Estimating the Size of a Large Network and its Communities from a Random Sample

Lin Chen[1,2], **Amin Karbasi**[1,2], **Forrest W. Crawford**[2,3]
[1]Department of Electrical Engineering, [2]Yale Institute for Network Science,
[3]Department of Biostatistics, Yale University
{lin.chen, amin.karbasi, forrest.crawford}@yale.edu

Abstract

Most real-world networks are too large to be measured or studied directly and there is substantial interest in estimating global network properties from smaller sub-samples. One of the most important global properties is the number of vertices/nodes in the network. Estimating the number of vertices in a large network is a major challenge in computer science, epidemiology, demography, and intelligence analysis. In this paper we consider a population random graph $G = (V, E)$ from the stochastic block model (SBM) with K communities/blocks. A sample is obtained by randomly choosing a subset $W \subseteq V$ and letting $G(W)$ be the induced subgraph in G of the vertices in W. In addition to $G(W)$, we observe the total degree of each sampled vertex and its block membership. Given this partial information, we propose an efficient PopULation Size Estimation algorithm, called PULSE, that accurately estimates the size of the whole population as well as the size of each community. To support our theoretical analysis, we perform an exhaustive set of experiments to study the effects of sample size, K, and SBM model parameters on the accuracy of the estimates. The experimental results also demonstrate that PULSE significantly outperforms a widely-used method called the network scale-up estimator in a wide variety of scenarios.

1 Introduction

Many real-world networks cannot be studied directly because they are obscured in some way, are too large, or are too difficult to measure. There is therefore a great deal of interest in estimating properties of large networks via sub-samples [15, 5]. One of the most important properties of a large network is the number of vertices it contains. Unfortunately census-like enumeration of all the vertices in a network is often impossible, so researchers must try to learn about the size of real-world networks by sampling smaller components. In addition to the size of the total network, there is great interest in estimating the size of different *communities* or sub-groups from a sample of a network. Many real-world networks exhibit community structure, where nodes in the same community have denser connections than those in different communities [10, 18]. In the following examples, we describe network size estimation problems in which only a small subgraph of a larger network is observed.

Social networks. The social and economic value of an online social network (e.g. Facebook, Instagram, Twitter) is closely related to the number of users the service has. When a social networking service does not reveal the true number of users, economists, marketers, shareholders, or other groups may wish to estimate the number of people who use the service based on a sub-sample [4].

World Wide Web. Pages on the World-Wide Web can be classified into several categories (e.g. academic, commercial, media, government, etc.). Pages in the same category tend to have more connections. Computer scientists have developed crawling methods for obtaining a sub-network of web pages, along with their hyperlinks to other unknown pages. Using the crawled sub-network and hyperlinks, they can estimate the number of pages of a certain category [17, 16, 21, 13, 19].

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

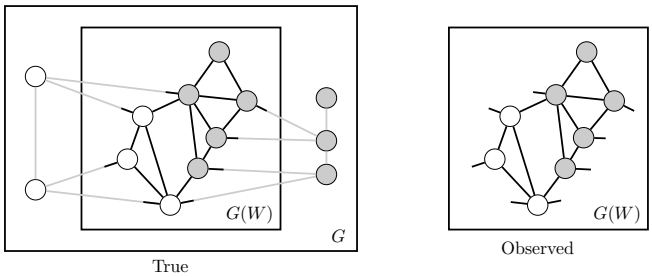

Figure 1: Illustration of the vertex set size estimation problem with $N = 13$ and $K = 2$. White vertices are type-1 and gray are type-2.

Size of the Internet. The number of computers on the Internet (the size of the Internet) is of great interest to computer scientists. However, it is impractical to access and enumerate all computers on the Internet and only a small sample of computers and the connection situation among them are accessible [24].

Counting terrorists. Intelligence agencies often target a small number of suspicious or radicalized individuals to learn about their communication network. But agencies typically do not know the number of people in the network. The number of elements in such a covert network might indicate the size of a terrorist force, and would be of great interest [7].

Epidemiology. Many of the groups at greatest risk for HIV infection (e.g. sex workers, injection drug users, men who have sex with men) are also difficult to survey using conventional methods. Since members of these groups cannot be enumerated directly, researchers often trace social links to reveal a network among known subjects. Public health and epidemiological interventions to mitigate the spread of HIV rely on knowledge of the number of HIV-positive people in the population [12, 11, 22, 23, 8].

Counting disaster victims. After a disaster, it can be challenging to estimate the number of people affected. When logistical challenges prevent all victims from being enumerated, a random sample of individuals may be possible to obtain [2, 3].

In this paper, we propose a novel method called PULSE for estimating the number of vertices and the size of individual communities from a random sub-sample of the network. We model the network as an undirected simple graph $G = (V, E)$, and we treat G as a realization from the stochastic blockmodel (SBM), a widely-studied extension of the Erdős-Rényi random graph model [20] that accommodates community structures in the network by mapping each vertex into one of $K \geq 1$ disjoint types or communities. We construct a sample of the network by choosing a sub-sample of vertices $W \subseteq V$ uniformly at random without replacement, and forming the induced subgraph $G(W)$ of W in G. We assume that the block membership and total degree $d(v)$ of each vertex $v \in W$ are observed. We propose a Bayesian estimation algorithm PULSE for $N = |V|$, the number of vertices in the network, along with the number of vertices N_i in each block. We first prove important regularity results for the posterior distribution of N. Then we describe the conditions under which relevant moments of the posterior distribution exist. We evaluate the performance of PULSE in comparison with the popular "network scale-up" method (NSUM) [12, 11, 22, 23, 8, 14, 9]. We show that while NSUM is asymptotically unbiased, it suffers from serious finite-sample bias and large variance. We show that PULSE has superior performance – in terms of relative error and variance – over NSUM in a wide variety of model and observation scenarios. All proofs are given in the extended version [6].

2 Problem Formulation

The stochastic blockmodel (SBM) is a random graph model that generalizes the Erdős-Rényi random graph [20]. Let $G = (V, E) \sim G(N, K, p, t)$ be a realization from an SBM, where $N = |V|$ is the total number of vertices, the vertices are divided into K types indexed $1, \ldots, K$, specified by the map $t : V \to \{1, \ldots, K\}$, and a type-$i$ vertex and a type-j vertex are connected independently with probability $p_{ij} \in [0, 1]$. Let N_i be the number of type-i vertices in G, with $N = \sum_{i=1}^{K} N_i$. The degree of a vertex v is $d(v)$. An edge is said to be of type-(i, j) if it connects a type-i vertex and a type-j vertex. A random induced subgraph is obtained by sampling a subset $W \subseteq V$ with $|W| = n$ uniformly at random without replacement, and forming the induced subgraph, denoted by $G(W)$. Let V_i be the number of type-i vertices in the sample and E_{ij} be the number of type-(i, j) edges in the sample.

For a vertex v in the sample, a *pendant* edge connects vertex v to a vertex outside the sample. Let $\tilde{d}(v) = d(v) - \sum_{w \in W} 1\{\{w, v\} \in E\}$ be the number of pendant edges incident to v. Let $y_i(v)$ be the number of type-$(t(v), i)$ pendant edges of vertex v; i.e., $y_i(v) = \sum_{w \in V \setminus W} 1\{t(w) = i, \{w, v\} \in E\}$. We have $\sum_{i=1}^{K} y_i(v) = \tilde{d}(v)$. Let $\tilde{N}_i = N_i - V_i$ be the number of type-i nodes outside the sample. We define $\tilde{N} = (\tilde{N}_i : 1 \leq i \leq K)$, $p = (p_{ij} : 1 \leq i < j \leq K)$, and $y = (y_i(v) : v \in W, 1 \leq i \leq K)$. We observe only $G(W)$ and the total degree $d(v)$ of each vertex v in the sample. Assume that we know the type of each vertex in the sample. The observed data $\mathbf{D}$ consists of $G(W)$, $(d(v) : v \in W)$ and $(t(v) : v \in W)$; i.e., $\mathbf{D} = (G(W), (d(v) : v \in W), (t(v) : v \in W))$.

Problem 1. Given the observed data $\mathbf{D}$, estimate the size N of the vertex set $N = |V|$ and the size of each community N_i.

Fig. 1 illustrates the vertex set size estimation problem. White nodes are of type-1 and gray nodes are of type-2. All nodes outside $G(W)$ are unobserved. We observe the types and the total degree of each vertex in the sample. Thus we know the number of pendant edges that connect each vertex in the sample to other, unsampled vertices. However, the destinations of these pendant edges are unknown to us.

3 Network Scale-Up Estimator

We briefly outline a simple and intuitive estimator for $N = |V|$ that will serve as a comparison to PULSE. The network scale-up method (NSUM) is a simple estimator for the vertex set size of Erdős-Rényi random graphs. It has been used in real-world applications to estimate the size of hidden or hard-to-reach populations such as drug users [12], HIV-infected individuals [11, 22, 23], men who have sex with men (MSM) [8], and homeless people [14]. Consider a random graph that follows the Erdős-Rényi distribution. The expected sum of total degrees in a random sample W of vertices is $\mathbb{E}\left[\sum_{v \in W} d(v)\right] = n(N-1)p$. The expected number of edges in the sample is $\mathbb{E}[E_S] = \binom{n}{2}p$, where E_S is the number of edges within the sample. A simple estimator of the connection probability p is $\hat{p} = E_S / \binom{n}{2}$. Plugging $\hat{p}$ into into the moment equation and solving for N yields $\hat{N} = 1 + (n-1)\sum_{v \in W} d(v)/2E_S$, often simplified to $\hat{N}_{NS} = n \sum_{v \in W} d(v)/2E_S$ [12, 11, 22, 23, 8, 14, 9].

Theorem 1. (Proof in [6]) *Suppose G follows a stochastic blockmodel with edge probability $p_{ij} > 0$ for $1 \leq i, j \leq K$. For any sufficiently large sample size, the NSUM is positively biased and $\mathbb{E}[\hat{N}_{NS} | E_S > 0] - N$ has an asymptotic lower bound $\mathbb{E}[\hat{N}_{NS} | E_S > 0] - N \gtrsim N/n - 1$, as n becomes large, where for two sequences $\{a_n\}$ and $\{b_n\}$, $a_n \gtrsim b_n$ means that there exists a sequence c_n such that $a_n \geq c_n \sim b_n$; i.e., $a_n \geq c_n$ for all n and $\lim_{n \to \infty} c_n / b_n = 1$. However, as sample size goes to infinity, the NSUM becomes asymptotically unbiased.*

4 Main Results

NSUM uses only aggregate information about the sum of the total degrees of vertices in the sample and the number of edges in the sample. We propose a novel algorithm, PULSE, that uses individual degree, vertex type, and the network structure information. Experiments (Section 5) show that it outperforms NSUM in terms of both bias and variance.

Given $p = (p_{ij} : 1 \leq i < j \leq K)$, the conditional likelihood of the edges in the sample is given by

$$L_W(\mathbf{D}; p) = \left(\prod_{1 \leq i < j \leq K} p_{ij}^{E_{ij}} (1 - p_{ij})^{V_i V_j - E_{ij}} \right) \times \left(\prod_{i=1}^{K} p_{ii}^{E_{ii}} (1 - p_{ii})^{\binom{V_i}{2} - E_{ii}} \right),$$

and the conditional likelihood of the pendant edges is given by

$$L_{\neg W}(\mathbf{D}; p) = \prod_{v \in W} \sum_{y(v)} \prod_{i=1}^{K} \binom{\tilde{N}_i}{y_i(v)} p_{i,t(v)}^{y_i(v)} (1 - p_{i,t(v)})^{\tilde{N}_i - y_i(v)},$$

where the sum is taken over all $y_i(v)$'s ($i = 1, 2, 3, \ldots, K$) such that $y_i(v) \geq 0, \forall 1 \leq i \leq K$ and $\sum_{i=1}^{K} y_i(v) = \tilde{d}(v)$. Thus the total conditional likelihood is $L(\mathbf{D}; p) = L_W(\mathbf{D}; p) L_{\neg W}(\mathbf{D}; p)$.

If we condition on p and y, the likelihood of the edges within the sample is the same as $L_W(\mathbf{D}; p)$ since it does not rely on y, while the likelihood of the pendant edges given p and y is

$$L_{\neg W}(\mathbf{D}; p, y) = \prod_{v \in W} \prod_{i=1}^{K} \binom{\tilde{N}_i}{y_i(v)} p_{i,t(v)}^{y_i(v)} (1 - p_{i,t(v)})^{\tilde{N}_i - y_i(v)}.$$

Therefore the total likelihood conditioned on p and y is given by $L(\mathbf{D}; p, y) = L_W(\mathbf{D}; p) L_{\neg W}(\mathbf{D}; p, y)$. The conditional likelihood $L(\mathbf{D}; p)$ is indeed a function of $\tilde{N}$. We may view this as the likelihood of $\tilde{N}$ given the data $\mathbf{D}$ and the probabilities p; i.e., $L(\tilde{N}; \mathbf{D}, p) \triangleq L(\mathbf{D}; p)$. Similarly, the likelihood $L(\mathbf{D}; p, y)$ conditioned on p and y is a function of $\tilde{N}$ and y. It can be viewed as the joint likelihood of $\tilde{N}$ and y given the data $\mathbf{D}$ and the probabilities p; i.e., $L(\tilde{N}, y; \mathbf{D}, p) \triangleq L(\mathbf{D}; p, y)$, and $\sum_y L(\tilde{N}, y; \mathbf{D}, p) = L(\tilde{N}; \mathbf{D}, p)$, where the sum is taken over all $y_i(v)$'s, $v \in W$ and $1 \leq i \leq K$, such that $y_i(v) \geq 0$ and $\sum_{i=1}^{K} y_i(v) = \tilde{d}(v)$, $\forall v \in W$, $\forall 1 \leq i \leq K$. To have a full Bayesian approach, we assume that the joint prior distribution for $\tilde{N}$ and p is $\pi(\tilde{N}, p)$. Hence, the population size estimation problem is equivalent to the following optimization problem for $\tilde{N}$:

$$\hat{\tilde{N}} = \arg\max \int L(\tilde{N}; \mathbf{D}, p) \pi(\tilde{N}, p) dp. \tag{1}$$

Then we estimate the total population size as $\hat{N} = \sum_{i=1}^{K} \hat{\tilde{N}}_i + |W|$.

We briefly study the regularity of the posterior distribution of N. In order to learn about $\tilde{N}$, we must observe enough vertices from each block type, and enough edges connecting members of each block, so that the first and second moments of the posterior distribution exist. Intuitively, in order for the first two moments to exist, either we must observe many edges connecting vertices of each block type, or we must have sufficiently strong prior beliefs about p_{ij}.

Theorem 2. (Proof in [6]) *Assume that $\pi(\tilde{N}, p) = \phi(\tilde{N})\psi(p)$ and p_{ij} follows the Beta distribution $\mathrm{B}(\alpha_{ij}, \beta_{ij})$ independently for $1 \leq i < j \leq K$. Let $\lambda = \min_{1 \leq i \leq K} \left(\sum_{j=1}^{K} (E_{ij} + \alpha_{ij}) \right)$. If $\phi(\tilde{N})$ is bounded and $\lambda > n + 1$, then the n-th moment of N exists.*

In particular, if $\lambda > 3$, the variance of N exists. Theorem 2 gives the minimum possible number of edges in the sample to make the posterior sampling meaningful. If the prior distribution of p_{ij} is $\mathrm{Uniform}[0, 1]$, then we need at least three edges incident on type-i edges for all types $i = 1, 2, 3, \ldots, K$ to guarantee the existence of the posterior variance.

4.1 Erdős-Rényi Model

In order to better understand how PULSE estimates the size of a general stochastic blockmodel we study the Erdős-Rényi case where $K = 1$, and all vertices are connected independently with probability p. Let N denote the total population size, W be the sample with size $|W| = V_1$ and $\tilde{N} = N - |W|$. For each vertex $v \in W$ in the sample, let $\tilde{d}(v) = y(v)$ denote the number of pendant edges of vertex v, and $E = E_{11}$ is the number of edges within the sample. Then

$$L_W(\mathbf{D}; p) = p^E (1 - p)^{\binom{|W|}{2} - E}, \qquad L_{\neg W}(\mathbf{D}; p) = \prod_{v \in W} \binom{\tilde{N}}{\tilde{d}(v)} p^{\tilde{d}(v)} (1 - p)^{\tilde{N} - \tilde{d}(v)}.$$

In the Erdős-Rényi case, $y(v) = \tilde{d}(v)$ and thus $L_{\neg W}(\mathbf{D}; p) = L_{\neg W}(\mathbf{D}; p, y)$. Therefore, the total likelihood of $\tilde{N}$ conditioned on p is given by

$$L(\tilde{N}; \mathbf{D}, p) = L_W(\mathbf{D}; p) L_{\neg W}(\mathbf{D}; p) = p^E (1 - p)^{\binom{|W|}{2} - E} \prod_{v \in W} \binom{\tilde{N}}{\tilde{d}(v)} p^{\tilde{d}(v)} (1 - p)^{\tilde{N} - \tilde{d}(v)}.$$

We assume that p has a beta prior $\mathrm{B}(\alpha, \beta)$ and that $\tilde{N}$ has a prior $\phi(\tilde{N})$. Let

$$L(\tilde{N}; \mathbf{D}) = \prod_{v \in W} \binom{\tilde{N}}{\tilde{d}(v)} \mathrm{B}(E + u + \alpha, \binom{|W|}{2} - E + |W|\tilde{N} - u + \beta),$$

where $u = \sum_{v \in W} \tilde{d}(v)$. The posterior probability $\Pr[\tilde{N} | \mathbf{D}]$ is proportional to $\Lambda(\tilde{N}; \mathbf{D}) \triangleq \phi(\tilde{N}) L(\tilde{N}; \mathbf{D})$. The algorithm is presented in Algorithm 1.

Algorithm 1 Population size estimation algorithm PULSE (Erdős-Rényi case)

Input: Data $\mathbf{D}$; initial guess for $\hat{N}$, denoted by $N(0)$; parameters of the beta prior, α and β

Output: Estimate for the population size $\hat{N}$

$\tilde{N}(0) \leftarrow N(0) - |W|$
$\tau \leftarrow 1$
repeat
  Propose $\tilde{N}'(\tau)$ according to a proposal distribution $g(\tilde{N}(\tau-1) \to \tilde{N}'(\tau))$
  $q \leftarrow \min\{1, \frac{\Lambda(\tilde{N}'(\tau);\mathbf{D})g(\tilde{N}'(\tau)\to\tilde{N}(\tau-1))}{\Lambda(\tilde{N}(\tau-1);\mathbf{D})g(\tilde{N}(\tau-1)\to\tilde{N}'(\tau))}\}$
  $\tilde{N}(\tau) \leftarrow \tilde{N}'(\tau)$ with probability q; otherwise $\tilde{N}(\tau) \leftarrow \tilde{N}(\tau-1)$
  $\tau \leftarrow \tau + 1$
until some termination condition is satisfied
Look at $\{\tilde{N}(\tau) : \tau > \tau_0\}$ and view it as the sampled posterior distribution for $\tilde{N}$
Let $\hat{\tilde{N}}$ be the posterior mean with respect to the sampled posterior distribution.

Algorithm 2 Population size estimation algorithm PULSE (general stochastic blockmodel case)

Input: Data $\mathbf{D}$; initial guess for $\tilde{N}$, denoted by $\tilde{N}^{(0)}$; initial guess for y, denoted by $y^{(0)}$; parameters of the beta prior, α_{ij} and β_{ij}, $1 \le i \le j \le K$.

Output: Estimate for the population size $\hat{N}$

$\tau \leftarrow 1$
repeat
  Randomly decide whether to update $\tilde{N}$ or y
  if update $\tilde{N}$ **then**
    Randomly selects $i \in [1, K] \cap \mathbb{N}$.
    $\tilde{N}^* \leftarrow \tilde{N}^{(\tau-1)}$
    Propose $\tilde{N}_i^*$ according to the proposal distribution $g_i(\tilde{N}_i^{(\tau-1)} \to \tilde{N}_i^*)$
    $q \leftarrow \min\{1, \frac{\Lambda(\tilde{N}^*,y;\mathbf{D})g_i(\tilde{N}_i^*\to\tilde{N}_i^{(\tau-1)})}{\Lambda(\tilde{N}^{(\tau-1)},y;\mathbf{D})g_i(\tilde{N}_i^{(\tau-1)}\to\tilde{N}_i^*)}\}$
    $\tilde{N}^{(\tau)} \leftarrow \tilde{N}^*$ with probability q; otherwise $\tilde{N}^{(\tau)} \leftarrow \tilde{N}^{(\tau-1)}$.
    $y^{(\tau)} \leftarrow y^{(\tau-1)}$
  else
    Randomly selects $v \in W$.
    $y^* \leftarrow y^{(\tau-1)}$
    Propose $y(v)^*$ according to the proposal distribution $h_v(y(v)^{(\tau-1)} \to y(v)^*)$
    $q \leftarrow \min\{1, \frac{L(\tilde{N},y^*;\mathbf{D})h_v(y(v)^*\to y(v)^{(\tau-1)})}{L(\tilde{N},y;\mathbf{D})h_v(y(v)^{(\tau-1)}\to y(v)^*)}\}$
    $y^{(\tau)} \leftarrow y^*$ with probability q; otherwise $y^{(\tau)} \leftarrow y^{(\tau-1)}$.
    $\tilde{N}^{(\tau)} \leftarrow \tilde{N}^{(\tau-1)}$
  end if
  $\tau \leftarrow \tau + 1$
until some termination condition is satisfied
Look at $\{\tilde{N}(\tau) : \tau > \tau_0\}$ and view it as the sampled posterior distribution for $\tilde{N}$
Let $\hat{N}$ be the posterior mean of $\sum_{i=1}^K \tilde{N}_i + |W|$ with respect to the sampled posterior distribution.

4.2 General Stochastic Blockmodel

In the Erdős-Rényi case, $y(v) = \tilde{d}(v)$. However, in the general stochastic blockmodel case, in addition to the unknown variables $\tilde{N}_1, \tilde{N}_2, \ldots, \tilde{N}_K$ to be estimated, we do not know $y_i(v)$ ($v \in W$, $i = 1, 2, 3, \ldots, K$) either. The expression $L_{\neg W}(\mathbf{D}; p)$ involves costly summation over all possibilities of integer composition of $\tilde{d}(v)$ ($v \in W$). However, the joint posterior distribution for $\tilde{N}$ and y, which is proportional to $\int L(\tilde{N}, y; \mathbf{D}, p)\phi(\tilde{N})\psi(p)dp$, does not involve summing over integer partitions; thus we may sample from the joint posterior distribution for $\tilde{N}$ and y, and obtain the marginal distribution for $\tilde{N}$. Our proposed algorithm PULSE realizes this idea. Let $L(\tilde{N}, y; \mathbf{D}) = \int L(\tilde{N}, y; \mathbf{D}, p)\psi(p)dp$. We know that the joint posterior distribution for $\tilde{N}$ and y, denoted by $\Pr[\tilde{N}, y|\mathbf{D}]$, is proportional to $\Lambda(\tilde{N}, y; \mathbf{D}) \triangleq L(\tilde{N}, y; \mathbf{D})\psi(\tilde{N})$. In addition, the conditional distributions $\Pr[\tilde{N}_i|\tilde{N}_{\neg i}, y]$ and $\Pr[y(v)|\tilde{N}, y(\neg v)]$ are also proportional to $L(\tilde{N}, y; \mathbf{D})\psi(\tilde{N})$, where $\tilde{N}_{\neg i} = (\tilde{N}_j : 1 \le j \le K, j \ne i)$, $y(v) = (y_i(v) : 1 \le i \le K)$ and $y(\neg v) = (y(w) : w \in W, w \ne v)$. The proposed algorithm PULSE is a Gibbs sampling process that samples from the joint posterior distribution (i.e., $\Pr[\tilde{N}, y|\mathbf{D}]$), which is specified in Algorithm 2.

For every $v \in W$ and $i = 1, 2, 3, \ldots, K$, $0 \le y_i(v) \le \tilde{N}_i$ because the number of type-$(i, t(v))$ pendant edges of vertex v must not exceed the total number of type-i vertices outside the sample. Therefore, we have $\tilde{N}_i \ge \max_{v\in W} y_i(v)$ must hold for every $i = 1, 2, 3, \ldots, K$. These observations put constraints on the choice of proposal distributions g_i and h_v, $i = 1, 2, 3, \ldots, K$ and $v \in W$; i.e., the support of g_i must be contained in $[\max_{v\in W} y_i(v), \infty) \cap \mathbb{N}$ and the support of h_v must be contained in $\{y(v) : \forall 1 \le i \le K, 0 \le y_i(v) \le \tilde{N}_i, \sum_{j=1}^K y_i(v) = \tilde{d}(v)\}$.

Let ω_i be the window size for $\tilde{N}_i$, taking values in $\mathbb{N}$. Let $l = \max\{\max_{v \in W} y_i(v), \tilde{N}_i^{(\tau-1)} - \omega_i\}$. Let the proposal distribution g_i be defined as below:

$$g_i(\tilde{N}_i^{(\tau-1)} \to \tilde{N}_i^*) = \begin{cases} \frac{1}{2\omega_i+1} & \text{if } l \le \tilde{N}_i^* \le l + 2\omega_i \\ 0 & \text{otherwise.} \end{cases}$$

The proposed value $\tilde{N}_i^*$ is always greater than or equal to $\max_{v \in W} y_i(v)$. This proposal distribution uniform within the window $[l, l + 2\omega_i]$, and thus the proposal ratio is $g_i(\tilde{N}_i^* \to \tilde{N}_i^{(\tau-1)})/g_i(\tilde{N}_i^{(\tau-1)} \to \tilde{N}_i^*) = 1$. The proposal for $y(v)$ is detailed in the extended version [6].

5 Experiment

5.1 Erdős-Rényi

Effect of Parameter p. We first evaluate the performance of PULSE in the Erdős-Rényi case. We fix the size of the network at $N = 1000$ and the sample size $|W| = 280$ and vary the parameter p. For each $p \in [0.1, 0.9]$, we sample 100 graphs from $G(N, p)$. For each selected graph, we compute NSUM and run PULSE 50 times (as it is a randomized algorithm) to compute its performance. We record the relative errors by the Tukey boxplots shown in Fig. 2a. The posterior mean proposed by PULSE is an accurate estimate of the size. For the parameter p varying from 0.1 to 0.9, most of the relative errors are bounded between -1% and 1%. We also observe that the NSUM tends to overestimate the size as it shows a positive bias. This confirms experimentally the result of Theorem 1. For both methods, the interquartile ranges (IQRs, hereinafter) correlate negatively with p. This shows that the variance of both estimators shrinks when the graph becomes denser. The relative errors of PULSE tend to concentrate around 0 with larger p which means that the performance of PULSE improves with larger p. In contrast, a larger p does not improve the bias of the NSUM.

Effect of Network Size N. We fix the parameter $p = 0.3$ and the sample size $|W| = 280$ and vary the network size N from 400 to 1000. For each $N \in [400, 1000]$, we randomly pick 100 graphs from $G(N, p)$. For each selected graph, we compute NSUM and run PULSE 50 times. We illustrate the results via Tukey boxplots in Fig. 2b. Again, the estimates given by PULSE are very accurate. Most of the relative errors reside in $[-0.5\%, 0.5\%]$ and almost all reside in $[-1\%, 1\%]$. We also observe that smaller network sizes can be estimated more accurately as PULSE will have a smaller variance. For example, when the network size is $N = 400$, almost all of the relative errors are bounded in the range $[-0.7\%, 0.7\%]$ while for $N = 1000$, the relative errors are in $[-1.5\%, 1.5\%]$. This agrees with our intuition that the performance of estimation improves with a larger sampling fraction. In contrast, NSUM heavily overestimates the network size as the size increases. In addition, its variance also correlates positively with network size.

Effect of Sample Size $|W|$. We study the effect of the sample size $|W|$ on the estimation error. Thus, we fix the size $N = 1000$ and the parameter $p = 0.3$, and we vary the sample size $|W|$ from 100 to 500. For each $|W| \in [100, 500]$, we randomly select 100 graphs from $G(N, p)$. For every selected graph, we compute the NSUM estimate, run PULSE 50 times, and record the relative errors. The results are presented in Fig. 2c. We observe that for both methods that the IQR shrinks as the sample size increases; thus a larger sample size reduces the variance of both estimators. PULSE does not exhibit appreciable bias when the sample size varies from 100 to 500. Again, NSUM overestimates the size; however, its bias reduces when the sample size becomes large. This reconfirms Theorem 1.

5.2 General Stochastic Blockmodel

Effect of Sample Size and Type Partition. Here, we study the effect of the sample size and the type partition. We set the network size N to 200 and we assume that there are two types of vertices in this network: type 1 and type 2 with N_1 and N_2 nodes, respectively. The ratio N_1/N quantifies the type partition. We vary N_1/N from 0.2 to 0.8 and the sample size $|W|$ from 40 to 160. For each combination of N_1/N and the sample size $|W|$, we generate 50 graphs with $p_{11}, p_{22} \sim \text{Uniform}[0.5, 1]$ and $p_{12} = p_{21} \sim \text{Uniform}[0, \min\{p_{11}, p_{22}\}]$. For each graph, we compute the NSUM and obtain the average relative error. Similarly, for each graph, we run PULSE 10 times in order to compute the average relative error for the 50 graphs and 10 estimates for each graph. The results are shown as heat maps in Fig. 2d. Note that the color bar on the right side of Fig.

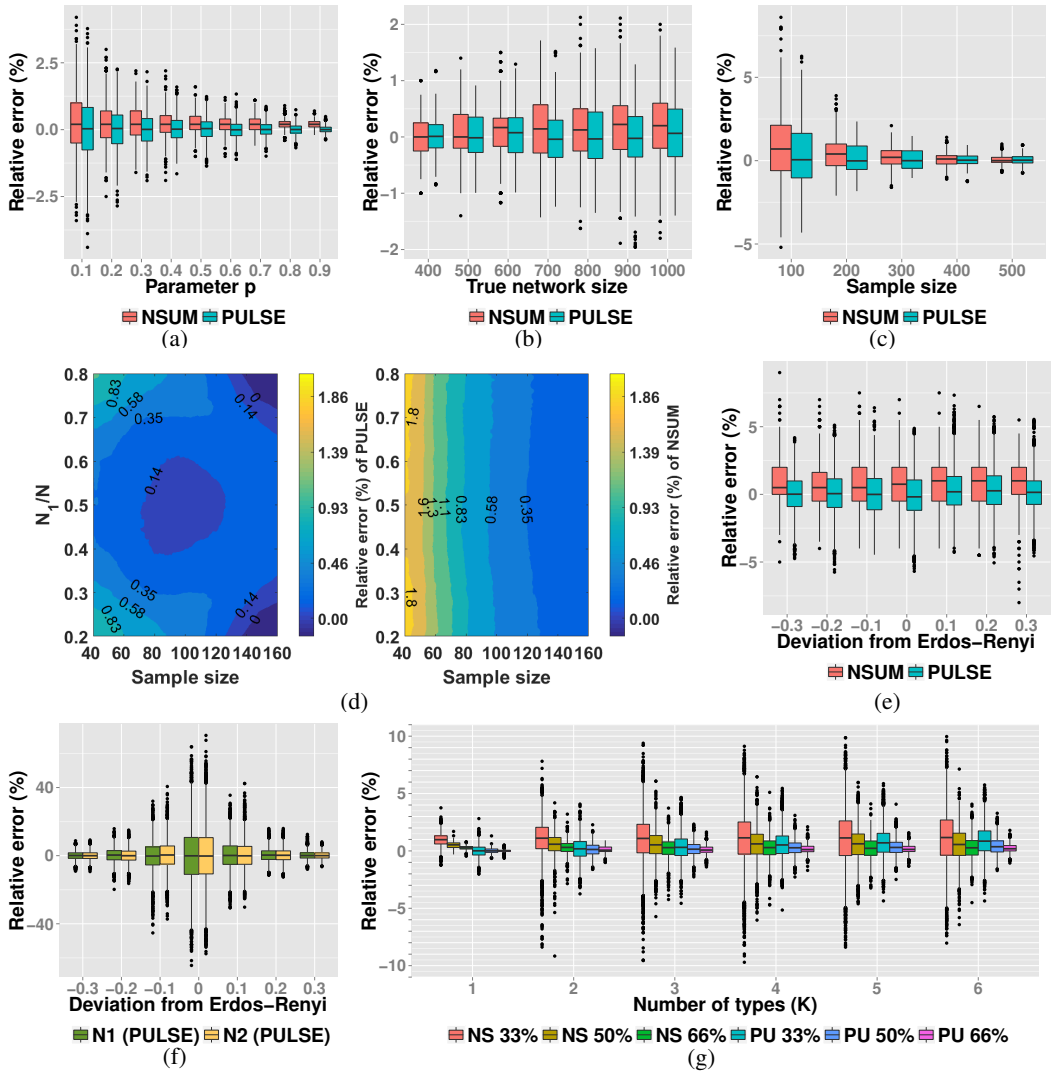

Figure 2: Fig. 2a, 2b and 2c are the results of the Erdős-Rényi case: (a) Effect of parameter p on the estimation error. (b) Effect of the network size on the estimation error. (c) Effect of the sample size on the estimation error. Fig. 2d, 2e, 2f and 2g are the results of the general SBM case: (d) Effect of sample size and type partition on the relative error. Note that the color bar on the right is on logarithmic scale. (e) Effect of deviation from the Erdős-Rényi model (controlled by ϵ) on the relative error of NSUM and PULSE in the SBM with $K = 2$. (f) Effect of deviation from the Erdős-Rényi model (controlled by ϵ) on the relative error of PULSE in estimating the number of type-1 and type-2 nodes in the SBM with $K = 2$. (g) Effect of the number of types K and the sample size on the population estimation. The percentages are the sampling fractions n/N. The horizontal axis represents the number of types K that varies from 1 to 6. The vertical axis is the relative error in percentage.

2d is on logarithmic scale. In general, the estimates given by PULSE are very accurate and exhibit significant superiority over the NSUM estimates. The largest relative errors of PULSE in absolute value, which are approximately 1%, appear in the upper-left and lower-left corner on the heat map. The performance of the NSUM (see the right subfigure in Fig. 2d) is robust to the type partition and equivalently the ratio N_1/N. As we enlarge the sample size, its relative error decreases.

The left subfigure in Fig. 2d shows the performance of PULSE. When the sample size is small, the relative error decreases as N_1/N increases from 0.2 to 0.5; when N_1/N rises from 0.5 to 0.8, the relative error becomes large. Given the fixed ratio N_1/N, as expected, the relative error declines when we have a larger sample. This agrees with our observation in the Erdős-Rényi case. However, when the sample size is large, PULSE exhibits better performance when the type partition is more homogeneous. There is a local minimum relative error in absolute value shown at the center of the subfigure. PULSE performs best when there is a balance between the number of edges in the sampled

induced subgraph and the number of pendant edges emanating outward. Larger sampled subgraphs allow more precision in knowledge about p_{ij}, but more pendant edges allow for better estimation of y, and hence each N_i. Thus when the sample is about half of the network size, the balanced combination of the number of edges within the sample and those emanating outward leads to better performance.

Effect of Intra- and Inter-Community Edge Probability. Suppose that there are two types of nodes in the network. The mean degree is given by $d_{\text{mean}} = \frac{2}{N}\left[\binom{N_1}{2}p_{11} + \binom{N_1}{2}p_{22} + N_1N_2p_{12}\right]$. We want to keep the mean degree constant and vary the random graph gradually so that we observe 3 phases: high intra-community and low inter-community edge probability (more cohesive), Erdős-Rényi , and low intra-community and high inter-community edge probability (more incohesive). We introduce a cohesion parameter ϵ. In the two-block model, we have $p_{11} = p_{22} = p_{01} = \tilde{p}$, where $\tilde{p}$ is a constant. Let's call ϵ the deviation from this situation and let $p_{11} = \tilde{p} + \frac{N_1N_2\epsilon}{2\binom{N_1}{2}}, p_{22} = \tilde{p} + \frac{N_1N_2\epsilon}{2\binom{N_2}{2}}, p_{12} = \tilde{p} - \epsilon$. The mean degree stays constant for different ϵ. In addition, p_{11}, p_{12} and p_{22} must reside in $[0, 1]$. This requirement can be met if we set the absolute value of ϵ small enough. By changing ϵ from positive to negative we go from cohesive behavior to incohesive behavior. Clearly, for $\epsilon = 0$, the graph becomes an Erdős-Rényi graph with $p_{11} = p_{22} = p_{01} = \tilde{p}$.

We set the network size N to 850, N_1 to 350, and N_2 to 500. We fix $\tilde{p} = 0.5$ and let ϵ vary from -0.3 to 0.3. When $\epsilon = 0.3$, the intra-community edge probabilities are $p_{11} = 0.9298$ and $p_{22} = 0.7104$ and the inter-community edge probability is $p_{12} = 0.2$. When $\epsilon = -0.3$, the intra-community edge probabilities are $p_{11} = 0.0702$ and $p_{22} = 0.2896$ and the inter-community edge probability is $p_{12} = 0.8$. For each ϵ, we generate 500 graphs and for each graph, we run PULSE 50 times. Given each value of ϵ, relative errors are shown in box plots. We present the results in Fig. 2e as we vary ϵ. From Fig. 2e, we observe that despite deviation from the Erdős-Rényi graph, both methods are robust. However, the figure indicates that PULSE is unbiased (as median is around zero) while NSUM overestimates the size on average. This again confirms Theorem 1.

An important feature of PULSE is that it can also estimate the number of nodes of each type while NSUM cannot. The results for type-1 and type-2 with different ϵ are shown in Fig. 2f. We observe that the median of all boxes agree with the 0% line; thus the separate estimates for N_1 or N_2 are unbiased. Note that when the edge probabilities are more homogeneous (i.e., when the graph becomes more similar to the Erdős-Rényi model) the IQRs, as well as the interval between the two ends of the whiskers, become larger. This shows that when we try to fit an Erdős-Rényi model (a single-type stochastic blockmodel) into a two-type model, the variance becomes larger.

Effect of Number of Types and Sample Size. Finally, we study the impact of the number of types K and the sample size $|W| = n$ on the relative error. To generate graphs with different number of types, we use a Chinese restaurant process (CRP) [1]. We set the total number of vertices to 200, first pick 100 vertices and use the Chinese restaurant process to assign them to different types. Suppose that CRP gives K types; We then distribute the remaining 100 vertices evenly among the K types. The edge probability p_{ii} $(1 \leq i \leq K)$ is sampled from $\text{Uniform}[0.7, 1]$ and p_{ij} $(1 \leq i < j \leq K)$ is sampled from $\text{Uniform}[0, \min\{p_{ii}, p_{jj}\}]$, all independently. We set the sampling fraction n/N to 33%, 50% and 66%, and use NSUM and PULSE to estimate the network size. Relative estimation errors are illustrated in Fig. 2g. We observe that with the same sampling fraction n/N and same the number of types K, PULSE has a smaller relative error than that of the NSUM. Similarly, the interquartile range of PULSE is also smaller than that of the NSUM. Hence, PULSE provides a higher accuracy with a smaller variance. For both methods the relative error decreases (in absolute value) as the sampling fraction increases. Accordingly, the IQRs also shrink for larger sampling fraction. With the sampling fraction fixed, the IQRs become larger when we increase the number of types in the graph. The variance of both methods increases for increasing values of K. The median of NSUM is always above 0 on average which indicates that it overestimates the network size.

Acknowledgements

This research was supported by Google Faculty Research Award, DARPA Young Faculty Award (D16AP00046), NIH grants from NICHD DP2HD091799, NCATS KL2 TR000140, and NIMH P30 MH062294, the Yale Center for Clinical Investigation, and the Yale Center for Interdisciplinary Research on AIDS. LC thanks Zheng Wei for his consistent support.

References

[1] D. J. Aldous. *Exchangeability and related topics*. Springer, 1985.

[2] H. Bernard, E. Johnsen, P. Killworth, and S. Robinson. How many people died in the mexico city earthquake. *Estimating the Number of People in an Average Network and in an Unknown Event Population. The Small World, ed. M. Kochen (forthcoming). Newark*, 1988.

[3] H. R. Bernard, P. D. Killworth, E. C. Johnsen, G. A. Shelley, and C. McCarty. Estimating the ripple effect of a disaster. *Connections*, 24(2):18–22, 2001.

[4] M. S. Bernstein, E. Bakshy, M. Burke, and B. Karrer. Quantifying the invisible audience in social networks. In *Proc. SIGCHI*, pages 21–30. ACM, 2013.

[5] L. Chen, F. W. Crawford, and A. Karbasi. Seeing the unseen network: Inferring hidden social ties from respondent-driven sampling. In *AAAI*, pages 1174–1180, 2016.

[6] L. Chen, A. Karbasi, and F. W. Crawford. Estimating the size of a large network and its communities from a random sample. *arXiv preprint arXiv:1610.08473*, 2016. `https://arxiv.org/abs/1610.08473`.

[7] F. W. Crawford. The graphical structure of respondent-driven sampling. *Sociological Methodology*, 46(1):187–211, 2016.

[8] S. Ezoe, T. Morooka, T. Noda, M. L. Sabin, and S. Koike. Population size estimation of men who have sex with men through the network scale-up method in Japan. *PLoS One*, 7(1):e31184, 2012.

[9] D. M. Feehan and M. J. Salganik. Generalizing the network scale-up method: a new estimator for the size of hidden populations. *Sociological Methodology*, 46(1):153–186, 2016.

[10] M. Girvan and M. E. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[11] W. Guo, S. Bao, W. Lin, G. Wu, W. Zhang, W. Hladik, A. Abdul-Quader, M. Bulterys, S. Fuller, and L. Wang. Estimating the size of HIV key affected populations in Chongqing, China, using the network scale-up method. *PLoS One*, 8(8):e71796, 2013.

[12] C. Kadushin, P. D. Killworth, H. R. Bernard, and A. A. Beveridge. Scale-up methods as applied to estimates of heroin use. *Journal of Drug Issues*, 2006.

[13] L. Katzir, E. Liberty, and O. Somekh. Estimating sizes of social networks via biased sampling. In *WWW*, pages 597–606. ACM, 2011.

[14] P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach. *Eval. Rev.*, 22(2):289–308, 1998.

[15] A. S. Maiya and T. Y. Berger-Wolf. Benefits of bias: Towards better characterization of network sampling. In *Proc. SIGKDD*, pages 105–113. ACM, 2011.

[16] L. Massoulié, E. Le Merrer, A.-M. Kermarrec, and A. Ganesh. Peer counting and sampling in overlay networks: random walk methods. In *Proc. PODC*, pages 123–132. ACM, 2006.

[17] B. H. Murray and A. Moore. Sizing the internet. *White paper, Cyveillance*, page 3, 2000.

[18] M. E. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8582, 2006.

[19] M. Papagelis, G. Das, and N. Koudas. Sampling online social networks. *TKDE*, 25(3):662–676, 2013.

[20] A. Rényi and P. Erdős. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[21] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *Proc. IMC*, pages 390–403. ACM, 2010.

[22] M. J. Salganik, D. Fazito, N. Bertoni, A. H. Abdo, M. B. Mello, and F. I. Bastos. 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*, 174(10):1190–1196, 2011.

[23] M. Shokoohi, M. R. Baneshi, and A.-a. Haghdoost. Size estimation of groups at high risk of HIV/AIDS using network scale up in Kerman, Iran. *Int'l J. Prev. Medi.*, 3(7):471, 2012.

[24] S. Xing and B.-P. Paris. Measuring the size of the internet via importance sampling. *J. Sel. Areas Commun*, 21(6):922–933, 2003.