Estimating the Number of HIV+ Latino MSM Using RDS, SS-PSE, and the Census

Nicholas Budzban, Katherine Silverio, and John Matta

Abstract—This paper presents a method for estimating the overall size of a hidden population using results from a respondent driven sampling (RDS) survey. We use data from the Latino MSM Community Involvement survey (LMSM-CI), an RDS dataset that contains information collected regarding the Latino MSM communities in Chicago and San Francisco. A novel model is developed in which data collected in the LMSM-CI survey serves as a bridge for use of data from other sources. In particular, American Community Survey Same-Sex Householder data along with UCLA's Williams Institute data on LGBT population by county are combined with current living situation data taken from the LMSM-CI dataset. Results obtained from these sources are used as the prior distribution for Successive-Sampling Population Size Estimation (SS-PSE) - a method used to create a probability distribution over population sizes. The strength of our model is that it does not rely on estimates of community size taken during an RDS survey, which are prone to inaccuracies and not useful in other contexts. It allows unambiguous, useful data (such as living situation), to be used to estimate population sizes.

Index Terms—RDS, hidden populations, public health.

I. Introduction

DETERMINING the size of a hard-to-reach or hidden population is of immense importance when planning for health interventions, estimating their success, or budgeting for costs associated with a disease. This is particularly true for a disease like HIV which affects minority populations disproportionately [1] and is likely to remain unreported due to stigma surrounding both men who have sex with men (MSM) populations and HIV. Respondent driven sampling (RDS) [2] is a survey technique in which members of a hidden population recruit other members into the survey. This technique has been used not only with HIV populations, but also to collect data on hidden populations such as MDMA users [3] and migrant workers [4].

Multiplier methods, which compare independent sources of data, are widely used to estimate the size of hidden populations. However, "multipliers based on different data sources can yield vastly different results" [5], and multiplier methods require random sampling, which is difficult or impossible when hidden members prefer to remain hidden.

Another approach to estimating the size of hard-to-count populations is *network scale-up* (NSUM) [6]–[8]. NSUM is a simple but powerful idea, where respondents from the hidden population are asked to estimate the number of members m_i

Nicholas Budzban, Katherine Silverio and John Matta are with the Computer Science Department, Southern Illinois University Edwardsville, Edwardsville, IL 62026 USA. Corresponding email:jmatta@siue.edu

of that population in their social circle, as well as to estimate the overall size of their social circle $\hat{c}_i$. This ratio is then scaled up by the size of the overall population N to produce an estimate $\hat{e}$ of the size of the hidden population, according to the formula

$$\hat{e} = \frac{\sum_i m_i}{\sum_i \hat{c}_i} \times N.$$

The use of NSUM requires planning when the survey data are collected, such that respondents are asked about the size of their network as well as the size of personally known sub-networks with sought-after characteristics. This requires asking respondents for estimates that may be error-prone. Additionally, the data collected are not useful in any other meaningful sense. The dataset used in this paper does not contain the respondent's overall network size information, so another method is developed.

We use the Latino MSM Community Involvement (LMSM-CI) dataset [9] to estimate the population size of Latino MSM with HIV in Cook County, (located in Chicago, Illinois USA), and San Francisco County (located in San Francisco, California, USA). This dataset is the work of Dr. Jesus Ramirez-Valles and was obtained for a study to determine whether community involvement reduced the risk of HIV in minority populations [10]. The theoretical framework for that study came from Ramirez-Valles' 2002 paper [11], in which he proposed guidelines for reducing HIV risk behavior.

In 2011 Gile introduced a successive sampling (SS) based estimator for population means that does not require knowledge of the subpopulation and uses data collected through respondent-driven sampling alone [12]. This method was improved upon by Handcock, Gile, and Mar [13], who implemented their methods in the RDS Analyst software package [14]. The estimation techniques described in this paper use RDS Analyst and its included successive sampling - population size estimation (SS-PSE) package. They are similar to the techniques used by Johnston, McLaughlin, Rouhani, and Bartels in [15]. In that paper RDS Analyst and the SS-PSE package are used, with experts providing a population estimate for RDS Analyst's posterior distribution tool. Here, however, instead of asking experts for a population estimate, we use a novel combination of data sources, including American Community Survey (ACS) census data and information on LGBT populations from the Williams Institute [16]. The single question of living situation, and in particular the response of "Same-Sex Householder" is

used as a bridge between the LMSM-CI survey and the corresponding census and Williams Institute data. We show that incorporating this additional data solves the multiplier method problem of requiring random data sources, as well as the NSUM problem of requiring survey-specific network estimation questions.

II. RELATED WORK

Many attempts have been made to estimate the size of hidden populations such as MSM and injecting drug users. One early method by Archibald et al. uses data from HIV testing results and combines it with HIV testing behavior data [17]. The size estimate for these populations is determined by dividing the number of the studied population (either MSM or injecting drug users) by the proportion of the respective group that reported being tested.

In Livak et al., the authors estimate the population of young Black MSM (YBMSM) living on the south side of Chicago, the area of the city where HIV is most prevalent [18]. They use three methods: an indirect approach, data from the National Survey of Family Growth, and a modified Delphi approach. They determine the crude average of these methods and estimate the population of YBMSM to be 5,578. Wesson, Handcock, McFarland, and Raymond also study African American MSM, but focus on San Francisco. They use the respondents' personal network sizes, collected as part of the survey, with RDS-Analyst to make their estimation [19]. The current study is similar to these in that we examine both Chicago and San Francisco. However, instead of concentrating on African American MSM, our target population is Latino MSM with HIV.

Safarnejad, Nga, and Son estimate the hard-to-reach population of MSM in Ho Chi Minh City and Nghe An province, Vietnam [20]. The authors use a multiplier method, social application technology, and internet surveys. Raymond, McFarland, and Wesson [21] update the estimated MSM population size in San Francisco [22] using multiple methods and data sources. The authors obtain this updated hidden population size estimate by using several estimates synthesized by the Anchored Multiplier method (a Bayseian method). The Anchored Multiplier method was developed by Wesson et al. [23].

In [24], the authors combine census and RDS data to estimate the size of female sex workers in a city in western China. The authors determine that census data tends to underestimate population sizes and could be used as a lower limit. They also find that multiplier methods could be used to determine population size estimates for larger geographic regions.

In [25], Handcock, Gile, and Mar study two different hard-to-reach populations in El Salvador. They use methods to estimate these population sizes from recruitment patterns obtained from RDS data.

In addition to specific population estimation techniques, there is a body of related literature that examines the techniques for strengths and weaknesses, as well as for accuracy. Fearon et al. study using multiplier methods, with an emphasis on examining the often large variance in the resulting estimates [26]. To obtain a more confident estimate of a hard-to-reach population when using RDS data, the authors suggest changes to survey collection methods, such as a longer period of coupon distribution, as well as beginning collection with enough seeds to adequately capture the diversity of the hidden population. An RDS survey is analyzed as a graph in [27].

Abdul-Quader, Baughman, and Hladik observe in [5] that estimation methods for crucial populations are inadequate. In their paper, they summarize and review six methods for estimating the size of a key population, including the single-survey method based on an RDS survey.

III. METHODS

A. Heckathorn's Respondent Driven Sampling

RDS is a data collection technique where samples are generated from a random walk along nodes in the underlying network with sampling probability proportional to the node's degree [13]. Heckathorn showed that even though the "seeds" of the network are chosen by the researchers and may be considered a convenience sample, the subsequent samples chosen by the participants become increasingly independent and disconnected from the seeds and bias of the researchers [2].

There are several components required for a successful RDS sampling process and its subsequent analysis. First, members of the target population must be able to identify and recruit other members of that population. This internal recruitment is one of the strengths of the RDS process, as members of hidden populations are often stigmatized and wish to keep their identities secret outside of the internal network. RDS allows an individual's participation in the study to be anonymous and unknown to anyone outside of their recruiter and their recruitees, and so RDS seems to be an ethical way to study populations whose individuals wish to remain hidden from those outside.

As for analysis, there are several assumptions being made for us to consider the collected data as an unbiased probability sample. First, the recruit's recruitment process must be done independently and uncorrelated with the study's variables – else the analysis of those variables will suffer from sampling bias. Furthermore, RDS assumes that a sample's recruitment probability is the inverse of their network "visibility" – the number of people who know them well enough to recruit them. This assumption about sampling probability yields the RDS-II probability estimator, the basis of Gile's SS estimator which we use in this study.

B. Gile's Successive Sampling Population Size Estimation

Successive Sampling Population Size Estimation (SS-PSE) is a technique developed by Gile and Handcock [13] to make inference on a network's size given the order and degree of nodes in the random-walk sampled network. The basic idea is that we expect to sample nodes with a higher degree earlier in the random walk process, and the prevalence of large nodes late in the random walk suggests we are only scratching

the surface of a large, untapped network. The reason this technique is so appealing for both our use and RDS studies in general is because RDS studies collect the SS-PSE required data by default [13], [25].

SS-PSE as implemented by RDS Analyst [14] has a few additional requirements. First, users must specify a prior estimate on the network's size in the form of a mean, median, mode, or 50% confidence interval. This information will be used by RDS Analyst to fit a prior beta distribution to the user's input. Furthermore, RDS Analyst will automatically fit an exponential degree distribution to the sampled network without any additional user input. From there, SS-PSE will undergo a Markov Chain Monte Carlo (MCMC) sampling process, collecting samples from random walks on the network of the known degree distribution, and using those samples to adjust the user's prior distribution into a posterior beta distribution. We consider SS-PSE's adjusted posterior beta distribution as our preferred estimate for the population size of Latino MSM who are HIV+ $Posterior N_{L,MSM,HIV+}$ and HIV unknown $Posterior N_{L,MSM,HIV?}$.

C. Ramirez-Valles's Latino MSM Community Involvement

Our primary data is the RDS network of individual respondents collected in the Latino MSM Community Involvement: HIV Protective Effects study [11]. This 2003-2004 survey features one of the original large-scale RDS sampling processes which collected a total of 643 samples of Latino MSM in the Chicago metro area (323 samples) and the San Francisco Bay area (320 samples). The study had several aims relevant to understanding the social determinants of HIV in the Latino MSM community, chief among them being whether a Latino MSM's sense of belonging or actions of involvement in their community had the so-called "HIV Protective Effects." We take special interest in the use of this RDS data combined with prior knowledge from other sources to make population size estimates for Latino MSM with HIV in Cook and San Francisco County.

D. American Community Survey's Same-Sex Households

While Latino MSM and GBT-identifying populations are considered "hidden" from statistical researchers, Latino Male Same-Sex Householders (SSH) and their domestic partners can be readily identified within the American Community Survey (ACS). SSH populations across genders, ethnicities, years, and places can be studied by filtering the millions of individual ACS Public-Use Microdata household samples released every year on the variables of "Sex" and "Relationship with Householder." The U.S. Census Bureau has been producing reports on SSH populations this way since the ACS began in 2005 [28] [29].

Interestingly, after a change in the graphic design of the survey form in 2008, there was a surprisingly significant drop in the number of SSH respondents across the nation. The U.S. Census Bureau concluded that the previous form resulted in respondents mischecking the box used to identify their sex. In 2011, after years of developing a statistical model relating a respondent's sex to their first name (first names and other personally identifying information are not included in the public data), the Census Bureau released their "preferred" estimates for SSH demographics and population sizes for states across the U.S.

E. William's Institute County-Level Estimates

The Williams Institute's LGBT Data Interactive leverages the aforementioned state-level estimates in combination with their own model to present finer-grained county-level estimates for the number of Same-Sex Householders N_{SSH}, and the probability that the SSH is Latino $P_{L|SSH}$ or Male $P_{M|SSH}$ for all counties in the U.S. [16].

We utilize their county-level estimates of Latino SSH and Male SSH and assume independence between $P_{L|SSH}$ and $P_{M|SSH}$ to produce an estimate of $P_{L,M|SSH}$. We combine that rate with their original estimate of overall population size N_{SSH} to arrive at our prior estimate for the number of Latino Male Same-Sex Householders in each county, $N_{L,M,SSH}$.

$$P_{M|SSH} \times P_{L|SSH} = \hat{P}_{L,M|SSH} \tag{1}$$

$$\hat{P}_{L,M|SSH} \times N_{SSH} = \hat{N}_{L,M,SSH} \tag{2}$$

F. Latino MSM's Living Situation

The question then is how to use the fairly reliable population size estimate of Latino Male Same-Sex Householders $\hat{N}_{L,M,SSH}$ to make an inference on the hidden Latino MSM population size, $\hat{N}_{L,MSM}$.

Our approach is to leverage Latino MSM answers to a single question in the Ramirez-Valles survey: "Which of the following best describes your living situation?" Respondents who answered "I am living with a domestic partner" $LivSit1$ or "I am living with a domestic partner and other people" $LivSit2$ are aggregated into the single population of Latino MSM who are in a domestic partnership, $P_{DomesPart|L,MSM}$.

$$\hat{P}_{LivSit1} + \hat{P}_{LivSit2} = \hat{P}_{DomesPart|L,MSM} \tag{3}$$

As the $LivSit$ question did not ask whether the respondent was the householder, we are left with the naive assumption that half of Latino MSM who are in a domestic partnership are also the householder, thus yielding a point estimate on $P_{SSH|L,MSM}$ - the critical ratio that bridges the known and hidden population size:

$$\frac{1}{2} \times \hat{P}_{DomesPart|L,MSM} = \hat{P}_{SSH|L,MSM} \tag{4}$$

$$\hat{N}_{L,M,SSH} / \hat{P}_{SSH|L,MSM} = \hat{N}_{L,MSM} \tag{5}$$

In other words, the less likely Latino MSM are to be SSH, the greater our estimate for $\hat{N}_{L,MSM}$.

G. Latino MSM's HIV Status

The Ramirez-Valles dataset also provides observations on the ratio of Latino MSM who are HIV+ $\hat{P}_{HIV+|L,MSM}$ and HIV unknown $\hat{P}_{HIV?|L,MSM}$ in both Cook and San Francisco Counties. We obtained point estimates and 95% confidence intervals for these variables using RDS Analyst.

TABLE I: Data sources and estimates leading a result for the number of Latino MSM who are HIV positive and HIV unknown using Respondent-Driven Sampling and county-level statistics.

Source	Symbol	Description	Cook County	95%	San Francisco County	95%
[16]	N_{SSH}	# of Same-Sex Householders, "SSH"	14,050	±?	10,450	±?
	$P_{M|SSH}$	% SSH are Male	68.33%	±?	82.05%	±?
	$P_{L|SSH}$	% SSH are Latino	12.55%	±?	10.32%	±?
Eq. (1)	$\hat{P}_{L,M|SSH}$	% SSH are Latino and Male	8.58%	±?	8.47%	±?
Eq. (2)	$\hat{N}_{L,M,SSH}$	# of Latino Male SSH	1,205	±?	885	±?
[9]	$\hat{P}_{LivSit1|L,MSM}$	% Latino MSM living with partner only	16.05%	±6.16%	14.79%	±5.71%
	$\hat{P}_{LivSit2|L,MSM}$	% Latino MSM living with partner and others	3.14%	±2.98%	3.50%	±3.22%
Eq. (3)	$\hat{P}_{DomesPart|L,MSM}$	% Latino MSM are domestic partner	19.19%	±6.84%	18.29%	±6.56%
Eq. (4)	$\hat{P}_{SSH|L,MSM}$	% Latino MSM are SSH	9.60%	±3.42%	9.15%	±3.28%
Eq. (5)	$Prior\hat{N}_{L,MSM}$	# of Latino MSM	12,552	±6,946	9,672	±5,405
[30]	$N_{L,M}$	Number of Latino Males	462,801	±?	54,251	±?
	$P_{MSM|L,M}$	% Latino Male are MSM	2.71%	±1.50%	17.83%	±9.96%
[11]	$\hat{P}_{HIV+|L,MSM}$	% Latino MSM are HIV positive	14.1%	±6.85%	34.6%	±9.8%
	$\hat{P}_{HIV?|L,MSM}$	% Latino MSM are HIV unknown	17.1%	±5.10%	10.1%	±6.86%
Eq. (6)	$\hat{N}_{L,MSM,HIV+}$	# of Latino MSM are HIV positive	1,770	±860	3,347	±948
	$\hat{N}_{L,MSM,HIV?}$	# of Latino MSM are HIV unknown	2,146	±640	976	±657

TABLE II: Gile's SS weighted population estimates for a variety of living situations and HIV statuses in two counties.

Living Situation	Cook County						San Francisco County					
	Point Est.(%)	95% Lower Bound	95% Upper Bound	Est. Design Effect	Std. Error	Sample Size	Point Est.(%)	95% Lower Bound	95% Upper Bound	Est. Design Effect	Std. Error	Sample Size
Alone in house / apartment	25.27	17.86	32.65	3.44	3.77	84	22.30	13.48	31.13	3.01	4.50	59
Homeless	0.89	-0.25	2.03	1.76	0.58	3	3.67	-0.24	7.60	2.91	2.00	6
School dormitory	0.41	-0.35	1.18	1.68	0.39	1						
Residential hotel	0.31	-0.17	0.79	0.89	0.24	1	10.96	3.21	18.72	4.13	3.96	21
Shelter / halfway home / rehabilitation facility	2.77	-1.59	7.15	8.45	2.23	6	5.29	1.36	9.24	2.07	2.01	15
With domestic partner / lover / boyfriend / others	3.14	0.16	6.13	3.49	1.52	10	3.50	0.28	6.70	2.04	1.64	9
With domestic partner / lover / boyfriend	16.05	9.89	22.18	3.34	3.14	45	14.79	9.08	20.45	1.72	2.90	41
With friend(s) or roommate(s)	25.55	18.96	32.20	2.74	3.38	92	33.30	24.35	42.25	2.42	4.57	84
With other relatives	14.04	7.72	20.33	3.92	3.22	36	3.86	0.00	7.73	2.70	1.97	10
With parents or guardians	11.57	6.00	17.18	3.64	2.85	36	2.33	-0.42	5.07	2.22	1.40	5
HIV-	68.80	60.55	77.17	3.83	4.24	205	55.30	46.41	64.31	2.17	4.57	142
HIV+	14.10	7.25	20.96	4.62	3.50	55	34.60	24.80	44.43	2.85	5.01	88
Unknown	17.10	12.00	22.06	2.12	2.56	54	10.10	3.24	16.81	3.40	3.46	20

H. RDS Analyst's Population Estimates

We use RDS Analyst to produce population proportion estimates using the Ramrirez-Valles RDS data weighted by Gile's Successive Sampling estimator (GSS) for the observed variables, $\hat{P}_{LivSit1}$ $\hat{P}_{LivSit2}$, $\hat{P}_{HIV+|L,MSM}$, and $\hat{P}_{HIV?|L,MSM}$.

I. RDS Analyst's SS-PSE

We provided RDS Analyst's SS-PSE with a prior estimate on Latino MSM population size in the form of a 50% confidence interval on $\hat{N}_{L,MSM}$. We obtained the 50% confidence interval under the assumption that it was 34.4% the width of our estimated 95% confidence interval. We then run SS-PSE on that prior distribution for N = 10 trials and take the mean of the means to be our posterior point estimate on $\hat{N}_{L,MSM}$.

Finally, we use those posterior estimates on $\hat{N}_{L,MSM}$ and the same population proportion estimates to produce posterior estimates on the number Latino MSM who are HIV+ and unknown, $\hat{N}_{L,MSM,HIV+}$ and $\hat{N}_{L,MSM,HIV?}$.

$$\hat{P}_{HIV|L,MSM} * \hat{N}_{L,MSM} = \hat{N}_{L,MSM,HIV} \quad (6)$$

(a) An example of a $N_{L,MSM}$ Posterior for Cook County   (b) An example of a $N_{L,MSM}$ Posterior for San Francisco County
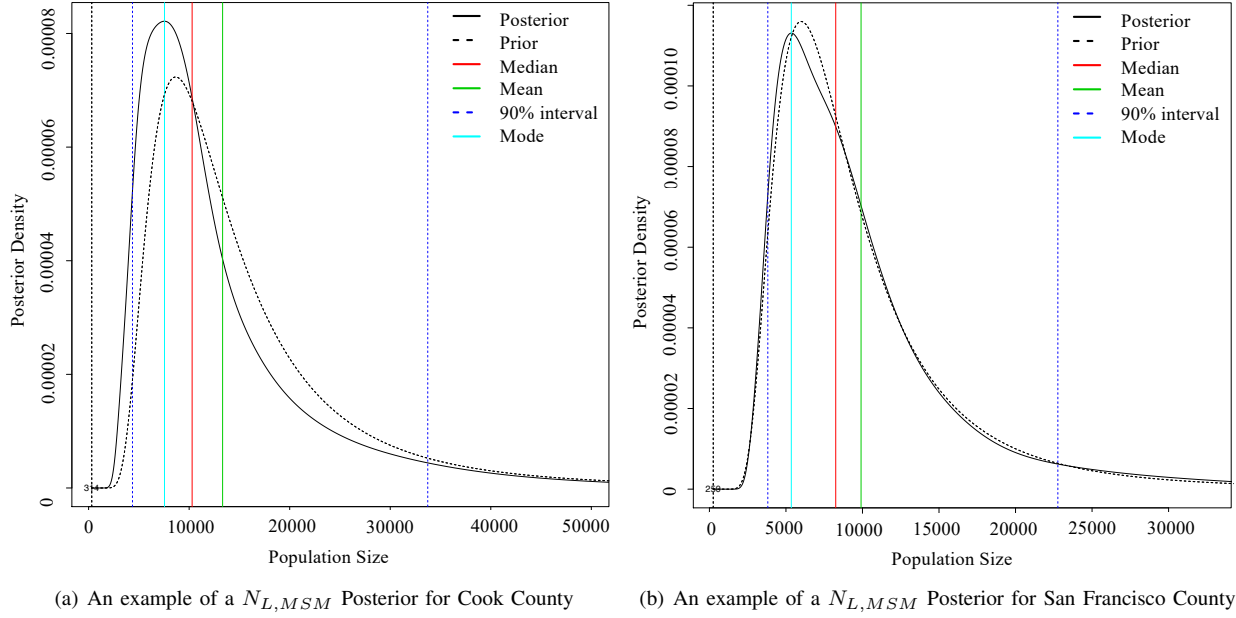
Fig. 1: Example $N_{L,MSM}$ posteriors. The dashed "Prior" curves are the probability distributions fit by RDS Analyst to our prior $N_{L,MSM}$ estimates. The solid "Posterior" curves are the probability distributions on $N_{L,MSM}$ produced by SS-PSE.

TABLE III: Summary of SS-PSE Posterior estimates. Priors from Table I are input into SS-PSE to estimate the mean $N_{L,MSM}$ across 10 trials. The resulting posterior estimates on $\hat{N}_{L,MSM,HIV+}$ and $\hat{N}_{L,MSM,HIV?}$ are re-applications of Eq. (6) to those new estimates on $N_{L,MSM}$.

	Mean	Median	Mode	25%	75%	90%	2.5%	97.5%
Cook County								
$Prior\,\hat{N}_{L,MSM}$	12878	10708	7772	7660	15598	22654	4458	34801
$Posterior\,\hat{N}_{L,MSM}$	10071.5	8545.3	6904.6	6013	12335.4	17607.9	2583.9	27020.4
Relative Change	-21.79%	-20.20%	-11.16%	-21.50%	-20.92%	-22.27%	-42.04%	-22.36%
$\hat{N}_{L,MSM,HIV+}$	1420.08	1204.89	973.55	847.83	1739.29	2482.71	364.33	3809.88
$\hat{N}_{L,MSM,HIV?}$	1722.23	1461.25	1180.69	1028.22	2109.35	3010.95	441.85	4620.49
San Francisco County								
$Prior\,\hat{N}_{L,MSM}$	9061	7722	5810	5642	10958	15475	3382	22986
$Posterior\,\hat{N}_{L,MSM}$	8469.8	7393.5	5925.8	5300.2	10416.5	14516.3	2448.5	21020.1
Relative Change	-6.52%	-4.25%	1.99%	-6.06%	-4.94%	-6.20%	-27.60%	-8.55%
$\hat{N}_{L,MSM,HIV+}$	2930.55	2558.15	2050.33	1833.87	3604.11	5022.64	847.18	7272.95
$\hat{N}_{L,MSM,HIV?}$	855.45	746.74	598.51	535.32	1052.07	1466.15	247.30	2123.03

IV. RESULTS

As shown in the first section of Table I, the William's Institute's published results on the number of Same-Sex Householders, $\hat{N}_{L,M,SSH}$, in Cook and San Francisco County are 14,050 and 10,450, respectively [16]. We could not find 95% intervals for these estimates.

The William's Institute also published estimates on the proportion of Latino SSH and Male SSH. Under the assumption of independence, we make an estimate on the proportion of Latino Male SSH for Cook and San Francisco counties to be 8.58% and 8.47%, respectively. The county proportion estimates are remarkably similar because, while San Francisco County has a larger Male SSH proportion, 82.05% vs. 68%, it has a lower Latino SSH proportion than Cook County, 10% vs. 12.5%. Multiplying the Latino Male

SSH proportion by the SSH population, our final estimate for the number of Latino Male SSH is 1,205 and 885 in Cook and San Francisco county.

Using the 2004 LMSM-CI study [9] we make estimates driving towards the percentage of Latino MSM who are SSH. Table II shows the categorical distribution across living situations for Latino MSM, with 95% intervals computed by RDS Analyst. By taking the sum of two key living situations as shown in Eq. (3), we estimate the proportion of Latino MSM who are in a domestic partnership to be 19.19% and 18.29% with 95% intervals $\pm\%6.85\%$ and $\pm6.56\%$. Assuming that half of domestic partners are SSH leads to the critical estimate for proportion of Latino MSM who are SSH, $\hat{P}_{SSH|L,MSM}$ to be 9.60% and 9.15% with 95% intervals $\pm3.42\%$ and $\pm3.28\%$, respectively.

As shown in Eq. (5), those ratios allow us to produce our prior estimates on the number of Latino MSM for Cook and San Francisco Counties to be 12,552 and 9,672 with 95% intervals $\pm 6,946$ and $\pm 5,405$, respectively.

An interesting detail that can be derived from the calculations in Table I is the percentage of Latino males who are MSM. The number of Latino Males is provided by the U.S. Census Bureau's county intercensal estimates in 2011 [30], as 462,801 and 54,251 for Cook and San Francisco County, respectively. The significantly smaller San Francisco Latino Male population implies a significantly larger probability that a Latino Male is a MSM in that county, with our estimates suggesting a probability of MSM given Latino Male to be about 1 in 6, 17.26% $\pm 9.96\%$ in San Francisco County vs. 1 in 36, 2.75% $\pm 1.5\%$ in Cook.

Table II shows population estimates produced by RDS Analyst complete with 95% intervals, sample sizes, and the estimated design effects. Using this table we can see that the sample sizes were close to zero for rarer living situations like "Homeless" and "School dormitory," a total of 9 Homeless and 1 School dormitory across both counties. Additionally, the rareness of a living situation varied between counties. In Cook County, Latino MSM are more likely to live with their parents, guardians or relatives, 25.61% $\pm 5.57\%$ vs. San Francisco's relatively rare 6.19% $\pm 4.73\%$. This may be an indication of a strong support network and familial roots for these men in Cook County vs. a population in San Francisco who tends to live in more non-traditional situations like residential hotels, shelters, or homeless, with 19.92% ± 3.95 in San Francisco vs. 1.61% ± 2.37 in Cook.

To use SS-PSE we supply the prior estimates as calculated and displayed in Table I, specifically the Cook and San Francisco County Latino MSM population sizes, N_{LMSM}. We run SS-PSE 10 times for each county, summarizing the SS-PSE estimates obtained from those trials in Table III.

Each trial produces a beta distribution, an example of which is shown in Figure 1. The dashed curve is the beta distribution fit by RDS Analyst to our prior estimates, which has been updated by SS-PSE to produce the posterior beta distribution shown as the solid curve. The vertical lines summarize the posterior beta distribution in terms of its mean (green), median (red), mode (light blue), and 90% confidence interval (dark blue).

We use the mean of the mean estimates as the best point estimate for population size. The mean of the means of the 10 SS-PSE trials for Cook and San Francisco County, as shown in Table III, are 10,072 and 8,470 with standard errors 241 and 127, respectively. The 95% bounds for the size of the Latino MSM population as computed by SS-PSE span from 2,584 to 27,020 for Cook County and 2,449 - 21,020 for San Francisco County.

Multiplying the same HIV proportions by the posterior estimates for the Latino MSM population size provided by SS-PSE, we arrive at our posterior estimates for the number of Latino MSM by HIV status in Cook and San Francisco counties to be 1,420 and 2,930 who are HIV positive and 1,722 and 855 who are HIV unknown, respectively.

V. CONCLUSION

Our method to estimate the total size of a hidden population featured several strengths and weaknesses. This is expected, as it is believed at this time that there is no single best method for estimating the size of a hard-to-reach population. This belief is based on work by Mauck et al. [31], in which the authors review and compare multiple methods for estimating population size, but specifically for men who have sex with men. They determine that there is no single best method at this time and that in order to obtain a robust estimate, multiple methods should be used.

In our study, we found that the methods available to us were expanded or limited by the inclusion of key variables in the RDS survey. We were fortunate to find a variable in our dataset which could be cross-referenced with data collected by the American Community Survey. This key variable, *Living Situation*, provided evidence suggesting that the odds of a Latino MSM being a Same-Sex Householder is roughly 1 in 10 in both San Francisco and Cook County. The reliable estimates provided by the U.S. Census Bureau in the sampling of Same-Sex Householders combined with the consistency in the Living Situation distribution between the two counties is an encouraging sign for the existence of a statistically reliable relationship between SSH and MSM populations.

On the other hand, our methods were limited by the exclusion of key variables. For example, we could not use the generalized network scale-up method because our RDS dataset was missing one variable, "Estimated Total Network Size." If our RDS survey included an estimation of the respondent's total network size along with their Latino MSM network size, we would have been capable of using NSUM as an additional technique to estimate the total Latino MSM population size.

We recommend future studies of MSM populations include a similar question to *Living Situation*, and the collection of other variables which can be cross-referenced with the Same-Sex Householder populations studied by the American Community Survey. Doing so could enable the analysis of variance in those variables between Same-Sex Householders and MSM populations in counties across the U.S. Furthermore, as demonstrated in this paper, estimates on the proportion of MSMs who are SSH enables an additional method of population size estimation based on the reliable information on Same-Sex Householders as provided by the U.S. Census Bureau.

REFERENCES

[1] K. L. Hess, S. D. Johnson, X. Hu, J. Li, B. Wu, C. Yu, H. Zhu, C. Jin, M. Chen, J. Gerstle, *et al.*, "Diagnoses of hiv infection in the united states and dependent areas, 2017," 2018.

[2] 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.

[3] 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.

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

[5] A. S. Abdul-Quader, A. L. Baughman, and W. Hladik, "Estimating the size of key populations: current status and future possibilities," *Current Opinion in HIV and AIDS*, vol. 9, no. 2, p. 107, 2014.

[6] H. R. Bernard, T. Hallett, A. Iovita, E. C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M. J. Salganik, T. Saliuk, O. Scutelniciuc, *et al.*, "Counting hard-to-count populations: the network scale-up method for public health," *Sexually transmitted infections*, vol. 86, no. Suppl 2, pp. ii11–ii15, 2010.

[7] W. Guo, S. Bao, W. Lin, G. Wu, W. Zhang, W. Hladik, A. Abdul-Quader, M. Bulterys, S. Fuller, and L. Wang, "Estimating the size of hiv key affected populations in chongqing, china, using the network scale-up method," *PloS one*, vol. 8, no. 8, p. e71796, 2013.

[8] 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.

[9] J. Ramirez-Valles, *Latino MSM Community Involvement: HIV Protective Effects*. 2013.

[10] J. Ramirez-Valles, D. Garcia, R. T. Campbell, R. M. Diaz, and D. D. Heckathorn, "Hiv infection, sexual risk behavior, and substance use among latino gay and bisexual men and transgender persons," *American Journal of Public Health*, vol. 98, no. 6, pp. 1036–1042, 2008.

[11] J. Ramirez-Valles, "The protective effects of community involvement for hiv risk behavior: a conceptual framework," *Health Education Research*, vol. 17, no. 4, pp. 389–403, 2002.

[12] K. J. Gile, "Improved inference for respondent-driven sampling data with application to hiv prevalence estimation," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 135–146, 2011.

[13] 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.

[14] M. S. Handcock, I. E. Fellows, and K. J. Gile, *RDS Analyst: Software for the Analysis of Respondent-Driven Sampling Data*. Los Angeles, CA, 2014. Version 0.42.

[15] L. G. Johnston, K. R. McLaughlin, S. A. Rouhani, and S. A. Bartels, "Measuring a hidden population: A novel technique to estimate the population size of women with sexual violence-related pregnancies in south kivu province, democratic republic of congo," *Journal of epidemiology and global health*, vol. 7, no. 1, pp. 45–53, 2017.

[16] "Lgbt demographic data interactive." https://williamsinstitute.law.ucla.edu/visualization/lgbt-stats/?topic=SS&area=6075&characteristic=hispanic#about-the-data. Accessed: 2020-12-16.

[17] C. P. Archibald, G. C. Jayaraman, C. Major, D. M. Patrick, S. M. Houston, and D. Sutherland, "Estimating the size of hard-to-reach populations: a novel method using hiv testing data compared to other methods," *Aids*, vol. 15, pp. S41–S48, 2001.

[18] B. Livak, S. Michaels, K. Green, C. Nelson, M. Westbrook, Y. Simpson, N. G. Prachand, N. Benbow, and J. A. Schneider, "Estimating the number of young black men who have sex with men (ybmsm) on the south side of chicago: towards hiv elimination within us urban communities," *Journal of Urban Health*, vol. 90, no. 6, pp. 1205–1213, 2013.

[19] P. Wesson, M. S. Handcock, W. McFarland, and H. F. Raymond, "If you are not counted, you don't count: estimating the number of african-american men who have sex with men in san francisco using a novel bayesian approach," *Journal of Urban Health*, vol. 92, no. 6, pp. 1052–1064, 2015.

[20] A. Safarnejad, N. T. Nga, and V. H. Son, "Population size estimation of men who have sex with men in ho chi minh city and nghe an using social app multiplier method," *Journal of Urban Health*, vol. 94, no. 3, pp. 339–349, 2017.

[21] H. F. Raymond, W. McFarland, and P. Wesson, "Estimated population size of men who have sex with men, san francisco, 2017," *AIDS and Behavior*, vol. 23, no. 6, pp. 1576–1579, 2019.

[22] H. F. Raymond, S. Bereknyei, N. Berglas, J. Hunter, N. Ojeda, and W. McFarland, "Estimating population size, hiv prevalence and hiv incidence among men who have sex with men: a case example of synthesising multiple empirical data sources and methods in san francisco," *Sexually transmitted infections*, vol. 89, no. 5, pp. 383–387, 2013.

[23] P. D. Wesson, A. Mirzazadeh, and W. McFarland, "A bayesian approach to synthesize estimates of the size of hidden populations: the anchored multiplier," *International journal of epidemiology*, vol. 47, no. 5, pp. 1636–1644, 2018.

[24] D. Zhang, L. Wang, F. Lv, W. Su, Y. Liu, R. Shen, and P. Bi, "Advantages and challenges of using census and multiplier methods to estimate the number of female sex workers in a chinese city," *AIDS care*, vol. 19, no. 1, pp. 17–19, 2007.

[25] M. S. Handcock, K. J. Gile, and C. M. Mar, "Estimating the size of populations at high risk for hiv using respondent-driven sampling data," *Biometrics*, vol. 71, no. 1, pp. 258–266, 2015.

[26] E. Fearon, S. T. Chabata, J. A. Thompson, F. M. Cowan, and J. R. Hargreaves, "Sample size calculations for population size estimation studies using multiplier methods with respondent-driven sampling surveys," *JMIR public health and surveillance*, vol. 3, no. 3, p. e59, 2017.

[27] 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.

[28] "Us census characteristics of same-sex couple households." https://www.census.gov/data/tables/time-series/demo/same-sex-couples/ssc-house-characteristics.html. Accessed: 2021-02-10.

[29] "Us census same-sex couples data tables." https://www.census.gov/topics/families/same-sex-couples/data/tables.All.html. Accessed: 2021-02-10.

[30] "Us census county intercensal estimates." https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-detail.html. Accessed: 2021-02-10.

[31] D. E. Mauck, M. T. Gebrezgi, D. M. Sheehan, K. P. Fennie, G. E. Ibañez, E. A. Fenkl, and M. J. Trepka, "Population-based methods for estimating the number of men who have sex with men: a systematic review," *Sexual health*, vol. 16, no. 6, pp. 527–538, 2019.