PLOS ONE

RESEARCH ARTICLE

One-step estimation of networked population size: Respondent-driven capture-recapture with anonymity

Bilal Khan[1]*, Hsuan-Wei Lee[1], Ian Fellows[2], Kirk Dombrowski[1]

1 Department of Sociology, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America, **2** Fellow Statistics, San Diego, California, United States of America

* bkhan2@unl.edu

OPEN ACCESS

Citation: Khan B, Lee H-W, Fellows I, Dombrowski K (2018) One-step estimation of networked population size: Respondent-driven capture-recapture with anonymity. PLoS ONE 13(4): e0195959. https://doi.org/10.1371/journal.pone.0195959

Editor: Ming Tang, East China Normal University, CHINA

Received: February 22, 2018

Accepted: March 31, 2018

Published: April 26, 2018

Data Availability Statement: All Java software developed and used is available at https://github.com/grouptheory/telefunken-support/tree/master/java. All R software developed and used is available at https://github.com/grouptheory/telefunkensupport/tree/master/R-v1. All data inputs and outputs are available at https://github.com/grouptheory/telefunken-support/tree/master/figures_and_data.

Funding: Research reported in this publication was supported by the National Institutes for Health,

Abstract

Size estimation is particularly important for populations whose members experience disproportionate health issues or pose elevated health risks to the ambient social structures in which they are embedded. Efforts to derive size estimates are often frustrated when the population is hidden or hard-to-reach in ways that preclude conventional survey strategies, as is the case when social stigma is associated with group membership or when group members are involved in illegal activities. This paper extends prior research on the problem of network population size estimation, building on established survey/sampling methodologies commonly used with hard-to-reach groups. Three novel one-step, network-based population size estimators are presented, for use in the context of uniform random sampling, respondent-driven sampling, and when networks exhibit significant clustering effects. We give provably sufficient conditions for the consistency of these estimators in large configuration networks. Simulation experiments across a wide range of synthetic network topologies validate the performance of the estimators, which also perform well on a real-world location-based social networking data set with significant clustering. Finally, the proposed schemes are extended to allow them to be used in settings where participant anonymity is required. Systematic experiments show favorable tradeoffs between anonymity guarantees and estimator performance. Taken together, we demonstrate that reasonable population size estimates are derived from anonymous respondent driven samples of 250-750 individuals, within ambient populations of 5,000-40,000. The method thus represents a novel and cost-effective means for health planners and those agencies concerned with health and disease surveillance to estimate the size of hidden populations. We discuss limitations and future work in the concluding section.

1 Introduction

Estimating the size of hidden and hard-to-reach populations is of critical importance to health officials seeking to mitigate the extent of health problems that may be concentrated within such populations [1], or when "reservoirs" of infection among a hidden population pose a

National Institute on Drug Abuse under Award Number R01 DA037117 and National Institute for General Medicine R01 GM118427, as well as National Science Foundation grants MMS-0851555 and SES-1357619.

Competing interests: The authors have declared that no competing interests exist.

health risk to the ambient population in which the hidden population is embedded [2, 3]. In the former, otherwise treatable maladies can remain unaddressed, multiplying eventual treatment costs when cases are discovered at more advanced stages. Such is the situation, for example, with mental illness among homeless and street dwelling populations [4–6]. An embedded "hidden" population can also frustrate intervention efforts that might otherwise be effective in the ambient population, preventing control of infection prevalence [7]. One example of this is the high prevalence of sexually transmitted disease among commercial sex workers [8–10]. In all such situations, health officials seek to estimate both the overall prevalence levels of maladies within a hidden population *and the size of the population itself*, in order to know the scope of treatment needs and overall social risk.

Efforts to ascertain prevalence and size estimates are frustrated by a range of factors that contribute to the "hiddenness" of the population. Such factors include heavy social stigma that inhibits the members of the hidden population from revealing their membership status. This is the case for people who inject drugs (PWID), who may be unwilling to self-identify as such under ordinary survey conditions [11, 12]. Hiddenness due to stigma can be further compounded when such activities are illegal, when they carry heavy personal costs (such as when self-identified heterosexual men also have sex with men), or when disease status is unknown (such as undiagnosed HIV infection rates among PWID). In these situations, conventional sampling is unreliable, and ordinary multiplier methods based on conventional sampling are rendered ineffective.

A number of techniques have been devised to address the problems of prevalence and population size estimation. These include capture-recapture [13, 14], chain referral [15, 16], venue-based sampling [17, 18], cluster sampling [19], and combinations thereof. Among the most popular is respondent-driven sampling (RDS) [20–22], which has been adapted for use in many situations, and which is employed widely in HIV surveillance efforts both within the United States and beyond [23]. RDS employs an incentivized chain referral process to recruit a sample of the hidden population. Under restricted but recognized conditions, RDS can be shown to result in a steady-state, "equilibrium" sample, and numerous methods have been derived for producing reasonable prevalence estimates from such a sample, while accounting for biases introduced in the referral process [24–29]. The ease of implementing RDS, the fact that it can operate under conditions of anonymity (via numbered coupons that track referrals), and its rigorous treatment under a range of statistical assumptions have made it a popular choice for researchers working with hidden populations [30]. While significant operational, design and analytical challenges frequently arise in deploying the RDS framework [31–33], the ability of the RDS-based methods to produce meaningful prevalence data remains, and presents considerable potential for use in population size estimation. Unfortunately, rigorous strategies for estimating the overall size of the hidden population from RDS data have been less successful, relying on simulation-based validation that fails to yield analytic insight, and generating widely varying estimates [34, 35]. While Berchenko and Frost have developed techniques that combine capture-recapture methods with RDS, their approach requires an initial degree-biased random sample and a second (independent) respondent driven sample [36]. Their hybrid schemes have been validated through simulations, and applied in the context of several field studies [37, 38]. In comparison, the approach we develop here requires only a single RDS sample, and is evaluated through both mathematical proofs and simulation experiments.

Other specialized methods have been developed to address size estimation for hidden populations, including capture-recapture procedures (sometimes called mark-recapture or multiplier procedures) [39, 40] and network scale-up methods (NSUM) [41]. Multiplier schemes typically use a sample of the hidden population and some external, often institutional knowledge-base (e.g. arrest records or hospital admissions) for estimation purposes [14, 42]. In these

methods, two assumptions must generally be met: (i) the sample is representative of the hidden population at large, and (ii) everyone in the hidden population is equally likely to be "captured" in the official statistics being used [43]. While representativeness can sometimes be assumed (as in the case of RDS), it is often difficult to establish the uniformity of the capture statistics. Frankly stated, police arrests and hospital admissions can seldom be assumed to draw randomly from the hidden population. Further, capture-recapture/multiplier methods often require that the sample be identifiable in the institutional record, implying that expectations of anonymity on the part of sample respondents be abandoned. When working with hidden and highly stigmatized populations, such a sacrifice can be highly detrimental to both recruitment and informant reliability [44].

Network scale-up methods are also used to establish the size of hidden populations, though work in this area remains at an early stage. Here members of the entire population (ambient plus hidden) are asked to report on the number of known associates who fit the hidden population criteria [45, 46]. This approach has the advantage of being employable under ordinary random sampling conditions that can make use of known sampling frames (i.e. mail surveys and/or random digit dialing) [47]. However, the technique requires that ordinary people know whom among their associates fit the criteria for inclusion in the hidden category [48, 49]. Such an assumption faces objections in many of the situations in which we might wish to apply the technique, as when we seek to estimate the size of populations of PWID or sex workers. In these types of settings, individuals from the hidden population may go to great lengths to hide their membership status from friends and associates. Such effects inject "transmission error" into NSUM calculations, a quantity that is difficult to both detect and measure.

In previous work, we presented a novel capture-recapture methodology for estimating the size of a hidden population from an RDS sample [50], referred to there as the "telefunken" method. The method could be easily integrated into a conventional RDS framework, allowing researchers to continue to take advantage of the wide body of work on RDS and its ability to yield reliable prevalence estimates. The method was adopted experimentally in the context of efforts to collect data on commercially sexually exploited children [51] and, later, users of methamphetamine [52]. Both these studies made use of RDS and took place in New York City. Subsequent implementations of the technique provided further evidence of its effectiveness and ease of implementation [34]. The *telefunken* method was so named because its application entailed asking each RDS respondent to report on others in the population known to them by providing an encoding of their associates' telephone number and demographic features (note that the technique is in no way related to the German apparatus company, Telefunken). In taking this approach, the method avoided reliance on official statistics (as needed in scale-up methods), and the requirement of drawing two independent samples (as needed by capture-recapture methods). Each individual's code was created by considering a protocol-specified number of digits of their phone number, in order from last to first, and encoding each digit as 0/1 based on whether it was even or odd, and again 0/1 based on whether it was low (0-4) or high (5-9); in this manner, each subject and associate was "identified" by means of a multibit binary code. This many-to-one encoding allowed for ongoing anonymity for both respondents and their reported associates, while enabling the matching of contacts across numerous respondent interviews. In essence, the telefunken method represents a "one-step" approach which lifts many assumptions normally associated with other capture-recapture methods, and can be achieved using a single RDS sample from the hidden population. If shown to be effective, such an approach lends simplicity and greater cost-effectiveness to the size estimation procedure, potentially allowing for widespread application.

Concerning the issue of anonymity, independently and in roughly the same time period, Fellows put forward a general framework of Privatized Network Sampling (PNS) design [53].

PNS addresses two of the major concerns with regard to RDS data, namely the assumption that coupons are passed at random among alters, and that subjects can accurately report the number of alters that they have. As PNS is closely related to RDS, the standard RDS estimators may be used on data collected with a PNS design.

Given the growing interest in telefunken and PNS-like techniques [26, 34, 54], this paper aims to provide a systematic exposition of its strategy for one-step, anonymity preserving, network-based population size estimation. In what follows we formally describe the technique, analyze its mathematical properties, and validate its performance through simulations under a variety of implementation conditions. The simulations show considerable promise for the technique in scenarios normally associated with research among "hidden populations". Limitations and next steps toward validation/extension are discussed at the end of the paper.

2 Background

Current network size estimation methods are based on quantifying the "repetition" or overlap observed across multiple samples [55]—where the category of objects sampled may be nodes, edges, distances, paths, motifs, or substructures [56, 57], depending on the specific approach in question.

- Node sampling methods often begin by taking independent uniform random samples of the population. In interpreting the overlap between samples [58, 59], these methods are based on the same principle as the well-studied "Coupon collector's problem" from probability theory, for which maximum likelihood estimators and conservative confidence intervals are well known [60]. This classic method considers two uniform independent random samples [61]; in ecology, the method is often referred to as the "mark and recapture" protocol. Within a population V, the protocol first selects a uniform random "capture" sample $S \subseteq V$, and then a second (and independent) uniform random "recapture" sample $R \subseteq V$. From independence assumptions one infers that

$$\frac{|V|}{|S|} \approx \frac{|R|}{|S \cap R|} \tag{1}$$

and hence

$$|V| \approx \frac{|S| \cdot |R|}{|S \cap R|}. \tag{2}$$

The right-hand-side expression in (2) is known as the Lincoln-Peterson estimator [62, 63]. Many extensions and improvements to this classical technique have been developed, such as those making use of weighted sampling techniques [64], or sampling that is biased by the degree distribution of network nodes [65].

- Edge sampling approaches to population size estimation have also been developed [66–68]. These methods not only consider a sampled set of nodes, but also elicit a sample of their network neighbors. While edge sampling encounters problems associated with a bias toward high degree nodes, these methods offer potential gains in efficiency in dense graphs and where independent random sampling of nodes is restricted.

- Lastly, sampling via random walks represents a practical approach that is commonly used in estimating the size of social networks. Random walk methods start from an arbitrary node, then move to a neighboring node uniformly at random, and iterate. A typical random walk visits every node with a frequency proportional to its degree, but this bias can be quantified

by Markov Chain analysis, and corrected to enable the derivation of an estimate of graph size from the frequency with which sampled nodes appear (and reappear) during the walk process. Random walk methods have largely used a sampling with replacement model, which may, in theory, introduce bias in estimates when the (fractional) size of the sample is large [24, 69]; however, there is some recent experimental evidence that such concerns may be overstated [70]. These methods are widely used to measure the size of online social networks, and are frequently employed in conjunction with a variety of web crawler data [71–75].

The approach developed here is inspired by and builds on several of the above strategies, including random walks and edge elicitation. An outline of this paper follows: In Section 3.1, we present a population estimator for uniform random samples. This estimator is extended for respondent-driven samples in Section 3.2. The two estimators are evaluated over a broad range of graph families (see Subsection 4.1) using a general experimental framework (see Subsection 4.2). The experimental results are presented in Sections 4.3 and 4.4. In Section 4.5, we adapt the estimators for use in networks with clustering, showing in Section 4.6 that the revised schemes continue to perform well on synthetic networks. In Section 5, we extend the network size estimation schemes to allow for protection of subject privacy. These anonymity-preserving extensions are evaluated through simulation experiments in Sections 5.2 and 5.3. The impact of non-uniformities is assessed in Section 6, with special consideration of degree bias in RDS seed selection, and bottlenecking due to community structure. The performance of the proposed estimators is evaluated on a real-world network in Section 7. Finally, discussion and limitations are presented in Section 8.

3 New population size estimators

We seek to generalize the Lincoln-Peterson framework of overlapping capture and recapture sets (2) to the context of networked populations, and describe it formally in the language of graphs. The following definition provides graph-theoretic notations which will be necessary in order to precisely define the proposed sampling and estimation processes.

Definition 1. *Let $G = (V, E)$ be a graph. For each $v \in V$, denote the degree of v in G as $d(v)$. Given $A \subseteq V$, denote the (arithmetic) mean degree of vertices in A as*:

$$\bar{d}(A) \coloneqq \frac{1}{|A|}\sum_{v\in A} d(v) \tag{3}$$

and the (harmonic) mean degree of vertices in A as

$$\tilde{d}(A) \coloneqq \frac{|A|}{\sum_{v\in A}\frac{1}{d(v)}}. \tag{4}$$

noting that the latter is more robust against the presence of high-degree outliers. If $H = (S, F)$ is a subgraph on $S \subseteq V$ with edge set $F \subseteq E \cap (S \times S)$, the "free neighborhood" of u (in G modulo H) is defined as

$$N(u, F) \coloneqq \{v \mid (u, v) \in E \setminus F\} \subseteq V. \tag{5}$$

Note that when G is allowed to have parallel edges (as is the case when it is obtained through configuration graph sampling), then $N(u, F)$ may be a multiset. The "free ends" of S (in G modulo H) are taken to be the disjoint union (multiset)

$$R(S, F) \coloneqq \coprod_{u\in S} N(u, F) \subseteq V \tag{6}$$

and the "matches" of S (in G modulo H) are taken to be the disjoint union (multiset)

$$M(S,F) \coloneqq \coprod_{u\in S}(N(u,F)\cap S) \subseteq V. \tag{7}$$

We denote the respective cardinalities of these multisets as

$$\langle R(S,F)\rangle \coloneqq \sum_{u\in S}|N(u,F)|$$
$$\langle M(S,F)\rangle \coloneqq \sum_{u\in S}|N(u,F)\cap S|.$$

Notation 1. *In the arguments that follow, graph-theoretic quantities (such as those formalized in Definition 1) will sometimes be considered simultaneously in the context of more than one graph—e.g. $G_1 = (V_1, E_1)$, and $G_2 = (V_2, E_2)$. To avoid ambiguity in such settings, we will make the context clear by appending the graph as a parameter—e.g. the arithmetic mean degree of vertices in G_1 is denoted $\bar{d}(V_1; G_1)$, while the harmonic mean degree of vertices in G_2 is expressed as $\tilde{d}(V_2; G_2)$.*

Notation 2. *Whenever we are considering a multiset X, we will denote to its multiset cardinality as $\langle X\rangle$, while its set cardinality will be written as $|X^*|$. For example, if $X = \{1, 1, 2, 8, 8, 8\}$ then $\langle X\rangle = 6$, while $|X^*| = 3$.*

Definition 2. *Given multisets of vertices $A, B \subseteq V$ we denote their characteristic functions as $\chi_A, \chi_B : V \to \mathbb{N}$ and define the multisets $A\backslash B$, $A \cap B$, $A \cup B$ by the respective characteristic functions*

$$\chi_{A\backslash B}, \chi_{A\cap B}, \chi_{A\cup B} : V \to \mathbb{N}$$

where for each $v \in V$

$$\chi_{A\backslash B}(v) \coloneqq \max\{0, \chi_A(v) - \chi_B(v)\}$$
$$\chi_{A\cap B}(v) \coloneqq \min\{\chi_A(v), \chi_B(v)\}$$
$$\chi_{A\cup B}(v) \coloneqq \chi_A(v) + \chi_B(v).$$

We say that $A \subseteq B$ are multisets, if $\forall v \in V$, we have $\chi_A(v) \leq \chi_B(v)$.

3.1 Population size from a uniform random sample

With the formalisms of Definition 1 in place, we can define the estimator n_1, which, given a uniform random subset of vertices $T \subseteq V$, yields an estimate of $|V|$.

Definition 3. *Given a graph $G = (V, E)$ and $T \subseteq V$, define*

$$n_1(T) \coloneqq \frac{|T| \cdot \langle R(T,\emptyset)\rangle}{\langle M(T,\emptyset)\rangle}. \tag{8}$$

Lemma 1 shows that as the sample size grows, n_1 converges to $|V|$.

Lemma 1. *Let $G = (V, E)$ be a graph and let $T_1 \subseteq T_2 \subseteq T_3 \subseteq \ldots \subseteq V$ be an ascending chain converging to $\bigcup_{i=1}^{\infty} T_i = V$. Then*

$$\lim_{i\to\infty} \frac{n_1(T_i)}{|V|} = 1.$$

Proof. Put $R_i \coloneqq R(T_i, \emptyset)$, $M_i \coloneqq M(T_i, \emptyset)$, and $\Delta_i \coloneqq R_i\backslash M_i$. Note that $R_1 \subseteq R_2 \subseteq R_3 \subseteq \ldots$ and $M_1 \subseteq M_2 \subseteq M_3 \subseteq \ldots$ are ascending chains of multisets, and $M_i \subseteq R_i$ ($i = 1, 2, \ldots$). Suppose

$u \in \Delta_i$ and $\chi_{R_i}(u) = a$; clearly $0 < a \leq d(u)$. Then since the ascending chain $(T_i)_{i=1,2,\ldots}$ converges to V, there exists a least $j_0 > i$ for which $\chi_{M_j}(u) = d(u)$ and therefore $\chi_{\Delta_j}(u) = 0$ for all $j \geq j_0$. It follows that

$$\bigcap_{i=1}^{\infty} R_i \backslash M_i = \emptyset$$

where multiset intersection and difference are as described in Definition 2, and thus

$$\lim_{i\to\infty} \frac{\langle R_i \rangle}{\langle M_i \rangle} = 1$$

which implies $\lim_{i\to\infty} n_1(T_i)/|T_i| = 1$, completing the proof.

The next proposition gives sufficient conditions under which uniform random samples $T \subseteq V$ produce consistent estimates $n_1(T) \sim |V|$ when $|V|$ is large. Concrete realizations of these conditions are presented in Corollary 1.

Proposition 1. *For $n = 1, 2, \ldots$ let $G_n = (V_n, E_n)$ be a graph on $|V_n| = f(n)$ vertices, where $f(n)$ grows unboundedly. Let $c_n \in (0, 1]$ and take $T_n \subseteq V_n$ to be a subset of size $|T_n| = \lfloor c_n \cdot f(n) \rfloor$ selected using uniform random sampling in V_n. If $c_n \cdot f(n)$ diverges as n goes to infinity while*

$$c_n^2 \cdot \bar{d}(V_n) \longrightarrow \Theta_1 \tag{9}$$

for some finite constant $\Theta_1 > 0$, then $\frac{n_1(T_n)}{|V_n|}$ necessarily converges to 1.

Proof. Define random variables

$$\bar{R}_n \coloneqq \frac{1}{f(n)} \langle R(T_n, \emptyset) \rangle = \frac{1}{f(n)} \sum_{u \in T_n} d(u) \tag{10}$$

$$\bar{M}_n \coloneqq \frac{1}{f(n)} \langle M(T_n, \emptyset) \rangle. \tag{11}$$

For uniform random $u \in V_n$, $E[d(u)] = \bar{d}(V_n)$. Since $|T_n| = \lfloor c_n \cdot f(n) \rfloor$ diverges, the law of large numbers and linearity of expectation imply that as n tends to infinity

$$\langle R(T_n, \emptyset) \rangle = \sum_{u \in T_n} d(u) \xrightarrow{p} \sum_{u \in T_n} \bar{d}(V_n) = |T_n| \cdot \bar{d}(V_n) \tag{12}$$

and thus

$$c_n \cdot \bar{R}_n = \frac{1}{f(n)} \langle R(T_n, \emptyset) \rangle \xrightarrow{p} c_n \cdot \frac{1}{f(n)} \cdot |T_n| \cdot \bar{d}(V_n) = c_n^2 \cdot \bar{d}(V_n) \xrightarrow{p} \Theta_1. \tag{13}$$

Now for each $u \in T_n$ we have $E[\langle N(u, F_n) \cap T_n \rangle] = d(u) \cdot |T_n|/f(n)$. Again, by the law of large numbers and linearity of expectation, as n tends to infinity

$$\bar{M}_n \xrightarrow{p} \bar{R}_n \cdot \frac{|T_n|}{f(n)} = \bar{R}_n \cdot c_n \xrightarrow{p} \Theta_1. \tag{14}$$

Considering (13) and (14) as preconditions of Slutsky's theorem [76], we conclude:

$$\frac{n_1(T_n)}{f(n)} = \frac{1}{f(n)} \cdot \frac{c_n \cdot f(n) \cdot \bar{R}_n}{\bar{M}_n} \xrightarrow{d} \frac{\operatorname{plim}_{n\to\infty} c_n \cdot \bar{R}_n}{\operatorname{plim}_{n\to\infty} \bar{M}_n} = \frac{\Theta_1}{\Theta_1} = 1.$$

The correspondence between Eq (8) in Definition 3 and our previous *telefunken* estimator is clear [77]. In addition, Eq (8) demonstrates a parallel structure with the Lincoln-Peterson estimator shown in expression (2): T represents the first assay (set); $R(T, \emptyset)$ stands for the second assay (a multiset); the multiset $M(T, \emptyset)$ is the subpopulation of the first assay that is recaptured by the second assay. Of course, in the present setting, the second assay $R(T, \emptyset)$ is far from independent of the first assay T, since the two sets are intrinsically linked through the network geometry of G. Nevertheless, the fact that T is a random subset of V is enough to neutralize the impact of this non-independence and enable consistent estimation of population size.

Corollary 1. *Several special cases of Proposition 1 are of interest. In each of these cases, it is straightforward to verify that as n goes to infinity,* $c_n \cdot f(n)$ *diverges, while* $c_n^2 \cdot \bar{d}(V_n)$ *tends to some finite strictly positive constant:*

- *When* $f(n) = O(n)$, $c_n = O(1)$ *is a constant, and* $\bar{d}(V_n) = O(1)$ *is a constant. In this case, we have a family of graphs of increasing size and constant average degree, in which we are taking uniform random samples whose size is a constant proportion of the entire population.*
- *When* $f(n) = O(n)$, $c_n = O(g(n)/n)$, *and* $\bar{d}(V_n) = O(n^{1-\epsilon}/g(n)^2)$, *where* $g(n)$ *is a function which diverges, and* $\epsilon > 0$ *is a constant. For example, if we take* $g(n) = n^{\epsilon}$, *then* $c_n = O(1/n^{1-\epsilon})$, *and* $\bar{d}(V_n) = O(n^{1-3\epsilon})$. *As* ϵ *tends to 0, we approach a family of graphs of increasing size and linear average degree, in which we are taking uniform random samples of an absolute constant size. This special limit case is manifested by Erdős-Rényi graphs* [78].

3.2 Population size from a respondent-driven sample

Although the n_1 estimator shows robust performance under uniform random sampling (see Section 4.3), random sampling is seldom a feasible strategy with hidden populations. As discussed above, sampling hard-to-reach populations presents considerable practical challenges [55], and many current surveys of hidden populations have come to depend on a tracked "peer referral" process known as respondent driven sampling [21].

For purposes of estimation, we consider a respondent-driven sample to be a random variable based on several parameters: an underlying networked population $G = (V, E)$, a specified number of seeds $|D|$, the number of recruiting coupons c to be given to each subject, and the target sample size r. In our simulation experiments, the sampling procedure begins by randomly choosing $|D|$ initial "seed" subjects in the network. For most of this paper, seeds are selected uniformly at random, though later, in Section 6, we will report on the differential impact of non-uniform RDS seed selection—specifically, seed selection that is biased by ego network size or restricted by the presence of community structures. Each seed subject is given c recruiting coupons and asked to participate in a "referral" process by distributing these among their study-eligible peers. Each subject v succeeds in recruiting between 0 and $\min\{c, d(v)\}$ individuals from their ego network, with the precise number being determined stochastically according to a specified distribution δ_R on $\{0, 1, \ldots, c\}$. Each referred peer is assumed to come in for their interview at a time that is offset from their recruiter's interview by an amount that random and exponentially distributed with rate λ_W. When one or more of the recruited peers come in for interview with the coupon given to them by their recruiter, they too are given c coupons and asked to participate in the referral process. The scheme proceeds recursively in this manner using a finite number of $3r$ depletable coupons, until all r individuals have been recruited and interviewed. If (and whenever) the referral process stalls before r subjects have been interviewed, a new seed is recruited. Participation incentives are arranged to

ensure that no subject will be the recipient of more than one coupon, and thus the process results in a collection of disjoint directed trees rooted at the seeds [79]. The precise values of the RDS parameters $|D|$, c, r and implementation parameters δ_R, λ_W for our simulation experiments are detailed in Assumption 2; the stochastic process used to generate the underlying synthetic networks $G(V, E)$ on which this RDS operates is described in Section 4.1.

Given the tendency of RDS to oversample high degree nodes, issues arise when estimation techniques attempt to make use of the degree statistics of a respondent driven sample. Special steps must be taken to account for differences between the average degree of an RDS sample and the average degree of the population from which the RDS sample is drawn. The simplifying assumption below is needed for our formal proofs of the proposed estimators' performance. We emphasize that this assumption is not enforced (and is often violated) within the synthetic networks we used in our simulations, through which the proposed estimators' performance was experimentally evaluated.

Assumption 1. *Whenever we are considering $H = (S, F)$ to be a subgraph on $S \subseteq V$ obtained through an RDS process inside graph $G = (V, E)$, we will assume $\tilde{d}(S) \sim \bar{d}(V)$. This assumption is justified in prior work [20, 22], is provably true for configuration graphs [24], and is reflective of the basic fact that the harmonic mean is robust against the presence of high-degree outliers, as we may expect to face when S is obtained via a non-uniform sampling process like RDS.*

The next estimator n_2, provides an estimate $|V|$ from a respondent driven sample $S \subseteq V$.

Definition 4. *Given a graph $G = (V, E)$, a set $S \subseteq V$, and $H = (S, F)$ a subgraph with edge set $F \subseteq E \cap (S \times S)$, define*

$$n_2(S, F) := \frac{\frac{\bar{d}(S)-1}{\tilde{d}(S)} \cdot |S| \cdot \langle R(S, F) \rangle}{\langle M(S, F) \rangle} \tag{15}$$

The next proposition gives sufficient conditions under which respondent-driven samples $S \subseteq V$ produce consistent estimates $n_2(T) \sim |V|$ when $|V|$ is large.

Proposition 2. *For $n = 1, 2, \ldots$ let $G_n = (V_n, E_n)$ be a graph obtained by configuration graph sampling via degree distribution $\mathcal{D}_n$, where the vertex set size $|V_n| = f(n)$ grows unboundedly. Let $c_n \in (0, 1]$, and take $S_n \subseteq V_n$ to be a subset of size $|S_n| = \lfloor c_n \cdot f(n) \rfloor$ selected using RDS sampling in G_n from $|D_n|$ seeds chosen uniformly at random. Define the random variable*

$$\Delta_n := \frac{\bar{d}(S_n) - 1}{\tilde{d}(S_n)}.$$

Accepting Assumption 1, if $c_n \cdot f(n)/D_n$ diverges as n goes to infinity, while

$$\Delta_n^2 \cdot c_n^2 \cdot \bar{d}(V_n) = \frac{(\bar{d}(S_n) - 1)^2 \cdot c_n^2}{\tilde{d}(S_n)} \xrightarrow{p} \Theta_2 \tag{16}$$

for some finite constant $\Theta_2 > 0$, then $\frac{n_2(S_n)}{|V_n|}$ necessarily converges to 1.

Proof. Let (S_n, F_n) be a subgraph produced by an RDS sampling process in G_n, and let $T_n \subseteq V_n$ be an equal-sized set of vertices chosen by uniform random sampling, i.e. $|T_n| = |S_n|$. For random $u \in S_n$ and $v \in T_n$, as n tends to infinity

$$\frac{|N(u, \emptyset)|}{\tilde{d}(S_n)} - \frac{|N(v, \emptyset)|}{\bar{d}(T_n)} = \frac{|N(u, \emptyset)|}{\bar{d}(S_n)} - \frac{|N(v, \emptyset)|}{\bar{d}(V_n)} = \frac{|N(u, \emptyset)|}{\bar{d}(S_n)} - \frac{|N(v, \emptyset)|}{\tilde{d}(S_n)} \xrightarrow{p} 0. \tag{17}$$

where the first equality stems from the law of large numbers, and the second from Assumption

1. Now S_n is an RDS sample and hence is the disjoint union of D_n many trees. It follows that

$$\frac{|F_n|}{|S_n|} = 1 - \frac{|D_n|}{\lfloor c_n \cdot f(n) \rfloor}.$$

Since $|S_n| = \lfloor c_n \cdot f(n) \rfloor$ diverges and $c_n \cdot f(n)/D_n$ diverges, we may conclude that

$$\lim_{n\to\infty} \frac{|F_n|}{|S_n|} = 1. \tag{18}$$

We note that $|N(u, F_n)| \leq |N(u, \emptyset)|$, and incorporating (18) back into the final expression in (17), we deduce

$$\frac{|N(u, F_n)|}{\bar{d}(S_n) - 1} - \frac{|N(v, \emptyset)|}{\tilde{d}(S_n)} \xrightarrow{p} 0. \tag{19}$$

Definition 1's Eq (6) and linearity of expectation then imply that as n tends to infinity

$$\langle R(S_n, F_n) \rangle \xrightarrow{p} \frac{\bar{d}(S_n) - 1}{\tilde{d}(S_n)} \cdot \langle R(T_n, \emptyset) \rangle. \tag{20}$$

The configuration graph sampling process dictates that as n tends to infinity, for uniformly random $u \in S_n$

$$E[\langle N(u, F_n) \cap S_n \rangle] = [\bar{d}(u) - 1] \cdot \frac{\langle R(S_n, F_n) \rangle}{2|E_n|} = [\bar{d}(u) - 1] \cdot \frac{\langle R(S_n, F_n) \rangle}{\bar{d}(V_n) \cdot f(n)}.$$

Definition 1's Eq (7), expression (20), the law of large numbers, and linearity of expectation, together imply that as n tends to infinity

$$\langle M(S_n, F_n) \rangle \xrightarrow{p} \frac{\langle R(S_n, F_n) \rangle^2}{\bar{d}(V_n) \cdot f(n)} \xrightarrow{p} \frac{1}{\bar{d}(V_n) \cdot f(n)} \cdot \left[\frac{\bar{d}(S_n) - 1}{\tilde{d}(S_n)} \right]^2 \cdot \langle R(T_n, \emptyset) \rangle^2. \tag{21}$$

Define the following random variables, closely related to (10) and (11) of Proposition 1:

$$R_n^* \coloneqq \langle R(S_n, F_n) \rangle \; / \; f(n) = \Delta_n \cdot \bar{R}_n \xrightarrow{p} \Delta_n \cdot c_n \cdot \bar{d}(V_n) \tag{22}$$

$$M_n^* \coloneqq \langle M(S_n, F_n) \rangle \; / \; f(n) = \Delta_n^2 \cdot \bar{R}_n^2 / \bar{d}(V_n) \xrightarrow{p} \Delta_n^2 \cdot c_n^2 \cdot \bar{d}(V_n) \tag{23}$$

From our assumptions on the convergence of $\Delta_n^2 \cdot c_n^2 \cdot \bar{d}(V_n)$, we see that as n tends to infinity

$$\Delta_n \cdot c_n \cdot R_n^* = \Delta_n^2 \cdot c_n^2 \cdot \bar{d}(V_n) \xrightarrow{p} \Theta_2 \tag{24}$$

$$M_n^* \xrightarrow{p} \Theta_2 \tag{25}$$

Considering (24) and (25) as preconditions of Slutsky's theorem [76], we conclude:

$$\frac{n_2(S_n)}{f(n)} = \frac{1}{f(n)} \cdot \frac{\Delta_n \cdot c_n f(n) \cdot R_n^*}{M_n^*} \xrightarrow{d} \frac{\operatorname{plim}_{n\to\infty} \Delta_n \cdot c_n \cdot R_n^*}{\operatorname{plim}_{n\to\infty} M_n^*} = \frac{\Theta_2}{\Theta_2} = 1.$$

4 Evaluating the n_1 and n_2 estimators

To evaluate the proposed estimators n_1 (8) and n_2 (15), we conducted simulation experiments on samples drawn from synthetic networks using uniform and respondent-driven sampling, respectively. Underlying networks were selected by configuration sampling techniques [80–82] relative to Lognormal, Poisson, and Exponential distributions. We also considered Barabási-Albert graphs [83], and Erdős-Rényi graphs [78].

4.1 Synthetic networks

The tendency of RDS to over-recruit high degree nodes is well known, and readily evidenced in experiments on idealized topologies. Attempts to model peer-referral or "snowball" recruitment processes point to the fact that the degree distribution of nodes can influence the performance of mean estimators [84], suggesting Bayesian approaches which make use of degree distribution data in the derivation of population size estimates [35, 85]. To validate the n_1 and n_2 estimators against a wide range of possible topologies, five idealized families of random graphs were used to perform initial experiments. In later sections, we take up the issue of clustering (Section 4.5), anonymity (Section 5), non-uniformity in the seed selection (Section 6), and performance on a real-world network (Section 7).

In what follows, configuration graphs were sampled (relative to a specified degree distribution) by first attaching the prescribed number of free half-edges to each node. Pairs of free half-edges were then chosen uniformly at random and bound together to form an edge, repeatedly, until no free half-edges remain. Note that this sampling process may yield graphs that have multiple parallel edges and self loops.

Definition 5. *Given a set V with $|V| = n$, for each $\lambda \in \mathbb{R}$, $\lambda > 1$, let distributions $\mathcal{D}_{\mathcal{L}(\lambda)}$, $\mathcal{D}_{\mathcal{P}(\lambda)}$, $\mathcal{D}_{\mathcal{X}(\lambda)}$, and $\mathcal{D}_{\mathcal{R}(\lambda)} : V \to \mathbb{N}$ be defined such that for each $v \in V$:*

- $\mathcal{D}_{\mathcal{L}(\lambda,n)}(v) = 1 + X$ *where X is a Lognormal random variable with mean $\lambda - 1$ and standard deviation 1.*
- $\mathcal{D}_{\mathcal{P}(\lambda,n)}(v) = 1 + X$ *where X is a Poisson random variable with rate parameter $\lambda - 1$.*
- $\mathcal{D}_{\mathcal{X}(\lambda,n)}(v) = 1 + X$ *where X is an Exponential random variable with mean $\lambda - 1$.*

Corresponding to each of the three distributions above, let $\mathcal{L}(\lambda, n)$, $\mathcal{P}(\lambda, n)$, $\mathcal{X}(\lambda, n)$, $\mathcal{R}(\lambda, n)$ be the sample spaces of configuration graphs $G = (V, E)$ where $|V| = n$. Note that a random graph drawn from these sample spaces will have expected mean vertex degree $E[\bar{d}(V)] = \lambda$.

Definition 6. *For each $\lambda \in \mathbb{R}$, $\lambda > 1$, let $\mathcal{B}(\lambda, n)$ be the sample space of n-vertex Barabási-Albert graphs $G = (V, E)$. Each such graph is the final output of a process which produces a sequence of graphs $G^i = (V^i, E^i)$ on $V^i := \{v_1, \ldots v_i\}$ with $\lambda \leq i \leq n$. The initial graph $G^\lambda = (V^\lambda, E^\lambda)$ is taken to be the complete graph on λ vertices, i.e. $E = V^\lambda \times V^\lambda$. At each stage $i > \lambda$ of the process, node v_i ($\lambda < i \leq n$) connects to a random number*

$$\Delta_i := |E_i \backslash E_{i-1}| = \begin{cases} \lfloor \lambda/2 \rfloor & \text{with probability } 1 + \lfloor \lambda \rfloor - \lambda \\ 1 + \lfloor \lambda/2 \rfloor & \text{otherwise.} \end{cases}$$

of pre-existing nodes $\{p_{i,1}, \ldots p_{i,\Delta_i}\} \subseteq V^{i-1}$. This set is constructed by sequential sampling without replacement, i.e. as $l = 1, \ldots, \Delta_i$, each of the candidates $w \in C_{i,l} := V^{i-1} \backslash \{v_{i,1}, \ldots v_{i,l-1}\}$ is chosen with a probability that reflects degree-biased preferential attachment

$$Prob(p_{i,l} = w) = \frac{1 + d(w; G^{i-1})}{\sum_{w' \in C_{i,l}} 1 + d(w'; G^{i-1})}.$$

Here $d(w; G^{i-1})$ denotes the degree of vertex w in graph $G^{i-1} = (V^{i-1}, E^{i-1})$. The final member of the resulting sequence $G^n = (V^n, E^n)$ is output as the sampled graph. Note that if $n \gg \lambda$, the process above results in a graph $G = (V, E)$, sampled from $\mathcal{B}(\lambda, n)$, and having expected mean vertex degree $E[\bar{d}(V)] \sim \lambda$.

Definition 7. *For each $\lambda \in \mathbb{R}, \lambda > 0$, let $\mathcal{E}(\lambda, n)$ be the sample space of n-vertex Erdős-Rényi graphs $G = (V, E)$, where $E \subseteq V \times V$ is a random subset constructed uniformly at random by taking:*

$$Prob((u, v) \in E) = \begin{cases} \lambda/(n-1) & u \neq v \\ 0 & u = v \end{cases}$$

for each $(u, v) \in V \times V$. Note that a random graph $G = (V, E)$ drawn from $\mathcal{E}(\lambda, n)$ will have expected mean vertex degree $E[\bar{d}(V)] \sim \lambda$.

4.2 Experimental framework

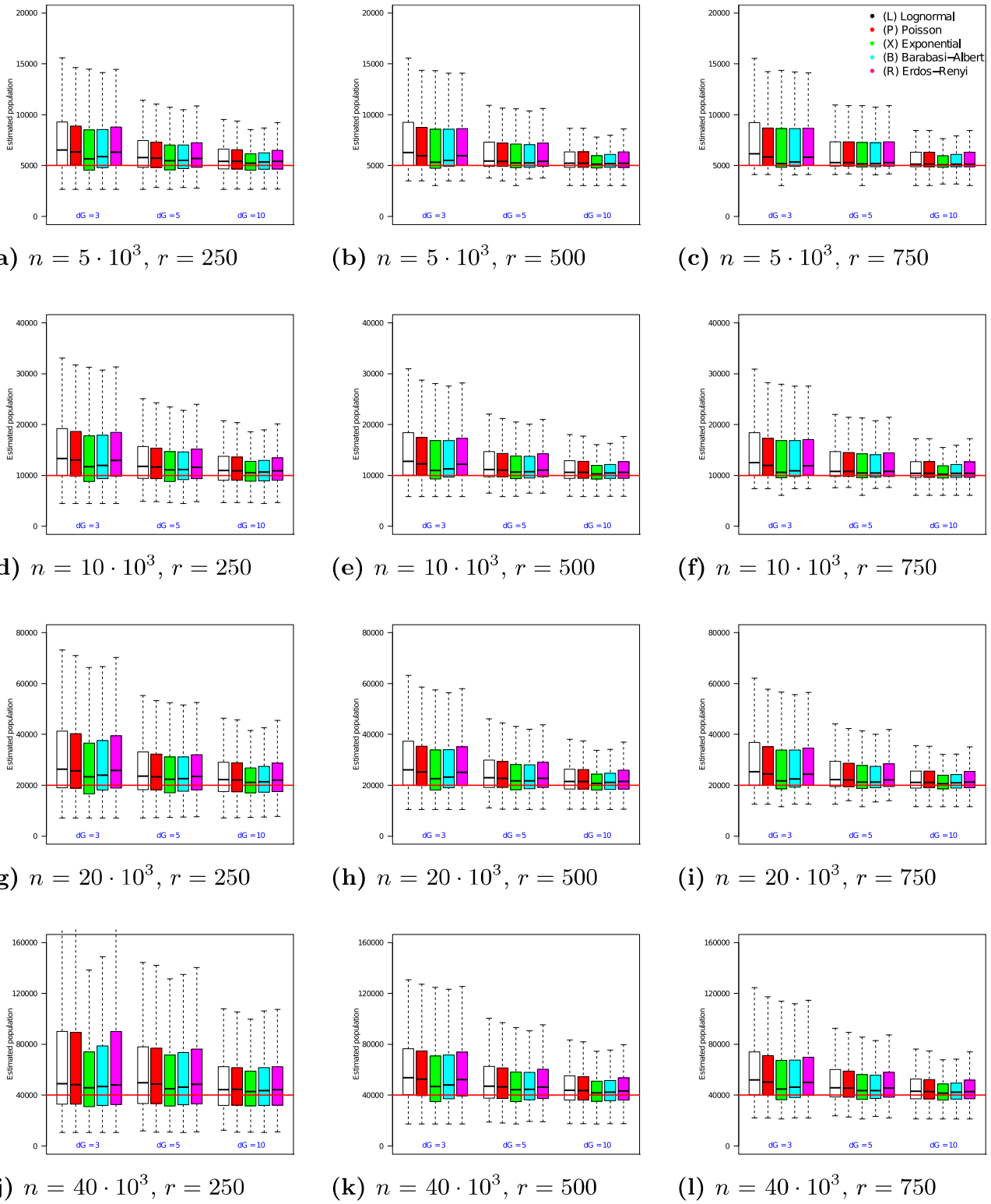
For each of the 5 families $\mathcal{L}(\lambda, n), \mathcal{P}(\lambda, n), \mathcal{X}(\lambda, n), \mathcal{B}(\lambda, n)$, and $\mathcal{E}(\lambda, n)$ defined in Section 4.1, we varied $\lambda = 3, 5, 10$; from each of these 15 concrete sample spaces, we used configuration graph sampling to select 30 random graphs of sizes $n = 5000$, $10K$, $20K$ and $40K$. In each of these $5 \times 3 \times 4 \times 30 = 1{,}800$ graphs, we generated 30 uniform and 30 RDS samples of size $r = 250$, 500 and 750. In this manner, a total of $1{,}800 \times 30 \times 3 \times 2 = 324{,}000$ simulations were conducted. Section 4.3 reports on simulation experiments in which n_1 was applied to uniform random samples; experiments in which n_2 was applied to respondent driven samples are presented in Section 4.4.

4.3 Evaluating n_1 on synthetic networks

The experiments here follow the framework described in Section 4.2 and use uniform random samples. The 12 graphs in Fig 1 present the performance of the n_1 estimator as the true population size n is varied from $5 \cdot 10^3$ to $40 \cdot 10^3$ (vertical axis of the grid) and the size of the uniform sample is varied from 250 to 750 (horizontal axis of the grid). In each of the 12 graphs, the x-axis varies the average degree λ from 3 to 10. For each choice of λ, the medians and quartile ranges of n_1 are given for each of the 5 graph families. Each of these is determined by 900 simulations (30 graphs times 30 uniformly drawn samples in each graph).

Fig 1 shows that as sample size increases, the medians of n_1 converge to the true population size. For example, when $n = 5 \cdot 10^3$ and $r = 250$, Exponential degree distribution graphs with $\lambda = 3$ have a median n_1 value of 5663 (a 13.3% offset from the true value of $n = 5 \cdot 10^3$). In comparison, when $r = 750$, the median for this family of graphs is 5204 (just 4.1% offset from the true value). As the sample size increases from $r = 250$ to $r = 750$, the error in the median estimate decreases by 9.2%. The benefit of increasing sample size diminishes as networks grow larger, however. For example, for a network of size $n = 40 \cdot 10^3$, increasing the sample size from $r = 250$ to $r = 750$ causes the error in the median n_1 estimate to undergo only a 2% change.

In addition, Fig 1 shows that as sample size increases, the interquartile range (IQR) of the estimates decreases. For example, when $n = 5 \cdot 10^3$ and $r = 250$, Lognormal degree distribution graphs with $\lambda = 10$ experience an interquartile range of 1950 in their n_1 estimates (35.9% of the median). In comparison, when $r = 750$, the interquartile range for this family of graphs decreases to 1425 (a 26.9% reduction). The magnitude of this effect increases as networks grow larger. For example, for a network of size $n = 40 \cdot 10^3$, increasing the sample size from $r = 250$ to $r = 750$ causes the interquartile range of the n_1 estimate to undergo a 48.6% decrease.



(a) $n = 5 \cdot 10^3$, $r = 250$ (b) $n = 5 \cdot 10^3$, $r = 500$ (c) $n = 5 \cdot 10^3$, $r = 750$

(d) $n = 10 \cdot 10^3$, $r = 250$ (e) $n = 10 \cdot 10^3$, $r = 500$ (f) $n = 10 \cdot 10^3$, $r = 750$

(g) $n = 20 \cdot 10^3$, $r = 250$ (h) $n = 20 \cdot 10^3$, $r = 500$ (i) $n = 20 \cdot 10^3$, $r = 750$

(j) $n = 40 \cdot 10^3$, $r = 250$ (k) $n = 40 \cdot 10^3$, $r = 500$ (l) $n = 40 \cdot 10^3$, $r = 750$

Fig 1. Estimator n_1 on uniform samples in populations of size $n = 5 \cdot 10^3$ to $40 \cdot 10^3$. In each box, the thick line indicates the sample median; the top of the box is the median of the upper half of the estimated values (75% quartile); the bottom of the box indicates the median of the lower half of the estimated values (25% quartile); and the whiskers indicate the full range of estimated values. No (finite) outliers were removed.

https://doi.org/10.1371/journal.pone.0195959.g001

4.4 Evaluating n_2 on synthetic networks

The experiments in this and all subsequent sections use respondent-driven samples. The precise values of the RDS parameters $|D|$, c, r and implementation parameters δ_R, λ_W are given below.

Assumption 2. *In all our experiments where RDS is used to generate samples, we take* $|D| = 7$ *random seeds drawn uniformly at random from V. Each subject was given* $c = 3$ *coupons. Depending on the experiment, the sample size r was either 250, 500, or 750. Reflecting our experiences in the field* [86], *we took the recruiting success distribution* δ_R *such that each subject had a 90% chance of recruiting 2 subjects randomly from their ego network, and a 10% chance of recruiting just 1. [Individuals with an ego network of size 1 were assumed to recruit that one individual with 100% probability, while individuals with an ego network of size 0 recruited no one]. The delay between recruiter and recruited subjects' interview times were assumed to be exponentially distributed with rate* $\lambda_W = 1$.

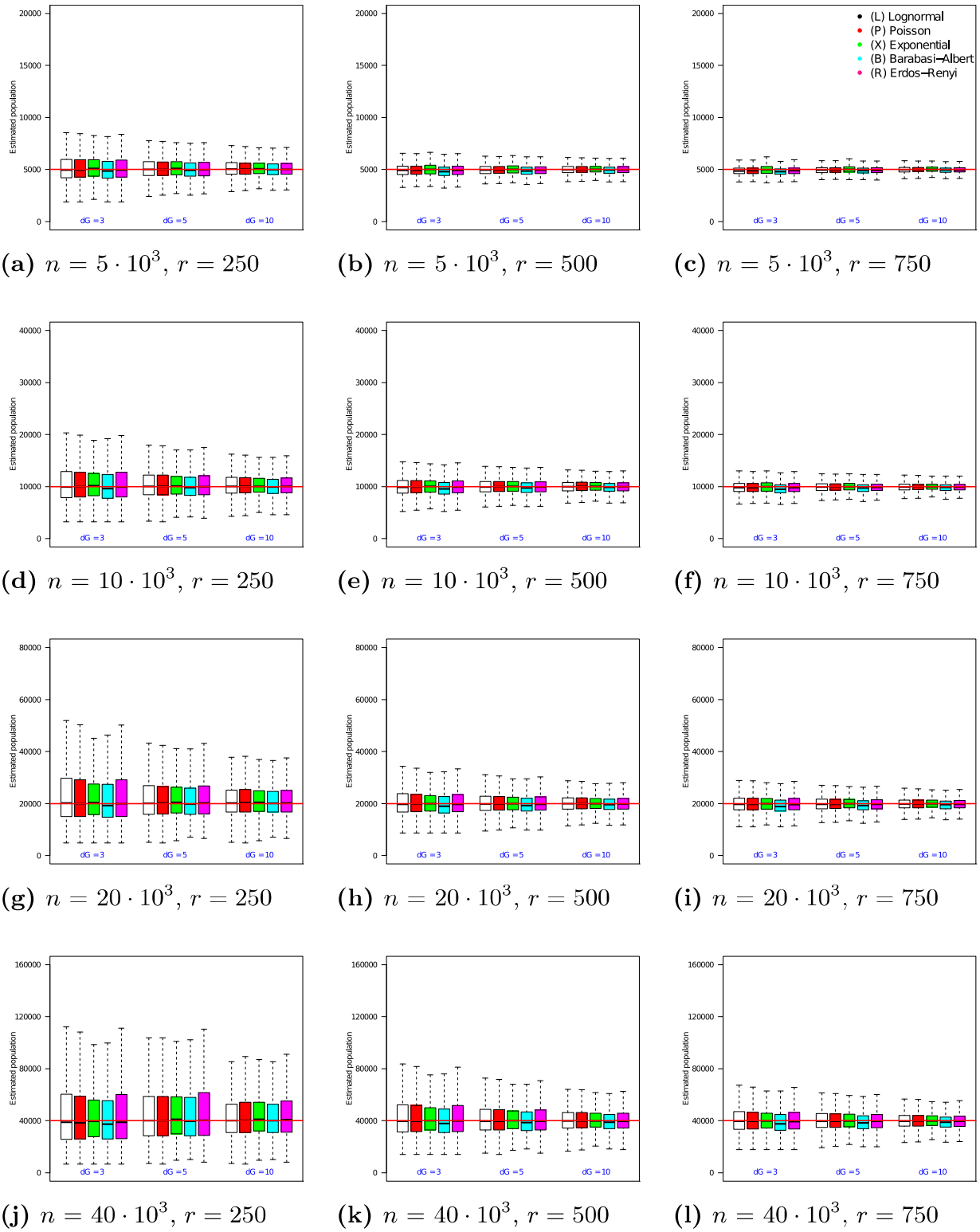
The 12 graphs in Fig 2 present the performance of the n_2 estimator as the true population size n is varied from $5 \cdot 10^3$ to $40 \cdot 10^3$ (vertical axis of the grid) and the size of the RDS sample is varied from 250 to 750 (horizontal axis of the grid). In each of the 12 graphs, the x-axis varies the average degree λ from 3 to 10. For each choice of λ, the medians and quartile ranges of n_2 are given for each of the 5 graph families. Each of these is determined by 900 simulations (30 graphs times 30 uniformly drawn samples in each graph).

Fig 2 shows that the median of n_2 converges to the true population size across a range of topologies, RDS sample sizes, and overall populations. In addition, Fig 2 shows that as sample size increases, the interquartile difference decreases. For example, when $n = 5 \cdot 10^3$ and $r = 250$, Poisson degree distribution graphs with $\lambda = 3$ experience an interquartile range of 1676 in their n_2 estimates (33.8% of the median). In comparison, when $r = 750$, the interquartile range for this family of graphs decreases to 524 (a 68.7% reduction). The magnitude of this effect decreases as networks grow larger, such that, for a network of size $n = 40 \cdot 10^3$, increasing the sample size from $r = 250$ to $r = 750$ causes the interquartile range of the n_2 estimate to undergo a 60.8% decrease. However, the total range of estimates as a proportion of the median decreases as sample size increases, indicating decreasing sample-based variance (a key concern in RDS sampling [28]).

4.5 Population size estimation in the presence of clustering

Beyond the oversampling of high degree nodes, RDS faces challenges when used in networks where network clustering is pronounced [49, 87]. While methods are available to assess the presence of clustering [25], and recent work has proposed new techniques to estimate and account for clustering from a single RDS sample [88], the effects of this phenomenon on population size estimation from RDS samples is seldom discussed. The root of the problem lies in the fact that RDS walks necessarily sample network neighborhoods. Where neighbors show high levels of network transitivity, counts of common edges will produce high numbers of "matches" that appear in the denominator of both n_1 and n_2. This will bias the estimates of overall population size derived from these estimators toward underestimation of the total network size.

In the context of random walk techniques, one approach to this problem is to *only* consider collisions among nodes that are *far away* from each other in the sampling chain when inferring a population size estimate [75]. A similar approach is taken here by considering neighbor overlap among respondents whose path distances in the RDS chains are above a specific threshold. For simplicity, here we take this threshold to be infinity, leaving the consideration of finite thresholds for consideration in future research. In short, we consider a modification of n_2 that



(a) $n = 5 \cdot 10^3$, $r = 250$ (b) $n = 5 \cdot 10^3$, $r = 500$ (c) $n = 5 \cdot 10^3$, $r = 750$

(d) $n = 10 \cdot 10^3$, $r = 250$ (e) $n = 10 \cdot 10^3$, $r = 500$ (f) $n = 10 \cdot 10^3$, $r = 750$

(g) $n = 20 \cdot 10^3$, $r = 250$ (h) $n = 20 \cdot 10^3$, $r = 500$ (i) $n = 20 \cdot 10^3$, $r = 750$

(j) $n = 40 \cdot 10^3$, $r = 250$ (k) $n = 40 \cdot 10^3$, $r = 500$ (l) $n = 40 \cdot 10^3$, $r = 750$

Fig 2. Estimator n_2 on RDS samples in populations of size $n = 5 \cdot 10^3$ to $40 \cdot 10^3$. In each box, the thick line indicates the sample median; the top of the box is the median of the upper half of the estimated values (75% quartile); the bottom of the box indicates the median of the lower half of the estimated values (25% quartile); and the whiskers indicate the full range of estimated values. No (finite) outliers were removed.

https://doi.org/10.1371/journal.pone.0195959.g002

discounts matched free ends within a single RDS sampling tree and, for purposes of estimation, only counts those matches that occur across distinct RDS trees. The next Definition introduces formalisms necessary to make this precise.

Definition 8. *Let $G = (V, E)$, take $S \subseteq V$, and let $H = (S, F)$ be a subgraph on $S \subseteq V$ with edge set $F \subseteq E \cap (S \times S)$ obtained by respondent driven sampling from a set of seeds $D \subseteq S$ where $|D| > 1$. Define the function $\gamma: S \to D$ associating each $u \in S$ with the unique seed $\gamma(u) \in D$ from which u was discovered through a sequence of referrals. For each $u \in S$, the component of u is denoted*

$$C_\gamma(u) \coloneqq \{v \mid \gamma(v) = \gamma(u)\} \subseteq S \tag{26}$$

while its complement is written $\tilde{C}_\gamma(u) \coloneqq S \backslash C_\gamma(u)$. Note that $C_\gamma(u) \cap \tilde{C}_\gamma(u) = \emptyset$. For each seed $s \in D$, we define the cross-seed matches from the $C_\gamma(u)$ component (in G modulo H) as the disjoint union (multiset)

$$X(s, F, \gamma) \coloneqq \coprod_{u \in C_\gamma(s)} (N(u, F) \cap \tilde{C}_\gamma(s)) \subseteq V \tag{27}$$

whose cardinality is denoted

$$\langle X(s, F, \gamma) \rangle \coloneqq \sum_{u \in C_\gamma(s)} |N(u, F) \cap \tilde{C}_\gamma(s)|.$$

The next estimator n_3, provides a revised estimate $|V|$ from a respondent driven sample $S \subseteq V$, discounting matches that occur within the same RDS component.

Definition 9. *Given a graph $G = (V, E)$, a set $S \subseteq V$, and $H = (S, F)$ a subgraph on $S \subseteq V$ with edge set $F \subseteq E \cap (S \times S)$. Take $D \subseteq S$ satisfying $|D| > 1$ and*

$$s_1 \neq s_2 \implies C_\gamma(s_1) \cap C_\gamma(s_2) = \emptyset.$$

Define

$$n_3(S, F, D, \gamma) \coloneqq \frac{\sum_{s \in D} \frac{\bar{d}(\tilde{C}_\gamma(s)) - 1}{\tilde{d}(S)} \cdot |\tilde{C}_\gamma(s)| \cdot \langle R(C_\gamma(s), F) \rangle}{\sum_{s \in D} \langle X(s, F, \gamma) \rangle}. \tag{28}$$

The next proposition gives sufficient conditions under which respondent-driven samples $S \subseteq V$ produce consistent estimates $n_3(T) \sim |V|$ when $|V|$ is large.

Proposition 3. *For $n = 1, 2, \ldots$, let $G_n = (V_n, E_n)$ be a graph on $|V_n| = f(n)$ vertices obtained by configuration graph sampling via degree distribution $\mathcal{D}_n$, where $f(n)$ grows unboundedly. Let $c_n \in (0, 1]$, and take $S_n \subseteq V_n$ to be a subset of size $|S_n| = \lfloor c_n \cdot f(n) \rfloor$ selected using RDS sampling in G_n from $|D_n| > 1$ seeds. Define the random variable*

$$\Delta_n \coloneqq \frac{\bar{d}(S_n) - 1}{\tilde{d}(S_n)}.$$

Accepting Assumption 1, if $c_n \cdot f(n)/D_n$ diverges as n goes to infinity, while

$$\Delta_n^2 \cdot c_n^2 \cdot \bar{d}(V_n) \cdot \frac{|D_n| - 1}{|D_n|} = \frac{(\bar{d}(S_n) - 1)^2 \cdot c_n^2}{\tilde{d}(S_n)} \cdot \frac{|D_n| - 1}{|D_n|} \xrightarrow{p} \Theta_3 \tag{29}$$

for some finite constant $\Theta_3 > 0$, then $\frac{n_3(S_n, F_n, D_n, \gamma)}{f(n)}$ necessarily converges to 1.

Proof. Since each seed $s \in D_n$ is chosen uniformly at random, and RDS recruits from all seeds concurrently, and $|S_n| = \lfloor c_n \cdot f(n) \rfloor$ diverges, for random $s \in D_n$, we know that

$$|C_\gamma(s)| \xrightarrow{p} \frac{1}{|D_n|} \cdot |S_n| = \frac{c_n \cdot f(n)}{|D_n|} \tag{30}$$

$$|\tilde{C}_\gamma(s)| \xrightarrow{p} \frac{|D_n| - 1}{|D_n|} \cdot |S_n| = \frac{|D_n| - 1}{|D_n|} \cdot c_n \cdot f(n) \tag{31}$$

$$\bar{d}(C_\gamma(s)), \quad \bar{d}(\tilde{C}_\gamma(s)) \xrightarrow{p} \bar{d}(S_n). \tag{32}$$

Combining (30) and (32), we conclude

$$\langle R(C_\gamma(s), F_n) \rangle \xrightarrow{p} \frac{\langle R(S_n, F_n) \rangle}{|D_n|}. \tag{33}$$

Sufficient reasoning about the configuration graph construction process tells us

$$\langle X(s, F_n, \gamma) \rangle \xrightarrow{p} \frac{1}{|D_n|} \cdot \langle M(S_n, F_n) \rangle \cdot \frac{|D_n| - 1}{|D_n|}. \tag{34}$$

Define the following random variables, closely related to (22) and (23) of Proposition 2:

$$\begin{aligned} R_n^\circ &:= \sum_{s \in D_n} \frac{\bar{d}(\tilde{C}_\gamma(s)) - 1}{\tilde{d}(S)} \cdot |\tilde{C}_\gamma(s)| \cdot \langle R(C_\gamma(s), F_n) \rangle / f(n) \\ M_n^\circ &:= \sum_{s \in D_n} \langle X(s, F_n, \gamma) \rangle / f(n). \end{aligned}$$

As n tends to infinity

$$\begin{aligned} R_n^\circ &\xrightarrow{p} \frac{\bar{d}(S_n) - 1}{\tilde{d}(S)} \left(\frac{|D_n| - 1}{|D_n|} \cdot c_n \cdot f(n) \right) \cdot R_n^*(S_n, F_n) \\ M_n^\circ &\xrightarrow{p} \frac{|D_n| - 1}{|D_n|} \cdot M_n^*(S_n, F_n). \end{aligned}$$

where

$$R_n^*(S_n, F_n) \xrightarrow{p} \Delta_n \cdot c_n \cdot \bar{d}(V_n)$$

as noted in (22), while

$$M_n^*(S_n, F_n) \xrightarrow{p} \Delta_n^2 \cdot c_n^2 \cdot \bar{d}(V_n)$$

as noted in (23). Thus

$$\begin{aligned} R_n^\circ &\xrightarrow{p} \Delta_n^2 \cdot c_n^2 \cdot \bar{d}(V_n) \cdot \frac{|D_n| - 1}{|D_n|} \cdot f(n) = \Theta_3 \cdot f(n) \\ M_n^\circ &\xrightarrow{p} \Delta_n^2 \cdot c_n^2 \cdot \bar{d}(V_n) \cdot \frac{|D_n| - 1}{|D_n|} = \Theta_3. \end{aligned}$$

By Slutsky's theorem [76], it follows that

$$\frac{n_3(S_n, F_n, D_n, \gamma)}{f(n)} = \frac{\frac{1}{f(n)} \cdot R_n^\circ}{M_n^\circ} \xrightarrow{d} \frac{\text{plim}_{n\to\infty} \frac{1}{f(n)} \cdot R_n^\circ}{\text{plim}_{n\to\infty} M_n^\circ} = \frac{\Theta_3}{\Theta_3} = 1. \tag{35}$$

4.6 Evaluating n_3 on synthetic networks

Prior to examining the performance of n_3 on empirical networks, we first look at its performance on the synthetic networks used to evaluate n_1 and n_2. The experiments shown in Fig 3 follow the framework described in Section 4.2 and use respondent driven samples, each obtained via an RDS process operating as specified in Assumption 2.

The 12 graphs in Fig 3 present the performance of the n_3 estimator as the true population size n is varied from $5 \cdot 10^3$ to $40 \cdot 10^3$ (vertical axis of the grid) and the size of the RDS sample is varied from 250 to 750 (horizontal axis of the grid). In each of the 12 graphs, the x-axis varies the average degree λ from 3 to 10. For each choice of λ, the medians and quartile ranges of n_3 are given for each of the 5 graph families. Each of these is determined by 900 simulations (30 graphs times 30 uniformly drawn samples in each graph).

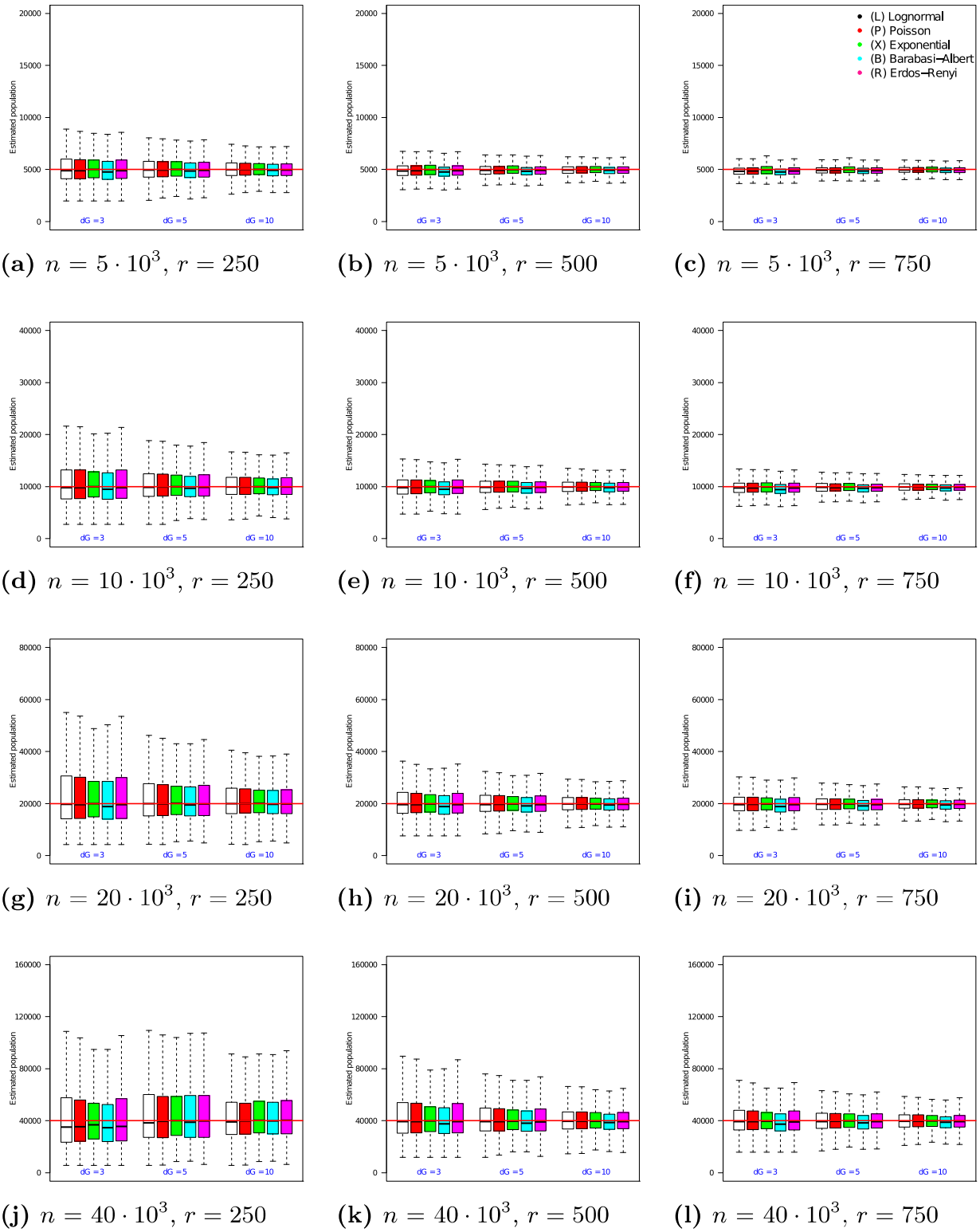
Fig 3 shows that the median of n_3 converge to the true population size, much like the performance of the n_2 estimator. In all the networks, the medians of n_3 estimates are all very close to the their true network populations, regardless the sample size, population size, and type of network topology. In addition, Fig 3 shows that as sample size increases, the interquartile range of the estimates decreases. For example, when $n = 5 \cdot 10^3$ and $r = 250$, Lognormal degree distribution graphs with $\lambda = 3$ experience a interquartile range of 1915 in their n_3 estimates (39.1% of the median). In comparison, when $r = 750$, the interquartile range for this family of graphs decreases to 604 (a 68.5% reduction). The magnitude of this effect decreases as networks grow larger. For example, in a network of size $n = 40 \cdot 10^3$, increasing the sample size from $r = 250$ to $r = 750$ causes the interquartile range of the n_3 estimate to undergo a (still sizable) 55.0% decrease.

5 Subject privacy through hashing

Significant obstacles arise in the direct application of estimators n_1, n_2, n_3 (see (8), (15), and (28), respectively). In many circumstances where RDS is used, researchers are often required to measure the sizes of stigmatized networked populations (e.g. people who inject drugs, sex workers, individuals engaged in specific types of illegal activity, etc.) and within social communities that naturally seek to remain "unidentified". In these circumstances, the membership of sets S and $R(S, F)$ is often not explicitly knowable because individuals are reluctant to unambiguously identify themselves or their social network peers.

To formalize and accommodate notions of privacy required under such circumstances within the estimation procedures described above, we assume that each individual in $V = \{v_1, v_2, \ldots, v_{|V|}\}$ has a unique ID; for simplicity we take the ID of $v_i \in V$ to be the integer i (for $i = 1, \ldots, |V|$). Towards ensuring anonymity, we imagine a *hashing* [89] function ψ: $V \to \Omega$ that assigns each individual's ID to a code in Ω. We thus follow the general framework of Privatized Network Sampling (PNS) design [53], mimicking the hash functions of telefunken-type [50].

By taking ψ to be a random (not necessarily 1-to-1) function that is difficult to invert, subjects are convinced that disclosing the hash code of an individual does not unambiguously identify the individual themselves, and so preserves their privacy.



(a) $n = 5 \cdot 10^3$, $r = 250$ (b) $n = 5 \cdot 10^3$, $r = 500$ (c) $n = 5 \cdot 10^3$, $r = 750$

(d) $n = 10 \cdot 10^3$, $r = 250$ (e) $n = 10 \cdot 10^3$, $r = 500$ (f) $n = 10 \cdot 10^3$, $r = 750$

(g) $n = 20 \cdot 10^3$, $r = 250$ (h) $n = 20 \cdot 10^3$, $r = 500$ (i) $n = 20 \cdot 10^3$, $r = 750$

(j) $n = 40 \cdot 10^3$, $r = 250$ (k) $n = 40 \cdot 10^3$, $r = 500$ (l) $n = 40 \cdot 10^3$, $r = 750$

Fig 3. Estimator n_3 on RDS samples in populations of size $n = 5 \cdot 10^3$ to $40 \cdot 10^3$. In each box, the thick line indicates the sample median; the top of the box is the median of the upper half of the estimated values (75% quartile); the bottom of the box indicates the median of the lower half of the estimated values (25% quartile); and the whiskers indicate the full range of estimated values. No (finite) outliers were removed.

https://doi.org/10.1371/journal.pone.0195959.g003

Assumption 3. *Suppose V is a set of individuals obtained via RDS referral tree F. While each $v_i \in V$ is unwilling to disclose their own ID i, and is secretive about the IDs of their peers $\{j | v_j \in N(v_i, \emptyset)\}$, they are readily willing to reveal (a) the own hash code $\psi(v_i)$; (b) the (multiset of) hash codes of their peers (outside the referral tree F):*

$$N_u^{\psi}(S,F) \coloneqq \coprod_{\substack{v \in N(u) \\ (u,v) \notin F}} \{\psi(v)\} \subseteq \Omega \tag{36}$$

and (c) their own network size $d(v_i) = \langle N_u^{\psi}(S,F) \rangle$, excluding the referral tree F.

Assumption 4. *To simplify our analysis, throughout what follows, we will assume ψ is a function chosen uniformly at random from the space of all functions from $V \to \Omega$. We will refer to such a ψ as a "random hash function" from V to Ω. The action of ψ on the V is illustrated in Fig 4. In Section 6.3, we describe ways to translate the results of this paper to settings where ψ is not a uniformly random hashing function.*

In practice, $\psi(v)$ might be an obtained by amalgamating a well-defined tuple of characteristics of v which are known to v's friends (e.g. v's gender, phone number, hair color, approximate age, racial category, etc.) and then encoding this using a cryptographic function. A related coding technique was used in our earlier work on estimating the size of the methamphetamine using population in New York City, where it was referred to as the *telefunken* code [50].

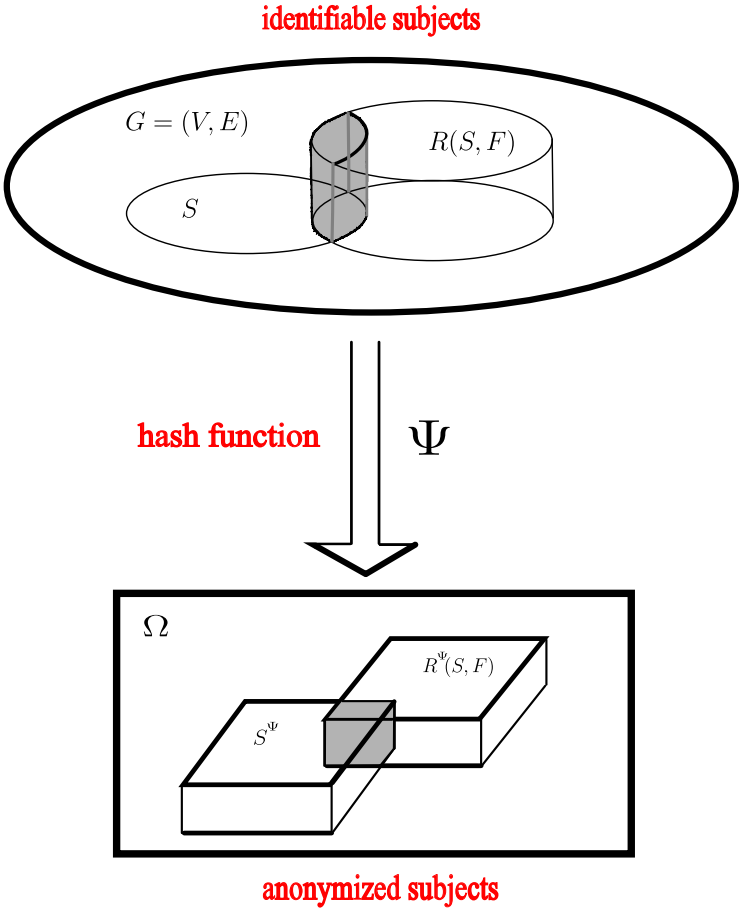


Fig 4. The action of ψ on V.

https://doi.org/10.1371/journal.pone.0195959.g004

5.1 Revised estimators incorporating hashing

We begin by "lifting" the terms introduced in the earlier Definition 1, to the hashing or PNS framework [53].

Definition 10. *Let $G = (V, E)$ be a graph, and ψ: $V \to \Omega$ a random hash function. Let $H = (S, F)$ be a subgraph on $S \subseteq V$ with edge set $F \subseteq E \cap (S \times S)$. The (multiset of) hash codes of the subjects is*

$$S^{\psi} \coloneqq \{\psi(v) \mid v \in S\} \subseteq \Omega. \tag{37}$$

*The ψ-*free ends *of S (in G modulo H) are taken to be the disjoint union (multiset)*

$$R^{\psi}(S, F) \coloneqq \coprod_{u \in S} N^{\psi}(u, F) \subseteq \Omega \tag{38}$$

*and the ψ-*matches *of (in G modulo H) are taken to be the disjoint union (multiset)*

$$M^{\psi}(S, F) \coloneqq \coprod_{u \in S} (N^{\psi}(u, F) \cap S^{\psi}) \subseteq \Omega. \tag{39}$$

We denote their respective multiset cardinalities as

$$\begin{aligned} \langle R^{\psi}(S, F) \rangle &\coloneqq \sum_{u \in S} |N^{\psi}(u, F)| \\ \langle M^{\psi}(S, F) \rangle &\coloneqq \sum_{u \in S} |N^{\psi}(u, F) \cap S^{\psi}|. \end{aligned}$$

The reader may wish to compare expressions (36), (38), *and* (39) *with the non-hashed analogues in Definition 1's expressions* (5), (6), *and* (7).

The next Lemma is foundational and justifies the proposed revised estimates n_1^{ψ}, n_2^{ψ}, and n_3^{ψ}, which will be presented subsequently.

Lemma 2. *Let $G = (V, E)$ a graph with $|V| = n'$, sampled from the space of all n'-vertex graphs by configuration sampling with respect to degree distribution $\mathcal{D}$. Let $S \subseteq V$ be an RDS sample collected as a subgraph $H = (S, F)$ be with edge set $F \subseteq E \cap (S \times S)$. Let $c \coloneqq |S|/|V|$, where $c \ll 1$. Accepting Assumption 1, take ψ: $V \to \Omega$ to be a random hash function.*

1. *Suppose $u \in S$ reports its own code $x \coloneqq \psi(u)$, the code $y \coloneqq \psi(v)$ of one of its neighbors $v \in N_u(S, F)$. If $w \in \psi^{-1}(y) \cap S$ is selected uniformly at random, and w has degree $d(w)$, then*

$$Prob(w = v) = \frac{1}{\frac{n'-1}{|\Omega|}\frac{\check{d}(S)}{(d(w)-1)} + 1}.$$

2. *For each code $y \in \Omega$, over the space of all random hash functions,*

$$E[\langle M^{\psi}(S, F) \rangle] = \hat{m}(y, n')$$

where

$$\hat{m}(y,n') \coloneqq \sum_{w\in\psi^{-1}(y)\cap S} \frac{1}{\frac{n'-1}{|\Omega|}\frac{\tilde{d}(S)}{(d(w)-1)}+1}$$
$$\hat{m}(n') \coloneqq \sum_{y\in M^{\psi}(S,F)} \hat{m}(y,n').$$

Proof. (1) Because ψ is a random function, for any $z \in \Omega$

$$E[|\psi^{-1}(z)|] = \frac{n'}{|\Omega|}.$$

The expected total number of free ends incident to some vertex in the set $\psi^{-1}(y)\backslash\{w\}$ is

$$\frac{(n'-1)(1-c)}{|\Omega|}\cdot\tilde{d}(S) + \frac{(n'-1)c}{|\Omega|}\cdot\left(\tilde{d}(S)-1\right)$$

and since $w \in S$, the expected number of free ends incident to w is $d(w) - 1$. So

$$Prob(w=v) = \frac{d(w)-1}{\frac{(n'-1)(1-c)}{|\Omega|}\cdot\tilde{d}(S) + \frac{(n'-1)c}{|\Omega|}\cdot\left(\tilde{d}(S)-1\right) + (d(w)-1)}.$$

dividing through by $d(w) - 1$, and considering $c \sim 0$, the Lemma is proved. Assertion (2) follows from (1) by linearity of expectation.

Definition 11. *If $f : \mathbb{R} \to \mathbb{R}$ is a real-valued function defined on the reals, then we denote $RootOf^{+}[f(x) = 0, x]$ to be (any one of the positive "roots") $x^* \in \mathbb{R}$ that satisfies the condition $f(x^*) = 0$, and $x^* > 0$.*

Definition 12. *Given a graph $G = (V, E)$, and ψ: $V \to \Omega$ a random hash function. Fix $S \subseteq V$, and $H = (S, F)$ a subgraph on $S \subseteq V$ with edge set $F \subseteq E \cap (S \times S)$. We define*

$$n_2^{\psi}(S,F) \coloneqq RootOf^{+}[f_2^{\psi}(n',S,F) - n' = 0, \quad n'] \tag{40}$$

where

$$f_2^{\psi}(n',S,F) \coloneqq \frac{\frac{\bar{d}(S)-1}{\bar{d}(S)}\cdot\langle S^{\psi}\rangle\cdot\langle R^{\psi}(S,F)\rangle}{\hat{m}(n')}$$

and $RootOf^{+}$ is the root operation described in Definition 11.

Definition 13. *Given a graph $G = (V, E)$, a set $S \subseteq V$, and $H = (S, F)$ a subgraph on $S \subseteq V$ with edge set $F \subseteq E \cap (S \times S)$. Let $D \subseteq S$ satisfying $|D| > 1$ and*

$$s_1 \neq s_2 \Longrightarrow C_{\gamma}(s_1) \cap C_{\gamma}(s_2) = \emptyset.$$

Take γ: $S \to D$ as described in Definition 8. The (multiset of) hash codes of vertices in the component of u are denoted

$$C_{\gamma}^{\psi}(u) \coloneqq \{\psi(v) \mid v \in C_{\gamma}(u)\} \subseteq S^{\psi} \tag{41}$$

while the codes of the complement set (inside S) are written as

$$\tilde{C}_{\gamma}^{\psi}(u) \coloneqq \{\psi(v) \mid v \in \tilde{C}_{\gamma}(u)\} \subseteq S^{\psi}.$$

Note that $C_\gamma^\psi(u) \cap \tilde{C}_\gamma^\psi(u)$ *may be non-empty. For each seed* $s \in D$*, we define the* cross-seed ψ-matches *from* $C_\gamma^\psi(s)$ *in G modulo H as the disjoint union (multiset)*

$$X^\psi(s, F, \gamma) \coloneqq \coprod_{u \in C_\gamma(s)} (N^\psi(u, F) \cap \tilde{C}_\gamma^\psi(s)) \subseteq \Omega. \tag{42}$$

The reader may wish to compare expressions (41) and (42) *with the non-hashed analogues in Definition 8's expressions* (26) and (27). We also define

$$\tilde{x}(y, s, \gamma, n') \coloneqq \sum_{w \in \psi^{-1}(y) \cap \tilde{C}_\gamma(s)} \frac{1}{\frac{n'-1}{|\Omega|} \frac{\tilde{d}(S)}{(d(w)-1)} + 1}$$

$$\hat{x}(s, F, \gamma, n') \coloneqq \sum_{y \in X^\psi(s,F,\gamma)} \tilde{x}(y, s, \gamma, n').$$

Definition 14. *Given a graph* $G = (V, E)$*, a set* $S \subseteq V$*, and* $H = (S, F)$ *a subgraph on* $S \subseteq V$ *with edge set* $F \subseteq E \cap (S \times S)$*. We define*

$$n_3^\psi(S, F) \coloneqq RootOf^+[f_3^\psi(n', S, F, D, \gamma) - n' = 0, \quad n'] \tag{43}$$

where

$$f_3^\psi(n', S, F, D, \gamma) \coloneqq \frac{\sum_{s \in D} \frac{\tilde{d}(\tilde{C}_\gamma(s))-1}{\tilde{d}(S)} \cdot \langle \tilde{C}_\gamma^\psi(s) \rangle \cdot \langle R^\psi(C_\gamma(s), F) \rangle}{\sum_{s \in D} \hat{x}(s, F, \gamma, n')}$$

and $RootOf^+$ *is the root operation described in Definition 11.*

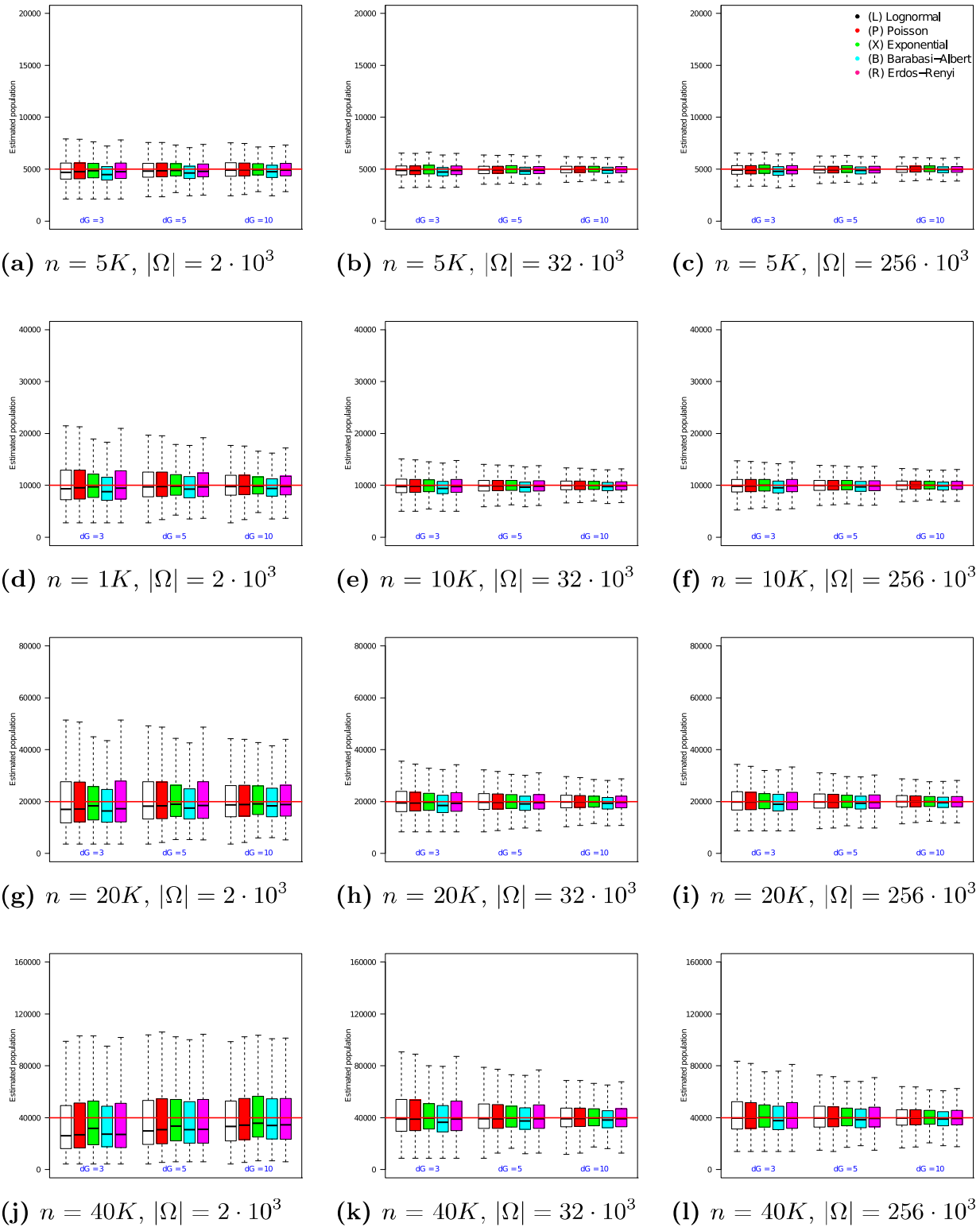
5.2 Evaluating n_2^ψ on synthetic networks

The experiments discussed here follow the framework used in prior experiments described above. Samples are derived using the RDS process operating as specified in Assumption 2. The hash space size used for the encoding of each agent's identity was varied from $|\Omega| = 2 \cdot 10^3$ to $256 \cdot 10^3$.

The 12 graphs in Fig 5 present the performance of the n_2^ψ estimator as the true population size n is varied from $5 \cdot 10^3$ to $40 \cdot 10^3$ (vertical axis of the grid), the sample size is fixed to $r = 500$ and the hash space size was varied from $|\Omega| = 2 \cdot 10^3$ to $256 \cdot 10^3$ (horizontal axis of the grid). In each of the 12 graphs, the x-axis varies the average degree λ from 3 to 10. For each choice of λ, the medians and quartile ranges of n_2^ψ are given for each of the 5 graph families. Each of these is determined by 900 simulations (30 graphs times 30 uniformly drawn samples in each graph).

Fig 5 shows that as hash space size increases, the medians of n_2^ψ converge to the true population size. For example, when $n = 5 \cdot 10^3$ and $|\Omega| = 2 \cdot 10^3$, Lognormal degree distribution graphs with $\lambda = 3$ have a median n_2^ψ value of 4705 (a 5.9% offset from the true value of $n = 5 \cdot 10^3$). In comparison, when $|\Omega| = 256 \cdot 10^3$, the median value for this family of graphs is 4901 (just 2.0% offset from the true value). As the hash space size increases from $|\Omega| = 2 \cdot 10^3$ to $|\Omega| = 256 \cdot 10^3$, the error in the median estimate decreases by 3.9%. The magnitude of this phenomenon increases as networks grow larger. For example for a network of size $n = 40 \cdot 10^3$, increasing the hash space size from $|\Omega| = 2 \cdot 10^3$ to $|\Omega| = 256 \cdot 10^3$ causes the error in the median n_2^ψ estimate to undergo a 33.9% change.

In addition, Fig 5 shows that as hash space size increases, the interquartile range of the estimates decreases. For example, when $n = 5 \cdot 10^3$ and $|\Omega| = 2 \cdot 10^3$, Poisson degree distribution



(a) $n = 5K$, $|\Omega| = 2 \cdot 10^3$ (b) $n = 5K$, $|\Omega| = 32 \cdot 10^3$ (c) $n = 5K$, $|\Omega| = 256 \cdot 10^3$

(d) $n = 1K$, $|\Omega| = 2 \cdot 10^3$ (e) $n = 10K$, $|\Omega| = 32 \cdot 10^3$ (f) $n = 10K$, $|\Omega| = 256 \cdot 10^3$

(g) $n = 20K$, $|\Omega| = 2 \cdot 10^3$ (h) $n = 20K$, $|\Omega| = 32 \cdot 10^3$ (i) $n = 20K$, $|\Omega| = 256 \cdot 10^3$

(j) $n = 40K$, $|\Omega| = 2 \cdot 10^3$ (k) $n = 40K$, $|\Omega| = 32 \cdot 10^3$ (l) $n = 40K$, $|\Omega| = 256 \cdot 10^3$

Fig 5. Estimator n_2^{ψ} on RDS samples of size $r = 500$ with $|\Omega| = 2 \cdot 10^3$ to $256 \cdot 10^3$. In each box, the thick line indicates the sample median; the top of the box is the median of the upper half of the estimated values (75% quartile); the bottom of the box indicates the median of the lower half of the estimated values (25% quartile); and the whiskers indicate the full range of estimated values. No (finite) outliers were removed.

https://doi.org/10.1371/journal.pone.0195959.g005

graphs with $\lambda = 3$ experience a interquartile range of 1522 in their n_2^{ψ} estimates (32.0% of the median). In comparison, when $|\Omega| = 256 \cdot 10^3$, the interquartile range for this family of graphs decreases to 793 (a 47.9% reduction). The magnitude of this effect increases as networks grow larger. For example for a network of size $n = 40 \cdot 10^3$, increasing the hash space size from $|\Omega| = 2 \cdot 10^3$ to $|\Omega| = 256 \cdot 10^3$ causes the interquartile range of the n_2^{ψ} estimate to undergo a 42.1% decrease.

5.3 Evaluating n_3^{ψ} on synthetic networks

A second set of experiments shows the performance of the n_3^{ψ} performance under identical hashing conditions used to test n_2^{ψ}. These experiments also follow the framework described in Section 4.2 and use samples derived from an RDS process operating as specified in Assumption 2. The hash space size was varied from $|\Omega| = 2 \cdot 10^3$ to $256 \cdot 10^3$.

The 12 graphs in Fig 6 present the performance of the n_3^{ψ} estimator as the true population size n is varied from $5 \cdot 10^3$ to $40 \cdot 10^3$ (vertical axis of the grid), the sample size is fixed to $r = 500$ and the hash space size was varied from $|\Omega| = 2 \cdot 10^3$ to $256 \cdot 10^3$ (horizontal axis of the grid). In each of the 12 graphs, the x-axis varies the average degree λ from 3 to 10. For each choice of λ, the medians and quartile ranges of n_3^{ψ} are given for each of the 5 graph families. Each of these is determined by 900 simulations (30 graphs times 30 uniformly drawn samples in each graph).

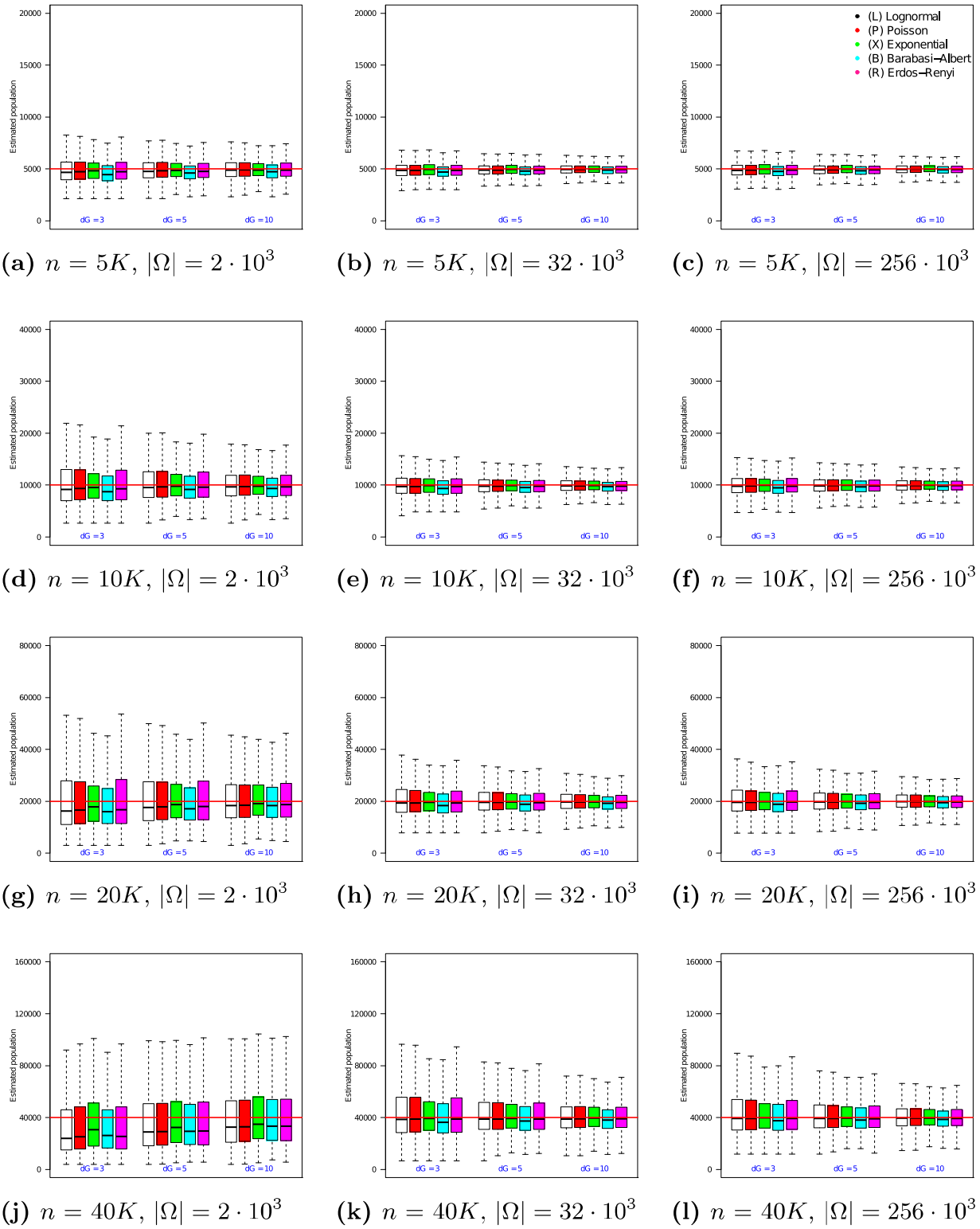
Fig 6 shows that as hash space size increases, the medians of n_3^{ψ} converge to the true population size. For example, when $n = 5 \cdot 10^3$ and $|\Omega| = 2 \cdot 10^3$, Lognormal degree distribution graphs with $\lambda = 3$ have a median n_3^{ψ} value of 4667 (a 6.7% offset from the true value of $n = 5 \cdot 10^3$). In comparison, when $|\Omega| = 256 \cdot 10^3$, the median for this family of graphs is 4865 (just 2.7% offset from the true value). As the hash space size increases from $|\Omega| = 2 \cdot 10^3$ to $|\Omega| = 256 \cdot 10^3$, the error in the median estimate decreases by 4.0%. The magnitude of this phenomenon increases as networks grow larger. For example for a network of size $n = 40 \cdot 10^3$, increasing the hash space size from $|\Omega| = 2 \cdot 10^3$ to $|\Omega| = 256 \cdot 10^3$ causes the error in the median n_3^{ψ} estimate to undergo a 38.4% change.

In addition, Fig 6 shows that as hash space size increases, the interquartile range of the estimates decreases. For example, when $n = 5 \cdot 10^3$ and $|\Omega| = 2 \cdot 10^3$, Exponential degree distribution graphs with $\lambda = 3$ experience a interquartile range of 1491 in their n_3^{ψ} estimates (31.0% of the median). In comparison, when $|\Omega| = 256 \cdot 10^3$, the interquartile range for this family of graphs decreases to 905 (a 39.3% reduction). The magnitude of this effect increases as networks grow larger. For example for a network of size $n = 40 \cdot 10^3$, increasing the hash space size from $|\Omega| = 2 \cdot 10^3$ to $|\Omega| = 256 \cdot 10^3$ causes the interquartile range of the n_3^{ψ} estimate to undergo a 43.0% decrease.

6 Impacts of non-uniformity

The experiments described in previous sections of this paper assumed an RDS process that begins with a set of seeds $D \subseteq V$ sampled *uniformly at random without replacement*. More precisely, $D = X_{|D|}$ is the last entry in sequence $X_0, X_1, \ldots X_{|D|}$, where $X_0 = \emptyset$ and $X_i = X_{i-1} \cup \{u_i\}$ with

$$Pr(u_i = u) = \begin{cases} \dfrac{1}{|V| - |X_{i-1}|} & u \in V \backslash X_{i-1} \\ 0 & \text{otherwise} \end{cases} \tag{44}$$



(a) $n = 5K$, $|\Omega| = 2 \cdot 10^3$ (b) $n = 5K$, $|\Omega| = 32 \cdot 10^3$ (c) $n = 5K$, $|\Omega| = 256 \cdot 10^3$

(d) $n = 10K$, $|\Omega| = 2 \cdot 10^3$ (e) $n = 10K$, $|\Omega| = 32 \cdot 10^3$ (f) $n = 10K$, $|\Omega| = 256 \cdot 10^3$

(g) $n = 20K$, $|\Omega| = 2 \cdot 10^3$ (h) $n = 20K$, $|\Omega| = 32 \cdot 10^3$ (i) $n = 20K$, $|\Omega| = 256 \cdot 10^3$

(j) $n = 40K$, $|\Omega| = 2 \cdot 10^3$ (k) $n = 40K$, $|\Omega| = 32 \cdot 10^3$ (l) $n = 40K$, $|\Omega| = 256 \cdot 10^3$

Fig 6. Estimator n_3^{ψ} on RDS samples of size $r = 500$ with $|\Omega| = 2 \cdot 10^3$ to $256 \cdot 10^3$. In each box, the thick line indicates the sample median; the top of the box is the median of the upper half of the estimated values (75% quartile); the bottom of the box indicates the median of the lower half of the estimated values (25% quartile); and the whiskers indicate the full range of estimated values. No (finite) outliers were removed.

https://doi.org/10.1371/journal.pone.0195959.g006

for each $u \in V$. While the uniform model allowed formal analysis of the estimators' properties to be tractable, many researchers have noted that practical deployments of RDS often exhibit bias in seed selection [90–92]. This bias originates in local features of the network topology (e.g. variation in node degrees) as well as global properties (e.g. the presence of community structures).

6.1 Degree-biased selection of RDS seeds

We begin by describing experimental findings on the differential impacts of degree-based bias in initial seed selection on the performance of the n_3^{ψ} estimator. Towards this, we define a new model of seed selection in which a real-valued parameter $\rho \in \mathbb{R}$ controls degree-based bias. In particular, expression (44) is generalized to

$$Pr(u_i = u) = \begin{cases} \frac{e^{\rho \cdot d(u)}}{\sum_{v \in V \setminus X_{i-1}} e^{\rho \cdot d(v)}} & u \in V \setminus X_{i-1} \\ 0 & \text{otherwise} \end{cases} \quad (45)$$

for each $u \in V$. Note that when $\rho = 0$ expression (45) reduces to the uniform random selection of seeds prescribed in (44). When $\rho > 0$, seed selection is biased towards the network's high degree vertices; when $\rho < 0$, low degree vertices are favored.

The first segment of Table 1 shows that as ρ is varied between -1 and +1, non-uniform seed selection has no discernable negative differential impact on the performance of RDS estimator

Table 1. Varying seed selection bias ($n = 10K$, $r = 500$, $|\Omega| = 256 \cdot 10^3$, $\bar{d}(V) = 5$).

	K	ρ	μ	n_3^{ψ} median	n_3^{ψ} I.Q.R.
ρ: Seed selection bias	1	-1	N.A.	9966.3	2253.0
		-0.5		10104.7	2374.9
		-0.4		10057.2	2313.4
		-0.3		9956.4	2267.2
		-0.2		9981.6	2231.6
		-0.1		9868.0	2254.3
		0		9909.0	2170.4
		0.1		9903.2	2155.1
		0.2		9963.4	2271.6
		0.3		9766.7	2277.9
		0.4		9921.4	2170.2
		0.5		9942.7	2307.1
		1		9793.4	2165.6
K: Number of components	1	0	0.5	10070.9	2325.5
	2			9977.7	2220.8
	4			9797.8	2136.7
	8			9312.6	2216.2
	16			8373.9	1900.0
μ: Cross component probability	8	0	0.1	3088.4	1102.6
			0.2	5526.1	1864.7
			0.3	7595.0	1961.6
			0.4	8639.9	2010.4
			0.5	9345.6	2033.6

https://doi.org/10.1371/journal.pone.0195959.t001

n_3^{ψ}. While the data in the first segment of Table 1 are based on 30 RDS samples ($r = 500$) on each of 30 graphs from $\mathcal{L}(\lambda = 5, n = 10^4)$, i.e. graphs with $10K$ nodes and a Lognormal degree distribution as described in Section 4.1, the conclusion for the other 5 graph families is similar.

6.2 Community structures

Next we consider the impact of community structures which can potentially create bottlenecks for RDS and restrict the reach of subject's self-reported ego networks [91, 92]. We quantify the impacts of such structures on the n_3^{ψ} estimator through simulation experiments, and towards this, extend each of the 5 families defined in Section 4.1 to support the controlled presence of community effects. Two new parameters are introduced: the number of communities K, and the cross-community connection probability μ. The space $\mathcal{L}(\lambda, n)$, for example, is thus extended to a space $\mathcal{L}(\lambda, n, K, \mu)$ consisting of graphs of size $K \cdot \lfloor n/K \rfloor$, i.e. approximately n, which is sampled from as follows:

1. Sample K graphs $G_1 = (V_1, E_1), \ldots G_k = (V_k, E_k)$ from $\mathcal{L}(\lambda, n)$ as defined in Section 4.1. Define $V \coloneqq \bigcup_{i=1}^{K} V_i$ to be the vertex set of our sampled graph. Take $E = \bigcup_{i=1}^{K} E_i$ to be our initial approximation of the edge set of our sampled graph, to be updated according to the rewiring process below.
2. For each $i \in \{1, 2, \ldots, K\}$, and each $u \in V_i$, with probability μ:
   a. Choose $j \in \{1, 2, \ldots, K\} \backslash \{i\}$ uniformly at random, and then choose $v \in V_j$ uniformly at random.
   b. Choose $u' \in N(u) \cap V_i$ uniformly at random.
   c. Choose $v' \in N(v) \cap V_j$ uniformly at random.
   d. Modify E by removing (u, u') and (v, v') from E.
   e. Modify E by adding (u, v) and (u', v') to E.
3. Completion of step (2) yields the sampled graph (V, E) on $K \cdot \lfloor n/K \rfloor$ vertices, having K communities each coming from family $\mathcal{L}(\lambda, n)$ and wired together so that roughly μ fraction of each community's members has a connection to some member of a different community (and the degree distribution of the graph as a whole is consistent with the bias of family $\mathcal{L}$).

The families $\mathcal{P}(\lambda, n, K, \mu)$, $\mathcal{X}(\lambda, n, K, \mu)$, $\mathcal{B}(\lambda, n, K, \mu)$, and $\mathcal{E}(\lambda, n, K, \mu)$, are defined analogously. When $\mu \sim 1$ or $K \sim 1$, community effects are insignificant. As $\mu \to 0^+$ or $K \gg 1$, the population consists of many effectively isolated communities. Whenever a set of seeds are to be selected from the network (e.g. to obtain a respondent driven sample), all seeds are chosen (uniformly at random) *from community 1*.

The second segment of Table 1 shows that as K is increased from 1 to 16 (while μ is held fixed at 0.5), increasing the number of communities causes n_3^{ψ} to slightly underestimate population size. For example, when the network consists of $K = 8$ communities, a median estimate falls short of the true value by 7%; for $K = 16$ communities the deficit becomes 16%. The third and final segment of Table 1 shows that as μ is decreased from 0.5 to 0.1 (while K is held fixed at 8), increasing community isolation causes n_3^{ψ} to significantly underestimate population size. For example, when the inter-community connection probability $\mu = 0.4$ the deficit is 14%, but when $\mu = 0.2$ the estimate produced is roughly 45% of the true value. While the data in the second and third segments of Table 1 are based on 30 trials on each of 30 graphs from $\mathcal{L}(\lambda, n)$, i.e.

graphs with Lognormal degree distribution as described in Section 4.1, the results for the other 5 graph families are quite similar.

6.3 Non-uniform hash functions

The experiments and analyses so far have considered a uniform random hashing function ψ, and have shown that the size of the hash space $|\Omega|$ has a significant impact on estimator variance. The uniform hashing assumption is reasonable when each individual's anonymity-preserving code is based on attributes that have been uniformly randomly assigned across the population. For example, it is reasonable to expect that a telephone company will assign numbers to customers randomly, and thus a code that is built from the parity and scale of the final 4 digits of each individual's phone number would constitute a uniform random hash function.

In this section, we describe how to translate the conclusions of previous experiments and analyses to settings where the hashing function is not uniformly random. This would likely be the case if ψ were built from each individual's demographic characteristics (e.g. age, height, hair color, and race) that are known to vary non-uniformly across the population. For example, if subjects and reports were encoded using 4 categories for age, 3 categories for height, 3 for hair color, and 5 categories for race, one could only say that the hash space size was $4 \times 3 \times 3 \times 5 = 180$ if *all combinations of these attributes were equally likely* to appear. Researchers employing such non-uniform hashing functions may want to know the *equivalent* uniform hash space size $|\Omega|$, so as to correctly translate the results of previous sections into reasonable expectations for the non-uniform situation at hand. The following Lemma will assist in defining this translation:

Lemma 3. *Let A, B be finite sets, and* $\psi: A \to B$ *be a uniformly random function. Then*

$$E|\psi(A)| = |B| \cdot \left[1 - \left(1 - \frac{1}{|B|}\right)^{|A|}\right].$$

Proof. We seek the expected number of distinct items obtained in sampling $|A|$ elements from B with replacement. Consider $x \in B$, then

$$Pr(\{x \in \psi(A)\}) = 1 - Pr(\{x \notin \psi(A)\}) = 1 - \left(1 - \frac{1}{|B|}\right)^{|A|}.$$

The result then follows by linearity of expectation.

Proposition 4. *Let A, B be finite sets, and* $\psi: A \to B$ *a uniformly random function. Suppose* $|A| = x$ *and* $|\psi(A)| = y$, *where* $x, y \gg 0$, *then the maximum likelihood estimator of* $|B|$ *is given by*

$$|B| = \frac{xy}{y \cdot W(-\frac{x}{y}(e^{-\frac{x}{y}})) + x} \tag{46}$$

where Lambert's W function is the inverse function of $f(W) = We^W$.

Proof. Applying Lemma 3, the maximum likelihood estimator is obtained by solving

$$y = |B| \cdot \left(1 - (1 - \frac{1}{|B|})^x\right) \tag{47}$$

for $|B|$. Since $|B| \geq y \gg 0$, we may approximate

$$\log\left(\frac{|B| - 1}{|B|}\right) \approx -\frac{1}{|B|}$$

when $|B|$ is large. Then we have

$$1-\left(1-\frac{1}{|B|}\right)^{x}=1-\exp\left(x\log\left(\frac{|B|-1}{|B|}\right)\right)\approx 1-\exp\left(-\frac{x}{|B|}\right).$$

Eq (47) now becomes

$$y=|B|\cdot\left(1-\exp\left(-\frac{x}{|B|}\right)\right),$$

which when solved for $|B|$ yields expression (46) above.

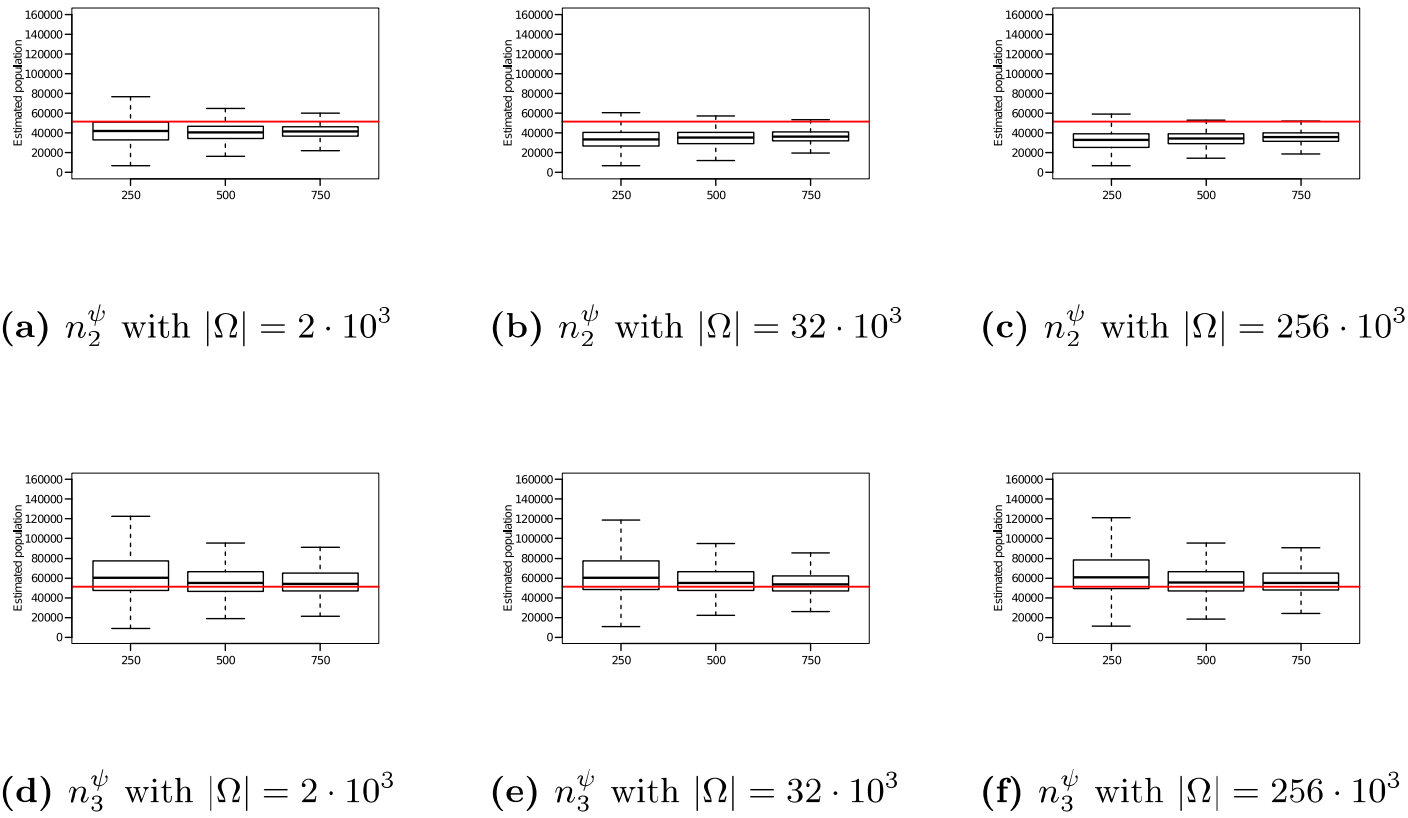
Proposition 4 tells us that the image of a set of size x is expected to have size y, provided the function is a uniform random map into a set whose size is given by expression (46). Such a combinatorial result can be used to compute the equivalent uniform hash space size in settings where the hash function is non-uniform. In particular, if we have $x = |A|$ subjects, who provide us with exactly $y = |\psi(A)|$ distinct codes ($y \leq x$), then the equivalent uniform hash space size $|\Omega|$ is given by expression (46) above.

7 Evaluating estimators on real networks

While a range of degree distributions and randomly occurring clusterings can be expected in idealized topologies, the performance of RDS-based estimators n_2^{ψ} and n_3^{ψ} on organically arising, natural human networks may vary. To test this possibility, we perform a number of random-start, RDS-based estimation experiments on the Brightkite data set. Brightkite was once a location-based social networking service provider where users shared their locations by checking-in. The friendship network was collected using their public API, and consists of $|V|$ = 58,228 nodes and $|E|$ = 214,078 edges [93]. Though originally a directed graph, we symmetrized the edges for the purposes of these experiments. Since not all users made a public check-in during the data collection period, the population we used here consists of 51,406 people. The average clustering coefficient in the network was 0.1723, while the fraction of closed triangles is 0.03979. The diameter (longest-shortest path in the symmetrized network) is 16, though the 90-percentile effective diameter is 6.

For purposes of the experiment we generated 900 respondent-driven samples of size r = 250, 500, 750 and hash space size from $|\Omega| = 2 \cdot 10^3$ to $|\Omega| = 256 \cdot 10^3$ within the Brightkite network, each obtained via an RDS process operating as specified in Assumption 2. The boxplot graphs in Fig 7(a)–7(c) show that estimator n_2^{ψ}—where no accommodation is made for the tendency of RDS to oversample tightly clustered network neighborhoods—underestimates the true population size of 51,406 in every case. Given the high clustering coefficient of the network (17.2%), it seems likely that, for a given sampling tree, the peer-discovery process necessarily walks across close pairs of nodes that shared one or more common vertices. Of note is that increasing the sample size and hash space size does little to correct for these effects.

Graphs (d-f) in Fig 7 present the boxplots of Brightkite population estimates using estimator n_3^{ψ}. As above, we generated 900 respondent-driven samples of size r = 250, 500, 750 and hash space size from $|\Omega| = 2 \cdot 10^3$ to $|\Omega| = 256 \cdot 10^3$ within the Brightkite network. We see that the three different hash space sizes show similar results, while increasing the sample size r from 250 to 500 and 750 improves the accuracy of the median estimate. Unlike the case in Fig 7(a)–7(c), we don't see a consistent pattern of underestimation, indicating that the cross-seed estimator n_3^{ψ} was successful in compensating for the clustering found in the network. As above, the overall size of the hash space has minimal effect on the accuracy of the median



(a) n_2^{ψ} with $|\Omega| = 2 \cdot 10^3$ (b) n_2^{ψ} with $|\Omega| = 32 \cdot 10^3$ (c) n_2^{ψ} with $|\Omega| = 256 \cdot 10^3$

(d) n_3^{ψ} with $|\Omega| = 2 \cdot 10^3$ (e) n_3^{ψ} with $|\Omega| = 32 \cdot 10^3$ (f) n_3^{ψ} with $|\Omega| = 256 \cdot 10^3$

Fig 7. Estimator n_2^{ψ} (above) and n_3^{ψ} (below) on Brightkite network; $|\Omega| = 2 \cdot 10^3$ to $256 \cdot 10^3$, with sample size $r = 250, 500, 750$. In each box, the thick line indicates the sample median; the top of the box is the median of the upper half of the estimated values (75% quartile); the bottom of the box indicates the median of the lower half of the estimated values (25% quartile); and the whiskers indicate the full range of estimated values. Data points that exceeded the third quartile boundary by over 1.5 times the interquartile range (IQR) were treated as outliers and removed.

https://doi.org/10.1371/journal.pone.0195959.g007

estimate, but we note that an increase in the RDS sample size improves the accuracy of the median estimate and produces smaller interquartile ranges.

8 Discussion

The results shown here indicate that size estimates for hidden and hard-to-reach populations can be derived from RDS samples across a range of topologies, and in the presence of significant network clustering. As important, this is accomplished under conditions of anonymity by way of identity hashing, e.g. using telefunken codes [50] or a Privatized Network Sampling (PNS) design [53]. The n_3^{ψ} estimator joins other successful, RDS-based population estimation procedures, such as those by Handcock and Gile [85], and Crawford, Wu, and Heimer [35]. Like Crawford et al, we make use of half-edge counts. However, our estimator invokes a different strategy—beginning with the original capture-recapture concept—and is shown to be robust across a wide range of topologies and assumptions.

A notable feature of the n_3^{ψ} estimator is that a lower level of variance can be expected at conventional RDS sample sizes. For $r = 500$ to 750, interquartile ranges were low relative to both the median estimate and true population size (See segments 1 and 2 of Table 2 which summarize a slice of the data in Fig 6).

Additionally, when hashing was employed towards ensuring subject anonymity, sufficiently large hash spaces ($32 \cdot 10^3$ or larger) and samples sizes (500 or above) produced a narrow

Table 2. A cross-section of the experimental findings in this paper.

	n	r	$\lvert\Omega\rvert$	$\bar{d}(V)$	n_3^{ψ} median	n_3^{ψ} I.Q.R.
n: Population size	5,000	750	$256 \cdot 10^3$	10	4934.3	342.1
	10,000				9927.2	1068.6
	20,000				20018.7	2731.1
	40,000				39964.7	9621.1
r: Sample size	5,000	250	$256 \cdot 10^3$	10	4972.7	1080.8
		500			4957.0	501.6
		750			4934.3	342.1
$\lvert\Omega\rvert$: Hash space size	5,000	750	$2 \cdot 10^3$	10	4875.8	945.1
			$32 \cdot 10^3$		4938.4	363.1
			$256 \cdot 10^3$		4934.3	342.1
$\bar{d}(V)$: Average degree	5,000	750	$256 \cdot 10^3$	3	4797.5	848.5
				5	4867.7	565.5
				10	4934.3	342.1

https://doi.org/10.1371/journal.pone.0195959.t002

range of estimates (See segment 3 of Table 2 which summarizes a slice of the data in Fig 6). Given concerns about RDS sample variance generally [28], these results indicate robustness against the faults of a single sample.

Another consistent feature observed in these experiments is the performance of the n_3^{ψ} estimator as graph density increases (See segment 4 of Table 2 which summarizes a slice of the data in Fig 6). In terms of the interquartile ranges, the estimator exhibits worse performance in sparse (i.e. $\bar{d}(V) = 3$) as opposed to dense networks (i.e. $\bar{d}(V) = 10$). Given the edge-sampling focus of our approach, this is not surprising. Fewer total edges suggest fewer total "matches" to discover, leading to greater variability depending on stochastic factors likely associated with the selection of RDS seeds and the random walk features of the RDS sampling process. These results suggest limits on the implementation of n_3^{ψ} estimator in sparse graphs.

As researchers increasingly turn to RDS methods for sampling hard-to-reach populations, these results should be of considerable interest to those concerned with what is often referred to as "the denominator problem". Where agencies and government administrations seek to understand both the scope of public health challenges and to measure the effectiveness of their intervention and promotion efforts, the ability to estimate population size (and with this, population prevalence) is widely needed. The results presented here indicate that "one step" methods are capable of providing such estimates. Along with the methods mentioned above, this work has the potential to provide public health officials and planners with means to more effectively promote the health of hidden populations—and thus the health of the larger populations in which they are embedded.

8.1 Limitations

In using uniform random samples to estimate population size, it is possible for the proposed n_1 estimator to "fail" if one finds that $\langle M(T, \emptyset)\rangle = 0$ in Definition 3. This happens with greater frequency as the sample size $r \ll n$ the population size. Fig 8(a) shows the mean failure rate (the fraction of the 13,500 trials where n_1 failed to produce a population estimate), for each choice of population size n (ranging from $5 \cdot 10^3$ to $40 \cdot 10^3$), and uniform sample size r (chosen to be 250, 500 or 750). We see from Fig 8(a) that the failure rate is non-linear in both r and

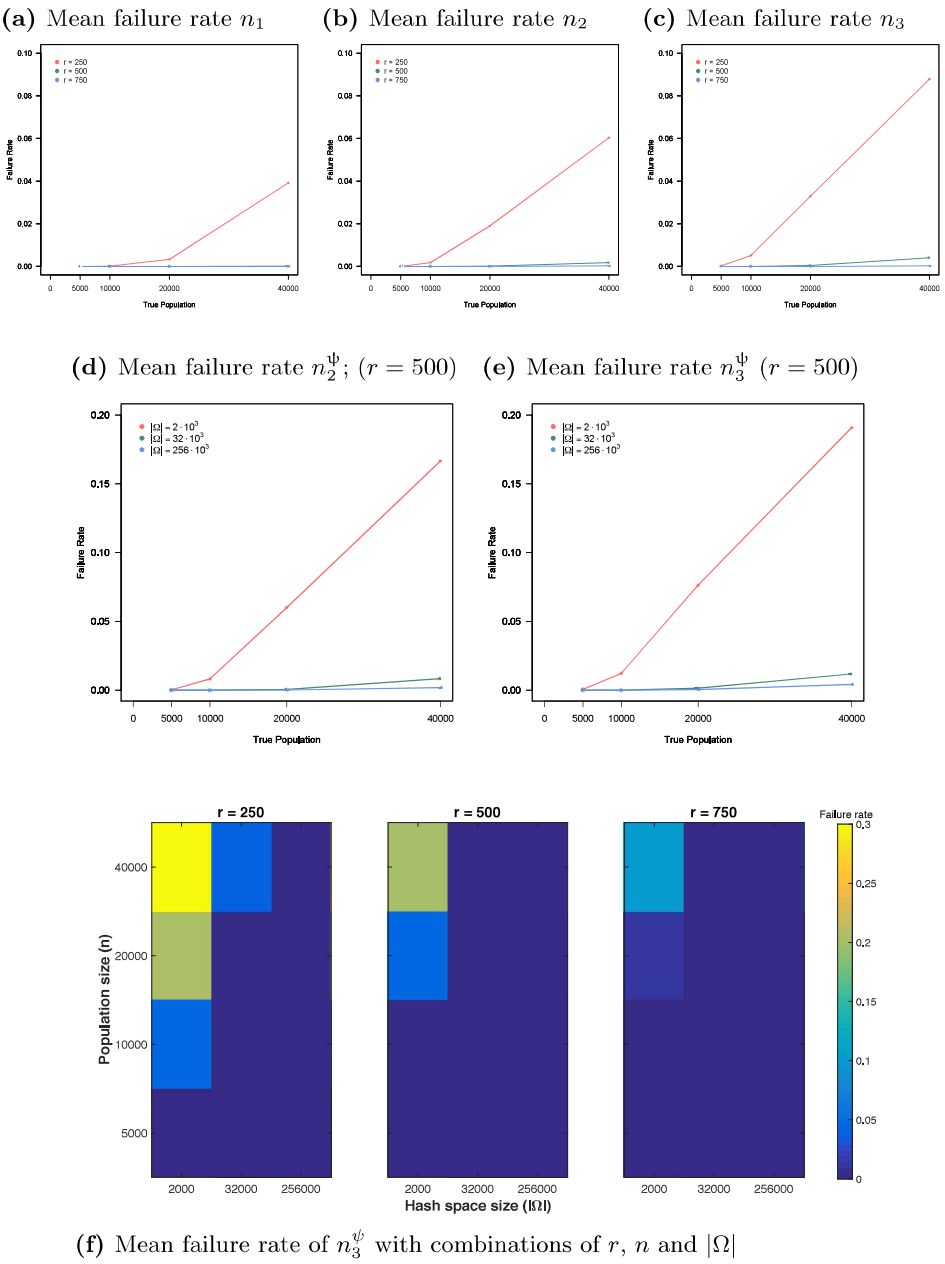


Fig 8. Mean failure rate analysis of the proposed estimators.

https://doi.org/10.1371/journal.pone.0195959.g008

n. For small uniform samples $r = 250$, the failure rate of n_1 is ~ 0 when $n = 10 \cdot 10^3$, but undergoes an inflection at $n = 20 \cdot 10^3$, and rises to 3.9% when the population size again doubles to $n = 40 \cdot 10^3$. Note that we considered each of 5 families $\mathcal{L}(\lambda, n)$, $\mathcal{P}(\lambda, n)$, $\mathcal{X}(\lambda, n)$, $\mathcal{B}(\lambda, n)$, and $\mathcal{E}(\lambda, n)$ defined in Section 4.1, and each $\lambda = 3, 5, 10$; from each of these 15 concrete sample spaces, we used configuration graph sampling to select 30 random graphs of size n. In each of these $5 \times 3 \times 30 = 450$ graphs, we generated 30 uniform samples (for n_1). In this manner, a total of $450 \times 30 = 13{,}500$ simulations were conducted.

Similarly, in using respondent-driven sampling to estimate population size, it is possible for the proposed n_2 (resp. n_3) estimators to "fail" if one finds that $\langle M(S, F) \rangle = 0$ in Definition 4

(resp. $\sum_{s\in D}\langle X(s,F,\gamma)\rangle = 0$ in Definition 9). Fig 8(b) shows the mean failure rate (the fraction of the 13,500 trials where n_2 failed to produce a population estimate), for each choice of population size n (ranging from $5 \cdot 10^3$ to $40 \cdot 10^3$), and RDS sample size r (chosen to be 250, 500 or 750). RDS samples of size $r = 250$ exhibit an n_2 failure rate of ~ 0 when $n = 5 \cdot 10^3$, but undergo an inflection at $n = 10 \cdot 10^3$; the mean failure rate rises to 6% when the population size again doubles to $n = 40 \cdot 10^3$. In examining the n_3 estimator, Fig 8(c) shows us that when it is used with RDS samples of size $r = 250$, it exhibits a failure rate of ~ 0 when $n = 5 \cdot 10^3$, but the failure rate undergoes an inflection at $n = 10 \cdot 10^3$, rising to 8.8% when the population size again doubles to $n = 40 \cdot 10^3$. For sample sizes that are $2X$ and $3X$ as large (i.e. $r = 500$ and $r = 750$) the inflection point is not yet reached at $n = 40 \cdot 10^3$ and mean failure rates remain below 0.1%. This indicates that our estimators based on RDS are more robust against failure than the n_1 uniform sampling estimator, and at typical RDS sample sizes ($500 \leq r \leq 750$), they are robust enough to be used in settings where the population size is expected to be on the order of $n \sim 40 \cdot 10^3$.

Fig 8(d)–8(e) explore the impact of hash space size on the mean failure rate. Here we consider a fixed sample size $r = 500$ and vary the size of hash space $|\Omega|$ between $2 \cdot 10^3$ and $256 \cdot 10^3$. We observe that the mean failure rates of n_2^{Ψ} and n_3^{Ψ} (again taken across 13,500 experiments) grow linearly as n increases, but that the rate of growth depends on $|\Omega|$. In particular, when $|\Omega|$ is too small (in this case $2 \cdot 10^3$ or smaller), the mean failure rate is seen to grow steeply, even for small networks. The two graphs (d-e) make evident the tradeoff between subject anonymity/privacy and the failure rates of the estimator. When the hash space size is sufficiently large ($32 \cdot 10^3$–$256 \cdot 10^3$), failure rates remain low, but smaller hash spaces (which provide for greater anonymity) may produce greater instability in the estimators. Finally, the three heatmaps in Fig 8(f) show how the failure rate of n_3^{Ψ} rises whenever the hash space size or sample size decreases.

Although $32 \cdot 10^3 - 256 \cdot 10^3$ may appear to be a very large hash space size, we note

$$10^4 \leq 32 \cdot 10^3 \leq 10^5 \leq 256 \cdot 10^3 \leq 10^6.$$

Thus, asking research subjects for the last 5 or 6 digits of their own telephone number and those digits of their friends' phone numbers would be sufficient to provide an accurate estimate (assuming that numerical digits are randomly assigned by phone service providers). In the event that research subjects remain reluctant to reveal precise digits of their own or their alter's phone numbers, a telefunken code could be constructed [50] or a Privatized Network Sampling (PNS) design [53] employed.

Acknowledgments

Research reported in this publication was supported by the National Institutes for Health, National Institute on Drug Abuse under Award Number R01 DA037117 and National Institute for General Medicine R01 GM118427, as well as NSF grants MMS-0851555 and SES-1357619. The authors would like to thank Meredith Dank, Ric Curtis, and Travis Wendel for feedback and discussions. We are grateful to the organizers and attendees of the PSE Consultation at CDC, where these results were presented in January 2017. We are indebted to the reviewers for their recommendations which greatly strengthened the exposition of these results.

All Java software developed and used for this research is publicly available at https://github.com/grouptheory/telefunken-support/tree/master/java.

All R software developed and used is available at https://github.com/grouptheory/telefunken-support/tree/master/R-v1.

All data inputs and outputs are available at https://github.com/grouptheory/telefunken-support/tree/master/figures_and_data.

Author Contributions

Formal analysis: Bilal Khan, Hsuan-Wei Lee, Ian Fellows, Kirk Dombrowski.

Investigation: Bilal Khan, Hsuan-Wei Lee, Ian Fellows, Kirk Dombrowski.

Methodology: Bilal Khan, Hsuan-Wei Lee, Ian Fellows, Kirk Dombrowski.

References

1. Magnani R, Sabin K, Saidel T, Heckathorn D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. AIDS. 2005; 19:S67. https://doi.org/10.1097/01.aids.0000172879.20628.e1 PMID: 15930843
2. Dombrowski K. Topological and Historical Considerations for Infectious Disease Transmission among Injecting Drug Users in Bushwick, Brooklyn (USA). World Journal of AIDS. 2013; 03(01):1–9. https://doi.org/10.4236/wja.2013.31001
3. Reluga T, Meza R, Walton Db, Galvani Ap. Reservoir interactions and disease emergence. Theoretical Population Biology. 2007; 72(3):400–408. https://doi.org/10.1016/j.tpb.2007.07.001 PMID: 17719617
4. Bonin JP, Fournier L, Blais R. A Typology of Mentally Disordered Users of Resources for Homeless People: Towards Better Planning of Mental Health Services. Administration and Policy in Mental Health and Mental Health Services Research. 2009; 36(4):223–235. https://doi.org/10.1007/s10488-009-0206-2 PMID: 19214733
5. Burt MR. Critical Factors in Counting the Homeless. American Journal of Orthopsychiatry. 1995; 65(3):334–339. https://doi.org/10.1037/h0085059 PMID: 7485418
6. Ivanich J, Welch-Lazoritz M, Dombrowski K. The Relationship between Survival Sex and Borderline Personality Disorder Symptoms in a High Risk Female Population. International Journal of Environmental Research and Public Health. 2017; 14(9):1031. https://doi.org/10.3390/ijerph14091031
7. Potterat JJ, Woodhouse DE, Rothenberg RB, Muth SQ, Darrow WW, Muth JB, et al. AIDS in Colorado Springs: is there an epidemic? AIDS (London, England). 1993; 7(11):1517–1521. https://doi.org/10.1097/00002030-199311000-00017
8. Abdul-Quader AS, Baughman AL, Hladik W. Estimating the size of key populations: current status and future possibilities. Current Opinion in HIV and AIDS. 2014; 9(2):107–114. https://doi.org/10.1097/COH.0000000000000041 PMID: 24393694
9. Law DCG, Serre ML, Christakos G, Leone PA, Miller WC. Spatial analysis and mapping of sexually transmitted diseases to optimise intervention and prevention strategies. Sexually Transmitted Infections. 2004; 80(4):294–299. https://doi.org/10.1136/sti.2003.006700 PMID: 15295129
10. Zohrabyan L, Johnston LG, Scutelniciuc O, Iovita A, Todirascu L, Costin T, et al. Determinants of HIV Infection Among Female Sex Workers in Two Cities in the Republic of Moldova: The Role of Injection Drug Use and Sexual Risk. AIDS and behavior. 2013; https://doi.org/10.1007/s10461-013-0460-x PMID: 23539186
11. Darke S. Self-report among injecting drug users: A review. Drug and Alcohol Dependence. 1998; 51(3):253–263. https://doi.org/10.1016/S0376-8716(98)00028-3 PMID: 9787998
12. Harwood EM, Horvath KJ, Courtenay-Quirk C, Fisher H, Kachur R, McFarlane M, et al. Sampling hidden populations: lessons learned from a telephone-based study of persons recently diagnosed with HIV (PRDH). International Journal of Social Research Methodology. 2012; 15(1):31–40. https://doi.org/10.1080/02650533.2011.573302
13. Larson A, Stevens A, Wardlaw G. Indirect estimates of 'hidden' populations: Capture-recapture methods to estimate the numbers of heroin users in the Australian capital territory. Social Science & Medicine. 1994; 39(6):823–831. https://doi.org/10.1016/0277-9536(94)90044-2
14. Vuylsteke B, Vandenhoudt H, Langat L, Semde G, Menten J, Odongo F, et al. Capture—recapture for estimating the size of the female sex worker population in three cities in Côte d'Ivoire and in Kisumu, western Kenya. Tropical Medicine & International Health. 2010; 15(12):1537–1543. https://doi.org/10.1111/j.1365-3156.2010.02654.x
15. Biernacki P, Waldorf D. Snowball Sampling: Problems and Techniques of Chain Referral Sampling. Sociological Methods & Research. 1981; 10(2):141–163. https://doi.org/10.1177/004912418101000205

16. Platt L, Wall M, Rhodes T, Judd A, Hickman M, Johnston LG, et al. Methods to recruit hard-to-reach groups: comparing two chain referral sampling methods of recruiting injecting drug users across nine studies in Russia and Estonia. Journal of Urban Health: Bulletin of the New York Academy of Medicine. 2006; 83(6 Suppl):i39–53. https://doi.org/10.1007/s11524-006-9101-2

17. Haley DF, Golin C, El-Sadr W, Hughes JP, Wang J, Roman Isler M, et al. Venue-based recruitment of women at elevated risk for HIV: an HIV prevention trials network study. Journal of Women's Health. 2014; 23(6):541–551. https://doi.org/10.1089/jwh.2013.4654 PMID: 24742266

18. Muhib FB, Lin LS, Stueve A, Miller RL, Ford WL, Johnson WD, et al. A venue-based method for sampling hard-to-reach populations. Public health reports. 2001; 116(Suppl 1):216. https://doi.org/10.1093/phr/116.S1.216 PMID: 11889287

19. Burnham G, Lafta R, Doocy S, Roberts L. Mortality after the 2003 invasion of Iraq: a cross-sectional cluster sample survey. The Lancet. 2006; 368(9545):1421–1428. https://doi.org/10.1016/S0140-6736(06)69491-9

20. Heckathorn DD. Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment. Sociological Methodology. 2007; 37(1):151–207. https://doi.org/10.1111/j.1467-9531.2007.00188.x

21. Heckathorn DD. Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. Social Problems. 2002;. https://doi.org/10.1525/sp.2002.49.1.11

22. Salganik MJ, Heckathorn DD. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. Sociological Methodology. 2004; 34(1):193–239. https://doi.org/10.1111/j.0081-1750.2004.00152.x

23. Johnston LG, Hakim AJ, Dittrich S, Burnett J, Kim E, White RG. A systematic review of published respondent-driven sampling surveys collecting behavioral and biologic data. AIDS and behavior. 2016; 20(8):1754–1776. https://doi.org/10.1007/s10461-016-1346-5 PMID: 26992395

24. Gile KJ, Handcock MS. Respondent-Driven Sampling: An Assessment of Current Methodology. Sociological Methodology. 2010; 40(1):285–327. https://doi.org/10.1111/j.1467-9531.2010.01223.x PMID: 22969167

25. Gile KJ, Johnston LG, Salganik MJ. Diagnostics for respondent-driven sampling. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2015; 178(1):241–269. https://doi.org/10.1111/rssa.12059

26. Mouw T, Verdery AM. Network Sampling with Memory A Proposal for More Efficient Sampling from Social Networks. Sociological Methodology. 2012; 42(1):206–256. https://doi.org/10.1177/0081175012461248 PMID: 24159246

27. Shi Y, Cameron CJ, Heckathorn DD. Model-Based and Design-Based Inference: Reducing Bias Due to Differential Recruitment in Respondent-Driven Sampling. Sociological Methods & Research. 2016; p. 0049124116672682. https://doi.org/10.1177/0049124116672682

28. Verdery AM, Mouw T, Bauldry S, Mucha PJ. Network structure and biased variance estimation in respondent driven sampling. PloS one. 2015; 10(12):e0145296. https://doi.org/10.1371/journal.pone.0145296 PMID: 26679927

29. Wejnert C. An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and Out-of-Equilibrium Data. Sociological Methodology. 2009; 39(1):73–116. https://doi.org/10.1111/j.1467-9531.2009.01216.x PMID: 20161130

30. Heckathorn DD, Cameron CJ. Network Sampling. Annual Review of Sociology. 2017; 43(1). https://doi.org/10.1146/annurev-soc-060116-053556

31. Abdul-Quader AS, Heckathorn DD, Sabin K, Saidel T. Implementation and analysis of respondent driven sampling: lessons learned from the field. Journal of Urban Health. 2006; 83(1):1–5. https://doi.org/10.1007/s11524-006-9108-8

32. Johnston LG, Malekinejad M, Kendall C, Iuppa IM, Rutherford GW. Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: field experiences in international settings. AIDS and Behavior. 2008; 12(1):131–141. https://doi.org/10.1007/s10461-008-9413-1

33. Goel S, Salganik MJ. Assessing respondent-driven sampling. Proceedings of the National Academy of Sciences. 2010; 107(15):6743–6747. https://doi.org/10.1073/pnas.1000261107

34. Sulaberidze L, Mirzazadeh A, Chikovani I, Shengelia N, Tsereteli N, Gotsadze G. Population Size Estimation of Men Who Have Sex with Men in Tbilisi, Georgia; Multiple Methods and Triangulation of Findings. PLoS ONE. 2016; 11(2):e0147413. https://doi.org/10.1371/journal.pone.0147413 PMID: 26828366

35. Crawford FW, Wu J, Heimer R. Hidden population size estimation from respondent-driven sampling: a network approach. Journal of the American Statistical Association. 2017;0(ja):0–0. https://doi.org/10.1080/01621459.2017.1285775

36. Berchenko Y, Frost SD. Capture-recapture methods and respondent-driven sampling: their potential and limitations; 2011.

37. Berchenko Y, White RG, Wejnert C, Frost SD. Analysis of a capture-recapture estimator for the size of populations with heterogenous catchability, and its evaluation on RDS data from rural Uganda. arXiv preprint arXiv:11111714. 2011;.

38. Paz-Bailey G, Jacobson J, Guardado M, Hernandez F, Nieto A, Estrada M, et al. How many men who have sex with men and female sex workers live in El Salvador? Using respondent-driven sampling and capture—recapture to estimate population sizes. Sexually transmitted infections. 2011; p. sti–2010.

39. Domingo-Salvany A, Hartnoll RL, Maguire A, Brugal MT, Albertin PA, Caylà JA, et al. Analytical considerations in the use of capture-recapture to estimate prevalence: case studies of the estimation of opiate use in the metropolitan area of Barcelona, Spain. American journal of epidemiology. 1998; 148(8):732–740. https://doi.org/10.1093/oxfordjournals.aje.a009694 PMID: 9786228

40. Kruse N, Frieda MTB, Vaovola G, Burkhardt G, Barivelo T, Amida X, et al. Participatory mapping of sex trade and enumeration of sex workers using capture-recapture methodology in Diego-Suarez, Madagascar. Sexually transmitted diseases. 2003; 30(8):664–670 PMID: 12897692

41. Bernard HR, Hallett T, Iovita A, Johnsen EC, Lyerla R, McCarty C, et al. Counting hard-to-count populations: the network scale-up method for public health. Sexually Transmitted Infections. 2010; 86(Suppl 2):ii11–ii15. https://doi.org/10.1136/sti.2010.044446 PMID: 21106509

42. Hay G, McKeganey N. Estimating the prevalence of drug misuse in Dundee, Scotland: an application of capture-recapture methods. Journal of Epidemiology and Community Health. 1996; 50(4):469–472. https://doi.org/10.1136/jech.50.4.469 PMID: 8882234

43. Jones HE, Hickman M, Welton NJ, De Angelis D, Harris RJ, Ades AE. Recapture or Precapture? Fallibility of Standard Capture-Recapture Methods in the Presence of Referrals Between Sources. American journal of epidemiology. 2014; p. kwu056.

44. Wolitski RJ, Pals SL, Kidder DP, Courtenay-Quirk C, Holtgrave DR. The effects of HIV stigma on health, disclosure of HIV status, and risk behavior of homeless and unstably housed persons living with HIV. AIDS and Behavior. 2009; 13(6):1222–1232. https://doi.org/10.1007/s10461-008-9455-4 PMID: 18770023

45. Ezoe S, Morooka T, Noda T, Sabin ML, Koike S. Population size estimation of men who have sex with men through the network scale-up method in Japan. PloS one. 2012; 7(1):e31184. https://doi.org/10.1371/journal.pone.0031184 PMID: 22563366

46. Guo W, Bao S, Lin W, Wu G, Zhang W, Hladik W, et al. Estimating the size of HIV key affected populations in Chongqing, China, using the network scale-up method. PloS one. 2013; 8(8):e71796. https://doi.org/10.1371/journal.pone.0071796 PMID: 23967246

47. Habecker P, Dombrowski K, Khan B. Improving the Network Scale-Up Estimator: Incorporating Means of Sums, Recursive Back Estimation, and Sampling Weights. PloS one. 2015; 10(12). https://doi.org/10.1371/journal.pone.0143406 PMID: 26630261

48. Killworth PD, McCarty C, Johnsen EC, Bernard HR, Shelley GA. Investigating the variation of personal network size under unknown error conditions. Sociological Methods & Research. 2006; 35(1):84–112. https://doi.org/10.1177/0049124106289160

49. Salganik MJ, Fazito D, Bertoni N, Abdo AH, Mello MB, Bastos FI. Assessing network scale-up estimates for groups most at risk of HIV/AIDS: evidence from a multiple-method study of heavy drug users in Curitiba, Brazil. American journal of epidemiology. 2011; 174(10):1190–1196. https://doi.org/10.1093/aje/kwr246 PMID: 22003188

50. Dombrowski K, Khan B, Wendel T, McLean K, Misshula E, Curtis R. Estimating the Size of the Methamphetamine-Using Population in New York City Using Network Sampling Techniques. Advances in Applied Sociology. 2012; 2(4):1–20. https://doi.org/10.4236/aasoci.2012.24032

51. Curtis R, Terry K, Dank M, Dombrowski K, Khan B. The commercial sexual exploitation of children in New York City, Volume 1: The CSEC population in New York City: Size, characteristics, and needs (NCJ Publication No. 225083). Bureau of Justice Statistics, Washington, DC. Final report submitted to the National Institute of Justice New York, NY: Center for Court Innovation and John Jay College of Criminal Justice Retrieved January. 2008;12:2012.

52. Wendel T, Khan B, Dombrowski K, Curtis R, McLean K, Misshula E, et al. Dynamics of Methamphetamine Markets in New York City: Final Technical Report to the National Institute of Justice; A Report to the National Institute of Justice (Award # 2007-IJ-CX-0110. vol. NIJ Document 236122; 2011.

53. Fellows I. Exponential Family Random Network Models. UCLA. Statistics 0891; 2012.

54. Merli MG, Verdery A, Mouw T, Li J. Sampling migrants from their social networks: The demography and social organization of Chinese migrants in Dar es Salaam, Tanzania. Migration studies. 2016; 4(2):182–214. https://doi.org/10.1093/migration/mnw004 PMID: 27746912

55. Heckathorn DD, Cameron CJ. Network Sampling: From Snowball and Multiplicity to Respondent-Driven Sampling. Annual Review of Sociology. 2017;(0). https://doi.org/10.1146/annurev-soc-060116-053556

56. Leskovec J, Faloutsos C. Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2006. p. 631–636.

57. Wang P, Lui J, Ribeiro B, Towsley D, Zhao J, Guan X. Efficiently estimating motif statistics of large networks. ACM Transactions on Knowledge Discovery from Data (TKDD). 2014; 9(2):8. https://doi.org/10.1145/2629564

58. Bawa M, Garcia-Molina H, Gionis A, Motwani R. Estimating aggregates on a peer-to-peer network. Stanford InfoLab; 2003.

59. Massoulié L, Le Merrer E, Kermarrec AM, Ganesh A. Peer counting and sampling in overlay networks: random walk methods. In: Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing. ACM; 2006. p. 123–132.

60. Finkelstein M, Tucker HG, Veeh JA. Confidence intervals for the number of unseen types. Statistics & Probability Letters. 1998; 37(4):423–430. https://doi.org/10.1016/S0167-7152(97)00146-6

61. Sekar CC, Deming WE. On a method of estimating birth and death rates and the extent of registration. Journal of the American Statistical Association. 1949; 44(245):101–115. https://doi.org/10.1080/01621459.1949.10483294

62. Lincoln FC. Calculating Waterfowl Abundance on the Basis of Banding Returns. United States Department of Agriculture Circular. 1930; 118:1–4.

63. Petersen CP. The Yearly Immigration of Young Plaice Into the Limfjord From the German Sea. Report of the Danish Biological Station. 1896; 6:5–84.

64. Dasgupta A, Kumar R, Sivakumar D. Social sampling. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2012. p. 235–243.

65. Katzir L, Liberty E, Somekh O. Estimating sizes of social networks via biased sampling. In: Proceedings of the 20th international conference on World wide web. ACM; 2011. p. 597–606.

66. Krishnamurthy V, Faloutsos M, Chrobak M, Lao L, Cui JH, Percus AG. Reducing large internet topologies for faster simulations. In: Networking. vol. 5. Springer; 2005. p. 328–341.

67. Kurant M, Butts CT, Markopoulou A. Graph size estimation. arXiv preprint arXiv:12100460. 2012;.

68. Dasgupta A, Kumar R, Sarlos T. On estimating the average degree. In: Proceedings of the 23rd international conference on World wide web. ACM; 2014. p. 795–806.

69. Gile KJ. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. Journal of the American Statistical Association. 2011; 106(493):135–146. https://doi.org/10.1198/jasa.2011.ap09475

70. Barash VD, Cameron CJ, Spiller MW, Heckathorn DD. Respondent-Driven Sampling—Testing Assumptions: Sampling with Replacement; 2016.

71. Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B. Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM; 2007. p. 29–42.

72. Ahn YY, Han S, Kwak H, Moon S, Jeong H. Analysis of topological characteristics of huge online social networking services. In: Proceedings of the 16th international conference on World Wide Web. ACM; 2007. p. 835–844.

73. Gjoka M, Kurant M, Butts CT, Markopoulou A. Walking in facebook: A case study of unbiased sampling of osns. In: Infocom, 2010 Proceedings IEEE. IEEE; 2010. p. 1–9.

74. Kurant M, Gjoka M, Butts CT, Markopoulou A. Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In: Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems. ACM; 2011. p. 281–292.

75. Hardiman SJ, Katzir L. Estimating clustering coefficients and size of social networks via random walk. In: Proceedings of the 22nd international conference on World Wide Web. ACM; 2013. p. 539–550.

76. Slutsky E. Uber stochastische asymptoten und grenzwerte. Metron. 1925; 5(3):3–89.

77. Dombrowski K, Khan B, Wendel T, McLean K, Misshula E, Curtis R. Estimating the Size of the Methamphetamine-Using Population in New York City Using Network Sampling Techniques. Advances in Applied Sociology. 2012; 2(4):245–252. https://doi.org/10.4236/aasoci.2012.24032 PMID: 24672746

78. Erdös P, Rényi A. On random graphs, I. Publicationes Mathematicae (Debrecen). 1959; 6:290–297.

79. Bollobas B. Modern Graph Theory. Springer; 1998.

80. Bender EA, Canfield ER. The asymptotic number of labeled graphs with given degree sequences. Journal of Combinatorial Theory, Series A. 1978; 24(3):296–307. http://dx.doi.org/10.1016/0097-3165(78)90059-6.

81. Bollobás B. A Probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs. European Journal of Combinatorics. 1980; 1(4):311–316. http://dx.doi.org/10.1016/S0195-6698(80)80030-8.

82. Newman MEJ, Strogatz SH, Watts DJ. Random graphs with arbitrary degree distributions and their applications. Phys Rev E. 2001; 64:026118. https://doi.org/10.1103/PhysRevE.64.026118

83. Albert R, Barabási AL. Statistical mechanics of complex networks. Rev Mod Phys. 2002; 74:47–97. https://doi.org/10.1103/RevModPhys.74.47

84. Illenberger J, Flötteröd G. Estimating network properties from snowball sampled data. Social Networks. 2012; 34(4):701–711. https://doi.org/10.1016/j.socnet.2012.09.001

85. Handcock MS, Gile KJ, Mar CM. Estimating hidden population size using respondent-driven sampling data. Electronic journal of statistics. 2014; 8(1):1491. https://doi.org/10.1214/14-EJS923 PMID: 26180577

86. Coronado-García M, Thrash CR, Welch-Lazoritz M, Gauthier R, Reyes JC, Khan B, et al. Using Network Sampling and Recruitment Data to Understand Social Structures Related to Community Health in a Population of People Who Inject Drugs in Rural Puerto Rico. Puerto Rico Health Sciences Journal. 2017; 36(2):77–83. PMID: 28622403

87. Verdery AM, Fisher JC, Siripong N, Abdesselam K, Bauldry S. New Survey Questions and Estimators for Network Clustering with Respondent-driven Sampling Data. Sociological Methodology. 2017; p. 0081175017716489. https://doi.org/10.1177/0081175017716489

88. Verdery AM, Siripong N, Pence BW. Social Network Clustering and the Spread of HIV/AIDS Among Persons Who Inject Drugs in 2 Cities in the Philippines. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2017; 76(1):26–32. https://doi.org/10.1097/QAI.0000000000001485 PMID: 28650399

89. Carter JL, Wegman MN. Universal classes of hash functions. Journal of Computer and System Sciences. 1979; 18(2):143–154. http://dx.doi.org/10.1016/0022-0000(79)90044-8.

90. McCreesh N, Johnston LG, Copas A, Sonnenberg P, Seeley J, Hayes RJ, et al. Evaluation of the role of location and distance in recruitment in respondent-driven sampling. International journal of health geographics. 2011; 10(1):56. https://doi.org/10.1186/1476-072X-10-56 PMID: 22008416

91. Rocha LE, Thorson AE, Lambiotte R, Liljeros F. Respondent-driven sampling bias induced by community structure and response rates in social networks. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2017; 180(1):99–118. https://doi.org/10.1111/rssa.12180

92. Sperandei S, Bastos LS, Ribeiro-Alves M, Bastos FI. Assessing respondent-driven sampling: A simulation study across different networks. Social Networks. 2017;.

93. Cho E, Myers SA, Leskovec J. Friendship and Mobility: User Movement in Location-based Social Networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'11. New York, NY, USA: ACM; 2011. p. 1082–1090. Available from: http://doi.acm.org/10.1145/2020408.2020579.