American Journal of Epidemiology

Vol. 187, No. 1
DOI: 10.1093/aje/kwx208
Advance Access publication: June 9, 2017

Practice of Epidemiology

Identification of Homophily and Preferential Recruitment in Respondent-Driven Sampling

Forrest W. Crawford*, Peter M. Aronow, Li Zeng, and Jianghong Li

* Correspondence to Dr. Forrest W. Crawford, Laboratory of Epidemiology and Public Health, 60 College Street Room 207, PO Box 208034, New Haven, CT 06510 (e-mail: forrest.crawford@yale.edu).

Initially submitted October 27, 2016; accepted for publication March 9, 2017.

Respondent-driven sampling (RDS) is a link-tracing procedure used in epidemiologic research on hidden or hard-to-reach populations in which subjects recruit others via their social networks. Estimates from RDS studies may have poor statistical properties due to statistical dependence in sampled subjects' traits. Two distinct mechanisms account for dependence in an RDS study: homophily, the tendency for individuals to share social ties with others exhibiting similar characteristics, and preferential recruitment, in which recruiters do not recruit uniformly at random from their network alters. The different effects of network homophily and preferential recruitment in RDS studies have been a source of confusion and controversy in methodological and empirical research in epidemiology. In this work, we gave formal definitions of homophily and preferential recruitment and showed that neither is identified in typical RDS studies. We derived nonparametric identification regions for homophily and preferential recruitment and showed that these parameters were not identified unless the network took a degenerate form. The results indicated that claims of homophily or recruitment bias measured from empirical RDS studies may not be credible. We applied our identification results to a study involving both a network census and RDS on a population of injection drug users in Hartford, Connecticut (2012–2013).

hidden population; link-tracing; network sampling; nonparametric bounds; social network

Abbreviation: RDS, respondent-driven sampling.

Epidemiologic studies of stigmatized or criminalized populations—such as drug users, men who have sex with men, or sex workers—can be challenging because potential subjects may be unwilling to participate if they fear exposure or persecution. Respondent-driven sampling (RDS) is a widely used procedure for recruiting members of hidden or hard-to-reach populations (1, 2). Starting with a set of initial participants called "seeds," subjects are interviewed and given a small number of coupons they use to recruit other members of the study population. Participants recruit others by giving them a coupon bearing a unique code and information about how to participate in the study. Each subject receives a reward for being interviewed and another for every new subject they recruit. Most methodological research on RDS assumes the existence of a social network connecting members of the target population, where recruitments take place across edges in that network (3–5). Subjects in an RDS study typically do not provide identifying information about their network alters; instead, researchers measure respondents' social network degree (also sometimes called their "egocentric network size") in the target population. Because RDS only reveals recruitment links, many links in the network of respondents remain unobserved.

Because the RDS recruitment process is network-based, the traits of recruiter and recruitee may not be independent (6, 7). Two mechanisms account for this dependence. First, homophily is the tendency for people to exhibit social ties with others who share their traits. Second, recruiters in RDS choose new recruits from among their network neighbors. Preferential recruitment of subjects with certain traits, conditional on existing social ties, can make RDS recruitment chains appear more homogeneous, even in the absence of network homophily. While homophily is a property of the target population social network, preferential recruitment is a property of the RDS recruitment process, conditional on that network.

Epidemiologists and public health researchers focus on homophily and preferential recruitment in RDS studies for 2 primary reasons. First, empirical and simulation results have repeatedly shown that these distinct forms of dependence can bias estimates

of population-level quantities in different ways (6–13). Gile et al. (14, Table 1) stated that 2 assumptions "required" by the most popular estimator of the population mean (4) are "homophily sufficiently weak" and "random referral" by recruiters (see (15) for more general identification conditions for the population mean). Lunagomez and Airoldi (16) showed that RDS is a nonignorable sampling design because the distribution of observed data depends on unobserved parts of the network; when the network exhibits homophily with respect to the trait of interest, estimates may be biased. Second, many studies of infectious disease outcomes explicitly seek characteristics of an epidemiologic contact network connecting members of the target population (e.g., 17). The topology of the contact network may govern the dynamics of infectious disease spread (18, 19); because RDS is a network-based recruitment method, it may reveal features of this contact network. Stein et al. (20, 21) treated RDS recruitments as epidemiologic contacts to estimate homophily (also called assortative mixing) in the close-contact network relevant for transmission of pathogens. Stein et al. (20, p. 18) suggested that from RDS data "correlations between linked individuals can be used to improve parameterization of mathematical models used to design optimal control" for epidemic management.

While epidemiologists generally recognize the importance of preferential recruitment and homophily in determining the statistical properties of estimators of population-level quantities (e.g., human immunodeficiency virus prevalence) from RDS data, they do not agree on the definitions of these sources of dependence. White et al. (11) observed that the term homophily has "inconsistent usage in the RDS community. Sometimes it is used to refer to the tendency for sample recruitments to occur between participants in the same social category and sometimes to refer to the tendency for relationships in the target population to occur between participants in the same social category." Often homophily is defined as a characteristic of recruiters' behavior during the study. Ramirez-Valles et al. (22, p. 388) defined homophily as "a tendency toward in-group recruitment." Uusküla et al. (23, p. 307) defined homophily as "the extent to which recruiters are likely to recruit individuals similar to themselves." Léon et al. (24, p. 3) defined homophily as a property of recruitment, while making reference to the variety of other definitions: "[w]hen an individual tends to recruit persons who resemble him or her, especially with regard to the variable of interest, this is classically defined as homophily, even if several definitions coexist." Likewise, many researchers have described preferential recruitment behavior during the study as a network property. Abramovitz et al. (25, p. 751) wrote that "[d]ifferential recruitment patterns are usually the result of individuals' tendencies to associate with other individuals who are similar to them, also known as homophily." Rudolph et al. (26, p. 2326) wrote that "[d]ifferential recruitment based on the outcome of interest may occur when 1) the outcome clusters by network or 2) network members cluster in space and the outcome is spatially clustered."

Even when homophily and preferential recruitment are well-defined, there are serious problems with traditional approaches for measuring these features and interpretation of their role in statistical properties of estimators. Several authors have recognized the observational ambiguity between preferential recruitment and homophily: Tomas and Gile (7, p. 911) pointed out that "it is not always possible to distinguish from the sample if differential recruitment exists, because its effect on the resulting sampling chain is similar to that of homophily." Fisher and Merli (27) introduced the term "stickiness," the tendency of recruitments to become stuck within a group of subjects with similar traits. In response to the difficulty of simultaneously estimating these features from RDS recruitment data alone, several authors have proposed alternative methods of data collection to elicit additional features of the network proximal to recruited subjects (13, 28–30). This extra information, along with suitable assumptions, may reveal topological information (e.g., nonrecruitment links between recruited subjects) that permits tighter inferences about homophily and preferential recruitment.

Nevertheless, many authors have claimed to measure homophily in the target population social network, or preferential recruitment in the recruitment chain, directly from traditional RDS data; there is widespread confusion about the difference between the RDS recruitment path and the underlying network (2, 26, 31–36). Some researchers have employed regression to estimate factors associated with "productive recruiting" of network alters (37) or "network risk factors" for outcomes (17). Two software tools for analysis of RDS data produce estimates of homophily or preferential recruitment—RDSAT (38) and RDS Analyst (39)—but the estimators used by these programs are not documented.

In this work, we show that homophily and preferential recruitment are generally not identified in RDS studies. We adapt ideas from partial identification (40) and confounding in social network studies (41) to show that these sample functionals are not point identified unless the recruitment tree is identical to the underlying subgraph. Fortunately, researchers can still make credible claims about homophily and preferential recruitment in empirical RDS studies: We describe nonparametric graph-theoretical bounds under minimal assumptions about the underlying network and recruitment process. To illustrate the nonparametric bounds and their implications for an important epidemiologic risk population, we analyzed data from an RDS survey of people who inject drugs (PWID) in Hartford, Connecticut, in which the subgraph of respondents and their network alters is known with near certainty. We compare the true values of homophily and preferential recruitment obtained by using the full network with the identification intervals computed using RDS data alone. Web Appendix 1 (available at https://academic.oup.com/aje) briefly describes an algorithm for finding the bounds.

RDS DATA

We first state some basic assumptions about the target population and RDS procedure (2–6). We assume that the social network connecting members of the target population exists and is an undirected graph $G = (V, E)$ with no parallel edges or self-loops. Members of the target population are vertices in V, and edges in E represent social ties between individuals. A seed is a vertex that is not recruited by a network peer but is chosen by some other mechanism. A recruiter is a vertex known to the study that has at least 1 coupon. A susceptible vertex is not yet known to the study but has at least 1 neighbor in G that

is a recruiter. Every vertex $i \in V$ has a binary attribute Z_i that is observed when i is recruited. RDS recruitments happen across edges in G, and no subject can be recruited more than once.

We now define the network data collected by typical RDS studies (42). The directed recruitment graph is $G_R = (V_R, E_R)$, where V_R is the set of n sampled vertices (including seeds), and a directed edge from i to j indicates that i recruited j. The recruitment-induced subgraph is an undirected graph $G_S = (V_S, E_S)$, where $V_S = V_R$ and $\{i, j\} \in E_S$ if and only if $i \in V_S$, $j \in V_S$, and $\{i, j\} \in E$. Subjects also report their network degree d_i for $i \in V_R$, the number of edges incident to i in G. Let $\mathbf{d}_R$ and $\mathbf{t}_R$ be the $n \times 1$ vectors of recruited vertices' degrees and times of recruitment in the order they entered the study, and let M be the set of seeds. Label the vertices in V_R by the time-order of their recruitment: $i < j$ if $t_i < t_j$. Furthermore, we observe a vector $\mathbf{Z}_R = (Z_1, \ldots, Z_n)$ of subjects' binary trait values. Researchers conducting an RDS study observe only G_R, $\mathbf{d}_R, \mathbf{t}_R$, and $\mathbf{Z}_R$.

Finally, let U be the set of unsampled vertices connected by at least 1 edge to a sampled vertex in V_R at the end of the study. Let E_U be the set of edges connecting vertices in U to sampled vertices in V_R. The augmented recruitment-induced subgraph is an undirected graph $G_{SU} = (V_{SU}, E_{SU})$, where $V_{SU} = V_S \cup U$ and $E_{SU} = E_S \cup E_U$. Note that G_{SU} does not contain edges between vertices in U. For each vertex $u \in U$, let d_u be its degree in G_{SU}.

HOMOPHILY AND PREFERENTIAL RECRUITMENT

Consider an RDS sample of size n with recruitment graph $G_R = (V_R, E_R)$, degrees $\mathbf{d}_R$, and univariate binary traits $\mathbf{Z}_R$. Let $G_{SU} = (V_{SU}, E_{SU})$ be the augmented subgraph for this sample, with traits $\mathbf{Z}_{SU}$. The observed traits $\mathbf{Z}_R$ are a subset of $\mathbf{Z}_{SU}$. Let $\mathbf{A} = \{A_{ij}\}$ be the adjacency matrix of G_{SU}. A standard definition of network homophily is the correlation between edges and trait values, known as the assortativity coefficient (43–45).

Definition 1: Homophily is the correlation between trait values of vertices connected by an edge,

$$h(G_{SU}, \mathbf{Z}_{SU}) = \frac{\sum_{i,j \in V_{SU}} (A_{ij} - d_i d_j / 2|E_{SU}|) Z_i Z_j}{\sum_{i,j \in V_{SU}} (d_i \delta_{ij} - d_i d_j / 2|E_{SU}|) Z_i Z_j}, \quad (1)$$

where δ_{ij}=1 when $i = j$ and 0 otherwise, and for $i \in U$, $d_i = |\{j \in V_S: \{i, j\} \in E_{SU}\}|$.

Let $S_i(t)$ be the set of susceptible neighbors of $i \in V_R$ just before time t (the set-valued function $S_i(t)$ is left-continuous). Let r_j be the recruiter of the sampled vertex $j \in V_S$. Under proportional (unbiased) recruitment, the probability of a recruiter recruiting a susceptible neighbor with the same trait value is proportional to the number of its susceptible neighbors with the same trait, $|\{k \in S_{r_j}(t_j): Z_{r_j} = Z_k\}|$. In other words, recruitment is uniformly at random among susceptible neighbors.

Definition 2: Preferential recruitment is the average deviation from proportional recruitment, given knowledge of $G_{SU}, G_R, \mathbf{t}_R$, and $\mathbf{Z}_{SU}$,

$$p(G_{SU}, G_R, \mathbf{t}_R, \mathbf{Z}_{SU}) = \frac{1}{n - |M|} \sum_{j \notin M} \left(1\{Z_j = Z_{r_j}\} - \frac{|\{k \in S_{r_j}(t_j): Z_{r_j} = Z_k\}|}{|S_{r_j}(t_j)|}\right), \quad (2)$$

where M is the set of seeds.

To ease notation, we will often use h and p to refer to $h(G_{SU}, \mathbf{Z})$ and $p(G_{SU}, G_R, \mathbf{t}_R, \mathbf{Z}_{SU})$ respectively.

IDENTIFICATION

The quantities h and p are features of the network proximal to the sampled vertices, and the observed recruitment process. The observed recruitment subgraph G_R and reported degrees $\mathbf{d}_R$ place strong topological restrictions on the structure of G_{SU} and hence imply restrictions on h and p.

Definition 3: Compatibility: The pair $(G_{SU}, \mathbf{Z}_{SU})$ is compatible with the observed data G_R, $\mathbf{d}_R$, *and* $\mathbf{Z}_R$ if 1) the recruitment graph is preserved ($G_R \subseteq G_{SU}$); 2) the set of recruited subjects' trait values is preserved ($\mathbf{Z}_R \subseteq \mathbf{Z}_{SU}$); 3) all unsampled vertices are connected to a recruited vertex (every $u \in V_{SU}$ with $u \notin V_S$ has an edge $\{v, u\}$ such that $v \in V_S$); and 4) total degree is preserved (for every $i \in V_R, d_i = \sum_{j \in V_{SU}} 1\{\{i, j\} \in E_{SU}\}$).

Let $\mathrm{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ be the set of pairs $(G_{SU}, \mathbf{Z}_{SU})$ compatible with the observed data in the sense of definition 3 (this is a finite set).

First, we examine whether G_S and G_{SU} are revealed by the observed data in RDS. Let d_i^r be the degree of subject i in the recruitment subgraph G_R.

Result 1: Suppose there exist $i, j \in V_R$ with $i < j, (i, j) \notin E_R, d_i^r < d_i$, and $d_j^r < d_j$. Then neither G_S nor G_{SU} are identified.

Proof is given in the Appendix. This result establishes the conditions under which statements about the population graph proximal to the sample can be made precise. Next, we define the information about h and p that is revealed by the observed data.

Definition 4: Identification region: The identification regions for h and p are given by the smallest intervals that contain $h(G_{SU}, \mathbf{Z}_{SU})$ and $p(G_{SU}, G_R, \mathbf{t}_R, \mathbf{Z}_{SU})$ for every $(G_{SU}, \mathbf{Z}_{SU}) \in \mathrm{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$.

When the identification region for h or p contains only a single point, that parameter is point identified. Figure 1 shows an example of the identification regions.

Result 2: Suppose there exist 2 vertices $i \in V_R$ and $j \in V_R$ such that $d_i > d_i^r, d_j > d_j^r$, and $Z_i \neq Z_j$. Then h is not point identified.

Result 3: Suppose there exists a vertex $i \in V_R$ that recruited at least 1 other vertex $j \in V_R, j \neq i$, and $d_i > d_i^r$. Then p is not point identified.

Proof is given in the Appendix. In practice, point identification of both homophily and preferential recruitment can be achieved only if the recruitment graph G_R is nearly identical to the augmented recruitment-induced subgraph G_{SU}. Because RDS recruitment is without replacement, the recruitment subgraph G_R is acyclic, so $G_R = G_{SU}$ means that the population

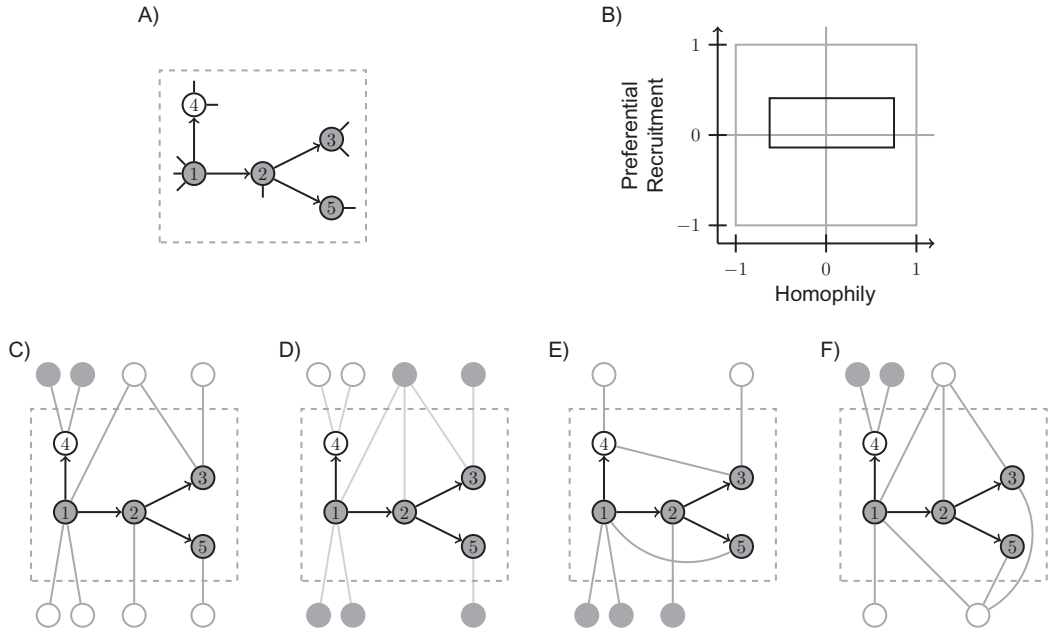

Figure 1. Example illustration of extrema for h and p with corresponding augmented subgraphs G_{SU} and $\mathbf{Z}_{SU}$. Vertices are shaded according to their binary type. A) G_R is shown with arrows indicating recruitments, and pendant edges implied by vertex degrees; B) the joint homophily/preferential recruitment space $[-1,1] \times [-1,1]$ (gray square), with the identification rectangle (shown in black) containing all values of (h, p) compatible with the observed data; C) a subgraph that achieves the minimum value of homophily, $h = -0.625$; D) maximum value of homophily, $h = 0.752$; E) minimum value of preferential recruitment, $p = -0.138$; and F) maximum value of preferential recruitment, $p = 0.408$. Some extremal subgraphs are not unique.

network proximal to recruited vertices is a tree, a situation that seems unlikely to occur in a real-world social network. Furthermore, results 1, 2, and 3 apply directly to the case where all vertices in the population have been sampled, and we have $V_R = V$: If pendant edges remain, then h and p may not be point identified.

APPLICATION: INJECTION DRUG USERS IN HARTFORD, CONNECTICUT

We now apply the ideas developed above to an RDS data set in which the augmented subgraph G_{SU} of an RDS sample G_R is known with near certainty. In the RDS-net study, researchers conducted an RDS survey of $n = |V_R| = 527$ injection drug users from $|M| = 6$ seeds in Hartford, Connecticut. Researchers simultaneously performed a census of the augmented recruitment-induced subgraph, consisting of $|V_{SU}| = 2{,}626$ unique injection drug users (46). Subjects were given $25 for being interviewed and $10 for recruiting another eligible subject (up to a maximum of 3). Subjects were required to be at least 18 years old, reside in the Hartford area, and report injecting illicit drugs in the last 30 days. The study was approved by the institutional review board for the Institute for Community Research, and informed consent was obtained from all subjects.

This study differs from typical RDS surveys because in addition to reporting their network degree, respondents also enumerated (nominated) their network alters—other people eligible for the study whom they knew by name and could possibly recruit. Unsampled injection drug users nominated by more than 1 participant were matched using identifying characteristics, including name (including aliases), photograph, multiple addresses, phone numbers, locations frequented, and social network links (47, 48). The resulting "nomination" network is the augmented subgraph G_{SU}. Figure 2 shows the nomination network G_{SU} with the recruitment graph G_R overlaid.

The RDS-net study included $|V_{SU}| = 2{,}626$ people, of whom $n = |V_R| = 527$ were either seeds or recruited subjects, and 2,099 were nominated but never-recruited subjects. There are $|E_{SU}| = 3{,}307$ edges in G_{SU}, of which $|E_S| = 1{,}180$ link recruited subjects to recruited subjects, and 2,127 link recruited subjects to unrecruited subjects. Demographic and trait data relevant to drug use were collected about each recruited subject. Each recruited subject also reported the traits of their nominees. Nominees who were never personally interviewed were assigned trait information as follows: If their nominating alters agreed on their trait value, that value was assigned to them. If there was disagreement, the modal value was assigned. When trait information for a recruited subject or an unrecruited alter was absent or contradictory, it was treated as missing. We selected 3 traits with the least missing data for analysis: sex, "crack" cocaine use, and homelessness. This information was fully observed for every recruited subject, but some values were missing for nominated but unrecruited subjects.

First we computed h and p using the full data (G_{SU}, G_R, $\mathbf{t}_R$, $\mathbf{Z}_{SU}$), where we omitted vertices in V_{SU} whose trait was missing (all recruited subjects' trait values were fully observed), any edges incident to these vertices, and the corresponding elements of $\mathbf{Z}_{SU}$. Second, we computed bounds for h and p using only

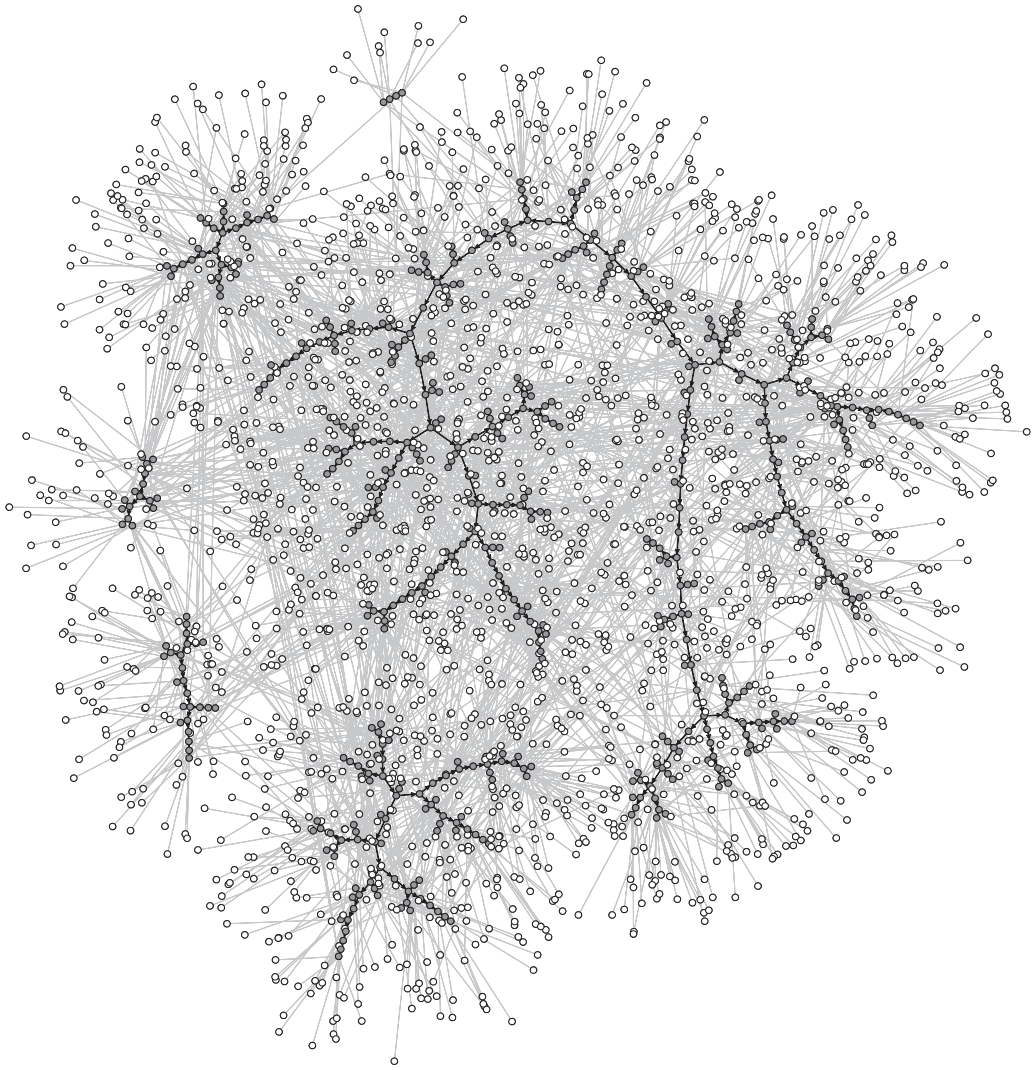

Figure 2. The nomination network and recruitment graph for injection drug users in the RDS-net study conducted in Hartford, Connecticut, 2012–2013. Recruited subjects are shown in dark gray, and recruitment edges are shown as a directed edge (arrow) from the recruiter to the recruitee. Nonrecruitment edges (linking recruited subjects to unrecruited subjects, or recruited to recruited subjects) are shown as light gray lines.

the data observed in the RDS portion of the study, (G_R, $\mathbf{d}_R$, $\mathbf{t}_R$, $\mathbf{Z}_R$) , using a procedure described in Web Appendixes 1 and 2. Table 1 shows values of h and p calculated using the full data. The intervals I_h and I_p give the identification bounds obtained using the observed RDS data alone. Point values (where unrecruited vertices with missing traits were excluded) always lay within the identification intervals. Values of homophily with respect to sex and crack use were positive, and they were negative for homelessness, while p was smaller in magnitude and positive for sex and homelessness but negative for crack use. The identification regions for each trait covered $h = 0$ and $p = 0$.

DISCUSSION

Statisticians, epidemiologists, and sociologists have conducted a wide variety of rigorous empirical and simulation studies devoted to understanding the influence of homophily and preferential recruitment on population-level estimates from RDS studies (6–10, 13, 49). Several researchers have correctly pointed out the definitional (11, 24) and observational ambiguity between these quantities (7, 27).

But to our knowledge, none have explained formally why these quantities are difficult to measure: In most empirical RDS studies, neither homophily nor preferential recruitment is identified from the observed data alone because the underlying network remains only partially observed. There is reason to be skeptical of claims that the social network of a particular target population surveyed by RDS exhibits homophily (26, 32–35) or that a particular RDS study suffers from selection bias due to preferential recruitment with respect to particular traits (36). The identification results described in this paper establish a formal basis for the skepticism expressed by Tomas and Gile (7) as well as others about estimates of homophily and preferential

Table 1. Homophily and Preferential Recruitment in the RDS-Net Study for the Traits of Sex, Crack Cocaine Use, and Homelessness Among Injection Drug Users in Hartford, Connecticut, During 2012–2013.

Trait	Homophily		Preferential Recruitment	
	h	I_h	p	I_p
Sex	0.2638	−0.7027, 0.8292	0.00397	−0.190, 0.534
Crack cocaine use	0.2711	−0.6918, 0.8231	−0.00051	−0.272, 0.453
Homelessness	−0.2154	−0.7179, 0.8242	0.04527	−0.221, 0.504

Abbreviations: h, homophily; p, preferential recruitment; RDS, respondent-driven sampling.
[a] Point values are given first, followed by identification intervals I_h and I_p, for h and p, respectively, obtained by using only the information observed in the recruitment portion of the study.

recruitment from RDS. Fortunately, researchers can still make credible claims about homophily and preferential recruitment without invoking additional assumptions, using the identification bounds introduced here.

The identification results depend on 3 assumptions: The network exists, subjects are recruited across its edges, and nobody can be recruited more than once. When these assumptions are met, the structure of data from RDS studies allows computation of credible bounds for h and p. However, the observed data impose strict limits on the precision of these estimates, and in practice bounds are often wide. Stronger assumptions about the topology of the network and dynamics of the recruitment process may yield narrower bounds, or point identification, at the cost of decreased credibility (40). Our identification results pertain to definitions of homophily and preferential recruitment for a univariate binary trait. Generalizations to multiple traits are possible, and similar identification results may continue to hold.

Alternative survey designs may reveal much more about homophily and recruitment behavior than RDS. Many researchers have traced social links of participants, sometimes alongside a traditional RDS study, to reveal the induced subgraph of respondents (50–53). Surveying respondents' egocentric networks can also reveal enough of the network to reliably estimate homophily (28, 30). Better estimates of preferential recruitment in RDS studies have been obtained by administering a follow-up questionnaire to subjects about their recruitment behavior (13, 14, 29, 54). Given auxiliary information and suitable assumptions, these data-collection procedures may yield tighter identification or even point identification. These approaches may also permit researchers to engage with generalizations of the definitions that we have considered here (e.g., proportional recruitment conditional on shared covariates). We believe that further methodological and empirical study of these approaches is of scientific interest.

Although it may be disappointing that homophily and preferential recruitment are not point identified in traditional RDS studies, researchers can still draw credible inferences about these parameters. The identification regions for sex, crack use, and homelessness in the RDS-net study were considerably smaller than the outcome space $[-1,1]$. Under some circumstances, it may be possible to deduce that homophily or preferential recruitment is strictly positive or negative in the augmented subgraph, even without exact knowledge of that subgraph. Informally, the more G_{SU} resembles G_R, the narrower the identification regions for (h, p) will be. Researchers who find the identification regions too wide for their scientific purposes may wish to consider alternative study designs that are more likely to reveal the quantities of interest.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut (Forrest W. Crawford, Peter M. Aronow, Li Zeng); Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut (Forrest W. Crawford); Department of Political Science, Yale University, New Haven, Connecticut (Peter M. Aronow); Yale School of Management, New Haven, Connecticut (Forrest W. Crawford, Peter M. Aronow); and Institute for Community Research, Hartford, Connecticut (Jianghong Li).

This research was supported by the Yale Center for Clinical Investigation, the Yale Center for Interdisciplinary Research on AIDS, Yale University Biomedical High Performance Computing Center, the Yale World Scholars Program, the China Scholarship Council, and the National Institutes of Health (grants DP2OD022614, KL2TR000140, P30MH062294, R01DA031594-03, RR19895, and RR029676-01).

We thank Dr. Gayatri Moorthi, Dr. Heather Mosher, Greg Palmer, Eduardo Robles, Mark Romano, Jason Weiss, and the staff at the Institute for Community Research for their work collecting and preparing the RDS-net data. We also thank Dr. Jacob Fisher, Dr. Krista Gile, Dr. Mark Handcock, Dr. Robert Heimer, Dr. Edward H. Kaplan, Lilla Orr, Jiacheng Wu, and Dr. Alexei Zelenev for helpful conversations and comments on the manuscript.

Conflict of interest: none declared.

REFERENCES

1. Broadhead RS, Heckathorn DD, Weakliem DL, et al. Harnessing peer networks as an instrument for AIDS prevention: results from a peer-driven intervention. *Public Health Rep*. 1998;113(suppl 1):42–57.

2. Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. *Soc Probl*. 1997;44(2):174–199.
3. Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol Methodol*. 2004;34(1):193–240.
4. Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. *J Off Stat*. 2008;24:79–97.
5. Gile KJ. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J Am Stat Assoc*. 2011;106(493):135–146.
6. Gile KJ, Handcock MS. Respondent-driven sampling: an assessment of current methodology. *Sociol Methodol*. 2010;40(1):285–327.
7. Tomas A, Gile KJ. The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electron J Stat*. 2011;5:899–934.
8. Liu H, Li J, Ha T, et al. Assessment of random recruitment assumption in respondent-driven sampling in egocentric network data. *Soc Netw*. 2012;1(2):13–21.
9. Rudolph AE, Fuller CM, Latkin C. The importance of measuring and accounting for potential biases in respondent-driven samples. *AIDS Behav*. 2013;17(6):2244–2252.
10. Rocha LEC, Thorson AE, Lambiotte R, et al. Respondent-driven sampling bias induced by clustering and community structure in social networks. *arXiv Preprints*. https://arxiv.org/abs/1503.05826v1. Published March 19, 2015. Accessed October 5, 2017.
11. White RG, Hakim AJ, Salganik MJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology for respondent-driven sampling studies: "STROBE-RDS" statement. *J Clin Epidemiol*. 2015;68(12):1463–1471.
12. Verdery AM, Mouw T, Bauldry S, et al. Network structure and biased variance estimation in respondent driven sampling. *PLoS One*. 2015;10(12):e0145296.
13. Verdery AM, Merli MG, Moody J, et al. Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in China. *Epidemiology*. 2015;26(5):661–665.
14. Gile KJ, Johnston LG, Salganik MJ. Diagnostics for respondent-driven sampling. *J R Stat Soc Ser A Stat Soc*. 2015;178(1):241–269.
15. Aronow PM, Crawford FW. Nonparametric identification for respondent-driven sampling. *Stat Probab Lett*. 2015;106:100–102.
16. Lunagomez S, Airoldi E. Bayesian inference from non-ignorable network sampling designs. *arXiv Preprints*. https://arxiv.org/abs/1401.4718. Published January 19, 2014. Accessed October 5, 2017.
17. Tsang MA, Schneider JA, Sypsa V, et al. Network characteristics of people who inject drugs within a new HIV epidemic following austerity in Athens, Greece. *J Acquir Immune Defic Syndr*. 2015;69(4):499–508.
18. Salathé M, Jones JH. Dynamics and control of diseases in networks with community structure. *PLoS Comput Biol*. 2010;6(4):e1000736.
19. Volz EM, Miller JC, Galvani A, et al. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. *PLoS Comput Biol*. 2011;7(6):e1002042.
20. Stein ML, van Steenbergen JE, Buskens V, et al. Comparison of contact patterns relevant for transmission of respiratory pathogens in Thailand and the Netherlands using respondent-driven sampling. *PLoS One*. 2014;9(11):e113711.
21. Stein ML, van Steenbergen JE, Chanyasanha C, et al. Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand. *PLoS One*. 2014;9(1):e85256.
22. Ramirez-Valles J, Heckathorn DD, Vázquez R, et al. From networks to populations: the development and application of respondent-driven sampling among IDUs and Latino gay men. *AIDS Behav*. 2005;9(4):387–402.
23. Uusküla A, Johnston LG, Raag M, et al. Evaluating recruitment among female sex workers and injecting drug users at risk for HIV using respondent-driven sampling in Estonia. *J Urban Health*. 2010;87(2):304–317.
24. Léon L, Des Jarlais D, Jauffret-Roustide M, et al. Update on respondent-driven sampling: theory and practical considerations for studies of persons who inject drugs. *Methodolog Innov*. 2016;9:1–9.
25. Abramovitz D, Volz EM, Strathdee SA, et al. Using respondent-driven sampling in a hidden population at risk of HIV infection: who do HIV-positive recruiters recruit? *Sex Transm Dis*. 2009;36(12):750–756.
26. Rudolph AE, Gaines TL, Lozada R, et al. Evaluating outcome-correlated recruitment and geographic recruitment bias in a respondent-driven sample of people who inject drugs in Tijuana, Mexico. *AIDS Behav*. 2014;18(12):2325–2337.
27. Fisher JC, Merli MG. Stickiness of respondent-driven sampling recruitment chains. *Netw Sci (Camb Univ Press)*. 2014;2(2):298–301.
28. Iguchi MY, Ober AJ, Berry SH, et al. Simultaneous recruitment of drug users and men who have sex with men in the United States and Russia using respondent-driven sampling: sampling methods and implications. *J Urban Health*. 2009;86(suppl 1):5–31.
29. Yamanis TJ, Merli MG, Neely WW, et al. An empirical analysis of the impact of recruitment patterns on RDS estimates among a socially ordered population of female sex workers in China. *Sociol Methods Res*. 2013;42(3):392–425.
30. Merli MG, Moody J, Smith J, et al. Challenges to recruiting population representative samples of female sex workers in China using respondent driven sampling. *Soc Sci Med*. 2015;125:79–93.
31. Wejnert C. Social network analysis with respondent-driven sampling data: a study of racial integration on campus. *Soc Networks*. 2010;32(2):112–124.
32. Gwadz MV, Leonard NR, Cleland CM, et al. The effect of peer-driven intervention on rates of screening for AIDS clinical trials among African Americans and Hispanics. *Am J Public Health*. 2011;101(6):1096–1102.
33. Simpson B, Brashears M, Gladstone E, et al. Birds of different feathers cooperate together: no evidence for altruism homophily in networks. *Sociol Sci*. 2014;1:542–564.
34. Rudolph AE, Crawford ND, Latkin C, et al. Subpopulations of illicit drug users reached by targeted street outreach and respondent-driven sampling strategies: implications for research and public health practice. *Ann Epidemiol*. 2011;21(4):280–289.
35. Wejnert C, Pham H, Krishna N, et al. Estimating design effect and calculating sample size for respondent-driven sampling studies of injection drug users in the United States. *AIDS Behav*. 2012;16(4):797–806.
36. Young AM, Rudolph AE, Quillen D, et al. Spatial, temporal and relational patterns in respondent-driven sampling: evidence from a social network study of rural drug users. *J Epidemiol Community Health*. 2014;68(8):792–798.
37. Forrest JI, Lachowsky NJ, Lal A, et al. Factors associated with productive recruiting in a respondent-driven sample of men

who have sex with men in Vancouver, Canada. *J Urban Health*. 2016;93(2):379–387.
38. Volz E, Wejnert C, Degani I, et al. *Respondent-Driven Sampling Analysis Tool (RDSAT) Version 7.1*. Ithaca, NY: Cornell University; 2012.
39. Handcock MS, Fellows IE, Gile KJ. RDS: Respondent-driven sampling (R package, version 0.5); 2013. https://CRAN.R-project.org/package=RDS. Published November 28, 2013. Accessed October 5, 2017.
40. Manski Charles F. *Partial Identification of Probability Distributions*. New York, NY: Springer; 2003.
41. Shalizi CR, Thomas AC. Homophily and contagion are generically confounded in observational social network studies. *Sociol Methods Res*. 2011;40(2):211–239.
42. Crawford FW. The graphical structure of respondent-driven sampling. *Sociol Methodol*. 2016;46(1):187–211.
43. Newman ME. Assortative mixing in networks. *Phys Rev Lett*. 2002;89(20):208701.
44. Newman ME. Mixing patterns in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2003;67(2 Pt 2):026126.
45. Newman ME. *Networks: An Introduction*. New York, NY: Oxford University Press; 2010.
46. Mosher HI, Moorthi G, Li J, et al. A qualitative analysis of peer recruitment pressures in respondent driven sampling: are risks above the ethical limit? *Int J Drug Policy*. 2015;26(9): 832–842.
47. Li J, Weeks MR, Borgatti SP, et al. A social network approach to demonstrate the diffusion and change process of intervention from peer health advocates to the drug using community. *Subst Use Misuse*. 2012;47(5):474–490.
48. Weeks MR, Clair S, Borgatti SP, et al. Social networks of drug users in high-risk sites: finding the connections. *AIDS Behav*. 2002;6(2):193–206.
49. Lu X. Linked ego networks: improving estimate reliability and validity with respondent-driven sampling. *Soc Networks*. 2013; 35(4):669–685.
50. McCreesh N, Frost SD, Seeley J, et al. Evaluation of respondent-driven sampling. *Epidemiology*. 2012;23(1):138–147.
51. Mouw T, Verdery AM. Network sampling with memory: a proposal for more efficient sampling from social networks. *Sociol Methodol*. 2012;42(1):206–256.
52. Mouw T, Chavez S, Edelblute H, et al. Binational social networks and assimilation: a test of the importance of transnationalism. *Soc Probl*. 2014;61(3):329–359.
53. Merli MG, Verdery A, Mouw T, et al. Sampling migrants from their social networks: the demography and social organization of Chinese migrants in Dar es Salaam, Tanzania. *Migr Stud*. 2016;4(2):182–214.
54. de Mello M, Araujo PA, Chinaglia M, et al. *Assessment of Risk Factors for HIV Infection Among Men Who Have Sex With Men in the Metropolitan Area of Campinas City, Brazil, Using Respondent-Driven Sampling*. Campinas, Brazil: Population Council, Horizons; 2008.

APPENDIX

Proof of Result 1: Call a recruitment-induced subgraph $G_S = (V_S, E_S)$ compatible with the observed data if $V_S = V_R$, $\{i, j\} \in E_R$ implies $\{i, j\} \in E_S$, and $\sum_{j \neq i} 1\{\{i, j\} \in E_S\} \leq d_i$ for each $i \in V_R$. Call an augmented recruitment-induced subgraph $G_{SU} = (V_{SU}, E_{SU})$ compatible with the observed data if conditions 1, 3, and 4 of definition 3 hold. Suppose $i \in V_R$ has $d_i^r < d_i$ and $j \in V_R$ has $d_j^r < d_j$, with $i < j$ and $(i, j) \notin E_R$. Let $G_{SU}^1 = (V_{SU}^1, E_{SU}^1)$ be any compatible subgraph with $\{i, u\} \in E_{SU}^1, \{j, u\} \in E_{SU}^1$, where $u \notin V_R$ is an unsampled vertex. Let $G_{SU}^2 = (V_{SU}^2, E_{SU}^2)$ be identical to G_{SU}^1 except that $\{i, j\} \in E_{SU}^2$, so neither i nor j is connected to u. If in the resulting subgraph u has no neighbors in V_R (i.e., there does not exist $k \in V_R$ such that $\{k, u\} \in E_{SU}^2$), then remove u from V_{SU}^2. Because there exist at least 2 augmented recruitment-induced subgraphs compatible with the observed data, G_{SU} is not identified. Let G_S^1 be the recruitment-induced subgraph obtained by removing any unsampled vertices (and edges connected to them) from G_{SU}^1. Define G_S^2 in the same way. Clearly both G_S^1 and G_S^2 are compatible with the observed data under the conditions of definition 3, but $\{i, j\} \notin G_S^1$ and $\{i, j\} \in G_S^2$. Because there exist at least 2 recruitment-induced subgraphs compatible with the observed data, G_S is not identified.

Proof of Result 2: Suppose the observed RDS data are G_R, $\mathbf{Z}_R$, $\mathbf{t}_R$, $\mathbf{d}_R$, and there exist distinct $i \in V_R$ and $j \in V_R$ with $d_i^r < d_i$, $d_j^r < d_j$, and $Z_i \neq Z_j$. Without loss of generality, suppose $Z_i = 0$ and $Z_j = 1$. We will exhibit $(G_{SU}^1, \mathbf{Z}_{SU}^1) \in \mathrm{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ and $(G_{SU}^2, \mathbf{Z}_{SU}^2) \in \mathrm{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$ such that $h(G_{SU}^1, G_R, \mathbf{Z}_{SU}^1) \neq h(G_{SU}^2, G_R, \mathbf{Z}_{SU}^2)$. Let $(G_{SU}^1, \mathbf{Z}_{SU}^1)$ be any compatible subgraph and trait set with the property that $\{i, u_1\} \in E_{SU}^1$ and $\{j, u_2\} \in E_{SU}^1$, where u_1 and u_2 are unsampled vertices with $Z_{u_1}^1 = 0$ and $Z_{u_2}^1 = 1$. Let $(G_{SU}^2, \mathbf{Z}_{SU}^2)$ be identical to $(G_{SU}^1, \mathbf{Z}_{SU}^1)$ except that the edges connecting i and j to u_1 and u_2 are swapped: $\{i, u_1\} \notin E_{SU}^2$,$\{j, u_2\} \notin E_{SU}^2$, and $\{i, u_2\} \in E_{SU}^2$ and $\{j, u_1\} \in E_{SU}^2$. Clearly $(G_{SU}^2, \mathbf{Z}_{SU}^2) \in \mathrm{C}(G_R, \mathbf{d}_R, \mathbf{Z}_R)$. Let $h_1 = h(G_{SU}^1, G_R, \mathbf{Z}_{SU}^1)$ and $h_2 = h(G_{SU}^2, G_R, \mathbf{Z}_{SU}^2)$ be the calculated values of homophily. The difference is

$$h_1 - h_2 = \frac{2\left(1 - \frac{d_j d_{u_2}}{2|E_{SU}|}\right) - 2\left(0 - \frac{d_j d_{u_2}}{2|E_{SU}|}\right)}{\sum_{k,l \in V_{SU}} (d_k \delta_{kl} - d_k d_l / 2|E_{SU}|) Z_k Z_l} = \frac{2}{\sum_{k,l \in V_{SU}} (d_k \delta_{kl} - d_k d_l / 2|E_{SU}|) Z_k Z_l} \neq 0. \tag{A1}$$

Therefore homophily is not point identified.

Proof of Result 3: Again suppose there exists $i \in V_R$ such that $d_i^r < d_i$ and i recruited $j \in V_R$, $j \neq i$. Without loss of generality, suppose $Z_i = 1$. Let $(G_{SU}^1, \mathbf{Z}_{SU}^1)$ be any compatible subgraph and trait set such that 1 edge connects i to an unsampled vertex u, where u has no other neighbors in V_R, and $Z_u = 1$. Let $K^1 = |\{k \in S_i(t_j) \colon Z_i = Z_k\}|$ under the trait values given by $\mathbf{Z}_{SU}^1$. Let $(G_{SU}^2, \mathbf{Z}_{SU}^2)$ be identical to $(G_{SU}^1, \mathbf{Z}_{SU}^1)$ except that $Z_u = 0$. Let $p_1 = p(G_{SU}^1, G_R, \mathbf{t}_R, \mathbf{Z}_{SU}^1)$ and $p_2 = p(G_{SU}^2, G_R, \mathbf{t}_R, \mathbf{Z}_{SU}^2)$. The difference is

$$p_1 - p_2 = \frac{1}{n - |M|}\left[\left(1\{Z_i = Z_j\} - \frac{K^1}{|S_i(t_j)|}\right) - \left(1\{Z_i = Z_j\} - \frac{K^1 - 1}{|S_i(t_j)|}\right)\right] = \frac{-1}{(n - |M|)|S_i(t_j)|} \neq 0. \tag{A2}$$

Therefore preferential recruitment is not point identified.