Appendix to "An empirical analysis of the impact of recruitment patterns on RDS estimates among a socially ordered population of female sex workers," by Thespina Yamanis, M. Giovanna Merli,[a] W. Whipple Neely, Felicia F. Tian, James Moody, Xiaowen Tu and Ersheng Gao, *Sociological Methods and Research* 2013

(a) corresponding author: M. Giovanna Merli, Sanford School of Public Policy and Department of Sociology, Duke University giovanna.merli@duke.edu

Bootstrap Procedure & Confidence Interval Estimation

The method we use for computing confidence intervals is based on the Markov model and sampling probability model used throughout the RDS literature (Salganik 2004). As such it is a variant of the bootstrap procedure developed by Salganik (2006), with the additional advantage that it makes it possible to explore the impact of preferential recruitment on RDS estimates produced with any of the RDS estimators. In this procedure one defines a resampling scheme that can be used to create a collection of samples that, according to the model, have a distribution similar to the actual sampling process that generated the data. To make this concrete, suppose we wish to estimate the proportion of the population who are in some subgroup A (say, for example that A represents members of the population who test positive for HIV, or some other sexually transmitted disease). In this situation we wish to estimate both the proportion of the population in A, and (in order to examine the statistical significance of the result) we need to compute a confidence interval. Salganik's (2006) procedure provides a recipe for computing such a confidence interval as follows. First, construct the following algorithm for computing a single bootstrap estimate:

1. Divide the sample into two subsets: A[rec] consisting of individuals recruited by members of A, and B[rec] consisting of individuals recruited by members in the complement of A.
2. Select a "bootstrap seed" from the data by selecting an observation of the sample at

random (i.e. make a random draw from the sample by selecting from amongst the observations so that each observation has equal probability of being drawn).

3. Starting with the bootstrap seed select a new observation by selecting at random (with equal probability) from either A[rec] or B[rec] depending on the group membership of the bootstrap seed. Continue this process recursively until one has selected a sequence of observations of the same size as the original sample. One then has a single bootstrap sample whose size is identical to the original data set. The data in the sample consist of an observed group membership and self-reported network size. A set of data constructed in this manner will be called a bootstrapped data set, or bootstrapped data for short.

Next we use 1-3 above to create many bootstrapped data sets. Thus the next phases of the algorithm are as follows.

4. Repeat 1-3 above until one has many versions of the bootstrapped data (i.e. one has many artificial data sets, each constructed by the procedure above).
5. To compute a confidence interval for an estimator of p[A] [in principal any estimator, though, at the time that Salganik was writing (2006), Volz and Heckathorn's V-H estimator (2008) had not be developed yet], apply the estimator to each member of the collection of data sets above to yield a collection of bootstrapped estimates. This collection can be used to compute confidence intervals by either computing the variance of the bootstrapped variance (which is what Salganik (2006) does) or by reporting 95% central quantiles of the bootstrapped estimates. The latter strategy has the advantages that it is a standard approach in the bootstrap literature (Davison and Hinkley 1997) and because it automatically provides intervals that are constrained to be within the interval [0,1] as is natural for a proportion estimate.

In order to examine the implications of using this procedure to compute confidence intervals, we recast this algorithm in a mathematically equivalent form that describes the model under which groups and degrees are simulated for the bootstrapped data. First, in order to simulate group memberships, we approximate the probability of transitions within and between the groups A and B by constructing a Markov transition matrix

$$\hat{P} = \begin{bmatrix} n_{AA}/(n_{AA}+n_{AB}) & n_{AB}/(n_{AA}+n_{AB}) \\ n_{BA}/(n_{BA}+n_{BB}) & n_{BB}/(n_{BA}+n_{BB}) \end{bmatrix}$$

where n_{A,B} is the number of observed transitions from group A to group B seen in the original data. The terms n_{AA} , n_{BA} and n_{BB} are similarly the number of observed transitions from the group indicated by the first subscript to the group indicated in the second subscript. This matrix gives the exact probabilities of the transitions between groups under Salganik's (2006) bootstrap. In terms of statistical modeling, $\hat{P}$ is the maximum likelihood estimate for the transition probabilities under a first order Markov model for the group transitions observed in the sample (see Anderson and Goodman 1957) for classical material on inference under this model, see either Volz and Heckathorn (2008) or Goel and Salganik (2010) for detailed discussions of $\hat{P}$ in the context of RDS). In Salganik's (2006) bootstrap, once groups have been simulated, we can simulate degrees by making a random draw from the observed degrees for the appropriate group. The entire process of creating a single bootstrapped data set can be described as follows.

1. Select a seed, y_0^{boot} by making a random draw from the observed sample, thus y_0^{boot} will be in A with probability $n_A/(n_A+n_B)$ and in B with probability $n_B/(n_A+n_B)$ where n_A and n_B are the number of observations in the sample from groups A and B respectively.
2. Select y_1^{boot} through y_n^{boot} iteratively by using the transition probabilities determined by $\hat{P}$

. In other terms, if y_i^{boot} is an A, then y_{i+1}^{boot} will be an A with probability $n_{AA}/(n_{AA}+n_{AB})$ and in B with probability $n_{AB}/(n_{AA}+n_{AB})$.

3. After selecting y_1^{boot} through y_n^{boot} , select bootstrapped degrees d_1^{boot} through d_n^{boot} by selecting d_i^{boot} randomly from the observed degrees in the group corresponding too y_i^{boot} . Again, to be concrete, if y_i^{boot} is A then we select d_i^{boot} by making a random draw from the observed degrees for group A. If y_i^{boot} is B then we select d_i^{boot} by making a random draw from the observed degrees for group B.

The description above can be summarized quite concisely by saying that the Salganik (2006) bootstrap (i) models group membership (in A or B) as a linear first order Markov chain of length n with transition probabilities $\hat{P}$ and (ii) models degrees as conditionally independent given group membership. Before we discuss the potential shortcomings of this approach we briefly describe the modeling assumptions behind the other variance estimation approach currently in use with RDS data.

There are two features of the above bootstrap procedure that are worth noting. First, the above procedure can be used to estimate the sampling distribution of any RDS estimator. This is because the bootstrap method is primarily a procedure for creating bootstrapped data. Thus, one uses the procedure to construct a large number of synthetic data sets whose distribution, one hopes, matches the sampling distribution of the actual RDS process. Then, in order to estimate the sampling distribution of a population estimator, one applies that estimator to each of the bootstrapped data sets in turn in order to create a large sample of bootstrapped population estimates. Consequently, one can apply this approach to any RDS estimator, including the V-H estimator developed by Volz and Heckathorn (2008).

The second feature that is worth noting is that in Salganik's (2006) procedure there are two factors that clearly influence the ability of the bootstrap to approximate the actual sampling distribution. The first is that one replaces the branching observations of the RDS sampling process with a linear chain. The second is that one samples the entire data set when selecting seeds. With regard to the first factor, one would expect that a bootstrap method that uses the same branching structure as the data collection process would do a better job of replicating the sampling distribution of RDS with the additional advantage that using the observed branching structure avoids underestimating variance which results from treating dependent observations as independent. With regard to the second factor, the seeds are drawn by the researcher from an accessible stratum of the target population and are surely not distributed in the same manner as the actual RDS sample. As a result, in our implementation of the bootstrap procedure (1) we have used the observed branching structure of our sample, rather than a linear structure – this is tantamount to treating the branching structure as a fixed feature of the survey design; (2) we have treated seeds as a fixed aspect of the sampling design because they are selected by the researcher. Were seeds selected at random, random selection of seeds should be incorporated into the bootstrap procedure. Thus our bootstrap algorithm can be described as follows:

1. Select bootstrapped seeds, y_0^{boot} by making these identical to the observed seeds.
2. For each wave in the data set, select member y_i^{boot} by using the transition probabilities in $\hat{P}$ and the value of the recruiter y_j^{boot} where j is the recruiter of alter i in the original data. In other terms, if y_j^{boot} is an A, then y_i^{boot} will be an A with probability $n_{AA}/(n_{AA}+ n_{AB})$ and in B with probability $n_{AB}/(n_{AA}+ n_{AB})$. In other terms we select y_i^{boot} based on the observed network structure.
3. After selecting y_1^{boot} through y_n^{boot} , select bootstrapped degrees d_1^{boot} through d_n^{boot} as

before. Thus we select d_i^{boot} randomly from the observed degrees in the group corresponding to y_i^{boot}. Again, to be concrete, if y_i^{boot} is A then we select d_i^{boot} by making a random draw from the observed degrees for group A. If y_i^{boot} is B then we select d_i^{boot} by making a random draw from the observed degrees for group B.

In order to obtain confidence intervals, we use the above method to simulate 100,000 bootstrapped data sets and apply the V-H estimator to each of these yielding 100,000 bootstrapped population estimates. The confidence intervals reported are thus the central 0.025% to 0.975% quantiles of the t bootstrapped population estimates.

Extending the bootstrap procedure to assess the impact of recruitment biases on RDS estimates

The bootstrap procedure described above can be used for studying the impact of different ego-centric network compositions, or more realistically different estimates of ego-centric network compositions. Recall that one of the key features of the bootstrap procedure is the matrix, $\hat{P}$, of estimated referral probabilities determined by the observed inter and intra group recruitments. This matrix gives an estimate (based on the observed RDS sample) of the probabilities of inter and intra group recruitments. If we accept the RDS assumption of non-preferential recruitment, then this matrix represents the maximum likelihood estimate of respondents' recruitment choices under the Markov model. However, in the presence of recruitment biases, the actual recruitments will not represent the egocentric network compositions of respondents. We can use alternative information on network composition to determine what recruitment probabilities would be under alternative mixing scenarios. In this paper, alternative information on network composition consistent with recruitment patterns that

would have resulted if respondents had recruited according to their self-reported network composition and with respondents' invited network composition was used to construct alternative versions of the recruitment probability matrix $\hat{P}$. We then used these alternative matrices in the bootstrap procedure to examine the impact of different network compositions on the resulting V-H estimates. In particular, this was done by replacing the actual network structure used to select y_i^{boot} with either the invited or the all-alter network structures. We then used the bootstrap procedure to generate samples that, according to the reasoning behind the RDS methodology, would have arisen under these alternative egocentric network structures. One main advantage of this method is that instead of focusing on a particular RDS estimator, we emphasize the process that is assumed to generate the RDS sample. As a result this method can be used to study the impact of recruitment biases on any RDS estimator.

References

Anderson, Theodore W. and Leo A. Goodman. 1957. "Statistical Inference about Markov Chains." *The Annals of Mathematical Statistics* 28(1):89–110.

Davison, Anthony C. and David V. Hinkley. 1997. *Cambridge Series on Statistical and Probabilistic Mathematics: Bootstrap Methods and their Application.* New York: Cambridge University Press.

Goel, Sharad and Matthew J. Salganik. 2010. "Assessing Respondent-Driven Sampling." *Proceedings of the National Academy of Sciences of the United States of America* 107(15):6743-6747.

Salganik, Matthew J. 2006. "Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling." *Journal of Urban Health* 83(s1):98-112.

Salganik, Matthew J. and Douglas D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34:193.

Volz, Erik and Douglas D. Heckathorn. 2008. "Probability Based Estimation Theory for Respondent-Driven Sampling." *Journal of Official Statistics* 24(1):79.