**eAppendix**

*Estimator Naming Conventions and Sources*

We examine seven respondent-driven sampling estimators in this paper. There is some variation in how these estimators are referred to in the literature, so we use either the conventions of the literature or otherwise identifying acronyms. To illustrate the theoretical similarities between two sets of respondent-driven sampling estimators, we prefix all of them, except the naïve estimator, with either RDS1- or RDS2-. Each RDS1-style estimator uses information from some measurement of the group to group transition matrix, whereas RDS2-style estimators ignore the group to group recruitment matrix and focus more on respondents’ network sizes. The seven estimators and their sources can be found in eTable 1.

*Sources of Simulated Data*

This paper makes use of simulated data generated for a previous study which assesses respondent-driven sampling performance (Merli et al. 2015). This study used data from the PLACE-RDS Comparison Study to generate synthetic population social networks and respondent-driven sampling chains running over these networks. The PLACE-RDS Comparison Study comprised two surveys conducted contemporaneously in the same population of female sex workers in Liuzhou, China: a respondent-driven sampling sample ($n=583$, with seven seeds) and a sample recruited through a venue-based approach called priorities for local AIDS control efforts ($n=161$) (cf. Weir et al. 2012 for more details on the design of the surveys). Both surveys had the same target population and used the same inclusion definitions – whether the respondent had exchanged sex for money in the last four weeks. Variables were also measured consistently across surveys – a rapid syphilis test screening measuring lifetime history of infection and a core activities questionnaire. The study also included a unique follow-up questionnaire where participants were asked about the attributes of all their social network contacts including those who were recruited in the sample and those who were not.

The procedure used by Merli et al. (2015) to generate the synthetic population social networks from these data sources is summarized as follows:

1) Construct a simulated set of urban venues within which members of a simulated population of female sex workers are placed.

2) Bootstrap sample 500 venues from lists of urban venues constructed for the priorities for local AIDS control efforts survey and distribute 7,500 simulated female sex workers across the simulated venues according to probabilistic rules[[1]](#footnote-1).

3) Use new methodologies in the social networks literature – case control logistic network regression (Smith 2012) and exponential random graph modeling (Hunter et al. 2008) – to construct synthetic social networks connecting members of this population of female sex workers[[2]](#footnote-2). Merli et al. parameterize the social network simulations with data drawn from both the respondent-driven sampling and priorities for local AIDS control efforts surveys, including the network module that collected respondent reports on social network peers which was previously tested among female sex workers in Shanghai (see Yamanis et al. 2013), on the assumption that these data represent the best estimates of the network structure linking the hidden population of female sex workers in Liuzhou.

In all, this approach produced a range of synthetic population social networks for the female sex worker population of Liuzhou which were taken to reflect uncertainty in (a) the estimates based on variability in the source of data that informed the demographic characteristics of female sex workers (respondent-driven sampling survey or priorities for local AIDS control efforts survey), (b) whether the simulated venue size was adjusted to reflect the number of sex workers in the priorities for local AIDS control efforts venues or not, and (c) variability in the effect of assumed geographic distance between a respondent and her peers on the probability of a network tie. The analyses we perform in the current paper focus on one of these variants, which we refer to as our focal population social network scenario. This scenario is informed by female sex workers characteristics from the priorities for local AIDS control efforts data, simulated venue size adjusted to reflect venue size reported in the priorities for local AIDS control efforts survey, and weak geographic distribution of ties where network peers reported further away than a 10-min walk were spread farther into the physical space (see Merli et al. 2015 for additional details). While we focus on only one of the population social networks, we note that our results would not substantively differ were we to have used a different network and that our conclusions are consistent with other considerations found in Merli et al. (2015). eFigure 1 provides comparisons of results across the different population social networks.

In the simulated networks which did not contain a single group of people wherein everyone could reach everyone else through some chain of connections, Merli et al. (2015) limited their analysis to the largest connected component in these undirected networks to comply with standard respondent-driven sampling assumptions that the network is connected (cf. Volz and Heckathorn 2008). eTable 2 contains a comparison of the largest connected component sizes and other items of interest in the eight network scenarios. In what follows and throughout the main text, we use only means from the largest connected component subset in constructing our measures of bias, etc.

To simulate the respondent-driven sampling process over these synthetic population social networks, Merli et al. (2015) performed stochastic replication of chains which trace links between individuals in the simulated population. Respondent-driven sampling chains were simulated without replacement (i.e. respondent-driven sampling participants can only appear once in any given chain) under different scenarios consistent with an incremental set of key theoretical features of the respondent-driven sampling process. Scenarios include (a) chains that mimic the size, demographics and branching structure of observed chains – both seeds and later respondents in each wave of the simulated chains have identical characteristics to those of the respondent-driven sample; (b) chains are initiated by seeds with the same characteristics of the actual convenience sample of seeds but the rest of the sampling process proceeds through the simulated network following the branching of the Liuzhou respondent-driven sampling study where participants each recruit a maximum of two other participants from her network peers at random; (c) chains run under idealized sampling conditions consistent with the theoretical respondent-driven sampling literature, with random seeds and random referrals. We focus on scenarios (a) and (c) in this paper because scenario (b) falls between these two extremes in its level of realism. While we ignore scenario (b) for space constraints, we note that preliminary analyses showed similar conclusions to those presented here (not shown). The respondent-driven sampling process under each of these scenarios was repeated 1,000 times on each population social network.

*Measurement of Network Types*

The PLACE-RDS Comparison Study and the Shanghai Women’s Health Survey data sets, used for the results section “Performance of Respondent-Driven Sampling Estimators in Empirical Samples” in our paper, included a unique follow-up questionnaire about the characteristics of network peers who were and were not recruited into the survey. This questionnaire was administered to the subset of participants who returned to the interview site to collect the secondary incentive for successful referrals, including 310 respondents in the PLACE-RDS Comparison Study and 271 in the Shanghai Women’s Health Survey. This follow-up questionnaire allows one to track the number and attributes of invited and uninvited peers in recruiting participants’ social networks. The number of invited peers who accepted an invitation was obtained from the recruiter’s response to the following question: “How many people who you have invited to participate by offering them a coupon have accepted it?” Recruiters were asked to provide detailed information on the individual attributes of up to four people who accepted the coupon. The number of those who refused an invitation was obtained from the recruiting participant’s response to the question: “How many people who you have invited to participate by offering them a coupon have refused?” Recruiting participants were asked to provide detailed information on the characteristics of up to four people whom they invited but rejected the coupons. As a unique variation on standard respondent-driven sampling practice, recruiting participants were also asked about network peers whom they did not invite to participate with the question: “How many female sex workers do you know whom you did not invite to participate?” We refer to this group as their uninvited peers.

A number of checks were performed on the quality of the network data collected in both surveys. For the PLACE-RDS Comparison Study data collected in Liuzhou we have calculated a number of measures that are directly relevant to this paper. Yamanis et al. (2013:209-210) report similar measures for the Shanghai Women’s Health Survey. First, we found that respondents’ reported network size in the PLACE-RDS Comparison Study was correlated at 0.91 and did not significantly differ (paired t-test: $t=-0.7,p=0.484$) in a test-retest experiment of the reports during the primary and follow-up interviews. Yamanis et al. reported test-retest correlations of 0.98 on network size in the Shanghai Women’s Health Survey. Next, we examined the accuracy of recruiting respondents’ reports on the attributes of their network peers who accepted the coupon in the PLACE-RDS Comparison Study. Recruiting participants were instructed to distribute coupons to network peers in the ascending order of the coupon numbers. In the follow-up interview, they were instructed to answer questions on attributes of their recruited peers in that same order. Thus we can assume that recruiters reported on their recruited peers in the order they were instructed. Using the coupon numbering system linking recruiters to their recruits we were able to evaluate recruiting participants’ reports on each of their recruited peers in the follow-up survey against recruited participants’ self-reports in the primary interview on the same attributes. These matched in 91% of the cases, a high level of agreement. Consistency of reports was also high on other attributes of recruited peers (e.g. age), with correlation coefficients of 0.93 and 0.97, respectively, measuring reports about recruiters and recruits. Yamanis et al. describe similar consistency in the Shanghai Women’s Health Survey.

Because of recruiting participants’ concerns with the length of the follow-up interview, especially if they had many social contacts, recruiting participants were asked to describe their uninvited peers as a group, rather than as individuals. To describe the venues of uninvited peers who did not work in the same venue as the focal participant, they were asked to select from a list of venue types and were allowed to choose multiple options to describe the diversity of this group with respect to venue type. These data constraints led us to adopt the following allocation procedure to allocate uninvited peers to a specific tier:

1. Initialize three variables. One to hold the number of peers that remain unassigned, which is initially defined as the self-reported network size of each respondent, and the remaining two to count the number of peers in the low tiers of sex work and the number of peers in the high tier of sex work, respectively, which are both initialized at zero.
2. Add the number of directly reported acceptors and rejecters who were in each tier to their respective variables, and subtract the total number of acceptors and rejecters from the unassigned peers variable. Because respondents reported directly on the characteristics of these peers we know their tier of sex work (to the extent that answers to the follow-up network questionnaire are accurate).
3. For the remaining unassigned peers, take advantage of the number of checks across multiple venue types. Because venue types were grouped into tiers, we can use the unique number of venue types checked within a tier as a minimum bound estimate on the number of peers in that tier. For instance, imagine a hypothetical respondent in the Shanghai Women’s Health Survey checked that their uninvited peers recruited clients in parks, hair salons, bath houses, and nightclubs. In this case, we would assign three peers to the low tiers of sex work category because parks, hair salons, and bath houses are venues associated with the low tiers of sex work. We would also assign one uninvited peer to the high tier because nightclubs are high tier venues. After this, we would deduct four peers from the count of unassigned peers. If a respondent made checks within only one type of tier – either all high tier or all in the low tiers of sex work – then we assigned all remaining peers to that tier.
4. For any remaining unassigned peers, assign them to the low tiers of sex work conditionally at random. This is done by assigning them proportionate to the naïve proportion of respondents in the whole survey who were in the low tiers of sex work conditional on their recruiter being in the low tiers of sex work or in the high tier.

 The aforementioned procedures were used in both surveys to guarantee consistency. While the respondent-level network data generated in this manner do not entirely conform to the format required by the RDS1-LEN estimator which calls for the tier composition of respondents, such as “what proportion of your network alters solicit in tier?” (Lu 2013), they yield a reasonable representation of the composition of personal networks of survey respondents. However, the effectiveness of these procedures varies across the samples. In the Liuzhou data, steps 1-3 listed above accounted for all peers of 80.3% of respondents, meaning only 19.7% of respondents had any peers assigned conditionally at random. By contrast, in Shanghai we were able to fully assign peers for only 46.5% of respondents. This discrepancy may owe to the fact that recruiting respondents in the Liuzhou survey were given the option to check 11 venue types compared with only 8 venue types offered to respondents in the Shanghai survey. This discrepancy may also explain the abnormal estimate for the Shanghai whole network reported in Table 2 of the main text.

***eAppendix References Not Cited in Text***

Gile, Krista J., and Mark S. Handcock. 2011. “Network Model-Assisted Inference from Respondent-Driven Sampling Data.” *arXiv:1108.0298 [stat]*, August. http://arxiv.org/abs/1108.0298.

**eTable 1. The seven respondent-driven sampling estimators evaluated in this paper.**

|  |  |
| --- | --- |
| **Estimator** | **Source** |
| 1. Naïve | None |
| 2. RDS1-SH | Salganik MJ, Heckathorn DD. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociol Methodol*. 2004;34(1):193–240. doi:10.1111/j.0081-1750.2004.00152.x. |
| 3. RDS1-DS | Heckathorn DD. Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. *Soc Probl*. 2002;49(1):11-34. doi:10.1525/sp.2002.49.1.11. |
| 4. RDS1-DG | Heckathorn DD. Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment. *Sociol Methodol*. 2007;37(1):151–207. doi:10.1111/j.1467-9531.2007.00188.x. |
| 5. RDS1-LEN | Lu X. Linked Ego Networks: Improving estimate reliability and validity with respondent-driven sampling. *Soc Netw*. 2013;35(4):669-685. doi:10.1016/j.socnet.2013.10.001. |
| 6. RDS2-VH | Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. *J Off Stat*. 2008;24(1):79. |
| 7. RDS2-SS | Gile KJ. Improved Inference for Respondent-Driven Sampling Data With Application to HIV Prevalence Estimation. *J Am Stat Assoc*. 2011;106(493):135-146. doi:10.1198/jasa.2011.ap09475. |

**eTable 2. Largest connected component size and proportions, and comparing population means to largest connected component means, by data scenario.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Source** | **Size Adjust** | **Geographic Distribution** | **No. in largest connected component** | **Percent in largest connected component** | **Population proportion in low tiers** | **Largest connected component proportion in low tiers** | **T-test p-value** |
| *Respondent-driven sampling* | *No* | *Strong* | 7500 | 100% | 0.547 | 0.547 | 1.000 |
| *Respondent-driven sampling* | *No* | *Weak* | 7500 | 100% | 0.547 | 0.547 | 1.000 |
| *Respondent-driven sampling* | *Yes* | *Strong* | 7500 | 100% | 0.547 | 0.547 | 1.000 |
| *Respondent-driven sampling* | *Yes* | *Weak* | 7500 | 100% | 0.546 | 0.546 | 1.000 |
| *Priorities for local AIDS control efforts* | *No* | *Strong* | 6837 | 91.2% | 0.476 | 0.484 | 0.347 |
| *priorities for local AIDS control efforts* | *No* | *Weak* | 6841 | 91.2% | 0.476 | 0.489 | 0.112 |
| *priorities for local AIDS control efforts* | *Yes* | *Strong* | 6833 | 91.1% | 0.473 | 0.483 | 0.219 |
| *priorities for local AIDS control efforts* | *Yes* | *Weak* | 6833 | 91.1% | 0.473 | 0.485 | 0.150 |

**Note: T-test p-value is from a t-test with unequal variances of the difference in means between whole population and the largest connected component subset. This table shows the differences in means between the total population and the LCC by data scenario. For the respondent-driven sampling data source, every respondent is in the largest connected component. For all generated networks from the priorities for local AIDS control efforts (PLACE) data source, approximately 91% was only in the largest subset. The means did not differ substantially or significantly between the total population and the subset in the largest connected component.**

**eFigure 1. Distributions of estimates of proportion working in low tiers of sex work in all population social networks, by estimator and seeding and recruitment scenario.**

****

**Notes: Population social networks are the following: 1) priorities for local AIDS control efforts data source, venue size adjusted, weak geographic distribution of ties (the focal scenario of Figures 1 and 2 and Table 1 in the main text); 2) priorities for local AIDS control efforts data source, venue size not adjusted, weak geographic distribution of ties; 3) priorities for local AIDS control efforts data source, venue size adjusted, strong geographic distribution of ties; 4) priorities for local AIDS control efforts (PLACE) data source, venue size not adjusted, strong geographic distribution of ties; 5) respondent-driven sampling data source, venue size adjusted, weak geographic distribution of ties; 6) respondent-driven sampling data source, venue size not adjusted, weak geographic distribution of ties; 7) respondent-driven sampling data source, venue size adjusted, strong geographic distribution of ties; 8) respondent-driven sampling data source, venue size not adjusted, strong geographic distribution of ties. Estimators are indicated with letters: a) Naïve, b) RDS1-SH, c) RDS1-DS, d) RDS1-DG, e) RDS1-LEN, f) RDS2-VH, g) RDS2-SS. The sources for these estimators are described in Table A1. Seeding and recruitment scenarios are indicated with lower case Roman numerals: i) matched seeds and recruitments, ii) random seeds and random recruitments. The vertical line indicates the population mean in that population social network scenario (see eTable 2).**

1. Estimates of the number of female sex workers and the number of venues are drawn from the Liuzhou Center for Disease Control. [↑](#footnote-ref-1)
2. In an unpublished paper, Gile and Handcock (2011) used a similar set of procedures to develop a “model assisted” respondent-driven sampling estimator. We do not focus on this estimator in this paper because code for its implementation is not in broad circulation, and because its approach to deriving estimates mirrors our approach to creating synthetic population social networks and simulating respondent-driven sampling samples on those networks. [↑](#footnote-ref-2)