Challenges to Recruiting Representative Samples of Female Sex Workers in China using Respondent Driven Sampling: How Much of the Network Do We See? [1]

M. Giovanna Merli,[a] Jim Moody,[a] Jeffrey Smith,[a] Jing Li,[ac] Sharon Weir[b] and Xiangsheng Chen[c]

a) Duke University
b) UNC Chapel Hill, MEASURE Evaluation
c) China National Center for STD Control

Short Abstract: We explore the network coverage of a Respondent Driven Sampling (RDS) sample of female sex workers (FSW) in China as part of an effort to evaluate RDS' claim of population representation with empirical data. We take advantage of unique information on the social networks of FSW obtained from two overlapping studies of FSWs --RDS and a venue-based sampling approach (PLACE) -- and use an exponential random graph modeling (ERGM) framework from local networks to construct the likely network from which our observed RDS is drawn. We then run recruitment chains over this simulated population and produce a sample with characteristics consistent with the observed RDS. We estimate population coverage rates by comparing population proportions and RDS sample proportions. We discuss the results in light of (a) potential estimation improvements implicit in network information, (b) strategies for improving coverage rates, and (c) multiple sources of potential variability in coverage.

Introduction

One of the challenges of studying HIV/STD infection and related risk behaviors in hidden populations like FSWs in China is the recruitment of a valid sample from which inference to the population can be drawn. FSWs are a hidden population because of the absence of complete sampling frames, social stigma and the illegal status of sex work. The organization of sex work in China also complicates efforts to recruit representative samples because it is structured around a semi-rigid hierarchy of tiers of sex work, ranging from high to low, which complicate researchers' efforts to access all strata of this population (Hershatter 1997; Huang, Henderson, Pan and Cohen. 2004; Lim 1998; Parish and Pan 2006; Rogers, Yin, Xin et al. 2002; Xia and Yang 2005).

[1] The analyses described here are funded by NICHD grant 1R01HD068523 to Duke University (Merli, PI). Funding for the collection of PLACE and RDS data was provided by USAID under the terms of cooperative agreements GPO-A-00-03-00003-00 and GPO-A-00-09-00003-0; by NICHD through the UNC R24 "Partnership for Social Science Research on HIV/AIDS in China" (Henderson, PI) and the Pre-Doctoral Research Program at the Carolina Population Center, University of North Carolina; by UNICEF, UNDP, World Bank, and WHO through the "WHO Rapid Syphilis Test Project (WHO A70577)"; by the Duke University and University of North Carolina Center(s) for AIDS Research; and by the National Center for STD Control in China. The PLACE-RDS Study was led by Sharon Weir (PI), with co-investigators, Xiangsheng Chen and Giovanna Merli. We thank the physicians and the outreach workers in the study areas for their hard work, and the study participants for their cooperation. The opinions expressed are those of the authors and do not necessarily reflect the views of any government.

Respondent Driven Sampling (RDS) has become an increasingly prominent sample recruitment tool for hidden and hard to reach population. It claims to provide a probability-based inferential structure to representations of these populations by capitalizing on their network structure to identify and interview subjects and to generate unbiased estimates of characteristics and behaviors. RDS encouraged study participation by exploiting social ties within the target population and efficiently and cost-effectively recruits large, diverse samples in a relatively short amount of time (Robinson, Risser, Goy et al. 2006; Kendall, Kerr, Gondim et al. 2008; Johnston, Sagin, Tian and Huong 2006; Carballo-Deguez, Balan, Marone et al. 2011). However, the capacity of RDS to represent faithfully hidden populations such as FSWs relies on strong, largely empirically untested assumptions regarding the unobserved referral process of participants to other participants, the characteristics of the underlying network and, more generally, the process of recruitment of potential respondents into the sample.

In this paper, we explore the network coverage of an RDS sample of a hidden population, female sex workers (FSW) in China. This work is part of a larger effort to evaluate Respondent Driven Sampling (RDS) which moves considerations regarding real-world network referral processes from the theoretical to the empirical realms. We combine information from two concurrently implemented surveys of female sex workers in Liuzhou (Guangxi Province): an RDS survey which includes unique information about attributes of respondents' social networks and a venue-based survey.

Respondent Driven Sampling

In RDS the target for representation is the hidden population of a well-defined geographic area (e.g. a city). RDS starts by recruiting survey respondents through a chain-referral approach that is initiated by the selection of a limited number of "seed" respondents known to the researchers administering the study and belonging to the population of interest. Seeds are interviewed and given a limited number of coupons which they are asked to distribute to their immediate social contacts in the target population as a means of recruiting other participants from among their social networks. Members of the seeds' social circles who receive coupons and then choose to participate in the study form the first "wave" of the sample. This process proceeds recursively (hence the term "chain-referral") through multiple waves until a desired sample size is reached.

RDS's relative ease of recruiting study participants comes at the cost of making two related assumptions about the sampling process used to select respondents: (1) all members of the target population can, in principle, be reached through the chain-referral process, and (2) an individual's sample inclusion probability is *exactly* proportional to the number of reciprocal ties she has with other members of the target population (her personal network size or degree), which we refer to as sampling with probability proportional to degree (SPPD) assumption. In order to assess each respondent's degree, RDS asks respondents to report the number of people they know in the target population. It uses this information to approximate sample inclusion probabilities as the basis for making population estimates. In this way, the RDS estimators (Salganik and Heckathorn 2004; Volz and Heckathorn 2008) attempt to compensate for the perceived tendency of RDS's chain referral strategy to over-sample individuals with large personal social networks.

The two assumptions described above are usually rationalized by thinking in terms of an idealized model for the *unobserved* underlying social network and coupon-based referral process.

Specifically one assumes (a) equal probability that a respondent, already contacted, will refer any of the individuals from her immediate social network; (b) reciprocity (the social ties between recruiters and their recruits are symmetric, that is, if individual a recruits b, then b would recruit a); (c) accurate self-report of how many members of the target population they know (degree); (d) the network must be sufficiently large that sampling without replacement can be treated as if it is equivalent to sampling with replacement.

Previous evaluations of RDS have quantified, with simulations, the effect of deviations from assumptions (a)-(d) on the RDS estimators (Gile and Handcock 2010; Neely 2009). A handful of studies have empirically investigated the RDS assumptions about the unobserved social network on non-hidden populations with known characteristics (Wejnert and Heckathorn 2008; Wejnert 2009; McCreesh, Frost, Seely et al 2012), with simulation on real-world network data sets (Goel and Salganik 2010) and with limited information on participants' recruitment behavior (Iguchi et al. 2009). Collectively these studies have shown that (1) violation of the characteristics of the underlying network and referral process assumptions can lead to considerable bias in the RDS estimates (Tomas and Gile 2010; Gile and Handcock 2010; Iguchi, Ober, Berry et al. 2009; Neely 2009; Wejnert 2009; McCreesh et al. 2012) and (2) the structure of real-world social networks may deviate so much from the idealized model assumed by RDS that the variance in population estimates may require sample sizes nearly ten times what has previously been assumed (Goel and Salganik 2010).

Despite increased interest and significant investments by CDC and the international public health community in RDS (Lansky, Abdul-Quader, Cribbin et al. 2007; Malekinejad, Johnston, Kendall et al. 2008), most RDS empirical studies are promotional. RDS assumptions have not been examined empirically among hidden or hard-to-reach populations with the result that we have very few empirical evaluations of the effectiveness of this sampling approach at keeping its promise of representation of hidden and hard to reach populations.[2]

In the context of FSWs in China, one can think of several reasons why the SPPD assumption may fail to be true, and why it may fail to yield valid estimates of population characteristics. First, one knows a priori that, in addition to degree, other factors may influence whether or not members of the population are included in the sample. Since sampling of female sex workers might start in a venue, an individual's venue attendance is likely to influence the probability of being recruited into the sample. Second, because it is difficult to accurately assess one's degree (McCormick, Salganik and Zheng. 2010), the second stage of the approximation of inclusion probabilities can potentially be quite coarse. Third, the chain-referral process may become trapped in a particular venue or tier of sex work overloading the sample with members of certain tiers of sex work or with respondents originating from a small number of venues. The RDS assumption of non-preferential recruitment of participants (assumption (a) above) constrains researchers from directing the referral process towards members of the population who are likely to be missed. The inability to redirect the chain referral process is a significant restriction and can prove particularly problematic for potential respondents to be recruited into the sample and for RDS to satisfactorily cover and represent the underlying hidden population.

[2] A few empirical evaluations of RDS to date have been performed on *non-hidden* populations with known characteristics (McCreesh et al 2012; Wejnert and Heckathorn 2008; Wejnert 2009).

Here, we take advantage of features of RDS surveys which are not typically utilized or collected in standard RDS protocols to drive the simulation of the likely network from which our observed RDS is drawn. We then run recruitment chains over this simulated population consistent with the observed RDS. We estimate population coverage rates by comparing population proportions and RDS sample proportions.

Data

The data come from a recent study of female sex workers in Liuzhou, China, the PLACE-RDS Comparison Study. This study was designed to compare two samples of female sex workers using two distinct approaches for hard to reach populations: RDS and PLACE (Priorities for Local AIDS Control Efforts). PLACE is a venue based sampling approach which focuses on the systematic identification of places where people meet new sexual partners and assesses HIV prevention coverage programs in those places (Weir, Hileman, Khan et al. 2005; Weir, Pailman, Mahlalela, et al. 2003; Weir, Tate, Zhusupov and Boerma. 2004) with the overall aim of gauging the true levels and distribution of a syphilis infection among female sex workers. The location of the study, Liuzhou, is a city with high levels of syphilis infection among high risk groups.

There are a number of unique features of these data that will inform our analyses. One feature is that two different sampling approaches have been used to sample the same population. Another feature is that participants were not only asked who they invited to participate (as is inherent in RDS), but also who they did not invite and why.

PLACE and RDS were conducted between November 2009 and January 2010. RDS recruited 583 participants. Eligibility for participation in the study was being at least 15 years old, first time participant and self-identified as a sex worker by responding affirmatively to the question: "Have you exchanged sex for money in the past month?." Seven seeds were recruited and 576 peer recruits were interviewed. Participants were given two coupons to recruit other participants but this number was reduced to one coupon as the desired sample size was approached. All except one of the seven seeds recruited other participants. The six productive seeds generated between 9 and 20 waves of recruitment. 310 out of 583 respondents were recruiting participants, while the remaining 273 did not recruit any participant, mostly because they were the terminal nodes of the branches of the recruitment trees. A primary incentive was provided for participation in the main survey interview and a secondary incentive for successfully recruiting other participants.

PLACE was implemented concurrently with RDS with some modifications over the standard PLACE protocol which were designed to recruit a large enough sample of female sex workers for comparison with the RDS sample. Respondents in PLACE, both venue staff and patrons, were drawn from within venues selected from a sampling frame of 971 unique venues based on information provided by 400 community informants. Names, addresses and GPS coordinates were collected for each venue. The final list of venues was selected according to a multi-stage stratified random sampling scheme of venues with strata formed according to the number of times a venue was cited by informants, the estimated number of sex workers working at the

venue and rural and urban location. Of the 971 venues named, 385 were selected for a venue visit and 64 venues (41 in urban districts and 23 in rural counties) were ultimately selected for workers' interviews of which 45 (27 urban and 18 rural) were in operation and agreed to participate. 680 workers were interviewed at these sites One-fourth of the female workers reported ever receiving cash or gifts in exchange for sex and 18.2% of the female workers (n=161) had done so in the last four weeks, thereby meeting the study definition for sex worker. FSWs had a lower age at first sex, lower education, more arrests and more sexual partners than other female workers (Weir, Merli, Li et al. forthcoming).

Participants in both surveys were asked about their individual demographic and socioeconomic characteristics as well as detailed questions on their sexual risky and preventive behaviors, health status, STD symptoms and exposure to HIV/AIDS prevention programs. In both RDS and PLACE studies, personal network size was measured with the question: "How many female sex workers do you know in this city? By knowing, I mean: you know their names and they know yours and you have met or contacted them in the past month." Both RDS and PLACE participants were also invited to provide blood samples for rapid syphilis test screening.

The Liuzhou RDS protocol included a social network module, previously piloted in an RDS study of FSWs in Shanghai (Merli, Neely, Tian et al. 2010; Yamanis, Merli, Li et al. 2011). When recruiting participants returned to the interview site to collect their secondary incentives, they were administered a brief follow-up questionnaire about their invited and non-invited alters. Invited alters were members of recruiters' networks invited to participate who accepted or rejected the invitation. Non-invited alters were members of recruiters' networks who were not invited to participate in the study. RDS recruiting participants were asked about attributes of their network alters and properties of their social ties with them. Because it was determined in advance that recruiting participants were not able to differentiate between attributes of alters they did not invite, attributes of these contacts and relation properties were treated as aggregate qualities of the group. Recruiters were allowed to select multiple options for each question to describe network alters whom they did not invite.[3]

Table 1 describes the information available in each of the two data sources which will be used to drive the simulations of the underlying network.

[3] The procedure to assign attributes of each recruiting participant's set of non-invited alters to individual characteristics categories is described in Yamanis, Merli, Neely et al. 2011; Merli, Moody, Li et al. 2012).

Table 1. Network data in the Liuzhou PLACE-RDS Study

			Recruiting participants' reports about invited and uninvited alters	Recruited participants' reports about their recruiter	Participant's report about self	
			RDS	RDS	RDS	PLACE
Individual network items	Individual attributes	Age	√	√	√	√
		Marital status, education	√	√	√	√
		Where/how solicits clients	√	√	√	√
		Condom use	√	√	√	√
	Properties of relation	Place where usually meet alter	√	√		
		Frequency of contact	√	√		
		Type	√	√		
		Intensity	√	√		
	Repertoire of relation	Reason why you invited this person?	√			
		Reason why you did not invite these person(s)	√			
Aggregate network items		# of known FSWs			√	√
		# of known FSWs you would invite to the study			√	
		# of known FSWs who solicit clients at your main site			√	√
		# of known FSWs who solicit clients elsewhere by site type and distance			√	√
		# of known sex workers who are also known by your recruiter			√	
Venue where solicits clients		Name			√	√
		Type			√	√
		Physical location (latitude and longitude)			√	√

Weir, Merli, Li et al. (forthcoming) compared RDS adjusted sample proportions with PLACE weighted sample proportions of demographic characteristics (e.g. type of venue, tier of sex work, age, marital status, education), risk behaviors (e.g. number of clients, frequency of condom use), and syphilis infection. RDS respondents were older than PLACE respondents, reported a higher monthly income from sex work, more RDS respondents reported consistent condom use, and fewer tested positive for a biomarker of syphilis infection. Because no rural seeds were selected for RDS and the RDS recruitment chains were not able to cross over to rural areas of Liuzhou, significantly fewer RDS respondents than PLACE respondents solicited in rural areas. No statistically significant differences in demographic characteristics and syphilis infection were found after limiting the comparison to urban areas, although the study was not powered to assess differences between rural and urban subgroups.

The rationale for this comparison was that since PLACE has the advantage of better known strategies to establish a sampling frame of venues and of known sampling weights, any disparities in characteristics may reveal differences in sample coverage of the underlying population and/or the ability of RDS to make valid inference from the sample to the population. It would be unwarranted, however, to say that the PLACE methodology provides a gold-standard against which to evaluate RDS. Furthermore, because there is no sampling frame of the FSW population, one does not know whether there are pockets of the population missed by both sampling approaches. For example, we already know that PLACE only covered physical venues while RDS was able to recruit non-venue based FSWs who solicited by phone or through the internet (Weir, Merli, Liet al. forthcoming; Li, Merli, Weir et al. 2012). In the simulations of the FSWs network we describe below, we will consider varying parameters across the simulations to provide a possible range of the distribution of characteristics of the true population against which the performance of the observed RDS chains can be assessed.

These preliminary findings provide further motivation for exploring the networks accessed by each of the two methods and identify any difference.

Methods

To understand the properties of the realistic network in this population and of the RDS chain referral process operating on the network and to gauge population coverage rates, we use a simulation approach to build the hidden population of FSWs, their social network structure and the RDS referral process. Because it was already shown that the Liuzhou RDS study mainly covered urban areas with very limited recruitments of rural FSWs, the simulations are driven by inputs drawn from the urban subsamples of the PLACE and RDS surveys.

First, we construct a base population of FSWs grounded in the Liuzhou-Place empirical sample. We begin by bootstrap sampling 500 venues from the PLACE dataset. For each venue, we know its physical location, tier (high, middle and low), as well as the distribution of age, education and marital status for its FSWs. We use 500 venues as this was the estimate provided by the local CDC office. We then sample FSWs from the PLACE dataset. Each sampled FSW is placed into a venue probabilistically, based on district, age, education, tier and size of venue. This maintains the correlation between demographic characteristics, tier, and physical location. The size of the FSW population is assumed to be 7,500, although this number could be varied in future analyses.

7,500 is an initial estimate of the true FSW population count based on estimates of the Liuzhou CDC.

Second, we will construct a social network over the observed population that is consistent with the social network information obtained through the Liuzhou RDS and PLACE surveys. While fixed attribute models for simulations that match exactly on given parameters have been used effectively (Newman, Strogatz and Watts 2001; Bearman, Moody and Stovel 2004; Merli, Moody, Mendelsohn and Gauthier 2012), a more flexible way to generate networks is to use the exponential random graph model (ERGM) framework. This is especially true when the network simulation is highly constrained (so that a large number of demographic dimensions inform the simulation). Given a population with known demographic distributions, we can use the STATNET (Handcock et al. 2003) simulation routines to generate networks from a pre-specified model. This has been used successfully for prior network simulation models in health (Morris, Kurth, Hamilton et al 2009), and provides an extremely flexible foundation for network population simulation.[4]

Of key interest for the purpose of this application, is that one can move directly from the statistics describing mixing in the observed data to equation-based probability models of the network. Thus, one can generate a network consistent with the social contacts between demographic groups observed in the empirical data. Specifically, we estimate the probability of a social tie between individuals with different age, tier, and marital status using the RDS social network module. We estimate the effect of physical distance on the probability of a tie from a separate question in the RDS survey. RDS respondents were asked how many FSWs they know in the same venue, less than 10 minutes walk away and more than 10 minutes walk away. This information is used to estimate how the probability of a tie decreases as physical distance increases. The baseline, relative to chance, comparison comes from the 7,500 FSWs seeded from the PLACE dataset (representing our population). We thus estimate the effect of demographic (or physical) distance on a social tie relative to chance expectations. The models are estimated using the case control approach discussed in McPherson, Smith and Smith-Lovin (2012) as well as Smith (2012). We then take those estimates and generate a network consistent with the observed mixing rates. In addition to terms for age, tier, marital status and physical distance, the simulated network also maintains the empirically observed degree distribution, as well as the correlation between demographic characteristics and degree.

The third step is to simulate the RDS referral process over the network. We begin by replicating[5] our observed referral chains on the simulated population. The first question is how well the observed chain covers the physical and social distribution represented in the simulated network. The analysis will be repeated multiple times to generate a distribution of RDS chains. The distribution of chains can be used to explore sampling variability in the RDS sample. To assess coverage and representativeness of the RDS sample, we will compare sample and simulated population proportions. We will also vary parameters across the simulations to provide a possible

[4] Although the ERGM framework to generate networks comes at the cost of computational efficiency, our goal in this application is not to simulate very large networks, but smaller, highly likely networks underlying the RDS referral chains, with edge-dependent features.

[5] Replication is stochastic, moving among nodes in the simulated population that are equivalent to the choices observed in the data.

range of the distribution of characteristics of the true population against which the performance of the observed RDS chains can be assessed. We will then compare the observed RDS chain to a chain run under ideal conditions, where all of the assumptions of RDS are met. Practically, this amounts to a random walk through the simulated network, where individuals randomly give out coupons to their friends. We will quantify any referral bias by comparing estimates from this idealized model to those based on the observed chain referral process.

Our final challenge will be to establish participants' recruitment rules and incentives that can remove as much of the referral bias as possible, so that the sample recruited accurately represents the hidden population. For example, in contrast with the assumption of nonpreferential recruitment posited by RDS, previous analyses of the Liuzhou RDS data (Merli, Moody, Li et al. 2012) suggested that RDS middle-tier respondents exhibit strong preferences in recruitment peers from this same tier. Thus, we may discover that population coverage is improved if we can provide incentives to move the chain out of local pockets of the social space (e.g. out of one single venue or tier of sex work). In general, the sampling adaptations will take the form of probabilistic encouragements to distribute coupons in a way that better represents the network composition. If, in fact, we find that the best RDS model is still insufficient to cover the population under the network structure parameters we observe, we will explore alternative data collection techniques that fall outside current RDS practices (such as varying coupon distribution, choice of seeds, etc.).