Social Networks 64 (2021) 109–121

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Bias decomposition and estimator performance in respondent-driven sampling

Antonio D. Sirianni [a,*], Christopher J. Cameron [b], Yongren Shi [c], Douglas D. Heckathorn [b]

[a] Dartmouth College, United States
[b] Cornell University, United States
[c] University of Iowa, United States

ARTICLE INFO

Keywords:
Respondent-driven sampling
Network sampling

ABSTRACT

Respondent-Driven Sampling (RDS) is a method of network sampling that is used to sample hard-to-reach populations. The resultant sample is non-random, but different weighting methods can account for the over-sampling of (1) high-degree individuals and (2) homophilous groups that recruit members more effectively. While accounting for degree-bias is almost universally agreed upon, accounting for recruitment-bias has been debated as it can further increase estimate variance without substantially reducing bias. Simulation-based research has examined which weighting procedures perform best given underlying population network structures, group recruitment differences, and sampling processes. Yet, in the field, analysts do not have *a priori* knowledge of the network they are sampling. We show that the RDS sample data itself can determine whether a degree-based estimator is sufficient. Formulas derived from the decomposition of a 'dual-component' estimator can approximate the 'recruitment component' (RC) and 'degree component' (DC) of a sample's bias. Simulations show that RC and DC values can predict the performance of different classes of estimators. Samples with extreme 'RC' values, a consequence of network homophily and differential recruitment, are better served by a classical estimator. The use of sample data to improve estimator selection is a promising innovation for RDS, as the population network features that should guide estimator selection are typically unknown.

1. Introduction

Respondent-driven sampling (RDS) has become the method of choice for studies of hidden and hard-to-reach populations. RDS leverages the networks of individuals, and the tendency for members of a hidden or hard-to-reach group to affiliate with one another, by paying respondents to recruit individuals they know within the population of interest. This allows for more effective recruiting of hard-to-reach populations than would be possible had sampling been carried out by professionals or other outsiders. However, peer-recruitment-based samples are inherently biased in most conditions because they reflect the patterns of social ties — including the tendency for like to associate with like. Hence, individuals in the population are not sampled with equal probability. RDS estimators are designed to account for this bias, by up-weighting individuals who have a lower probability of being chosen into the sample, and down-weighting those who have a higher probability of being chosen into the sample.

RDS weights and estimators account for two main sources of bias: one due to differences in nodal degree (how many individuals an individual is connected to in the population), and another due to group-level differences in how actively individuals recruit members of their network. Information on degree is obtained from self-report information about the number of connections an individual has in the target population. Recruitment information is obtained by looking at the rates at which individuals in different groups recruit one another and is often expressed as a 'recruitment matrix'.

Different RDS estimators account for these two sources of bias in different ways, and some will ignore one source altogether. Estimators that use only degree information are termed "degree-based" (Volz and Heckathorn, 2008; Gile, 2011); estimators that use only recruitment-based information are termed "recruitment-based" (Heckathorn, 1997); and estimators that use both are termed "classical" or "dual-component" estimators (Heckathorn, 2002, 2007; Salganik and Heckathorn, 2004). A naïve estimate, which gives all individuals equal weight, ignores both.

Estimators that focus purely on the 'degree-based' component are

* Corresponding author at: Blunt Alumni Center, Department of Sociology, Dartmouth College, United States.
E-mail address: antonio.d.sirianni@dartmouth.edu (A.D. Sirianni).

https://doi.org/10.1016/j.socnet.2020.08.002

Available online 11 September 2020

typically lower in variance and should be preferred when homophily and differential recruitment activity by group are not jointly present. Yet if certain groups are more likely to recruit and are disproportionately connected to one another, then estimates will not be adequately corrected by a degree-based estimator, and the higher-variance classical estimators will provide estimates with less bias. It has been suggested that in cases where extensive homophily and clustering is thought to be an issue *a priori*, that respondent-driven sampling simply not be used because of the variance issues (Gile and Handcock, 2010) . We propose that the recruitment matrix itself can be used to identify when biased recruitment is an issue *a posteriori*, and that the typically higher-variance 'classical estimators' provide a more accurate estimate in these situations. Conversely, the recruitment matrix can also indicate when the problems of differential recruitment activity and homophily are laregely absent and 'degree-based' estimators should be preferred.

In this article we offer recommendations for how different estimators can be chosen based on the observed recruitment matrix. After providing a review of the RDS approach and different RDS estimators, we offer a novel two-part assessment of the performance of different RDS estimators. First, we demonstrate two principled ways of analytically isolating the level of recruitment bias from the level of degree bias by using two different decompositions. One is based on factoring out the degree-based estimator from the classical (dual-component) estimator, and the other is based on factoring out the recruitment-based estimator from the classical estimator. This is less intuitive than it seems, because dual-component weights are more than just the product of the recruitment-based and degree-based weights. This procedure allows us in both cases to reduce the sampling weight into two numbers: one based on recruitment patterns between groups, the recruitment component (RC); and the other based on the degree distribution among individuals, the degree component (DC). Second, we simulate an array of RDS-samples on an empirically observed set of social networks, calculate the RC and DC for each sample, and measure performance for a set of classical, recruitment-based, and degree-based estimators. We then examine how performance varies by estimator type and the values of 'RC' and 'DC' that are calculated from the sample.

Instead of looking at the direct relationship of the population, social network, and sampling procedure on estimator performance, our analysis focuses on both (1) the relationship between different properties of the target population, network, and sampling procedure on the observed RC and DC of a sample, and (2) the relationship between a sample's location in 'RC-DC space' and the bias/performance of different RDS estimators. The first relationship sheds further light on the mathematical mechanisms that cause different RDS-estimators to be more appropriate for different populations. The second relationship enables field analysts to select RDS estimators based exclusively on data from peer recruitments and self-reported network sizes in their sample, as opposed to speculation about the social processes underlying the recruitment process and network of participants. In a final empirical section, we demonstrate how this method of assessing estimator-selection may be used in the field with a real example of RDS data. We then briefly discuss the implications of our findings.

2. An overview of respondent-driven sampling

RDS provides a means of drawing probability samples for populations which cannot be effectively sampled using traditional population survey methods because they lack a sampling frame, and because these populations are hard for outsiders to penetrate due to stigma or privacy concerns. Many RDS studies have largely focused on populations of relevance to public health, such as injection drug users (IDUs), men who have sex with men (MSM), and commercial sex workers (CSW); on populations of relevance to arts and culture such as jazz Heckathorn and Jeffri, 2001) and visual artists (Jeffri et al., 2011) ; on hard-to-reach or rare general populations such as low-wage workers (Bernhardt et al., 2013) and Canadian urban aboriginals (Smylie et al., 2011); and on vulnerable populations such as under age sex trafficking victims (Curtis et al., 2008). A 2009 survey analyzed the results of 128 studies drawn from more than 28 countries (Johnston et al., 2013). RDS has been employed in studies funded by agencies including CDC, CDC/Global AIDS, Gates India, USAID, NSF and NIH institutes including NIDA, NIMH, NICHD and NINR.

The recruitment methods and statistical techniques developed by RDS researchers has also been extended to study of non-hidden populations. RDS data can also be used to gain a better understanding of the network features of non-hidden populations, such as rates of racial homophily on college campuses (Wejnert, 2010), and to generate demographic point estimates of fully connected populations (Wejnert and Heckathorn, 2008).

2.1. Recruitment procedure

Drawing an RDS sample involves several steps. First, when sampling from a hidden population, one begins with a convenience sample of initial respondents who serve as "seeds." Seeds can be identified by key informants who are drawn from organizations where the target population congregates, or they may self-identify by volunteering for the study. Second, the initial respondents each recruit several peers, who compose the sample's first "wave." Third, the first-wave recruits each recruit several peers, who form the sample's second wave. The sample expands in this recursive manner, wave by wave, with the prior wave's recruits becoming the recruiters of the subsequent wave, until the desired sample size has been reached.

One essential feature of RDS sampling is keeping track of who recruited whom. This is important because affiliation patterns (e.g., members of a racial/ethnic group tending to recruit members of the same racial/ethnic group) affect the composition of the sample. A second essential feature is asking each respondent how many members of the target population they know as acquaintances, friends, or closer than friends, or employing an alternative means such as a scale-up method for assessing personal network size. RDS estimators employ information about respondents' degrees to correct for a source of bias inherent in chain referral sampling: all else equal, nodes with larger personal networks are more likely to be sampled. In the computation of any RDS estimator, the estimated size of each of the population's subgroups is inflated or deflated based on whether the subgroup was judged to be under or over-sampled. In sum these two features of the sampling process allow RDS estimators to function as a corrective lens that compensates for network-based sources of bias in the sampling process.

The popularity of RDS derives in part from proofs showing that when the assumptions of the method are satisfied, population estimates are asymptotically unbiased (Heckathorn, 2007; Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008). This means that bias is only on the order of $1/n$, where n is the sample size, so bias is trivial in samples of significant size. Unbiased estimates depend on four conditions: (1) the network connecting the population is dense enough to form a single component; (2) recruiters know one another (as acquaintances, friends, or those closer than friends) and their relationships are reciprocal, so any individual's in-degree and out-degree are equivalent; (3) respondents recruit as though they are selecting randomly from their neighborhoods; and (4) sampling occurs with replacement.

2.2. RDS estimators and selection

RDS estimators use data from the sample to compute the mean level of a certain variable within the target population. In this article we focus on estimating the prevalence of categorical variables. The "naïve" estimate of a population proportion is simply the number of people in the RDS sample that belong to the group of interest, divided by the total number of people in the sample. There are two main sources of bias that undermine the accuracy of this estimate, and both are accounted for in different RDS estimators (at the cost of higher levels of variance).

In a random walk on a network, individual nodes are encountered with a probability that is proportional to their degree (the count of other nodes that they are connected to with an edge, or neighbors). For example, an individual with degree 6 in a network will be encountered twice as frequently as an individual with degree 3. To the extent that a network-sample can be thought of as a random walk, degree-based estimates use the self-reported degree of respondents (the number of other individuals who they know in the target population) to correct for this bias. We refer to estimators that exclusively account for differences in degree as "degree-based" estimators. In especially small samples, this procedure can be inadequate. Individuals cannot be selected to a sample more than once, which means the relationship between the probability of being selected into the sample and nodal degree becomes less linear as the sampling fraction (the proportion of the target population in the sample) increases. The 'Successive-Sampling' estimate offers a way to correct for this problem and is useful when the target population from which the sample is drawn is thought to be small (Gile, 2011), although this estimate requires an analyst estimate of the size of the target population.

The other source of bias comes from differential recruitment within a graph: individuals may be homophilous (or heterophilous) with regards to the variable of interest, meaning they are more (or less) likely to affiliate with individuals who are similar on the variable of interest. Homophily is one of the most robust findings in all of sociology (McPherson et al., 2001) and can cause clustering in the social network that a network-sampling procedure "walks" through. This by itself will cease to be a problem in large enough samples, as a true random walk on a graph will ultimately select each node with a probability proportional to degree, however, the more homophilous a network, the longer it will take the biased selection of nodes in a random walk to purely reflect differences in degree. Furthermore, if individuals in different groups are more likely to actively recruit other participants, this can amplify the effects of homophily in a way that cannot be accounted for by a large sample or with information on respondent network sizes.

An early RDS-estimator accounted for this by calculating the Markov equilibrium distribution of different groups given rates of cross recruitment between them (Heckathorn, 1997). Later estimators have attempted to account for both differential recruitment and differences in degree simultaneously (Heckathorn, 2002, 2007; Salganik and Heckathorn, 2004). We refer to these as 'classical' estimators.

The type of estimator that is most reliable varies from situation to situation and survey to survey. Respondent-driven samples can vary along several dimensions. Some of these dimensions are controlled by the design of the survey (e.g., how large is the sample, how many seeds should be used, and how many peers should each person be allowed to recruit), but many dimensions are beyond the control of the researcher. The social structure of the population being sampled is significant (e.g., how strongly segmented is the population along different dimensions of interest such as race or gender, and how do the network sizes of individuals vary within and among these dimensions). The recruitment behavior of different groups within the target population is also significant (e.g., members of some groups are more likely to recruit their peers than members of other groups).

Though several assessments of RDS estimators have been conducted, to our knowledge none combine known population baselines with realistic sampling behavior and empirically observed network structures. Several studies have used simulated samples on networks to assess the effectiveness of RDS in general (Goel and Salganik, 2010), and to compare the performance of different RDS estimators by systematically varying different features of the sampling process (Gile and Handcock, 2010; Tomas and Gile, 2011) Other assessments of RDS have used data from actual RDS studies and compared RDS estimates to other population estimates (McCreesh et al., 2012). The simulation approach allows researchers to make sure the sampling process does not violate the assumptions built in to RDS estimators, but often relies on simulated networks. Studies using actual RDS samples operate on organic social networks, but do not guarantee that assumptions of RDS estimators are met – making theoretical evaluations about estimator performance problematic.

We proceed through a formal analysis and then simulate RDS-samples across-empirically observed networks to determine when different RDS estimators are more appropriate. First, we analytically decompose the sources of bias in an RDS sample. This is done by using a decomposition of RDS sampling weights (i.e., W) into a degree component, (i.e., DC), and a recruitment component, (i.e., RC) where the product of the two components yields the sampling weight (DC * RC═W)5. DC captures variation in sampling weights due to difference in groups' mean network sizes; and RC captures variation in sampling weights due to different recruitment patterns among groups. We propose two different decompositions of a classical (dual-component) sampling weight that incorporates both of these effects. These can be used to create two different sets of 'RC-DC spaces', which are useful in determining which estimators are appropriate. Data from each sample can be translated into a single coordinate in each RC—DC space. We also identify conditions where the RC is neutral, and thus can be ignored.

Second, we examine the predictive validity of each RC—DC space using a simulation-based experiment. We run 250,000 random RDS samples across empirically observed networks drawn from the "Facebook 100″ data set (Traud et al., 2012). For each simulation, we randomly select one of the 100 observed univesity networks, a target class (freshman, sophomore, junior or senior) group within the university, a level of branching (i.e., the quota for peer recruitments in the sampling design), and a level of differential recruitment. For each sample the difference between each estimators' prediction and the true population value can be mapped to the sample's observed RC and DC. We show that a sample's location in RC—DC space is a strong predictor of estimator buas and performance. Finally, we illustrate the application of our findings using data from two real-world studies of Latino GLBT in Chicago and San Francisco. This underscores a main practical aim of our study: the identification of principles that can be applied in the field to select the optimal RDS estimator.

3. Formal models of recruitment component and degree component

Our formal analysis of recruitment component and degree component consists of four steps. First, we will introduce how data from RDS samples have been used to create sampling weights that are based on degree biases, recruitment biases, or both (dual-component or classical sampling weights) (Eqs. 1–6). Second, we will show how the weight based on recruitment bias can be isolated from the dual-component weight to create an estimate of the recruitment component (RC) and degree component (DC) of sample bias (Eq.s 7–13). Third, we will show how the weight based on degree bias can be isolated from the dual-component weight to create an alternative estimate of the recruitment component (RC`) and degree component (DC`) of sample bias (Eq.s 14–19). And fourth, we will use this to determine when RC is 'neutral' (RC = 1) in a sample (Eq.s 20–24). This is important because correcting for recruitment biases unnecessarily may introduce more variance into sample estimates.

3.1. Constructing sampling weights and estimates from RDS data

In any respondent-driven sample where the estimate of concern is the true proportion of group X within a target population, a two-by-two recruitment matrix of who recruits whom can be produced between members of group X, and those not in group X, who will be referred to as group Y. R_{XX} is the number of recruitments of members of X by members of X, R_{XY} is the number of recruitments of members of Y be members of X, and so on. The sample can be defined by a two-by-two recruitment matrix, R, the sum of whose cells equals the sample size (less the 'seeds' in the sample who are not recruited by another respondent), N.

$$R = \begin{bmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{bmatrix}, N = R_{XX} + R_{XY} + R_{YX} + R_{YY} \quad (1)$$

This recruitment matrix can also be transformed into a transition matrix, S, where each cell is the probability that a member of a specific group recruits a member of another group

$$S = \begin{bmatrix} \frac{R_{XX}}{R_{XX}+R_{XY}} & \frac{R_{XY}}{R_{XX}+R_{XY}} \\ \frac{R_{YX}}{R_{YX}+R_{YY}} & \frac{R_{YY}}{R_{YX}+R_{YY}} \end{bmatrix} = \begin{bmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{bmatrix} \quad (2)$$

In the absence of any weighting procedures that account for differences in sampling probability between nodes, the estimate of the true portion of group X within a target population is simply the composition of the sample. This is the composition estimate, C_X, also referred to as the "Naïve Estimate".

$$C_X = \frac{R_{XX} + R_{YX}}{N} \quad (3)$$

Cross recruitments between members of group X and non-members (group Y), can be used to estimate the equilibrium proportion of both groups. The proportion of group X in equilibrium, E_X, is the proportion of recruitments from members of Y being directed to X, S_{YX}, divided by the sum of this and the proportion of recruitments directed from X to Y, S_{XY}. (Heckathorn, 1997)

$$E_X = \frac{S_{YX}}{S_{YX} + S_{XY}} \quad (4)$$

Subsequent work includs information on the network sizes of members of both groups, and how that effects their probability of being included in the sample. The new population estimate, P_x, is defined as:

$$P_X = \frac{D_Y S_{YX}}{D_Y S_{YX} + D_X S_{XY}} \quad (5)$$

P_X is considered a "dual-component" or "classical" estimator, as it considers both the effects of differences in degree between groups and differential rates of cross-recruitment.

The sample weight, W_x, the factor by which we must adjust our composite estimate of X to account for sampling biases, can be defined as P_X divided by the composition point estimate (the simple proportion of individuals in the sample who belong to X), C_X.

$$W_X = \frac{P_X}{C_X} \quad (6)$$

3.2. Decomposing the sample weight

The total sample weight, W_X, can be decomposed into a degree component (DC_X) and recruitment component (RC_X). (Heckathorn, 2007) Part of the sampling weight accounts for differences in network size by group, and the remainder accounts for recruitment patterns between groups.

$$W_X = DC_X RC_X \quad (7)$$

The RC is defined as the ratio between the equilibrium estimate, E_X, and the composition estimate, $C_{x.}$.

$$RC_X = \frac{E_X}{C_X} \quad (8)$$

We can now define the recruitment component of group X as a function of the recruitment matricies R and S, by inserting Eqs. (3) and (4) into Eq. (8):

$$RC_X = \frac{\frac{S_{YX}}{S_{YX}+S_{XY}}}{\frac{R_{XX}+R_{YX}}{N}} = \frac{\left(\frac{R_{YX}}{R_{YX}+R_{YY}}\right)}{\frac{R_{XX}+R_{YX}}{R_{XX}+R_{XY}+R_{YX}+R_{YY}}} \quad (9)$$

The degree component of group X is obtained by inserting Eq. (8) and Eq. (6) into Eq. (7):

$$\frac{P_X}{C_X} = DC_X \frac{E_X}{C_X} \quad (10)$$

Which means that DC_X can be defined as:

$$DC_X = \frac{C_X}{E_X}\frac{P_X}{C_X} = \frac{P_X}{E_X} \quad (11)$$

Using S and R the probability of selection for both groups according to network size (D_X and D_Y), we can calculate the DC as follows:

$$DC_X = \frac{\frac{D_Y S_{YX}}{D_Y S_{YX} + D_X S_{XY}}}{\frac{S_{YX}}{S_{YX}+S_{XY}}} \quad (12)$$

This term, while thought to account for the residual left after differential recruitment is accounted for, still includes many terms that address patterns of cross recruitment. Algebraically this simplifies to[i]:

$$DC_X = \frac{\frac{S_{YX}}{S_{XY}} + 1}{\frac{S_{YX}}{S_{XY}} + \frac{D_X}{D_Y}} \quad (13)$$

From this equation we can see that while the degree component will be neutral (= 1) if $D_X = D_Y$, DC_X will otherwise be a function of cross recruitment patterns if $D_X \neq D_Y$.

3.3. An alternative decomposition

Alternatively, the sampling weight W_X can be decomposed by expressing degree component as the ratio of the degree-based estimate of the proportion of X, G_x, to the composition estimate, and by expressing recruitment component as the ratio of the dual-component estimate, P_X, to the degree-based estimate, G_X. We name these new components DC_X' and RC_X'.

$$W_X = DC'_X RC'_X \quad (14)$$

$$DC'_X = \frac{G_X}{C_X} \quad (15)$$

$$RC'_X = \frac{P_X}{G_X} \quad (16)$$

The degree-based estimate is simply the naïve estimate of X adjusted by the average (or harmonic average) of the network sizes of groups X and Y, which are represented by D_x and D_y.

$$G_X = \frac{(R_{XX} + R_{YX})}{N}\frac{D_Y}{D_X} \quad (17)$$

This gives us a straightforward formula for DC_X':

$$DC'_X = \frac{\left(\frac{(R_{XX}+R_{YX})}{N}\frac{D_Y}{D_X}\right)}{\left(\frac{(R_{XX}+R_{YX})}{N}\right)} = \frac{D_Y}{D_X} \quad (18)$$

And the following expression for RC_X':

[i] This formula is acquired through the following rearranging: $\frac{\frac{D_Y S_{YX}}{D_Y S_{YX}+D_X S_{XY}}}{\frac{S_{YX}}{S_{YX}+S_{XY}}} = \frac{D_Y S_{YX}}{D_Y S_{YX}+D_X S_{XY}} * \frac{S_{YX}+S_{XY}}{S_{YX}} = \frac{D_Y(S_{YX}+S_{XY})}{D_Y S_{YX}+D_X S_{XY}} = \frac{D_Y S_{YX}+D_Y S_{XY}}{D_Y S_{YX}+D_X S_{XY}} = \frac{S_{YX}+S_{XY}}{S_{YX}+\frac{D_X}{D_Y}S_{XY}} = \frac{\frac{S_{YX}}{S_{XY}}+1}{\frac{S_{YX}}{S_{XY}}+\frac{D_X}{D_Y}}$

$$RC_X^{'} = \frac{\frac{D_Y S_{YX}}{D_Y S_{YX} + D_X S_{XY}}}{\frac{(R_{XX}+R_{YX})}{N}\frac{D_Y}{D_X}} = \frac{\frac{D_X S_{YX}}{D_Y S_{YX} + D_X S_{XY}}}{\frac{(R_{XX}+R_{YX})}{N}} \quad (19)$$

This expression, unfortunately, is now dependent on the network sizes of group X and group Y. However, there are now 4 single dimensions (RC_x, DC_x, RC_x', and DC_x'), and two 2-dimensional spaces (RC_X x DC_X, and RC_X' x DC_x') that can be used to predict the accuracy of various RDS estimators. These are summarized in Table 1.

3.4. Conditions of recruitment component neutrality

The equations listed in Table 1 can also help us determine when estimation might benefit from ruling out the recruitment component entirely. In these situations, a degree-based estimator would provide an unbiased estimator, and accounting for recruitment components would be an unnecessary source of variance. From the equation in Table 1, we can set the denominator equal to the numerator and see that RC_X will be neutral when the following equality holds:

$$\frac{S_{YX}}{S_{YX}+S_{XY}} = \frac{R_{XX}+R_{YX}}{N} \quad (20)$$

This equation can be re-expressed purely in terms of cells from the Recruitment Matrix, R:

$$\frac{\frac{R_{YX}}{R_{YX}+R_{YY}}}{\frac{R_{YX}}{R_{YX}+R_{YY}}+\frac{R_{XY}}{R_{XX}+R_{XY}}} = \frac{R_{XX}+R_{YX}}{R_{XX}+R_{XY}+R_{YX}+R_{YY}} \quad (21)$$

And further simplified to:

$$\frac{R_{YX}(R_{XX}+R_{XY})}{R_{YX}(R_{XX}+R_{XY})+R_{XY}(R_{YX}+R_{YY})} = \frac{R_{XX}+R_{YX}}{R_{XX}+R_{XY}+R_{YX}+R_{YY}} \quad (22)$$

Upon further inspection, this equation holds when one of the three following conditions are true[ii]:

$$RC_X = 1 \text{ if } R_{XX}R_{YY} = R_{XY}R_{YX} \text{ or } R_{XY} = R_{YX} \quad (23)$$

These conditions correspond to two situations: one where there is no association between the group membership of the recruiters and the recruited, and a second situation where the number of cross recruitments from each group to the other are equal. This second type of "RC neutrality" has an interesting implication: deviance from neutrality is limited by the number of seeds and the level of branching in the sample: an RDS study with one seed where each recruiter can only recruit one participant can not produce a difference in cross-recruitment cells that exceeds one.

The conditions that lead to RC neutrality in the second decomposition ("RC' neutrality") can also be calculated. In this case, the recruitment component is neutral when the following equality holds:

Table 1
Formulas for both the degree component and recruitment component from decompositions of the dual component sampling weight, using both the equilibrium estimator and a purely degree-based estimator.

	Decomposition with E_X	Decomposition with G_X
Degree Component	$DC_X = \frac{\frac{S_{YX}}{S_{XY}}+1}{\frac{S_{YX}}{S_{XY}}+\frac{D_X}{D_Y}}$	$DC_X^{'} = \frac{D_Y}{D_X}$
Recruitment Component	$RC_X = \frac{\frac{S_{YX}}{S_{YX}+S_{XY}}}{\frac{R_{XX}+R_{YX}}{N}}$	$RC_X^{'} = \frac{\frac{D_X S_{YX}}{D_Y S_{YX}+D_X S_{XY}}}{\frac{(R_{XX}+R_{YX})}{N}}$

$$\frac{(R_{XX}+R_{YX})}{N} = \frac{D_X S_{YX}}{D_Y S_{YX} + D_X S_{XY}} \quad (24)$$

While there is one condition where this equality will hold when all cells of the recruitment matrix are greater than 0, it does not lend itself to any clear intuitions about RC' neutrality[ii].

4. Simulation experiments on empirically observed networks

We now simulate a series of respondent-driven samples on observed empirical networks to observe the relationship between estimator perofmrance and RC—DC space. The data comes from the "Facebook 100″ sample, which consists of the networks of Facebook friendships from 100 American universities as observed during a single day in September 2005 (Traud et al., 2012). This data set has been used in other studies of RDS estimators (Verdery et al., 2015). In each simulation, the proportion of a college population that is in a class (i.e., freshman, sophomore, junior, or senior) is estimated by each of the aforementioned estimators. The differing structures of each school's network, the differing positions of each class in the overall school structure, and 'experimentally' varied levels of branching and differential recruitment are used to draw samples that have varying recruitment and degree components. Having a range of samples from different areas of RC-DC space allows us to map where different estimators will perform more favorably.

In other simulation-based studies of RDS estimators, the bias of different estimators is considered a direct function of the target population's underlying social network, the constraints of the sampling procedure, and the recruitment behaviors of simulated participants. However, in the field the true nature of the target population's network and recruitment behaviors can only be estimated from the sample. Our simulation analysis decomposes results into two distinct parts: analyzing the recruitment matrix and degree distribution (in terms of DC and RC) as a function of the underlying traits of the network and the sampling procedure, and then assessing the performance of different estimators on the basis of observed levels of DC and RC in the recruitment matrix. This augments prior simulation-based assessments of RDS estimators because it highlights the mathematical mechanisms that connect the social reality of the target population to the performance of different RDS estimators. More importantly, this helps analysts in the field because it illuminates how they might directly select an RDS estimator based purely on the details of data from the sample – as opposed to speculating on the ground truth characteristics of their target population and choosing an RDS estimate accordingly(This process is outlined in Fig. 1).

5. Procedure

While most comparisons of respondent-driven sampling simulations systematically vary one or more parameters and compare estimator performance, the approach taken here is modified to ensure that coverage is spread across "RC-DC" space. First, we estimate the proportion of a certain class year in a college's Facebook community using 250,000 simulated RDS samples. The procedure is as follows: a school is randomly selected from the Facebook100 dataset; one of the four classes of the university is randomly chosen; one of three different levels of differential recruitment (i.e., in the simulation the target class recruits one half as much, equally, or twice as much as the mean for other classes, making the value of differential recruitment 0.5, 1, or 2) is randomly chosen; and one of 4 different levels of branching is chosen (i.e., each respondent has a maximum recruitment quota of 1, 2, 3, or 4). Next, beginning with 5 randomly selected seeds, a respondent-driven sample of 300 individuals is drawn. Individuals cannot be sampled more than once so the procedure involves sampling without replacement. In the Facebook data set, different schools exhibit varying levels of homophily by class year, and also differ in terms of overall network density. Also, different class years within each school tend to have

[ii] The condition is the following: $\frac{D_Y}{D_X} = \frac{R_{XX}^2 R_{YX} + R_{XX}R_{XY}R_{YX} - R_{XX}R_{XY}R_{YY} + R_{XX}R_{YX}^2 + R_{XX}R_{YX}R_{YY} + R_{XY}^2 R_{YX}}{R_{YX}(R_{XX}+R_{XY})(R_{XX}+R_{YX})}$

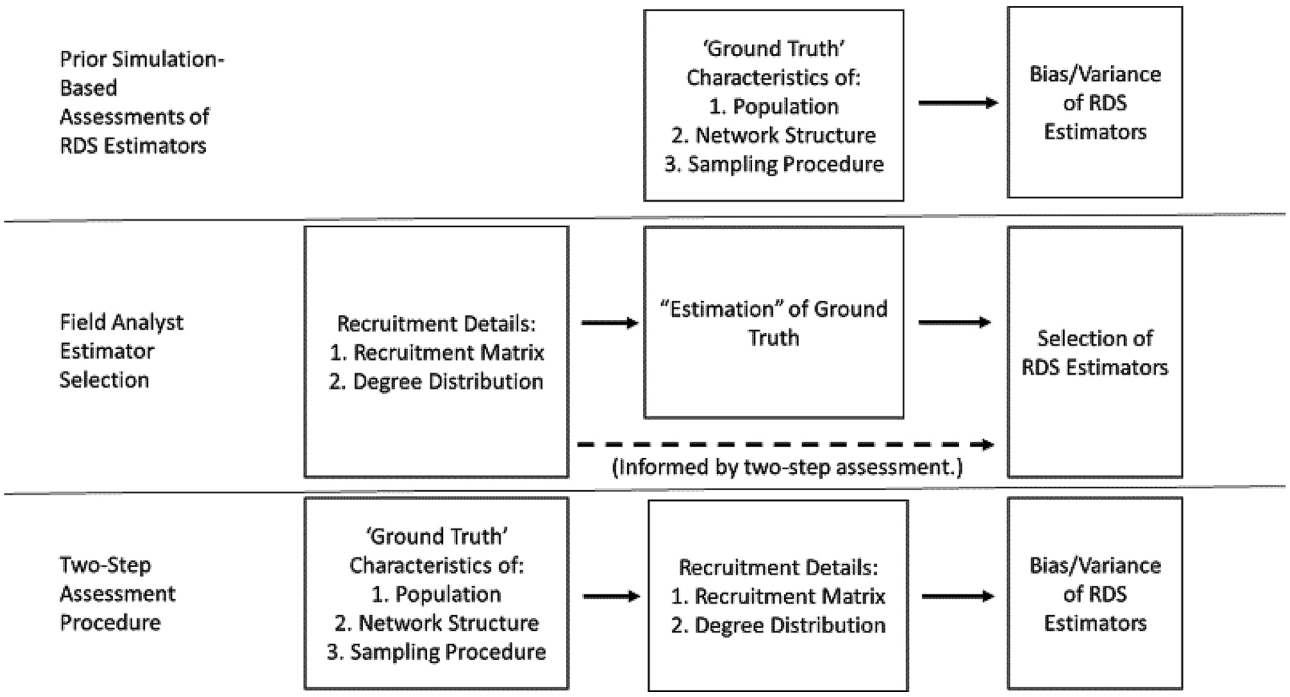

Fig. 1. Comparison of prior simulation-based assessments of RDS estimators (top), which relate estimator performance to unobservable details of the network and population, how analysts in the field ideally select estimators (middle) (the dotted line indicates how this process may be improved by our approach), and the two-step assesement procedure conducted in this article (bottom).

different positions within the overall network structure (i.e., freshman tend to be more isolated, juniors tend to be more central, etc.). Because the size of the sample is fixed at 300, the sampling fraction is also varied by randomly selecting from university networks of different sizes. For each of the 250,000 samples, the proportion of the class year of interest is estimated using a series of RDS estimators as well as the naïve estimator (the unweighted sample proportion).

5.1. RC and DC as a function of the underlying network, population, and sampling procedure

Of the 250,000 simulated samples, roughly 96 % (239,637) are successful, which is to say that none of the recruitment chains stemming from each of the 5 initial seeds "stall" before reacheing (300/5 – 1 =) 59 other recruits. Stalling occurs when every branch of a recruitment chain is incapable of recruiting an individual who has not yet been recruited into the study.[iii] Fig. 2depicts how each of three variables (differential recruitment, branching, and class year) influence both decompositions of the recruitment component and degree component.

Differential recruitment, as would be expected, decreases the recruitment component of the target class when the target class is twice as likely to recruit, and increases the recruitment component of the target class when they are half as likely to recruit. Differential recruitment, however, does not have an apparent influence on the degree component of the target class.

Branching, as predicted in the formal analysis of neutral recruitment-component matricies, increases the variance of the recruitment component. A sample with branching equal to 1 will have off-diagonal cells in the recruitment matrix that are nearly equal to each other, keeping the sample very close to Type II neutrality (the total of recruitments from outside of a group to inside a group are equal to the number of recruitments from inside of a group to the outside of a group) in all situations. While this prediction derives from an analysis of the first decomposition of RC—DC Space, increased variance in both RC_X and $RC_{X'}$ is associated with increased levels of branching.

Varying class year effects both degree component and recruitment component because each class year in the network has a different position in the network. Individuals are generally homophilous on the basis of class year, and certain class years have a higher average degree than others. Freshman, who typically have the lowest average degree (presumably because they have had the least amount of time to establish social networks), tend to have a higher degree component on average. Their smaller networks cause them to be under-sampled, so a larger degree component inflates their estimated population size to compensate for this bias in the sample. Freshman also show the highest variance in recruitment component, perhaps because they tend to be more 'clustered' (they direct a higher proportion of ties to themselves) than the other classes, causing more deviance from Type I neutrality (which holds when both groups recruit from other groups randomly).

The variable levels of the degree component and the recruitment component in the samples allow us to relate the performance of each RDS estimator to two variables that are easy to obtain from an RDS sample. This has profound practical implications, because in the field analysts may not be able to get a sense for the true underlying network structure or the social processes that are driving recruitment. Yet in our simulations DC and RC can be directly observed, and in the field these values can easily be calculated. Mapping estimator performance onto a directly observable 'RC-DC' space will be useful to the analyst trying to determine which estimates are most trustworthy in different situations.

6. Results

For each of the samples, six RDS estimators are used to compute predictions of the population proportion for a randomly chosen class year. The first is the simple naïve estimator (the proportion of the class year of interest observed in the sample), and the second is the 'equilibrium estimator', which is based solely on cross-group recruitment (Heckathorn, 1997). These two estimators are neither 'classical' nor 'degree-based' but provide a useful baseline for comparison. There are two 'classical' or 'dual-component' estimators that use information about average nodal degree and cross-group recruitment levels, one weights the equilibrium estimator by factoring in the arithmetic mean

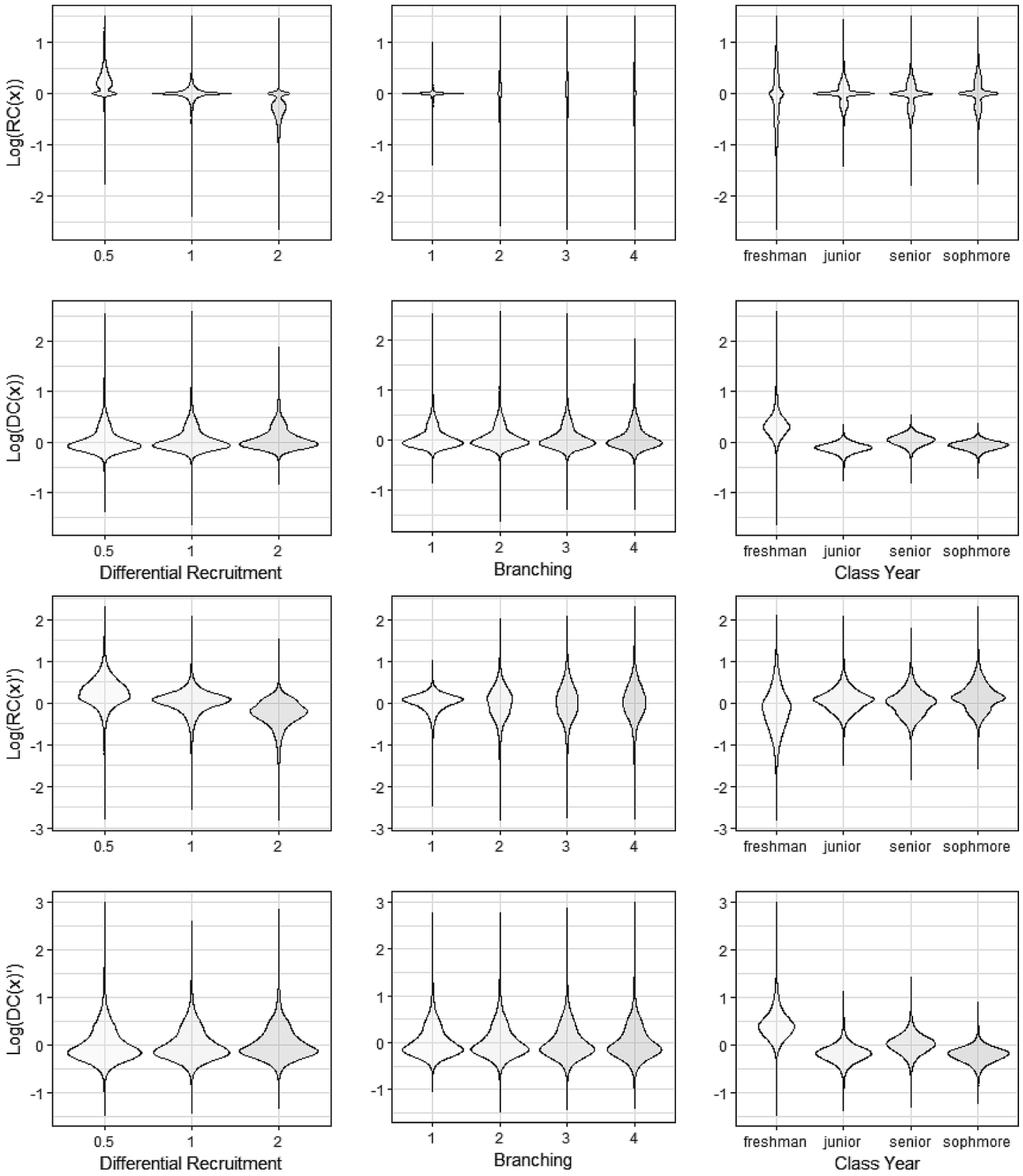

Fig. 2. A & 2B - Variance in the recruitment component and degree component (decomposed by the equilibrium estimator in 2A (top), and the degree-based estimator in 2B (bottom) are induced by variations in differential recruitment (the amount the class of interest recruits compared to other classes), branching (the target number of recruits per recruiter), and class year (which roughly corresponds to a group's overall position in the network structure). Variation in these three variables produces variation in RC, DC, RC' and DC' .

degree of each group (Heckathorn, 2002), and the more recent estimator takes degree component into account through a more sophisticated measure (the total members of the group divided by the sum of the inverse of each member's degree count, i.e., the geometric mean). We will refer to these estimators as "Classical I" and "Classical II", respectively. Finally, we include two degree-based estimators. One weights individuals by the inverse of their degree count (Volz and Heckathorn, 2008), the other uses a "successive sampling" procedure which better accounts for the probability of repeat sampling (Gile, 2011).This latter estimator requires an estimate of overall population size.We provide the known size of the school's student population so the estimator's performance will be at its theoretical peak. We will subsequently refer to these estimators as "Degree-Based" and "Degree-Based: SS"

Of the 239,647 successful samples, the results presented will focus on

samples located within a conservative subsection of the two RC-DC spaces: where both RC and DC are between 2/3 (0.67) and 3/2 (1.5). In the first RC-DC space, this results in the inclusion of 177,610 of the 239,647 (74.1 %) simulated samples, and in RC'-DC' space this results in the inclusion of 156,662 (65.4 %) of samples. This is done to avoid drawing inferences from portions of the RC-DC space where the data is thin, and to focus attention on scenarios where the observed RC and DC are not extreme enough to otherwise rouse the suspicions of the analyst.

6.1. Bias by estimator in RC-DC and RC'-DC' space

The mean bias for each area in the RC-DC space across the 6 estimators of interest in both subspaces is shown in Fig. 3. After log transforming the recruitment component and the degree component,

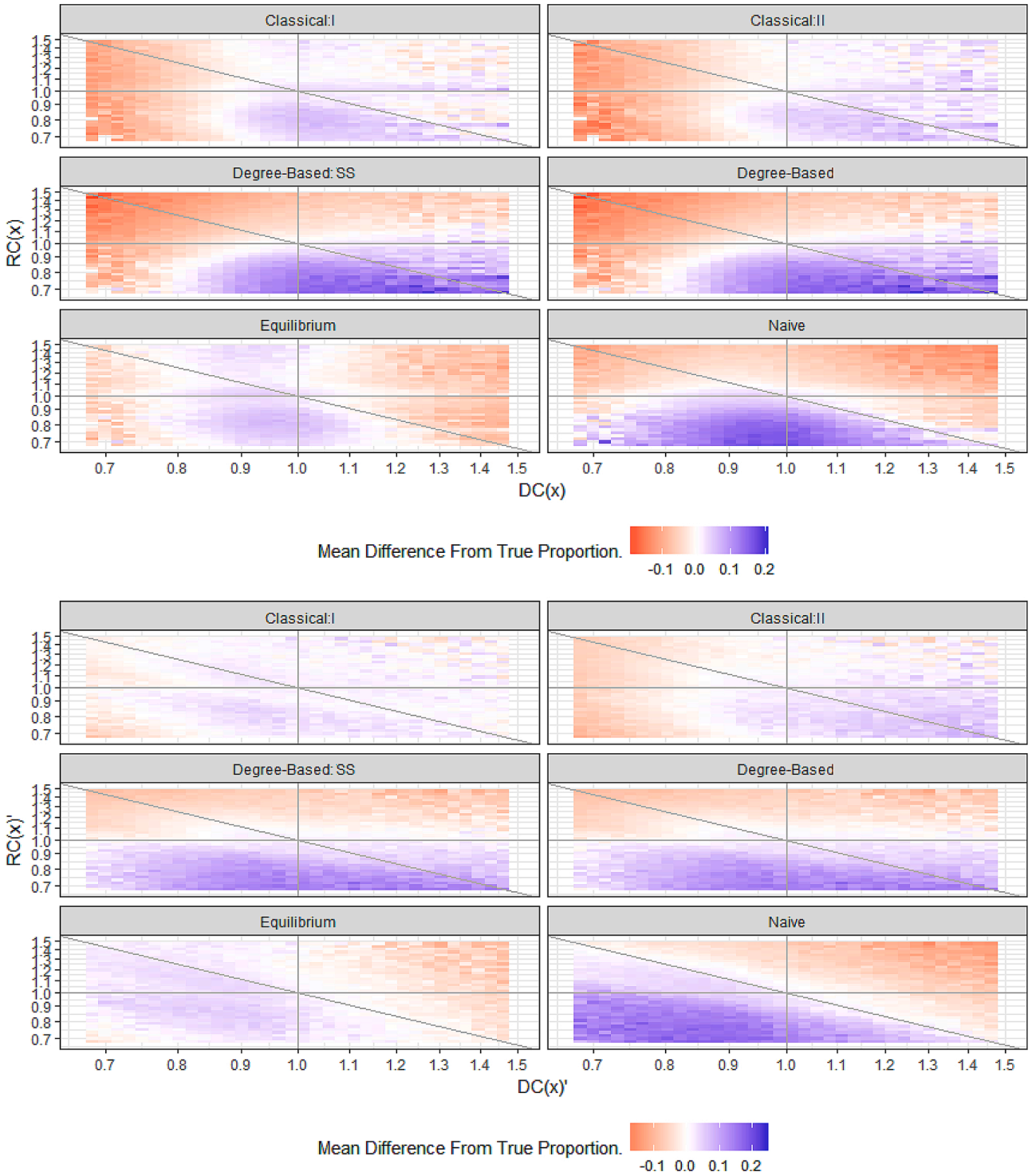

Fig. 3. A & 3B - Heat plots showing the mean overestimation or underestimation of the class year of interest in a school's Facebook network for each of six estimators. Each cell corresponds to a 0.01*0.01 region of log transformed RC-DC space (Top, Figure 3A) and RC'-DC' space (Bottom, Figure 3B). In Figure 3B, the two degree-based estimators, seen in the middle row, both categorically overestimate in the case of low recruitment component and underestimate in the case of a large recruitment component. This suggests that degree-based estimators are inappropriate where the sample has a substantially non-neutral recruitment component.

estimates are binned into .01 * .01 cells. The two degree-based estimators both have a mean bias that is very tightly coupled with the observed recruitment component. In RC-DC space (Fig. 3A), the mean bias of degree-based estimators is very close to zero when RC is neutral or positive, with a hint of slight overestimation when the RC is greater than one. The estimators drastically underestimate, however when the RC is less than one. In RC'-DC' space, the bias of the degree-based estimates align more closely with predictions: estimates are unbiased when RC is neutral, overestimate when RC is negative, and underestimate when RC is positive.

Naïve estimator performance also aligns more closely with theoretical predictions in RC'-DC' space than in RC-DC space. The bias is closest to zero along the diagonal axis of RC'-DC' space, where both of these components should offset one another. The trough of neutrality

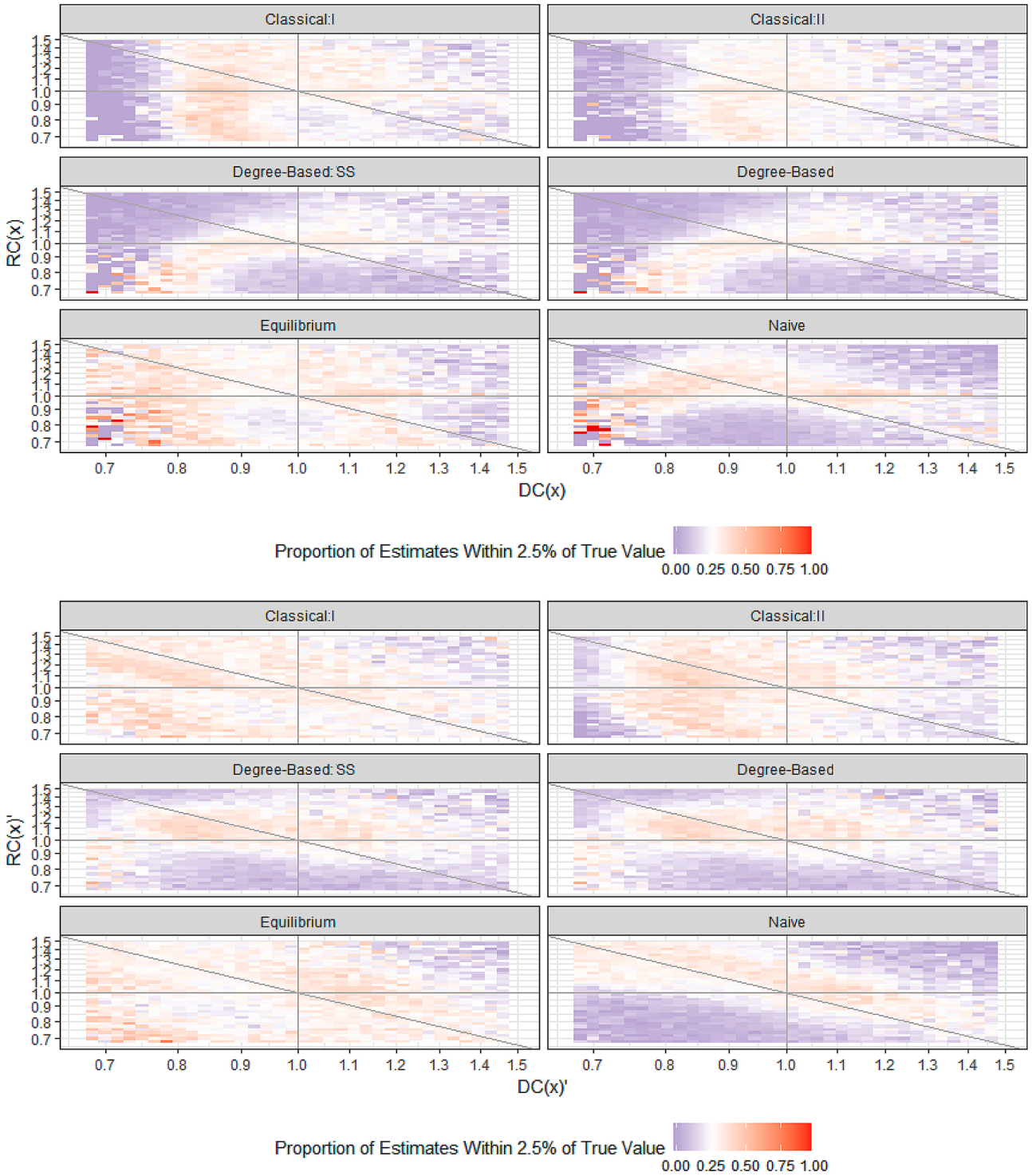

Fig. 4. A & 4B – Heat maps showing, for each of the six estimators and two 2D spaces of interest, the proportion of estimates that fall within 2.5 % of the true population proportion within each bin. The performance of the degree-based and the Naïve estimators tend to align better with theoretical predictions in 4B (bottom), where the theoretical space is generated by decomposition with the degree-based estimator, as opposed to the space shown in 4A (top). This figure also demonstrates that samples with a non-neutral recruitment component are better estimated by an estimator that controls for recruitment bias.

extending through RC-DC space on the other hand, is more curved, and extends through the left-half of the DC axis instead of the diagonal axis.

The equilibrium estimator tends to be more determined by the DC-axis than the RC-axis. However, the bias in general tends to be much less strongly associated with coordinate position. The two classical estimators also exhibit fewer "regional" biases in both RC-DC spaces, as predicted. In general, the levels of bias also tend to be influenced more by the value of either DC or DC', similar to the equilibrium estimator. But in the case of the "Classical I" estimate's level of bias in RC'-DC' space, it is very hard to discern any systematically high or low areas of

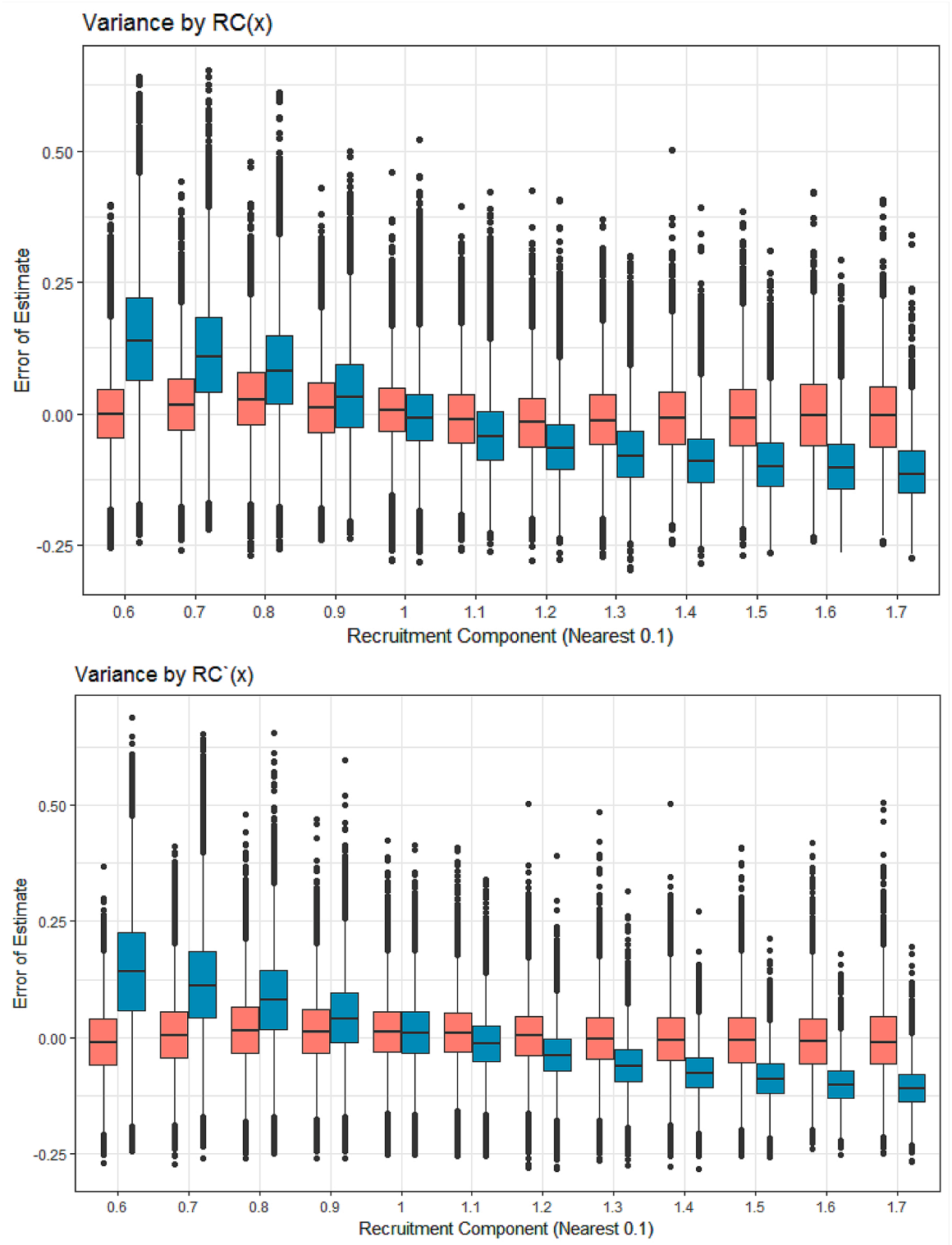

Fig. 5. 5a (top) and 5b (bottom) show the distribution of errors for the Classical I Estimator (red/left) and the Degree-Based Estimator (blue/right) across the range of observed RC (5a) and RC` values (5b). Each RC bin corresponds to a range of 0.1 (the bin '1′ for example, corresponds to 0.950 to 1.049). The bias of the degree-based estimator is sensitive to the RC, whereas the Classical I Estimator remains fairly consistent. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

bias.

6.2. Estimator accuracy in RC-DC and RC'-DC' space

The mean difference from the true population proportion is not necessarily a perfect predictor of which estimator will perform best in each scenario. An unbiased estimator with high variance around the true estimate may not perform as well as an estimate that is systematically biased but has low variance. The overall performance of an estimator is perhaps better judged by the proportion of times that it falls within a certain range of the true proportion.

Fig. 4A and B show the proportion of the estimates in each bin that falls within 2.5 % of the true proportion of the class of interest (e.g.,: if a school is 25 % freshman, then a positive outcome would be estimating between 22.5 % and 27.5 % freshman). This corresponds to the standard error for the estimate of a 25 % proportion within a 300-person sample. ($\sqrt{\frac{0.25(1-0.25)}{300}} = 0.025$) The high performing areas for each estimator reflect the analysis of previous mean bias by estimator: within RC'-DC' space, the degree-based estimators perform better when RC is neutral, and the naïve estimator performs well along the diagonal axis. In the RC-DC space the high-performing areas for these three estimators are more curvilinear. Classical estimators and the equilibrium estimator tend to show a wider range of high-performing areas in RC -DC space, although they are somewhat more clouded perform around values of DC-neutrality.

Taken together, these analyses suggest that the decomposition of sample weights into a recruitment and degree component, and more specifically a calculation of RC'_X, can be very useful in determining which estimators are appropriate for RDS samples.

In order to replicate how variance has been shown by estimator type in other assessments of RDS performance (e.g. Gile and Hancock 2010), Fig. 5A and B show boxplots for the estimated error by calculated RC. We see that for both measures of RC, as the RC level goes from less than 1 to greater than 1, the bias for the degree-based estimates shifts from positive to negative while the variance decreases. It is clear from the diagrams, that a higher proportion of classically derived estimates are within a small margin of error. Furthermore, for recruitment component values that are less than 1, the purely degree-based estimates are inferior in terms of both variance and bias.

To summarize the formal analysis and simulation-based results to this point, we have (1) shown how to calculate the potential bias from recruitment using sample data by calculating a recruitment component (RC) (2) shown that non-neutral recruitment components come from a combination of homophily and differential recruitment, and can be enhanced by branching, (3) used simulations to verify that the RC calculated from the sample is indeed effected by these three factors (see Fig. 2A and B), and (4) shown that a non-neutral RC can help us predict when a classical estimate is the optimal choice. We now turn to an example that can illustrate how important this technique could be in applied RDS research.

7. Two RDS studies: latino MSM in Chicago and San Francisco

In this section we relate the findings of the prior analytical and simulation-based findings to two real RDS datasets. We consider studies of HIV prevalence and socioeconomic factors among Latino gay and bisexual men and transgender persons in Chicago and San Francisco. The study was conducted in the summer and fall of 2004 (Ramirez-Valles et al., 2008). Data were collected from a sample of 643 individuals (Chicago: n = 320; San Francisco: n = 323) using computer-assisted self-administered interviews (CASI).

We focus on two variables from each of the two data sets: the proportion of individuals who are reported as HIV positive versus HIV negative or unknown, and the proportion of individuals who report speaking Spanish at home versus those who speak English at home. Both sets of recruitment and degree components for both groups in both settings are reported in Table 2. The samples are plotted in both RC—DC and RC`-DC` space in Fig. 6.

If we are interested in the proportion of the population in Chicago that speaks Spanish, or the proportion of the San Francisco population that has HIV-positive status, weighting may prove unnecessary, as the dual-component weight is very close to 1. In the case of estimating what proportion of the target population in San Francisco speaks Spanish, a degree-based estimator may be most appropriate: the RC` value is nearly neutral, but the value of DC` suggests there is a small but non-trivial bias that may result in the overestimation of this population.

Analysis of the sample suggests that estimation of the proportion of HIV-positive men in Chicago is prone to substantial bias. In this case, the highly non-neutral value of RC` and DC` both indicate that an estimate based purely on the Markov Equilibrium or the relative network neighborhood sizes of both populations would drastically over-estimate the proportion of men who have HIV. It is also worth noting that in this case, the way in which the dual-component weight is decomposed (the choice of RC-DC and RC`-DC` space) is of consequence. RC-DC space indicates that most of the bias is degree-based, and as such the improvements in variance from using a purely degree-based estimator may offer more accurate estimates. However, decomposing the sampling weight using the degree-based estimator shows that this is not the case: there are substantial differences between the dual-component sampling weight and the degree-based sampling weight. These differences emerge from recruitment bias, or the interaction of recruitment bias with the difference in nodal degrees between HIV-positive and HIV-negative members of the sample.

In the example data drawn on here, the differences between a purely degree-based estimate and dual-component estimates correspond to 1.2 %, 0.0 %, 2.9 %, and 3.0 % relative to the entire population. The difference between the two estimates as a proportion of the degree-based estimates are 3.2 %, 0.0 %, 20.3 % and 7.6 % respectively. The relative difference of 20.3 % corresponds to the HIV case rate in Chicago.

Differences of this magnitude can have large implications for public funding. For example, about 1 out of 40 Americans live in the Chicago area, and roughly $21.5 billion dollars was spent on domestic care and treatment for HIV in 2019 according to the Kaiser Family Foundation,[iii] which would correspond to roughly $500 million dollars for patients in the Chicago Area assuming that Chicago has HIV rates that reflect the United States as a whole. A 20 % difference in an estimate, like the one we just found for patients in Chicago, corresponds to about $100 million dollars annually in funding. While the allocation of funding is certainly more complicated, this does give us a sense of proportion for the importance of improving these estimates by even a small percentage.

8. Conclusion

Uncertainties regarding estimatorperformance have motivated many studies comparing the relative strengths of alternative estimators. The analytic and simulation components of this article suggest that the calculation of degree component and recruitment component can help analysts determine which estimators will minimize bias.

This is important because it suggests that directly observable information from the recruitment matrix and self-reported degree size can be used to select an appropriate estimator, regardless of the mechanisms that generate bias in the recruitent process. It may not be possible for a field analyst to know the true levels of homophily among groups of interest in a population, or how disproportionately likely one group is to recruit compared to another, but a sample's location in RC—DC and RC`-DC` space is easily calculated and immediately useful for selecting an appropriate estimator.

[iii] https://www.kff.org/hivaids/fact-sheet/u-s-federal-funding-for-hivaids-trends-over-time/

Table 2
The recruitment component, degree component, and dual component of two population proportions (Spanish Speaking and HIV Positive) each from two sample populations of Latino MSM in Chicago and San Francisco.

Population and Proportion of Interest	Total Sample Size	Naïve Estimate	RC_X	DC_X	RC'_X	DC'_X	Dual Component (RC*DC or RC'*DC')	VH Estimate (Naïve * DC')	Equilibrium Estimate (Naïve * RC)	Classical Estimate (Naïve * RC*DC)
San Francisco – HIV Positive	297	.380	1.006	1.009	1.033	0.982	1.014	0.373	0.382	0.385
San Francisco – Spanish Speaking	323	.260	1.031	1.069	0.998	1.104	1.102	0.287	0.268	0.287
Chicago – HIV Positive	265	.215	1.036	0.510	0.794	0.666	0.529	0.143	0.223	0.114
Chicago – Spanish Speaking	320	.416	1.072	0.950	1.076	0.947	1.019	0.394	0.446	0.424

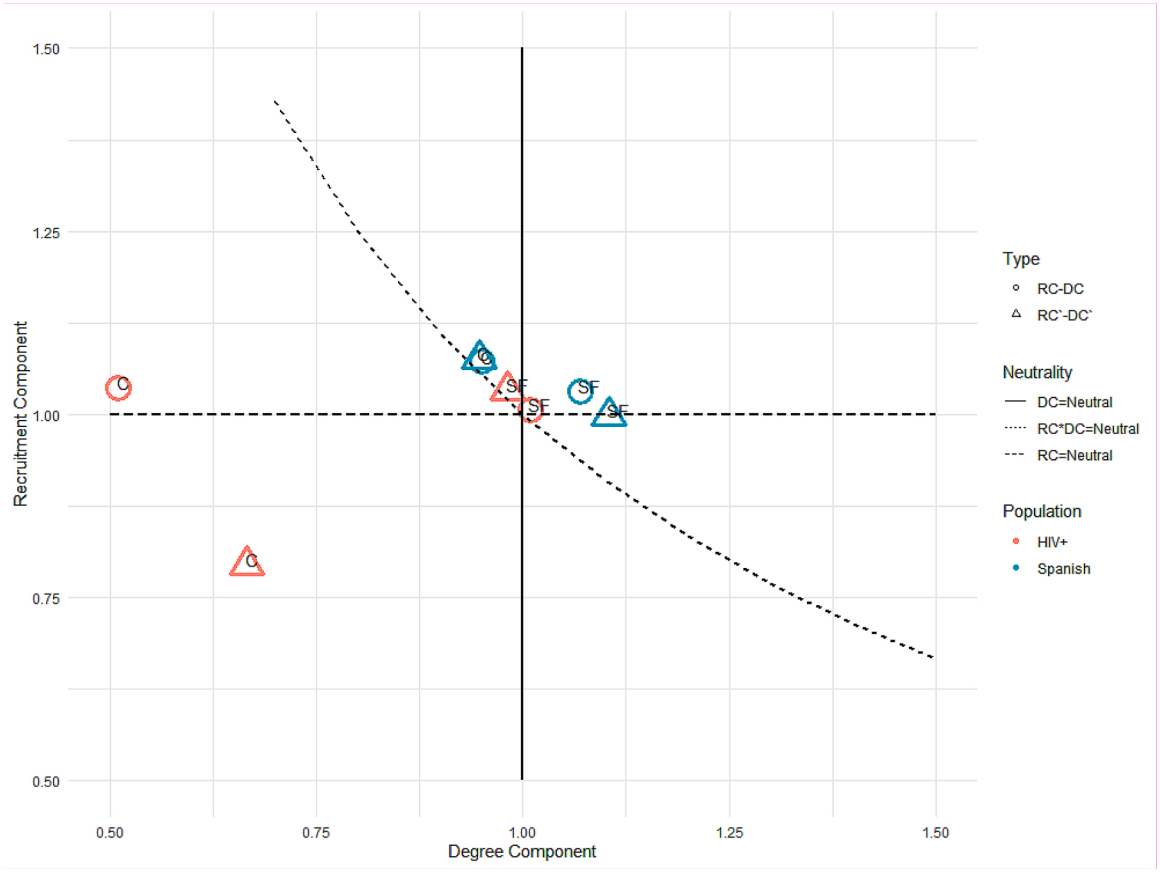

Fig. 6. Image showing each of the four samples' recruitment component and degree component within both RC-DC and RC`-DC` space.

The recruitment and degree components of the sample are estimates of how sample degree and bias from differential recruitment contribute to the overweighting or underweighting of groups in the sample. Counter-intuitively, the dual-component weight is more than simply the product of the recruitment component and the degree component. This is why two decompositions were necessary. One is created by separating the recruitment component from the remainder of the dual-component weight, and another is created by separating the degree component from the dual-component weight. For reasons we do not know, the latter decomposition seems to provide a clearer indication of when certain types of estimates will be systematically biased.

The fact that the dual-component weight cannot be isolated into a pure degree component and pure recruitment component suggest that the two main sources of bias in RDS interact with one another. The higher predictive power of the second RC—DC space suggests that even when differences in degree are accounted for directly, differences in degree by group inflate or deflate recruitment bias in ways that are not directly accounted for by ratios of cross recruitment. This is an important area for further study. A deeper analytical understanding of these two sources of bias and how they are intertwined will advance both methods of respondent-driven sampling and potentially our understanding of network structure more generally.

Several recent advances in RDS estimation have focused on how to account for individual network size and network structure when determining appropriate sampling weights for individuals. More sophisticated attempts to account for the influence of network size on inclusion in the sample have been seen in the successive sampling estimator (which was included in the estimators analyzed in this draft), and the use of network model-assisted inference (Gile and Handcock, 2015). Furthermore, asking respondents more questions about the structure of their ego-networks (Verdery et al., 2015) or what they know about the network size or similarity of their recruiter (Verdery et al., 2017), and keeping track of how many times specific individuals have been targeted for recruitment (Mouw and Verdery, 2012), can give us a better sense of

the micro-structural phenomena that bias RDS estimates towards certain individuals. These improvements will no doubt improve the overall efficacy of the method, but they focus primarily on how individuals are sampled as a function of the structure of their immediate network. Our formal analysis indicates both when group-level differences in recruitment activity and network segregation may interact to cause recruitment bias and how it can be identified from the sample, and our simulations show the conditions where accounting for recruitment bias may be worth the additional variance from a classical estimator.

A limitation of our analysis is that the empirically observed networks do not exhibit extreme degree components. The analysis of RC-DC space will hopefully be extended into network structures that have more skewed degree distributions. The Facebook100 data provided a wide range of potential variation in RC because there is a large degree of homophily by class year that is induced by the structure of universities (students often live with or take classes with those who share the same class year). However, the degree distributions of the largest components of these networks are fairly normal, as most students have a sizable number of social connections that are strong enough to merit acknowledgement on the chosen social networking site.

Stigmatized populations that are often the target of RDS studies may have a more extreme degree distribution. For example, a population of drug users may feature a small number of dealer "hubs" with many contacts and a larger number of people who know only a few other users. This would likely introduce more extreme degree components. Furthermore, the simulations in this paper focused on the estimation of proportional variables as opposed to continuous variables. A similar analysis comparing the performance of classical, equilibrium, naïve, and degree-based estimators across RC-DC space with regards to a continuous variable may also be informative.

While many avenues for research remain, it is encouraging that the information contained in an RDS sample itself (recruitment patterns and self-reported degree information) can inform estimator selection. At this point in the development of RDS, developing insights that inform estimator selection is just as important as the creation and refinement of existing estimators.

Funding

This research was made possible by a grant from the National Institutes of Health/National Institute for Nursing Research (1R21NR10961). We thank members of the Symposium on Respondent-Driven Sampling, Department of Mathematics, Stockholm University for helpful comments and advice.

Data availability statement

The data used in this paper is featured in peer-reviewed articles by Ramirez-Valles et al. (2008) and Traud et al. (2012), and can be made available upon reasonable request.

References

Bernhardt, A., Spiller, M.W., Polson, D., 2013. All work and no pay: violations of employment and labor laws in Chicago, Los Angeles and New York City. Soc. Forces 91 (3), 725–746.
Curtis, R., Terry, K., Dank, M., Dombrowski, K., Khan, B., Muslim, A., Labriola, M., Rempel, M., 2008. The commercial sexual exploitation of children in New York City. New York: Center for Court Innovation.
Gile, K.J., 2011. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. J. Am. Stat. Assoc. 106 (493), 135–146.
Gile, K.J., Handcock, M.S., 2010. Respondent-driven sampling: an assessment of current methodology. Sociol. Methodol. 40 (1), 285–327.
Gile, K.J., Handcock, M.S., 2015. Network model-assisted inference from respondent-driven sampling data. J. R. Stat. Soc. Ser. A Stat. Soc. 178 (3), 619–639.
Goel, S., Salganik, M.J., 2010. Assessing respondent-driven sampling. Proc. Natl. Acad. Sci. 107 (15), 6743–6747.
Heckathorn, D.D., 1997. Respondent-driven sampling: a new approach to the study of hidden populations. Soc. Probl. 44 (2), 174–199.
Heckathorn, D.D., 2002. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. Soc. Probl. 49 (1), 11–34.
Heckathorn, D.D., 2007. Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. Sociol. Methodol. 37 (1), 151–207.
Heckathorn, D.D., Jeffri, J., 2001. Finding the beat: using respondent-driven sampling to study jazz musicians. Poetics 28 (4), 307–329.
Jeffri, J., Heckathorn, D.D., Spiller, M.W., 2011. Painting your life: a study of aging visual artists in New York City. Poetics 39 (1), 19–43.
Johnston, L.G., Chen, Y.H., Silva-Santisteban, A., Raymond, H.F., 2013. An empirical examination of respondent driven sampling design effects among HIV risk groups from studies conducted around the world. AIDS Behav. 17 (6), 2202–2210.
McCreesh, N., Frost, S., Seeley, J., Katongole, J., Tarsh, M.N., Ndunguse, R., Jichi, F., Lunel, N.L., Maher, D., Johnston, L.G., Sonnenberg, P., 2012. Evaluation of respondent-driven sampling. Epidemiology 23 (1), 138–147, 2012.
McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. Annu. Rev. Sociol. 27 (1), 415–444.
Mouw, T., Verdery, A.M., 2012. Network sampling with memory: a proposal for more efficient sampling from social networks. Sociol. Methodol. 42 (1), 206–256.
Ramirez-Valles, J., Garcia, D., Campbell, R.T., Diaz, R.M., Heckathorn, D.D., 2008. HIV infection, sexual risk behavior, and substance use among Latino gay and bisexual men and transgender persons. Am. J. Public Health 98 (6), 1036–1042, 2008.
Salganik, M.J., Heckathorn, D.D., 2004. Sampling and estimation in hidden populations using respondent-driven sampling. Sociol. Methodol. 34 (1), 193–240.
Smylie, J., Firestone, F., Cochran, L., Prince, C., Maracle, S., Morley, M., Mayo, S., Spiller, T., McPherson, B., 2011. Our Health Counts. Urban Aboriginal Health Database Research Project. Community Report: First Nations Adults and Children. Toronto, ON.
Tomas, A., Gile, K.J., 2011. The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. Electron. J. Stat. 5, 899–934.
Traud, A.L., Mucha, P.J., Porter, M.A., 2012. Social structure of Facebook networks. Phys. A Stat. Mech. Its Appl. 391 (16), 4165–4180.
Verdery, A.M., Mouw, T., Bauldry, S., Mucha, P.J., 2015. Network structure and biased variance estimation in respondent driven sampling. PLoS One 10 (12), e0145296.
Verdery, A.M., Fisher, J.C., Siripong, N., Abdesselam, K., Bauldry, S., 2017. New survey questions and estimators for network clustering with respondent-driven sampling data. Sociol. Methodol. 47 (1), 274–306.
Volz, E., Heckathorn, D.D., 2008. Probability based estimation theory for respondent driven sampling. J. Off. Stat. 24 (1), 79–97.
Wejnert, C., 2010. Social network analysis with respondent-driven sampling data: a study of racial integration on campus. Soc. Networks 32 (2), 112–124.
Wejnert, C., Heckathorn, D.D., 2008. Sociol. Methods Res. 37 (1), 105–134.