2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Preserving Multiple Homophilies in a Network Configuration Model

Derek Lopez, Bhuvaneshwar Mohan, Lyric Boone, and John Matta

Abstract— **Respondent-driven sampling (RDS) is a popular method for surveying hidden populations based on friendships and existing social network connections. In such a survey the underlying hidden network remains largely unknown. However, it is useful to estimate its size as well as the relative proportions of surveyed features. The fact that linked network participants are likely to share common features is called homophily, and is an important property in understanding the topology of social networks. In this paper we present a methodology that scales up RDS data to model the underlying hidden population in a way that preserves multiple homophilies among different features. We test our model using 46 features of the population sampled by the SATHCAP RDS survey. Our network generation methodology successfully preserves the homophilic associations in a randomly generated Barabasi-Albert network. Having created a realistic model of the expanded SATHCAP network, we test our model by simulating RDS surveys over it, and comparing the resulting sub-networks with SATHCAP. In our generated network, we preserve 85% of homophilies to under 2% error. In our simulated RDS surveys we preserve 85% of homophilies to under 15% error.**

I. Introduction

Hidden populations such as drug users or people infected with HIV are difficult to detect via traditional statistical sampling methods, both because their absolute numbers are small, and also because of the stigma associated with identifying as a member of these groups. Public health authorities need to know the size and composition of these hidden populations to plan and execute interventions to protect them and those who interact with them. One way of obtaining data is through respondent-driven sampling (RDS) [1]. RDS has been used to capture data on a variety of targeted hidden populations, including groups of MDMA users [2], jazz musicians [3], and migrant populations [4].

RDS utilizes a phenomenon known as homophily. Homophily is the idea that individuals with similar behaviors in similar populations will regularly interact with one another. An RDS survey begins by recruiting seed nodes [5], which are individual members of the target population. These seed nodes are then asked to recruit more people from the target population, often by handing out survey coupons to friends or acquaintances.

An RDS survey represents only a fraction of a hidden population, and it is often desirable to know the population's true size. Scale up methods can be used to estimate a population's size based on a sample, providing a more accurate estimate when looking at the social anomalies within a network. There are different methods for scaling up networks, such as the service multiplier method [6] or the NSUM generalized scale-up estimator [7]. These methods yield size estimates for hard-to-reach populations, and often these estimates can guide researchers and medical experts in prevention and treatment efforts.

RDS is incompatible with scale-up methods that rely on random samples, which an RDS survey does not produce. However, one method that has shown positive results with RDS data is successive-sampling population size estimation (SS-PSE) [8]. SS-PSE relies on changing trends across successive waves of recruitment to estimate the fraction of the total population that was captured by the RDS survey. SS-PSE can be used to estimate the total size of the network without the need for outside studies or other data. [9].

In this paper we demonstrate a Monte-Carlo algorithm that preserves homophilies in a scaled-up version of an RDS network, using the estimation of the size of the hidden population to create a model that lets us evaluate the relative importance and distribution of the homophilic features. The algorithm successfully deals with both extremely rare and common features, such as those occurring in almost all or almost none of the population, and is able to handle large and complex sets of homophilies between nodes with many simultaneous features, occurring at frequencies that are present in real-world data.

It is difficult to know how well an RDS survey reflects the homophilies of the underlying network. We test this empirically by simulating an RDS survey over our generated scaled-up network. We show that our algorithm preserves multiple homophilies well. We also show that while individual RDS samples add considerable noise the the homophily measurements, the mean homophily over many RDS samples still reflects the homophily of the underlying network.

The data used in this paper are from SATHCAP [10], an RDS survey which was conducted from 2006 to 2008, primarily involving men who have sex with men (MSM), drug users (DU) and injected drug users (IDU). Participants were asked almost 1500 questions (referred to as *features*) concerning their sexual habits, drug-related habits, and demographic and other information. Participants were then given recruitment coupons and instructed to give them to individuals with whom they had participated in potential HIV-spreading behavior, such as sex or needle-sharing. The study was conducted in four cities, but in our analysis we focus on the Chicago component of the data.

The data have been obtained through the National Addiction and HIV Data Archive Program (NAHDAP), ac-

Derek Lopez, Bhuvaneshwar Mohan, Lyric Boone, and John Matta are with the Computer Science Department, Southern Illinois University Edwardsville, Edwardsville, IL 62025. Corresponding email: jmatta@siue.edu.

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

cessible online[1]. This research was conducted under the approval of the Southern Illinois University Edwardsville IRB. Code used in this paper is publicly available at https://github.com/derek200pz/homophily-config-graph.

II. Related Work

Homophily is an important and often defining property of social networks [11], [12]. Preserving social tendencies like homophilies within a network is crucial because these hidden populations can contain important information such as disease reservoirs, which are groups with similar traits being affected by a disease [13]. Observing data about disease reservoirs can aid in identifying communities with higher risk of developing those specific diseases [14]. Additionally, studying homophilies can help us further interpret how these diseases are spread.

RDS was introduced as a way to create samples of hidden populations that are externally valid, i.e. reasonably representative of the entire hidden population. RDS was proposed as a solution to known biases in other sampling methods that target hidden populations [15]. Prior to Heckathorn's proposal of RDS, snowball sampling was widely used for similar purposes. While similar in methodology to RDS, snowball sampling does not produce a probability sample, meaning it has little value for estimating statistics about the population as a whole [16], [17]. RDS is also useful for its ability to capture a portion of the population's network structure [18], [19], which is a property that is explored in this paper, though there is evidence that this captured network structure does not represent the structure of the population well [20].

III. Methods

A. Cleaning and Curation of Dataset

The process of cleaning, normalizing, and converting the RDS data for use with this study is discussed fully in [21], as is the process of generating a network representation of the data. The SATHCAP dataset includes data from three different collection sites, with no cross-connections between sites. For this work, we only use the data collected in Chicago, consisting of 2739 participants. Based on coupon numbers included with the data, it is possible to reconstruct the recruitment network, which consists of 132 components (resulting from 132 seed nodes), and has a largest connected component of 949 nodes. The survey consisted of 1488 questions (also referred to as features). We narrowed the data by limiting it to features missing fewer than 5% of responses. The focus was narrowed further to features that provide societal, behavioral and economic details which would affect a recruiter's choice of partners and recruits. This resulted in 46 features, which are listed in Table I. These features are expected to display high homophily. Note that a feature can have a low prevalence, but still display high homophily. Information contained in the features of interest include gender, sexual orientation, sexual behavior, education, income, living situation, drug use, and acquisition of infections such as HIV, gonorrhea, syphilis and chlamydia.

B. Determination of Population Size and Homophily

The goal of this paper is to use an RDS sample to extrapolate a larger network with a similar distribution of features and homophilies. To determine the size of the underlying population, we used the RDS Analyst software and its included SS-PSE package [22].

SS-PSE [9] uses a Bayesian inference method to estimate the total size of the population based on an RDS sample. It utilizes the distribution of the reported network sizes for respondents in the sample combined with information about how late in the recruitment chain each respondent appears. In successive waves of recruitment, hub nodes with many connections will tend to be recruited less frequently in the later waves because they have already been recruited in an earlier wave. Based on how quickly the network sizes dwindle as successive samples are taken to construct the RDS sample, it is possible to estimate what proportion of the total population has been included in the current sample, and by extension, the total size of the population.

For the SATHCAP sample, this method gave an estimate with a 90% confidence interval of [11,290 23,118], a median of 15,491 and a mean of 16,175. Based on these estimates, for the networks generated in section IV-A, we use a network size of 15,000 nodes.

Reproducing the homophilies present in the RDS sample increases the usefulness of a scaled-up generated network by better reflecting the biases portrayed by individuals in associating with others that share similar features. For this paper, RDS Analyst was used to calculate both recruitment homophily, which is the tendency for respondents within the RDS survey to recruit others with similar traits, and population homophily, which is the estimated homophily of the underlying social network.

The method used to calculate the recruitment and population homophilies for a given feature f is as follows. To describe respondents and their responses, we notate the response of a respondent x as $\lambda(x)$. A recruit is represented by β, and $\alpha(x)$ represents the recruiter of x.

For feature f, $\Omega_{\text{recruitment}}(f)$ represents the number of recruits that share the same response as their recruiter for the given feature f. This can be described as the sum of responses:

$$\Omega_{\text{recruitment}}(f) = \sum_{i=1}^{N} (\lambda(\beta_i) = \lambda(\alpha(\beta_i))|f).$$

The expected number of $\alpha(\beta)$-β-similar responses to feature f that would be received if homophily were not present in the respondent dataset is represented by $\Omega_{\text{expected-RDS}}(f)$. Recruitment homophily for the feature K is then calculated as the ratio of $\Omega_{\text{recruitment}}(f)$ to $\Omega_{\text{expected-RDS}}(f)$:

$$\text{Recruitment Homophily}(f) \equiv \frac{\Omega_{\text{recruitment}}(f)}{\Omega_{\text{expected-RDS}}(f)}.$$

[1] https://www.icpsr.umich.edu/icpsrweb/NAHDAP/index.jsp

Similarly, the population homophily is the ratio of expected pairs sharing response f in the underlying social network to the number of pairs sharing response f that would be expected in a network without homophily. We use population homophily as the target homophily for the networks we generate.

A homophily of 1 implies that the number of homophilous pairs calculated for the feature is equal to the number of expected pairs over a random network, and therefore does not hold significance. A homophily greater than 1 implies a larger tendency for pairs to share trait f than would be observed by chance, while homophily less than 1 implies heterophily, or that pairs are less likely to share attribute f than would be expected if pairs were matched randomly [22].

C. Algorithm for Configuration Network

We have developed a methodology by which RDS data can be scaled up to a representation of the underlying hidden network, preserving pairwise homophily (for multiple, overlapping feature sets) at the rate observed in the RDS sample. Our algorithm takes three parameters:

- G, a network whose nodes will be randomly assigned features and whose edges will be re-wired to attain a specified homophily for those features.
- P, a list of the frequencies at which features occur. For three features, for example, *residence.mine*, *residence.shelter*, and *education.college*, this list might be P = [residence.mine:0.372, residence.shelter:0.098, education.college:0.025].
- H, a list of homophily *targets*. For the previous example, this might be H = [residence.mine:1.048, residence.shelter:1.740, education.college:0.733].

Our algorithm begins by compiling a list F of boolean features, matching the features in P and H. Each individual node $i \in G$ is assigned its own corresponding list of true/false values F_i indicating whether it is included in that feature. For example, if F = [residence.mine, residence.shelter, education.college], a specific node might have a feature set F_i = [residence.mine:0, residence.shelter:1, education.college:1], indicating that node i does not stay at their own residence, does live in a homeless shelter, and does have a college education. A node's inclusion for each feature $f \in F$ is determined randomly with probability $p_f \in P$. We lay out the algorithm as a set of steps followed in order:

1) For each feature $f_i \in F_i$ of node $i \in G$, randomly assign f_i either a 1 with probability $p_f \in P$ or a 0 with probability $1 - p_f$
2) For each feature $f \in F$, create sets S_f and $\bar{S}_f$ of nodes that are positive or negative for feature f ($2|F|$ sets).
3) Repeat the following steps a user-defined number of times, converging the homophilies of the network to the targets in H:
    a) For each feature $f \in F$, place edges attached to the nodes in set S_f into two sets, E_f and X_f:
        - E_f contains homophilous edges (i, j), where $f_i = f_j = 1$
        - X_f contains heterophilous edges (i, j), where $f_i \neq f_j$
    b) For each feature $f \in F$, calculate the number of re-wires r_f which must be performed (and the direction, homophilous or heterophilous) to reach the target homophily.

    $$r_f = p_f h_f |E_f + X_f| - |E_f|$$

    c) If r_f is positive, randomly choose r_f edges from X_f and remove them from the network. Then, choose r_f pairs of nodes i and j such that $i \in S_f$ and $j \in \bar{S}_f$ and add the edges e_{ij} to the network.
    d) If r_f is negative, randomly choose r_f edges from E_f and remove them from the network. Then, choose r_f pairs of nodes i and j such that $i \in S_f$ and $j \in S_f$ and add the edges e_{ij} to the network.

Because the re-wiring process for one feature may add or remove edges that affect the homophily of another feature, the re-wiring must be run several times to allow the homophily metrics to converge to the desired homophily for every feature.

D. Simulation of RDS

To simulate a respondent-driven sample over the generated network, we use an approach similar to [23], but with some important differences. The algorithm used is as follows.

1) Select α seeds at random from all the nodes in the network and add them to set A.
2) Create an empty set R of *responsive* nodes.
3) Pop one node from A and add it to R.
4) Pop a random node i from R and add it to the simulated sample S. If R is empty, pop a random node from B (The list of recruited nodes that did not respond, See step 5a), or if B is empty, from A (unused seed nodes)
5) Randomly choose between 0 and 6 nodes adjacent to i for it to *recruit*.
    a) For each *recruited* node i_r, choose with probability p_r whether to add it to R. These nodes are considered to have *responded*. If a node does not respond, it is added to a list B to ensure it is not recruited again.
6) If $\frac{|S|}{\omega} \geq \frac{\alpha - |A|}{\alpha}$, pop another seed from A and add it to R (Where ω is the rds sample size goal, α is the number of seeds selected in step 1, $|S|$ is the size of the sample so far, and $|A|$ is the number of remaining unused seeds.)
7) repeat step 4 until the sample reaches its size goal ω.

Note that seeds do not all respond at the beginning of the sampling chain. Instead, one seed responds at the beginning, and the rest are added linearly as the sample grows in size. If the recruitment chain ends before the desired sample size is reached, nodes from B respond to continue the chain. This is a notable difference between our simulation and [23], in which additional seeds are added when faced with this situation.

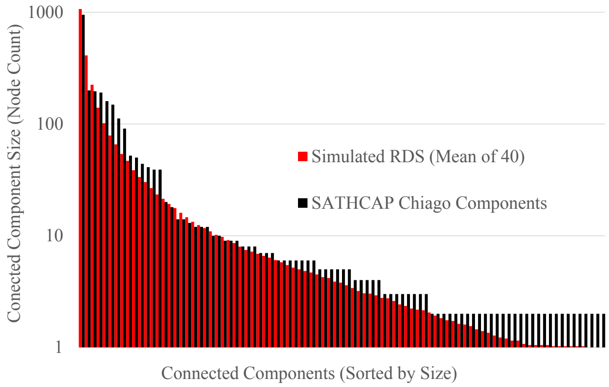

Fig. 1. Average size distribution of 40 simulated RDS samples vs SATHCAP

IV. Results

A. Homophily Configuration Networks Generated

Our algorithm was used to create 15,000-node multi-homophily configuration networks with 46 binary features. These features were one-hot encoded from 9 categorical features in the original data. We used a scale-free Barabasi-Albert graph as input, because previous work [21] suggests that the population network underlying SATHCAP is scale-free. Thirteen passes of the re-wiring portion of the algorithm were required.

For parameters P and H, we used the feature frequencies and homophilies estimated for SATHCAP's underlying social network. These values can be seen in Table I, in columns labeled "Freq." and "Target."

We changed the feature assignment portion of the algorithm to better model the population sampled by SATHCAP. In the algorithm described in section III-C, the first step is to randomly assign the features according to the probabilities in P, *independently* determining based on probability $p_f \in P$ whether node i will be positive for feature f. Our features were one-hot encoded from SATHCAP questions, and are therefore not independent. An example question is "Are you (choose one): male, female, trans male to female, trans female to male?" These responses were converted to four boolean features, *sex.male*, *sex.female*, *sex.transFTM* and *sex.transMTF*. These features should not be assigned using independent probabilities because there is a possibility that they will coincide, which is not a good model of the original SATHCAP survey data. To avoid contradictory collisions, we grouped the features by original question (shown in bold in Table I) and randomly assigned only one response to a given node, based on their calculated frequencies.

B. Comparison of Homophily Configuration Networks against Predicted Population Homophilies

Table I shows population homophily results for the 46 features we use for our homophily configuration networks. The "Target" column shows the estimated population homophily, and the "Mean" column shows the mean of the homophily (by feature) from 200 independently-generated 15,000-node homophily configuration graphs. Note that there are two features where the error is approximately 15%, four where it is approximately 4%, and for the other 40, the error is less than 2%. The root-mean-square error for all 46 homophilies (target homophily vs mean achieved homophily) is 2.99%

There is an inverse correlation between the frequency of a feature's occurrence (p_i for feature i) and the error in that feature's homophily. Both the standard deviation and the percent error in the median are higher for features such as *sex.transFTM* where the frequencies are very low.

For example, *sex.transFTM* has a frequency of 0.001, which implies that in a graph of 15,000 nodes, only 15 are transMTF. Because there are only a small number of positive nodes for the feature, every edge added or removed from these nodes creates a large change in homophily, resulting in the relatively large errors seen in the configured homophily. If you weigh the root-mean-square (RMS) error by frequency (essentially taking the RMS homophily error by node instead of by feature) the RMS error drops to 0.454%.

It is important to distinguish between homophilic and heterophilic behavior. Homophilic behavior is indicated by homophily greater than 1, and heterophilic by homophily less than 1. Feature *orientation.hetero* with a homophily greater than 1 shows that heterosexual participants did in fact recruit other heterosexual participants, whereas feature *sex.male* with a homophily less than one indicates that men have a slight tendency to recruit people who are not men. An inaccuracy that flips a homophily to a heterophily or vice versa would indicate the wrong behavior, even if the percent error were low. Thus, it is worthwhile to note that this particular error does not occur for any of the mean homophilies of the homophily configuration networks generated in Table I.

C. Simulated RDS Samples Generated

We tested our generated networks by simulating an RDS survey on them. Our simulated samples are generated using $\alpha = 132$ seeds, because we are trying to mimic the Chicago specific sub-graph of the SATHCAP network, which recruited 132 seeds. When nodes who have "responded" choose to recruit between 0 and 6 adjacent nodes, this choice is weighted to create graph components with a specific structure. The weights we use are based on the pattern of seed node recruitment in the original dataset. By using only the seed nodes, we eliminate bias toward the recruitment tendencies of the larger components.

In section III-D, it is explained that not all seeds are immediately available for expansion in the RDS simulation. The reason we choose to slowly activate the seeds this way is to better match the component size distribution of the SATHCAP network.

A comparison of our simulated RDS and SATHCAP component sizes is shown in Fig. 1. The figure shows the sizes of the largest 87 connected RDS components from SATHCAP (in black) along with the mean size of the largest

TABLE I

HOMOPHILY MEAN AND STANDARD DEVIATIONS FOR 200 HOMOPHILY-CONFIGURATION NETWORKS AND 200 SIMULATED RDS SAMPLES

Feature	Freq. (P)	Target (H)	Hom. Conf. Mean	Std. Dev.	RDS Mean	Std. Dev.
sex						
male	0.619	0.782	0.782	0.005	0.843	0.018
female	0.376	0.785	0.784	0.008	0.698	0.030
transMTF	0.004	1.192	1.153	0.202	1.135	7.086
transFTM	0.001	NULL	1.049	2.733	0.750	2.634
residence						
mine	0.372	1.048	1.051	0.011	1.047	0.039
family	0.320	1.050	1.051	0.013	1.058	0.047
partner	0.066	1.209	1.211	0.037	1.230	0.295
friend	0.056	1.060	1.061	0.038	1.090	0.359
hotel	0.045	1.153	1.150	0.048	1.177	0.457
shelter	0.098	1.740	1.741	0.043	1.737	0.230
street	0.026	1.296	1.301	0.057	1.281	0.842
other	0.018	0.965	0.970	0.061	0.991	1.133
homeless						
yes	0.386	1.132	1.132	0.012	1.242	0.040
no	0.614	1.132	1.133	0.007	1.078	0.015
education						
none	0.014	1.435	1.438	0.104	1.248	1.607
slf2	0.350	1.112	1.113	0.013	1.096	0.040
highschool	0.363	1.050	1.051	0.012	1.060	0.041
slf4	0.240	1.068	1.070	0.015	1.081	0.067
college	0.025	0.744	0.754	0.042	0.720	0.707
gradschool	0.009	0.888	0.891	0.086	0.766	2.039
work						
unable	0.261	1.119	1.120	0.015	1.209	0.066
unemployed	0.561	1.141	1.140	0.009	1.081	0.016
fulltime	0.060	0.968	0.969	0.031	1.085	0.337
parttime	0.084	1.172	1.175	0.034	1.311	0.257
homemaker	0.018	0.950	0.954	0.058	1.102	1.355
student	0.005	1.219	1.203	0.180	0.938	4.035
retired	0.011	0.478	0.491	0.047	0.447	1.483
insured						
yes	0.321	1.032	1.032	0.013	1.075	0.051
no	0.679	1.032	1.033	0.006	1.016	0.011
treated						
yes	0.707	0.894	0.894	0.005	0.946	0.014
no	0.293	0.894	0.894	0.011	0.782	0.040
orientation						
homo	0.035	1.848	1.850	0.074	2.121	0.896
bi	0.141	0.988	0.986	0.019	1.117	0.138
hetero	0.622	1.128	1.128	0.007	1.053	0.013
downlow	0.040	0.656	0.654	0.026	0.812	0.435
samegender	0.028	0.838	0.849	0.041	1.001	0.830
messingaround	0.030	0.889	0.888	0.041	0.929	0.791
transMTF	0.007	1.380	1.373	0.149	1.119	3.259
transFTM	0.004	0.596	0.724	0.259	1.604	8.049
nolabel	0.093	0.810	0.812	0.020	0.942	0.222
sexbehavior						
onlymen	0.252	0.935	0.937	0.013	0.969	0.063
mostlymen	0.147	1.019	1.019	0.018	1.019	0.117
equal	0.034	0.781	0.785	0.030	0.821	0.529
mostlywomen	0.101	1.142	1.142	0.028	1.150	0.183
onlywomen	0.462	1.059	1.058	0.008	1.036	0.025
virgin	0.004	0.951	0.943	0.218	0.809	4.369

87 components of 40 simulated RDS samples (in red). Sizes are plotted on a logarithmic scale to improve visibility.

In SATHCAP Chicago results, the rate of recruitment into the study as a percentage of the number of coupons handed out is 31.619%. Thus, in our simulation, once a node has been recruited, it has a $p_r = 31.619\%$ chance of responding.

D. Comparison of Simulated RDS with Actual RDS

The simulated RDS samples, created by running an RDS simulation over the generated homophily configuration networks, are similar to the SATHCAP sample in component size distribution (Fig. 1) and component structure.

In general, the homophilies of individual simulated RDS samples vary significantly from sample to sample. Individual samples are not precise predictors of the homophilies of the networks they survey, nor of the homophily targets. However, as the *RDS Mean* column in table I shows, the mean sample homophily across 200 independent samples roughly approximates the target homophily, with error less than 15% for 85% of features (39 out of 46 features). Similar to section IV, the largest error values coincide with the smallest frequencies. The root-mean-square error for all 46 homophilies (target homophily vs mean RDS sample homophily) is 16.89%.

V. CONCLUSION

In this paper we successfully demonstrate a novel Monte-Carlo algorithm for re-wiring networks to preserve multiple homophilies. We empirically demonstrate that this algorithm can preserve many homophilies between simultaneously occurring features with a high degree of accuracy.

We use this algorithm to create a model of the population of low-income injection-drug users and MSM in Chicago, i.e. the SATHCAP population. In future work, such models could be used for expanded study of hidden populations, including clustering analysis, identification of high-betweenness nodes, and resilience testing.

We test our model by simulating RDS samples over the generated networks, and comparing these against the SATHCAP and against the predicted population homophilies. We find that RDS samples introduce noise into homophily measurements, but that there is an empirically valid connection between RDS sample homophilies and the homophilies of the population underlying them, which further validates our model.

The presented model could potentially be useful in other areas of network science, such as the study of clustering algorithms, for its ability to create networks with predetermined groups of high homophily. The relatively high accuracy of this model with 46 features makes it useful for modeling networks with many overlapping communities.

REFERENCES

[1] L. G. Johnston and K. Sabin, “Sampling hard-to-reach populations with respondent driven sampling,” *Methodological innovations online*, vol. 5, no. 2, pp. 38–48, 2010.

[2] J. Wang, R. G. Carlson, R. S. Falck, H. A. Siegal, A. Rahman, and L. Li, “Respondent-driven sampling to recruit mdma users: a methodological assessment,” *Drug and alcohol dependence*, vol. 78, no. 2, pp. 147–157, 2005.

[3] D. D. Heckathorn and J. Jeffri, “Finding the beat: Using respondent-driven sampling to study jazz musicians,” *Poetics*, vol. 28, no. 4, pp. 307–329, 2001.

[4] G. Tyldum and L. Johnston, *Applying respondent driven sampling to migrant populations: Lessons from the field.* Springer, 2014.

[5] D. D. Heckathorn, “Respondent-driven sampling: a new approach to the study of hidden populations,” *Social problems*, vol. 44, no. 2, pp. 174–199, 1997.

[6] L. G. Johnston, D. Prybylski, H. F. Raymond, A. Mirzazadeh, C. Manopaiboon, and W. McFarland, "Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations: case studies from around the world," *Sexually transmitted diseases*, vol. 40, no. 4, pp. 304–310, 2013.

[7] D. M. Feehan and M. J. Salganik, "Generalizing the network scale-up method: a new estimator for the size of hidden populations," *Sociological methodology*, vol. 46, no. 1, pp. 153–186, 2016.

[8] L. G. Johnston, K. R. McLaughlin, H. El Rhilani, A. Latifi, A. Toufik, A. Bennani, K. Alami, B. Elomari, and M. S. Handcock, "Estimating the size of hidden populations using respondent-driven sampling data: case examples from morocco," *Epidemiology (Cambridge, Mass.)*, vol. 26, no. 6, p. 846, 2015.

[9] M. S. Handcock, K. J. Gile, and C. M. Mar, "Estimating hidden population size using respondent-driven sampling data," *Electronic journal of statistics*, vol. 8, no. 1, p. 1491, 2014.

[10] M. Y. Iguchi, A. J. Ober, S. H. Berry, T. Fain, D. D. Heckathorn, P. M. Gorbach, R. Heimer, A. Kozlov, L. J. Ouellet, S. Shoptaw, *et al.*, "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*, vol. 86, no. 1, p. 5, 2009.

[11] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[12] J. M. McPherson and L. Smith-Lovin, "Homophily in voluntary organizations: Status distance and the composition of face-to-face groups," *American sociological review*, pp. 370–379, 1987.

[13] I. E. Fellows, "Respondent-driven sampling and the homophily configuration graph," *Statistics in medicine*, vol. 38, no. 1, pp. 131–150, 2019.

[14] M. J. Salganik, D. Fazito, N. Bertoni, A. H. Abdo, M. B. Mello, and F. I. Bastos, "Assessing network scale-up estimates for groups most at risk of hiv/aids: evidence from a multiple-method study of heavy drug users in curitiba, brazil," *American journal of epidemiology*, vol. 174, no. 10, pp. 1190–1196, 2011.

[15] R. Magnani, K. Sabin, T. Saidel, and D. Heckathorn, "Review of sampling hard-to-reach and hidden populations for hiv surveillance," *Aids*, vol. 19, pp. S67–S72, 2005.

[16] P. Biernacki and D. Waldorf, "Snowball sampling: Problems and techniques of chain referral sampling," *Sociological methods & research*, vol. 10, no. 2, pp. 141–163, 1981.

[17] M. S. Handcock and K. J. Gile, "Comment: On the concept of snowball sampling," *Sociological Methodology*, vol. 41, no. 1, pp. 367–371, 2011.

[18] M. L. Stein, J. E. Van Steenbergen, V. Buskens, P. G. Van Der Heijden, C. Chanyasanha, M. Tipayamongkholgul, A. E. Thorson, L. Bengtsson, X. Lu, and M. E. Kretzschmar, "Comparison of contact patterns relevant for transmission of respiratory pathogens in thailand and the netherlands using respondent-driven sampling," *PloS one*, vol. 9, no. 11, p. e113711, 2014.

[19] M. L. Stein, J. E. Van Steenbergen, C. Chanyasanha, M. Tipayamongkholgul, V. Buskens, P. G. van der Heijden, W. Sabaiwan, L. Bengtsson, X. Lu, A. E. Thorson, *et al.*, "Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in thailand," *PloS one*, vol. 9, no. 1, p. e85256, 2014.

[20] F. W. Crawford, P. M. Aronow, L. Zeng, and J. Li, "Identification of homophily and preferential recruitment in respondent-driven sampling," *American journal of epidemiology*, vol. 187, no. 1, pp. 153–160, 2018.

[21] J. Grubb, D. Lopez, B. Mohan, and J. Matta, "Identifying biomarkers for important nodes in networks of sexual and drug activity," in *International Conference on Complex Networks and Their Applications*, pp. 357–369, Springer, 2020.

[22] M. S. Handcock, I. E. Fellows, and K. J. Gile, *RDS Analyst: Software for the Analysis of Respondent-Driven Sampling Data*. Los Angeles, CA, 2019. Version 0.71.

[23] M. Griffin, K. J. Gile, K. I. Fredricksen-Goldsen, M. S. Handcock, and E. A. Erosheva, "A simulation-based framework for assessing the feasibility of respondent-driven sampling for estimating characteristics in populations of lesbian, gay and bisexual older adults," *The annals of applied statistics*, vol. 12, no. 4, p. 2252, 2018.