From theoretical to real-world networks: Using Empirical Samples of Female Sex Workers in China to Evaluate Respondent Driven Sampling[1]

M. Giovanna Merli[ab], Jim Moody[b], Jing Li[ac], Jake Fisher,[b] W. Whipple Neely,[d] Sharon Weir,[e] and Xiangsheng Chen [c]

(a) Duke University Sanford School of Public Policy

(b) Duke University Department of Sociology

(c) China National Center for STD Control

(d) Independent Consultant

(e) Department of Epidemiology, School of Public Health, University of North Carolina Chapel Hill

Introduction

This paper evaluates Respondent Driven Sampling (RDS) by moving considerations regarding real-world network referral processes from the theoretical to the empirical realms in the context of a hard to reach population, female sex workers in China.

RDS is an increasingly popular chain-referral sampling method used to recruit samples of hidden and hard-to-reach populations. It seeks to provide a probability-based inferential structure for representations of population groups with status characteristics that are not likely to be revealed by omnibus survey research because they are rare and socially stigmatized and/or illegal. 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 is potentially quite valuable. It is able to reach enhanced coverage of the hard-to-reach pockets of the population of interest, to encourage study participation (by exploiting social ties

[1] The analyses described here are funded by NICHD grant 1R01HD068523 to Duke University (Merli, PI). Funding for the China PLACE-RDS Comparison Study 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.

within the target population) and to efficiently and cost-effectively recruit large numbers of survey respondents in a relatively short amount of time (Robinson et al. 2006; Kendall et al. 2008; Qun et al. 2008; Johnston et al. 2006).

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), or 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 on the RDS estimators of deviations from assumptions (Gile and Handcock 2010; Neely 2009). Few empirical evaluations of RDS assumptions have been produced with simulations on a population with *known* characteristics (Weijnert 2009; Weijnert and Heckathorn 2008), or on real-world network data sets (Goel and Salganik 2010). 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 (Neely 2009; Gile and Handcock 2010; Wejnert 2009) 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 significant investments by CDC and similar organizations in RDS, its assumptions have not been examined empirically among hidden or hard-to-reach populations with the result that we don't have empirical evaluations of the effectiveness of this sampling approach at keeping its promise of representation of hidden and hard to reach populations.

In this paper, we use a combination of data sources recently collected among female sex workers in China to observe RDS real-world network referral processes. In the context of sex workers 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 et al. 2010),

the second stage of the approximation of inclusion probabilities can potentially be quite coarse. Third, in the particular case of FSWs in China, the chain-referral process may become trapped in a particular venue or tier of sex work. 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 the ability of a potential respondent to be recruited in the sample and for RDS to satisfactorily cover and represent the population.

Here, we take advantage of features of RDS surveys which are not typically utilized or collected in standard RDS protocols to (1) evaluate the reliability of self reports on personal network size and on characteristics of personal networks; (2) map RDS respondents' recruitment chains and characterize the mixing patterns of recruitment; (3) model the recruitment process through dyad-level logistic choice models of recruitment to characterize the mixing patterns of recruitment and identify sources of bias in RDS recruitment; (4) quantify the amount of bias in the RDS estimates with a new bootstrap methodology used to approximate the distribution of RDS estimates under various recruitment scenarios consistent with the data.

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 et al. 2003, 2004, 2005) with the overall aim of gauging the true levels and distribution of a syphilis infection among female sex workers. The location of the study was a city with high levels of syphilis infection among high risk groups.

PLACE and RDS were conducted simultaneously 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 simultaneously to RDS with some modifications over the standard PLACE protocol (Weir et al. 2005) 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 972 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. This sampling design yielded 42 venues in urban districts and 21 in rural districts, a list that is considered representative of the 972 venues cited by informants. All female staff and female patrons who self identified as sex workers were selected at the urban sites and up to five self-identified sex workers were selected at the rural sites for a total of 683 female staff and 227 female patrons. 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.

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 the RDS study, 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.

Our RDS protocol included a personal network data module. 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 noninvited 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 the pilot phase of the study revealed 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. This information can be used to characterize the patterns of recruitment of RDS participants and the recruitment biases entailed in the RDS recruitment processes.

Table 1 describes the type of network data available in each of the three data sources which will be used to describe the patterns of recruitments of RDS participants and the underlying network

An additional advantage of our PLACE-RDS protocol is that it also provides the means to independently assess respondents’ degree and recruiting participants’ networks and corroborate their reports on objective individual attributes as well as on subjective attributes of the relationship. RDS participants were asked venue names and addresses of the places where they had solicited the four most recent clients. This information can be linked to the venue information in PLACE to identify the venues implicated in both arms of the RDS-PLACE study. With this information, we will be able to independently map the social networks of RDS participants. In addition, questions on the number of female sex workers known to respondents by type and geographic location of the venue were asked of both RDS and PLACE participants. For network alters not at the respondent’s venue, information was collected on the type of site where these alters solicited clients and distance of the site relative to the site of the respondent.

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

			Recruiting participants' reports about recruited and non-recruited 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 address			√	√
		# of known sex workers who are also known by your recruiter			√	
Venue where solicits clients		Name			√	√
		Type			√	√
		Address			√	√

Methods

Our analyses will proceed as follows:

(1) ***Evaluate the reliability of degree self reports and of self reports about characteristics of one's network*** by comparing respondents' self reports and objective information on the size and attribute composition of RDS respondents' personal networks gleaned from PLACE data on venues and venue respondents.

(2) ***Characterize the mixing patterns of recruitment among FSWs and identify sources of bias in RDS respondents' recruitment of network alters***. Since we know the characteristics of recruited alters, those invited who refused to participate, and those who were not invited from recruiters' self reports, we can compare invited and non-invited alters to identify any over/under representation of particular mixing combinations. For this comparison, we will carry out two distinct types of analyses. First, at the aggregate descriptive level, we will compare the two groups with respect to socially relevant characteristics (e.g proportion by type of venue). Second, since we know the size of the personal network of each RDS respondent, the attributes of those invited and the distributions of attributes for those non-invited, we will construct actor-level datasets from this information to build records that are consistent with the observed distributions. Once these distributions are created, we will construct attribute-based mixing matrices (counts of the number of people of a row category linked to numbers of a column category) for both the referral chains and the all-alter networks, and ascertain how similar they are. If recruitment were non preferential, the mixing odds based on the RDS chain should match those from the all-alters information. At a more micro inferential level, we can build a dyad-level logistic choice model of recruitment over all alters that has the form:

$$P(j \text{ invited by } i) = \alpha + \beta 1 Xi + \beta 2 Xj + \beta 3(sim(Xi,Xj)) + \varepsilon i$$

where *i* indexes each individual and *j* the people in their local network they could have invited, and the dependent variable is the choice of offering a coupon or not. α represents the covariate-adjusted recruitment rate. $\beta 1$ represents respondent characteristics telling us if respondents of acertain type are more/less likely to recruit. $\beta 2$ are alter characteristics in the first column of Table 1 that tell us how much more/less likely particular sorts of alters are to be chosen. $\beta 3$ represents dyad-specific effects on recruitment that capture "homophily" effects based on the similarity of each actor.[2] Any significant similarity (sim) coefficients represent preferential recruitment. The simple dyad-level logistic regression model does not account for dependence of recruiter cases (the *i*'s in the above equation) or the limited choice imposed by the coupon distribution schedule. To address these complications, we will estimate a limited-choice, network autocorrelation form of this model.

(3) Gauge the impact of recruitment bias on the RDS estimates with a new bootstrap procedure which uniquely incorporates the sample's observed branching structure and is applied to study the impact of alternative ego-centric network compositions on the RDS estimates.

[2] Model identification issues are simplified since alters are not repeated across respondents. In fact, unlike in global network data where all *j's* are also *i's*, the similarity terms need not be sum functions of the node-level terms.