Implications for researchers employing web-based Respondent Driven Sampling

by

Heather RC Franklin

FINAL THESIS

Presented to the Department of Public Health and Preventive Medicine
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of
Master of Public Health

August 2010

Department of Public Health and Preventive Medicine

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Heather Rebecca-Craig Franklin

has been approved

Jodi Lapidus, PhD – Chair

Sean Schafer, MD

Mark Loveless, MD

Rochelle Fu, PhD

Table of Contents

List of Tables **Page**

List of Figures **Page**

Glossary

Bootstrapping: statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample, with the purpose of deriving robust estimates of standard errors and confidence intervals of a population parameter

Cluster: degree to which people in a network group together

Equilibrium: The equilibrium sample population proportions indicate each group's size after the proportions have converged to their equilibrium value. This occurs when further recruitment waves do not change the population proportion by a significant amount.

Ergodic: every person in the target group has a non-zero chance of being recruited

Markov Chain: is a discrete random process with the property that the next recruitment wave depends only on the current wave

Social Network Size: total number of people that a person interacts with in a specific network

Observed Proportion: also called the ""I" estimates of population proportions. The term naïve is used because the proportion is a simple ratio of how many of a particular group was recruited to the total number of recruits. It is not adjusted for any statistical biases.

Reciprocity: the absolute links from group m to group n in the population equals the number from group n to group m

Reciprocity Prevalence: estimate of the proportional size of the MSM population based on two sources of data: transitional probabilities derived from the analysis of recruitment patterns (Markov process), and self-reported personal network size.

Respondent Driven Sampling: Respondent-driven sampling (RDS), comb"nes "snowball sam"ling" with a mathematical model that weights the sample to compensate for the fact that the sample was collected in a non-random way

Seed: RDS starts with a small number of peers (called 'seeds') and expands through successive "waves" of peer recruitment

Snowball Sampling: a technique for developing a research sample where existing study subjects recruit future subjects from among their acquaintances. Thus the sample group appears to grow like a rolling snowball. As the sample builds up, enough data is gathered to be useful for research

Transitional Probability: probability of one group recruiting another

Wave: seeds recruit first-**wave** respondents, first-wave respondents recruit second-**wave** respondents, and this process continues until the desired sample size is reached.

Abstract

Estimating prevalence proportion and incidence rate of HIV/AIDS in concealed or hard-to-reach populations such as men who have sex with men (MSM) is hampered by the lack of accurate estimates of at-risk population size. Estimates of MSM populations have most commonly been made by substituting self-reported sexual orientation data collected via random household surveys for sexual behavior data. Respondent-Driven Sampling (RDS) is a recently developed tool used successfully to sample hard-to-reach populations typically missed by traditional sampling frames. Our recent, failed attempt to employ an anonymous, web-based model lead to a simulation study of RDS scenarios to evaluate how transitional probability structures, average social network sizes and seed composition interact to affect: (1) when/if equilibrium was attained; (2) composition of the final sample; (3) observed prevalence; (4) reciprocity. Researcher employing a RDS tool will benefit from: collecting pilot data in order to estimate transitional probabilities and average network sizes; ensuring that the seed composition reflects pilot data; and implementing a careful, deliberate incentive structure.

Background

The need for accurate estimates of the number of men who have sex with other men.

Preventing new HIV infections depends on systematic study of both determinants of infection and outcomes of new interventions. Both of these activities require that at-risk populations are comprehensively accessed. Unfortunately, when infection is associated with stigmatized behavior such as same-sex sexual activity or illegal activity such as injection drug use, counting, studying and intervening effectively among these groups is difficult.

The Centers for Disease Control and Prevention estimates that 5-7% of the U.S. male population is comprised of men who have sex with men (MSM); yet, MSM account for two thirds of new HIV infections. The precise number of men who have sex with men (MSM) is unknown; as a substitute, the proportion of men who report gay, bisexual, or homosexual sexual orientation via telephone survey is multiplied by census estimates of males in the general population. This method of estimating the size of the MSM population is probably inaccurate because of traditional flaws in population-based surveys such as low response rates and also because homosexual activity is probably underreported by telephone survey because of the attendant social stigma. In addition, survey questions ask about sexual orientation and not about actual sexual behaviors and frequency, meaning that some men who report being gay or homosexual might not have actually engaged in sex with other men.

Snowball sampling (aka chain-referral sampling) has been another approach taken to identify and quantify attributes of hidden populations. This approach efficiently identifies and includes members of hidden populations by having ‘seeds’ who are members of the target population and will recruit other members. Investigators have tried to achieve representativeness by purposeful selection of diverse seeds. However the approach ultimately lacks validity or precision and cannot be generalized to the target populations because the

composition of the sample is dependent upon the choice of seeds (initial recruits) and length of recruitment chains.

Respondent Driven Sampling

In an effort to representatively sample hidden populations such as injection drug users, Heckathorn proposed and piloted Respondent Driven Sampling (RDS) and developed theoretic and empiric arguments for its validity when used to estimate population proportions. (Heckathorn 1997; Heckathorn et al. 2002). Respondent-driven sampling combines snowball sampling with a mathematical model that validly adjusts for non-random collection, and mitigating bias inherent in the snowball approach. Unlike traditional random sampling methods, RDS does not require a sampling frame to calculate the probability of selecting a sampling unit. Rather than estimating from the population, RDS relies on information about the social network to draw population estimates (Heckathorn 1997 & Heckathorn 2002).

RDS starts with a non-randomly selected group of participants (i.e. 'seeds') from the target population. It differs from snowball sampling in that subjects are not asked to identify peers for subsequent recruitment by investigators but to recruit them directly into the study. This distinction is particularly important because it reduces masking (protection of others) by allowing peers to decide for themselves whether they wish to participate. RDS is premised on the assumption that group members can most efficiently and effectively recruit other members. In accordance, highly networked, motivated individuals are selected as the initial seeds.

To prevent imbalances generated by differential recruiting effectiveness, Heckathorn suggests implementing a quota of recruitment incentive coupons, typically 5 or fewer. Each coupon was identified by a unique code that related the recruit to his recruiter. Seeds were then asked to recruit members of the target population from their social network to participate in the survey. After consenting and completing the survey, each recruit was in turn given a quota of

coded coupons and asked to become recruiters themselves. This recruitment process continued for multiple waves, either until the pre-determined sample size was reached or until the sample composition stabilizes with respect to the target variables. This was known as equilibrium.

Heckathorn showed that upon attainment of equilibrium the sample composition is independent of the seeds from which sampling began. He showed that equilibrium can be estimated from the tendency to recruit others who share the trait in question, known as within group recruiting. He also showed that the underlying prevalence of the attribute in the population is a function of the average size of respondent social networks, and within group recruiting. Since both within group recruiting and the size of each respondent's social network can be measured, the underlying population prevalence of the trait can be calculated. Confidence intervals around the estimates can be calculated by a bootstrap-like procedure that uses hypothetical seeds and recreates the sample randomly using within group recruitment probabilities as parameters. Though valid estimates can be obtained from relatively short recruiting chains, longer recruitment chains increase precision and sensitivity for obscure or infrequent traits. (Salganik & Heckathorn 2004).

An important design element in RDS involves techniques for increasing the lengths of referral chains. Traditionally this has been addressed by providing dual incentives: (1) upon completion of the survey and (2) upon meeting recruitment quotas.

Although final RDS sample composition is independent of seed composition, choice of seeds can affect the rate which equilibrium is reached and the rate of sampling. The speed with which sampling goals are attained is dependent upon whether initial seeds will be productive initiating the recruitment chain.

The RDS Sampling Frame

RDS provides a means of mapping relationships by determining the proportional size of a population and its internal structure (DiMaggio 2000). In contrast to traditional sampling methods, in RDS the sampling frame is created after sampling is complete. Recruiter-recruit chains are documented which allows potential sampling biases to be quantified. Respondents were asked how many other members of the target population (i.e. adult males that reside in the Portland-Metro region, they know). The inclusion of an individual is proportional to their network (number of people in the target population that he personally knows). Once a sample reaches equilibrium all contacts within the target population have equal probability of being used for recruitment.

As originally presented by Heckathorn (1997) RDS can be modeled as a Markov process with two essential characteristics. First, the process must assume a limited number of states (here MSM or Non-MSM). Second, the process is state dependent; the probability of moving from state to state depends on transitional probabilities. Thus, the probability that the next recruit will come from a given group depends on the group from which the current recruiter comes (Salganik & Heckathorn 2004).

Heckathorn (1997) has shown that a Markov model of recruitment has several important and unique assumptions. First, the process is memory less; recruitment patterns depend only on the recruiter, not the recruiters' recruiter. This designates RDS as a first order Markov process. Second, the process is ergodic meaning that there is a non-zero probability that any state will never occur. This means that recruitment cannot become trapped within a single group (i.e. whites recruiting only whites). When RDS meets these two criteria it follows a regular Markov process

Ensuring that equilibrium is attained is vital to ensure that the sample's characteristics reflect the population of interest rather than simply the initial seeds. Heckathorn (2002) graphically demonstrated that as the sampling process progresses, the effect of the starting point weakens until its' effect is negligible. Convergence of Markov chains occurs at a geometric rate; equilibrium has been typically approximated within six waves (Heckathorn 1997).

Social Network Estimates

RDS analysis requires information on the focal variable (i.e. MSM) as well as two additional types of information that provide the sampling frame from which population estimates are calculated: (1) transitional probabilities (aka cross-group recruitment); and (2) self-estimated mean social network size. If one group is oversampled estimates will be exaggerated. Additionally, groups with larger networks sizes are also oversampled.

Equilibrium

By the law of large numbers for regular Markov chains, as the recruitment process continues from wave to wave, an equilibrium state will be attained that is independent of the characteristics of the initial seeds. Allowing recruitment to operate until equilibrium is reached avoid the central problem in snowball sampling – that the final sample merely reflects the initial sample (Heckathorn 2002).

If one group recruits more effectively Equilibrium can be calculated using the proportion of crosscutting ties (S) of both MSM and Non-MSM respondents. Heckathorn (2002) has found that equilibrium should be reached within 6 waves. Equilibrium can be used as a population estimator – if and only if – the proportion of cross-cutting ties and social network sizes are equal. Variance in either of these social network estimates requires correction with the reciprocity model.

Reciprocity Prevalence Estimate

The final analysis in RDS is to calculate prevalence using the Network size (N) and the proportion of crosscutting ties (S) to estimate the population. The reciprocity model provides an estimate of the proportional size of the MSM population based on two sources of data: transitional probabilities derived from the analysis of recruitment patterns (Markov process), and self-reported personal network size (Heckathorn 2002).

Lesson from failure

Public Health interventions in Portland, Oregon are in need of a precise estimate of the size and structure of the MSM population in the area. Encouraged by Heckathorns' work, an anonymous, web-based RDS study was developed with the goal of attaining a non-biased estimate of MSM in the Portland metro area.

Five males, diverse by age, race and sexual orientation were recruited by the author from her own social and professional network as seeds. They were chosen intentionally as they are motivated and enthusiastic people from various Portland area colleges, community-based organizations, and public health agencies.

After the author informed the seeds about the project, and they consented to participate, the author sent them an e-mail officially informing them of the project's purpose and recruitment process. The email directed them to visit the project website, where they found a project identifier and unique identifier code. While there, they completed the study questionnaire and agreed to recruit five additional participants from their own social network.

Seeds were asked to recruit other men from their network who resided in the Portland metropolitan region and were aged greater than 18 years but otherwise without regard to recruit race, age or sexual orientation. Seeds were not restricted to any communication medium such as email, telephone or direct conversation for recruiting others to participate. Seeds were

directed by the project website to give their unique identifier code and the web address of the project website to each of their recruits and to ask the recruits to visit the site, consent to participation, and complete the questionnaire. Recruits who visited the site and completed the questionnaire were each give new unique identifier code that linked them to the person that recruited them and asked to become recruiters themselves. The coupon number given to the recruit by the recruiter was required to complete the questionnaire.

The questionnaire was developed using the survey software Inquisite, a secure web-based tool housed at the Oregon State Health Division, Center of Health Statistics. It was developed to be easy to follow and took respondents no more than 5 minutes to complete.

Section I of the questionnaire gathered basic demographic questions such as: age, race, gender, county of residence and self-disclosed sexual orientation. Additionally, participants were asked to estimate the size of their social network.

Section II of the questionnaire addressed sexual behavior. Questions were designed to assess current sexual risk behavior, primarily MSM. A typical question may read, "In the last 5 years, have you ever been sexually intimate with another man to the point of orgasm?" There were no more than fifteen yes/no questions in this part of the survey. Those who reported no same sex behavior answered considerably less as a series of skips were built in.

In order to track who recruited who without collecting personal identifiers, a web-based tool was developed that collected the respondents recruitment code and then provided the respondent with a recruiter code to be used in subsequent recruitment of peers. Upon receiving their new code they were forwarded to the questionnaire.

The homepage of the online survey informed the participant of the studies general purpose and the nature of the questions. The homepage also served as the consent form for all anonymous participants (wave 1 and beyond); continuing on with the survey implied consent.

The five initial seeds completed the survey and were provided with new recruiter codes. All five seeds reported having attempted to recruit five additional peers, but only two of the seeds successfully recruited additional participants. After a month, the author identified eight additional seeds. None of the seeds recruited more than one peer.

As implemented, we conclude that our web-based RDS approach to estimating the size of the Portland metro area MSM population was infeasible. Though we ultimately recruited 13 enthusiastic seeds, none of these were able to successfully recruit more than 2 additional participants to the website to complete the questionnaire. Anecdotal evidence from the seeds indicated that general apathy and suspicion about web surveys and the lack of a financial incentive were substantial barriers to successful recruiting.

Study aims

RDS has shown promise for researchers as a sampling tool in collecting data from hidden or hard-to-reach populations. We wondered: what do public health researchers need to know about RDS in order to employ it most effectively and efficiently.

This question and our failed attempt to employ an anonymous, web-based model lead to a simulation study of to investigate how the following parameters affect RDS sampling:

1. Transitional probability structures: tightly knit vs. less tightly knit clusters
2. Average network sizes: equal, larger MSM and smaller MSM
3. Seed composition

In each case we evaluated:

a. Number of waves to equilibrium

b. Composition of the final sample

c. Observed proportion of MSM

d. Reciprocity estimate of MSM

Methods

A priori assumptions

The use of RDS to estimate an unbiased prevalence estimator requires researchers to measure (and potentially adjust for): transitional probabilities & social network size. Transitional probabilities are simply the probability that a person of one group will recruit a member of another group. For example, if male participants recruit male peers 60% of the time and females 40% of the time and female participants recruit male peers 30% of the time and females 70% of the time then the transitional probabilities according to gender are: 0.60/0.40/0.30/0.70. This measurement is necessary to establish the equilibrium estimate. In the following models three different transitional probabilities were modeled for the hidden population - MSM. For the sake of simplicity and consistency in the simulations, Non-MSM recruited MSM 5% of the time and Non-MSM 95% of the time in all scenarios.

The first transitional probability represents what one might expect from a tightly networked hidden population. In this case MSM participants recruited MSM 90% of the time and Non-MSM 10% of the time. This scenario is unlikely in the case of urban MSM males but may be reasonable for other hidden populations who tend to interact primarily with members of their own group (e.g. commercial sex workers or IDU). Assuming the CDC estimate of 5-7%, MSM are 13-18 times more likely to recruit another MSM (rather than Non-MSM) male than if chosen at random.

The second transitional probability represents a moderately networked hidden population. Here, MSM participants recruited MSM 50% of the time and Non-MSM 50% of the time. Using the CDC estimate that MSM prevalence is between 5-7% of the general male population in this scenario MSM are 7-10 times more likely to recruit another MSM (rather than Non-MSM) male than if chosen at random.

The third transitional probability represents a slightly networked hidden population where MSM participants recruited MSM 30% of the time and Non-MSM 70% of the time. Using the same CDC estimate, in this scenario MSM are 4-6 times more likely to recruit another MSM (rather than Non-MSM) male than if chosen at random.

In addition to transitional probabilities, social network size estimates must be collected from participants in order to calculate unbiased prevalence estimates. If groups have unequal social network sizes the group with the larger network size will be oversampled. In order to evaluate the effect that network has on prevalence estimates each of the transitional probabilities were modeled with: (1) equal average social network sizes of 12, (2) larger MSM social networks (16-MSM vs. 12-Non-MSM), and (3) smaller MSM social networks (12-MSM vs. 16-Non-MSM).

As previously discussed, the final sample is unaffected by the original composition of the seeds but the choice of initial seeds can affect the speed and efficiency of attaining equilibrium. To examine how the choice of initial seeds effect equilibrium and final size each of the nine scenarios were modeled with: (1) 5 MSM and 0 Non-MSM seeds, (2) 3 MSM and 2 Non-MSM seeds, (3) 1 MSM and 4 Non-MSM seeds and (4) 0 MSM and 5 Non-MSM seeds.

Programming

All simulations were conducted and analyzed using SAS 9.2. The following steps were taken to construct the defined models:

1. 'Wave 0' was created with the following variables:
    a. Participant ID numbers 1 -5
    b. MSM type defined according to the model (e.g. 5 MSM/0 Non-MSM, 3 MSM/2 Non-MSM, etc.)

c. Network size was generated as a random variable with a normal distribution with a standard deviation of 6 according to the model parameters (e.g. 12 MSM/12 Non-MSM, 16 MSM/12 Non-MSM, 12 MSM/16 Non-MSM)

2. 'Wave 1' was then created by utilizing information from the previous wave. Variables created included:

   a. The number of recruits each participant from 'Wave 0' was randomly generated by instructing SAS 9.2 to assign each participant (ID 1-5) a uniform variable between 0 and 5. This number determined how many people each person successfully recruited.

   b. Each of the recruited participants was issued ID numbers created by multiplying their recruiters ID (defined in the previous wave) by 10 and then adding the number corresponding to their chronological assignment. For example, if participant 2 recruited 4 participants (as defined in step 2b) they would be assigned ID numbers 21, 22, 23 & 24.

   c. Using the MSM status of each participant's recruiter MSM status assigned based on the appropriate transitional probability. For example in the tightly networked models were MSM recruit MSM 90% of the time, if participant 2 in Wave 0 was MSM then in Wave 1 a random number would be generated between 0-1 for each of his recruits. If the number was less than or equal to 0.90 then the recruit would be MSM, if the number was greater than 0.90 the recruit would be Non-MSM.

   d. Network size was generated as a random variable with a normal distribution with standard deviations of 6 according to the model parameters (e.g. 14 MSM/14 Non-MSM, 16 MSM/12 Non-MSM, etc.).

3. 'Wave 2'- 'Wave 10' were then created by utilizing information corresponding previous waves in the same fashion as 'Wave 1".
4. Data set 'All' was compiled using the variables generated in 'Wave 1'-'Wave 10' including:
    a. Wave number
    b. ID
    c. Recruiter ID
    d. Network size
    e. MSM status
    f. Recruiter MSM status
5. Analysis was then conducted on the composite data set 'All'

Data was generated through wave 10 or until the proportion of MSM in the sample was within 2% of the estimated equilibrium for two consecutive waves. Analysis was conducted with SAS 9.2.

Analysis

The aims of the analysis are to:

1. Monitor the proportion of MSM, wave by wave, until the proportion reaches *calculated* equilibrium
2. Compare the calculated equilibrium to the observed equilibrium
3. Compare the observed proportion of MSM to the reciprocity estimate and examine any inconsistencies
4. Evaluate the effect that network size has on the final sample
5. Examine how seed composition affects waves to equilibrium and total sample size

Equilibrium

Observed equilibrium occurs when further recruitment waves do not change the population proportion by a significant amount (+/- 2%). Expected equilibrium is calculated using the proportion of crosscutting ties (S) of both MSM and Non-MSM respondents (see Equation 1). E_m is the proportion of MSM in the sample at equilibrium; S_{nm} is the proportion of crosscutting ties where a Non-MSM recruits an MSM; S_{mn} is the proportion of crosscutting ties where a MSM recruits a Non-MSM

$$E_m = \frac{S_{nm}}{S_{nm} + S_{mn}}$$

Equation 1. Expected Equilibrium

Reciprocity Estimate

Proportional size of the MSM population based on two sources of data: (1) the proportion of cross-cutting ties (S) of MSM and Non-MSM respondents and (2) self-reported personal network size (N) of each group (see Equation 2). P_m is the proportion of MSM in the sample at adjusted for in-group recruitment and average social network size; S_{nm} is the proportion of crosscutting ties where a Non-MSM recruits an MSM; S_{mn} is the proportion of crosscutting ties where a MSM recruits a Non-MSM; N_n is the average social network size of Non-MSM and N_m is the average social network size MSM.

$$P_m = \frac{S_{nm} N_n}{S_{nm} N_n + S_{mn} N_m}$$

Equation 2. Reciprocity Prevalence Estimate

Results

Results are organized into three sections according to the transitional probabilities: (1) tightly networked clusters; (2) moderately networked clusters and (3) loosely networked clusters. Each section includes a table with summary information and graphs that depict wave-by-wave proportions of MSM.

Tables are organized according to social network structure and initial seed composition. Organizing the results in this fashion allows examination of how many waves should be expected, the approximate total sample size, the observed prevalence at the point of equilibrium and the reciprocity prevalence estimate of each scenario (36 in total).

Corresponding graphs are arranged by social network structure to illustrate the effect of varying seed composition on the sample structure.

Tightly networked clusters

In the following scenarios, MSM recruit MSM 90% of the time and Non-MSM the remaining 10% of the time; Non-MSM recruit Non-MSM 95% of the time and MSM, the remaining 5% of the time.

Regardless of how network structure and seed composition is varied, because the clusters are so tightly knit, equilibrium is not reached after 10 waves (Table 1). The sample sizes attained ranged from 179 to 89,929, highly unreasonable for most 'hidden' populations. The observed MSM prevalence at the 10^{th} wave ranged between 22.4% and 50.4% and the

reciprocity MSM prevalence ranged between 27.2% and 40%. Given that the national CDC MSM estimate is between 5% and 7% we can safely conclude that these estimates are highly inaccurate and the simulation parameters are not realistic.

Table 1. Tightly Networked Clusters

Tight Clusters 90/10/95/05					
Seed Composition	Network Structure	Number Waves to equilibrium	Sample Size	Prevalence MSM at equilibrium	Reciprocity estimate MSM
5 MSM 0 Non-MSM	Equal	> 10 (not reached)	65,661	0.478	0.333
3 MSM 2 Non-MSM	Equal	> 10 (not reached)	48,598	0.512	0.333
1 MSM 4 Non-MSM	Equal	4	179	0.358	0.333
0 MSM 5 Non-MSM	Equal	> 10 (not reached)	15,122	0.224	0.333
5 MSM 0 Non-MSM	Larger MSM	> 10 (not reached)	89,934	0.505	0.272
3 MSM 2 Non-MSM	Larger MSM	> 10 (not reached)	88,123	0.469	0.272
1 MSM 4 Non-MSM	Larger MSM	> 10 (not reached)	73,379	0.261	0.272
0 MSM 5 Non-MSM	Larger MSM	8	8,996	0.298	0.272
5 MSM 0 Non-MSM	Smaller MSM	> 10 (not reached)	20,439	0.470	0.400
3 MSM 2 Non-MSM	Smaller MSM	> 10 (not reached)	38,202	0.476	0.400
1 MSM 4 Non-MSM	Smaller MSM	> 10 (not reached)	48,841	0.242	0.400
0 MSM 5 Non-MSM	Smaller MSM	> 10 (not reached)	27,984	0.265	0.400

Figure 1 illustrates the effect of seed composition in a tightly networked cluster where MSM and Non-MSM have equal networks sizes of 12. It is evident from the graph that beginning with a composition of seeds that roughly reflects the expected prevalence reduced the number of waves required to reach equilibrium and increases the likelihood that the observed prevalence will correspond to the reciprocity prevalence estimate.

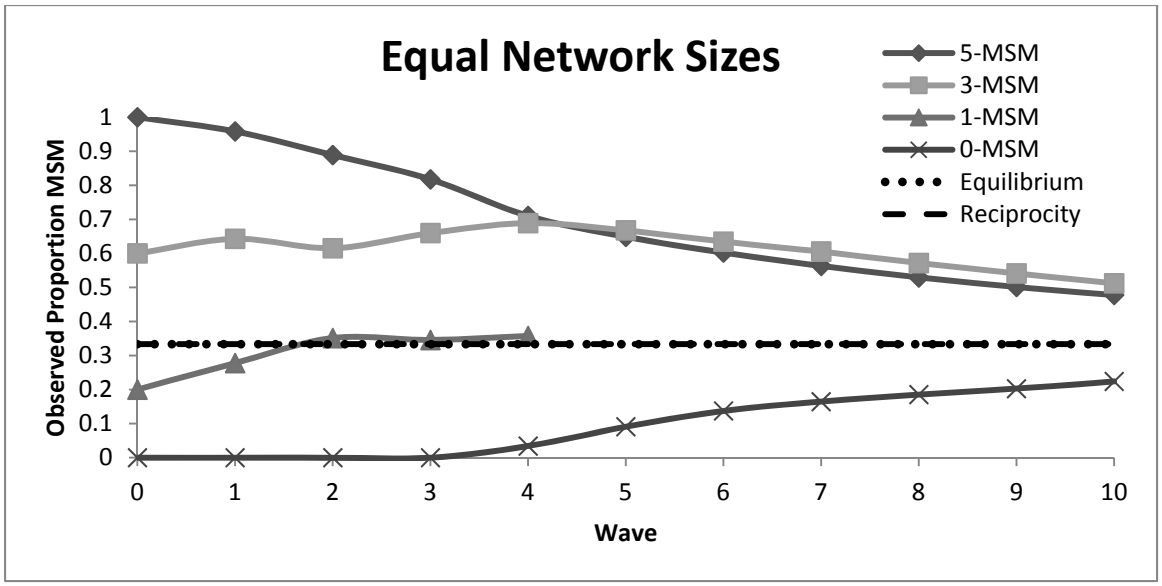

Figure 1. Comparison of seed compositions by wave with equilibrium and reciprocity prevalence estimates in tightly networked clusters with equal social network sizes.

Figure 2 illustrates the effect of seed composition in a tightly networked cluster where MSM have larger social network sizes than Non-MSM (16 & 12 respectively). It is evident from the graph that beginning with a composition of seeds that roughly reflects the expected prevalence reduced the number of waves required to reach equilibrium and increases the likelihood that the observed prevalence will correspond to the reciprocity prevalence estimate.

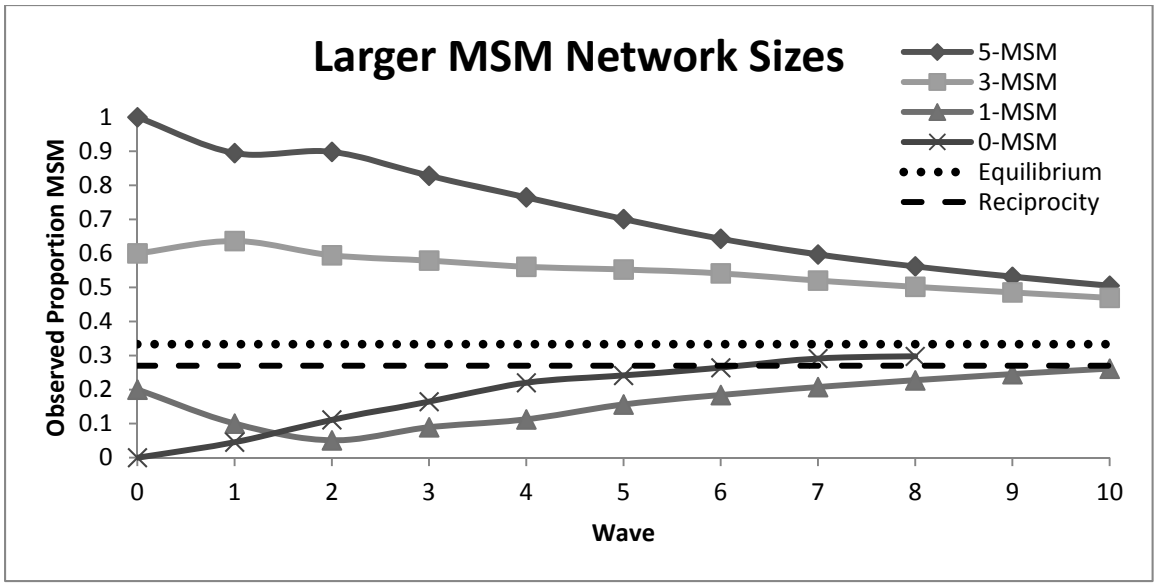

Figure 2. Comparison of seed compositions by wave with equilibrium and reciprocity prevalence estimates in tightly networked clusters with larger MSM social network sizes.

Figure 3 illustrates the effect of seed composition in a tightly networked cluster where MSM have larger social network sizes than Non-MSM (12 & 16 respectively). Although they are approaching, none of the scenarios reached equilibrium or reciprocity after 10 waves.

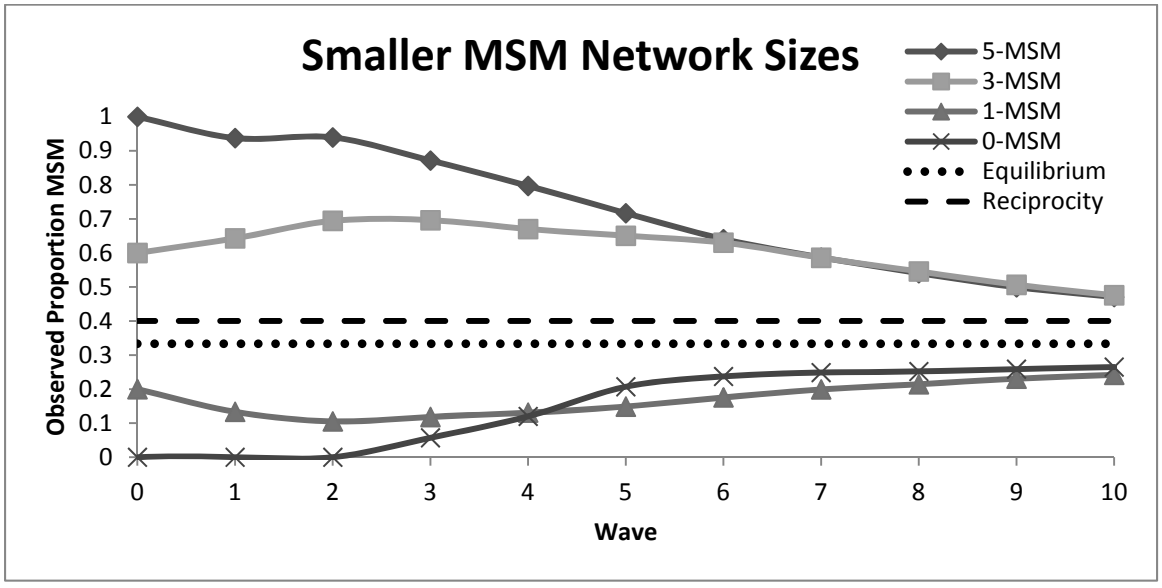

Figure 3. Comparison of seed compositions by wave with equilibrium and reciprocity prevalence estimates in tightly networked clusters with smaller MSM social network sizes.

Moderately networked clusters

In the following scenarios, MSM recruit MSM 50% of the time and Non-MSM the remaining 50% of the time; Non-MSM recruit Non-MSM 95% of the time and MSM the remaining 5% of the time. A scenario like this may be found in a hard-to-reach population that represents a small proportion of the total population and who interact equally with people of both networks.

In the following scenarios (Table 2) all, except one, reach equilibrium within 9 waves, varying according to seed composition and network structure. In those that reached equilibrium, sample sizes attained ranged from 591 to 12,998 and observed MSM prevalence ranged between 7.7% and 9.7% and the reciprocity MSM prevalence ranged between 7.0% and 11.8%.

Table 2. Moderately Tight Clusters 50/50/95/05

Seed Composition	Network Structure	Number Waves to equilibrium	Sample Size	Prevalence MSM at equilibrium	Reciprocity estimate MSM
5 MSM 0 Non-MSM	Equal	8	11,345	0.104	0.091
3 MSM 2 Non-MSM	Equal	7	3,858	0.095	0.091
1 MSM 4 Non-MSM	Equal	10	48,841	0.047	0.091
0 MSM 5 Non-MSM	Equal	7	2,727	0.097	0.091
5 MSM 0 Non-MSM	Larger MSM	7	4,533	0.097	0.070
3 MSM 2 Non-MSM	Larger MSM	5	619	0.095	0.070
1 MSM 4 Non-MSM	Larger MSM	5	591	0.095	0.070
0 MSM 5 Non-MSM	Larger MSM	7	5,071	0.077	0.070
5 MSM 0 Non-MSM	Smaller MSM	9	2,891	0.083	0.118
3 MSM 2 Non-MSM	Smaller MSM	6	1,324	0.091	0.118
1 MSM 4 Non-MSM	Smaller MSM	6	664	0.092	0.118
0 MSM 5 Non-MSM	Smaller MSM	9	12,998	0.086	0.118

Figure 4 illustrates that in a moderately networked cluster with equal network sizes, equilibrium is reached within ten waves. Beginning with one or three MSM seeds required seven waves; beginning with five MSM seed required eight waves and beginning with all Non-MSM seeds required ten waves to reach equilibrium. Therefore, it is best to start with a proportion of seeds similar to the expected prevalence, in this case one or three MSM.

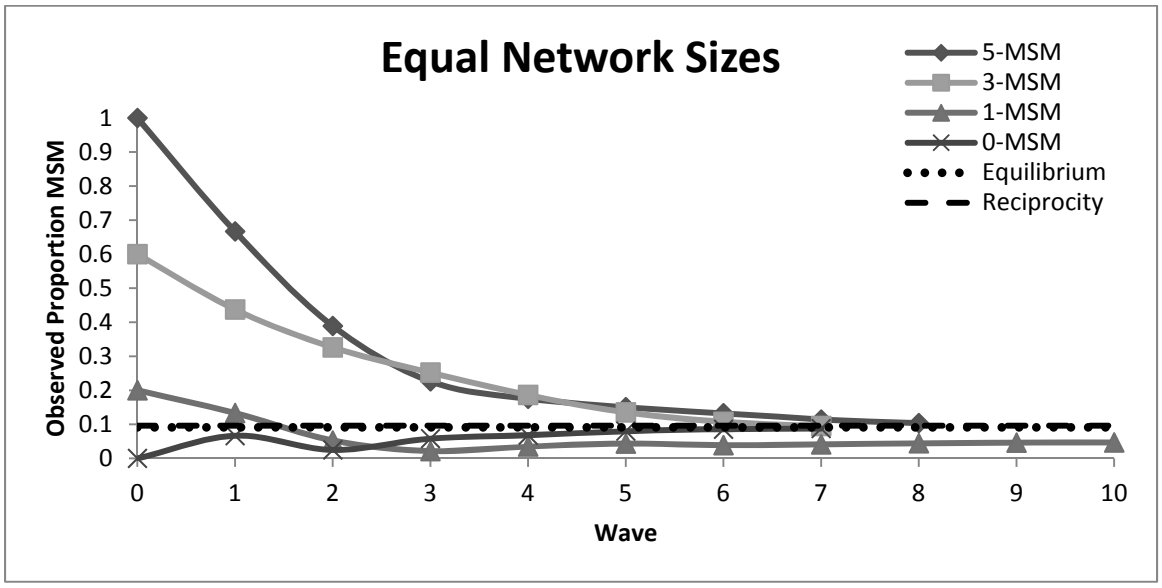

Figure 4. Comparison of seed compositions by wave with equilibrium and reciprocity prevalence estimates in moderately networked clusters with equal social network sizes.

Figure 5 illustrates the effect of seed composition in a moderately networked cluster where MSM have larger social network sizes than Non-MSM (16 & 12 respectively). Equilibrium is attained in all scenarios although, beginning with a mix of MSM and Non-MSM seeds reduced the number of waves and sample size required to reach equilibrium and increases the likelihood that the observed prevalence will correspond to the reciprocity prevalence estimate.

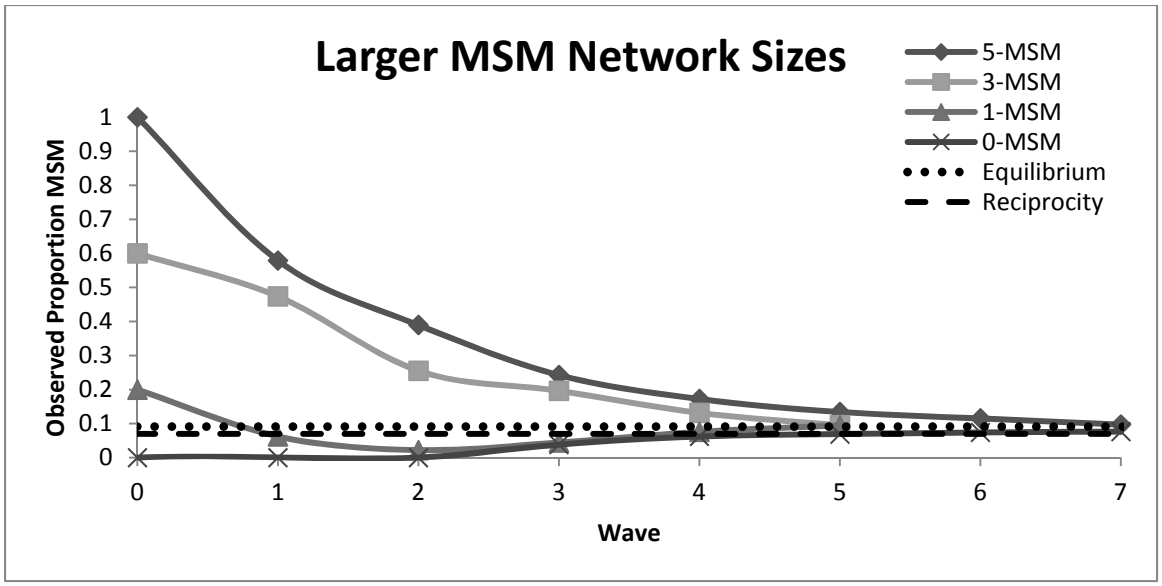

Figure 5. Comparison of seed compositions by wave with equilibrium and reciprocity prevalence estimates in moderately networked clusters with larger MSM social network sizes.

Figure 6 illustrates the effect of seed composition in a moderately networked cluster where MSM have smaller social network sizes than Non-MSM (12 & 16 respectively). Equilibrium is attained in all scenarios although, beginning with a mix of MSM and Non-MSM seeds reduced the number of waves and sample size required to reach equilibrium.

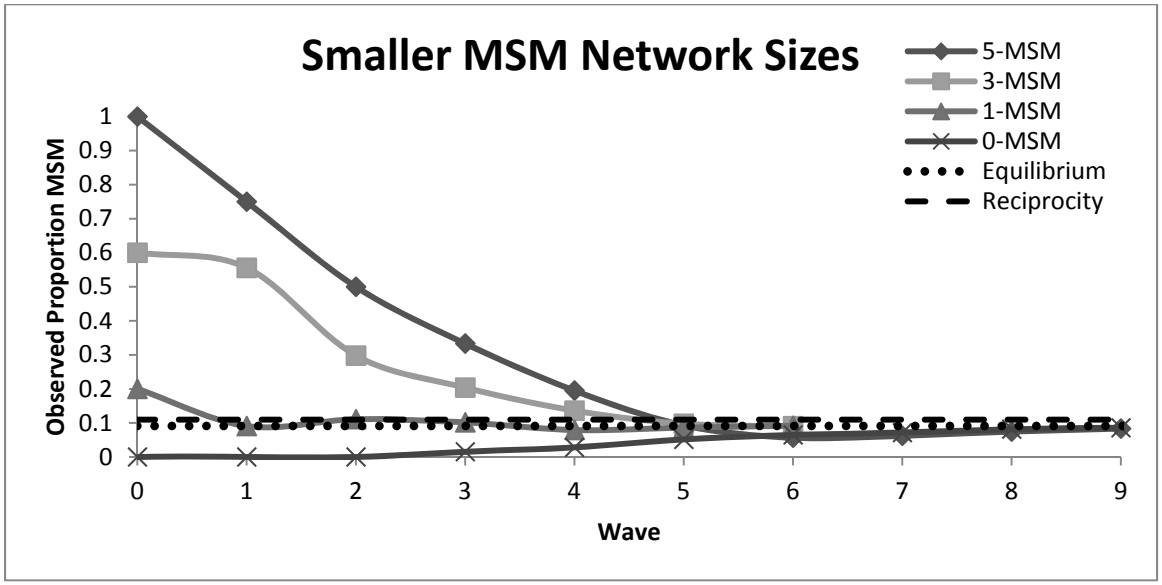

Figure 6. Comparison of seed compositions by wave with equilibrium and reciprocity prevalence estimates in moderately networked clusters with smaller MSM social network sizes.

Loosely networked clusters

In the following scenarios, MSM recruit MSM 30% of the time and Non-MSM the remaining 70% of the time; Non-MSM recruit Non-MSM 95% of the time and MSM the remaining 5% of the time. A scenario like this may be found in a hard-to-reach population that represents a small proportion of the total population and who interact with people of both networks.

In the following scenarios (Table 3) all reach equilibrium within eight waves, varying according to seed composition and network structure. The sample sizes attained ranged from 259 to 2,492. The observed MSM prevalence at the equilibrium ranged between 5.7% and 7.7% and the reciprocity MSM prevalence ranged between 5.1% and 8.7%. This scenario is closest to what we expected to find had our original study been successful.

Table 3. Looser Clusters 30/70/95/05

Seed Composition	Network Structure	Number Waves to equilibrium	Sample Size	Prevalence MSM at equilibrium	Reciprocity estimate MSM
5 MSM 0 Non-MSM	Equal	8	4,694	0.068	0.067
3 MSM 2 Non-MSM	Equal	6	1,632	0.073	0.067
1 MSM 4 Non-MSM	Equal	5	1,526	0.071	0.067
0 MSM 5 Non-MSM	Equal	4	306	0.072	0.067
5 MSM 0 Non-MSM	Larger MSM	7	2,492	0.071	0.051
3 MSM 2 Non-MSM	Larger MSM	7	2,350	0.072	0.051
1 MSM 4 Non-MSM	Larger MSM	4	259	0.062	0.051
0 MSM 5 Non-MSM	Larger MSM	7	2,235	0.057	0.051
5 MSM 0 Non-MSM	Smaller MSM	5	950	0.069	0.087
3 MSM 2 Non-MSM	Smaller MSM	5	740	0.077	0.087
1 MSM 4 Non-MSM	Smaller MSM	4	262	0.074	0.087
0 MSM 5 Non-MSM	Smaller MSM	4	127	0.074	0.087

Figure 7 illustrates that in a loosely networked cluster with equal network sizes, equilibrium is reached within eight waves. Beginning with five MSM seeds required eight; beginning with three MSM seeds required six waves; beginning with one MSM seed required five waves and beginning with all Non-MSM seeds required four waves to reach equilibrium. Therefore, it is best to start with a proportion of seeds similar to the expected prevalence.

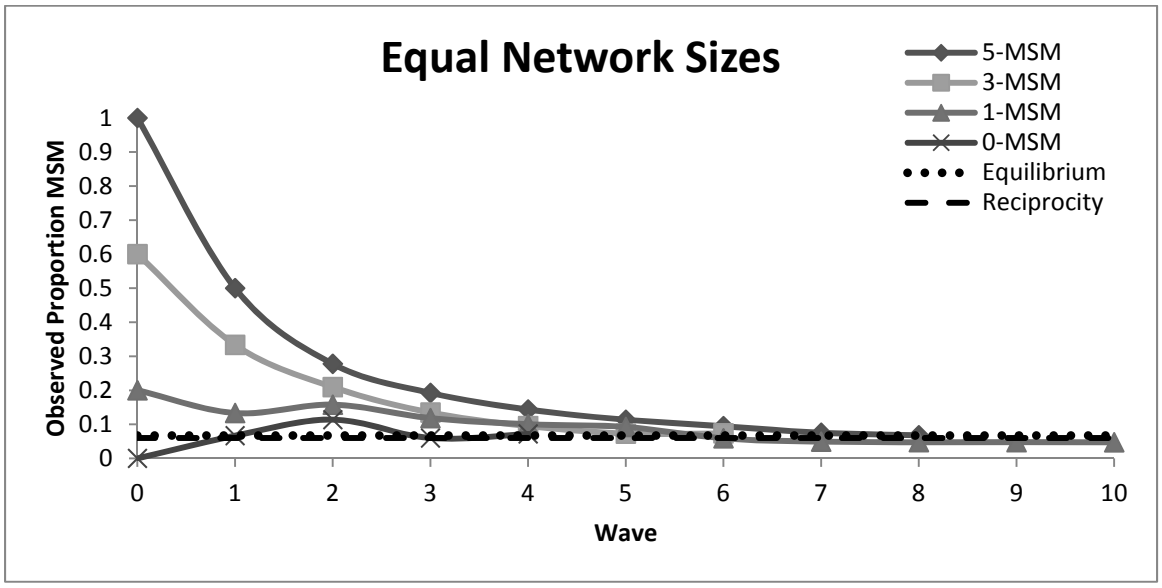

Figure 7. Comparison of seed compositions by wave with equilibrium and reciprocity prevalence estimates in loosely networked clusters with equal social network sizes.

Figure 8 illustrates the effect of seed composition in a loosely networked cluster where MSM have larger social network sizes than Non-MSM (16 & 12 respectively). Equilibrium is attained in all scenarios although beginning with fewer one MSM seed reduced the total sample size.

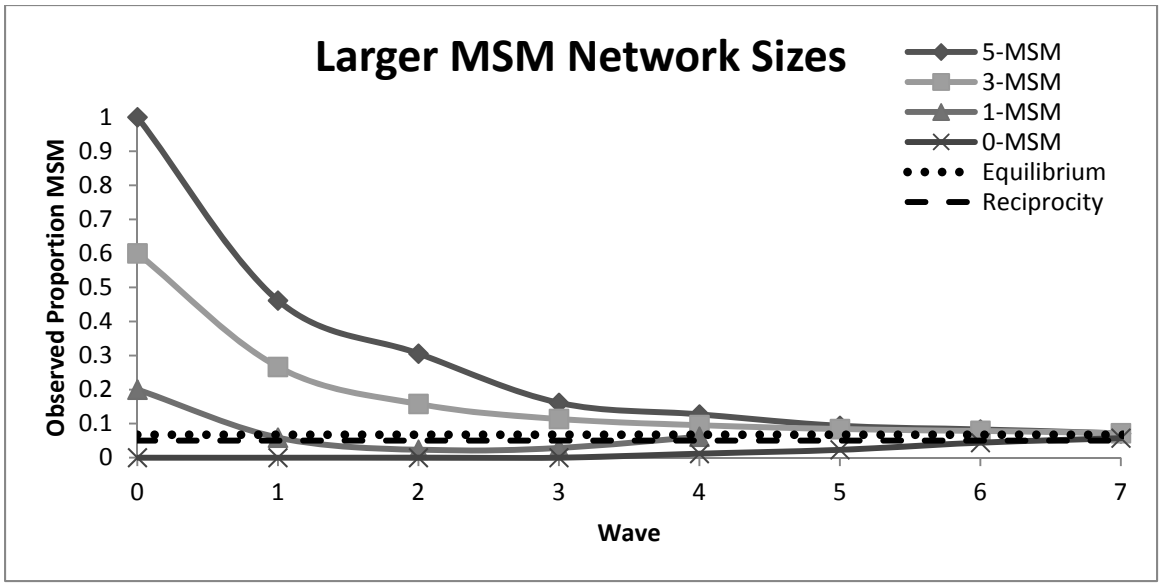

Figure 8. Comparison of seed compositions by wave with equilibrium and reciprocity prevalence estimates in loosely networked clusters with larger MSM social network sizes.

Figure 9 illustrates the effect of seed composition in a tightly networked cluster where MSM have larger social network sizes than Non-MSM (12 & 16 respectively). Equilibrium is attained in all scenarios although, beginning fewer (1 or 0) MSM seeds reduced the number of waves and sample size required to reach equilibrium.

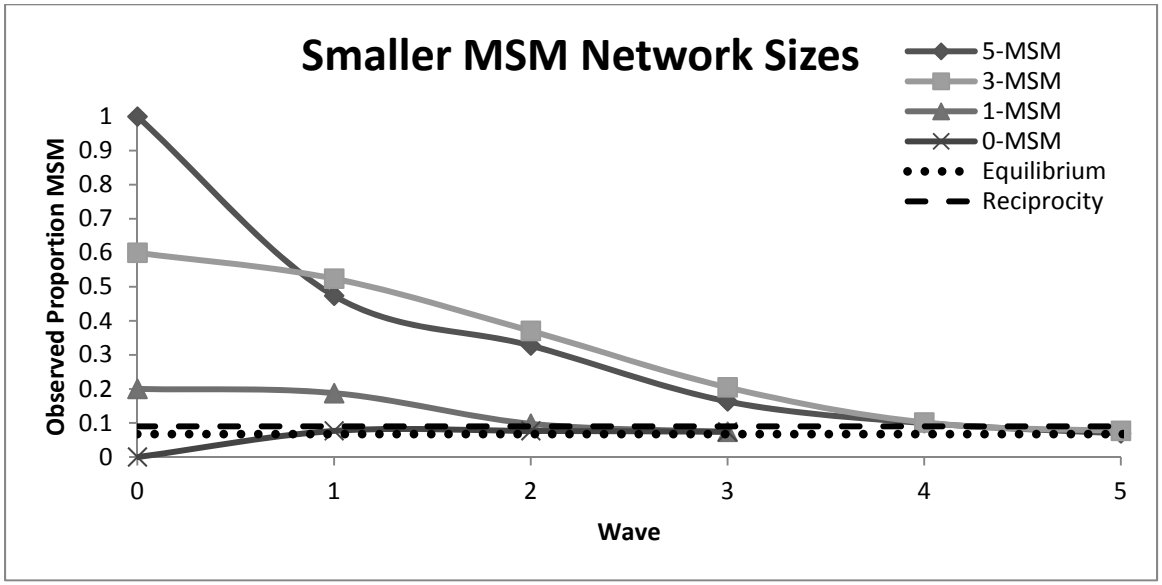

Figure 9. Comparison of seed compositions by wave with equilibrium and reciprocity prevalence estimates in loosely networked clusters with smaller MSM social network sizes.

Saturated Networks

The previous scenarios are premised on the assumption that there is an infinite pool of MSM from which to recruit. In order to address this, a modification was made on two of the previous models (graph 1 & graph 7) to limit the size of the potential MSM sample. At wave 3 the maximum number of recruits that a MSM was allowed decreased from five to two.

The first (figure 10) models saturated MSM networks where, because MSM represent a small proportion and are recruiting almost entirely within group, very quickly the number of new recruits diminishes. The result is observed prevalence below the calculated equilibrium and reciprocity MSM estimate because MSM are recruiting at a diminished capacity.

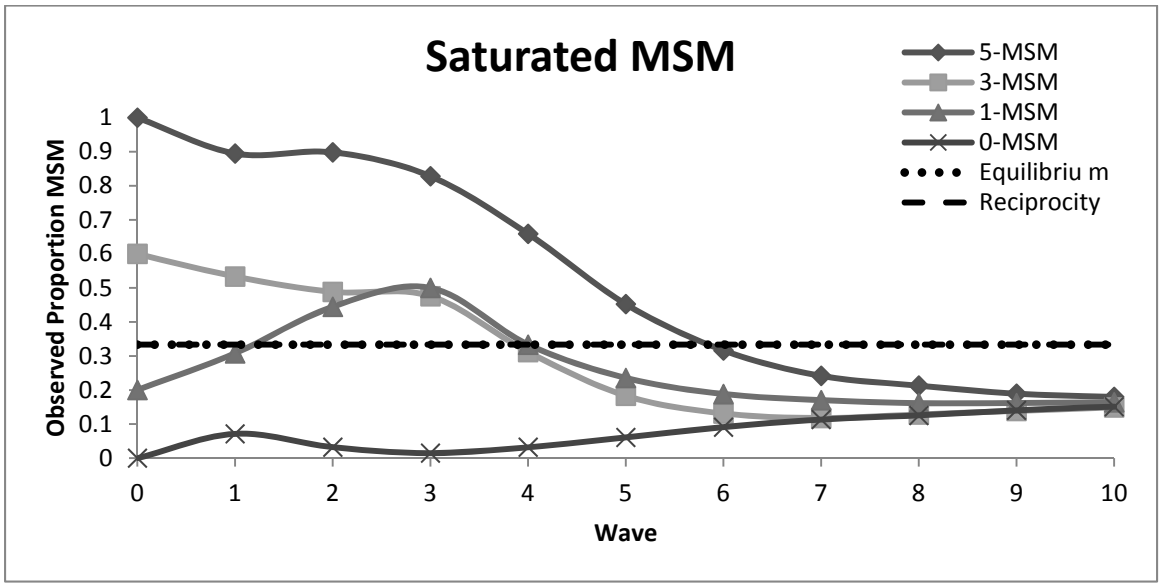

Figure 10. Comparison of seed compositions by wave with equilibrium and reciprocity prevalence estimates in tightly networked clusters with equal social network sizes where the pool of MSM is 'saturated'

The second (figure 11) models a saturated MSM network where, because MSM represent a small proportion and are recruiting almost entirely within group; very quickly the number of new recruits diminishes.

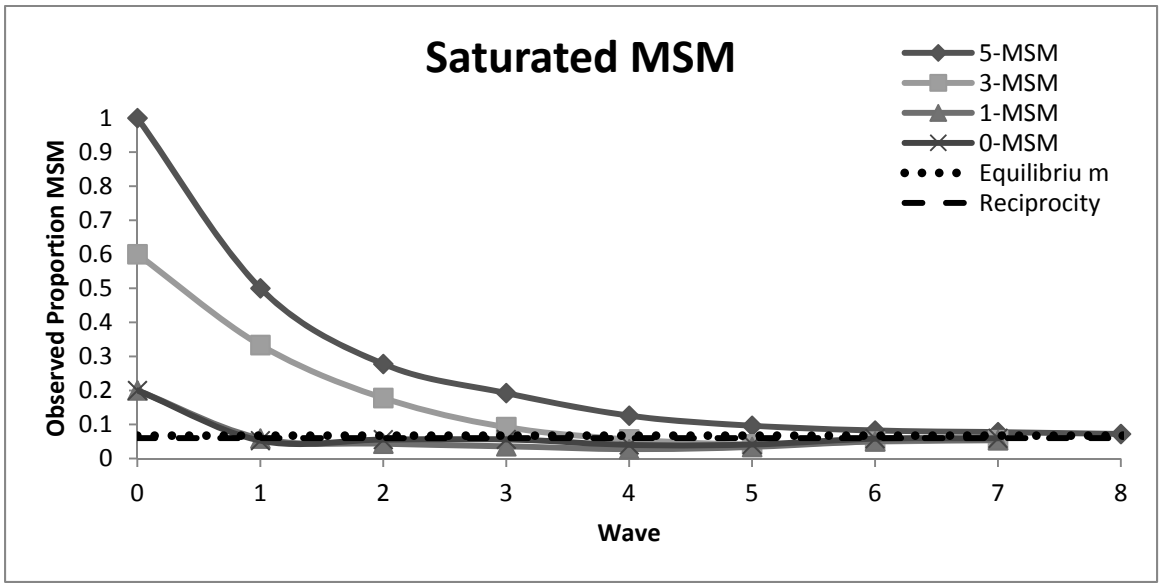

Figure 11. Comparison of seed compositions by wave with equilibrium and reciprocity prevalence estimates in loosely networked clusters with equal social network sizes where the pool of MSM is 'saturated'.

The two saturated models above demonstrate what may happen when MSM find they have fewer available peers to recruit after a few waves because so many of their peers have already been sampled. The transitional probabilities have not changed, just the number of peers that each MSM after wave 3 can recruit.

Another way to examine the idea of saturated networks would have been to change the transitional probabilities, wave by wave. For instance, MSM may *want* to recruit MSM 90% of the time but after a few waves they find less and less available because their peers have already participated in the study. Since their peers have already been sampled they now have to recruit more Non-MSM. This dynamic will also change the number of waves required to equilibrium and the final structure of the sample. In some scenarios, the changing transitional probabilities

may violate the reciprocity assumption central to the Markov process. This is methodological exercise that should be explored in the near future by public health researchers.

Discussion

Study Design

Respondent Driven Sampling has been demonstrated to be the best available method to effectively and efficiently sample hard to reach or hidden populations. RDS is attractive both methodologically because it provides a framework to reach groups that are small relative to the general population, for which no exhaustive list of population members is available and analytically because it is now possible to draw statistically valid samples of previously unreachable groups. RDS survey data is collected in a convenient process in which the researcher is (except for the initial recruits) completely removed. Then a mathematical model of the recruitment process is created to weight the sample and compensate for non-random recruitment patterns. RDS, when used properly, can be an inexpensive and accurate tool in sampling hard to reach populations.

Traditionally, sampling hard-to-reach or hidden populations is difficult due to the stigmatized behavior in which they participate. RDS overcomes most of these challenges but there may remain a tendency to protect the most vulnerable members of the population by not recruiting them into the study. For instance, MSM recruiters may avoid recruiting peers who do not openly identify as MSM in an effort to help mask their behavior. The tendency to mask or protect may be addressed by creating an appropriate incentive structure in which recruiters are given a higher incentive if they recruit a peer from the protected group.

Researchers should be very thoughtful in determining when to employ in an epidemiological study. RDS is a great tool for reaching and estimating hard to reach populations but is not intended for use on populations that already have a well-known sampling frame. This

problem arises when priority is placed on coverage rather than statistical validity. Network-based methods can provide comprehensive coverage of the target population due to the principle of "six degrees of separation," so total coverage is possible, at least theoretically.

However, this approach is prey to a host of biases. Most people recruit those whom they resemble in race, ethnicity, education, income, and religion and well-connected individuals tend to be over-sampled because many recruitment paths lead to them, so the peer recruitment upon which network-based sampling is based is anything but random (Heckathorn 2002).

Although the RDS estimator is asymptotically unbiased variance may be very large there is significant dependence between all samples, increasing the variance of the RDS estimates.. This increased variability of the RDS estimates means that a larger sample is required (approximately 2x as large) to reach the same level of precision compared with a simple random sample. (Salganik and Heckathorn 2004). When a known frame work exists traditional sampling is more appropriate.

Web-based implementation

The appeal of using a web-based design over a person/location based RDS design is that (1) data can be collected much faster because participants do not have to travel to a survey site; (2) less man hours are required for data collection and entry; (3) data can be collected anonymously and (4) the design of the survey can be adaptive to the person's response resulting in better data. Web-based implementation at least in theory provides many advantages over traditional RDS but little evidence yet exists to how well the web-based design works in practice.

Further exploration is needed regarding the web-based RDS format. Mathematically, person-based and web-based analyses are identical. Qualitatively issues such as participant

motivation, sampling without replacement, confidentiality concerns and incentive structure must be further addressed.

Developing a web-based tool that is confidential and easy to use is vital to promote participation. In our failed attempt, feedback suggested general apathy and lack of interest in completion. This may have been avoided with financial incentive but providing incentive would have negated anonymity – central to our design. Researchers employing an anonymous web-based RDS survey in the future may be able to avoid this problem by providing electronic gift certificates to be used at online stores. If the choice is made to use electronic incentive participants should be assured that the use of the gift certificate is in no way recorded and cannot be linked to their survey response.

In RDS, it is imperative that researchers track the recruiter-recruit network in order to measure and adjust for sampling bias. Using the Internet as a sampling tool makes it much more difficult for researchers to verify who people are. If a participant is motivated and able to deceive a sampling tool by misreporting or taking the survey multiple times, it is unlikely that the researcher would be ever become aware.

Providing enough incentive to encourage recruits to participate and then to recruit new members varies from study to study and is largely dependent on the required time and involvement of the participant. As noted, in small tightly networked clusters, providing additional incentive to recruit out-of-group is necessary in order to reach equilibrium. In a web-based format this may motivate participants to misreport personal information for monetary gain.

One of central estimation mathematical assumptions of RDS is that sampling is with replacement; although in design researchers plan for sampling without replacement (we do not

want people to respond more than once). This quandary is justified by proving that the limit of difference between the two sampling methods is zero. Following this same logic one can assume that any potential deception factor will have a limited effect on RDS estimates. Furthermore, it is reasonable to assume that this potential bias is non-differential.

Upon determining that RDS is an appropriate survey tool the choice of whether to use a traditional person-based or an unconventional web-based design is a matter of methodological preference. The types of considerations a researcher should take in making this decision are: (1) geographically dispersion of the population of interest; (2) internet usage rate; (3) feasibility of the incentive structure/delivery; (4) whether anonymity important; (5) how recruiter-recruit waves will be mapped; (6) how to ensure valid collection of RDS variables (e.g. social network size).

Recommendations

1. Collect pilot data in order to estimate:
    a. Transitional probabilities
    b. Average network sizes
    c. Expected proportions
2. If pilot data demonstrates that the hidden population is tightly networked they should be encouraged to recruit from outside of their group.
3. Thoughtfully design incentive structure. The incentives could be doubled for cross-group recruiting. This will decrease the density of the clusters and the time to which equilibrium is attained.
4. Carefully format instructions so that they are easy to follow and leave no room for miss-interpretation or error.

References

1. CDC. *HIV/AIDS Surveillance Report, 2004* (Vol. 16). Atlanta: US Department of Health and Human Services, CDC; 2005:1–46

2. CDC. Subpopulations estimates from the HIV/AIDS surveillance report, United States 2006 MMWR: September 12, 2008 / 57(36);985-989

3. Centers for Disease Control and Prevention (CDC). *Behavioral Risk Factor Surveillance System Survey Data*. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2006.

4. Goldbaum G., Perdue T.,Wolitski R., Rietmeijer C., Hedrich A., Wood R., Fishbein M., (1998). Differences in risk behavior and sources of AIDS information among gay, bisexual, and straight-identified men who have sex with men. *AIDS and Behavior* Vol 2(1), Mar 1998. pp. 13-21

5. Heckathorn D.D. (1997). Respondent-Driven Sampling: A new approach to the Study of Hidden Populations. *Social Problems*, 44(2), 174-99.

6. Heckathorn D.D. (2002). Respondent driven sampling II: Deriving valid population estimates from chain referral samples of hidden populations. *Social Problems*, 49(1), 11–34.

7. Heckathorn D.D & Jeffri J. Finding the beat: using Respondent-Driven Sampling to study jazz musicians. *Poetics.* 2001; 28: 307-329.

8. Johnston L., Khanam R., Reza M., Khan S., Banu S., Alam S., Rahman M., & Azim T. (2007). The Effectiveness of Respondent Driven Sampling for Recruiting Males who have Sex with Males in Dahka, Bangladesh. *AIDS Behav* (2008) 12:294-304

9. Magnani R., Sabin K., Saidel T., & Heckathorn D. (2005). Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS* 2005, 19 (suppl 2): S67-S72

10. McKnight C., Jarlais D., Bramson H., Tower L., Abdul-Quadar A., Nemeth C., & Heckathorn D. (2006). Respodent-Driven Sampling in a Study of Drug Users in New York City: Notes from the Field. *Journal of Urban Health* 2006, Vol 83, No. 7: i54-i58

11. Miller, Serner M, Wagner M, (2005) Sexual diversity among black men who have sex with men in an inner-city community. *Journal Of Urban Health* 2005 Mar; Vol. 82 (1 Suppl 1), pp. i26-34

12. Millett GA, Flores SA, Peterson JL, Bakeman R, (2007) Explaining disparities in HIV infection among black and white men who have sex with men: a meta-analysis of HIV risk behaviors. *AIDS* 2007 Oct 1; Vol. 21 (15), pp. 2083-91

13. Salganick M. (2006). Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling. *Journal of Urban Health* 2006, Vol. 83, No. 7: i98-i112

14. Salganik M.J. & Heckathorn D.D. (2003). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Presented at various conferences in 2003.*

15. Wejnert C. & Heckathorn D. (2008) Web-based network sampling: efficiency and efficacy of respondent-driven sampling for online research. (forth-coming)

16. Wood, R; Wolitski, R; Rietmeijer, C; Perdue, T; Hedrich, A; Goldbaum, G; Fishbein, M.(1998) Differences in risk behavior & sources of AIDS information among gay, bisexual, & straight-identified men who have sex with men. *AIDS & Behavior*, 1 (March), 1998, Vol. 2, p 13-21, 9p; (*AN SDS-011058*)

17. Yeka, W., Maibani-Michie, G., Prybylski, D., & Colby, D. (2006). Application of respondent driven
sampling to collect baseline data on FSWs and MSM for HIV risk reduction intervention in two urban centers in Papua New Guinea. *Journal of Urban Health*, 83(Suppl. 7), 1–5.

18. Binson D, Michaels S, Stall R, et al. Prevalence and social distribution of men who have sex with men: United States and its urban centers. *Journal of Sex Research* 1995; 32:245–254.