Supporting Information:
Diagnostics for Respondent-driven Sampling

Krista J. Gile
University of Massachusetts, Amherst, MA, USA

Lisa G. Johnston
Tulane University, New Orleans, LA, USA
and University of California, San Francisco, San Francisco, CA, USA

Matthew J. Salganik
Microsoft Research, New York, NY USA
and Princeton University, Princeton, NJ, USA
**Authorship alphabetical; all authors contributed equally to the paper.*

S1. With-replacement Sampling

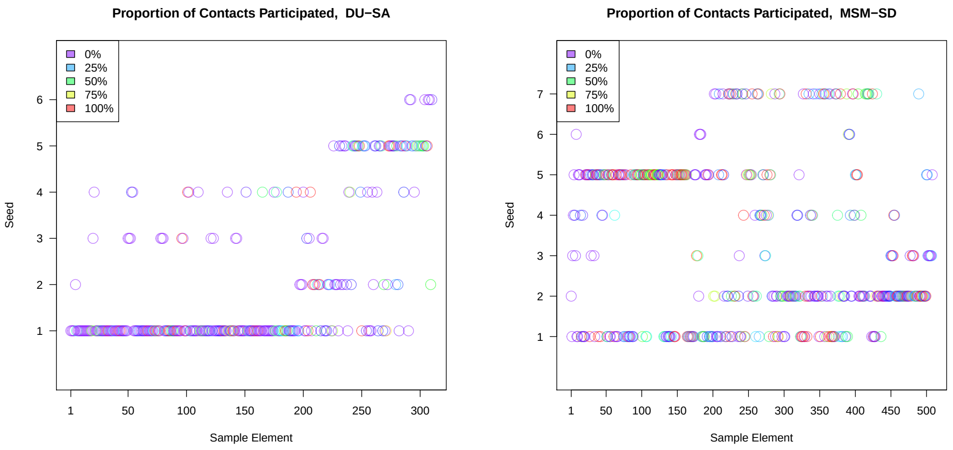
S1.1. Multiple Connections to Survey Participants

In Section 4.3 we presented results about the proportion of respondents' contacts who had already participated in the study. It may also be of interest to visualize these trends. Figs. S1(a) and S1(b) show the reported proportions that already participated for each respondent, by seed, over time. In Fig. S1(a), we can see that within seed, particularly seed 1, periods of low proportion already sampled are often followed by periods of higher proportion already sampled. This may be indicative of the exhaustion of local subgroups. Fig. S1(b) shows less evidence of a positive trend in proportion already sampled over time. Finally, Fig. S2 shows the fitted linear trends for all 12 sites.

S1.2. Decreasing Degree over Time in Sample

Under a broad range of assumptions, link-tracing samples result in higher draw-wise sampling probabilities for people with higher degrees (Gile, 2011). Thus, as the sample begins to deplete the target population, we would expect higher-degree nodes to be sampled earlier, followed by lower-degree nodes, suggesting that a decreasing trend in degree over time could be an indication of finite population effects on sampling. We compared several options for evaluating the trend of degree over time. We used time-order in the study to measure time in these analyses, although results were robust to using survey date. These approaches grouped roughly into two families: those sensitive to a small number of outliers (linear regression, Poisson regression), and those robust to a small number of outliers (regression on log degree, robust regression approaches such as least trimmed squares, M regression, and median regression, as well as rank-based Kendall's Tau and Spearman's Rho). Approaches within each family tended to produce similar results. Because of the unknown dependence in the data structure, we considered only the sign of the coefficient of time in each model. Surprisingly, we find little evidence of decreasing degree over time

Address for correspondence: Krista J. Gile, Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003-9305, U.S.A.
E-mail: gile@math.umass.edu



(a) (b)

Fig. S1. Proportion of alters already participated, by seed.

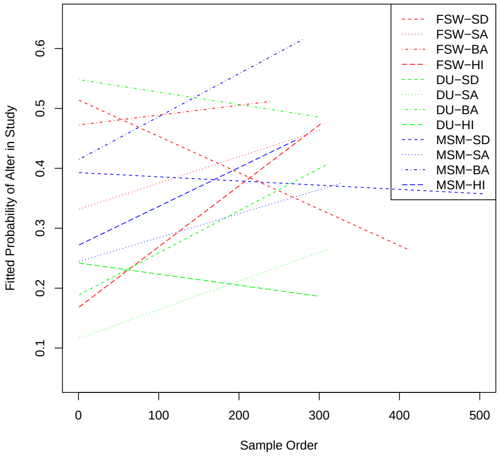
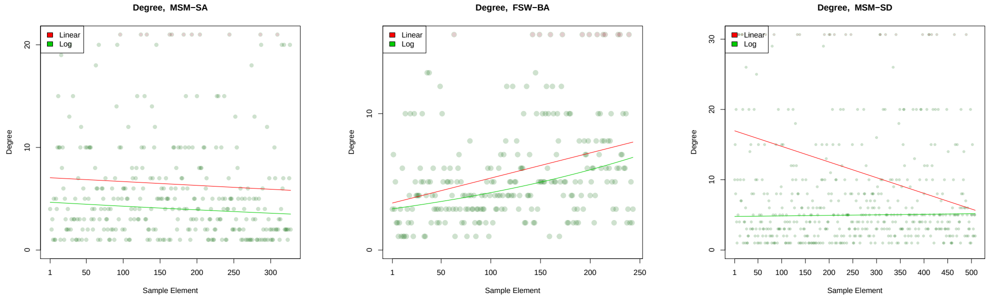


Fig. S2. Fitted linear trends for 12 sites for the proportion of respondents' contacts who had already participated in the study.

with either the non-robust (5 of 12 flagged for linear slope with the linear model) or robust methods (1-3 of 12 flagged).

Fig. S3 illustrates the fitted linear relationship between degree and sample order, as well as the linear relationship fitted to log degree for three sites. In Fig. S3(a) (MSM-SA), both approaches found a negative relationship between degree and sample order; in Fig. S3(b) (FSW-BA) both approaches found a positive relationship; and in Fig. S3(c) (MSM-SD) the two approaches found differing trends, likely driven by the few high responses early in the sample.

Because we have more confidence in the more robust methods, we conclude that this



(a) (b) (c)

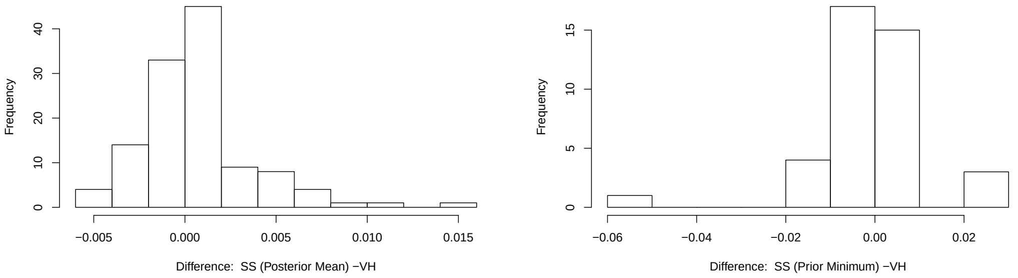
Fig. S3. Degree of respondents over time, with fitted linear model and linear model for log of degree. For visualization, the highest responses were truncated and represented in red at the tops of the plots.

indicator clearly suggests finite population effects for MSM-SA (flagged by all indicators) and perhaps MSM-SD (flagged by most robust indicators). It is surprising to us that all the other populations, including the three known to have not reached their target sample size (FSW-BA, MSM-BA, MSM-HI), suggested positive or null trends in sample degree over time. Because we have strong theoretical reasons (Gile, 2011) to expect negative trends in these cases, we hope future research, with other data sets, will help explain this phenomenon.

S1.3. Successive Sampling Estimation of Finite Population Bias

If researchers have an estimate of the size of the target population, they can compare the SS estimator (Gile, 2011) to the VH estimator (Volz and Heckathorn, 2008) in order to assess finite population effects on estimates. As is typically the case, however, there were no existing estimates of the sizes of our target populations. Therefore, we use the RDS data itself in order to estimate the sizes of our target populations using the approach introduced in Handcock et al. (2012) and implemented in the `R` (R Core Team, 2012) package `size` (Handcock, 2011).

The method of Handcock et al. (2012) requires specifying a prior distribution for the size of the population. To specify the prior distribution for populations of MSM, we drew on a meta-analysis of Caceres et al. (2006), which provides broad bounds on the proportion of men who have had sex with another man in the past year. The estimate for the Dominican Republic (and all of Latin America) is 1-8% of the sexually active adult male population, which we assume to constitute 15-64 year olds. Combining this information with information on the number of males between 15-64 in each city from the Dominican Republic's National Statistical Office (Oficina Nacional de Estadistica, 2009), we created a conservative upper and lower bound for the size of the MSM population in each city. These bounds are then used to define the lower and upper quartiles of a prior distribution. For DU and FSW, no comparable meta-analyses existed so we used broad ranges, consisting of 1-10% of the 15-64 year old total population (DU), or population of women (FSW). As with the MSM, we used these ranges, combined with information from the Dominican Republic's National Statistical Office (Oficina Nacional de Estadistica, 2009) to create prior distributions. When setting the priors in this manner, the method of Handcock et al. (2012) results in posterior mean MSM population size estimates within the original range for SD, SA, and HI, and just above the higher end of the range in Barahona. For DU and FSW, this procedure produced



(a) Posterior Mean

(b) Prior Minimum (MSM only)

Fig. S4. Histograms of the differences between Volz-Heckathorn and Successive Sampling estimates over many traits of interest, with SS estimate based on (a) the posterior mean and (b) the prior minimum (MSM populations only).

6 estimates consistent with the ranges specified in the prior, one (FSW in BA) higher than the 10% number, and two (DU in SD and SA) lower than 1%.

When using the SS estimator, therefore, we used three plausible low population sizes:

- The posterior mean (best point estimate from the population size estimation)
- The lower bound of the posterior highest probability density (HPD) region (lowest plausible estimate from the population size estimation)
- For MSM populations, 1% of the 15-64 year old male population (lower bound of the plausible region from the meta-analysis of Caceres et al. (2006)).

Using each of these estimates of population size, we estimate prevalence of each of the characteristics described in Section S2 using the SS estimator, as well as using the VH estimator. A histogram of the differences based on the lower bound of the HPD region is given in the main text (Fig. 4). Corresponding plots for the other two population size estimates are given in Figure S4. For an single site, a plot like those in Figure S5 may be more useful. These *Population Size Sensitivity Plots* summarize the differences across several population size estimates for the traits of interest in an individual study. For completeness, all items with difference greater than .01 are summarized in Table S1.

S2. Seed dependence

We recommend visual inspection of Convergence Plots and Bottleneck Plots, but in cases where there are many study sites and many traits of interest, it may be difficult to monitor all of these plots. Therefore, we develop a set of procedures that enable researchers to automatically flag plots for further inspection.

For the Convergence Plots, further inspection is called for if the estimates seem to be changing at the end of the sample. That is, a trait should be flagged if

$$\text{there exists} \quad t < \tau \text{ such that } \mid \hat{p}_{(n-t)} - \hat{p}_{(n)} \mid > \ \epsilon \tag{1}$$

where τ is a parameter that sets how much of the trace will be examined and ϵ represents the maximum allowable difference between the estimate at time t and the final estimate. We suspect that the desired values of τ and ϵ will vary from study to study, but in this case we set $\tau = 50$ and $\epsilon = 0.02$. In other words, we ask whether there are any of the final 50 estimates that have a difference of more than 0.02 from the final estimate. We run

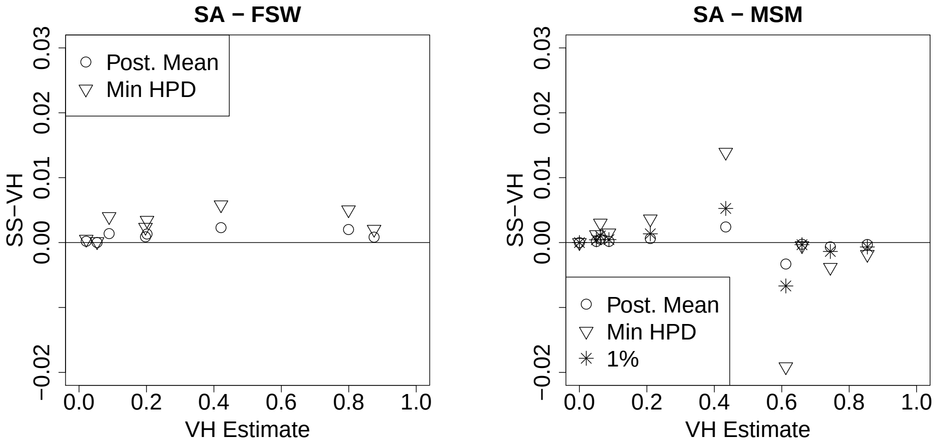


Fig. S5. *Population Size Sensitivity Plots* comparing Volz-Heckathorn estimates and Successive Sampling estimates of prevalence of many traits of interest in two target populations.

Table S1. Prevalence estimates based on Successive Sampling and Volz-Heckathorn estimators for each trait with maximum absolute difference greater than .01.

		Trait	VH	Max HPD	Post. Mean	Min HPD	1%
FSW	HI	Last Client Brothel	0.306	0.316	0.321	0.334	-
FSW	HI	Been In Program	0.345	0.349	0.351	0.356	-
DU	SD	Main Drug Crack	0.263	0.267	0.270	0.275	-
DU	SD	Use Drugs Every Day	0.378	0.385	0.388	0.397	-
DU	SA	Use Drugs Every Day	0.360	0.364	0.367	0.374	-
DU	SA	Been Imprisoned	0.370	0.374	0.376	0.382	-
DU	BA	Use Drugs Every Day	0.391	0.397	0.400	0.410	-
DU	HI	Main Drug Cocaine	0.422	0.418	0.416	0.410	-
DU	HI	Use Drugs Every Day	0.406	0.410	0.413	0.419	-
DU	HI	Been Imprisoned	0.259	0.263	0.265	0.271	-
MSM	SA	Had HIV Test	0.434	0.434	0.436	0.448	0.439
MSM	SA	Bisexual	0.612	0.612	0.609	0.593	0.605
MSM	BA	HIV+	0.087	0.087	0.086	0.085	0.074
MSM	BA	Had HIV Test	0.331	0.330	0.328	0.321	0.277
MSM	BA	Working	0.711	0.712	0.712	0.716	0.735
MSM	BA	Use Drugs	0.607	0.608	0.609	0.613	0.633
MSM	BA	Sex With Woman	0.858	0.859	0.859	0.863	0.884
MSM	HI	Had HIV Test	0.503	0.503	0.501	0.493	0.491
MSM	HI	Used Condom	0.790	0.790	0.787	0.776	0.773
MSM	HI	Sex With Woman	0.834	0.834	0.833	0.825	0.823

this procedure on 120 group × trait × city combinations shown in Fig. S6, and we find the most convergence flags in MSM data: 37.5% of traits were flagged, as compared with 25% of traits for DU and 22% for FSW. Increasing ϵ to 0.05 results in flagging only two traits, both in MSM populations: Bisexual in Santiago and Use Drugs in Higuey. The convergence problems that we detected could be caused by the network structure in the population, the method of seed selection, and the interaction between the two.

For the Bottleneck Plots, further inspection is called for if the estimates from each seed

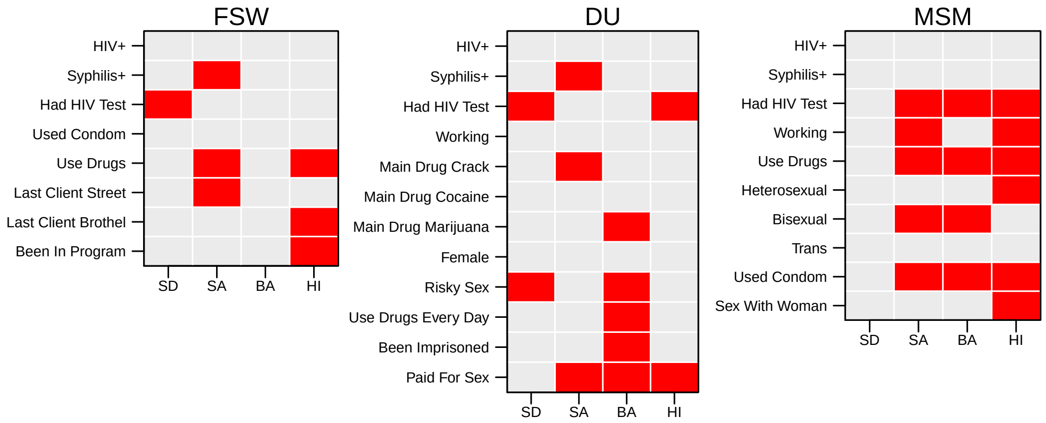


Fig. S6. Convergence test results for $\tau = 50$ and $\epsilon = 0.02$. Red cells represent traits flagged for possible lack of convergence.

deviate from the overall estimate. We formalize that intuition by calculating a weighted squared deviation:

$$WSD = \sum_s n_s \cdot (\widehat{p}_s - \widehat{p})^2 \qquad (2)$$

where $\widehat{p}_s$ is the estimate using sample only from the tree originating at seed s and n_s is the corresponding sample size (not including the seed itself and not including cases with missing data on the trait of interest or degree). In order to assess whether this statistic is unusual, we perform a permutation procedure where the structure of trees are preserved (including weights) while the individual traits are permuted. We then calculate the WSD for the permuted data, and we repeat this procedure 10,000 times. We flag a trait for further investigation if the observed WSD is greater than 90% of the permuted WSD values; this threshold can be adjusted for desired sensitivity.

We ran this procedure on the same 120 group × trait × city combinations examined previously and found that the rates of flagging were highest among FSW (41%) followed by MSM (30%) and then DU (23%) (Fig. S7). Although no trait was flagged in all four cities, these results suggest that likely sources of bottlenecks for FSW are based on sources of clients (e.g., brothel vs. street), drug use, and disease status (HIV and Syphilis); for DU based on type of drug used (Marijuana), employment status, and gender; and for MSM based on self-identification (e.g., bi-sexual and transsexual). These results also suggest that bottlenecks can occur across traits that are not visible to respondents (e.g., disease status) possibly because these traits are correlated with other traits (e.g., age or risky behavior) that do affect social tie formation. Finally, it is important to note that some target populations (e.g., MSM in Santo Domingo) appear to have bottlenecks along many traits.

We wish to emphasize that our flagging procedure for Bottleneck Plots may not always match the intuition of experienced RDS researchers. While it does correctly flag obvious cases of bottlenecks (Fig. 8(a)) and it does not flag in cases where there do not appear to be bottlenecks (Fig. 8(b)), there are a small proportion of edge cases where our flagging procedure did not match our expectations. For example, our procedure flags HIV status for MSM in Barahona (Fig. 8(c)) although this is caused by two chains of length 1, and is therefore probably not cause for concern. On the other hand, our procedure does not flag

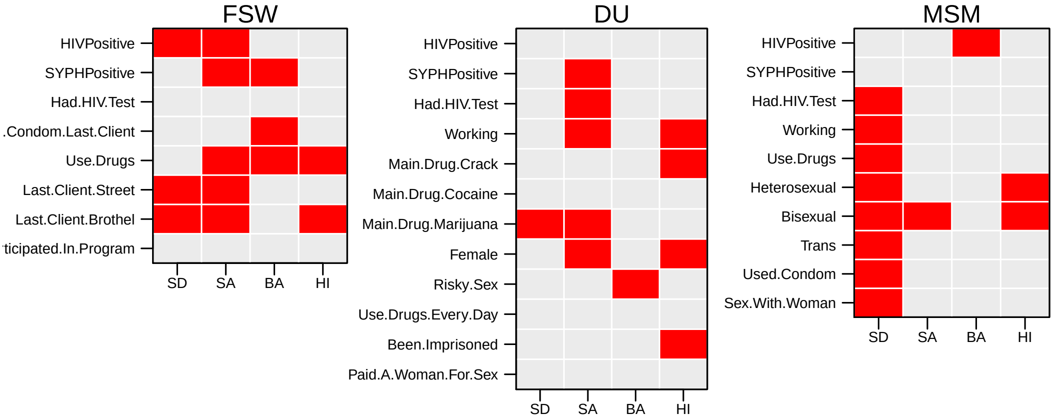


Fig. S7. Bottleneck test results. Red cells represent traits flagged for possible bottlenecks.

HIV status for MSM in Santiago (Fig. 8(d)) even though a review of the plot seems to call for further investigation into the difference between the long chain with approximately 15% estimated prevalence to the other chains with close to zero estimated prevalence. Thus, while we find this procedure a useful heuristic, we hope that future researchers will develop a more formal approach.

S3. Reciprocation

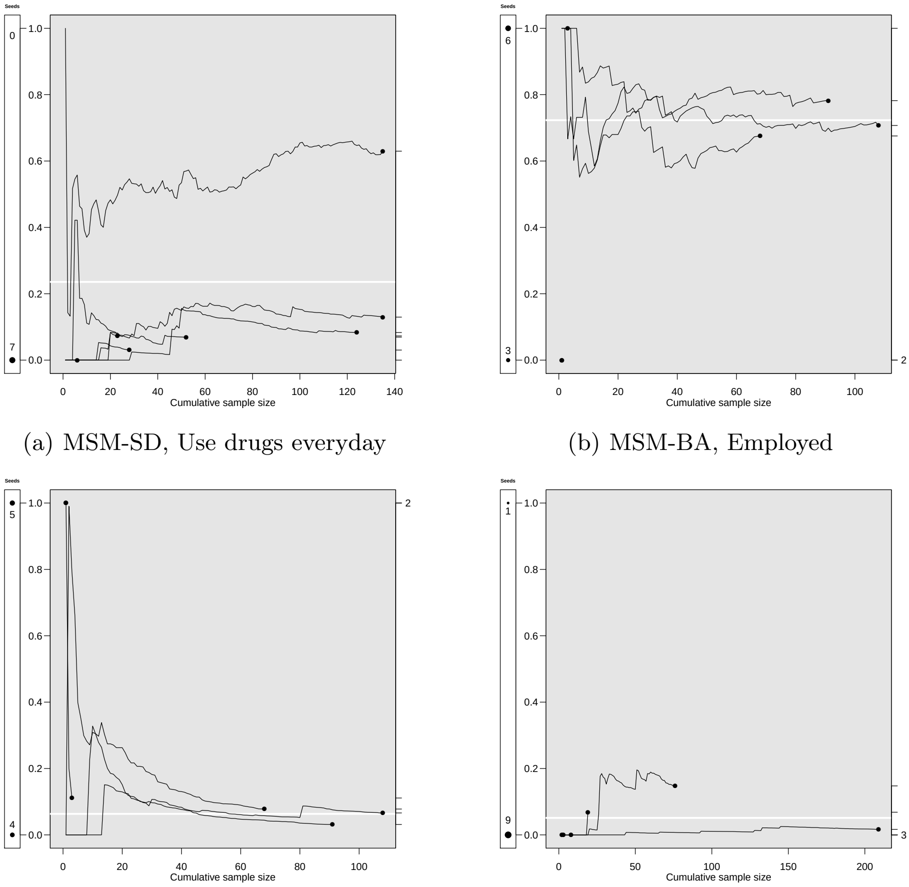
In this section, we introduce a measure of reciprocation of all network ties, rather than just the ties associated with coupon-passing. Although the recall task associated with reporting these data is more complicated than asking only about the recruiter, it is the reciprocation of all ties, not just those involved in coupon-passing that is necessary for estimating sampling probabilities. This is because the estimation of sampling probabilities in RDS relies on the self-reported number of network connections. If all relationships are reciprocated, then the number of network connections is related to a respondent's sampling probability. Otherwise, it is the respondent's in-degree, or number of incoming relations, that is related to sampling probability. Unfortunately, reporting numbers of incoming relations is very difficult. Current estimators for RDS data therefore require reciprocation for two reasons. First, out-ties are easier to self-report and therefore more often recorded, while in-ties are more directly related to sampling probabilities. If all ties are reciprocated, then self-reported out-degree is the same as in-degree. Furthermore, if all ties are reciprocated, the sampling process more closely approximates a random walk on an undirected graph, a common assumption of estimators used.

During the initial visit, participants were asked the following questions about their alters (MSM versions; other groups were analogous):

(Q) How many of them (repeat the number in F) know you well enough that they could give you a coupon within a week if they had been in this study?

(R) If we were to give you as many coupons as you wanted, how many of them (repeat the number in F) could you give a coupon to?

(S) If we were to give you as many coupons as you wanted, how many of these MSM (repeat the number in R) do you think you could give a coupon to by this time next week?



(a) MSM-SD, Use drugs everyday

(b) MSM-BA, Employed

(c) MSM-BA, HIV+

(d) MSM-SA, HIV+

Fig. S8. *Bottleneck Plots:* The left panel in each plot reports the composition of the seeds and the tick marks on the right axis show the final estimates. If there is more than one tree with the same final estimate, that number is also shown on the right axis (see (c) and (d)).

Among all 3,860 respondents who responded to all of these questions, 29.7% gave the same answer for both questions Q and S, 46.7% reported they could give more coupons than they might receive, and 23.6% reported the opposite. The median difference between responses to these questions was 0 and the mean difference of 1.5 more coupons that could be given out than received. Larger differences are positively associated with larger maximum response to either question. For this reason, we also consider normalized difference values, computed as follows:

$$\frac{|Q - S|}{max(Q, S)}.$$

Using these normalized values, the median difference is still 0, with mean 0.40 and third quartile 0.67. This approach is conceptually closer to the full requirement of the reciprocity assumption, but it is also subject to larger concerns of reporting accuracy. Therefore, we prefer the approaches described in Section 6.

S4. Measurement of Degree

S4.1. Time dynamics

We conducted three analyses to check whether the one week time frame in question G was reasonable (see Sec. 7.1). First, for each respondent, we calculated the proportion of his or her alters (based on question F) that could be reached in a specific time frame (based on questions H and I). Fig. S9 depicts the average proportion of alters reachable in each period, by site, with logically inconsistent results excluded.† With the exception of the DU in Santiago, almost all alters were reachable within seven days. The average rate across sites was 92% reachable within one week. Within one day, the across-site-average percent reachable was 62%.

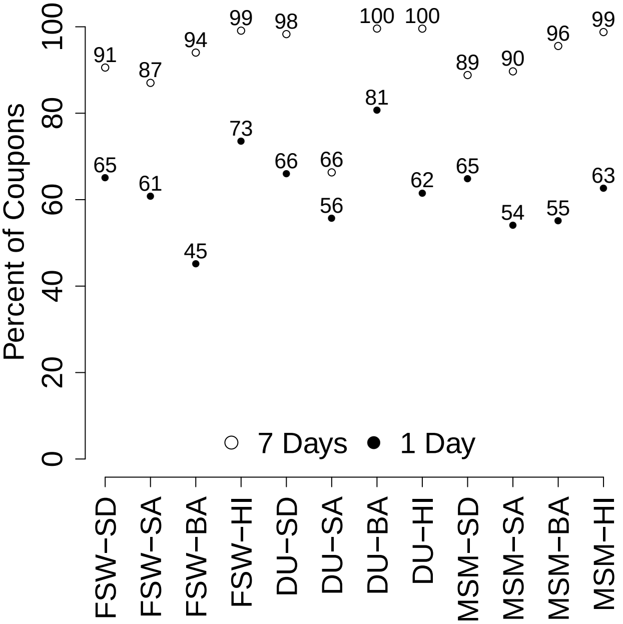


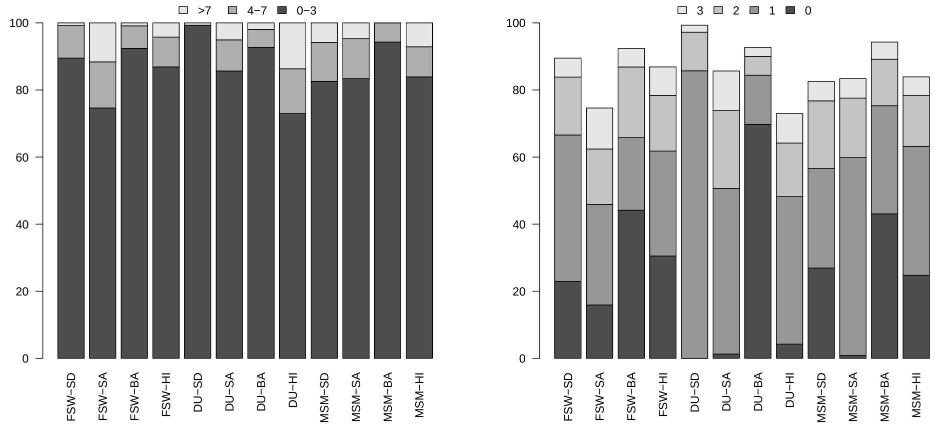
Fig. S9. Proportion of reported contacts that respondents could get a coupon to in 1 or 7 days.

Second, we considered the self-reported number of days each respondent took to distribute his or her coupons (asked at follow-up). Fig. S10 illustrates that across sites, over half (64%) of coupons were distributed in one day and almost all (95%) within seven days.

Finally, we examined the difference between the interview dates for each recruiter-recruit pair, a measure of time dynamics that does not rely on respondent's reports (but which may, unfortunately, be influenced by the capacity for study sites to process interviews during high demand.) Fig. S11 shows that in each site, a substantial majority (79% overall) of interviews occur within a week of the recruiter's interview.

Overall, these three results suggest that restricting social network recall to people a respondent has seen within the last week appears reasonable in this study. Nearly all coupons

†Responses were deemed logically inconsistent and therefore were excluded if a respondent reported being able to reach more contacts than (s)he knew (F). Three sites had high numbers of logically inconsistent responses: FSW-SD (56, 21) (7 days, 1 day), DU-SA (65, 39), MSM-SD (63, 40). A total of 42 responses were inconsistent across the remaining 7 sites.



(a) All days (b) Days 0 - 3

Fig. S10. Percent of coupons distributed by number of days, by site. Most coupons were distributed within 3 days, and nearly all within 7 (a). Among DU in Barahona most coupons were distributed in one day (b).

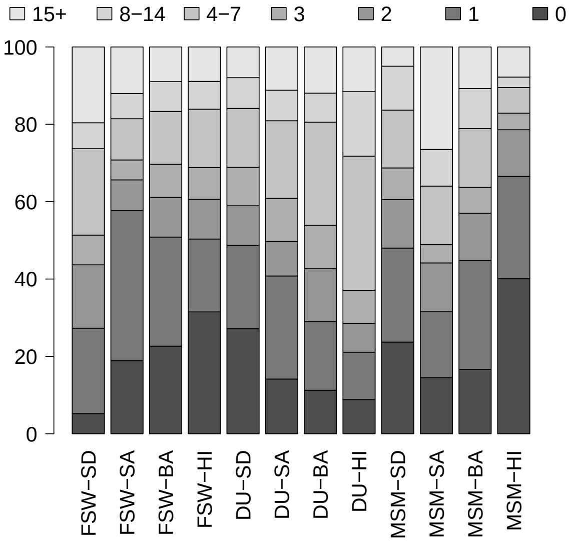


Fig. S11. Distribution of difference between recruiter's interview date and recruit's interview date, by site.

were distributed within a week, and aside from the DU in Santiago, most respondents reported being able to reach nearly all social contacts within a week. Because most coupons were distributed within a shorter period of a few days, it might even make sense to further restrict the recall period to two or three days. Note that the validity of this measure, however, relies on the assumption that coupons were distributed to people incidentally encountered, rather than sought out. Further study is necessary to determine whether respondents seek

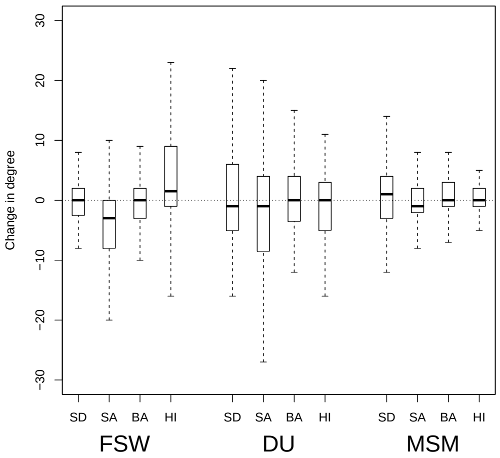
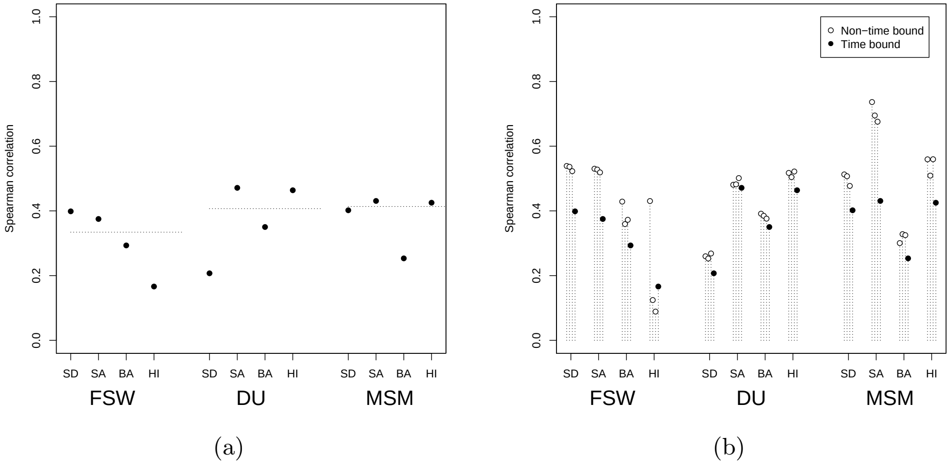


Fig. S12. Boxplots of the test-retest difference in reported degree, measured by Question (G), by group and city. There is no general pattern of increase or decrease. In order to show the median, 25th and 75th percentiles more clearly, this plot does not include points outside of the whiskers.



(a)

(b)

Fig. S13. (a) Spearman rank correlation between test and retest measures for the main degree question (G) with the median value for each group marked by the horizontal line. (b) The measures that are not time-bounded (D, E, F) have higher correlation.

out their recruits, or select them from among incidentally encountered alters.

S4.2. Test-retest reliability

In order to assess the test-retest reliability of the degree questions, questions D-G were included in both the initial and follow-up interviews. The median difference in degree (G) at interview and follow-up is 0 (25^{th} percentile = -3, 75^{th} percentile = 3). Further, Fig. S12 shows that results were similar across study site and target population.

Fig. S13(a) shows the test-retest reliability for the degree question used for estimation (question G). One potential reason for the low test-retest reliability of question G is that it refers to a seven day time frame. Therefore, even if respondents are perfectly accurate in their responses there could be test-retest variation because of week-to-week variation. This issue of time-bounded questions has come up in other test-retest studies (e.g., van Groenou et al. (1990)), but is difficult to resolve because it is not reasonable to ask respondents at the follow-up interview about their experiences in the one week proceeding their initial interview as that is often about three weeks in the past. However, one way to roughly gauge how much extra variability is introduced by this time frame is by examining the first three network size questions which are not time-bounded. Fig. S13(b) shows that the test-retest reliability is higher for the non time-bounded questions, but only slightly so.

Finally, we note that when considering measures of test-retest reliability, it is critical to consider any potential sources of dependence between the measures. Here interviewers at the follow-up visit did not know respondent's answers from the initial visit. Further, since the time period between interviews was generally around three weeks, it is extremely unlikely that respondents remembered their original responses to the degree questions. One possible source of dependence that did exist in this study is that the respondents may have been interviewed by the same interviewer at the initial and follow-up visits, thus possibly increasing perceived test-retest reliability.

S5. Testing Recruitment Bias

Most RDS inference relies on the assumption that recruits are selected at random from among the contacts of each recruiter. Under this assumption, successful recruits should constitute a simple random sample of the personal networks of respondents. In most cases, reviewing a Recruitment Bias Plot (e.g., Fig. 10) should be sufficient to inform researchers' intuition about whether recruitment bias is a concern. In some cases, researchers may want to test whether the observed recruitment patterns are consistent with random recruitment. However researchers should note that statistical significance is not the same as estimator bias, so even a perfect test would not be a good judge of whether a recruitment bias is strong enough to impact estimates.

With no recruitment bias, the coupons should be passed to a simple random sample of the recruiter's contacts, and the coupons returned should be returned by a simple random sample of those receiving coupons. To test these assumptions non-parametrically, we compare the (unweighted) count of employed at each stage to a null distribution approximated by simulated simple random sampling from the reported composition of the relevant subset of each recruiter's eligible alters. To test for biased coupon passing, for example, we simulate the coupon-recipients of each recruiter by drawing n_i^c samples from among F_i units, including J_i employed, where n_i^c is the reported number of coupons distributed by i, F_i is i's reported number of contacts, and J_i is i's reported number of employed alters. Non-parametric null distributions for returning coupons and for overall recruitment were constructed similarly, with test statistics and reference distributions described in Table S2.

Our tests show very small p-values, suggesting the reported recruitment patterns are very unlikely absent recruitment bias (see Table S3). However, one reason for these extreme findings could be poor data quality, perhaps due to a desirability bias of employment status reporting. Many data points are logically inconsistent, with more employed alters receiving coupons than were originally reported known, or with more employed recruits than coupons given to employed people. In addition, there is no evidence that those with more reported employed contacts tend to recruit more employed people (the correlation between

Table S2. Test statistics and reference distributions for testing for Recruitment Bias at three levels: in the passing of coupons, in returning coupons, and overall.

Test	Test statistic	Reference Distribution
Coupon Passing	Count of employed coupon-recipients	SRS from contact composition of each recruiter
Returning Coupons	Count of employed recruits	SRS from composition of coupon recipients of each recruiter
Overall	Count of employed recruits	SRS from contact composition of each recruiter

these proportions is negative in many of the samples). Therefore, while we feel this test is mathematically appropriate, we suggest caution in its use, or the use of earlier tests relying on self-reported network compositions.

Finally, we note that other approaches have been used previously to compare reported network composition to actual sample recruits. These approaches can be roughly divided into two categories: those that make standard distributional assumptions, including independence, to perform chi-square and t-tests of recruitment patterns, and those that test for the impact on estimators, using standard RDS confidence intervals. While each of these other approaches certainly adds to information available to researchers, we prefer our non-parametric testing approach because it relies on neither parametric assumptions, nor the validity of standard RDS confidence intervals.

The earliest work in assessing non-random recruitment is in Heckathorn et al. (2002), which looks at the correlation between implied population proportions across several groups under random sampling and observed cross-group recruitment patterns. This approach is not ideal because we would like to test whether the compositions are the same, not just correlated. Wang et al. (2005) therefore extend this approach by using a t-test to compare the sample proportion of the observed data to the reported network compositions. This approach relies on a binomial approximation to the distribution of the estimated proportions. Wejnert and Heckathorn (2008) introduce a chi-square test to compare the expected referral matrix under random referral to the observed referral matrix. This approach also relies on distributional assumptions, in particular an assumption of independence of observations. It is unclear from the Wang et al. (2005) and Wejnert and Heckathorn (2008) papers which form of weighting is used to estimate the composite alter characteristics. Rudolph et al. (2011) also use chi-square tests and t-tests to test for non-random sampling, but they do not give details on their methods. Finally, Jenness et al. (2014) use a t-test to compare the geographical distance to sampled contacts to (a proxy for) the distance to contacts in general.

Iguchi et al. (2009) compare population proportion estimates directly, by substituting reported network composition for composition of referrals in RDS estimators. Their test is based on comparing the usual and network-composition-based population proportion estimates, using the RDS uncertainty estimates under the two conditions. This approach has the advantage of placing the comparison on the scale of the measure of interest, but is only applicable for non-random recruitment on the feature to be estimated, and can only be used with estimators relying on estimates of proportions of cross-group ties (e.g. Salganik and Heckathorn (2004)). Liu et al. (2012) also propose a test relying on standard RDS estimates for some (five) visible attributes and their confidence intervals, and also make use of information collected on the composition of the social network alters of the respondents. They compare the estimates to the value of the "population proportions of five visible

Table S3. P-values for non-parametric tests of recruitment bias based on employment status on three levels: Which contacts are given coupons, which coupon recipients return coupons to become recruits, and overall, which contacts become recruits. P-values suggest the reported recruitment patterns are very unlikely absent recruitment bias. The second section records the proportion of cases in each setting in which the number of employed persons selected was larger than the number available. This suggests the apparently strong recruitment bias may be due to data quality issues.

	P-value				Proportion Inconsistent			
	SD	SA	BA	HI	SD	SA	BA	HI
Coupon Passing	0.369	$< .0001$	$< .0001$	$< .0001$	0.094	0.137	0.201	0.216
Returning Coupons	$< .0001$	$< .0001$	$< .0001$	$< .0001$	0.519	0.286	0.352	0.243
Overall	$< .0001$	$< .0001$	$< .0001$	$< .0001$	0.202	0.143	0.192	0.206

variables among the total drug-using alters from which the RDS sample was drawn." It is not clear from their paper how they compute this proportion, or precisely how it relates to the population proportion.

The work of Yamanis et al. (2013), concurrent with this work, provides descriptive analyses most similar to our work. These authors also compare the inferred characteristics of invited recruits, successful recruits, and the full population of all alters, in a tabular representation similar to a *Recruitment Bias Plot*, and broken down by recruiter characteristics. Their descriptives are also slightly different from ours because their measurement strategy is different from ours, asking respondents only collectively about the features of their uninvited alters. Rather than testing for non-random recruitment directly, they suggest a test for the impact of non-random recruitment on inference, similar to the approach of Iguchi et al. (2009). They suggest a modified bootstrap procedure to compute theoretical proportion estimates under random sampling from invited recruits, successful recruits, and the full population of alters. They derive statistical significance based on the relative values of these estimates as compared to their bootstrap confidence intervals.

S6. Non-response to Follow-up

We use data collected from the follow-up interview to provide some evidence for some of our diagnostics, but only about half of participants completed a follow-up interview (Fig. 2). Therefore, we compare the participants who did and did not complete a follow-up questionnaire on several characteristics, and summarize those results in Table S4. Table S5 presents comparisons of each population type, aggregated across the four cities, as well as a grand aggregation (total) over all 12 sites. Chi-square and t-tests for statistical significance were conducted for each of the 12 sites, each of the 3 groups, and the study total, resulting in 16 tests for each trait; tests that were significant at the .05 level are listed in table S5.

We first compare the degrees of respondents and non-respondents across the 16 comparisons, the mean degrees of follow-up respondents and non-respondents were only significantly different in one site, Drug Users in Santo Domingo.

Many respondents return to the study center to receive secondary incentives for successful recruitments. For this reason, we might expect higher rates of recruitment among follow-up respondents. Indeed, this is the case in our data, with significantly higher average recruitments among follow-up respondents in all 16 comparisons. A comparison of the proportions with no recruits continues this pattern, with a significantly higher proportion with no recruits among follow-up non-respondents. Finally, we also compare the average number of recruits for follow-up respondents and non-respondents, excluding those with no recruits. Again, all sites have significantly higher average numbers of recruits among follow-up respondents.

Table S4. Comparison of respondents (Resp.) to follow-up and non-respondents (Non-Resp) to follow-up, based on average values of various variables for each population and for the full sample.

	FSW Resp	FSW Non-Resp	DU Resp	DU Non-Resp	MSM Resp	MSM Non-Resp	Total Resp	Total Non-Resp
N	558	698	557	665	562	826	1677	2189
Mean Degree	9.08	7.68	14.04	14.61	9.68	6.82	10.93	9.46
# Recruits	1.57	0.50	1.47	0.64	1.62	0.56	1.55	0.57
No Recruits	0.22	0.68	0.23	0.57	0.18	0.63	0.21	0.63
# Recruits/0	2.02	1.57	1.89	1.49	1.97	1.54	1.96	1.53
Study Day	24.04	35.53	15.23	19.71	28.43	34.33	22.59	30.27
Study Day/0	22.21	35.39	13.23	16.43	27.51	31.17	21.13	26.79
HIV Positive	0.04	0.05	0.05	0.08	0.09	0.06	0.06	0.06
For Incentive	0.03	0.03	0.07	0.12	0.14	0.14	0.08	0.10
For HIV Test	0.85	0.86	0.59	0.65	0.12	0.14	0.52	0.53
For Any Test	0.86	0.87	0.61	0.66	0.37	0.38	0.61	0.62

Table S5. Listing of populations showing significant differences between follow-up respondents and non-respondents for each variable considered in Table S4. Significant results for the full sample are indicated by "Total", and for aggregated data by population by "FSW," "DU," and "MSM." Average differences for these populations can be found from Table S4. For individual sites, numbers in parentheses are the difference in means (Mean_Resp-Mean_Non-Resp). Statistical significance based on t-test and chi-square tests at level 0.05.

Term	Significant Sites
Mean Degree	DU-SD (-7.83)
# Recruits	Total, FSW, DU, MSM, FSW-SD (1.98), FSW-SA (0.88), FSW-BA (0.92), FSW-HI (0.57), DU-SD (0.87), DU-SA (0.78), DU-BA (0.88), DU-HI (0.82), MSM-SD (0.97), MSM-SA (1.45), MSM-BA (0.86), MSM-HI (1.13)
No Recruits	Total, FSW, DU, MSM, FSW-SD (-0.75), FSW-SA (-0.35), FSW-BA (-0.41), FSW-HI (-0.26), DU-SD (-0.29), DU-SA (-0.29), DU-BA (-0.43), DU-HI (-0.38), MSM-SD (-0.39), MSM-SA (-0.56), MSM-BA (-0.45), MSM-HI (-0.47)
# Recruits/0	Total, FSW, DU, MSM, FSW-SD (0.66), FSW-SA (0.45), FSW-BA (0.54), FSW-HI (0.24), DU-SD (0.55), DU-SA (0.45), DU-BA (0.28), DU-HI (0.25), MSM-SD (0.37),, MSM-SA (0.89), MSM-BA (0.29), MSM-HI (0.5)
Study Day	Total, FSW, DU, MSM, FSW-SD (-10.7), FSW-SA (-19.77), FSW-BA (-9.33), FSW-HI (-4.27), DU-SD (-4.56), DU-BA (-7.82), MSM-SA (-12.72), MSM-BA (-5.28), MSM-HI (-11.73)
Study Day/0	Total, FSW, DU, MSM, FSW-SD (-12.63), FSW-SA (-21.37), FSW-BA (-10.8), DU-SD (-6.81), MSM-SA (-16.89), MSM-HI (-16.82)
HIV Positive	DU, FSW-BA (-0.08), MSM-HI (0.08)
For Incentive	DU
For HIV Test	DU, MSM-SA (-0.1)
For Any Test	none

It is also possible there is a censoring effect deterring response to follow-up among respondents completing their primary interviews closer to the end of the study. Indeed, we find that follow-up non-respondents tend to have their initial interviews significantly later in the study than follow-up respondents in all but 3 sites. Part of this effect may be attributable to the fact that respondents at the very end of the study are not given coupons and are therefore unable to make recruitments. Comparing the average study date for follow-up respondents and non-respondents excluding those with no recruits reveals remaining significant differences in all but 6 sites.

Finally, the second contact with the interview site also serves to deliver HIV and other test results. We therefore compare follow-up respondents and non-respondents on HIV status as well as based on their stated motivations for participating in the study. We find significant differences in HIV status in only 3 cases (drug users overall, female sex workers in Barahona, and men who have sex with men in Higuey), without a consistent pattern in the direction of the association. In terms of motivations for participation, drug users overall (but not in any particular site) are less likely to participate in the follow-up study if their primary stated motivation was for the financial incentive. Surprisingly, drug users whose primary stated motivation was HIV test results were less likely to participate in follow-up, while MSM in Higuey were more likely to participate in follow-up when motivated by HIV test results. There were no significant differences in other sites. We also compared the proportions who mentioned any disease test as a primary motivation, and found no significant differences.

Overall, then, our follow-up non-respondents seem to differ from follow-up respondents primarily based on their rates of recruitment (follow-up respondents recruited more often), and study date (earlier participants more likely to follow-up). Where relevant to our conclusions, the impacts of these results are noted in the text.

References

Caceres, C., K. Konda, M. Pecheny, A. Chatterjee, and R. Lyerla (2006). Estimating the number of men who have sex with men in low and middle income countries. *Sexually Transmitted Infections 82*(suppl 3), iii3–iii9.

Gile, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association 106*(493), 135–146.

Handcock, M. S. (2011). size: Estimating hidden population size using respondent driven sampling data. R package version 0.20.

Handcock, M. S., K. J. Gile, and C. M. Mar (2012). Estimating hidden population size using respondent-driven sampling data. *Working paper*.

Heckathorn, D., S. Semaan, R. Broadhead, and J. Hughes (2002). Extensions of respondent-driven sampling: A new approach to the study of injection drug users aged 18-25. *AIDS and Behavior 6*(1), 55–67.

Iguchi, M. Y., A. J. Ober, S. H. Berry, T. Fain, D. D. Heckathorn, P. M. Gorbach, R. Heimer, A. Kozlov, L. J. Ouellet, S. Shoptaw, and W. A. Zule (2009). Simultaneous recruitment of drug users and men who have sex with men in the United States and Russia using respondent-driven sampling: Sampling methods and implications. *Journal of Urban Health 86*(S1), 5–31.

Jenness, S. M., A. Neaigus, T. Wendel, C. Gelpi-Acosta, and H. Hagan (2014). Spatial recruitment bias in respondent-driven sampling: Implications for HIV prevalence estimation in urban heterosexuals. *AIDS and Behavior*, forthcoming.

Liu, H., J. Li, T. Ha, and J. Li (2012). Assessment of random recruitment assumption in respondent-driven sampling in egocentric network data. *Social Networking 1*(2), 13–21.

Oficina Nacional de Estadistica (2009). Poblacion estimada y proyectada region provincia y municipio 2000-2010. September 18, 2009; Accessed August 2, 2012.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org/.

Rudolph, A. E., N. D. Crawford, C. Latkin, R. Heimer, E. O. Benjamin, K. C. Jones, and C. M. Fuller (2011). Subpopulations of illicit drug users reached by targeted street outreach and respondent-driven sampling strategies: Implications for research and public health practice. *Annals of Epidemiology 21*(4), 280–289.

Salganik, M. J. and D. D. Heckathorn (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology 34*(1), 193–240.

van Groenou, M. B., E. v. Sonderen, and J. Ormel (1990). Test-retest reliability of personal network delineation. In T. Antonucci and C. Knipscheer (Eds.), *Social Network Research: Substantive Issues and Methodological Questions*, pp. 121–136. Amsterdam: Swets and Zeitlinger.

Volz, E. and D. D. Heckathorn (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics 24*(1), 79–97.

Wang, J., R. G. Carlson, R. S. Falck, H. A. Siegal, A. Rahman, and L. Li (2005). Respondent-driven sampling to recruit MDMA users: A methodological assessment. *Drug and Alcohol Dependence 78*(2), 147–157.

Wejnert, C. and D. D. Heckathorn (2008, August). Web-based network sampling efficiency and efficacy of respondent-driven sampling for online research. *Sociological Methods & Research 37*(1), 105–134.

Yamanis, T. J., M. G. Merli, W. W. Neely, F. F. Tian, J. Moody, X. Tu, and E. Gao (2013, August). An empirical analysis of the impact of recruitment patterns on RDS estimates among a socially ordered population of female sex workers in China. *Sociological Methods & Research 42*(3), 392–425.