Estimating properties from snowball sampled networks

Johannes Illenberger* Gunnar Flötteröd†

January 17, 2011

Working Paper 11-01

Abstract

This article addresses the estimation of topological network parameters from data obtained with a snowball sampling design. An approximate expression for the probability of a vertex to be included in the sample is derived. Based on this sampling distribution, estimators for the mean degree, the degree-degree correlation, and the mean vertex clustering coefficient are derived. The performance of these estimators and their sensitivity with respect to the response rate are validated through Monte-Carlo simulations on several networks. The results indicate a good estimation quality of the mean degree and the mean vertex clustering coefficient, and they show reasonable results for the degree-degree correlation.

1 Introduction

The increasing availability of large data sets has enabled great advances in the empirical research on social networks. Electronic databases, such as the Internet Movie Database (www.imdb.com) or the scientific paper database arxiv.org, are available as proxy-data sources. The inferred networks, which can be in the order of up to 10^5 vertices, are usually embedded in an institutional setting or in a specific community, with regard to the above examples: movie actors [1] and authors of scientific papers [2]. Large social networks outside of such settings are rather hard to obtain since appropriate proxy-data is rare and even if existing, privacy regulations make its access nearly impossible. The researcher then needs to turn to the traditional "paper and pencil" survey to directly sample a social network.

A straightforward approach would be to draw random respondents, denoted as *egos*, and ask them about their social contacts, denoted as *alters*. This so-called "ego-centric" network sampling approach [3] produces star-like networks, which provide insights into the relations between egos and alters. Higher topological networks properties, however, remain unrevealed. In principle, it would be possible to draw a sufficiently large sample such that the ego-centric networks become connected. Practically, such an approach would be prohibitively expensive.

The *snowball sampling* approach, also called chain-referral or link-tracing, addresses this issue. In snowball sampling, an initial set of respondents, denoted as *seeds*, is enquired to report their alters. These alters are then invited to participate in the survey and to report their alters in turn. This procedure is repeated for a given number of iterations (also denoted as *waves* or *stages* [4]) or until the desired number of vertices is sampled. Clearly, snowball sampling reveals more complex network structures than the ego-centric approach because it is not constrained to first degree relations.

A drawback of snowball sampling is that it bears several possible sources of bias [5]. Since the recruiting of new respondents is done, or at least influenced, by the respondents themselves, the researcher has only limited control over which individuals are included in the sample. Furthermore, if strong homophily exists between individuals, there is a danger that the snowball is caught in a homogeneous cluster.

Another source of bias is the underlying network topology, which governs the progress of the snowball. Well-connected individuals, i.e., vertices with a large degree, have a higher

*Berlin Institute of Technology, illenberger@vsp.tu-berlin.de
†Ecole Polytechnique Fédédral de Lausanne, gunnar.floetteroed@epfl.ch

probability to be revealed in a snowball sample than less strongly connected individuals. Since well-connected vertices are overrepresented in the sample, any inference of statistical network properties needs to correct for this bias.

Several methods to account for the bias in snowball sampling have been proposed in the past [4, 6, 7, 8, 9, 10]. However, since snowball sampling can be implemented in quite different variants, each specification requires its own estimation method. This article treats a snowball sampling design that is targeted at revealing structural properties of social networks. A correction for the approximately unbiased estimation of the network's mean degree, its degree-degree correlation, and its transitivity is presented and validated with Monte Carlo simulations on several networks. The interested reader is referred to [11] for a real-world application of the proposed sampling and estimation techniques.

The remainder of this article is organised as follows. Section 2 defines the considered snowball sampling design and gives an overview of related work. Section 3 derives an approximation of the inclusion probability of a vertex in the sample and presents the resulting sampling corrections. Section 4 evaluates the proposed estimators in a large number of settings and on several networks. Finally, Section 5 concludes the article.

2 Definitions and related work

This section defines the considered snowball sampling design and relates it to the existing literature.

2.1 Snowball sampling design

Consider an undirected and unweighted graph without self-loops. Let $\mathcal{V}$ be the set of vertices, and let N be its size. Further consider the graph to consist of one giant component, i.e, each vertex in the graph is reachable by each other vertex. The considered snowball sampling algorithm proceeds as follows:

1. Initialise an empty set $\mathcal{S}$ of sampled ego-vertices.
2. Set iteration counter i to 0.
3. Initialise an empty set $\mathcal{R}^{(i)}$ of recruited alter-vertices.
4. Draw $n^{(i)}$ vertices (seeds) uniformly and without replacement from $\mathcal{V}$. Add those vertices to $\mathcal{R}^{(i)}$.
5. Repeat until $\mathcal{S}$ contains at least the desired number of vertices or i has reached some maximum value:
   (a) Initialise an empty set $\mathcal{R}^{(i+1)}$ of recruited alter-vertices.
   (b) Ask each vertex in $\mathcal{R}^{(i)}$ to report its neighbours (alters). Add those neighbours to $\mathcal{R}^{(i+1)}$.
   (c) Move all vertices of $\mathcal{R}^{(i)}$ that did respond to the enquiry to $\mathcal{S}$.
   (d) Remove all vertices from $\mathcal{R}^{(i+1)}$ which are already in $\mathcal{S}$.
   (e) Increase iteration counter i by one.

The sampled network consists of the vertices in $\mathcal{S}$, denoted as *ego-vertices*, and the vertices in $\{\mathcal{R}^{(i)}\}_i$, denoted as *alter-vertices*, which are those vertices that either did not respond or were not asked because the snowball was aborted. The differentiation between ego-vertices and alter-vertices is crucial since some vertex properties, such as the degree, are only known for ego-vertices. The above algorithm specifies that sampling is done without replacement, i.e., an ego-vertex is never enquired twice, and thus the sampled graph does not contain double edges. For simplicity, it is assumed that the response probability of a vertex is constant, and once a vertex is non-responding it maintains this state throughout the remaining sampling process.

Table 1: Comparison of different snowball sampling designs. Notation for braching rules: k_i represents degree of vertex i to be expanded; "k_i" = all neighbours are reported; "$\propto k_i$" = number of reported neighbours is propotional to vertex degree; "k^*" = number of reported neighbours is constant for all vertices.

reference	graph	sampling design for seeds	branching rule	replacement
Goodman [4]	undirected	uniform	k_i; all vertices have same degree	without
Frank [6]	directed	"	k_i; only one iteration	"
Johnson et al.[12]	"	"	$\min\{k^*, k_i\}$	"
Snijders [7]	both	"	$\propto k_i$	"
Lee et al. [13]	undirected	uniform (n=1)	k_i	"
Volz and Heckathorn [10]	"	non-uniform	$k^* = 1$	with
present study	"	uniform	k_i; alters can be non-responding	without

2.2 Related work

This section clarifies the difference between several snowball sampling designs from the literature and the approach presented in Sec. 2.1. For this, the following characteristic aspects of a snowball sampling design are identified:

- Does the snowball run on a directed or an undirected graph?
- What is the sampling distribution for the seed-vertices?
- How is the branching rule defined? Are all alters recruited or, for instance, is there a recruiting probability for each alter.

A selective overview of snowball sampling designs with respect to these criteria is given in Tab. 1 and discussed in the following.

One of the first authors who uses the term snowball sampling is Goodman [4]. He focuses on the estimation of undirected edges given a snowball sample that is conducted for a given number of iterations. Quit differently from later studies, Goodman defines the underlying graph to be regular such that each vertex has the same predefined degree.

Frank [6] and later also Snijders [7] address the estimation of the inclusion probabilities of vertices and edges. Knowledge about the inclusion probabilities allows for unbiased estimates of population totals and means. Both authors show that the inclusion probabilities for a snowball sampling that is run only to the first iteration can be directly calculated. Snijders [7] also considers snowball samples with multiple iterations. If a snowball sample is run for $2i - 1$ iterations, then the inclusion probabilities of a vertex can be calculated since the number of vertices with geodesic distance (number of edges of shortest path) $\leq i$ is known and thus each possible recruiting path can be identified. However, this requires to perform $i-1$ more iterations just to calculate the shortest paths and further requires that the branching rule is defined such that all vertices reported by an ego-vertex participate in the survey, i.e., each vertex is fully expanded. It is questionable that this requirement can be met in reality.

The problem of estimating the vertex in-degree from a snowball sample is addressed with Monte Carlo simulations by Johnson et al. [12]. They investigate the effects of the number of seeds, number of iterations, and maximum number of neighbours each vertex is allowed to report on the estimated in-degrees. Johnson et al. highlight that the probability of being included in the sample increases with increasing in-degree and thus results in smaller errors when estimating this quantity. They state that the number of iterations accounts for most of the estimation errors, whereas the number of seeds has only a minor effect.

A comparison of snowball sampling with the ego-centric sampling approach and link-sampling, i.e., a random draw of edges, is presented by Lee et al. [13]. Naturally, the latter sampling approach is only applicable of edges are observable. They conduct numerical simulations on real-world networks, including a protein interaction network, the Internet at the

autonomous systems level, and a co-authorship network. Their results indicate that snowball sampling underestimates several topological network properties such as the exponents of the power-law degree, the betweenness distribution, and the degree-degree correlation.

A common application of snowball sampling is to access specific populations that are difficult or even impossible to reach through direct sampling. Such applications are addressed by Frank and Snijders [8] and Heckathorn [14, 9]. Heckathorn's approach, known as Respondent-Driven Sampling (RDS), is probably the most common real-world application of snowball sampling. Especially, in medical research RDS is of interest as it allows to access hidden or hard-to-reach populations such as drug-users or HIV infected people.

In RDS, the selection of seeds is typically non-uniform but aims at individuals who are somehow related to the target population. RDS requires a respondent to recruit only one neighbour. Hence, the sampling process constitutes a random walk on a graph with the transition probability from vertex v to vertex w being $p_{vw} = 1/k_v$ where k_v denotes the degree of v [10]. With each additional step, i.e., with each additional sample, this process converges to a known equilibrium distribution from which the selection probability of a vertex can be derived. Thus the error of the estimates decreases with increasing sample size. However, this implies that the sampling process is with replacement, i.e., an individual can be recruited multiple times – an aspect in which RDS differs from the above sampling designs. A comprehensive review of the RDS methodology including a detailed discussion of the strengths and weaknesses is given by Gile and Handcock [15].

The present study focuses on the unbiased identification of structural network properties. The snowball is initialised with a uniform sample of seeds. Each vertex is assumed to report all of its neighbours, however, neighbours may be non-responding with a constant probability. The number of iterations is only constrained by the size of the underlying network.

3 Estimation

Snowball sampling selects vertices with unequal inclusion probabilities. However, in contrast to other sampling strategies such as importance sampling, the inclusion probabilities are not deliberately chosen but are, except for the initial and first iteration, unknown. While all inclusion probabilities are equal in the zero-th iteration, they scale with the vertex degree in the first iteration because each neighbour is a potentially recruiting vertex. In succeeding iterations, the inclusion probability of a vertex does not only depend on its degree but also on the degrees of its neighbours.

Even though this effect is undesired for network parameter estimation, it is of advantage for immunisation strategies: Randomly selecting a person to immunise and also immunising her contacts increases the chance of reaching persons with higher connectivity and hence higher exposure to infectious contacts (see, for instance, [16]).

3.1 Inclusion probability

In the remainder of this article, the following notation is used: Quantities that are calculated based on different iterations of the snowball sampling are written with the iteration index in parentheses in the superscript. For instance, the number of ego-vertices sampled in iteration i is denoted by $n^{(i)}$, and the number of ego-vertices that have been sampled up to and including iteration i is denoted by $n^{(\leq i)}$. Symbols without an iteration index refer to the complete sample. For example, π_v is the inclusion probability of vertex v in the entire sample.

To obtain estimators for the population total, mean, and variance of a quantity of interest, one requires the π-expanded values y_v/π_v where y_v is the quantity of interest for a sampled vertex v. The inclusion probabilities π_v are unknown a priori, but they can be estimated from the data.

Denote by $\pi_v^{(\leq i)}$ the probability that vertex v is included in a snowball sample that has been run up to and including iteration i. Given a 100 percent response rate, this equals the probability that one of v's neighbours has been sampled in or before the previous iteration $i-1$. Observing that the probability that a vertex v is *not* sampled in or before iteration i is the joint probability that none of its neighbours w has been sampled in or before the previous iteration $i-1$, and assuming that the events of being not sampled are independent, one obtains

$$\pi_v^{(\leq i)} = 1 - \prod_{w \sim v} \left(1 - \pi_w^{(\leq i-1)}\right) \tag{1}$$

where $w \sim v$ reads as "w is a neighbour of v". The probability $\pi_w^{(\leq i-1)}$ is, however, just as unknown as $\pi_v^{(\leq i)}$. A simple assumption, which will turn out later to yield quite satisfactory results, is to assume that all neighbours of v are included in the sample up to iteration $i-1$ independently and with equal probabilities. This assumptions implies that a candidate vertex v reveals no information about the sampling probabilities of its neighbours. Since these probabilities actually depend on the degrees of the neighbours, an implicit assumption is that there is no degree correlation in the network. In other words, this π-estimator treats the sample as obtained from a snowball conducted only up to iteration 1 with $n^{(\leq i-1)}$ randomly drawn seeds (see also [6] and [7]). The implications of this simplification are experimentally investigated in the next section.

Based on the above independence assumption, the inclusion probability of a neighbour is approximated by

$$\pi_w^{(\leq i-1)} \approx \frac{n^{(\leq i-1)}}{N}. \tag{2}$$

The resulting estimator of $\pi_v^{(\leq i)}$ in Eq. (1) becomes

$$\hat{\pi}_v^{(\leq i)} := 1 - \prod_{w \sim v} \left(1 - \frac{n^{(\leq i-1)}}{N}\right). \tag{3}$$

Since the factors in Eq. (3) are equal for all neighbours, one obtains

$$\hat{\pi}_v^{(\leq i)} := \hat{\pi}^{(\leq i)}(k_v) := 1 - \left(1 - \frac{n^{(\leq i-1)}}{N}\right)^{k_v} \tag{4}$$

where k_v is the degree of vertex v and $\hat{\pi}^{(\leq i)}(k)$ is the arguably most simple estimator of the inclusion probability that only depends on the degree of a considered vertex. This estimator is applicable for $i > 0$; in the zero-th iteration, samples are drawn uniformly such that $\hat{\pi}^{(0)}(k) = \pi^{(0)}(k) = n^{(0)}/N$.

3.2 Population mean

Given the estimated inclusion probabilities $\hat{\pi}_v$, one obtains

$$\hat{t}_y := \sum_{v \in \mathcal{S}} \frac{y_v}{\hat{\pi}_v} \tag{5}$$

as an estimator for the population total of a quantity of interest y, where $\mathcal{S}$ is the set of sampled vertices. Hence,

$$\hat{\bar{y}} := \frac{\hat{t}_y}{N} \tag{6}$$

constitutes an estimator of the population mean. It is known as the Horwitz-Thompson estimator [17]. Since this estimator requires knowledge of the population size N, this information can be further exploited to improve the estimation of the inclusion probabilities. Noting that $\sum_{v \in \mathcal{S}} 1/\hat{\pi}_v$ is an estimator of the total population size N, the inclusion probabilities $\hat{\pi}_v$ are uniformly scaled by a factor κ such that an unbiased estimator of the population size results:

$$\sum_{v \in \mathcal{S}} \frac{1}{\kappa \hat{\pi}_v} \stackrel{!}{=} N \tag{7}$$

such that

$$\kappa = \frac{\sum_{v \in \mathcal{S}} 1/\hat{\pi}_v}{N}. \tag{8}$$

Replacing $\hat{\pi}_v$ by $\kappa \hat{\pi}_v$ in Eq. 5, substituting Eq. 8, and evaluating Eq. 6 results in:

$$\hat{\bar{y}}' := \frac{\sum_{v \in \mathcal{S}} y_v/\hat{\pi}_v}{\sum_{v \in \mathcal{S}} 1/\hat{\pi}_v}, \tag{9}$$

which is known as the weighted sample mean [18]. Although it has been derived based on the additional constraint (7) that makes use of the known population size N, the final estimator does not require knowledge of this quantity. Intuitively, it can be expected that $\hat{\bar{y}}'$ performs better than $\hat{\bar{y}}$ since it exploits the additional condition Eq. 7. It does so without knowing the population size by evaluating the unscaled (and possibly biased) inclusion probabilities $\hat{\pi}_v$ both in the numerator and the denominator, which can be expected to have a compensatory effect on the overall estimation result. The following experiments shed more light on this effect.

4 Simulation

4.1 Simulation setup

To validate the performance of several estimators based on Sec. 3, a series of numerical experiments implementing the snowball sampling design according to Sec. 2.1 is conducted. The following networks are considered in order to investigate the estimators' performance for different topologies:

1. an Erdős-Rényi [19] random network with 36'458 vertices and a mean degree of 9;
2. a Barabási-Albert [20] network with 36'461 vertices, a power-law degree distribution and a mean degree of 6;
3. the giant component extracted from a co-authorship network of physicists [2]. The giant component represents a network with 36'458 vertices, a mean degree of 9.4, and exhibits a positive degree-degree correlation [21].

Accounting for the stochasticity in the simulations, each experiment is repeated 1000 times. Each simulation run is initialised with ten randomly drawn seed-vertices and is run until the complete network or all from the seed-vertices reachable vertices are sampled. The average performance of the different estimators over the number of snowball iterations is evaluated. For now, it is assumed that all vertices are responding. Experiments with response rates less than one are presented in Sec. 4.3.

4.2 Experiments with response rate of one

4.2.1 Mean degree

From Eq. 6 and Eq. 9, one obtains two estimators for the mean degree:

$$\hat{\bar{k}} = \frac{1}{N} \sum_{v \in \mathcal{S}} \frac{k_v}{\hat{\pi}_v} \tag{10}$$

and

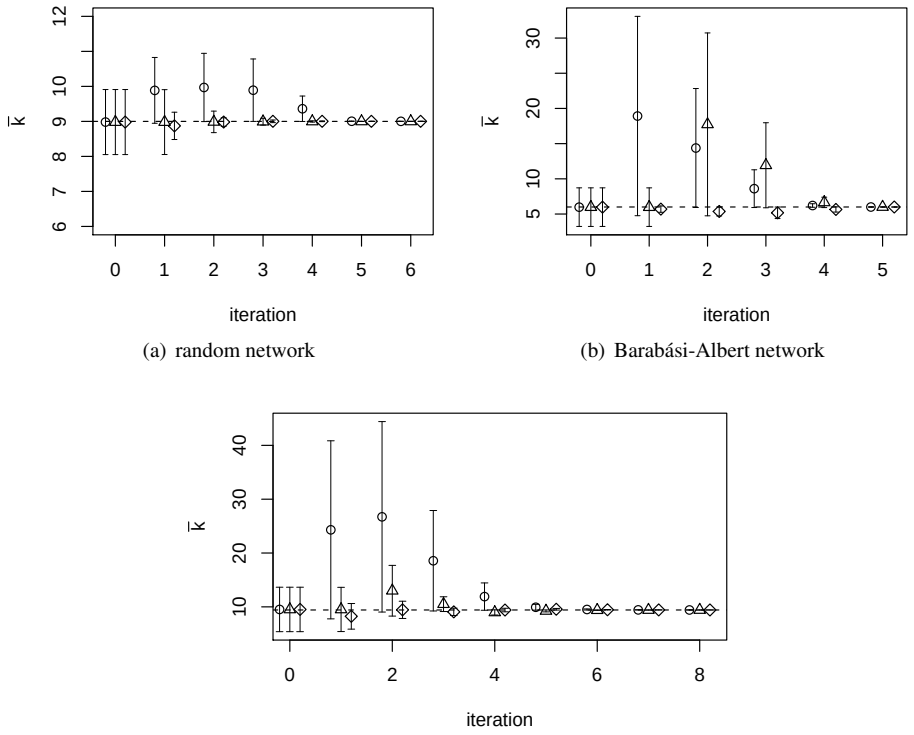
$$\hat{\bar{k}}' = \frac{\sum_{v \in \mathcal{S}} k_v / \hat{\pi}_v}{\sum_{v \in \mathcal{S}} 1 / \hat{\pi}_v} \,. \tag{11}$$

Note that only the ego-vertices in $\mathcal{S}$ are accounted for. Alter-vertices (remaining vertices in $\{\mathcal{R}^{(i)}\}_i$) are not considered because their degree is unknown.

Figure 1 shows, for all three networks, the estimated mean degree for both estimators as well as for a naive estimator where the sampling correction is omitted. The naive estimator reveals the bias of the snowball sampling in that its values are permanently above the true mean degree. The bias is the strongest in early iterations (except, of course, iteration 0) since in those iterations vertices with high degrees are heavily overrepresented in the sample. However, the bias of the naive estimator behaves differently for different network topologies. While for the random network it is only of small magnitude, it is much more pronounced for the Barabási-Albert and the co-authorship network. Considering the Barabási-Albert network this can be explained by the much broader degree distribution. Additionally to the broader degree distribution the positive degree-degree correlation in the co-authorship network introduces a positive feedback in the growth of high-degree vertices in the sample.

Both estimators $\hat{\bar{k}}$ and $\hat{\bar{k}}'$ perform well for the random network. The weighted sample mean $\hat{\bar{k}}'$ performs slightly better than the Horwitz-Thompson estimator $\hat{\bar{k}}$ (Fig 1). This difference becomes much more distinct in the Barabási-Albert and the co-authorship network. While $\hat{\bar{k}}'$ provides quite precise estimates of the real mean degree, $\hat{\bar{k}}$ is in some situations even worse than the naive estimator, cf. iterations 2 and 3 in Fig. 1(b).

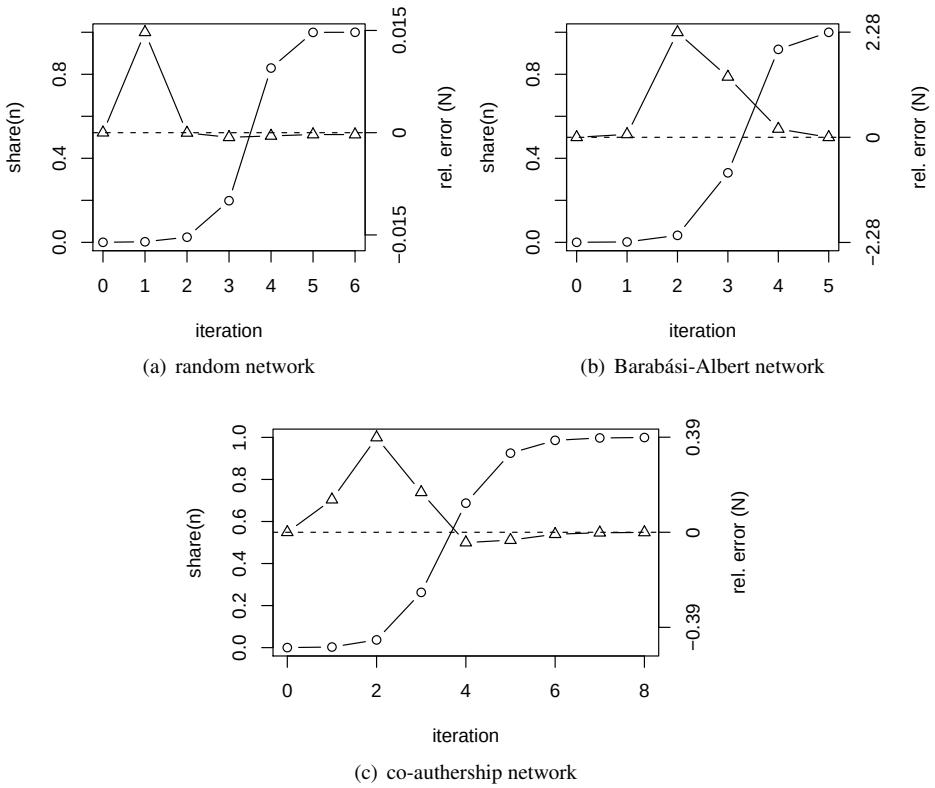
The superiority of the weighted sample mean can be explained by the fact that it is based on an unbiased estimation of the population size, cf. Sec. 3.2. Figure 2 shows the estimated population size $\sum_{v \in \mathcal{S}} 1/\hat{\pi}_v$, i.e., without the scaling coefficient κ that ensures unbiasedness. The population size is well estimated for the random network, which indicates that the unscaled inclusion probabilities $\hat{\pi}_i$ are reasonably well estimated in this case. This is plausible because the random network exhibits no degree correlation, which is neglected in the estimation of the inclusion probabilities as well.



(a) random network (b) Barabási-Albert network

(c) co-authership network

Figure 1: Mean degree calculated each time an iteration is completed; numbers are averaged over the simulation ensemble. ◯ = naive estimator, △ = $\hat{\bar{k}}$ estimated with Hortwitz-Thompson-Estimator Eq. 10, ◊ = $\hat{\bar{k}}'$ estimated with weighted sample mean Eq. 11. The dotted line indicates the true mean degree. Error bars indicate the root mean square error.



(a) random network (b) Barabási-Albert network

(c) co-authership network

Figure 2: Share of sampled ego-vertices (circles) and relative error of estimated population size $\hat{N}$ (triangles).

The population size is vastly overestimated in the two networks with broad degree distributions, which indicates that the inclusion probabilities in particular of high degree nodes are underestimated in this case. This effect has already been explained: the positive degree correlation introduces a feedback in the snowball sampling in that a larger share of high-degree vertices within one wave reaches even more high-degree vertices in the next wave. Since the weighted sample mean implicitly corrects for this bias through the estimated population size in the denominator, cf. Eq. 9, it performs well even for networks with a positive degree correlation. This indicates that the *unscaled* version of $\hat{\pi}_v$ does not yield particularly good estimates, but it captures the *relative* values of the inclusion probabilities quite well.

4.2.2 Degree-degree correlation

The degree-degree correlation can be quantified by the Pearson correlation coefficient of the degrees of the vertices on either side of all edges in the network:

$$r = \frac{\mathrm{Cov}\left(\mathcal{E}\right)}{\sqrt{\mathrm{Var}\left(\mathcal{E},v\right)}\sqrt{\mathrm{Var}\left(\mathcal{E},w\right)}} \tag{12}$$

where

$$\mathrm{Cov}\left(\mathcal{E}\right) = \frac{\sum_{e\in\mathcal{E}} k_v(e)k_w(e)}{M-1} - \frac{\sum_{e\in\mathcal{E}} k_v(e)\sum_{e\in\mathcal{E}} k_w(e)}{M\left(M-1\right)} \tag{13}$$

denotes the covariance of the degrees k_v and k_w of the two adjacent vertices of an edge e (in arbitrary yet unique order) and $\mathcal{E}$ the set of all edges in the network with $M = |\mathcal{E}|$ is its size.

$$\mathrm{Var}\left(\mathcal{E},v\right) = \frac{\sum_{e\in\mathcal{E}} \left(k_v(e)\right)^2}{M-1} - \frac{\left(\sum_{e\in\mathcal{E}} k_v(e)\right)^2}{M\left(M-1\right)} \tag{14}$$

is the variance of the degrees of a vertex v adjacent to an edge e.

To properly determine the degree-degree correlation for a sampled network we evaluate Eq. 12 only for the set of sampled edges $\mathcal{T}$ with size $m = |\mathcal{T}|$, where an edge is denoted as sample if both vertices are in $\mathcal{S}$:

$$\tilde{r} = \frac{\mathrm{Cov}\left(\mathcal{T}\right)}{\sqrt{\mathrm{Var}\left(\mathcal{T},v\right)}\sqrt{\mathrm{Var}\left(\mathcal{T},w\right)}} \,. \tag{15}$$

In contrast to one-point properties, such as the degree, the sample of interest is now an edge. The inclusion probability π_e of an edge $e = (vw)$ follows from the observation that the probability that an edge is sampled before or in iteration i equals the probability that at least one of its adjacent vertices v or w is sampled before or in iteration $i-1$. Hence,

$$\hat{\pi}_{(e)}^{(\leq i)} = \left(\hat{\pi}_v^{(<i)} + \hat{\pi}_w^{(<i)}\right) - \left(\hat{\pi}_v^{(<i)}\hat{\pi}_w^{(<i)}\right) \,, \tag{16}$$

where again independence of the sampling events is assumed. An estimator $\hat{r}'$ of the degree-degree correlation can now be obtained by (i) estimating M, the total number of edges, by $\hat{M} = \sum_{e\in\mathcal{T}} 1/\pi_e$ and (ii) estimating the edge inclusion probabilities according to Eq. 16 from the approximate vertex inclusion probabilities:

$$\hat{r}' = \frac{\hat{\mathrm{Cov}}\left(\mathcal{T}\right)}{\sqrt{\hat{\mathrm{Var}}\left(\mathcal{T},v\right)}\sqrt{\hat{\mathrm{Var}}\left(\mathcal{T},w\right)}} \tag{17}$$

where

$$\hat{\mathrm{Cov}}\left(\mathcal{T}\right) = \frac{\sum_{e\in\mathcal{E}} \frac{k_v(e)k_w(e)}{\hat{\pi}_{(e)}}}{\hat{M}-1} - \frac{\sum_{e\in\mathcal{E}} \frac{k_v(e)}{\hat{\pi}_{(e)}}\sum_{e\in\mathcal{E}} \frac{k_w(e)}{\hat{\pi}_{(e)}}}{\hat{M}\left(\hat{M}-1\right)} \tag{18}$$

and

$$\hat{\mathrm{Var}}\left(\mathcal{T},v\right) = \frac{\sum_{e\in\mathcal{T}} \left(\frac{k_v(e)}{\hat{\pi}_{(e)}}\right)^2}{\hat{M}-1} - \frac{\left(\sum_{e\in\mathcal{T}} \frac{k_v(e)}{\hat{\pi}_{(e)}}\right)^2}{\hat{M}\left(\hat{M}-1\right)} \,. \tag{19}$$

The resulting estimator again does not require knowledge of M, which is typically unknown (or even the quantity of interest) in real applications.

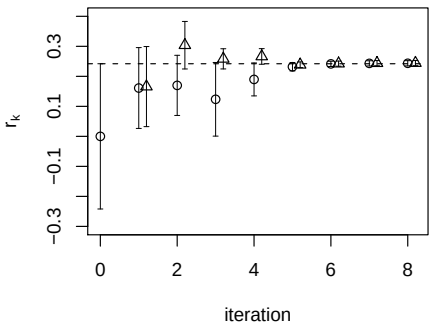


Figure 3: Degree-degree correlation of the co-authership network calculated each time an iteration is completed and averaged over the simulation ensemble. ◯ = naive estimator $\tilde{r}$ of degree-degree correlation, △ = degree-degree correlation estimator $\hat{r}'$. Error bars indicate root mean square errors. The dotted line shows the real value of r.

Random networks have by definition a degree correlation of zero. It has also been shown that the model of Barabási and Albert exhibit no degree correlation [21]. Hence, the following analysis concentrates on the co-authorship network, which exhibits a positive degree correlation of $r = 0.24$.

Fig. 4.2.2 shows the estimation results for the naive estimator $\tilde{r}$ that does not correct for the sampling and for the proposed estimator $\hat{r}'$. The naive estimator underestimates the degree-degree correlation until the majority of vertices is sampled in iteration 5. The sampling correction in the $\hat{r}'$ estimator removes this undesired effect substantially and performs better than the naive estimator as from iteration two.

4.2.3 Transitivity

Network transitivity can be quantified with the clustering coefficient, which comes in two versions. A global definition [22] is

$$C = \frac{3 \cdot n\,(\text{triangles})}{n\,(\text{connected triples})}\,. \tag{20}$$

where $n(\cdot)$ reads as "number of $\cdot$". The alternative definition by Watts and Strogatz [23] is the average over a local vertex parameter:

$$\bar{C} = \frac{1}{N} \sum_{v \in \mathcal{V}} \frac{2m_v}{k_v\,(k_v - 1)}\,, \tag{21}$$

where m_v is number of edges that connect neighbours of v. Both definitions can lead to quite different results as small-degree vertices have a small denominator in Eq. 21 and thier contribution is weighted more heavily [22].

In the estimation of this quantity, only ego-vertices all neighbours of which are also ego-vertices are accounted for because the neighbourhood of alter-vertices is unknown (in particular edges between alter-vertices are missing). This means that only those ego-vertices that have been sampled strictly before the last iteration are considered.

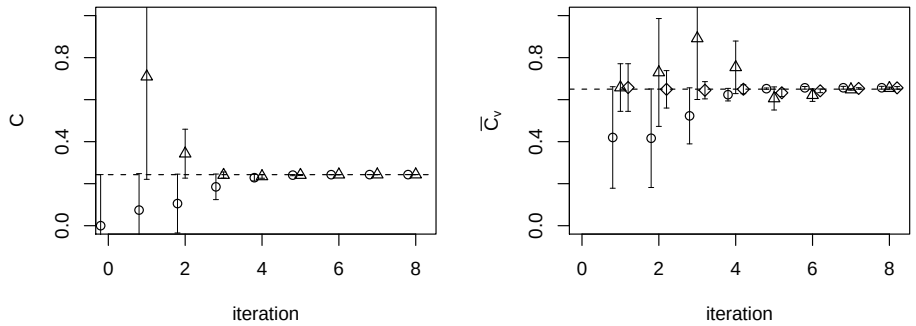
The already derived sampling correction based on vertex inclusion probabilities is clearly better applicable for the estimation of $\bar{C}$ from a sample than for the estimation of C: Since $\bar{C}$ is a vertex-local property, its population mean can be directly estimated according to the Horwitz-Thompson estimator

$$\hat{\bar{C}}^{(\leq i)} = \frac{1}{N} \sum_{v \in \mathcal{S}^{(<i)}} \frac{2m_v}{k_v\,(k_v - 1)} \cdot \frac{1}{\hat{\pi}_v^{(<i)}} \tag{22}$$

or according to the weighted sample mean

$$\hat{\bar{C}}_v'^{(\leq i)} = \frac{1}{\sum_{v \in \mathcal{S}^{(<i)}} 1/\hat{\pi}_v^{(<i)}} \sum_{v \in \mathcal{S}^{(<i)}} \frac{2m_v}{k_v\,(k_v - 1)} \cdot \frac{1}{\hat{\pi}_v^{(<i)}}\,. \tag{23}$$

Figure 4: Clustering coefficient of the co-authership network.



(a) (b)

Figure 5: (a) Naive estimator of the network clustering coefficient C defined in Eq. 20 and calculated each time an iteration is completed and averaged over the simulation ensemble. ◯ = connected triples and triangles counted from ego- and alter-vertices, △ = connected triples and triangles counted only from ego-vertices. (b) Mean clustering coefficient as defined in Eq. 21 and estimtated each time an iteration is completed and averaged over the simulation ensemble. ◯ = naive estimator without sampling correction, △ = $\hat{\bar{C}}$ estimated with the Horwitz-Thompson estimator according to Eq. 22, ◇ = $\hat{\bar{C}}'$ estimated with the weighted sample mean according to Eq. 23. (a) and (b): The dotted line indicates the true value of C and $\bar{C}$, respectively. Error bars indicate the root mean square error.

Since the random and Barabási-Albert network exhibit no or insignificant clustering, the investigation focuses on the co-authorship network, analogously to Sec. 4.2.2.

Figure 5(a) shows the results obtained with a naive estimator of the network clustering coefficient (Eq. 20). Since the alter-vertices of the current iteration do predominantly contribute to connected triples but only rarely to triangles, the network clustering coefficient is underestimated. Accounting only for ego-vertices in the calculation does not add much of an improvement.

The local clustering coefficient (Eq. 21) is underestimated by a naive estimator that does not correct for the sampling bias up to iteration 4 (Fig. 5(b)). This is so because in the co-authorship network the values for the local clustering coefficient correlate negatively with the degree. Thus, vertices with a low local clustering coefficient are overrepresented in the samples of early iterations. The weighted sample mean performs very well as from iteration one, which is the first iteration where an estimation of the local clustering coefficient is possible. The Horwitz-Thompson estimator does perform much worse, most likely due to the same reasons given for its inferiority when estimating the mean degree.

4.3 Experiments with response rate below one

In the above sections, it has been assumed that the response rate is one, i.e., that all inquired vertices report all of their neighbours. However, in real applications researchers are faced with considerably lower response rates. In an application of the presented snowball sample design, a response rate of approximately 25 % is observed [24].

In the following, the proposed estimation approach is extended in order to account for a response rate below one. The sensitivity of the estimator with respect to variations in the response rate is also investigated. It is assumed that a fraction of α vertices is non responding. In the snowball simulations, these vertices are selected from a uniform distribution before the sampling starts. The tagged vertices are not expanded during the snowball iterations. This approach implies the assumption that the the response rate is equally distributed over all vertices and does not change throughout the sampling process.

4.3.1 Inclusion probability and population mean

The estimated inclusion probabilities are straightforward to extend in order to account for the response rate α:

$$\hat{\pi}_{v,\alpha} = \alpha \hat{\pi}_v. \tag{24}$$

The response rate α can be obtained directly from the survey data through

$$\alpha^{(\leq i)} = \frac{n^{(\leq i)}}{n^{(<i)} + a^{(<i)}} \tag{25}$$

where $n^{(\leq i)}$ denotes the number of ego-vertices sampled up to and including iteration i, $n^{(<i)}$ denotes the number of ego-vertices sampled strictly before iteration i, and $a^{(<i)}$ denotes the number of alter-vertices sampled strictly before iteration i. In words, the numerator corresponds to the number of all vertices that have responded to an inquiry before or in iteration i, an the denominator corresponds to the number of all vertices that have been inquired strictly before iteration i.

The previously developed estimators can be applied for response rates below one by replacing $\hat{\pi}_v$ with $\hat{\pi}_{v,\alpha}$. In particular, the response rate strikes out in the accordingly adopted weighted sample mean:

$$\hat{\bar{y}}' = \frac{1}{\sum_{v \in \mathcal{S}} 1/\left(\alpha \hat{\pi}_v\right)} \sum_{v \in \mathcal{S}} \frac{y_v}{\alpha \hat{\pi}_v} \,, \tag{26}$$

clearly an additional advantage of $\hat{\bar{y}}'$ over $\hat{\bar{y}}$.

4.3.2 Degree and degree-degree correlation

Figure 6 shows the estimated mean degree for all three types of investigated networks. For the estimation of the population mean we only consider the weighted sample mean since the above results clearly showed that it performs superior compared to the Horwitz-Thompson estimator. The response rate is varied from 0.1 to 0.5 in 0.05 steps. Values of $\alpha > 0.5$ are not considered since it is unlikely that such high rates are archived in reality. Different from the preceding sections the sampled network is not analysed after the completion of an iteration, but after a certain number of ego-vertices are sampled. Different response rates result in different sample sizes per iteration which makes a comparison based on iterations less meaningful. Moreover, in practical applications the extend of the survey is usually constrained by the costs it takes to sample a vertex rather than the number of iterations conducted. The sampled network is analysed each time 100 additional ego-vertices are sampled up to a total population of 1000 sampled ego-vertices and then each time after 1000 additional vertices are sampled.

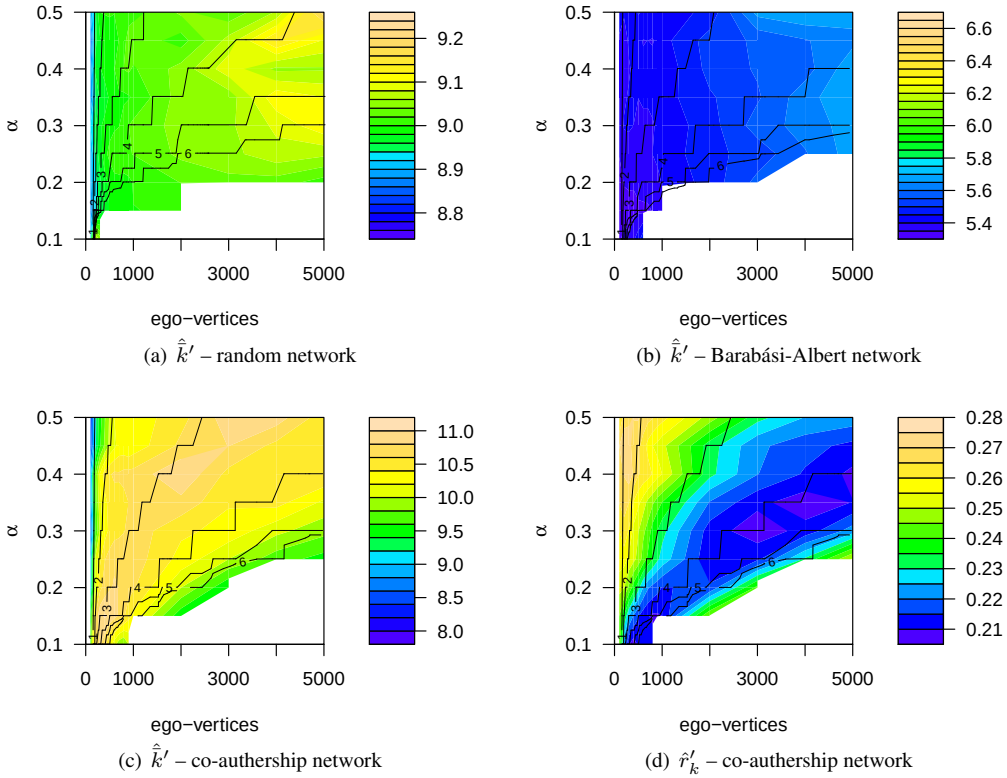
The estimated mean degree $\hat{\bar{k}}'$ for the random network is rather unaffected by the response rate. The estimator performs quite well and shows only minor sensibility towards the sample size (Fig. 6(a)). The same behaviour is observed with the Barabási-Albert network: The estimated mean degree $\hat{\bar{k}}'$ shows only sensibility towards the number of ego-vertices but not towards α (Fig. 6(b)). This means that for both networks only the sample size affects the estimation of the mean degree, regardless the response rate and number of iterations conducted.

A completely different picture is drawn with the co-authorship network (Fig. 6(c). The values of $\hat{\bar{k}}'$ correlate with the iteration as the "relief" is stretched with increasing α. The characteristic behind this effect is the positive degree-degree correlation. It appears that the bias is dependent on the number of iterations rather than on the number of sampled ego-vertices. Yet, after the bias has been "overcome" $\hat{\bar{k}}'$ provides a good approximation of $\bar{k}$. In this regard, it is even beneficial if the response rate is low since then one can conduct more iterations with the same number of samples and hence faster overcome the bias. The estimated degree-degree correlation exhibits the same effect (Fig. 6(d)).

4.3.3 Transitivity

Section 4.2.3 already mentions the issue of missing edges m_v between alter-vertices. With low response rates this issue becomes more distinct. If the neighbours of ego-vertex v are non-responding it is likely that the sample data misses existing edges between the neighbours, i.e., the value of m_v in Eq. 21 is likely to be underestimated. Here, a rather simple approach is chosen to estimate m_v.

Consider n_v as the number of neighbours of ego-vertex v that are in $\mathcal{S}$. Denote $k_v - n_v$ as the number of neighbours of v that are not in $\mathcal{S}$, i.e., neighbours of v that are either non-responding or have not yet been enquired. Further denote p_e as the probability of an edge connecting two neighbours of v. The estimated number of missing edges is p_e times the



(a) $\hat{\bar{k}}'$ – random network

(b) $\hat{\bar{k}}'$ – Barabási-Albert network

(c) $\hat{\bar{k}}'$ – co-authership network

(d) $\hat{r}'_k$ – co-authership network

Figure 6: (a) – (c): Estimated mean degree $\hat{\bar{k}}'$. (d): Estimated degree-degree correlation $\hat{r}'_k$. The colors indicate the values of $\hat{\bar{k}}'$ and $\hat{r}'_k$, respectively. Green color inidicates the true value of $\bar{k}$ and r_k, respectively. Black lines indicate the iteration transitions (averaged over the simulation ensemble). For better visibility only iteration one to six are drawn.

number of possible edges between unsampled neighbours $k_v - n_v \left(k_v - n_v - 1\right)/2$. The estimated number of edges $\hat{m}_v$ is then the sum of actually observed edges and the estimated number of missing edges:

$$\hat{m}_v = m_v + p_e \frac{1}{2} \left(k_v - n_v\right) \left(k_v - n_v - 1\right) \ . \tag{27}$$

Probability p_e can be obtained from the hitherto sampled data. Denote M_v as the possible number of edges between neighbours of v that can be observed given the response rate α. Note that M_v is not $k_v \left(k_v - 1\right)/2$ since edges between non-responding neighbours cannot be observed. Instead, M_v is the number of possible edges that can occur between responding neighbours and between responding and non-responding neighbours:

$$M_v = \frac{1}{2} n_v \left(n_v - 1\right) + n_v \left(k_v - n_v\right) \ , \tag{28}$$

and thus

$$p_e = \frac{\sum_{v \in \mathcal{S}} m_v}{\sum_{v \in \mathcal{S}} M_v} \ . \tag{29}$$

Probability p_e is an average over all ego-vertices to avoid artefacts if the degree of a v is small or the response rate is very low.

The alters of the last iteration, i.e., those that are not enquired yet, can be treated as non-responding vertices. This allows to estimate m_v also for the ego-vertices of the last iteration. The mean clustering coefficient can thus be calculated as the mean over all ego-vertices in the sample, contrary to Sec. 4.2.3 where ego-vertices of the last iteration are excluded.

Figure 7 shows the estimated local clustering coefficient $\hat{\bar{C}}'_v$ which exhibits a surprisingly low sensibility towards the response rate. Even at low response rates it requires only about 300 ego-vertices to get a very precise estimated of the clustering coefficient. Moreover, the estimates show to be independent of the number of iterations conducted, such as the estimated mean degree of the random and Barabási-Albert network.

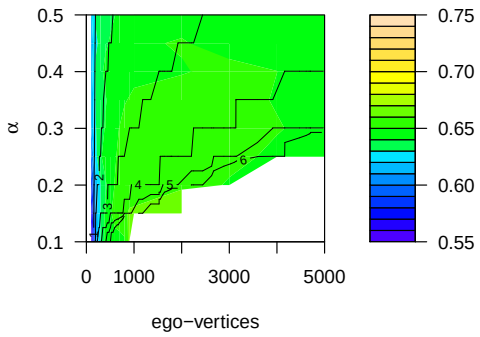


Figure 7: Estimated mean clustering coefficient $\hat{\bar{C}}'_v$. The colors indicate the values of $\hat{\bar{C}}'_v$. Green color indicates the true value of $\bar{C}_v$. Black lines indicate the iteration transitions (averaged over the simulation ensemble). For better visibility only iteration one to six are drawn.

5 Conclusion

This article addresses the estimation of topological network properties from data obtained with a snowball sampling design. We present an estimator for the probability of a vertex to be included in the sample which allows for the estimation of the mean degree, degree-degree correlation and mean clustering coefficient. The estimation methodology treats the sample as obtained from a snowball that has been run up the first iteration and is thus fairly easy to calculate. Although, it is arguable that this assumption is to simple we show that the estimator is rather powerful and robust. From the simulation studies four major conclusion can be drawn:

- The mean vertex clustering coefficient is estimated fairly precise even with small sample sizes. Moreover, the performance of the estimator is sensible only towards the sample size. The influence of the response rate or the number of iterations conducted is negligible.
- Considering networks without degree-degree-correlations the estimator for the mean degree performs well. Its sensitivity towards the sample size, response rate and number of iterations shows the same characteristics as the estimator of the mean clustering coefficient.
- Considering networks with positive degree-degree-correlation the estimator of the mean degree as well as the estimator of the degree-degree-correlation show to be sensible towards the number of snowball iterations conducted. Both estimators show that a low response rate can be of advantage as it allows to conduct more iterations with same sample size and thus to ”overcome“ the bias.
- The performance of the estimator scales with the width of the degree distribution. The broader the distribution the worse the performance of the estimator.

Two assumptions that have been made to simplify the simulation studies are arguable: First, the total number of vertices N is usually unknown but is required for the estimation of the inclusion probability (Eq. 4). Considering large networks and relative small sample sizes and educated guess of N is sufficient. Since N is in the denominator variations to this quantity have negligible effect. Second, the response rate is assumed to be equally distributed and constant throughout the entire sampling process. In real-world applications it is observed that the response rate of respondents in later iteration decreases [24]. Adapting the estimator for a descending response rate is possible. However, difficulties will arise if the response rate correlates with other vertex properties. For instance if people with large personal networks are to busy to participate the survey.

An aspect that is still open for further research is the estimation of network global parameters such as the network diameter, closeness or betweenness. An estimation of such parameters will be quite challenging. Even the estimation of a two-point property such as the degree-degree correlation turned out to be by no means trivial. Some work in this direction has been done by Lee et al. [13] but more insights would in particular provide a sound basis for the modelling of the spreading of diseases or rumours. Finally, snowball sampling

is the designated tool for such studies as it provides an effective method to obtain connect ego-centric networks (see for instance [11]).

6 Acknowledgement

We thank Kai Nagel for helpful suggestions and support. This work was funded by the VolkswagenStiftung within the project "Travel impacts of social networks and networking tools".

References

[1] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, "Classes of small-world networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 21, pp. 11149–11152, 2000.

[2] M. E. J. Newman, "Scientific collaboration networks. i. network construction and fundamental results," *Physical Review E*, vol. 64, no. 016131, 2001.

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

[4] L. A. Goodman, "Snowball sampling," *The Annals of Mathematical Statistics*, vol. 32, no. 1, pp. 148–170, 1961.

[5] R. Atkinson and J. Flint, "Accessing hidden and hard-to-reach populations: Snowball research strategies," *Social Research Update*, vol. 33, 2001.

[6] O. Frank, *Estimation of population totals by use of snowball samples*, pp. 319–346. New York: Academic Press, 1979.

[7] T. A. B. Snijders, "Estimation on the basis of snowball samples: How to weight," *Bulletin de Méthodologie Sociologique*, vol. 36, pp. 59–70, 1992.

[8] O. Frank and T. Snijders, "Estimating the size of hidden population using snowball sampling," *Journal of Official Statistics*, vol. 10, no. 1, pp. 53–67, 1994.

[9] D. D. Heckathorn, "Respondent-driven sampling ii: Deriving vaild population estimates from chain-referral samples of hidden populations," *Social Problems*, vol. 49, no. 1, pp. 11–34, 2002.

[10] E. Volz and D. D. Heckathorn, "Probability based estimation theory for Respondent Driven Sampling," *Journal of Official Statistics*, vol. 24, no. 1, pp. 79–97, 2008.

[11] J. Illenberger, M. Kowald, K. W. Axhausen, and K. Nagel, "Insights into a spatially embedded social network from a large-scale snowball-sample," VSP Working Paper 10-10, TU Berlin, Transport Systems Planning and Transport Telematics, 2010. See www.vsp.tu-berlin.de/publications.

[12] J. Johnson, J. Boster, and D. Holbert, "Estimating relational attributes from snowball samples through simulation," *Social Networks*, vol. 11, pp. 135–158, 1989.

[13] S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Physical Review E*, vol. 73, no. 016102, 2006.

[14] D. D. Heckathorn, "Respondent-driven sampling: A new approach to the study of hidden populations," *Social Problems*, vol. 44, no. 2, pp. 174–199, 1997.

[15] K. J. Gile and M. S. Handcock, "Respondent-driven sampling: An assessment of current methodology," *Sociological Methodology*, vol. 40, no. 1, pp. 285–327, 2010.

[16] M. Andre, K. Ijaz, J. D. Tillinghast, V. E. Krebs, L. A. Diem, B. Metchock, T. Crisp, and P. D. McElroy, "Transmission network analysis to complement routine tuberculosis contact invertigations," *American Journal of Public Health*, vol. 96, no. 11, pp. 1–8, 2006.

[17] D. G. Horwitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 663–685, 1952.

[18] C.-E. Särndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling*. Springer-Verlag, 1992.

[19] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[20] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[21] M. E. J. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, 2002.

[22] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[23] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 440-442, 1998.

[24] M. Kowald, A. Frei, J. Hackney, J. Illenberger, and K. Axhausen, "Collecting data on leisure travel: The link between leisure acquaintances and social interactions," *Procedia — Social and Behavioral Sciences*, vol. 4, no. 1, pp. 38–48, 2010.