Adaptive Web Sampling

Leonardo Agate (1945534)
Francesco Natali (1945581)

Università degli Studi La Sapienza,
Statistical Methods and Applications – Data Analyst

May 29, 2025

Abstract

This study explores advancements in adaptive sampling strategies for network and spatial settings. Link-tracing methods, such as adaptive cluster sampling, snowball sampling, and targeted random walk designs, improve upon conventional approaches but face challenges, including inflexible sample placement, variable sample sizes, and limited efficiency in estimating population parameters. Adaptive web sampling (AWS), introduced by [20], addresses these limitations by offering enhanced flexibility through design variations. This study analyzes simulation results from [20] and [23], focusing on the Strategy 1 of AWS applied to two empirical populations: an HIV/AIDS at-risk population and a blue-winged teal bird population. The results demonstrate that AWS estimators, particularly $\hat{\mu}_1$ and $\hat{\mu}_4$ for HIV/AIDS and $\hat{\mu}_2$ and $\hat{\mu}_3$ for the bird population, achieve low mean squared errors, with Rao-Blackwellization further improving precision. AWS proves to be a versatile and efficient method for complex sampling scenarios.

Keywords: Adaptive sampling, Link-tracing designs, Markov chain Monte Carlo, Network sampling, Spatial sampling, Estimation methods.

1 Introduction

Adaptive web sampling (AWS) is a versatile class of sampling designs developed for populations with network or spatial structures. This method employs sequential selections based on a mixture distribution, guided by an evolving active set that adapts as sampling progresses. Selections are influenced by network or spatial relationships and observed sample values, enabling AWS to allocate sampling effort disproportionately toward high-value or interesting population segments. Unlike traditional link-tracing methods, AWS provides precise control over sample sizes and the balance between adaptive and conventional sampling efforts. The design incorporates efficient inference techniques, accounting for all possible sample paths consistent with the minimal sufficient statistic, with Markov chain resampling ensuring computational feasibility.

1.1 Network Sampling

In a network-structured population, units (or nodes) represent entities such as individuals, and edges (or links) denote relationships, such as social connections. Units are labeled $1, 2, \ldots, N$, with each unit i associated with an observable variable of interest y_i, which may take any numerical value in general settings. In specific cases, y_i is an indicator variable defined as:

$$y_i = \begin{cases} 1 & \text{if unit } i \text{ is a unit of interest} \\ 0 & \text{otherwise} \end{cases}$$

A unit i is deemed a "unit of interest" based on specific traits, such as possessing a particular characteristic. For each ordered pair of units (i, j), the variable w_{ij} quantifies the existence or strength of their relationship, defining the graph structure. Generally, w_{ij} can be any numerical value, representing measures like physical distance or interaction frequency. In simpler settings, it is an indicator variable:

$$w_{ij} = \begin{cases} 1 & \text{if there is a link from unit } i \text{ to unit } j \\ 0 & \text{otherwise} \end{cases}$$

For simplicity, we set $w_{ii} = 0$ for all $i = 1, 2, \ldots, N$. A sample consists of selected nodes and pairs of nodes, with observed values y_i and w_{ij}. Additionally, the out-degree w_{i+}, the number of links emanating from node i, is recorded for sampled units.

In studies of hard-to-reach populations, such as individuals at high risk for HIV, link-tracing designs like AWS are often the most practical approach to obtain sufficiently large samples. These designs follow social links from current sample members to recruit new members, adapting selections based on observed node and link values. Fig. 1 illustrates a simulated network population with 147 members, comprising injection and non-injection drug users. Dark nodes, representing injection drug users, are units of interest ($y_i = 1$), while light nodes have $y_i = 0$. Symmetric links exist between nodes sharing drug paraphernalia or sexual contact, with link probabilities following a logistic distribution based on node status and distance [10].

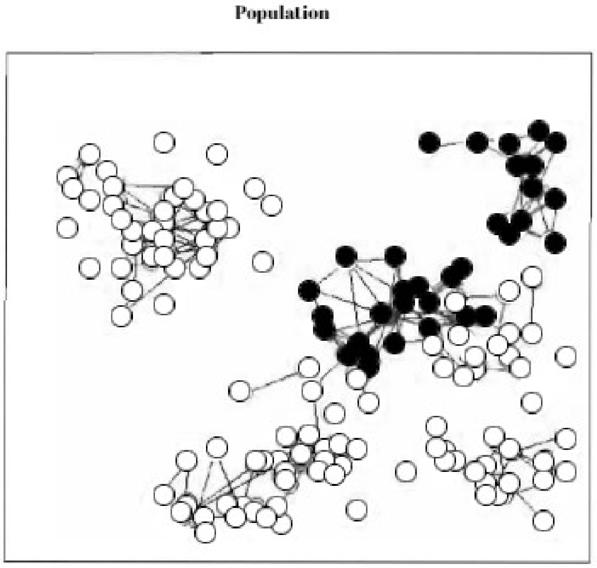


Fig. 1: A simulated network population with 147 nodes, where dark nodes represent injection drug users ($y_i = 1$) and light nodes represent non-injection drug users ($y_i = 0$). Links are symmetric and follow a logistic distribution.

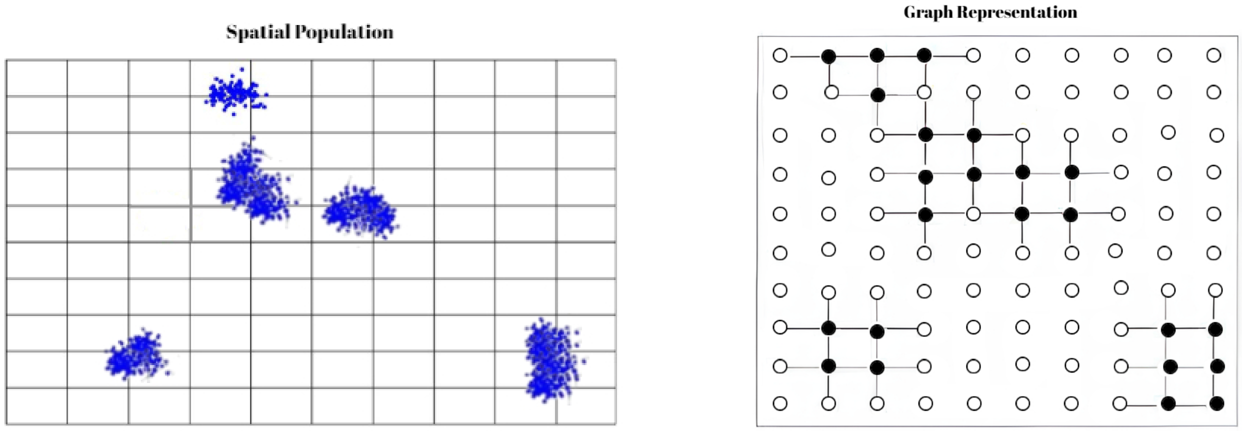
1.2 Spatial Sampling

In a spatial setting, a geographical area is partitioned into units, such as grid squares. For instance, in the simulated spatial population shown in Fig. 2, each square represents a unit, with y_i denoting the count of point-objects (e.g., animals) within it. The link variable w_{ij} is defined as:

$$w_{ij} = \begin{cases} 1 & \text{if units } i \text{ and } j \text{ are adjacent and unit } i \text{ is a unit of interest} \\ 0 & \text{otherwise} \end{cases}$$

Units i and j are adjacent if i is directly above, below, left, or right of j. Symmetry in links occurs only when both units are of interest and adjacent. Units of interest are those with at least one point-object ($y_i \geq 1$), depicted as dark nodes in the graph representation.

Fig. 2 presents a simulated spatial population divided into 100 plots, with point-objects clustered around randomly selected centers following a symmetric Gaussian distribution (standard deviation 0.03). The left panel shows the spatial distribution, while the right panel illustrates the graph representation, highlighting units of interest and one-way links. This setup models populations with clustering behavior, such as plants, fish, or deer, where y_i represents the number of individuals in each plot. Subsequent sections will estimate the mean number of point-objects per plot.



(a) Spatial distribution

(b) Graph representation

Fig. 2: A simulated spatial population with 100 plots. The left panel shows the spatial distribution of point-objects, and the right panel depicts the graph representation, with dark nodes indicating units of interest ($y_i \geq 1$) and directed links showing adjacency.

2 Adaptive Web Sampling

2.1 Sampling Setup

A population of interest comprises units labeled $1, 2, \ldots, N$, each associated with a variable of interest y_i. In a network or graph setting, additional structure is defined by link variables w_{ij}, which describe relationships between pairs of units i and j. Typically, w_{ij} is an indicator variable:

$$w_{ij} = \begin{cases} 1 & \text{if there is a link from unit } i \text{ to unit } j \\ 0 & \text{otherwise} \end{cases}$$

More generally, w_{ij} represents a weight, such as the strength of a relationship. These link variables, along with node variables y_i, determine the population's graph structure and are observed only through sampling.

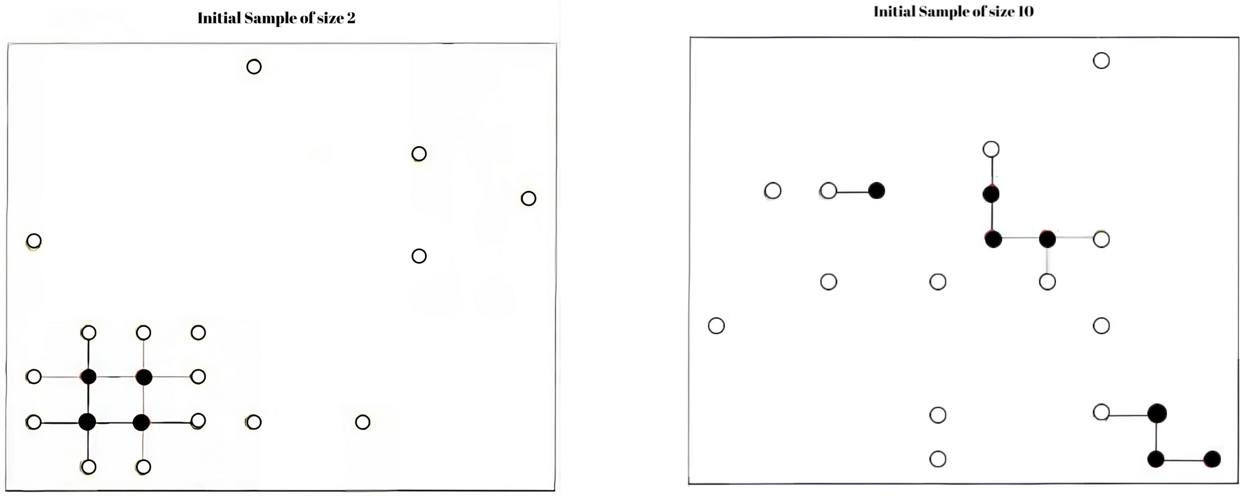
A sample s consists of a subset of units, divided into a node sample $s^{(1)}$, where y_i values are observed, and a pair sample $s^{(2)}$, where w_{ij} values are observed. A design is adaptive if selections depend on observed values of y_i or w_{ij}. The original data d_0 include the sequence of selected unit labels and their y_i values, with possible repetitions in with-replacement designs. The minimal sufficient statistic, however, is the reduced data $d_r = \{(i, y_i), ((j, k), w_{jk}) : i \in s^{(1)}, (j, k) \in s^{(2)}\}$, comprising distinct units and pairs with their associated values [8, 1, 22]. For instance, if node $i \in s^{(1)}$ but $j \notin s^{(1)}$, y_i is known, but y_j is not; yet, w_{ij} may be observed if $(i, j) \in s^{(2)}$. The out-degree $w_{i+} = \sum_j w_{ij}$, a node variable, is also recorded for sampled units.

2.2 Designs

Adaptive web sampling begins with an initial sample s_0 of size n_0, selected with probability p_0 using a conventional design, such as simple random sampling (SRS). At step k, the next sample segment s_k is chosen based on values associated with an active set $a_k \subseteq s_{ck}$, where $s_{ck} = \bigcup_{i=0}^{k-1} s_i$ is the current sample. The active set's flexibility, which may include all sampled nodes or only those with high interest and external links, distinguishes AWS from random walk designs, where selections are restricted to the most recently sampled unit.

To avoid over-sampling clustered nodes, AWS employs a mixture distribution for selection probabilities. This distribution combines two components: an adaptive component, based on observed values in the active set, and a conventional component, typically SRS. These are weighted by d and $1 - d$, respectively, where d, the dampening value, controls the emphasis on link-tracing. With probability d, a link from the active set is randomly selected, adding the connected unit to the sample; with probability $1 - d$, a unit is chosen randomly from the remaining population (without-replacement) or the entire population (with-replacement). The value of d may vary based on active set values, enhancing flexibility.

The initial sample size n_0 and dampening value d significantly influence the final sample's composition. Larger n_0 ensures broader population coverage, reducing bias, while smaller n_0 with high d focuses on clustered nodes. Fig. 3 illustrates this trade-off in a simulated spatial population with six clusters, using $d = 0.9$ and final sample size $n = 20$. The left panel, with a smaller initial sample ($n_0 = 5$), samples all nodes in one cluster, missing others. The right panel, with a larger initial sample ($n_0 = 15$), covers nodes from four clusters, demonstrating wider exploration.



(a) Small initial sample ($n_0 = 5$)

(b) Large initial sample ($n_0 = 15$)

Fig. 3: Two adaptive web samples from a simulated spatial population with six clusters, using $d = 0.9$ and final sample size $n = 20$. The left sample focuses on one cluster, while the right sample covers multiple clusters.

Selections can occur unit-by-unit or in waves, where the active set remains constant for multiple selections. Unlike snowball or adaptive cluster sampling, AWS does not require sampling all links or connected components, allowing precise control over sample size and exploration depth.

Mathematically, at step k, the current sample is $s_{ck} = \bigcup_{i=0}^{k-1} s_i$, with active set a_k, containing n_{ak} units, and n_{ck} units in s_{ck}. The next set s_k is selected with probability $q_k(s_k|a_k, y_{ak}, w_{ak})$. A common design selects one node i per wave with probability proportional to the number of links from a_k to i. For without-replacement sampling:

$$q_{ki} = d\frac{w_{a_k i}}{w_{a_k +}} + (1-d)\frac{1}{N - n_{s_{ck}}},$$

where $w_{a_k i} = \sum_{j \in a_k} w_{ji}$ is the number of links from a_k to node $i \notin s_{ck}$, and $w_{a_k +} = \sum_{i \in a_k, j \notin s_{ck}} w_{ij}$ is the total links from a_k to unsampled units. For with-replacement sampling:

$$q_{ki} = d\frac{w_{a_k i}}{w_{a_k +}} + (1-d)\frac{1}{N},$$

where $w_{a_k +} = \sum_{i \in a_k, j=1,\dots,N} w_{ij}$. If no links exist from a_k, then $q_{ki} = \frac{1}{N - n_{s_{ck}}}$ (without-replacement) or $q_{ki} = \frac{1}{N}$ (with-replacement).

Fig. 4 illustrates a without-replacement design at step k, where the active set has four outgoing links, two to node i. The selection probability is $q_{ki} = d\frac{2}{4} + (1-d)\frac{1}{N-6}$.

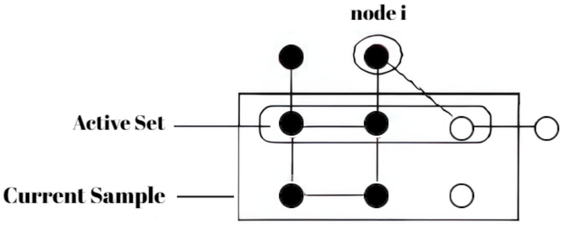


Fig. 4: Example of inclusion probability for a node i given the current active set with four outgoing links, two to i, in a without-replacement design.

For a sample s with one node per step, the overall selection probability is:

$$p(s) = \prod_{k=1}^{n-n_0} q_{k i_k},$$

where i_k is the node selected at step k. For $n_k > 1$ nodes in wave k, the selection probability for the t-th unit is:

$$q_{kti} = d\frac{w_{a_k i}}{w_{a_{kt} +}} + (1-d)\frac{1}{N - n_{s_{ckt}}},$$

(without-replacement) or:

$$q_{kti} = d\frac{w_{a_k i}}{w_{a_{kt} +}} + (1-d)\frac{1}{N},$$

(with-replacement), where $w_{a_{kt}+} = \sum_{i \in a_k, j \notin s_{ckt}} w_{ij}$. The overall sample probability is:

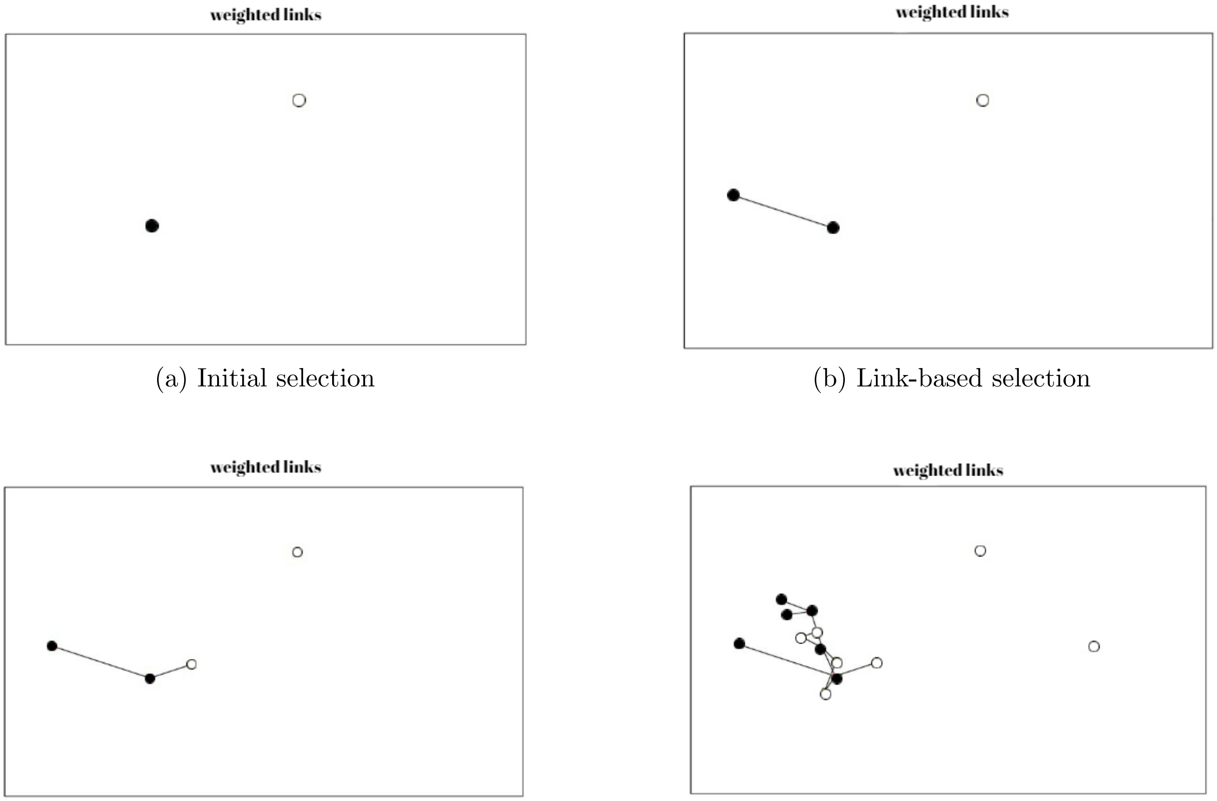
$$p(s) = \prod_{k=1}^{K} \prod_{t=1}^{n_k} q_{kti}.$$

Sampling can stop when sufficient coverage is achieved, allowing flexible sample sizes. AWS's ability to prioritize high-value links or auxiliary variables is particularly effective for hard-to-reach populations, such as those at risk for HIV [20].

2.2.1 Example of Selection Process in Adaptive Web Sampling

To illustrate AWS, consider a sequence of selections in a network setting, where black nodes represent units of interest (e.g., HIV-positive individuals) and white nodes represent others. An initial sample is chosen via SRS, followed by adaptive selections based on an active set, typically a subset of sampled nodes. At each step, a link from the active set is followed with probability d, or a node is selected randomly with probability $1 - d$. Sampling continues until a specified sample size or stopping criterion is met, with or without replacement.

Fig. 5a shows an initial sample of two nodes selected without replacement. Fig. 5b depicts the next selection, where a link from one initial node is followed. Unlike random walk designs, AWS allows links from any active set node, as shown in Fig. 5c, where a link from an earlier node is chosen. Fig. 5d highlights the design's flexibility, branching in multiple directions or selecting nodes randomly. In this example, links from high-risk (black) nodes are prioritized, with $d = 0.9$.



(a) Initial selection

(b) Link-based selection

(c) Previous node link

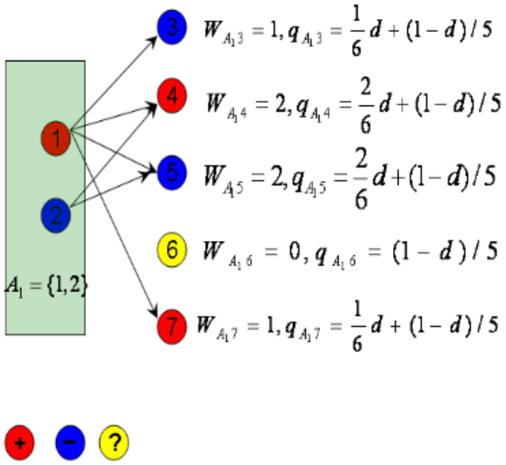
(d) Flexible branching

Fig. 5: Steps in an adaptive web sampling process. (a) Two nodes are selected randomly. (b) A new node is added via a link from the active set. (c) A link from an earlier node is followed. (d) The sample branches flexibly, with possible random selections.

Fig. 6 illustrates a sample from a population with a link matrix w_{ij}, where 1 indicates a link and 0 indicates none. Nodes 1 and 2 are initially selected, revealing links to nodes 3, 4, 5, and 7. Red nodes (e.g., HIV-positive) are of interest, blue nodes are not, and node 6 is isolated ($w_{a_1 6} = 0$). Transition probabilities are calculated as $w_{a_1 3} = 1$, $w_{a_1 4} = 2$, etc. Unlike snowball sampling, AWS allows selection of isolated nodes, enhancing flexibility.

$$\left(\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
2 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
3 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\
4 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\
5 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\
6 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
7 & 0 & 0 & 0 & 1 & 0 & 0 & 0
\end{array}\right)$$

(a) Link matrix



1
2
$A_1 = \{1,2\}$
3 $W_{A_1 3} = 1, q_{A_1 3} = \frac{1}{6}d + (1-d)/5$
4 $W_{A_1 4} = 2, q_{A_1 4} = \frac{2}{6}d + (1-d)/5$
5 $W_{A_1 5} = 2, q_{A_1 5} = \frac{2}{6}d + (1-d)/5$
6 $W_{A_1 6} = 0, q_{A_1 6} = (1-d)/5$
7 $W_{A_1 7} = 1, q_{A_1 7} = \frac{1}{6}d + (1-d)/5$
+
−
?

(b) Sample illustration

Fig. 6: Adaptive web sampling example. (a) Link matrix w_{ij} for a population. (b) Sample with nodes 1 and 2 initially selected, showing weighted links to nodes 3 and 4, with red nodes indicating interest.

3 Estimation

This section presents four estimators for the population mean μ, the average value of the variable of interest y_i, in adaptive web sampling (AWS), with Rao-Blackwellization used to enhance precision. In network sampling, the data consist of a node sample $s^{(1)}$, containing units i in their selection order with observed values y_i, and a pair sample $s^{(2)}$, containing link variables $w_{i,j}$ for pairs where unit $i \in s^{(1)}$ and j is any unit in the population. In with-replacement sampling, $s^{(1)}$ may include repeated units. The minimal sufficient statistic (m.s.s.) is the reduced data $d_r = \{(i, y_i), ((j,k), w_{j,k}) : i \in s^{(1)}, (j,k) \in s^{(2)}\}$, capturing distinct units, their y_i values, and observed link variables [22]. Combined with the Rao-Blackwell theorem [16, 2], this m.s.s. enables improved estimators by computing the conditional expectation of preliminary estimators given d_r. Four estimators, developed by Thompson [20], are described below for without-replacement designs, with modifications for with-replacement designs discussed in subsequent subsections. None of these estimators uniformly minimizes mean squared error due to the m.s.s. not capturing all population information in design-based sampling [19]. Let $\mathbf{s}$ denote the ordered sample and $s = r(\mathbf{s})$ the set of distinct units, where r eliminates order (without-replacement) or multiplicity (with-replacement). Link variables $w_{i,j}$ between sampled units are observed directly in the sample.

3.1 Estimation Based on Initial Sample Mean

Consider an initial sample s_0 of size n_0, where unit i has inclusion probability π_i. The Horvitz-Thompson estimator for the population mean $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$ is:

$$\hat{\mu}_{01} = \frac{1}{N}\sum_{i \in s_0} \frac{y_i}{\pi_i}.$$

If s_0 is selected via simple random sampling (SRS) without replacement, $\pi_i = n_0/N$, and $\hat{\mu}_{01} = \bar{y}_0$, the initial sample mean. This estimator is unbiased for all $0 \leq d \leq 1$. The Rao-Blackwellized estimator is:

$$\hat{\mu}_1 = \mathbb{E}(\hat{\mu}_{01} | d_r) = \sum_{\mathbf{s}: r(\mathbf{s}) = s} \hat{\mu}_{01}(\mathbf{s}) p(\mathbf{s} | d_r),$$

where the conditional probability is:

$$p(\mathbf{s}|d_r) = \frac{p(\mathbf{s})}{\sum_{\mathbf{s}: r(\mathbf{s}) = s} p(\mathbf{s})}.$$

For without-replacement sampling, the summation covers all $n!$ reorderings of the n sampled nodes. For designs with an initial SRS of size n_0, the expectation is over all $\binom{n}{n_0}$ initial sample combinations and $(n - n_0)!$ reorderings of subsequent selections. This estimator remains unbiased, weighting each reordering by its selection probability.

3.2 Estimation Based on Conditional Selection Probabilities

The second estimator, resembling a Hansen-Hurwitz estimator, accounts for conditional selection probabilities at each wave. Let $\hat{\tau}_{s_0}$ be an unbiased estimator of the population total $\tau = \sum_{i=1}^{N} y_i$ based on the initial sample. For SRS without replacement, $\hat{\tau}_{s_0} = \frac{N}{n_0} \sum_{i \in s_0} y_i = N\bar{y}_0$; for unequal probability designs, $\hat{\tau}_{s_0} = \sum_{i \in s_0} \frac{y_i}{\pi_i}$. For the i-th selection after the initial sample, with current sample $s_{ck} = (s_0, \ldots, s_{k-1})$, define:

$$z_i = \frac{y_i}{q_{ki}},$$

where q_{ki} is the conditional selection probability. Each z_i is an unbiased estimator of τ for $0 \le d < 1$. The composite estimator is:

$$\hat{\mu}_{02} = \frac{1}{Nn} \left[n_0 \hat{\tau}_{s_0} + \sum_{i=n_0+1}^{n} z_i \right].$$

This is a weighted average of the initial sample estimator and the average of $n - n_0$ conditional estimators, unbiased for $0 \le d < 1$. If $d = 1$ (following only links), $\hat{\mu}_{02}$ is biased, estimating only the total of accessible nodes. The Rao-Blackwellized estimator is:

$$\hat{\mu}_2 = \mathbb{E}(\hat{\mu}_{02} | d_r) = \sum_{\mathbf{s}: r(\mathbf{s}) = s} \hat{\mu}_{02}(\mathbf{s}) p(\mathbf{s} | d_r).$$

For binary y_i (0 or 1), $\hat{\mu}_2$ may exceed 1, prompting the use of ratio estimators below.

3.3 Composite Conditional Generalized Ratio Estimator

An unbiased estimator of population size N is obtained by setting $y_i = 1$ in the total estimator: $\hat{N}_0 = \sum_{i \in s_0} \frac{1}{\pi_i}$. For SRS, $\pi_i = n_0/N$, so $\hat{N}_0 = N$. For subsequent selections, set $y_i = 1$ in z_i, yielding $\hat{N}_i = \frac{1}{q_{ki}}$ for $i = n_0 + 1, \ldots, n$, unbiased for N when $0 \le d < 1$. The composite estimator is:

$$\hat{N} = \frac{1}{n} \left[n_0 \hat{N}_0 + \sum_{i=n_0+1}^{n} \hat{N}_i \right].$$

The generalized ratio estimator is:

$$\hat{\mu}_{03} = \frac{\hat{\mu}_{02}}{\hat{N}/N}.$$

Since $\hat{\mu}_{03}$ is a ratio, it may be biased. The Rao-Blackwellized estimator is:

$$\hat{\mu}_3 = \mathbb{E}(\hat{\mu}_{03}|d_r) = \sum_{\mathbf{s}:r(\mathbf{s})=s} \hat{\mu}_{03}(\mathbf{s})p(\mathbf{s}|d_r),$$

with the same bias but reduced variance compared to $\hat{\mu}_{03}$.

3.4 Composite Conditional Mean-of-Ratios Estimator

Ratio estimates for μ are formed by dividing the initial estimator by $\hat{N}_0$ and each z_i by $\hat{N}_i$. The mean-of-ratios estimator is:

$$\hat{\mu}_{04} = \frac{1}{n}\left[\frac{n_0\hat{\tau}_{s_0}}{\hat{N}_0} + \sum_{i=n_0+1}^{n} \frac{z_i}{\hat{N}_i}\right].$$

The Rao-Blackwellized estimator is:

$$\hat{\mu}_4 = \mathbb{E}(\hat{\mu}_{04}|d_r) = \sum_{\mathbf{s}:r(\mathbf{s})=s} \hat{\mu}_{04}(\mathbf{s})p(\mathbf{s}|d_r).$$

None of $\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4$ uniformly minimizes mean squared error due to the m.s.s.'s incompleteness. While $\hat{\mu}_2$ is unbiased, it may produce large values for small q_{ki}. For binary y_i, $\hat{\mu}_2$ can exceed 1, whereas $\hat{\mu}_3$ and $\hat{\mu}_4$, though biased, yield estimates between 0 and 1.

3.5 Variance Estimation

For an estimator $\hat{\mu} = \mathbb{E}(\hat{\mu}_0|d_r)$, the variance is:

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}(\hat{\mu}_0) - \mathbb{E}[\mathrm{Var}(\hat{\mu}_0|d_r)].$$

An unbiased variance estimator is:

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \mathbb{E}[\widehat{\mathrm{Var}}(\hat{\mu}_0)|d_r] - \mathrm{Var}(\hat{\mu}_0|d_r).$$

For $\hat{\mu}_1$, with an initial SRS of n_0 units, the initial variance estimator is:

$$\widehat{\mathrm{Var}}(\hat{\mu}_{01}) = \frac{N-n_0}{Nn_0}\frac{v_0}{N},$$

where $v_0 = \frac{1}{n_0-1}\sum_{i\in s_0}(y_i - \bar{y}_0)^2$ is the sample variance. Thus:

$$\widehat{\mathrm{Var}}(\hat{\mu}_1) = \mathbb{E}[\widehat{\mathrm{Var}}(\hat{\mu}_{01})|d_r] - \mathrm{Var}(\hat{\mu}_{01}|d_r).$$

For $\hat{\mu}_2$, with $n_0 = 1$ and $\hat{\mu}_{02} = \frac{1}{Nn}\sum_{i=1}^{n} z_i$, the variance estimator is:

$$\widehat{\mathrm{Var}}(\hat{\mu}_{02}) = \frac{1}{n(n-1)N^2}\sum_{i=1}^{n}(z_i - N\hat{\mu}_{02})^2,$$

unbiased following Raj [15], Murthy [13], as the z_i's are uncorrelated. Thus:

$$\widehat{\mathrm{Var}}(\hat{\mu}_2) = \mathbb{E}[\widehat{\mathrm{Var}}(\hat{\mu}_{02})|d_r] - \mathrm{Var}(\hat{\mu}_{02}|d_r).$$

A practical approach, recommended by Thompson [20], involves selecting m independent samples, each yielding an estimate $\hat{\mu}_k$. The mean is:

$$\hat{\mu} = \frac{1}{m}\sum_{k=1}^{m}\hat{\mu}_k,$$

with variance estimator:

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{m(m-1)}\sum_{k=1}^{m}(\hat{\mu}_k - \hat{\mu})^2.$$

Approximate $100(1-\alpha)\%$ confidence intervals are:

$$\hat{\mu} \pm t_{m-1,\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\mu})},$$

where $t_{m-1,\alpha/2}$ is the upper $\alpha/2$-point of the Student's t-distribution with $m-1$ degrees of freedom.

4 Markov Chain Monte Carlo

Computing Rao-Blackwellized estimators $\hat{\mu}_1$, $\hat{\mu}_2$, and their variances under adaptive web sampling (AWS) designs requires evaluating all possible reorderings of the sample selection sequence, each weighted by its conditional probability $p(\mathbf{s}|d_r)$. For small sample sizes ($n \leq 10$), direct enumeration of permutations (up to $n!$) is feasible, calculating the estimator and variance for each reordering. However, for larger samples, the number of permutations becomes computationally prohibitive. To address this, a Markov chain Monte Carlo (MCMC) resampling approach, based on a Metropolis-Hastings algorithm [9, 20], is employed to approximate the expected values efficiently.

The goal is to generate a Markov chain $X_0, X_1, X_2, \ldots$ with stationary distribution $p(x|d_r)$, where x represents a permutation of the ordered sample $\mathbf{s}$ (without replacement) or a sequence of length n consistent with the reduced data d_r (with replacement). The following algorithm ensures the chain remains in the stationary distribution:

4.1 MCMC Resampling Algorithm

1. **Initialization**: Set $X_0 = \mathbf{s}$, the original ordered sample, ensuring the chain starts in the stationary distribution. At step $k-1$, suppose the chain is at X_{k-1}, a permutation j.

2. **Candidate Generation**: Generate a tentative permutation t_k from a candidate distribution p_t. This distribution consists of permutations of the n sampled units, obtained by applying the AWS design to these units as if they form the entire population.

3. **Acceptance Probability**: Compute the acceptance probability:

$$\alpha = \min\left\{\frac{p(t_k)p_t(X_{k-1})}{p(X_{k-1})p_t(t_k)}, 1\right\},$$

   where $p(\cdot)$ is the design probability of the permutation. With probability α, set $X_k = t_k$; otherwise, set $X_k = X_{k-1}$.

4. **Iteration**: Return to step 2 and repeat.

This algorithm produces a Markov chain with the desired stationary distribution $p(x|d_r)$. Starting at $\mathbf{s}$, the chain remains in the stationary distribution at each step, allowing resampling of permutations consistent with d_r.

Suppose n_r permutations are resampled, with $\hat{\mu}_{0j}$ denoting the initial estimator (e.g., $\hat{\mu}_{01}$ or $\hat{\mu}_{02}$) for the j-th permutation. The Rao-Blackwellized estimator $\hat{\mu} = \mathbb{E}(\hat{\mu}_0|d_r)$ is approximated by:

$$\tilde{\mu} = \frac{1}{n_r}\sum_{j=1}^{n_r}\hat{\mu}_{0j}, \quad \text{for } j = 1, 2, 3, 4.$$

Similarly, the expected variance and conditional variance are approximated as:

$$\tilde{E}(\widehat{\text{Var}}(\hat{\mu}_0)|d_r) = \frac{1}{n_r}\sum_{j=1}^{n_r}\widehat{\text{Var}}(\hat{\mu}_{0j}),$$

$$\tilde{\text{Var}}(\hat{\mu}_0|d_r) = \frac{1}{n_r}\sum_{j=1}^{n_r}(\hat{\mu}_{0j} - \tilde{\mu})^2.$$

The additional variance due to resampling, $\text{Var}(\tilde{\mu}|d_r)$, can be estimated by dividing the n_r resampled permutations into L groups of size K (where $n_r = L \cdot K$). Let $\bar{y}_i$ be the mean of the i-th group. The sample variance of the group means is:

$$s_{\bar{y}}^2 = \frac{1}{L(L-1)}\sum_{i=1}^{L}(\bar{y}_i - \tilde{\mu})^2.$$

Since resampling is computationally inexpensive compared to actual sampling, large n_r values are recommended to minimize this additional variance [9].

Bayesian model-based inference with AWS designs also relies on MCMC methods, except in simple cases where explicit posterior distributions are derivable [4]. Typically, MCMC involves updating model parameters and, via data augmentation, generating complete population network realizations from the predictive posterior distribution conditioned on observed data [11, 12]. These realizations enable flexible inference about various population characteristics, such as network structure or node attributes.

5 Empirical Studies

This section presents simulation results for adaptive web sampling (AWS) applied to two real-world populations: a population at risk for HIV/AIDS and a blue-winged teal bird population. The studies, primarily based on Vincent (2008) [23] with additional analyses from Thompson (2006) [20], evaluate four estimators ($\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4$) under the Strategy 1 design of AWS. Mean squared error (MSE) scores are reported for preliminary and improved (Rao-Blackwellized) estimates, derived from 2000 simulation runs with 10,000 MCMC resamples per run.

5.1 At Risk for HIV/AIDS Population

Adaptive sampling methods are effective for hidden populations like those at risk for HIV/AIDS, where conventional sampling often yields biased results due to rarity and cost constraints [20]. The dataset comes from the Colorado Springs Study [14, 6], comprising 595 interviewed individuals within a network of 8762. Nodes represent people, with edges indicating drug-sharing relationships, and the variable of interest is the proportion of injection drug users (population mean $\mu = 0.575$).

Simulations used Strategy 1 of AWS [23] with an initial sample size of $n_0 = 15$, final sample size of $n = 30$, and dampening probability $d = 0.9$ in a without-replacement design. Initial samples were selected via simple random sampling, and links were followed randomly.

	Estimator 1	Estimator 2	Estimator 3	Estimator 4
Preliminary Estimate	0.0160	0.0775	0.0306	0.0114
Improved Estimate	0.0154	0.0725	0.0281	0.0113

Table 1: MSE scores of the four estimators for the HIV/AIDS population, Strategy 1 [23].

Table 1 shows that $\hat{\mu}_1$ and $\hat{\mu}_4$ have the lowest MSE scores (0.0154 and 0.0113 for improved estimates), with $\hat{\mu}_4$ slightly better but exhibiting bias [23]. Rao-Blackwellization reduces MSE for all estimators, especially $\hat{\mu}_3$.

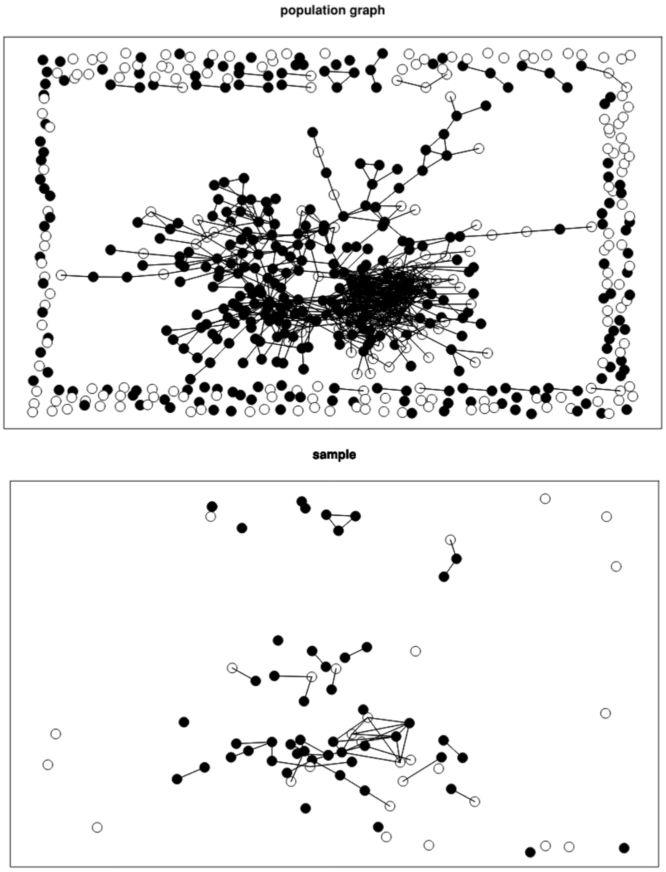


Fig. 7: Top: HIV/AIDS population network [14]. Dark nodes indicate injection drug users; links show drug-sharing relationships. Bottom: AWS sample with $n_0 = 15$, $n = 30$, $d = 0.9$, random link selection [23].

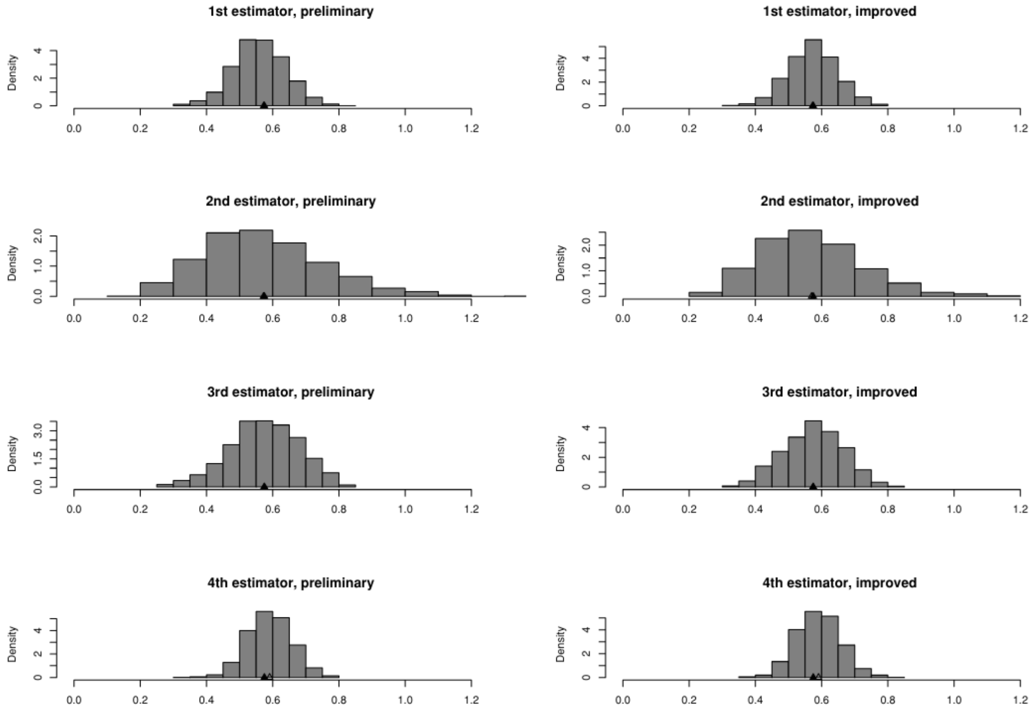


Fig. 8: Histograms of preliminary and improved estimators for the HIV/AIDS population, Strategy 1, with true mean $\mu = 0.575$ (solid triangle). Based on 2000 simulations, 10,000 resamples per run [23].

An additional analysis from Thompson (2006) [20] examines degree distributions for this population, using a design with $n_0 = 10$, $n = 20$, $m = 4$, and $d = 0.9$. The population degree distribution (Figure 9) shows a near-linear structure on a logarithmic scale, suggesting a power-law approximation. AWS samples over-represent high-degree nodes, with a sample mean degree of 42.5 versus the population mean of 2.5. The unbiased estimator $\hat{\mu}_1$ corrects this bias, as shown in the sampling distribution of mean degree.

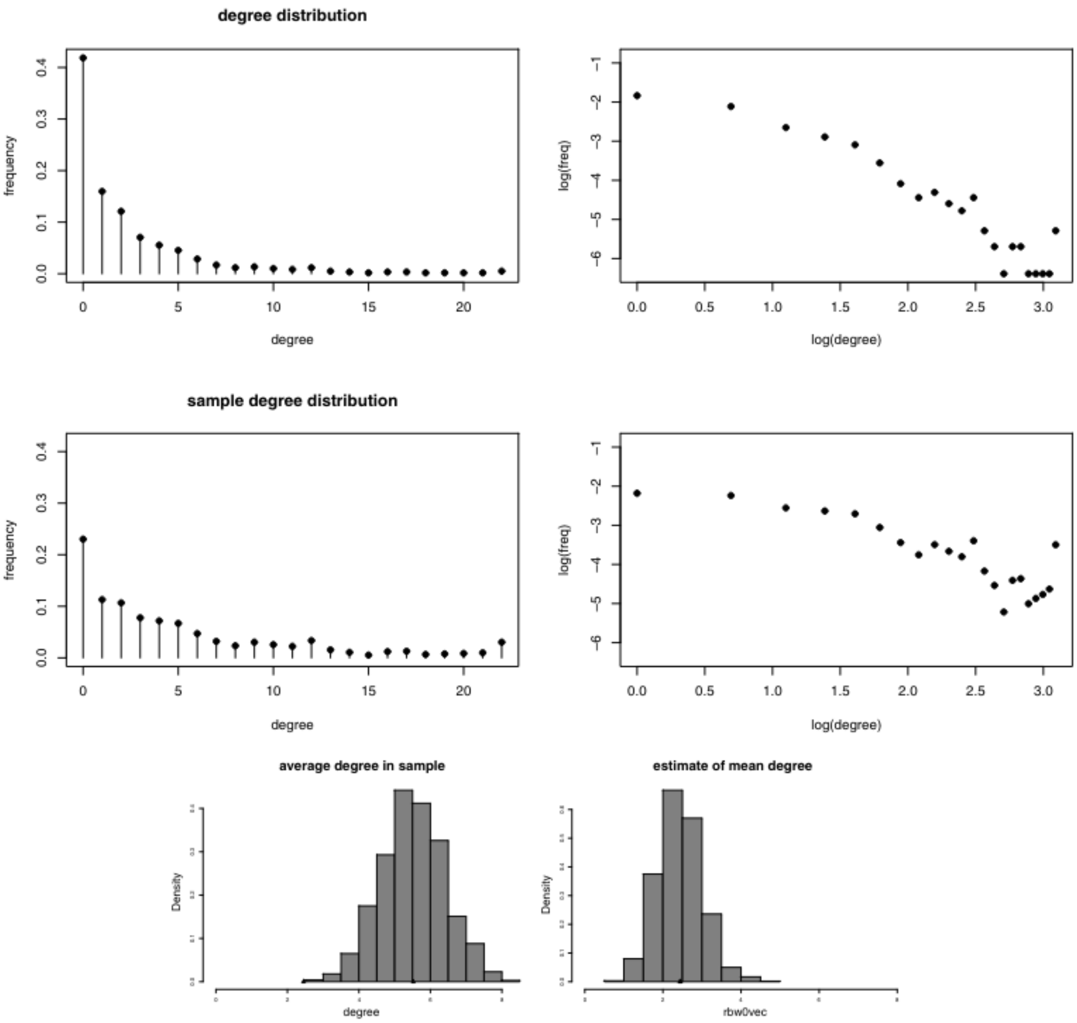


Fig. 9: Population and sample degree distributions for the HIV/AIDS population, natural (top left) and logarithmic (top right) scales. Sample mean degree distribution (bottom left) and unbiased estimator $\hat{\mu}_1$ for mean degree (bottom right) [20].

5.2 Wintering Waterfowl Population

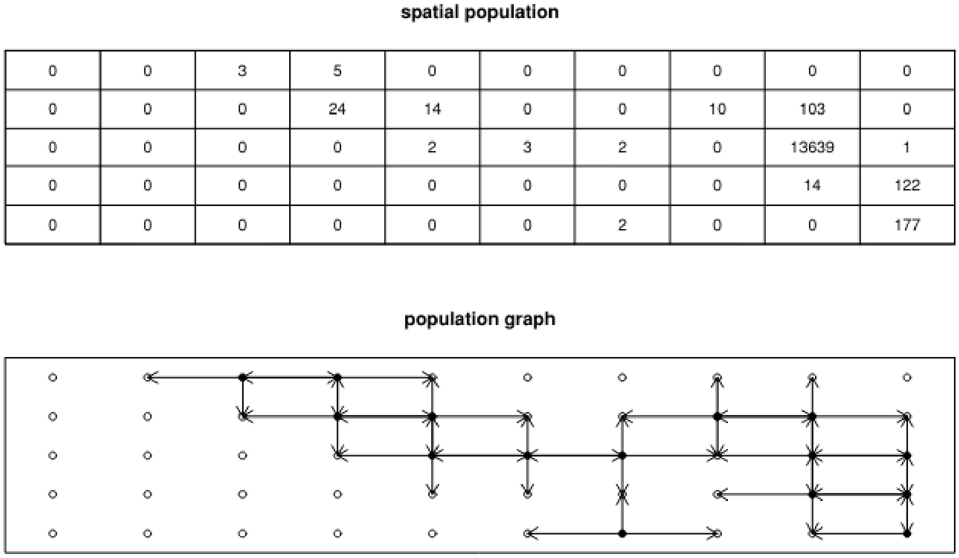
Blue-winged teal populations are spatially clustered, challenging conventional sampling [18]. The dataset from Smith et al. (1995) divides a Florida wildlife refuge into 50 plots, estimating the mean number of birds per plot ($\mu = 282.42$). Nodes are plots, with directed edges based on bird counts.

Simulations used Strategy 1 of AWS [23] with $n_0 = 15$, $n = 30$, and $d = 0.9$ in a without-replacement design. Initial samples were random, with random link selection.

	Estimator 1	Estimator 2	Estimator 3	Estimator 4
Preliminary Estimate	174431.02	84922.88	84972.93	63469.09
Improved Estimate	36093.01	20238.76	22132.54	26827.33

Table 2: MSE scores of the four estimators for the blue-winged teal population, Strategy 1 [23].

Table 2 shows that $\hat{\mu}_2$ and $\hat{\mu}_3$ have the lowest improved MSE (20238.76 and 22132.54). Rao-Blackwellization significantly reduces MSE, but $\hat{\mu}_4$ is biased [23].



spatial population

0	0	3	5	0	0	0	0	0	0
0	0	0	24	14	0	0	10	103	0
0	0	0	0	2	3	2	0	13639	1
0	0	0	0	0	0	0	0	14	122
0	0	0	0	0	0	2	0	0	177

population graph

Fig. 10: Blue-winged teal population spatial counts and graph structure [18].

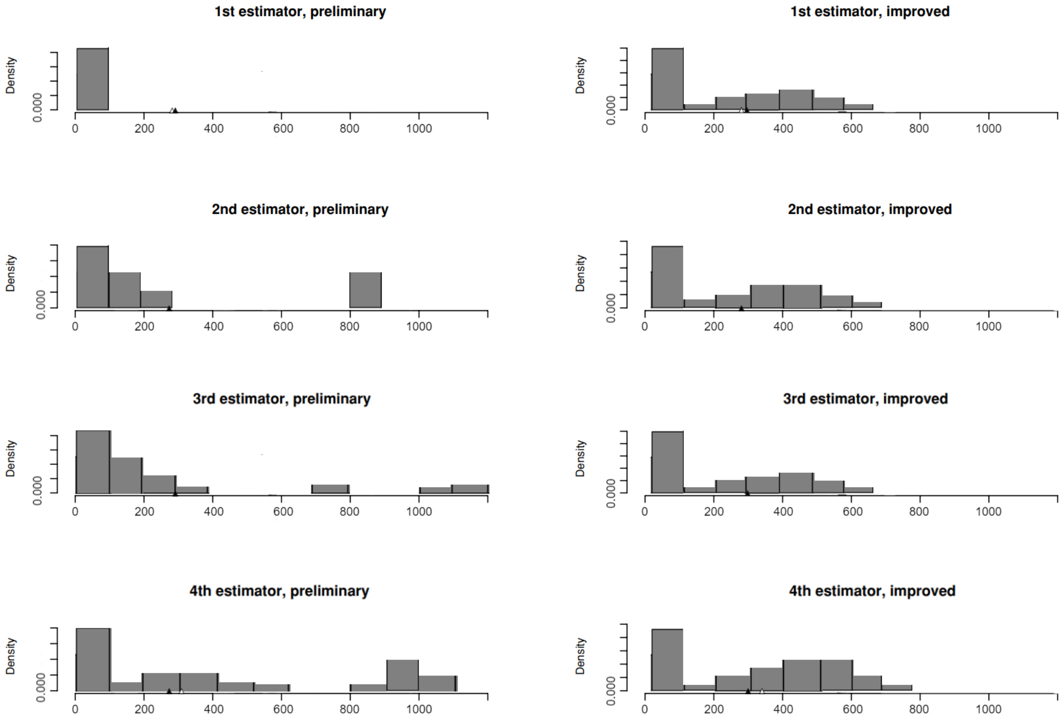


Fig. 11: Histograms of preliminary and improved estimators for the blue-winged teal population, Strategy 1, with true mean $\mu = 282.42$ (solid triangle). Based on 2000 simulations, 10,000 resamples per run [23].

Thompson (2006) [20] analyzes optimal initial sample sizes with a fixed total sample size ($n = 20$), varying n_0, and $d = 0.9$. Figure 12 shows minimum MSE at $n_0 = 13$–14, with a 75% efficiency gain over simple random sampling, outperforming adaptive cluster sampling [18].

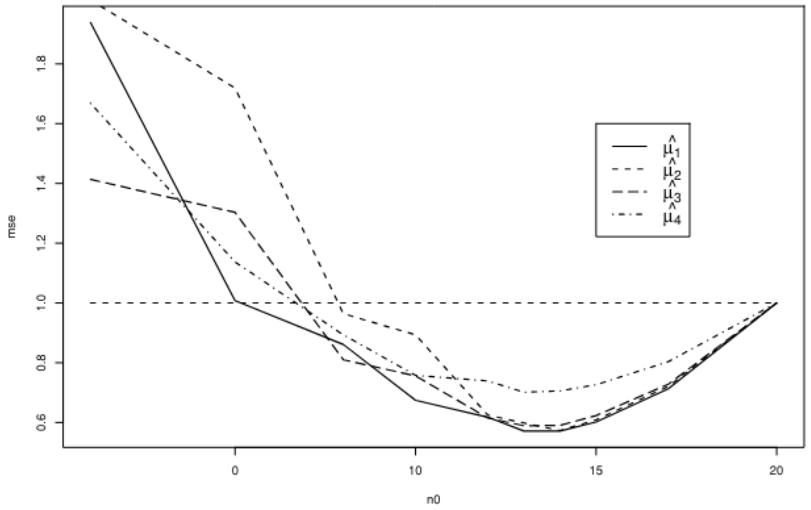


Fig. 12: Standardized MSE of estimators for varying initial sample sizes n_0, with total sample size $n = 20$, $d = 0.9$, for the blue-winged teal population [20].

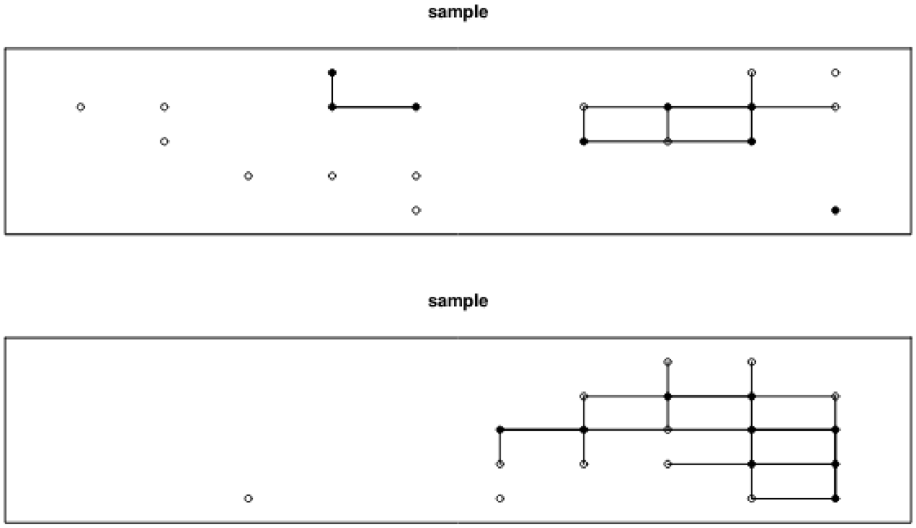


Fig. 13: Two AWS samples for blue-winged teal population ($n = 20$, $n_0 = 13$ (top), $n_0 = 1$ (bottom), $d = 0.9$ [20].

6 Conclusions

Adaptive web sampling (AWS) significantly enhances sampling strategies for populations with network or spatial structures, particularly those that are difficult to access, such as the HIV/AIDS at-risk population and the blue-winged teal population studied in Section 5 [23, 20]. The simulations presented demonstrate that AWS offers several advantages over traditional methods like snowball sampling, adaptive cluster sampling, and random walk designs.

First, unlike adaptive cluster sampling, AWS does not require complete sampling of connected components, enabling more efficient resource allocation [20]. Second, AWS provides precise control over sample size, a feature not always available in snowball sampling [7] or adaptive cluster sampling [17, 3]. This control allows researchers to balance deep exploration (following links over multiple waves) with broad coverage (fewer waves, wider reach). By using a mixture distribution, AWS flexibly allocates effort between conventional and adaptive sampling, allowing prioritization of high-value units or links [20]. Compared to random walk designs, where unit selection depends solely on the most recent unit, AWS defines a more flexible active set, which may include recent selections, the entire current sample, or units close in geographic or graph distance [20].

The empirical studies in Section 5 highlight the effectiveness of AWS. For the HIV/AIDS population, estimators $\hat{\mu}_1$ and $\hat{\mu}_4$ achieved the lowest mean squared errors (MSEs) (0.0154 and 0.0113, respectively), though $\hat{\mu}_4$ showed slight bias [23]. For the blue-winged teal population, $\hat{\mu}_2$ and $\hat{\mu}_3$ performed best, with improved MSEs of 20238.76 and 22132.54 [23]. Rao-Blackwellization consistently reduced estimator variance, enhancing precision across both populations [23, 20]. Likelihood and Bayesian inference methods can further improve estimation with AWS [21, 5]. The choice of initial sample size n_0 and link-tracing probability d is critical. In the blue-winged teal example, allocating 65–70% of the total sample ($n_0 = 13$–14 for $n = 20$) to the initial random selection minimized MSE, achieving a 75% efficiency gain over simple random sampling [20]. However, optimal values of n_0 and d depend on population characteristics and study objectives. Setting $d = 1$ eliminates random jumps unless the sampling reaches a dead end, preserving the unbiasedness of $\hat{\mu}_1$ but not $\hat{\mu}_2$ [20]. Choosing $d < 1$ and $n_0 > 1$ ensures broader population exploration, preventing entrapment in a single network component.

Future research should explore dynamic adjustments to d, such as increasing d as the sample grows to focus on high-value areas, potentially improving efficiency. Additionally, testing n_0 and d across diverse populations will clarify their impact on estimator performance. The flexibility of AWS, combined with the promising results from this study, suggests that further refinements will enhance its applicability in network and spatial sampling.

In conclusion, AWS is a versatile and powerful tool, offering improved efficiency and flexibility over traditional methods. Its ability to adapt to complex population structures makes it ideal for challenging sampling scenarios, with significant potential for future advancements.

References

[1] D. Basu. Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā: The Indian Journal of Statistics, Series A*, 31(4):441–454, 1969.

[2] David Blackwell. Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, 18(1):105–110, 1947.

[3] Jonathan A. Brown and Bryan F. J. Manly. Restricted adaptive cluster sampling. *Environmental and Ecological Statistics*, 5(1):49–63, 1998.

[4] M. Chow and S. K. Thompson. Bayesian inference for adaptive sampling designs. *Unpublished manuscript*, 2003.

[5] Ming T. Chow and Steven K. Thompson. Estimation with link-tracing sampling designs: A bayesian approach. *Survey Methodology*, 29(2):197–205, 2003.

[6] William W. Darrow, John J. Potterat, Richard B. Rothenberg, Donald E. Woodhouse, Stephen Q. Muth, and Alden S. Klovdahl. Using knowledge of social networks to prevent human immunodeficiency virus infections: The colorado springs study. *Sociological Focus*, 32(2):143–158, 1999.

[7] Ove Frank and Tom Snijders. Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10(1):53–67, 1994.

[8] V. P. Godambe. A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society: Series B*, 17(2):269–278, 1955.

[9] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. doi: 10.1093/biomet/57.1.97.

[10] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460): 1090–1098, 2002. doi: 10.1198/016214502388618906.

[11] M. Kwanisai. *Bayesian Inference for Adaptive Sampling Designs.* PhD thesis, Simon Fraser University, 2005.

[12] M. Kwanisai. Bayesian adaptive sampling. *Unpublished manuscript*, 2006.

[13] M. N. Murthy. Ordered and unordered estimators in sampling without replacement. *Sankhyā: The Indian Journal of Statistics*, 18(3/4):379–390, 1957.

[14] John J. Potterat, Donald E. Woodhouse, Stephen Q. Muth, Richard B. Rothenberg, William W. Darrow, Alden S. Klovdahl, and Franklyn N. Judson. Network dynamism: History and lessons of the colorado springs study. *AIDS and Behavior*, 1993.

[15] Des Raj. Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association*, 51(274):269–284, 1956.

[16] C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

[17] Mohammad M. Salehi and George A. F. Seber. Adaptive cluster sampling with fixed sample size. *Biometrics*, 53(3):959–990, 1997.

[18] David R. Smith, Jonathan A. Brown, and Nancy C. H. Lo. Application of adaptive cluster sampling to biological populations. *Environmental and Ecological Statistics*, 2 (2):91–108, 1995.

[19] Steven K. Thompson. *Sampling*. Wiley, 2 edition, 2002.

[20] Steven K. Thompson. Adaptive web sampling. *Biometrics*, 62(4):1224–1234, 2006. doi: 10.1111/j.1541-0420.2006.00654.x.

[21] Steven K. Thompson and Ove Frank. Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26(1):87–98, 2000.

[22] Steven K. Thompson and George A. F. Seber. *Adaptive Sampling*. Wiley, 1996.

[23] Kyle Shane Vincent. Design variations in adaptive web sampling. Master's project, Simon Fraser University, 2008.