Targeted Random Walk Designs

To appear in *Survey Methodology,* 2006

Steven K. Thompson[1]

Simon Fraser University

Abstract

Hidden human populations, the Internet, and other networked structures conceptualized mathematically as graphs are inherently hard to sample by conventional means, and the most effective study designs usually involve procedures that select the sample by adaptively following links from one node to another. Sample data obtained in such studies are generally not representative at face value of the larger population of interest. However, a number of design and model based methods are now available for effective inference from such samples. The design based methods have the advantage that they do not depend on an assumed population model, but do depend for their validity on the design being implemented in a controlled and known way, which can be difficult or impossible in practice. The model based methods allow greater flexibly in the design, but depend on modeling of the population using stochastic graph models and also depend on the design being ignorable or of known form so that it can be included in the likelihood or Bayes equations. For both the design and the model based methods, the weak point often is the lack of control in how the initial sample is obtained, from which link-tracing commences. The designs described in this paper offer a third way, in which the sample selection probabilities become step by step less dependent on the initial sample selection. A Markov chain "random walk" model idealizes the natural design tendencies of a link-tracing selection sequence through a graph. This paper introduces

uniform and target walk designs in which the random walk is nudged at each step to produce a a design with the desired stationary probabilities. A sample is thus obtained that in important respects is representative at face value of the larger population of interest, or that requires only simple weighting factors to make it so.

Key words: Adaptive sampling, link-tracing designs, Markov chain Monte Carlo, network sampling, random walk, respondent-driven sampling, sampling in graphs, sampling hidden population.

1 Introduction

Populations with linkage or network structure are conceptualized as graphs, with the nodes of the graph representing the units of the population and the edges or arcs of the graph representing the relationships or links between the units in the population. A central problem of studies in graph settings is that for many of the populations of interest it is difficult or impossible to obtain samples using conventional designs, and the samples obtained may be at face value highly unrepresentative of the larger population of interest. In practice, often the only practical methods of obtaining the sample involve following links from sample nodes to add more nodes and links to the sample. For example, in studies of hidden human populations such as injection drug users, sex workers, and others at risk for HIV/AIDS or hepatitis C, social links are followed from initially identified respondents to add more research participants to the sample. Similarly, in investigations of the characteristics of the Internet, the usual procedure is to obtain a sample of web sites by following links from initial sites to other sites.

Klovdahl (1989) used the term "random walk" to describe a procedure for obtaining a sample from a hidden population by asking a respondent to identify several contacts, one

[1]Address for correspondence: Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6. Email: thompson@stat.sfu.ca

of whom is selected at random to be the next respondent, with the pattern continuing for a number of steps. Heckathorn (1997) described methods of "respondent-driven sampling" using procedures of this type. The motivation for using designs like this in practice is to penetrate deeper into the hidden population to obtain respondents who are more "representative" of the population than the more conspicuous initial respondents may be. In studies of the Internet, the parallel idea is that of the "random surfer," who selects a web page at random, clicks at random on one of the links on that page, thus moving to another page, and so on (Brin and Page 2002). The random walk design can be conceptualized as a Markov chain (Heckathorn, 1997, 2002, Henzinger et al 2000, Salganik and Heckathorn 2004). In this paper some modifications of these Markov chain designs are described, with the object of obtaining stationary probabilities of equal or specified values in order to obtain simple estimates of characteristics of the population graph of interest.

Approaches to inference from samples in a graph setting include design-based, model-based, and combination methods. In the design based approach, all values of node and link variables in the graph are considered fixed or given, and inference is based on the design-induced probabilities involved in selecting the sample. In the model based approach, the population is itself viewed as a realization of a stochastic graph model, which provides the joint probability distribution of all the node and link variables. Previous design-based approaches include the methods of network or multiplicity sampling (Birnbaum and Sirken 1965), adaptive cluster sampling applied in a graph setting (Thompson and Collins 2002), and a few of the methods in the snowball sampling literature (Frank 1977, 1978, Frank and Snijders 1994). A method combining design and model based approaches is used in Felix-Medina and Thompson (2004) for studying a hidden population in which link-tracing follows from a probability survey sample from a frame that covers only part of the population.

The advantage of design-based methods is that populations such as socially networked hidden human populations are difficult to model realistically, and the design-based inference does not rely on modeling assumptions for properties such as unbiasedness and consistency of esti-

mators. Design-based inference methods do rely on the design being implemented according to plan, however, and exact implementation of a given design may be a very great challenge in studies of hidden human populations. This was the motivation for the development of a range of model-based methods for inference from samples in graphs, including maximum likelihood and Bayes techniques (Thompson and Frank 2000, Chow and Thompson 2003). Assuming that the initial sample is "ignorable" in the likelihood sense (Rubin 1976), or that the design is of known form so that it can be included in the likelihood and Bayes equations, these methods work for a very wide range of link-tracing sampling procedures, including most variations of the snowball and network methods. In reality, however, the initial sample may be selected in a fashion that is anything but ignorable, with selection probabilities depending on node value, node degree, and other factors. The pervasive problem of initial sample selection in link-tracing studies has been remarked upon by Spreen(1992) among others.

The approach pursued in the present paper does not assume total control over all design possibilities, but rather seeks to work with the way samples naturally tend to get selected in networked populations, whether by ethnographers or other social scientists, members of the population themselves, or automated web crawlers. Starting with those natural selection processes, we introduce iterative modifications to obtain sampling procedures that step by step approach desired selection probabilities.

Although the underlying structure of the designs in this paper depends on Markov chains, the estimators and quantities of most interest to investigators may not in fact be Markovian. For example, while the sequence of selections of sample units may depend at each step only on the most recently selected unit, the sequence by which distinct units are added to the sample depends on all units selected thus far. For this reason, the properties of a number of alternative estimators with different designs are examined using simulation, by repeatedly selecting samples from stochastic graph realizations and from an empirical population from a study of a people at high risk for HIV/AIDS transmission.

Random walk designs are described in Section 2. Uniform and targeted walk designs are

introduced in Sections 3 and 4 respectively. Examples are worked in Section 5, including an illustrative example using as the population a realization of a stochastic graph model and an empirical example using data from a study of a population at high risk for HIV/AIDS.

2 Random Walk

The population of interest is a graph, given by a set of N nodes with labels $U = \{1, 2, \ldots, N\}$ and values $\mathbf{y} = (y_1, ..., y_N)$ and an $N \times N$ matrix $\mathbf{A}$ indicating relationships or links between nodes. An element a_{ij} of $\mathbf{A}$ is one if there is a link from node i to node j and zero otherwise. The diagonal elements a_{ii} are assumed to be zero. For node i, the row sum $a_{i\cdot}$ is the out-degree or number of nodes to which i has a link and the column sum $a_{\cdot i}$ is the in-degree or number of nodes which link to i. With an undirected graph, the matrix $\mathbf{A}$ is symmetric and the in-degree of any node equals its out-degree.

Let W_k denote the unit or node of the graph that is selected at the kth wave. If i is the node selected at the kth wave, then for wave $k+1$ one of the nodes linked from i is selected at random. Thus, $\{W_0, W_1, W_2, \ldots\}$ is a Markov chain with

$$P(W_{k+1} = j \,|\, W_k = i) = a_{ij}/a_{i\cdot} \tag{1}$$

Let $\mathbf{Q}$ denote the transition matrix of the chain with elements $q_{ij} = P(W_{k+1} = j \,|\, W_k = i)$. The chain is a random walk in that at each step, one of the neighboring states of the present state is selected at random.

If the graph consists of a single connected component, that is, if every node of the graph is reachable from every other node by some path, then the chain is irreducible and its stationary probabilities $(\pi_1, \ldots, \pi_N)$ satisfy $\pi_j = \sum \pi_i q_{ij}$ for $j = 1, \ldots, N$. In fact, with the simple random walk design in a connected undirected graph the stationary probabilities can be shown (Salganik and Heckathorn 2002) to be

$$\pi_j \propto a_{\cdot j}$$

That is, for an undirected graph consisting of only one connected component, the long term selection frequency for any node is proportional to its in-degree, which, for a nondirected graph, equals the out-degree.

Suppose one wishes to estimate a characteristic of the population graph, such as the population mean of the node values $\mu_y = \sum_{i=1}^{N} y_i/N$ using data from a random walk sample. The sample mean $\bar{y} = \sum_{i \in s} y_i$ is in general not unbiased because the value y_i of a node may be related to its degree and hence to its probability of being selected. However, one can obtain an approximately unbiased estimate by weighting each sample y-value by the reciprocal of its in-degree, assuming that that information is available from the data (Salganik and Heckathorn 2002).

2.1 Random walk with random jumps

In a graph with separate components or with unconnected nodes, the simple random walk just described does not have the property that every node can be eventually reached from every other node. Without this property, the limiting distribution of the random walk is sensitive to the starting distribution, since the limiting probability for a node depends on the initial probability of starting in the component that contains that node. A modification of the design which overcomes this problem allows for a jump with small probability to a node at random from the whole graph. At each step, this random walk follows a randomly selected link with probability d and, with probability $1 - d$, jumps to another node in the graph at random or with specified probability. In the Internet search literature, d is referred to as the "damping factor," since a value of d less than one damps the effect of the out-degree of a given node (Brin and Page 1998).

The transition probabilities for the random walk with jumps are given by

$$q_{ij} = \begin{cases} (1-d)/N + da_{ij}/a_{i\cdot} & \text{if } a_{i\cdot} > 0 \\ 1/N & \text{if } a_{i\cdot} = 0 \end{cases} \tag{2}$$

With the small probability $1-d$ of a random jump at any step, the Markov chain walk can potentially reach any node in the graph from any other, so that the chain is irreducible. Further, the random jumps, which include the possibility of going to node i from node i, ensure that the chain is aperiodic so that the stationary probabilities are limiting probabilities. With $d < 1$ the stationary probability of node i is not a simple function of its own in-degree, but depends also on the stationary probabilities of the nodes that link to it.

More generally, the jumps can be made with any specified probabilities $\mathbf{p} = (p_1, \ldots, p_N)$ and the probability of a jump can depend on the current state, so that the transition probabilities are

$$q_{ij} = \begin{cases} (1-d_i)p_j + d_i a_{ij}/a_{i\cdot} & \text{if } a_{i\cdot} > 0 \\ 1/N & \text{if } a_{i\cdot} = 0 \end{cases}$$

Estimates which are approximately design-unbiased for population graph characteristics can be obtained by weighting sample values inversely proportional to the limiting Markov chain selection probabilities, but with the additional problem that these limiting probabilities are unknown and must be estimated from the sample data (see Henzinger, et al. 2001 for an approach to this).

For the remainder of this paper, "random walk" or "ordinary random walk" will refer to the random walk with jumps unless it is specifically stated to be a random walk without the option of jumps.

3 Uniform Walk

In this section a modification of the random walk design is proposed which leads to uniform stationary probabilities $\boldsymbol{\pi} = (\pi, \ldots, \pi)$.

Consider first the case of the population graph consisting of only one connected component. Let $\mathbf{Q}$ be the transition matrix for the simple random walk with transition probabilities q_{ij} given by (1). Suppose that at step k the state of the process is i. A tentative selection is made using the transition probabilities in the ith row of $\mathbf{Q}$. Suppose that the tentative selection is node j. If the out-degree $a_{j\cdot}$ of node j is less than the out-degree $a_{i\cdot}$ of node i, then the selection for the next wave is node j, that is, $W_{k+1} = j$. If, on the other hand, the out degree of node j is greater than the out degree of node i, then a uniform random number Z is selected from the unit interval. If $Z < a_{i\cdot}/a_{j\cdot}$, then $W_{k+1} = j$. Otherwise, $W_{k+1} = i$.

Using the Hastings-Metropolis method (Hastings 1970), the transition matrix for the modified walk in the connected graph is constructed with elements

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{for } i \neq j$$

and

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

where

$$\alpha_{ij} = \min\left\{\frac{a_{i\cdot}}{a_{j\cdot}}, 1\right\}$$

With a population graph containing separate components or isolated nodes, the random walk with jumps, having transition matrix $\mathbf{Q}$ given by (2), can be modified to give

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{for } i \neq j$$

and

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

where

$$\alpha_{ij} = \min\left\{\frac{q_{ji}}{q_{ij}}, 1\right\}$$

Thus, for two mutually connected nodes i and j, the acceptance probability for a transition from i to j is

$$\alpha_{ij} = \min\left\{\frac{(1-d)/N + d/a_{j\cdot}}{(1-d)/N + d/a_{i\cdot}}, 1\right\}$$

For a transition from an isolated unit to one in a component larger than one node, the acceptance probability is $\alpha_{ij} = 1 - d$. Other acceptance probabilities have $\alpha_{ij} = 1$. Note also that for a directed graph, the acceptance probability for following an asymmetric link would be zero.

The uniform walk is implemented, when the current state is i, by selecting a candidate next state, say j, using the transition probabilities in the ith row of $\mathbf{Q}$. A standard uniform random number Z is selected and, if $Z < \alpha_{ij}$, the next state is j, whereas otherwise the walk stays at i for one more step.

The quantity α_{ij} with the uniform walk designs depends on the known transition probabilities of the basic random walk, so does not require estimation for implementation.

4 Targeted Walk

The same approach can be used to construct a walk having any specified stationary probabilities, for example selecting nodes with high y values with higher probabilities or selecting nodes to have probabilities strictly proportional to degree, even when the graph contains separate connected components. Let $\pi_i(y)$ denote the desired stationary selection probability for the ith node as a function of its y value. For example, in a study of a hidden human population at risk for HIV/AIDS, suppose it is desired to sample injection drug users ($y_i = 1$) with twice the probability of noninjectors ($y_i = 0$). The relevant transition probabilities for the value-targeted walk, using again the Hastings-Metropolis method, are

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{for } i \neq j$$

and

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

where

$$\alpha_{ij} = \min\left\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right\}$$

Note that the basic transition probability is known, since it depends only on out-degree of observed nodes, the chosen probability d, and the specified ratio π_j/π_i.

For a walk in which the relative selection probability depends on y value, the ratio $\pi_j(y_j)/\pi(y_i)$ is specified and

$$\alpha_{ij} = \min\left\{\frac{\pi_j(y_j)q_{ji}}{\pi_i(y_i)q_{ij}}, 1\right\}$$

As another example of a targeted walk, the target distribution could be to have nodes selected proportional to their out-degree, that is, the number of links out. Since the degree for an isolated node is zero, one possibility, referred to as the "degree+1" targeted walk, simply adds one to each degree, so that $\pi_i \propto a_{i\cdot} + 1$ is the target selection probability.

A slightly different choice, referred to simply as the degree-targeted walk, adds one only to the degree of isolated nodes, so that $\pi_i \propto \max(a_{i\cdot}, 1)$. For a degree-targeted walk of this type, the acceptance probability for a transition between two mutually connected nodes is

$$\alpha_{ij} = \min\left\{\frac{a_{j\cdot}(1-d)/N + 1}{a_{i\cdot}(1-d)/N + 1}, 1\right\}$$

For a transition between an isolated node and one with positive degree, the probability is

$$\alpha_{ij} = \min(a_{j\cdot}(1-d), 1)$$

The transition probability between two nodes each having positive degree is

$$\alpha_{ij} = \min\left\{\frac{a_{j\cdot}}{a_{i\cdot}}, 1\right\}$$

In that case

$$\alpha_{ij} = \min\left\{\frac{a_{j\cdot}q_{ji}}{a_{i\cdot}q_{ij}}, 1\right\}$$

Since isolated nodes, without any links to other nodes, have degree zero, to give them a positive selection probability their degree can arbitrarily be assigned the value "1" in the degree-targeted walk calculation, or the value 1 can be added to the degree of every node.

5 Nonreplacement walk designs

The limiting distribution results of the previous sections apply exactly to walk designs with replacements, so that the selection of nodes can proceed indefinitely through the finite population. Some of the estimators used in the examples to follow, are based however on the sequence of distinct units selected through that process. The sequence of distinct units, which in effect provides a walk sample without replacement, can add new units only until the number of distinct nodes in the sample equals that of the finite population, at which point the sample mean and the population mean coincide.

A different procedure for selecting a walk sample without replacement is to directly confine the selection of the next unit at any step from the set of units not already selected, as with the "self-avoiding random walk" (Lovász 1993). If a select-reject procedure is used as with the targeted walks, the next selection is made from the set of units not having been tentatively selected at all, whether or not the unit was accepted.

6 Estimators based on the values of the accepted nodes

With a uniform random walk with replacement the draw-by-draw sample mean of the sequence of accepted values is asymptotically unbiased for the mean of the population, because the limiting selection probabilities are all equal. The draw-by-draw sample mean is the nominal mean including repeat values, so a node's value is weighted by the number of times it is selected. With a without-replacement design this same estimator is not precisely asymptotically unbiased because the limiting probabilities are not exactly equal. The standard variance estimator based on a within-walk sample variance is not unbiased because of the dependencies within walks. Variance estimators are examined empirically in the examples.

With a targeted walk in which the limiting probability π_i of node i is proportional to c_i, an asymptotically consistent estimator, based on the limiting probabilities, is provided by the

generalized ratio estimator

$$\hat{\mu} = \frac{\sum_{\mathbf{s}_a} y_i/c_i}{\sum_{\mathbf{s}_a} 1/q_i}$$

Note that the Horvitz-Thompson estimator can not be used because the proportionality constant in the inclusion probabilities is unknown, whereas in the generalized ratio estimator it cancels out. Again the limiting probabilities on which the estimator is based hold exactly for the with-replacement design. For the without-replacement variation, the estimator is examined empirically in the examples.

7 Examples

7.1 Realized stochastic graph

Figure 1 depicts first a small simulated population having 60 nodes. Nodes having value $y = 1$ are colored dark and nodes with value $y = 0$ are light. The entire realization is taken to be our population of interest. The model producing the realization is a stochastic block model in which the probability of a link between any two nodes depends on the values of the nodes. Links are more likely between nodes of the same type, and the dark nodes are more highly connected than the light nodes. For example, it may be of interest to estimate the proportion of positive nodes (that is, nodes with $y = 1$) in the graph. In the population graph, 24 of the 60 nodes are positive, so the true proportion is 0.4. To the right is shown the same graph but with node sizes proportional to the random walk limiting selection probabilities. Because of the higher linkage tendencies of the positive nodes, many of them have higher than average selection probabilities.

In the bottom row of Figure 1 a random walk and a uniform walk selected from the population are shown. Each starts from the same randomly selected node, labeled "1," and proceeds until five distinct nodes are selected. The arrows show the direction of following links and a jump to a new node selected at random from the graph is shown as a dotted line. Note that the random walk backtracks from the third selected node to the second one before

following a new link to the fourth sample node. From the first sample node, the uniform walk passes up the higher-probability node selected by the random walk, accepting instead another of the nodes linked to it. Either of these walks can at any time take a random jump, though in the examples illustrated only the uniform walk happens to take one, in the transition from the third to the fourth sample node.

7.2 Empirical population

Data from a study on the heterosexual transmission of HIV/AIDS in a high-risk population in Colorado Springs (Potterat et al. 1993, Rothenberg et al. 1995) are shown in figures 2 and 3. The 595 people interviewed in the study population are represented by the nodes of the graph, and the reported sexual relationships between the respondents are shown as links between nodes. (Additional sexual links from any of the 595 to persons who were not subsequently interviewed are not shown.) The study population includes at-risk people including injecting drug users, sex workers, their sexual and drug-use partners and other close social contacts. The node variable depicted indicates sex work, with a positive value ($y = 1$) colored dark. Only sexual links are shown, though many coincide with the drug-related links. The largest sexually connected component of the graph contains 219 of the people. The next largest connected component contains 12 people, followed by a number of components of four, three and two people. The remaining nodes represent people without reported sexual contacts within the interviewed population.

The observed pattern of this population, with one connected component very much larger than the others, has been described by researchers as not atypical of studies of hidden, at-risk populations. We are using this population solely as an empirical population from which to select samples to compare sampling designs and estimators.

Figure 3 shows the same population with node size drawn proportional to random walk limiting selection probability.

Each plot of Figure 4 shows a cumulative sample mean of a single walk with is continued

until 120 distinct nodes have been selected. The actual proportion of positive (1-valued) nodes in the empirical population (.2235) is shown by the horizontal line in each plot.

In the top row of Figure 4, an ordinary random walk with a randomly selected starting node is shown. The left plot shows the cumulative sample mean of the distinct units. The lower plot shows the same data but with the draw-by-draw sample mean, which includes repeat selections of the same node, so that each node value is weighted by the number of times that node was selected during the random walk.

In the bottom row of Figure 4 the same two types of sample mean are shown for a uniform walk that is continued until 120 distinct nodes are selected. Notice that, for the ordinary random walk, the sample mean wanders mainly above the actual mean, representing the positive bias resulting from the preferential selection of the more highly connected, high-risk people in the population. For the uniform walk, the sample mean wanders closer to the actual value, sometimes above and sometimes below. Each of these plots also gives indication of the autocorrelation present within a single Markov chain.

The plots in Figure five show the expected node value as a walk progresses wave by wave, for different types of walks and with different initial distributions from which the first node is selected, for the empirical population with 595 nodes. Thus, for the kth wave, the plots show $\mathrm{E}(Y_k)$, where Y_k is the value of the node selected at the kth wave. The dashed line shows the actual mean for the Colorado Springs population (.2235). The other three lines represent three different starting distributions. In all cases, the line that starts out the lowest is the uniform initial distribution, since the mean for the initial randomly selected node equals the mean for the population. The value-dependent initial distribution, in which positive nodes ($y = 1$) have twice the initial selection probability of zero nodes ($y = 0$), gives the expected value line that is in all cases mostly in the middle initially and shows the strongest tendency toward initial periodicity. The degree-based initial distribution, in which initial probability of selection for a node is proportional to its degree (plus one, since isolated nodes have zero degree), forms the top line in each of the plots.

The six plots in Figure 5 show the expected values for six different types of walks. For a random walk that follows links only, without the possibility of random jumps, the long term distribution is dependent on which component the walk starts in, which depends on the initial distribution. The three separate lines in the first figure reflect the sensitivity to the initial distribution. The random walk with jumps, on the other hand, enables any node to be reached from any other so that a limiting distribution is approached quite rapidly whatever the initial distribution. With the uniform random walk, the walk that starts with the uniform distribution stays in the uniform distribution wave after wave, and the walks that start with either of the unequal distributions depicted approach this distribution fairly rapidly. Each of the value-dependent and degree-dependent walks also approaches its limiting distribution fairly rapidly, with the expected node value considerably higher than the average node value in the population. The "degree +1" walk approaches a distribution with selection probabilities proportional to one plus the degree for each node, while the "degree" walk has limiting probabilities proportional to the actual degree except that isolated nodes are assigned degree one.

Tables 1 and 2 show the calculated values of the expected value of y for the Colorado Springs study population for each type of walk, wave by wave, and with different starting distributions for the node selections. Results for ordinary random walks are in Table 1 and for uniform walks are in Table 2. The expected values are shown for the initial selections, waves 1,2,3,4,5,6,8,16, and 32, and for the limit as the number of waves approaches infinity. The three initial distributions, for the selection of the first node of a walk, are random, selection in which positive nodes have twice the probability of zero-valued nodes, and selection proportional to in-degree of each node plus one. Note that, with k independent walks of a given design, the expectations at wave j would apply to the sample mean of the k y-values at wave j from each of the walks.

For the ordinary random walks, starting with the initial sample, the observed value is unbiased for the population value only for the initial selection, and thereafter the bias rapidly

rises to its limiting value of .3303787-.223594. With the initial samples biased toward the positive nodes, the bias changes less as the walk progresses.

For the uniform walk, an initial random selection coincides with the stationary distribution, so that the walk continues to be unbiased wave after wave. With the initial selection in which positive nodes have twice the selection probability of zero-valued nodes, the bias is greatly reduced with each of the first few waves and the selected node values approach their unbiased limiting state. With the initial selection proportional to in-degree plus one, the bias requires a few more waves to become small. The rapid initial approach of the expected value toward the limiting value suggests that it may be desirable to have an initial "burn in" period which is not used in the estimation part. Even a very short burn in of one to three waves could substantially reduce the bias of estimators based on short walks.

Figures 6-9 show the sampling distributions of sample means and weighted estimators for different walk designs with the Colorado Springs data set. Each histogram is based on 1000 simulations of the sampling design applied to the empirical population. For the designs in Figures 6 and 7, each sample consists of 24 walks, each having length 5, that is, continuing until 5 distinct nodes are selected. Figure 5 shows the distributions of sample means for random walks (top row) and uniform walks (bottom row). The distribution of the mean of the 24 sample means of 5 distinct units is given on the left. On the right, the mean of the 24 draw-by-draw means, incorporating repeat selections, is given.

The actual proportion (.2235) of the y values in the empirical population is indicated by the solid triangle, while the mean of the sampling distribution is indicated by the hollow triangle. The sample means for the random walks are biased upward, while the sample means for the uniform walk are nearly unbiased. Neither is precisely unbiased, because of the way the walk continues until a fixed number of distinct nodes is selected, instead of proceeding for a fixed number of waves.

Figure 7 shows the distribution of the generalized ratio estimator for the targeted walks having stationary probabilities related to node value and to degree (node degree plus one).

For comparison purposes, each of these walks was started in its own stationary distribution, in effect giving the distributions of the estimators after "burn in." These estimators are not unbiased, since effective sample size is fixed, which affects the actual probabilities with with distinct nodes are selected in sequence, and because the denominator of the estimator is random, being the sum of the sample weights.

Figures 8 and 9 show the distributions of the same estimators and designs as in Figures 6 and 7, but with each sample consisting on one long walk of 120 distinct nodes.

Tables 3-6 summarize the expected values and mean square errors of the estimators with the various strategies, based on the 1000 simulation runs with the Colorado Springs data set serving as the population.

Tables 7 and 8 give the variance and the expected values of between-walk sample variances, where available, and of within-walk sample variances for the uniform walk designs.

8 Acceptance Rates

The principal advantages of the controlled Markov chain sampling designs, such as the uniform and targeted walks, are (1) they make the limiting selection probabilities known from the data so that they can be used in estimation; (2) the limiting probabilities are chosen, so that certain types of nodes or graph characteristics may be preferentially selected; (3) the estimates are design based and so certain of their key properties do not depend on assumptions, which might turn out to be incorrect, about the population graph itself; and (4) with increasing chain length, the expected values of estimates tend to move toward the corresponding graph quantities even when the initial selection distribution is different from the limiting one. Further, the uniform walk design produces a sample that, without weighting or analysis, is at face value "representative" in some respects of the larger population.

An important practical concern with the uniform and targeted walks is the acceptance rate, that is, the average probability a tentatively selected node is accepted. Tentatively selected

nodes that are rejected do not contribute to the simple estimators. For a population such as the Internet, in which tentative selections and accept/reject decisions can be automated and made quickly, the acceptance rate may not be critical. Sampling simply continues until a suitable number of nodes are accepted. For studies of hidden human populations, sample sizes tend to be small. Members of the population are difficult to find and interviews may be time consuming. In some studies, however, the decision to accept or reject, based on a tentatively selected person's out degree, may be fairly quickly ascertained through a short screening interview. Even so, it is desirable to have a sampling method with as high an acceptance rate as possible.

The random walks have acceptance probability equal to one, but do not in general have known or controlled limiting probabilities. If one thinks of the underlying random walk as the natural, uncontrolled walk through a population, then a controlled walk having a limiting distribution close to the natural random walk of the population would be expected to have a higher acceptance rate than a controlled having a limiting distribution very different from the natural random walk. That is, a controlled walk with a stationary distribution not far from the underlying random walk distribution should require less modification through the rejection of tentatively selected nodes than one with stationary distribution far from the natural random walk tendencies.

As mentioned earlier, the stationary probabilities for an ordinary random walk in a nondirected graph with a single component are proportional to the degrees of the nodes. When there is more than one connected component, the random jump innovation is necessary to ensure that every node is reachable and to produce a single stationary distribution not dependent on the starting distribution, and the limiting probabilities are influenced by, but not strictly proportional to, the node degrees. Even with the random jump innovation and the induced acceptance probabilities, the targeted walks producing stationary probabilities proportional to node degrees may be the closer than the other controlled walks under consideration to the natural random walk distribution. Indeed, in Figure 5 it is evident that, for the empirical

population, the equilibrium distribution of the expected node value for the degree+1 walk is closer to the equilibrium for the random walk with jumps than is any of the other controlled designs studied.

For the empirical population from the HIV/AIDS heterosexual transmission study, the acceptance rates for the different designs are given in Table 9. For the uniform walk design, the acceptance rate was 62 percent. For the value walk, giving twice the limiting probability for the high risk as for the low risk people, the acceptance rate was 60 percent. For the degree walk, in which the limiting probability was proportional to the degree plus one, the acceptance rate was 85 percent. For the degree walk with ond added only for the degree of the isolated nodes, the acceptance rate was 88 percent.

9 Discussion

The uniform and target walk sampling designs serve to make the limiting selection probabilities known from the data so that they can be used in estimation. Further, the limiting probabilities are chosen, so that certain types of nodes or graph characteristics may be preferentially selected. Dependence on the initial selection, which may be uncontrolled, decreased step by step.

The estimators used in this paper with the uniform and targeted walk designs can be said to be design based. Even though the exact design based selection probabilities may be unknown if they are unknown in the initial selection, the stationary selection probabilities are used in the estimators. With increasing chain length, these probabilities become more accurate and the expected values of estimates move toward the corresponding graph quantities. The design based estimation methods have the advantage that certain of their properties, such as design unbiasedness or consistency, do not depend on model based assumptions that would possibly be incorrect. The design based estimates have the additional attractive quality that they are very simple and easy to understand and explain, and can even produce data that can be

presented without analysis or interpretation as representative in important characteristics of the wider population of interest.

The use of Markov Chain Monte Carlo algorithms for data analysis with complicated models is common in statistics. The methods described here are unusual in that the Markov Chain methods are applied to real-world populations to actually obtain the data, with the result that the data thus obtained can be easily analyzed by hand. In fact, one could go a step farther and construct a complex Bayes stochastic graph model for the population, using Markov Chain Monte Carlo methods in the conventional fashion in analyzing the data as well as in their collection.

The uniform or targeted walk designs are useful to obtain samples of accepted nodes that have certain desirable properties in relation to the population, that provide very simple estimators of population quantities, or that could provide an initial sample for another design. It should be noted that nodes that were observed but then "rejected" under the design are actually still part of the data. Their values can still be incorporated into estimates if desired using the Rao-Blackwell method applied once the chain has reached approximate equilibrium, though the estimates then are computationally complex.

Another alternative is to use model based methods such as Bayes estimates. The model based methods require, in addition to adequate stochastic graph modeling of the population, an ignorable initial selection procedure, which is not in general satisfied with initial selections biased by node or degree values, or else adequate modeling of the nonignorable selection procedure as part of the likelihood. Targeted walk designs producing an asymptotic distribution unrelated to the nonignorable selection procedure and hence approximately unrelated to node or degree values outside of the sample could provide the initial selections for a sample with which model based inference methods could then be applied.

10 Acknowledgements

Support for this work was provided by funding from the National Center for Health Statistics, the National Science Foundation (DMS-9626102), and the National Institutes of Health (R01-DA09872). I would like to thank John Potterat and Steve Muth for advice and use of the data from the Colorado Springs study.

11 References

Birnbaum, Z.W., and Sirken, M.G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. *Vital and Health Statistics,* Ser. 2, No.11. Washington:Government Printing Office.

Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference.* Elsevier, 107-117.

Chow, M. and Thompson, S.K. (2003). Estimation with link-tracing sampling designs—a Bayesian approach. *Survey Methodology* **20** 197-205.

Felix-Medina, M.H. and Thompson, S.K. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics,* **20** 19-38.

Frank, O. (1977). Survey sampling in graphs. *Journal of Statistical Planning and Inference* **1** 235-264.

Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks* **1** 91-101.

Frank, O., and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics* **10** 53-67.

Hastings, W.K. (1970). Monte-Carlo sampling methods using Markov chains and their application. *Biometrika* **57** 97-109.

Heckathorn, D. D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems* **44** 174-199.

Heckathorn, D. D. (2002). Respondent driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems* **49** 11-34.

Henzinger, M.R., Heydon, A., Mitzenmacher, M., and Najork, M. (2000). On near-uniform URL sampling. *Proceedings of the Ninth International World Wide Web Conference.* Elsevier. pp. 295-308.

Klovdahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. In M. Kochen, ed., *The Small World,* Norwood, NJ: Ablex Publishing, 176-210.

Lovász, L. (1993). Random walks on graphs: A survey. In Miklós, D., Sós, D., and Szöni, T., eds., *Combinatorics, Paul Erdös is Eighty,*, Vol. 2, pp. 1-46. János Bolyai Mathematical Society, Keszthely, Hungary.

Potterat, J. J., Woodhouse, D. E., Rothenberg, R. B., Muth, S. Q., Darrow, W. W., Muth, J. B. and Reynolds, J. U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS* **7** 1517-1521.

Rothenberg, R.B., Woodhouse, D.E., Potterat, J.J., Muth, S.Q., Darrow, W.W. and Klovdahl, A.S. 1995. Social networks in disease transmission: The Colorado Springs study. In Needle, R.H., Genser, S.G., and Trotter, R.T. II, eds., *Social Networks, Drug Abuse, and HIV Transmission.* NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse. 3-19.

Rubin, D.B. (1976), Inference and missing data. *Biometrika* **63** 581-592.

Salganik, M. J. and Heckathorn, D.D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling". *Sociological Methodology* **34** 193-239.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: what and why? *Bulletin de Methodologie Sociologique* **36** 34-58.

Thompson, S.K. and Collins, L.M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence* **68** S57-S67.

Thompson, S.K. and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology* **26** 87-98.

Table 1. Random walks: Expected value of y for waves 0, 1, 2, 3, 4, 5, 6, 8, 16, 32, and infinite. Wave 0 is the initial selection. Three different initial selection probability assumptions are used: Initial random selection ($\pi_0 = 1/N$ for all nodes), nodes with value $y = 1$ have twice the selection probability of nodes with value $y = 0$ ($\pi_0 \propto y + 1$), and initial selection probability proportional to in-degree plus one ($\pi_0 \propto a_{\cdot j}$). The actual mean of the node values for this population is 0.2235294

wave	$\pi_0 = 1/N$	$\pi_0 \propto y + 1$	$\pi_0 \propto a_{\cdot j}$
0	0.2235294	0.3653846	0.3349894
1	0.2998771	0.2752690	0.3560839
2	0.3005446	0.3587093	0.3507451
3	0.3273606	0.3082865	0.3570490
4	0.3177081	0.3594697	0.3500041
5	0.3320705	0.3179675	0.3528395
6	0.3231213	0.3542086	0.3469835
8	0.3256034	0.3490933	0.3440449
16	0.3291087	0.3372548	0.3363884
32	0.3302606	0.3313908	0.3315119
∞	0.3303787	0.3303787	0.3303787

Table 2. Uniform walks: Expected value of y for waves 0, 1, 2, 3, 4, 5, 6, 8, 16, 32, and infinite, with three different initial selection assumptions.

wave	$\pi_0 = 1/N$	$\pi_0 \propto y + 1$	$\pi_0 \propto a_{\cdot j}$
0	0.2235294	0.3653846	0.3349894
1	0.2235294	0.2590239	0.2903147
2	0.2235294	0.2741356	0.2877974
3	0.2235294	0.2447258	0.2761270
4	0.2235294	0.2511473	0.2707929
5	0.2235294	0.2372440	0.2646280
6	0.2235294	0.2420866	0.2600923
8	0.2235294	0.2371714	0.2522952
16	0.2235294	0.2285370	0.2352150
32	0.2235294	0.2243635	0.2256228
∞	0.2235294	0.2235294	0.2235294

Table 3. Means and mean square errors for sample means of distinct units and draw-by-draw means for random walks and uniform walks. The design uses 24 walks each continuing until 5 distinct nodes are included.

design:	random walk	random walk	uniform walk	uniform walk
estimator:	sample mean	draw mean	sample mean	draw mean
mean	0.3008000	0.2994872	0.2423000	0.2289125
m.s.e.	0.007617465	0.007608868	0.002016378	0.001974826

Table 4. Means and mean square errors for weighted means (generalized ratio estimator), using the distinct units in each walk or the draw-by-draw selections for value-dependent walks and degree-dependent walks. The design uses 24 walks each continuing until 5 distinct nodes are included.

design:	value walk	value walk	degree walk	degree walk
estimator:	distinct units	draw by draw	distinct units	draw by draw
mean	0.1805114	0.2144555	0.2235257	0.1994530
m.s.e.	0.002546968	0.001195507	0.001807981	0.004382568

Table 5. Means and mean square errors for sample means of distinct units and draw-by-draw means for random walks and uniform walks. The design uses one walk continuing until 120 distinct nodes are included.

design:	random walk	random walk	uniform walk	uniform walk
estimator:	sample mean	draw mean	sample mean	draw mean
mean	0.3274083	0.3325171	0.2379333	0.2232534
m.s.e.	0.012004961	0.014902382	0.001777285	0.002442825

Table 6. Means and mean square errors for weighted means (generalized ratio estimator), using the distinct units in each walk or the draw-by-draw selections for value-dependent walks and degree-dependent walks. The design uses one walk continuing until 120 distinct nodes are included.

design:	value walk	value walk	degree walk	degree walk
estimator:	distinct units	draw by draw	distinct units	draw by draw
mean	0.1652275	0.2254267	0.2404622	0.1835336
m.s.e.	0.003952703	0.001578039	0.002115518	0.003951540

Table 7. Variance of estimators and expected values of between-walk and within-walk sample variances for the uniform random walk, for the design with 24 walks of 5 distinct nodes each.

estimator:	sample mean	draw-by-draw mean
variance of estimator:	0.001665709	0.001947796
E(between-walk variance)	0.001584203	0.001919005
E(average within-walk variances)	0.001515521	0.001231983

Table 8. Variance of estimators and expected values of within-walk sample variance for the uniform random walk, for the design with a single walk of 120 distinct nodes. (No between-walk sample variance is available for this design.)

estimator:	sample mean	draw-by-draw mean
variance of estimator:	0.001571384	0.002445194
E(average within-walk variances)	0.001510515	0.001429126

Table 9. Acceptance rates for the uniform and targeted walks in the empirical population.

design:	uniform walk	value walk	degree+1 walk	degree walk
acceptance rate	0.62	0.60	0.85	0.88

Figure 1: Top left: Population is realization of stochastic block graph model. Top right: The random walk limit probabilities of the nodes. Bottom left: Random walk of 5 steps. Bottom right: Uniform walk of 5 steps. Arbitrary axes scales are provided as a visual aid in identifying sample nodes with population nodes.

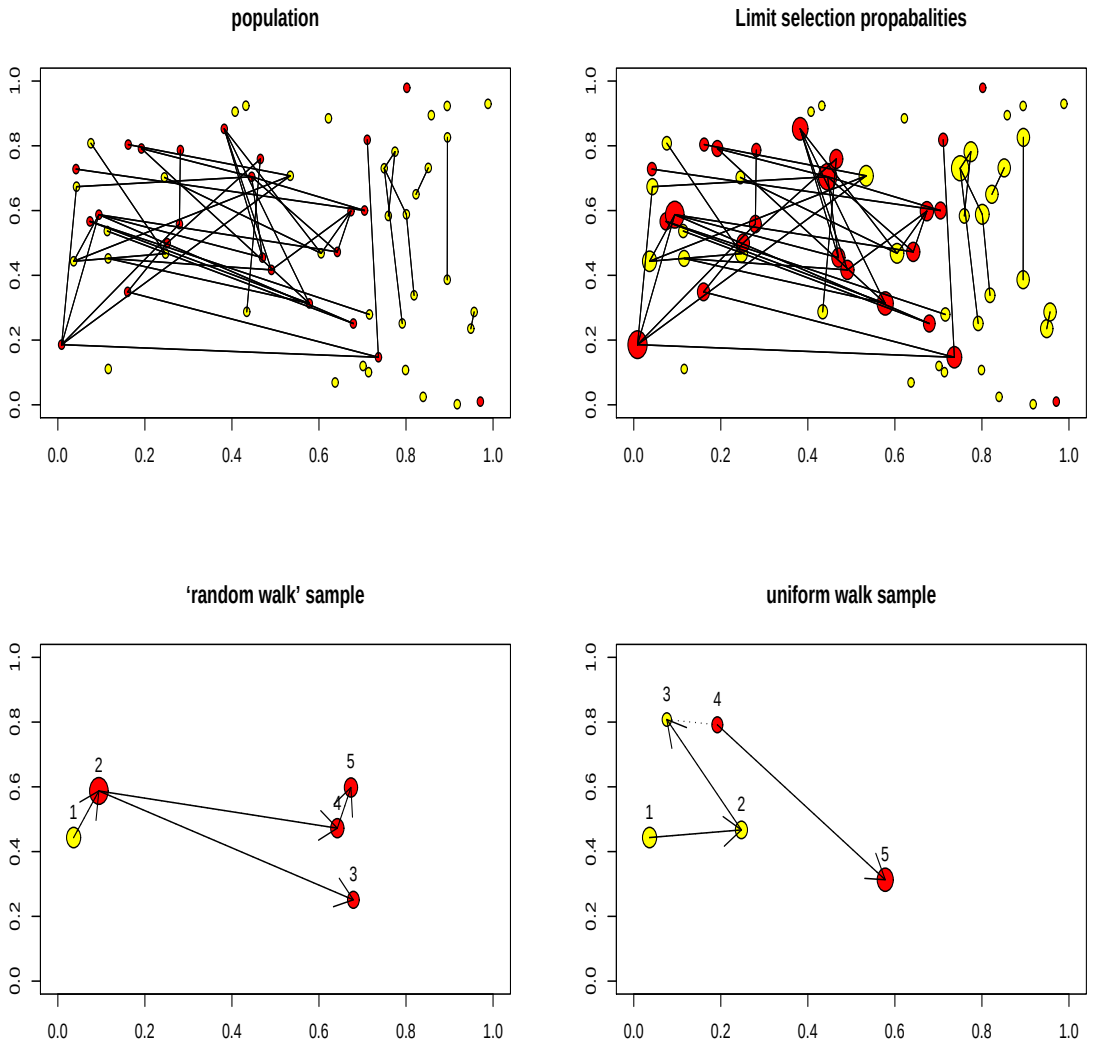

Figure 2: High-risk population in Colorado Springs study on the heterosexual transmission of HIV/AIDS (Potterat et al. 1993, Rothenberg et al. 1995, and personal communications). Red circles represent highest-risk individuals, in this case those who have exchanged sex for money. Links shown between individuals are sexual and drug injecting partnerships.

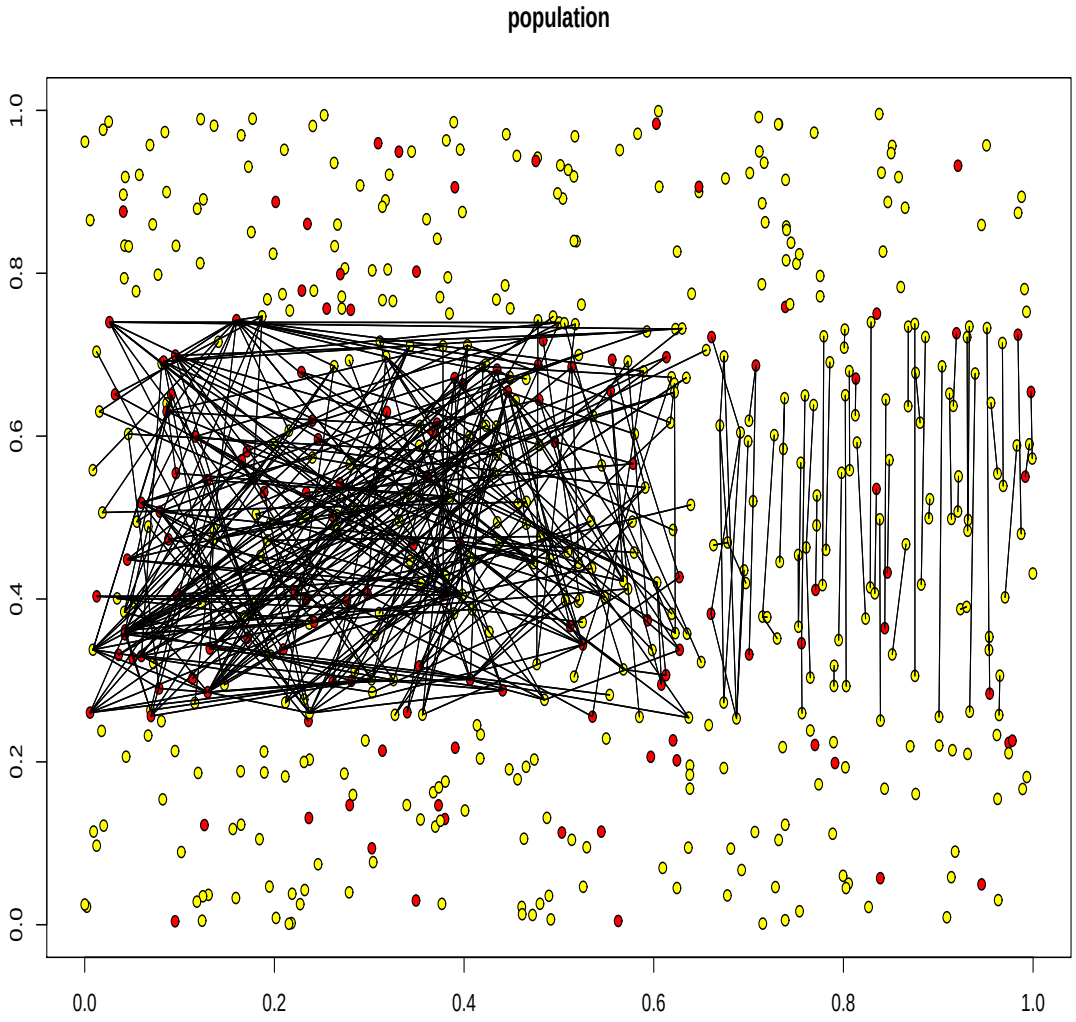

Figure 3: Limiting random walk selection probabilities for Colorado Springs population. Notice that in the real population many of the individuals with the highest-risk behavior also have high selection probabilities with the ordinary random walk, and so will tend to be over-represented in a sample.

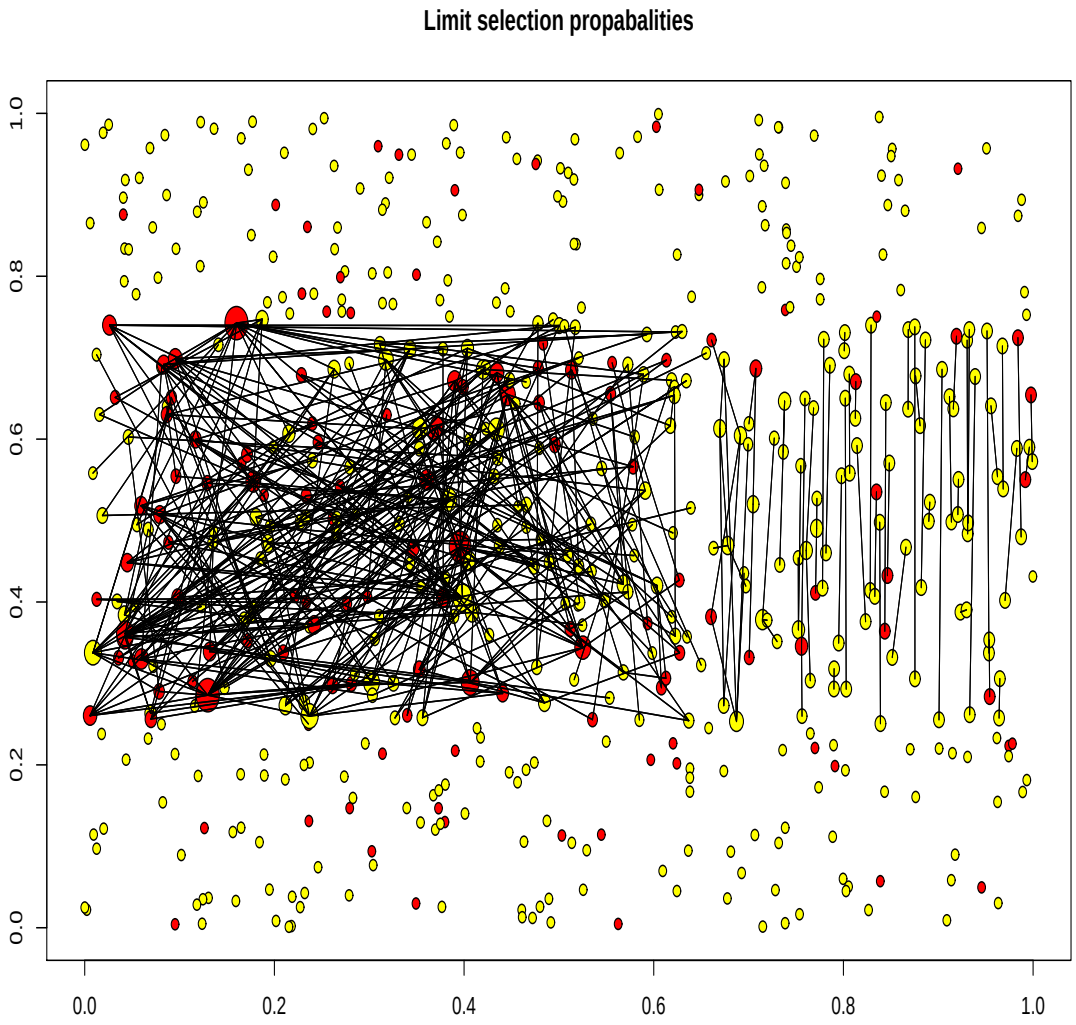

Figure 4: Sample paths of sample means for a single random walk of length 120 nodes. The top two plots are with an ordinary random walk, while the bottom two are with a uniform walk. Sample mean of the distinct units, up to the wave given by the x-axis, is plotted on the left. On the right is the sample mean of the nominal draws, so that node value is weighted by the number of times the node is selected.

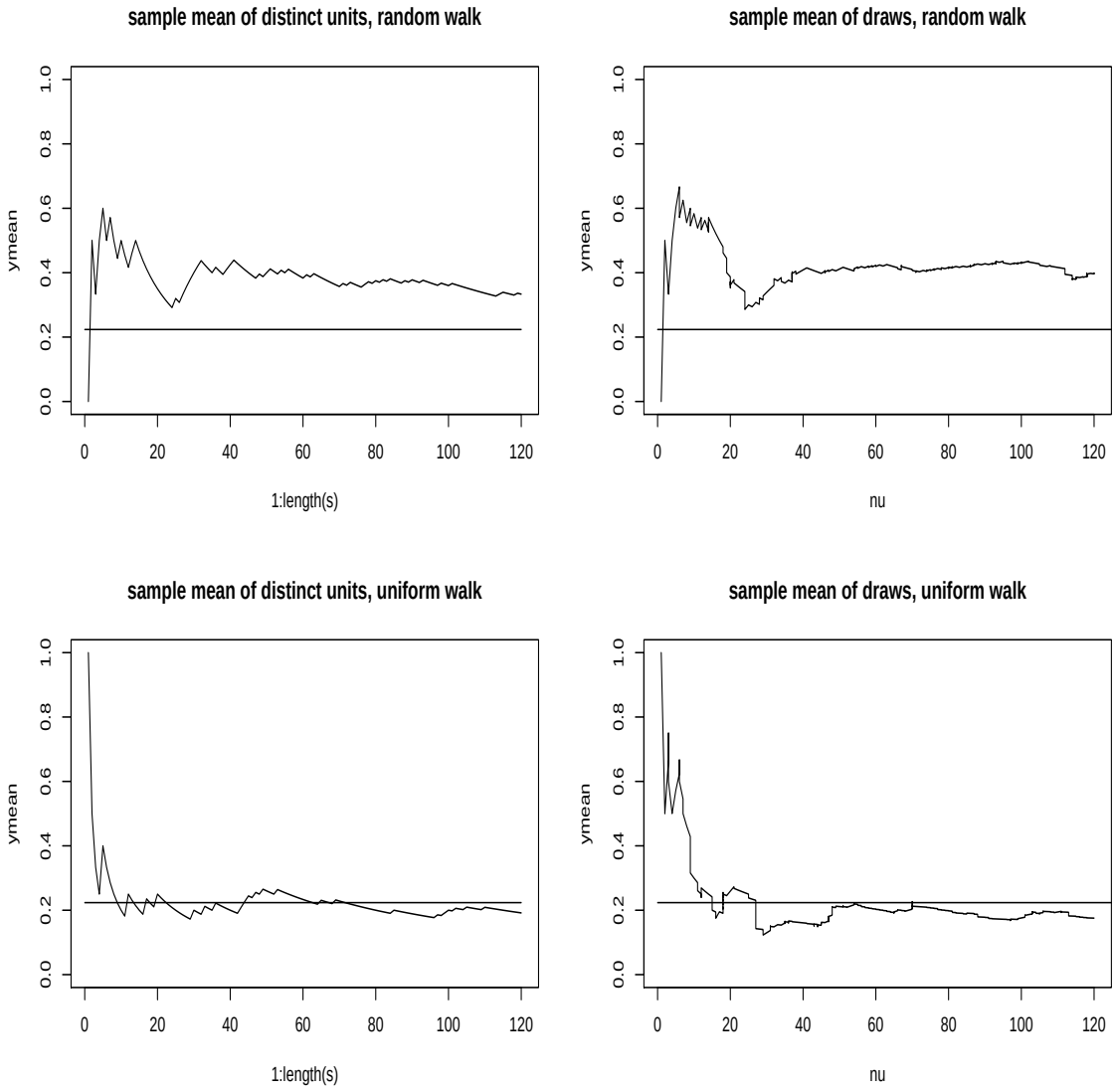

Figure 5: Expected value of node by wave for different walk designs with the Colorado Springs empirical population. Each plot shows one walk design. The dashed line is the actual mean. The other three lines show expected value for three different starting distributions. In each case the lower of the three lines starts with the uniform distribution, the middle line with the value 2/1 distribution, and the top line with the degree distribution.

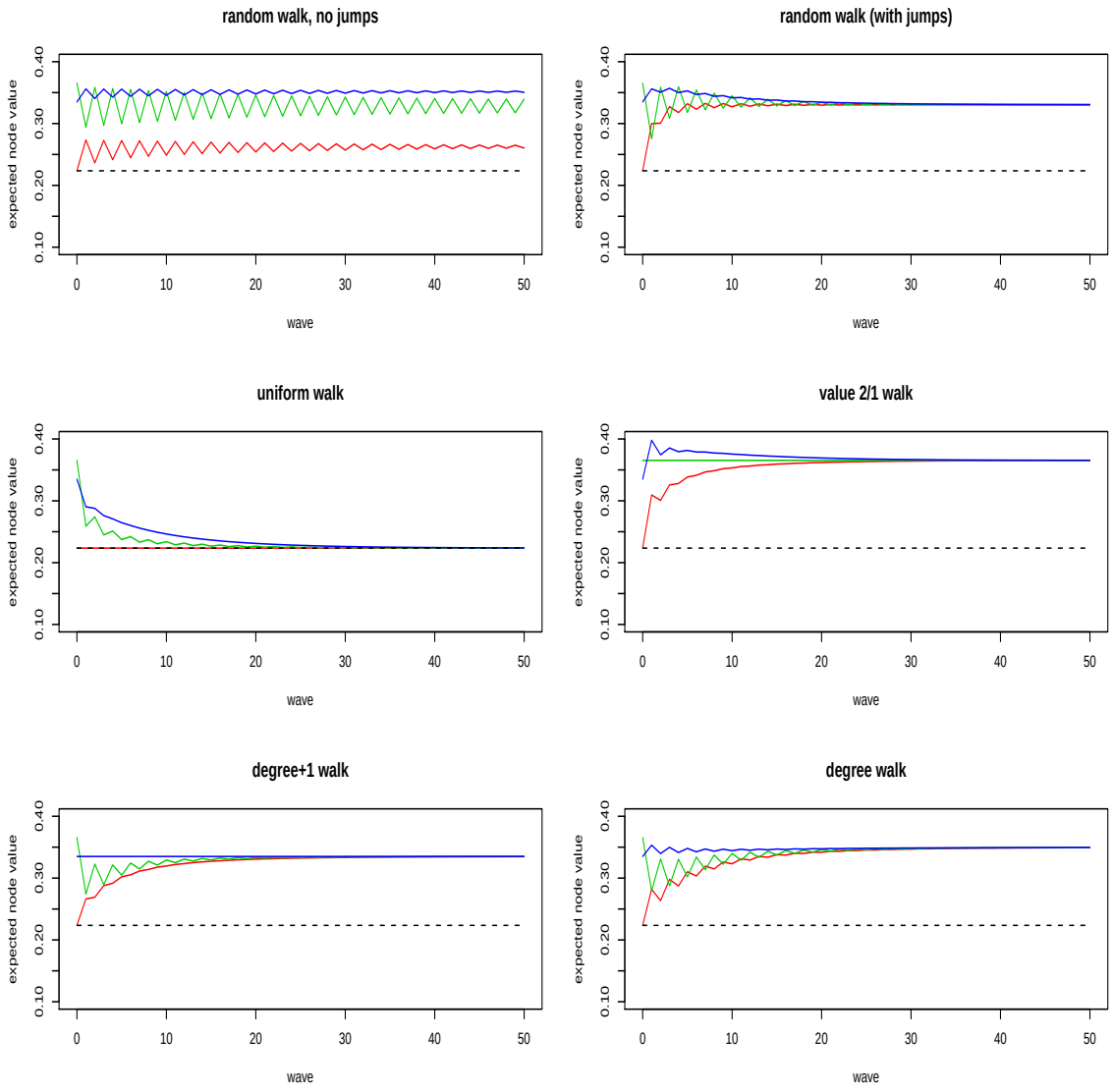

Figure 6: Distributions of sample means as estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with random and uniform walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means or ordinary random walks. Random walks are at top, uniform walks at bottom. Design was 24 walks, each of length 5, with all 120 observations used in the estimator. The number of realizations for the simulation was 1000.

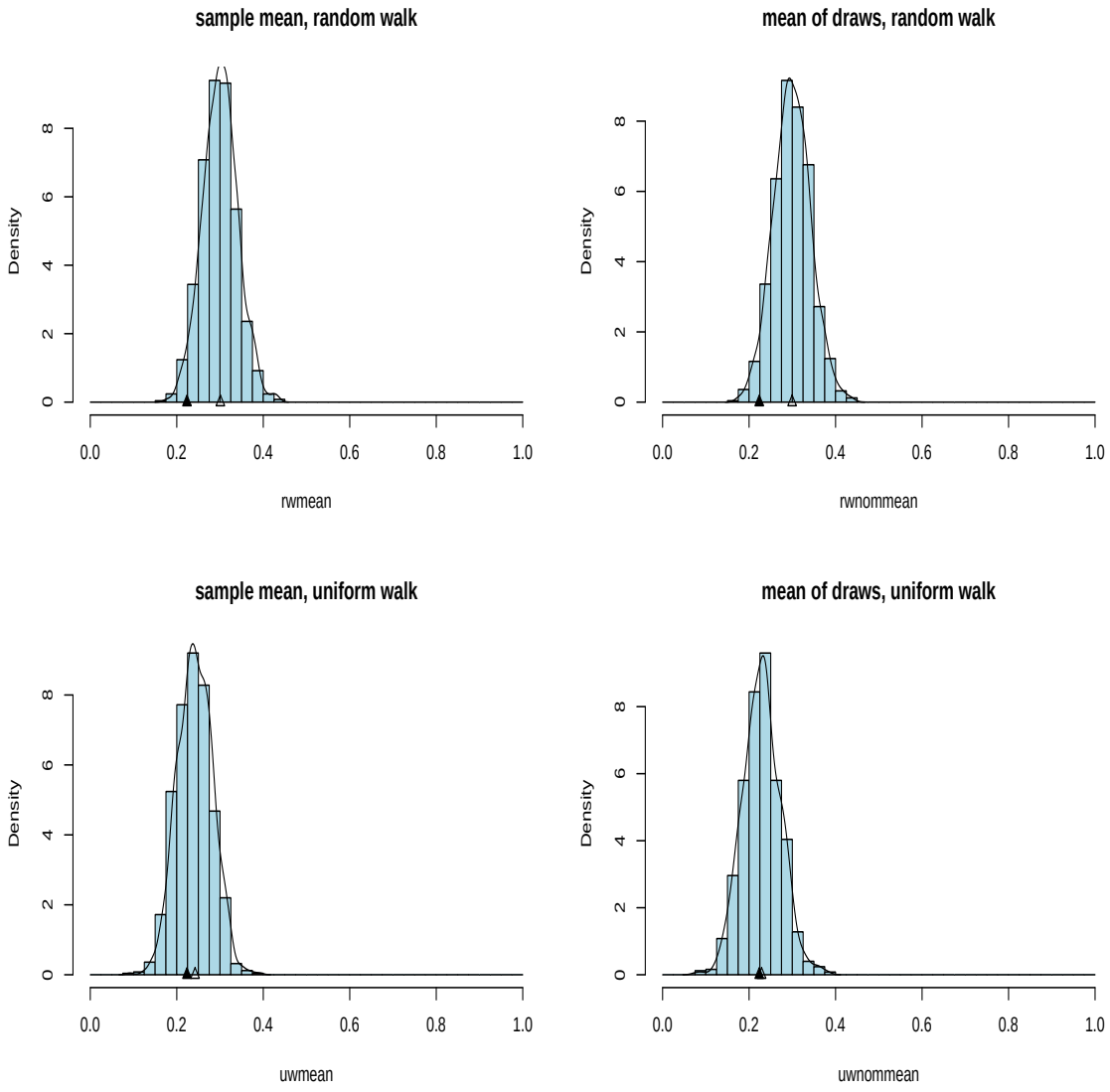

Figure 7: Distributions of generalized ratio estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with targeted walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means or ordinary random walks. Random walks are at top, uniform walks at bottom. Design was 24 walks, each of length 5, with all 120 observations used in the estimator. The number of realizations for the simulation was 1000.

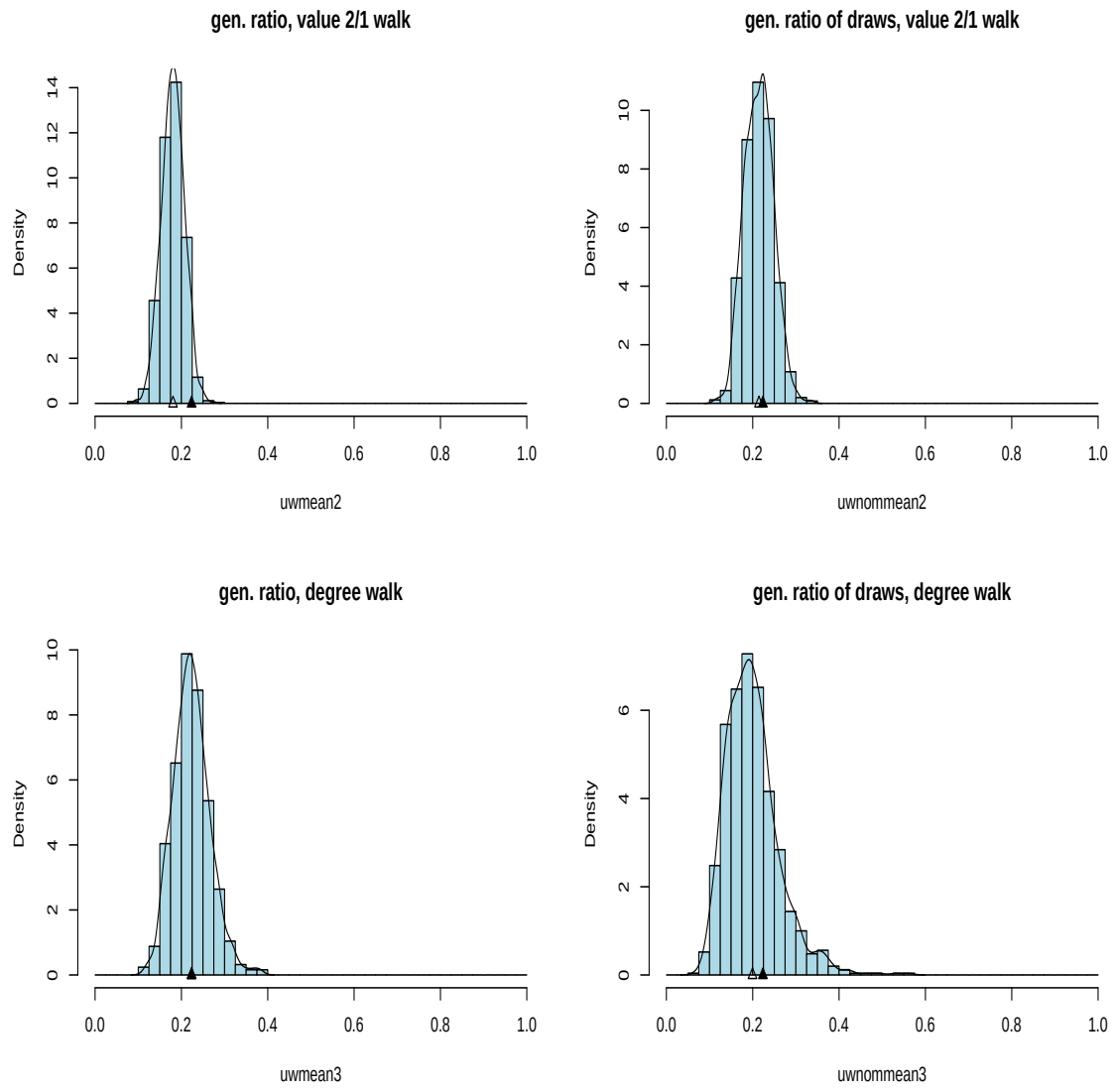

Figure 8: Distributions of sample means as estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with random and uniform walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means or ordinary random walks. Random walks are at top, uniform walks at bottom. Design was a single walk of length 120. The number of realizations for the simulation was 1000.

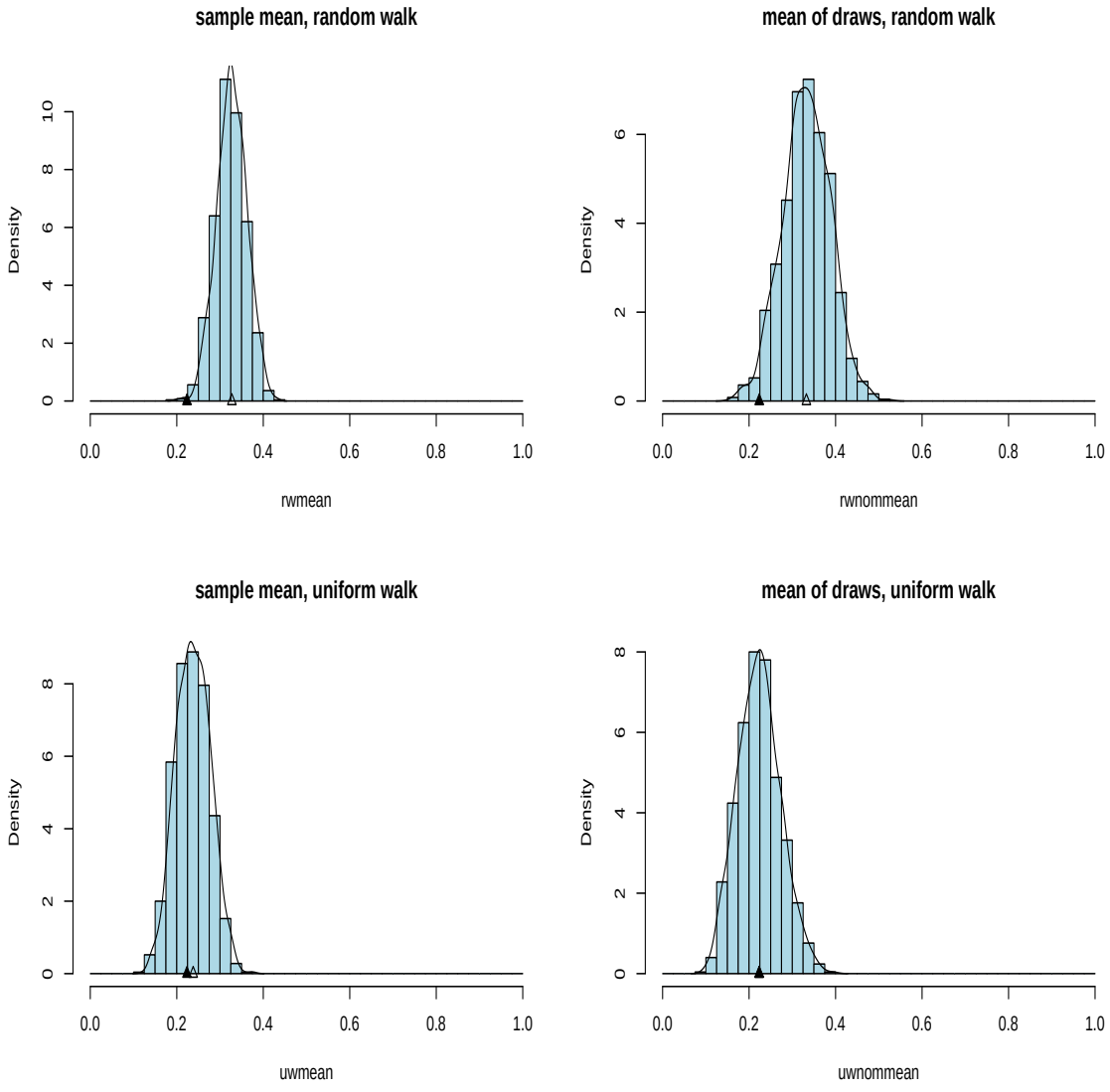

Figure 9: Distributions of generalized ratio estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with targeted walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means or ordinary random walks. Random walks are at top, uniform walks at bottom. Design was a single walk of length 120. The number of realizations for the simulation was 1000.