get

$$\sum_{ij} \frac{A_{ij}}{k_i k_j} \simeq \sum_k [Np(k)] k \sum_{\ell} \frac{\ell p(\ell)}{k\ell\mu_1} = \frac{N}{\mu_1} \sum_k p(k) \sum_{\ell} p(\ell) = \frac{N}{\mu_1}. \quad (7)$$

In Figures 1a, 1b, and 1c, we compare the value of $\sum_{ij} \frac{A_{ij}}{k_i k_j}$ with that of $\frac{N}{\mu_1}$ for an ER network of 1000 nodes, for three different cases: the probability of existence of each link equals 0.05, 0.2, and 0.7, respectively. In each case, the ratio of $\frac{N}{\mu_1}$ to $\sum_{ij} \frac{A_{ij}}{k_i k_j}$ is obtained for 1000 Monte Carlo trials individually, and the distribution of this ratio is depicted. It can be observed that in all cases, this ratio is concentrated around unity, with very small deviation. We then plot the distribution of this ratio for the Newman-Watts small world model, where the network is constructed as follows: we take a $2d$ -regular ring of N nodes, in which each node is connected to d nearest-neighbors from each side, and then establish each non-existing link with some fixed probability p . We denote such a network by $\text{SW}(p, d, N)$. Figure 1d pertains to $\text{SW}(0.02, 10, 500)$, Figure 1e pertains to $\text{SW}(0.03, 5, 1000)$, and Figure 1f pertains to $\text{SW}(0.02, 10, 1000)$. Similar to the case of ER networks, we observe high accuracy of the approximation employed in (7). In Figure 1g, we consider Barabási-Albert (BA) networks constructed via the preferential attachment mechanism. The network is constructed sequentially as follows: starting from a small initial graph (whose properties does not affect the structure of the network in the long-time limit), we introduce new nodes one at a time, and each new incoming node selects m ¹ existing nodes to connect to with probabilities proportional to their current degrees. Such a network with size N is denoted by $\text{BA}(m, N)$. Figure 1g pertains to the example case of $\text{BA}(20, 4000)$. Again, the reasonable accuracy of (7) is observed. In Figure 1h, we assess (7) for a random recursive tree (RRT) of 1000 nodes with $m = 5$. An RRT is constructed similar to the BA network, with the distinction that each new incoming node chooses from existing nodes with uniform probabilities, instead of degree-proportional probabilities. Finally, Figure 1i depicts the results for a network grown under the shifted-linear preferential attachment mechanism, which is a combination of BA and RRT: each new incoming node, selects from existing nodes with probabilities proportional to $k + \theta$, where k is the degree of the existing node and θ is a positive constant. We consider the example case of $m = 5$ and $\theta = 20$ for illustrative purposes.

¹In the BA model, m is a constant, but can also be random with any distribution with finite mean and variance; in the long time limit, only the expected value of m will matter

In all the cases discussed above, we observe high accuracy of the approximate result presented in (7), hence we employ it hereinafter. Plugging the result of (7) into (5), we get

$$\mathbb{E}(\ell_{ee}) = \frac{q^2 N c}{2} \left(2 - \frac{c}{\mu_1} \right). \quad (8)$$

Now let us find the expected value of ℓ_{ea} , the number of links that connect a node in V_0 to a node in V_1 . For a given link that belongs to G to be observed in $\tilde{G}$, exactly one of its nodes should be selected into V_0 , and the selected node should mention the other one (hence putting it into V_1). We have

$$\mathbb{E}(\ell_{ea}) = \frac{1}{2} \sum_{ij} \left[s_i(1 - s_j)\alpha_{ij} + s_j(1 - s_i)\alpha_{ji} - 0 \right] A_{ij}. \quad (9)$$

The third term is zero, unlike in (3), because the intersection of the two events are empty: once a node is in V_1 , it cannot mention the other node, since it is—by definition—not interviewed. For the simple Bernouli selection that is considered in this paper, we have

$$\mathbb{E}(\ell_{ea}) = \frac{q(1 - q)}{2} \sum_{ij} \left(\frac{c}{k_i} + \frac{c}{k_j} \right) A_{ij}. \quad (10)$$

Using the result obtained in (4), we have

$$\mathbb{E}(\ell_{ea}) = N c q (1 - q). \quad (11)$$

Finally, let us find the expected value of n_e . Given the simple Bernouli selection considered in this paper, this is straightforward. The result is:

$$\mathbb{E}(n_e) = q N. \quad (12)$$

Equating the expected values of ℓ_{ea} , ℓ_{ee} and n_e with thier respective observed values, we need to solve the following system of equations:

$$\begin{cases} \ell_{ee} = \frac{q^2 N c}{2} \left(2 - \frac{c}{\mu_1} \right) \\ \ell_{ea} = N c q (1 - q) \\ n_e = q N. \end{cases} \quad (13)$$

There are three unknowns: q , N , and μ_1 , to be obtained by solving this system of equations, which involves straightforward algebraic steps. The solutions are:

$$\begin{cases} \hat{N} = \frac{n_e}{1 - \frac{\ell_{ea}}{c n_e}} \\ \hat{q} = 1 - \frac{\ell_{ea}}{c n_e} \\ \hat{\mu}_1 = \frac{c}{2} \frac{n_e c - \ell_{ea}}{n_e c - \ell_{ea} - \ell_{ee}} \end{cases} \quad (14)$$

Fig. 1: The distribution of the ratio of the right hand side to the left hand side of Equation (7) for different graph topologies. An Erdős-Rényi graph of size N with link-existence probability of p is denoted by $ER(p, N)$. A Newman-Watts small-world network of size N which results from taking a $2d$ -regular ring and creating each non-existing link with probability p is denoted by $SW(p, d, N)$. A Barabási-Albert network of size N for which the initial number of links of incoming nodes is m is denoted by $BA(m, N)$. Similarly, $RRT(m, N)$ pertains to a random recursive tree. Finally, a shifted preferential attachment network of size N for which each incoming node has m initial links and for which the probability that an existing node with degree k receives a link is proportional to $k + \theta$ is denoted by $PA(\theta, m, N)$.

Fig. 2: The distribution of the ration of estimated values, as given by Equation (14), to true values of N , q , and μ_1 as a function of N . The hollow markers denote the median of the distribution. The underlying network is $\text{SW}(25, 0.005, N)$ for different values of N . The value of q is 0.05.

In Figure 2a, the distribution of $\frac{\hat{N}}{N}$ is plotted against N , for fixed sampling proportion $q = 0.05$, for a $\text{SW}(25, 0.005, N)$ network. In Figures 2b and 2c, the distributions of $\frac{\hat{q}}{q}$ and $\frac{\hat{\mu}_1}{\mu_1}$ are plotted against N for the same setup, respectively. We expect these ratios to be concentrated around unity, which would evince absence of bias for the estimators. The estimators are unbiased, as is visible from the figures. Diminishing variance evinces the consistency of the estimators.

To estimate the standard errors, we can use the delta method. We first need to calculate the variance of ℓ_{ee} , ℓ_{ea} and n_e . We begin by n_e —because they are the observables, and every estimator is expressed as a function of these observables, and the delta method relates the variance of each estimator to the variance of these observables, which need to be known. Since n_e follows a Binomial distribution with parameters N, q , we readily have

$$\sigma_{n_e}^2 = Nq(1 - q). \quad (15)$$

Inserting the estimated values for N and q into this equation, we arrive at

$$\hat{\sigma}_{n_e}^2 = \frac{\ell_{ea}}{c}. \quad (16)$$

Now let us find the expected value of ℓ_{ea}^2 , which is invoked in the calculation of $\sigma_{\ell_{ea}}^2$. We need to find:

$$\mathbb{E}(\ell_{ea}^2) = \mathbb{E}\left(\left[\frac{1}{2} \sum_{ij} \left(s_i(1 - s_j)I_{ij} + (1 - s_i)s_j I_{ji} - 0\right) A_{ij}\right]^2\right). \quad (17)$$

This sum is nontrivial because the presence of distinct links in $\tilde{G}$ is only independent if the links are not incident on the same node. Instead of finding this sum explicitly, we approximate it as follows. A given link in G which is incident to a node with degree k_1 and a node with degree k_2 will connect a node in V_0 to a node in V_1

with probability $q(1-q)\frac{c}{k_1} + q(1-q)\frac{c}{k_2}$. The expected value of this probability equals $2q(1-q)\frac{c}{\mu_1}$. There are $\frac{N\mu_1}{2}$ links in total, and each of them contributes to ℓ_{ee} with approximate probability $2q(1-q)\frac{c}{\mu_1}$. This yields a binomial distribution. Its mean value is simply $\frac{N\mu_1}{2} \times 2q(1-q)\frac{c}{\mu_1}$, which coincides with (11). The variance is

$$\sigma_{\ell_{ea}}^2 = \frac{N\mu_1}{2} \times 2q(1-q)\frac{c}{\mu_1} \times \left[1 - 2q(1-q)\frac{c}{\mu_1}\right] = Ncq(1-q) \left[1 - 2q(1-q)\frac{c}{\mu_1}\right]. \quad (18)$$

Using our estimated values for N , q and μ_1 , which are given in (14), we arrive at the estimated variance for ℓ_{ea} as follows:

$$\hat{\sigma}_{\ell_{ea}}^2 = \frac{m1}{c^2 n_e^2} \left[(cn_e - 2\ell_{ea})^2 + 4\ell_{ee}\ell_{ea} \right]^2. \quad (19)$$

The last variance we need to estimate is that of ℓ_{ee} , in order to be able to apply the delta method and estimate the variance of $\hat{N}$, $\hat{q}$ and $\hat{\mu}_1$. In principle, one would need to calculate

$$\mathbb{E}(\ell_{ee}^2) = \mathbb{E} \left(\left[\frac{1}{2} \sum_{ij} (s_i s_j I_{ij} + s_i s_j I_{ji} - s_i s_j I_{ij} I_{ji}) A_{ij} \right]^2 \right). \quad (20)$$

Similar to what we had above, this is tedious to compute explicitly. Hence we approximate it. A given link in G which is incident to a node with degree k_1 and a node with degree k_2 will connect a node in V_0 to another node in V_0 with probability $q^2 \frac{c}{k_1} + q^2 \frac{c}{k_2} - q^2 \frac{c^2}{k_1 k_2}$, whose expected value equals $2q^2 \frac{c}{\mu_1} - q^2 \frac{c^2}{\mu_1^2}$. The network has $\frac{N\mu_1}{2}$ links in total to choose from, and each of them contributes to ℓ_{ee} with approximate probability $2q^2 \frac{c}{\mu_1} - q^2 \frac{c^2}{\mu_1^2}$. Similar to above, we have a binomial distribution. Its mean value is simply $\frac{N\mu_1}{2} \times \left[2q^2 \frac{c}{\mu_1} - q^2 \frac{c^2}{\mu_1^2} \right]$, which is identical to what we have in (8). The variance is

$$\sigma_{\ell_{ee}}^2 = \frac{N\mu_1}{2} \times \left[2q^2 \frac{c}{\mu_1} - q^2 \frac{c^2}{\mu_1^2} \right] \times \left[1 - 2q^2 \frac{c}{\mu_1} + q^2 \frac{c^2}{\mu_1^2} \right] = \frac{q^2 Nc}{2} \left(2 - \frac{c}{\mu_1} \right) \left(1 - 2q^2 \frac{c}{\mu_1} + q^2 \frac{c^2}{\mu_1^2} \right). \quad (21)$$

Using the estimated values obtained in (14), we arrive at the following estimate for the variance of ℓ_{ee} :

$$\hat{\sigma}_{\ell_{ee}}^2 = \frac{m0}{c^2 n_e^2} \left[(cn_e - 2\ell_{ee})^2 + 4\ell_{ee}\ell_{ea} \right]^2. \quad (22)$$

For the variance of $\hat{N}$, we have:

$$\begin{aligned} \hat{\sigma}_{\hat{N}}^2 &\simeq \left(\frac{\partial \hat{N}}{\partial n_e} \right)^2 \sigma_{n_e}^2 + \left(\frac{\partial \hat{N}}{\partial \ell_{ea}} \right)^2 \sigma_{\ell_{ea}}^2 = \left[1 - \frac{m1^2}{(cn_e - \ell_{ea})^2} \right]^2 \sigma_{n_e}^2 + \left[\frac{cn_e^2}{(cn_e - \ell_{ea})^2} \right]^2 \sigma_{\ell_{ea}}^2 \\ &= \left[1 - \frac{m1^2}{(cn_e - \ell_{ea})^2} \right]^2 \frac{\ell_{ea}}{c} + \left[\frac{cn_e^2}{(cn_e - \ell_{ea})^2} \right]^2 \frac{m1}{c^2 n_e^2} \left[(cn_e - 2\ell_{ea})^2 + 4\ell_{ee}\ell_{ea} \right]^2. \end{aligned} \quad (23)$$

Simplifying (23), we arrive at

$$\hat{\sigma}_{\hat{N}}^2 \simeq n_e^2 \ell_{ea} \frac{4\ell_{ee}\ell_{ea} + (1+c)(cn_e - 2\ell_{ea})^2}{(cn_e - m1)^4}. \quad (24)$$

For the variance of $\hat{q}$, we have

$$\begin{aligned}\hat{\sigma}_{\hat{q}}^2 &\simeq \left(\frac{\partial \hat{q}}{\partial n_e}\right)^2 \sigma_{n_e}^2 + \left(\frac{\partial \hat{q}}{\partial \ell_{ea}}\right)^2 \sigma_{\ell_{ea}}^2 = \left(\frac{\ell_{ea}}{cn_e^2}\right)^2 \sigma_{n_e}^2 + \left(\frac{1}{cn_e}\right)^2 \sigma_{\ell_{ea}}^2 \\ &= \left(\frac{\ell_{ea}}{cn_e^2}\right)^2 \frac{\ell_{ea}}{c} + \left(\frac{1}{cn_e}\right)^2 \frac{m1}{c^2 n_e^2} \left[(cn_e - 2\ell_{ea})^2 + 4\ell_{ee}\ell_{ea}\right]^2.\end{aligned}\quad (25)$$

After some simplifying algebraic steps, we obtain

$$\hat{\sigma}_{\hat{q}}^2 \simeq \ell_{ea} \frac{4\ell_{ee}\ell_{ea} + c\ell_{ea}^2 + (cn_e - 2\ell_{ea})^2}{c^4 n_e^4}.\quad (26)$$

Finally, for the variance of $\widehat{\mu}_1$, we have

$$\begin{aligned}\hat{\sigma}_{\widehat{\mu}_1}^2 &\simeq \left(\frac{\partial \widehat{\mu}_1}{\partial n_e}\right)^2 \sigma_{n_e}^2 + \left(\frac{\partial \widehat{\mu}_1}{\partial \ell_{ea}}\right)^2 \sigma_{\ell_{ea}}^2 + \left(\frac{\partial \widehat{\mu}_1}{\partial \ell_{ee}}\right)^2 \sigma_{\ell_{ee}}^2 \\ &= \left[\frac{c^2 n_e}{2(cn_e - \ell_{ea} - \ell_{ee})^2}\right]^2 \sigma_{n_e}^2 + \left[\frac{c\ell_{ee}}{2(cn_e - \ell_{ea} - \ell_{ee})^2}\right]^2 \sigma_{\ell_{ea}}^2 + \left[\frac{c(cn_e - \ell_{ea})}{2(cn_e - \ell_{ea} - \ell_{ee})^2}\right]^2 \sigma_{\ell_{ee}}^2 \\ &= \left[\frac{c^2 n_e}{2(cn_e - \ell_{ea} - \ell_{ee})^2}\right]^2 \frac{\ell_{ea}}{c} \\ &\quad + \left[\frac{c\ell_{ee}}{2(cn_e - \ell_{ea} - \ell_{ee})^2}\right]^2 \frac{m1}{c^2 n_e^2} \left[(cn_e - 2\ell_{ea})^2 + 4\ell_{ee}\ell_{ea}\right]^2 \\ &\quad + \left[\frac{c(cn_e - \ell_{ea})}{2(cn_e - \ell_{ea} - \ell_{ee})^2}\right]^2 \frac{m0}{c^2 n_e^2} \left[(cn_e - 2\ell_{ee})^2 + 4\ell_{ee}\ell_{ea}\right]^2.\end{aligned}\quad (27)$$

This can be simplified and expressed as follows:

$$\hat{\sigma}_{\widehat{\mu}_1}^2 \simeq \frac{c^4 n_e^4 - 2c^3 n_e^3 (2\ell_{ee} + \ell_{ea}) + c^2 n_e^2 [4\ell_{ee}^2 + (c + 13)\ell_{ee}\ell_{ea} + \ell_{ea}^2 - 8cn_e\ell_{ea}\ell_{ee}(\ell_{ee} + 2\ell_{ea}) + 8\ell_{ee}\ell_{ea}^2(\ell_{ee} + \ell_{ea})]}{4n_e^2 (cn_e - \ell_{ea} - \ell_{ee})^4}.\quad (28)$$

B. Network Density

Network density, which is a measure of cohesion, is defined as the ratio of the number of existing ties to the total number of possible ties. The latter is simply $\binom{N}{2}$, so density is given by

$$\rho = \frac{L}{\frac{N(N-1)}{2}} = \frac{2L}{N(N-1)}.\quad (29)$$

The sum of degrees of all nodes in the network equals twice the number of links (this is the case because the network is considered to be undirected, otherwise no factor of 2 would be needed). The sum of degrees of all nodes can be written as $N\mu_1$. So we can express density equivalently as follows

$$\rho = \frac{\mu_1}{N-1}.\quad (30)$$

So the estimated value for network density becomes

$$\hat{\rho} = \frac{\widehat{\mu}_1}{\widehat{N} - 1}.\quad (31)$$

Inserting the estimated values from (14) into (31), we arrive at

$$\hat{\rho} = \frac{c}{2} \frac{(cn_e - m1)^2}{(cn_e - \ell_{ea} - \ell_{ee})[cn_e(n_e - 1) + m1]}. \quad (32)$$

Fig. 3: The distribution of the ratio of estimated values, as given by Equation (32), to true value ρ . The underlying network is $\text{SW}(25, 0.005, N)$ for different values of N . The value of q is 0.05.

In Figure 3, the distribution of $\frac{\hat{\rho}}{\rho}$ is plotted against N , for fixed sampling proportion $q = 0.05$, for a $\text{SW}(25, 0.005, N)$ network.

The variance of this estimator is given by

$$\begin{aligned} \hat{\sigma}_{\hat{\rho}}^2 &\simeq \left(\frac{\partial \hat{\rho}}{\partial n_e} \right)^2 \sigma_{n_e}^2 + \left(\frac{\partial \hat{\rho}}{\partial \ell_{ea}} \right)^2 \sigma_{\ell_{ea}}^2 + \left(\frac{\partial \hat{\rho}}{\partial \ell_{ee}} \right)^2 \sigma_{\ell_{ee}}^2 \\ &= \left[\frac{c^2(cn_e - \ell_{ea})}{2} \frac{n_e(cn_e - 2\ell_{ea})(cn_e - \ell_{ea}) - \ell_{ee}(\ell_{ea} + (2\ell_{ea} - c)n_e)}{(cn_e - \ell_{ee} - \ell_{ea})^2[cn_e(n_e - 1) + \ell_{ea}]^2} \right]^2 \frac{\ell_{ea}}{c} \\ &\quad + \left[\frac{c(cn_e - \ell_{ea})}{2} \frac{\ell_{ee}(\ell_{ea} + cn_e(2n_e - 1)) - cn_e^2(cn_e - \ell_{ea})}{(cn_e - \ell_{ee} - \ell_{ea})^2[cn_e(n_e - 1) + \ell_{ea}]^2} \right]^2 \frac{m1}{c^2 n_e^2} \left[(cn_e - 2\ell_{ea})^2 + 4\ell_{ee}\ell_{ea} \right]^2 \\ &\quad + \left[\frac{c}{2} \frac{(cn_e - \ell_{ea})^2}{(cn_e - \ell_{ee} - \ell_{ea})^2[cn_e(n_e - 1) + \ell_{ea}]} \right]^2 \frac{m0}{c^2 n_e^2} \left[(cn_e - 2\ell_{ee})^2 + 4\ell_{ee}\ell_{ea} \right]^2. \end{aligned} \quad (33)$$

C. Second moment of the degree distribution: A Rudimentary Approach

Now we provide a rough estimate μ_2 , which is defined as $\mu_2 = \frac{1}{N} \sum_x k_x^2 = \sum_k k^2 p(k)$. A better estimators will be obtained through triad census, which is discussed afterwards. The second moment of the degrees is of

theoretical significance in various contexts. Aside from giving an estimate of how dispersed the degree distribution is, μ_2 is invoked in the theoretical analysis of the spread of infectious disease on networks, and helps estimate the epidemic threshold [7]–[9]. It is also invoked in estimating the percolation threshold [1], which is required in the studying of resilience and robustness of networks against random node and link removals [2]–[4], and estimating the largest eigenvalue of the adjacency matrix of undirected networks [20]—which is a quantity of crucial importance for the theoretical analysis of network synchronization [10], consensus dynamics [21], and epidemics [22]. The second moment of the degree distribution is also employed in quantitative analysis of the friendship paradox in social networks [23].

We begin by finding a moment estimate for n_1 , the size of V_1 —which, as mentioned above, are nodes who are mentioned by nodes in V_0 , but are not themselves in V_0 . We will observe that μ_2 , for which we seek an estimator, will be evoked. For node x to contribute to n_1 , two things must happen simultaneously: it should not be selected into V_0 , AND it should have been mentioned by at least one of its neighbors. The former happens with probability $1 - q$. Let us find the probability of the latter event. Consider node y , a neighbor of x , and let us find the probability that y does not mention x . For y not to have mentioned x , one of the following disjoint events must have occurred: either y should not have been selected into V_0 in the first place, or y should have been selected into V_0 AND not have mentioned x as an alter. The former happens with probability $1 - q$. The latter happens with probability $q \frac{\binom{k_y-1}{c}}{\binom{k_y}{c}} = q \frac{k_y - c}{k_y}$. This means that y does not mention x with probability $1 - q + q \frac{k_y - c}{k_y} = 1 - \frac{qc}{k_y}$. The probability that this happens for all neighbors of node x , that is, no neighbor of node x mentions it, is given by $\prod_{y \in N_x} \left(1 - \frac{qc}{k_y}\right)$. The probability that at least one neighbor of x mentions it equals 1 minus the said probability. So the probability that node x contributes to n_1 is given by $(1 - q) \left[1 - \prod_{y \in N_x} \left(1 - \frac{qc}{k_y}\right)\right]$. Since y is a neighbor of x , its expected degree can be approximated as $\frac{\mu_2}{\mu_1}$, as discussed above. This means that the probability that node x contributes to n_1 can be approximated by $(1 - q) \left[1 - \left(1 - \frac{qc\mu_1}{\mu_2}\right)^{k_x}\right]$. So we have

$$\mathbb{E}(n_1) = \sum_x (1 - q) \left[1 - \left(1 - \frac{qc\mu_1}{\mu_2}\right)^{k_x}\right] = \sum_x (1 - q) - (1 - q) \sum_x \left(1 - \frac{qc\mu_1}{\mu_2}\right)^{k_x}. \quad (34)$$

Now note that $qc \frac{\mu_1}{\mu_2} \ll 1$ for almost all realistic settings. Let us Taylor-expand the summand up to the first order of the said quantity. We have:

$$\mathbb{E}(n_1) = N(1 - q) - (1 - q) \sum_x \left[1 - k_x q \frac{c\mu_1}{\mu_2}\right] = N(1 - q) - (1 - q) \left(N - Nqc \frac{\mu_1^2}{\mu_2}\right) = Nq(1 - q)c \frac{\mu_1^2}{\mu_2}. \quad (35)$$

This is straightforward to solve. The result is:

$$\hat{\mu}_2^{(1)} = \frac{Nq(1 - q)c\hat{\mu}_1^2}{\tilde{n}_1}. \quad (36)$$

This estimator is assessed in Figure 4.

Fig. 4: TDistribution of the ratio of $\hat{\mu}_2^{(1)}$ to the true value. The underlying network is $\text{SW}(25, 0.005, N)$. The value of q is 0.05.

D. Number of Triangles and Triads

We intend to find the expected values of the number of triads and the number of triangles that will be observed in $\tilde{G}$. A triad in $\tilde{G}$ can either have been a triad or a triangle in G (that is, for example, three unconnected nodes in G will not suddenly become connected in $\tilde{G}$). Similarly, a triangle that is observed in $\tilde{G}$ must have been a triangle in G as well. In other words, a triad in $\tilde{G}$ comes from either a triangle or a triad of G , and a triad in $\tilde{G}$ can only have come from a triad of G . Consider an open triad (that is, a triad that is not a triangle) which exists in G . Let us call the central node in the triad by x , and the other two nodes by y and z . With certain probabilities, it will be observed as a triad in $\tilde{G}$. There are various ways through which this triad can be observed. For example, x can be selected into V_0 as a respondent and mention both the names of y and z as two of its G alters. Or, both z and y can be selected into V_0 as respondents and can both mention x as their alters. There are 17 different ways that this triad can be observed in $\tilde{G}$, which are presented in Figure 5. The two examples mentioned above pertain to the first and second diagrams in this figure, respectively.

Similarly, a triangle that exists in G can be observed as a triad in $\tilde{G}$ under certain events, and as a triangle under other events. Suppose nodes x , y , and z form a triangle. Then if, for example, only x is selected into V_0 , and x mentions both y and z , then this triangle will be observed as a triad. Conversely, if y and z are selected into V_0 and x is not, then if they both mention x , this triangle will again be observed as a triad. All 17 types of triads can originate from a triangle as well. However, there are many ways that this triangle can be indeed observed as a triangle. The least probable is that all the three nodes are selected into V_0 , and each of them mentions the

other two. The most probable is that two of them are selected into V_0 , they both mention the other node (which is not selected into V_0), and one of them also mentions the other one. There are 36 ways this triangle can be observed, which are depicted in Figure 6.

Fig. 5: Possible ways that open triads in G can be observed as triads in $\tilde{G}$.

The diagrams presented in Figure 5 have different probabilities of occurrence. Let us compute the probabilities. First, we need to compute some auxiliary probabilities that will facilitate notation and enable us avoid tedium. Consider some node x and two of its neighbors, y and z . Suppose that x is selected into V_0 . Let us denote the degree of node x by k_x . Let us denote the probability that x mentions both y and z among its c claimed alters by b_2 , the probability that it mentions only y by b_1 , and the probability that it mentions neither of them by b_0 . In the first case, node x has to choose both y and z , and then $c - 2$ nodes from its $k_x - 2$ neighbors. In the second case, node x has to select y as an alter, and then $c - 1$ alters from $k_x - 2$ neighbors (because z is forbidden). In the third case, node x has to select all of its c alters from $k_x - 2$ neighbors, because y and z are both forbidden. For a typical node, we can approximate its degree with μ_1 , thus we have:

Fig. 6: Possible ways that triangles in G can be observed as triads or triangles in $\tilde{G}$.

TABLE I: Probabilities corresponding to events in Figure 6. Under events 1-17, a triangle in G will be observed as a triad in $\tilde{G}$, and under the rest of the events, it will be observed as a triangle in $\tilde{G}$.

$\pi_1 = q(1-q)^2b_2$	$\pi_2 = q^2(1-q)b_1^2$	$\pi_3 = q^2(1-q)b_1^2$	$\pi_4 = q^2(1-q)b_2b_0$
$\pi_5 = q^2(1-q)b_1b_2$	$\pi_6 = q^2(1-q)b_1^2$	$\pi_7 = q^2(1-q)b_2b_0$	$\pi_8 = q^2(1-q)b_1b_2$
$\pi_9 = q^3b_0b_1^2$	$\pi_{10} = q^3b_1^2b_0$	$\pi_{11} = q^3b_1^3$	$\pi_{12} = q^3b_1^2b_0$
$\pi_{13} = q^3b_2b_0^2$	$\pi_{14} = q^3b_0b_1b_2$	$\pi_{15} = q^3b_1^3$	$\pi_{16} = q^3b_0b_1b_2$
$\pi_{17} = q^3b_2b_1^2$	$\pi_{18} = q^2(1-q)b_2b_1$	$\pi_{19} = q^2(1-q)b_2b_1$	$\pi_{20} = q^2(1-q)b_2^2$
$\pi_{21} = q^2(1-q)b_2b_1$	$\pi_{22} = q^2(1-q)b_2b_1$	$\pi_{23} = q^2(1-q)b_2^2$	$\pi_{24} = q^2(1-q)b_2b_1$
$\pi_{25} = q^2(1-q)b_2b_1$	$\pi_{26} = q^2(1-q)b_2^2$	$\pi_{27} = q^3b_0b_1b_2$	$\pi_{28} = q^3b_0b_1b_2$
$\pi_{29} = q^3b_2^2b_0$	$\pi_{30} = q^3b_1^3$	$\pi_{31} = q^3b_0b_1b_2$	$\pi_{32} = q^3b_2b_1^2$
$\pi_{33} = q^3b_2b_1^2$	$\pi_{34} = q^3b_2b_1^2$	$\pi_{35} = q^3b_2^2b_1$	$\pi_{36} = q^3b_0b_1b_2$
$\pi_{37} = q^3b_1^3$	$\pi_{38} = q^3b_1^2b_2$	$\pi_{39} = q^3b_0b_1b_2$	$\pi_{40} = q^3b_0b_1b_2$
$\pi_{41} = q^3b_2b_1^2$	$\pi_{42} = q^3b_2^2b_0$	$\pi_{43} = q^2b_1^2b_2$	$\pi_{44} = q^3b_2^2b_1$
$\pi_{45} = q^3b_1^2b_2$	$\pi_{46} = q^3b_1^2b_2$	$\pi_{47} = q^3b_2^2b_1$	$\pi_{48} = q^3b_1^2b_2$
$\pi_{49} = q^3b_2^2b_0$	$\pi_{50} = q^3b_2^2b_1$	$\pi_{51} = q^3b_2^2b_1$	$\pi_{52} = q^3b_2^2b_1$
$\pi_{53} = q^3b_2^3$			

$$\left\{ \begin{array}{l} b_2 = \frac{\binom{\hat{\mu}_1 - 2}{c - 2}}{\binom{\hat{\mu}_1}{c}} = \frac{c(c-1)}{\hat{\mu}_1(\hat{\mu}_1 - 1)} \\ b_1 = \frac{\binom{\hat{\mu}_1 - 2}{c - 1}}{\binom{\hat{\mu}_1}{c}} = \frac{c(\hat{\mu}_1 - c)}{\hat{\mu}_1(\hat{\mu}_1 - 1)} \\ b_0 = \frac{\binom{\hat{\mu}_1 - 2}{c}}{\binom{\hat{\mu}_1}{c}} = \frac{(\hat{\mu}_1 - c)(\hat{\mu}_1 - c - 1)}{\hat{\mu}_1(\hat{\mu}_1 - 1)}. \end{array} \right. \quad (37)$$

Note that these are conditional probabilities. That is, the selection into V_0 or V_1 is assumed as given. To calculate probabilities of occurrence of each event depicted in Figure 6, we need to account for selections as well. The event probabilities are as given in Table I. Note that in practice, all q s must be replaced by $\hat{q}$.

We can undertake similar steps for open triads. Consider an open triad, with node x at the center and y and z connected to x . Let us compute the probabilities that correspond to events depicted in Figure 5. The probability

TABLE II: Probabilities corresponding to the events depicted in Figure 5. If any of these events occur, an open triad in G will be observed as a triad in $\tilde{G}$.

$\phi_1 = q(1-q)^2b_2$	$\phi_2 = q^2(1-q)a_1^2$	$\phi_3 = q^2(1-q)a_1b_1$	$\phi_4 = q^2(1-q)b_2a_0$
$\phi_5 = q^2(1-q)a_1b_2$	$\phi_6 = q^2(1-q)a_1b_1$	$\phi_7 = q^2(1-q)b_2a_0$	$\phi_8 = q^2(1-q)a_1b_2$
$\phi_9 = q^3a_0a_1b_1$	$\phi_{10} = q^3a_1^2b_0$	$\phi_{11} = q^3a_1^2b_1$	$\phi_{12} = q^3a_0a_1b_1$
$\phi_{13} = q^3b_2a_0^2$	$\phi_{14} = q^3a_0a_1b_2$	$\phi_{15} = q^3a_1^2b_1$	$\phi_{16} = q^3a_0a_1b_2$
$\phi_{17} = q^3b_2a_1^2$			

that x is selected into V_0 and mentions both, one, or neither of the other two nodes are b_2 , b_1 and b_0 , respectively, which are calculated above. However, for nodes y and z , the calculations differ. Suppose that y is chosen into V_0 . Let us denote the probability that y mentions x by a_1 , and the probability that y does not mention x by a_0 . We have:

$$\left\{ \begin{array}{l} a_1 = \frac{\binom{\hat{\mu}_1 - 1}{c - 1}}{\binom{\hat{\mu}_1}{c}} = \frac{c}{\hat{\mu}_1} \\ a_0 = \frac{\binom{\hat{\mu}_1 - 1}{c}}{\binom{\hat{\mu}_1}{c}} = \frac{(\hat{\mu}_1 - c)}{\hat{\mu}_1}. \end{array} \right. \quad (38)$$

The probabilities that pertain to the events depicted in Figure 5 are given in Table II.

Let us denote the number of triangles in G by Δ and the number of triads by τ . Note that the number of open triads (non-triangles) is $\Lambda = \tau - 3\Delta$. The expected value of the number of open triads in $\tilde{G}$, which we denote by Λ^* , is

$$\mathbb{E}(\Lambda^*) = \Lambda \times \sum_{i=1}^{17} \hat{\phi}_i + \Delta \times \sum_{i=1}^{17} \hat{\pi}_i. \quad (39)$$

Similarly, the expected number of observed triangles is given by

$$\mathbb{E}(\Delta^*) = \Delta \times \sum_{i=18}^{53} \hat{\pi}_i. \quad (40)$$

This readily yields the estimator for the number of triangles:

$$\hat{\Delta} = \frac{\Delta^*}{\sum_{i=18}^{53} \hat{\pi}_i}. \quad (41)$$

Fig. 7: Distribution of estimates of number of triads and triangles as a function of N . The underlying network is $\text{SW}(25, 0.005, N)$. The value of q is fixed at 0.05.

Inserting this into the right hand side of (39), we arrive at the following estimator for the number of open triads:

$$\hat{\Lambda} = \frac{\Lambda^* \sum_{i=18}^{53} \hat{\pi}_i - \Delta^* \sum_{i=1}^{17} \hat{\pi}_i}{\sum_{i=18}^{53} \hat{\pi}_i \sum_{i=1}^{17} \hat{\phi}_i}. \quad (42)$$

Finally, the total number of triads is given by:

$$\hat{\tau} = \hat{\Lambda} + 3\hat{\Delta} = \frac{\Lambda^* \sum_{i=18}^{53} \hat{\pi}_i + \Delta^* \left(3 \sum_{i=1}^{17} \hat{\phi}_i - \sum_{i=1}^{17} \hat{\pi}_i \right)}{\sum_{i=18}^{53} \hat{\pi}_i \sum_{i=1}^{17} \hat{\phi}_i}. \quad (43)$$

The accuracy of this estimate, in addition to (41), can be observed in Figure 7. The underlying network is $\text{SW}(25, 0.005, N)$. The value of q is fixed at 0.05. Finally, these estimators for the number of triads and the number of triangles can be used to find estimators for the clustering coefficient of the network, which is the triad closure ratio:

$$C \stackrel{\text{def}}{=} \frac{3T}{\tau} \implies \hat{C} = \frac{3\hat{T}}{\hat{\tau}}. \quad (44)$$

E. Second moment of the degree distribution

Using these results, we can also find an estimator for μ_2 , the second moment of the degree distribution. A node with degree k is adjacent to $\frac{k(k-1)}{2}$ triads centered at that node. Summing this quantity over all nodes, we

get $\sum_i \frac{k_i(k_i-1)}{2} = \frac{N}{2}(\mu_2 - \mu_1)$. We can also estimate the total number of triads (closed plus open) by $\hat{\tau}$, as given by (43). Equating these two and solving for μ_2 , we obtain the following estimator for the second moment of the degrees:

$$\hat{\mu}_2 = \hat{\mu}_1 + \frac{2\hat{\tau}}{\hat{N}}, \quad (45)$$

as given in the main text.

III. DISCUSSION ABOUT SETTINGS WITH SMALL NETWORK SIZE

There are contexts where the population under study is small, such as high school friendship network studies. To test how the estimators perform for smaller populations, we tested them on networks with size 100 to 500. Figure 8a shows the results for the small world model, and Figure 8b shows the results for the BA model. As expected, the performance is poorer for small sizes, but it is still within reasonable limits.

It is also worth noting that in high school contexts, the approximation used in the present analysis might be less reasonable to employ. As discussed in the future-extensions section of the main text, a more detailed model that incorporated tie heterogeneity might be better in the school context, which distinguishes between links within and between classes, or links within or between schools.

IV. DETAILS OF SIMULATIONS FOR SYNTHETIC NETWORK MODELS AND NETWORK DATA SETS

For each family of networks we used 5000 Monte Carlo trials. Network size was randomly generated between 10K and 15K. For ER networks [24], the probability of tie creation was randomly chosen in a way that the average degree would not exceed 200 and was not below 50 (to keep the average degree in a realistic range). For HK networks [25], the triad-formation probability was randomly chosen between 0 and 1, and initial number of connections were randomly chosen in a way that the resulting network has average degree in the mentioned range. For KE networks [26] too, μ was uniformly chosen in $[0, 1]$ with the said constraint. For FF networks [27], forward and backward probabilities were uniformly chosen in $[0, 1]$ with the said constraint. For BA networks [28], m was chosen uniformly between 25 and 75 (since the average degree is $2m$). For RGG networks [29], we considered the unit square and the threshold Euclidean distance for tie formation was randomly generated in $[0, 0.2]$, and the constraint for average degree was imposed as above.

It must be noted that to assess the proposed estimators, there is a lack of benchmark networks to be used as ground truths. We would need a (large) *fully-observed* offline social network data set, which we can synthetically sample with degree cap and assess the estimators. Such a data set is not common, if not non-existing. Every existing social network data (including the pioneering ones cited above) are already sampled ones. There do exist studies that do not explicitly impose a cutoff on the number of alters in the questionnaire, but we do not know with what degree of certitude we could claim they are *full* social networks. Perhaps for contexts where

Fig. 8: assessing the performance of the estimators for small network sizes

the name generators typically targets few individuals (such as money-borrowing) or asks about intimate relations (such as secret sharing) this can be reasonably assumed. But for weak and/or numerous ties, it is unclear to us. Nevertheless, to have some network data in addition to the synthetic ones for assessing the performance of the estimators, we included a few data sets with social origins. These are networks that are fully observed, and we used them as ground-truth, artificially fixed-choice-sampled them many times, and compared the estimates with true values over the distribution produced over these trials.

Note that the setup of the paper assumes that the type of tie under consideration is such that every node has more than c links. All the real data sets considered in the paper have many nodes with degree 1 and 2. So their nature is different from what the premise of the paper is. In data cleaning, we discarded nodes with degree 1 and acquired networks with minimum degree 2, and set $c = 2$, which is still very small compared to actual values of the cutoff. Much information is lost. Yet, the errors presented in Figure 3.b of the main text are reasonably small. Not pruning the data (retaining degree-1 nodes) and running the tests with $c = 1$ gives slightly poorer results, but the median is still within the 20% margin of the true value and the order of magnitude is captured correctly in almost all trials. The value of q used for synthetic sampling of real data sets is $q = 0.1$, and the distributions are depicted for 1000 trials.

V. STUDIES THAT USE DEGREE CAPS IN THE NAME GENERATOR

Fixed-choice sampling is common in network data collection. In Table III, we present a number of example studies along with their corresponding cutoff values (i.e., the maximum number of alters each respondent is asked to mention to the interviewer):

TABLE III: Examples studies that have employed degree caps

Reference	[30]–[35]	[36]–[40]	[41]–[69]	[70]–[75]	[76]	[77]	[78]–[89]	[90], [91]	[92]
Cutoff value (c)	3	4	5	6	7	8	19	12	20

REFERENCES

- [1] R. Cohen, D. Ben-Avraham, and S. Havlin, "Percolation critical exponents in scale-free networks," *Physical Review E*, vol. 66, no. 3, p. 036113, 2002.
- [2] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin, "Breakdown of the internet under intentional attack," *Physical review letters*, vol. 86, no. 16, p. 3682, 2001.
- [3] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin, "Resilience of the internet to random breakdowns," *Physical review letters*, vol. 85, no. 21, p. 4626, 2000.
- [4] L. K. Gallos, R. Cohen, P. Argyrakis, A. Bunde, and S. Havlin, "Stability and topology of scale-free networks under attack and defense strategies," *Physical Review Letters*, vol. 94, no. 18, p. 188701, 2005.
- [5] Y. Moreno, M. Nekovee, and A. F. Pacheco, "Dynamics of rumor spreading in complex networks," *Physical Review E*, vol. 69, no. 6, p. 066130, 2004.
- [6] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, "Theory of rumour spreading in complex social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 374, no. 1, pp. 457–470, 2007.
- [7] R. Pastor-Satorras and A. Vespignani, "Epidemics and immunization in scale-free networks," *Handbook of graphs and networks: from the genome to the internet*, pp. 111–130, 2005.
- [8] M. Boguná, R. Pastor-Satorras, and A. Vespignani, "Absence of epidemic threshold in scale-free networks with degree correlations," *Physical review letters*, vol. 90, no. 2, p. 028701, 2003.
- [9] Y. Moreno, R. Pastor-Satorras, and A. Vespignani, "Epidemic outbreaks in complex heterogeneous networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 26, no. 4, pp. 521–529, 2002.
- [10] J. G. Restrepo, E. Ott, and B. R. Hunt, "Emergence of coherence in complex networks of heterogeneous dynamical systems," *Physical review letters*, vol. 96, no. 25, p. 254103, 2006.
- [11] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer, "Approximating pagerank from in-degree," in *Algorithms and models for the web-graph*, pp. 59–71, Springer, 2008.
- [12] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer, "On local estimations of pagerank: a mean field approach," *Internet Mathematics*, vol. 4, no. 2-3, pp. 245–266, 2007.
- [13] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks*. Cambridge University Press, 2008.
- [14] J. Noh, "Loop statistics in complex networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 66, no. 2, pp. 251–257, 2008.
- [15] M. Boguñá, C. Castellano, and R. Pastor-Satorras, "Langevin approach for the dynamics of the contact process on annealed scale-free networks," *Physical Review E*, vol. 79, no. 3, p. 036110, 2009.
- [16] S. C. Ferreira, R. S. Ferreira, and R. Pastor-Satorras, "Quasistationary analysis of the contact process on annealed scale-free networks," *Physical Review E*, vol. 83, no. 6, p. 066113, 2011.
- [17] E. Pugliese and C. Castellano, "Heterogeneous pair approximation for voter models on networks," *EPL (Europhysics Letters)*, vol. 88, no. 5, p. 58004, 2009.
- [18] V. Sood and S. Redner, "Voter model on heterogeneous graphs," *Physical review letters*, vol. 94, no. 17, p. 178701, 2005.
- [19] M. Newman, *Networks: an introduction*. Oxford University Press, 2010.
- [20] J. G. Restrepo, E. Ott, and B. R. Hunt, "Approximating the largest eigenvalue of network adjacency matrices," *Physical Review E*, vol. 76, no. 5, p. 056119, 2007.
- [21] R. Olfati-Saber, A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.
- [22] P. Van Mieghem, "Epidemic phase transition of the sis type in networks," *EPL (Europhysics Letters)*, vol. 97, no. 4, p. 48004, 2012.

- [23] Y.-H. Eom and H.-H. Jo, "Generalized friendship paradox in complex networks: The case of scientific collaboration," *Scientific reports*, vol. 4, 2014.
- [24] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, vol. 5, pp. 17–61, 1960.
- [25] P. Holme and B. J. Kim, "Growing scale-free networks with tunable clustering," *Physical review E*, vol. 65, no. 2, p. 026107, 2002.
- [26] K. Klemm and V. M. Eguiluz, "Growing scale-free networks with small-world behavior," *Physical Review E*, vol. 65, no. 5, p. 057102, 2002.
- [27] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 177–187, ACM, 2005.
- [28] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [29] M. Penrose, *Random geometric graphs*. No. 5, Oxford University Press, 2003.
- [30] C. B. Kenny, "Political participation and effects from the social environment," *American Journal of Political Science*, pp. 259–267, 1992.
- [31] S. T. Ennett and K. E. Bauman, "Peer group structure and adolescent cigarette smoking: A social network analysis," *Journal of Health and Social Behavior*, pp. 226–236, 1993.
- [32] M. Schneider, P. Teske, C. Roch, and M. Marschall, "Networks to nowhere: Segregation and stratification in networks of information about schools," *American Journal of Political Science*, pp. 1201–1223, 1997.
- [33] D. A. Reingold, "Social networks and the employment problem of the urban poor," *Urban Studies*, vol. 36, no. 11, pp. 1907–1932, 1999.
- [34] R. E. Rice, L. Donohew, and R. Clayton, "Peer network, sensation seeking, and drug use among junior and senior high school students," *Connections*, vol. 25, no. 2, pp. 32–58, 2003.
- [35] T. W. Valente, P. Gallaher, and M. Mouttapa, "Using social networks to understand and prevent substance use: A transdisciplinary perspective," *Substance use & misuse*, vol. 39, no. 10-12, pp. 1685–1712, 2004.
- [36] L. Nir, "Ambivalent social networks and their consequences for participation," *International Journal of Public Opinion Research*, vol. 17, no. 4, pp. 422–442, 2005.
- [37] C. M. Lakon, S. T. Ennett, and E. C. Norton, "Mechanisms through which drug, sex partner, and friendship network characteristics relate to risky needle use among high risk youth and young adults," *Social Science & Medicine*, vol. 63, no. 9, pp. 2489–2499, 2006.
- [38] H.-P. Kohler, J. R. Behrman, and S. C. Watkins, "Social networks and hiv/aids risk perceptions," *Demography*, vol. 44, no. 1, pp. 1–33, 2007.
- [39] C. Coohy, "Social networks, informal child care, and inadequate supervision by mothers," *Child Welfare*, vol. 86, no. 6, p. 53, 2007.
- [40] A. J. Omalley, S. Arbesman, D. M. Steiger, J. H. Fowler, and N. A. Christakis, "Egocentric social network structure, health, and pro-social behaviors in a national panel study of americans," *PLoS One*, vol. 7, no. 5, p. e36250, 2012.
- [41] L. McCallister and C. S. Fischer, "A procedure for surveying personal networks," *Sociological Methods & Research*, vol. 7, no. 2, pp. 131–148, 1978.
- [42] E. M. Rogers and D. L. Kincaid, "Communication networks: toward a new paradigm for research.," 1981.
- [43] A. V. Horwitz, R. C. Tessler, G. A. Fisher, and G. M. Gamache, "The role of adult siblings in providing social support to the severely mentally ill," *Journal of Marriage and the Family*, pp. 233–241, 1992.
- [44] V. A. Haines, J. S. Hurlbert, and J. J. Beggs, "Exploring the determinants of support provision: Provider characteristics, personal networks, community contexts, and support following life events," *Journal of Health and Social Behavior*, pp. 252–264, 1996.

- [45] J. J. Beggs, V. A. Haines, and J. S. Hurlbert, "Revisiting the rural-urban contrast: Personal networks in nonmetropolitan and metropolitan settings¹," *Rural sociology*, vol. 61, no. 2, pp. 306–325, 1996.
- [46] T. W. Valente, S. C. Watkins, M. N. Jato, A. Van Der Straten, and L.-P. M. Tsitsol, "Social network associations with contraceptive use among cameroonian women in voluntary associations," *Social science and medicine*, vol. 45, no. 5, pp. 677–687, 1997.
- [47] A. Munch, J. M. McPherson, and L. Smith-Lovin, "Gender, children, and social contact: The effects of childrearing for men and women," *American Sociological Review*, pp. 509–520, 1997.
- [48] D. Ruan, "The content of the general social survey discussion networks: an exploration of general social survey discussion name generator in a chinese context," *Social Networks*, vol. 20, no. 3, pp. 247–264, 1998.
- [49] T. W. Valente, J. B. Unger, and C. A. Johnson, "Do popular students smoke? the association between popularity and smoking among middle school students," *Journal of Adolescent Health*, vol. 37, no. 4, pp. 323–329, 2005.
- [50] S. D. McClurg, "The electoral relevance of political talk: Examining disagreement and expertise effects in social networks on political participation," *American Journal of Political Science*, vol. 50, no. 3, pp. 737–754, 2006.
- [51] M. McPherson, L. Smith-Lovin, and M. E. Brashears, "Social isolation in america: Changes in core discussion networks over two decades," *American sociological review*, vol. 71, no. 3, pp. 353–375, 2006.
- [52] M. C. Aalsma, J. D. Fortenberry, M. A. Sayegh, and D. P. Orr, "Family and friend closeness to adolescent sexual partners in relationship to condom use," *Journal of Adolescent Health*, vol. 38, no. 3, pp. 173–178, 2006.
- [53] K. L. Fiori, J. Smith, and T. C. Antonucci, "Social network types among older adults: A multidimensional approach," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 62, no. 6, pp. P322–P330, 2007.
- [54] L. Mercken, M. Candel, P. Willems, and H. De Vries, "Disentangling social selection and social influence effects on adolescent smoking: the importance of reciprocity in friendships," *Addiction*, vol. 102, no. 9, pp. 1483–1492, 2007.
- [55] K.-H. Choi, Z. Ning, S. E. Gregorich, and Q.-c. Pan, "The influence of social and sexual networks in the spread of hiv and syphilis among men who have sex with men in shanghai, china," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 45, no. 1, pp. 77–84, 2007.
- [56] B. Cornwell, E. O. Laumann, and L. P. Schumm, "The social connectedness of older adults: A national profile," *American Sociological Review*, vol. 73, no. 2, pp. 185–203, 2008.
- [57] B. Cornwell, "Network bridging potential in later life life-course experiences and social network position," *Journal of aging and health*, vol. 21, no. 1, pp. 129–154, 2009.
- [58] B. Cornwell, L. P. Schumm, E. O. Laumann, and J. Graber, "Social networks in the nshap study: rationale, measurement, and preliminary findings," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 64, no. suppl 1, pp. i47–i55, 2009.
- [59] K. G. Provan, K. Huang, and H. B. Milward, "The evolution of structural embeddedness and organizational social outcomes in a centrally governed health and human services network," *Journal of Public Administration Research and Theory*, vol. 19, no. 4, pp. 873–893, 2009.
- [60] J. K. Goeree, M. A. McConnell, T. Mitchell, T. Tromp, and L. Yariv, "The 1/d law of giving," *American Economic Journal: Microeconomics*, pp. 183–203, 2010.
- [61] A. I. Balsa, J. F. Homer, M. T. French, and E. C. Norton, "Alcohol use and popularity: Social payoffs from conforming to peers' behavior," *Journal of Research on Adolescence*, vol. 21, no. 3, pp. 559–568, 2011.
- [62] V. A. Haines, J. J. Beggs, and J. S. Hurlbert, "Neighborhood disadvantage, network social capital, and depressive symptoms," *Journal of Health and Social Behavior*, vol. 52, no. 1, pp. 58–73, 2011.
- [63] E. Y. Cornwell and L. J. Waite, "Social network resources and management of hypertension," *Journal of Health and Social Behavior*, vol. 53, no. 2, pp. 215–231, 2012.

- [64] K.-H. Choi, G. Ayala, J. Paul, R. Boylan, and S. E. Gregorich, "Social network characteristics and hiv risk among african american, asian/pacific islander, and latino men who have sex with men," *Journal of acquired immune deficiency syndromes (1999)*, vol. 64, no. 5, 2013.
- [65] C. L. Martin, O. Kornienko, D. R. Schaefer, L. D. Hanish, R. A. Fabes, and P. Goble, "The role of sex of peers and gender-typed activities in young children's peer affiliative networks: A longitudinal analysis of selection and influence," *Child Development*, vol. 84, no. 3, pp. 921–937, 2013.
- [66] J. Bello and M. Rolfe, "Is influence mightier than selection? forging agreement in political discussion networks during a campaign," *Social Networks*, vol. 36, pp. 134–146, 2014.
- [67] M. Hellard, E. McBryde, R. S. Davis, D. A. Rolls, P. Higgs, C. Aitken, A. Thompson, J. Doyle, P. Pattison, and G. Robins, "Hepatitis c transmission and treatment as prevention—the role of the injecting network," *International Journal of Drug Policy*, vol. 26, no. 10, pp. 958–962, 2015.
- [68] M. Forster, T. J. Grigsby, A. Bunyan, J. B. Unger, and T. W. Valente, "The protective role of school friendship ties for substance use and aggressive behaviors among middle school students," *Journal of school health*, vol. 85, no. 2, pp. 82–89, 2015.
- [69] K.-H. Chu, E. Hoepfner, T. Valente, and L. Rohrbach, "A social network analysis of a coalition initiative to prevent underage drinking in los angeles county," in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pp. 4356–4365, IEEE, 2015.
- [70] N. Shulman, "Network analysis: a new addition to an old bag of tricks," *Acta Sociologica*, vol. 19, no. 4, pp. 307–323, 1976.
- [71] R. E. Mitchell, "Social networks and psychiatric clients: The personal and environmental context," *American Journal of Community Psychology*, vol. 10, no. 4, pp. 387–401, 1982.
- [72] L. Michell and A. Amos, "Girls, pecking order and smoking," *Social science & medicine*, vol. 44, no. 12, pp. 1861–1869, 1997.
- [73] E. Macdonald, R. Hayes, and A. Baglioni, "The quantity and quality of the social networks of young people with early psychosis compared with closely matched controls," *Schizophrenia research*, vol. 46, no. 1, pp. 25–30, 2000.
- [74] T. W. Valente, "Social network influences on adolescent substance use: An introduction," *Connections*, vol. 25, no. 2, pp. 11–16, 2003.
- [75] C. Steglich, T. A. Snijders, and M. Pearson, "Dynamic networks and behavior: Separating selection from influence," *Sociological methodology*, vol. 40, no. 1, pp. 329–393, 2010.
- [76] K. Fujimoto and T. W. Valente, "Multiplex congruity: Friendship networks and perceived popularity as correlates of adolescent alcohol use," *Social Science & Medicine*, vol. 125, pp. 173–181, 2015.
- [77] R. Iyengar, C. Van den Bulte, and T. W. Valente, "Opinion leadership and social contagion in new product diffusion," *Marketing Science*, vol. 30, no. 2, pp. 195–212, 2011.
- [78] A. E. Kazak, "Families with disabled children: Stress and social networks in three samples," *Journal of abnormal child psychology*, vol. 15, no. 1, pp. 137–146, 1987.
- [79] J. H. Stevens Jr, "Social support, locus of control, and parenting in three low-income groups of mothers: Black teenagers, black adults, and white adults," *Child Development*, pp. 635–642, 1988.
- [80] K. A. Urberg, S. M. Değirmencioglu, J. M. Tolson, and K. Halliday-Scher, "The structure of adolescent peer networks," *Developmental psychology*, vol. 31, no. 4, p. 540, 1995.
- [81] S. R. Friedman, M. Bolyard, P. Mateu-Gelabert, P. Goltzman, M. P. Pawlowicz, D. Z. Singh, G. Touze, D. Rossi, C. Maslow, M. Sandoval, *et al.*, "Some data-driven reflections on priorities in aids network research," *AIDS and Behavior*, vol. 11, no. 5, pp. 641–651, 2007.
- [82] S. A. Haas, D. R. Schaefer, and O. Kornienko, "Health and the structure of adolescent social networks," *Journal of Health and Social Behavior*, vol. 51, no. 4, pp. 424–439, 2010.

- [83] S. M. Reich, K. Subrahmanyam, and G. Espinoza, "Friending, iming, and hanging out face-to-face: overlap in adolescents' online and offline social networks.," *Developmental psychology*, vol. 48, no. 2, p. 356, 2012.
- [84] D. A. Shoham, L. Tong, P. J. Lamberson, A. H. Auchincloss, J. Zhang, L. Dugas, J. S. Kaufman, R. S. Cooper, and A. Luke, "An actor-based model of social network influence on adolescent body size, screen time, and playing sports," *PloS one*, vol. 7, no. 6, p. e39795, 2012.
- [85] C. L. Apicella, F. W. Marlowe, J. H. Fowler, and N. A. Christakis, "Social networks and cooperation in hunter-gatherers," *Nature*, vol. 481, no. 7382, pp. 497–501, 2012.
- [86] E. Rice, I. W. Holloway, A. Barman-Adhikari, D. Fuentes, C. H. Brown, and L. A. Palinkas, "A mixed methods approach to network data collection," *Field methods*, p. 1525822X13518168, 2014.
- [87] D. R. Schaefer and S. D. Simpkins, "Using social network analysis to clarify the role of obesity in selection of adolescent friends," *American journal of public health*, vol. 104, no. 7, pp. 1223–1229, 2014.
- [88] W. Jacobs, K. C. Jeon, P. Goodson, and T. W. Valente, "Whats love got to do with it? adolescent romantic networks and substance use," *International Journal of Adolescence and Youth*, pp. 1–9, 2015.
- [89] M. Y. Delgado, A. V. Ettekal, S. D. Simpkins, and D. R. Schaefer, "How do my friends matter? examining latino adolescents friendships, school belonging, and academic achievement," *Journal of youth and adolescence*, pp. 1–16, 2015.
- [90] M. Jennes-Coussens, J. Magill-Evans, and C. Koning, "The quality of life of young men with asperger syndrome a brief report," *Autism*, vol. 10, no. 4, pp. 403–414, 2006.
- [91] M. Pearson, C. Sieglic, and T. Snijders, "Homophily and assimilation among sport-active adolescent substance users.," *Connections*, vol. 27, no. 1, pp. 47–63, 2006.
- [92] A. Bowling, M. Farquhar, and P. Browne, "Life satisfaction and associations with social network and support variables in three samples of elderly people," *International journal of geriatric psychiatry*, vol. 6, no. 8, pp. 549–566, 1991.