Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations

DOUGLAS D. HECKATHORN, *Cornell University*

Researchers studying hidden populations—including injection drug users, men who have sex with men, and the homeless—find that standard probability sampling methods are either inapplicable or prohibitively costly because their subjects lack a sampling frame, have privacy concerns, and constitute a small part of the general population. Therefore, researchers generally employ non-probability methods, including location sampling methods such as targeted sampling, and chain-referral methods such as snowball and respondent-driven sampling. Though nonprobability methods succeed in accessing the hidden populations, they have been insufficient for statistical inference. This paper extends the respondent-driven sampling method to show that when biases associated with chain-referral methods are analyzed in sufficient detail, a statistical theory of the sampling process can be constructed, based on which the sampling process can be redesigned to permit the derivation of indicators that are not biased and have known levels of precision. The results are based on a study of 190 injection drug users in a small Connecticut city.

Non-probability sampling methods were once considered appropriate only for pilot studies. This has changed because of the AIDS epidemic (Laumann, et al. 1989) and decreases in the accuracy of the U.S. census (Brown, et al. 1999). Efforts to address both problems have focused attention on sampling *hidden populations*, such as injection drug users, men who have sex with men, and the homeless. These populations lack a sampling frame, and sampling is further complicated by privacy concerns based on the stigma associated with membership in the population. Traditional methods for sampling these groups are often inapplicable (MacKellar, et al. 1996). For example, one cannot sample injection drug users through household surveys because injectors sometimes hide their habit from those with whom they live, including parents, roommates, and sexual partners. Similarly, the homeless cannot be reached through household surveys or random digit dialing, nor can these methods reach persons with unstable living arrangements, as occurs when several families live in an apartment, but only one name appears on the lease (Sudman and Kalton 1986). Studies of drug treatment programs, prisons, and homeless shelters provide access to some hidden populations, but when the institutions do the sampling, drawing a representative sample is not a priority.

Two methods currently dominate studies of hidden populations. *Location sampling*, the best-known form of which is *targeted sampling* (Watters and Biernacki 1989), is suitable when the target population is geographically concentrated. It generally involves two basic steps. Researchers ethnographically map the target population and then conduct interviews at the

This research was made possible by grants from the Centers for Disease Control and Prevention (U62/CCU114816-01) and the National Institute on Drug Abuse (RO1 DA08014). I thank Salaam Semaan, James Carey, Matthew Salganik, James Hughes, Robert Broadhead, Jennifer Tiffany Yael van Hulst, Eugene Johnsen, Denise Anthony, and several anonymous reviewers for their helpful comments and advice. This paper was presented at the School of Public Health, University of Illinois at Chicago, December 9, 1999. Direct correspondence to: Douglas D. Heckathorn, Dept. of Sociology, 344 Uris Hall, Cornell University, Ithaca, NY 14853-7601. E-mail: douglas.heckathorn@cornell.edu

SOCIAL PROBLEMS, Vol. 49, No. 1, pages 11–34. ISSN: 0037-7791; online ISSN: 1533-8533

sites identified by the ethnographic mapping. Location sampling varies based on the characteristics of the locations being sampled. If the locations where the population can be found are well defined and public, as in studies of nonhidden populations such as patrons of a shopping mall, respondents can be drawn randomly. However, this is generally not possible when sampling a hidden population, because sampling is limited in several ways (Watters and Biernacki 1989). First, by time of day. For example, in studies of injection drug users (IDUs), safety concerns generally limit interviews to daylight hours even though this is not when many drug scenes are most active. Second, by researchers' recruitment strategies. For example, researchers have explored techniques for increasing response rates for brief screening interviews (MacKellar, et al. 1996). And third—but most significantly—by location. Private settings, such as individuals' homes, and small, geographically dispersed settings are often excluded because accessing them would be prohibitively costly or impossible. Therefore, members of the population who do not frequent large public settings tend to be excluded.

The second method for sampling hidden populations, *chain-referral sampling* (Erickson 1979), is suitable when members of the target population know one another and are densely interconnected. It is not suitable when the population members do not know one another as members of the population. The method is further limited by the form of referral. When, as in the study reported here, interviews take place at a single location, the sample is limited to those in relatively close proximity to that site. However, were telephone interviews used, samples of less geographically concentrated populations could be drawn.

Sampling begins with a set of initial subjects who serve as seeds for an expanding chain of referrals, with subjects from each wave referring subjects of the subsequent wave. The best-known form is snowball sampling (Goodman 1961): The seeds, drawn randomly from the population, provide researchers with the names and contact information of other potential subjects; the researchers then select a fixed number of names from each list; and this process continues until the desired number of waves is reached. Of course, drawing a random initial sampling is not feasible when sampling a hidden population (Spreen 1992), so Frank and Snijders (1994) recommend beginning with ethnographic mapping to select a maximally diverse set of initial subjects, and then conducting only a single wave to preserve the diversity of the initial sample and avoid the unknown biases that would arise from multiple waves. In Klovdahl's (1989) "random-walk" approach, the number of referrals is limited to one and the number of waves is limited to three.

Interest in chain-referral methods has been fueled by recognition of their power to access members of hidden populations. As demonstrated in the literature on the "small world," even in a nation as large as the United States, every person is indirectly associated with every other person through approximately six intermediaries (Killworth and Bernard 1978–1979). Therefore, everyone in the country could hypothetically be reached by the sixth wave of a maximally expansive chain-referral sample. This ability to reach even those who shun public locations makes chain-referral sampling potentially powerful as a method for sampling hidden populations.

Because of biases associated with the method, however, chain-referral sampling is generally considered a form of convenience sampling for which no claims of representativeness can be made. The first bias derives from the choice of the *initial sample*. As Erickson (1979:299) states, "inferences about individuals must rely mainly on the initial sample, since additional individuals found by tracing chains are never found randomly or even with known biases." This issue is important because in the contexts where chain-referral methods are used, the initial sample cannot be drawn randomly.

Second, chain-referral samples tend to be biased by *volunteerism* (Erickson 1979), in which more cooperative subjects agree to participate in larger numbers. The initial subjects are especially likely to be subject to this bias because they frequently make themselves known to researchers, but subsequent waves of recruits can also be affected by this source of bias.

Third, bias depends on the manner in which chain-referrals take place. In Goodman's snowball sampling and Klovdahl's random walk procedure, respondents provide the names and contact information for population members, and the researcher randomly chooses a fixed number from that list. This procedure introduces several problems. First, contact information is frequently inadequate, so the attrition rate is high (Klovdahl 1989). Knowing how to go to a friend's house does not mean one knows the address. This procedure also requires that respondents violate the confidentiality of other population members, so respondents may protect their peers by refusing to refer them, a procedure called "masking." Finally, asking respondents to put their peers at risk by disclosing their membership in a stigmatized hidden population is so ethically unacceptable that the Institutional Review Boards, which govern federally funded research in the United States, forbid this form of referral. Therefore, referrals now generally involve recruitment by respondents rather than by researchers: In accepting peer recruitment, respondents choose to become known to researchers. Although this resolves the ethical and masking problems, it introduces another potential source of bias, *differential recruitment*. If one group recruits more peers than other groups, its recruitment pattern will be overrepresented in the sample.

Fourth, these samples are subject to a *homophily* bias. As emphasized by Erickson (1979), referrals are made nonrandomly. Subjects refer those with whom they have social ties, such as friends, relatives, and other associates; and recruitment patterns reflect these affiliations. As Galton recognized more than a century ago, affiliations tend to form among those who are similar in age, education, prestige, social class, and race and ethnicity (McPherson and Smith-Lovin 1987). Hence, the composition of each wave biases the subsequent wave.

Fifth, referrals occur through network links, so the sample overrepresents those with large personal networks because the number of potential recruitment paths leading to them is greater. Thus, the most gregarious and socially central subjects are drawn differentially into the sample, and more socially isolated members of the population are neglected.

The aim of this paper is to show that, despite those biases, indicators computed from chain-referral sampling data can provide the basis for valid statistical inference. Statistical inference is generally based exclusively on probability sampling methods, in which the probability of each population element being selected for the sample is known. As Kalton (1983:90) stated,

> The major strength of probability sampling is that the probability selection mechanism permits the development of statistical theory to examine the properties of sample estimators. Thus, estimators with little or no bias can be used, and estimates of the precision of sample estimates can be made.

He went on to say that one cannot estimate the precision of estimators from nonprobability samples; precision can be assessed only by "subjective evaluation." However, this is not necessarily the case. For if a nonprobability sampling process is modeled in sufficient detail, a statistical theory can also be constructed, based on which unbiased indicators can be constructed, and estimates of the precision of sample indicators can be made. The following analysis is based on the premise that *when chain-referral methods are statistically modeled in sufficient detail, it is possible to derive statistically valid indicators and quantitatively determine their precision.*

The analysis builds on a recent paper (Heckathorn 1997) that introduced a method termed *respondent-driven sampling (RDS)*, which included two components—a subject recruitment mechanism termed participant-driven recruitment in which lengthy referral chains were produced by a combination of incentives for peer recruitment and recruitment quotas; and a theoretic model of the sampling process from which population indictors were computed. It showed that the first two sources of bias, choice of initial subjects and volunteerism, can be reduced by redesigning the sampling process. That paper also specified the condition under which the fourth source of bias, homophily, would cancel out, and demonstrated through a sensitivity analysis that even when it did not cancel out, the resulting bias tended to be modest. This paper expands the RDS method in two ways. First, it introduces new means for computing indicators that are not biased by either differences in homophily or network size.

Second, it shows how a modification of a bootstrapping procedure can be employed to analyze the variability of indicators and thereby compute standard errors for population estimates.

Part I summarizes the analytic models that provided the technical basis for RDS (Heckathorn 1997) and analyzes their limitations. Part II extends the analysis to show how sources of bias that remained uncontrolled in the original presentation of RDS can be controlled. Part III introduces means for computing standard errors and analyzes the conditions under which RDS-derived population estimates are most and least statistically efficient. Finally, the conclusion discusses remaining limitations of the method, and potential further refinements and applications.

I. Limitations of Respondent-Driven Sampling

As originally presented (Heckathorn 1997), respondent-driven sampling was based on two analytic models. The first, the Markov model provided a statistical model of the sampling process; the second, the homophily model, provided a statistical model for an important source of bias, the tendency of respondents to refer those who are similar. This section summarizes the limitations of these models. Because the focus of the discussion is methodological rather than substantive, details regarding the larger context of the study (e.g., its role as part of a HIV-prevention intervention termed a *peer-driven intervention* that was developed with Robert Broadhead) are omitted (see Broadhead and Heckathorn 1994; Heckathorn 1990; Heckathorn, et al. 1999).

Sampling as a Markov Process

A chain-referral sample can be viewed as a stochastic process in which the social characteristics of each recruiter affect the characteristics of the recruits. In the case of race and ethnicity, this means that recruiters of each ethnic group[1] generate a distinct ethnic mix of recruits. For example, Table 1A reports recruitment by race and ethnicity in a study of injection drug users. A tendency toward within-group recruitment is readily apparent. For example, non-Hispanic whites, on average, recruit 81% from within, 8% Hispanics, 6% non-Hispanic blacks, and 5% others. Hence, recruiting occurs preponderantly within the ethnic group, but cross-ethnic recruitment also occurs. Similarly, Hispanics recruit, on average, 45% other Hispanics, 43% whites, 10% blacks, and 2% others, so the tendency toward within-group recruitment persists. This pattern continues among blacks, whose in-group recruitment rate is 36%, and the number of acts of recruitment by "others" is too small (n = 7) to reveal any consistent pattern. What is clear is that ethnicity affects recruitment as does gender (see Table 1B).

Recruitment can be modeled as a Markov process, a form of stochastic process with two essential characteristics. First, the process can assume a limited number of states, e.g., four ethnic groups. Second, the process is state dependent, where the probability of moving from state to state depends on a transition probability matrix, e.g., when Table 1's recruitment proportions are interpreted as probabilities (i.e., see the shaded portion of Table 1A), the probability that the next recruit will come from a given group depends on the group from which the current recruiter comes. Thus, the probability of the next recruit being Hispanic is 45% if the current recruiter is Hispanic, 14% if that recruiter is black, 8% if the recruiter is white, and 6% if the recruiter comes from the "other" group.

Analysis showed that the recruitment process had several characteristics. First, it was found to be a memoryless process, in that recruitment patterns depended only on the recruiter, not on the recruiter's recruiter. This means that recruitment corresponded to what is termed a first-order Markov process (Heckathorn 1997:183). Second, no groups recruited exclusively from within. Therefore, recruitment was "ergodic." A process is termed ergodic if,

1. Here and elsewhere in this paper, Blau's (1977) use of the term *group* is adopted, in which this term refers both to groups in the standard sense (i.e., sets of affiliated individuals) and to collectivities (i.e., sets of individuals who share a common demographic or other characteristic such as members of a racial or ethnic group).

Table 1 • ***Recruitment by Race/Ethnicity and Gender***

A: Race and Ethnicity	*Race and Ethnicity of Recruit*				
	White	*Hispanic*	*Black*	*Other*	*Total*
Race and ethnicity of person who recruited					
Non-Hispanic white					
Recruitment count	102	10	8	6	126
Selection proportion, S	.810	.079	.063	.048	1
Adjusted count	107.657	10.555	8.444	6.333	132.988
Hispanic					
Recruitment count	18	19	4	1	42
Selection proportion, S	.429	.452	.095	.024	1
Adjusted count	13.73	14.493	3.051	.763	32.036
Non-Hispanic black					
Recruitment count	7	2	5	0	14
Selection proportion, S	.5	.143	.357	0	1
Adjusted count	8.94	2.554	6.386	0	17.881
Other					
Recruitment count	3	5	0	0	8
Selection proportion, S	.375	.625	0	0	1
Adjusted count	2.661	4.435	0	0	7.096
Total distribution of recruits	130	36	17	7	190
Sample distribution, SD	.684	.189	.089	.037	1
Equilibrium, E	.7	.169	.094	.037	
Mean network size, N	55.2	38.4	63.3	76.7	
Homophily, H	.362	.317	.301	−1	
Population estimate, P (linear least squares)	.702	.198	.08	.02	
Standard error of P	.048	.038	.029	.013	

B: Gender	*Gender of Recruit*		
	Female	*Male*	*Total*
Gender of recruiter			
Female			
Recruitment count	21	29	50
Selection proportion, S	.42	.58	1
Adjusted count	24.98	34.496	59.475
Male			
Recruitment count	37	103	140
Selection proportion, S	.264	.736	1
Adjusted count	34.496	96.029	130.525
Total distribution of recruits	58	132	190
Sample distribution, SD	.305	.695	1
Equilibrium, E	.313	.687	
Mean network size, N	37.5	57.1	
Homophily, H	.018	.355	
Population estimate, P	.41	.59	
Standard error of P	.039	.039	

as a process moves from state to state, any state can recur, and there is a zero probability that any state will never recur. When applied to a chain-referral sample, the states refer to the characteristics of the subjects, the movement from state to state refers to a recruiter with one set of characteristics recruiting another subject with the same or different characteristics, and that any state can recur means that after one or more recruitment waves a recruit can have the same characteristics as the earlier recruiter. In essence, this means that recruitment cannot become trapped within a single group or set of groups, as would occur if once the recruitment chain entered that group, no exit (i.e., no outside recruitment) were possible. Thus, analyses revealed that recruitment corresponded to what is termed a "regular" Markov process.

This modeling of the recruitment process is relevant to understanding the reliability of indicators drawn from respondent-driven samples because of two deductions regarding regular Markov processes. First, the *law of large numbers for regular Markov chains* (Kemeny and Snell 1960:73) implies the following (Heckathorn 1997):

> *Theorem 1:* As the recruitment process continues from wave to wave, an equilibrium mix of recruits will eventually be attained that is independent of the characteristics of the subject or set of subjects from which recruitment began.

Thus, allowing recruitment to operate until equilibrium is reached in a sample corresponding to a regular Markov process avoids the central problem for sampling hidden populations—that the sample's characteristics merely reflect the initial sample. Instead, the sample composition is wholly independent of the initial subjects. For example, Figure 1 shows what would be expected over the course of eight waves had recruitment begun with seeds of a single ethnicity. Figure 1A projects the expected course of recruitment by wave had it begun with all Hispanics. The composition of the waves ultimately reach a stable equilibrium at 70% white, 17% Hispanic, 9% black, and 4% other. Figure 1B projects the expected course of recruitment had it begun with all non-Hispanic whites, and exactly the same equilibrium is attained. This shows graphically how the sample can reach the same equilibrium point, irrespective of the subjects with whom recruitment began. What happens, in essence, is that as sampling progresses, the effect of the starting point become progressively weaker until it becomes negligible.

The law of large numbers for regular Markov chains provides a means for computing the equilibrium analytically (see Kemeny and Snell 1960:72). The equilibrium state ($E = E_a, E_b, \ldots, E_n$) for a system with N types of subjects is found by solving a system of N linear equations:

$$\begin{aligned} 1 &= E_a + E_b + \ldots + E_n \\ E_a &= S_{aa}E_a + S_{ba}E_b + \ldots + S_{na}E_n \\ E_b &= S_{ab}E_a + S_{bb}E_b + \ldots + S_{nb}E_n \\ &\vdots \\ E_{n-1} &= S_{a,n-1}E_a + S_{b,n-1}E_b + \ldots + S_{n,n-1}E_n \end{aligned} \tag{1}$$

where $E_a, E_b, \ldots E_n$ are the equilibrium proportions for groups A, B, to N, respectively, and S_{xy} is the probability that a subject of type X will recruit a subject of type Y. The first equation states that the equilibrium proportions must sum to one. The subsequent equations express the groups' equilibrium sizes as a function of the equilibrium sizes of the groups and the groups' proportional recruitment of each group. Because this is a system of N linear equations with N unknowns, there is a unique solution.

For example, a two-group system is defined by the equation system,

$$\begin{aligned} 1 &= E_a + E_b \\ E_a &= S_{aa}E_a + S_{ba}E_b \end{aligned} \tag{2}$$

Solving this system yields,

$$\begin{aligned} E_a &= \frac{S_{ba}}{1 - S_{aa} + S_{ba}} \\ E_b &= 1 - E_a \end{aligned} \tag{3}$$

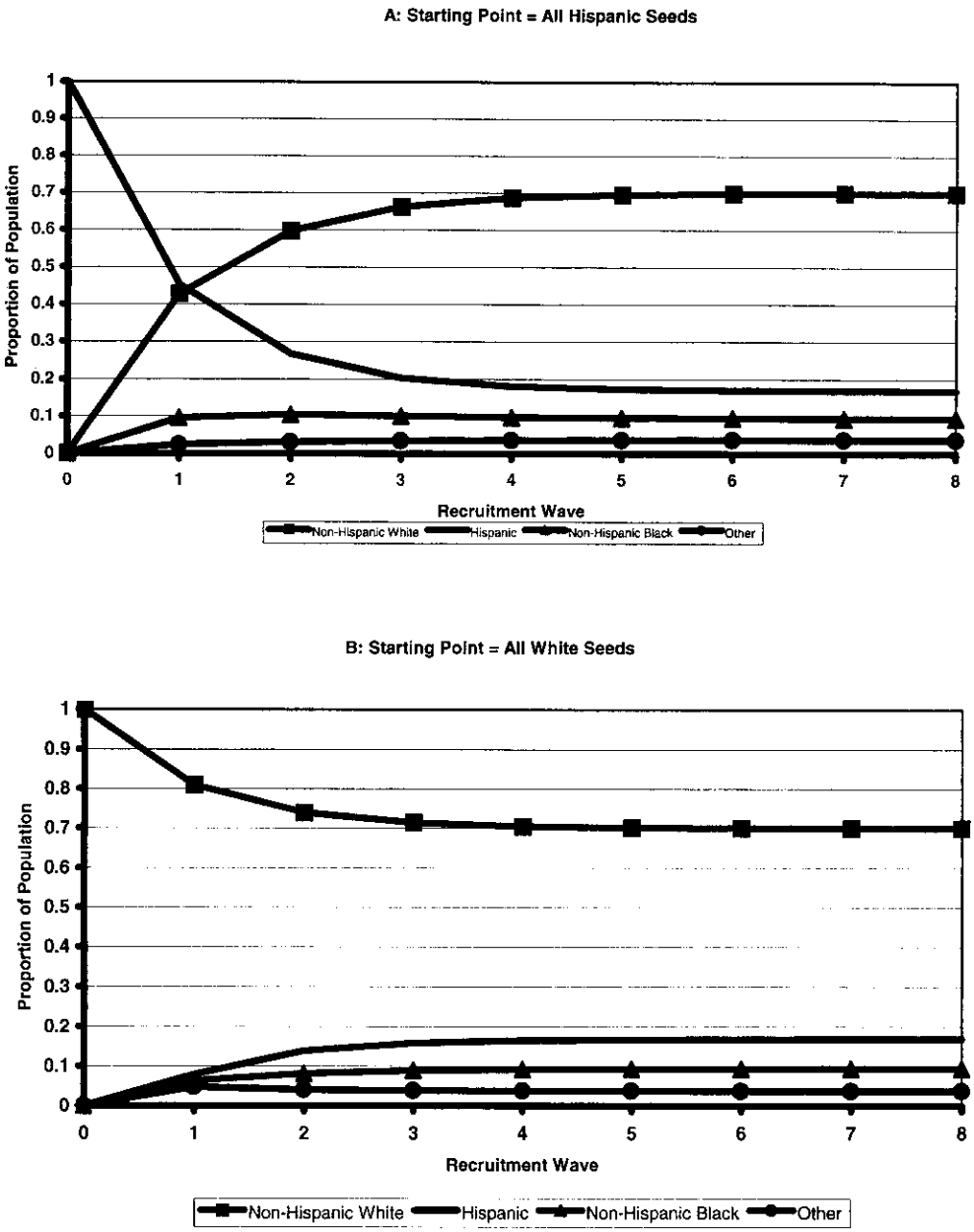

Figure 1 • *Race and Ethnicity of Recruits Expected by Wave in a Respondent-Driven Sample, Beginning with only White or Hispanic Seeds*

For example, in Table 1B, where $S_{ba} = .264$ is the proportion of females recruited by males and $S_{aa} = .42$ is the proportion of females recruited by females, the equilibrium distribution of female IDUs is $E_a = .264/(1 - .42 + .264) = .313$, and for males $E_b = 1 - .313 = .687$. Therefore, the equilibrium comprises 31% female IDUs and 69% males.

Regular Markov chains are characterized by what Kemeny and Snell (1960:72) describe as "a very fast kind of convergence." This conclusion is based on a deduction stating that convergence occurs at a geometric rate. The implication was described as follows (Heckathorn 1997):

> *Theorem 2:* The set of subjects generated by a respondent-driven sampling process approaches equilibrium at a rapid (i.e., geometric) rate.

Thus as the sample expands from wave to wave, the mean composition of the subjects in the leading wave approaches equilibrium at a geometric rate. However, because the overall sample includes both current and preceding waves, it approaches equilibrium at less then geometric, but nonetheless rapid rate. For example, in the race/ethnicity analysis, the leading wave approximates equilibrium within 2% after only three waves, but when respondents recruit three peers, the overall sample requires six waves. The size of this lag depends on the number of recruits per subject. For example, if this number is increased to seven, only five waves are required. This occurs because the greater the number of recruits per subject, the greater the rate at which the number of subjects in each wave increases, and hence the less will be the influence of earlier waves on sample composition. In previous applications of RDS, equilibrium was approximated within six or fewer waves (Heckathorn 1997; Heckathorn, et al. 1999).

If recruitment takes the form of the simplest Markov process, that is, a linear chain beginning with a single seed, the implication of this theorem is that recruitment chains should be long. For only then will most of the subjects be drawn from waves after which equilibrium was attained. However, producing linear chains faces a problem. For example, in Klovdahl's (1989) random walk approach, which employs linear recruitment chains, but most chains are shorter than the maximum of three steps, due to attrition when a respondent fails to provide a valid referral. Therefore, the approach employed in many chain-referral sampling methods allows respondents to produce multiple referrals, thereby generating a tree-shaped recruitment network. This approach has two advantages. First, it helps to resolve the attrition problem, because after a seed has produced multiple referrals, no single subject's failure to produce valid referrals can stop the sampling process. Therefore, given ample time and resources, recruitment chains of virtually any length can be generated. Second, multiple referrals help to reduce bias from the choice of initial subjects, because as recruits become more socially distant from the seed (i.e., as the sample expands wave by wave), they also become more numerous. This occurs because the number of respondents in each wave expands geometrically at a rate determined by the number of referrals produced by each respondent.

There is also a potential complication from the introduction of multiple referrals, because the tree-shaped recruitment structure produced by multiple referrals does not correspond to the linear structure assumed by the Markov-chain model. To formally analyze the implications of this structural difference for the applicability of the Markov model to chain-referral sampling would exceed the scope of this paper, but two comments can be offered. First, a tree-shaped referral structure can be analyzed as a *set* of linear structures, e.g., a respondent from the final wave can be seen as the product of a linear chain beginning with the seed and whose links correspond to intermediate waves. It seems reasonable to suppose that an analysis that is valid for a linear chain will also be valid for a set of such chains. Second, whether the Markov model fits the data should be determined empirically, by comparing the actual sample composition with the sample composition that will be theoretically expected if the sampling process corresponds to a Markov process, that is, the equilibrium. For example, in a previous application of RDS (Heckathorn 1997:188–189), a large (17.1%) discrepancy was found between a theoretically-computed equilibrium and a sample mean, and an examination revealed that recruitment was not ergodic. This problem was solved by dividing the sample into two sub-samples, each of which was ergodic. In contrast, in the analyses reported in Table 1, the correspondence between the equilibrium and sample means is close. The mean absolute differences between the equilibrium and sample means for the race/ethnicity and gender analyses are only 1%. Therefore, the fit with the Markov model appears to be good. This suggests not only that the ergodic and

other assumptions of the Markov model were approximated to a reasonable extent, but also that other factors, such as the nonlinear recruitment structure, did not serve as a confounding factor.

It might seem that long referral chains would be unnecessary, given that the equilibrium can be computed from the transition probability matrix, and this matrix depends not on the length of chains, but on the number of referrals. Thus, one could follow Frank and Snijders (1994) recommendation and begin with a large number of seeds chosen for their diversity, and conduct only a single wave. However, this approach has two disadvantages. First, the sample would lack sociometric depth. The part of the hidden population accessible to researchers may not be representative of the full population. Even when chosen for diversity, the seeds constitute a mere convenience sample. If only a single wave is conducted, all of the subjects will lie within a single link of those respondents who are accessible to researchers. More socially distant sectors of the population will therefore not appear in the sample. In contrast, when recruitment chains are a dozen or more steps in length, consistent with the above-noted literature on "six degrees of separation" all members of the population should be reachable. Second, long referral chains are efficient, because respondents who are neither a seed nor a member of the final wave play a dual role, as both the source and the product of referral. The longer the referral chains, all else equal, the greater is the number of these intermediate respondents, and therefore the greater is the ratio of referrals to respondents.

In RDS, long referral chains are produced in two ways. First, respondents are rewarded for recruiting; rewards averaged $14 each. Second, quotas were imposed on recruitment so that no small subset of recruiters could monopolize recruitment rights; the quota was set at three recruits after the initial interview and each follow up interview. Quotas on recruitment were implemented using a coupon system, in which potential recruiters are given dollar-bill-sized coupons to give their recruits. The coupon includes the study name, a phone number to call to make an appointment for an interview, and a map to the interview site. The coupon also includes a serial number that documents the link between the recruiter to whom it was given and the recruit who returns it to the project (Heckathorn, Broadhead, and Sergeyev 2001). Combining recruitment incentives and quotas yielded recruitment chains that were very long by the standards of the chain-referral literature; in some cases a single seed initiated a chain of referrals that resulted in more than one hundred recruits over the course of more than a dozen waves. A second reason for introducing quotas was to avoid order effects in recruitment. As shown in the network literature on name generators (Burt 1986), first named persons differ from last named persons. Limiting recruitment coupons to three ensured that all recruits were near the beginning of the queue of potential recruits.

The introduction of recruitment incentives not only lengthened referral chains, it also reduced bias due to volunteerism (Heckathorn 1997). Recruitment incentives harness peer pressure by motivating the potential subject's peers to employ their social influence. In essence, recruitment incentives serve as transformers that convert material incentives (i.e., the reward to recruiters) into peer-based symbolic incentives (i.e., the social influence exercised by recruiters). A comparison of recruitment patterns and self-reported network composition based on race and ethnicity, gender, and homelessness showed a strong association (Heckathorn, et al. in press). Therefore, respondents appeared to recruit as though they were selecting randomly from their personal networks.

In sum, chain referral samples can yield reliable indicators, in that sample means reach the same equilibrium independent of starting point. What is required is that recruitment chains be long enough for equilibrium to be approximated. In the RDS method, such referral chains were produced by a recruitment mechanism termed Participant-Driven Recruitment that combines recruitment incentives and quotas. However, an important limitation of the analysis was the absence of means to assess the variability of indicators *quantitatively.* Section III below removes this limitation by introducing means for computing standard errors for population estimates.

The Homophily Model

Reliability is a necessary, but not sufficient condition for an effective indicator, for a reliable indicator may be biased. The original presentation of RDS (Heckathorn 1997) analyzed one type of systematic bias in data from chain-referral samples, that due to homophily. Drawing on Fararo and Sunshine's (1964), Blau's (1977, 1994), and Rapoport's (1979) models, homophily was formally defined as follows: Perfect homophily, in which all ties are formed within the group, is assigned the value +1; and no homophily, in which ties are formed without regard to group membership, is assigned the value zero. When an individual forms ties within the group, say, a third of the time, and forms ties randomly, without regard to group membership, two thirds of the time, the level of homophily is plus one third. For example, as shown in Table 1A, Hispanic homophily is .317. This implies that they recruit from within 31.7% of the time; the other 68.3% of the time they recruit randomly, and given that there are an estimated 19.8% Hispanics in the population, nonhomophily governed recruitments produce an additional 13.5% (68.3% × 19.8%) Hispanic recruits, for a total of 45.2% (31.7% + 13.5%) recruits from within.

The concept of homophily can be extended to cover the case where a bias exists against forming ingroup ties. This is termed *heterophily.* When all ties are formed outside the group, homophily is assigned the value −1. Intermediate levels of negative homophily are defined in a way parallel to intermediate positive levels: If, for example, ties are formed with those outside the group, one third of the time, and ties are formed randomly, without regard to group membership, two thirds of the time, homophily is negative one third.

To further specify the homophily model, consider the case of a system comprising two groups, A and B, where homophily is positive. The probability that a member of group A will select from the in-group, S_{aa}, is the sum of the probability that selection is controlled by homophily, an event with probability H_a, and the probability that homophily does not govern choice $(1 - H_a)$, weighted by the proportion of members of group A in the population, P_a, i.e.,

$$S_{aa} = H_a + (1 - H_a)P_a \tag{4}$$

By similar principles, the probability that a member of group B will select a member of group A is the probability that homophily does not govern B's choice $(1 - H_b)$, weighted by the proportion of members of group A in the population, i.e.,

$$S_{ba} = (1 - H_b)P_a \tag{5}$$

Extending this model to cases where homophily is negative requires slight modifications in these expressions. For example, when a group is heterophilous, forming an in-group tie requires the conjunction of two events, that heterophily not govern tie formation, an event with probability $1 + H_a$, and that a member of the in-group then be selected irrespective of group identity, an event whose probability depends on the group's proportional size, P_a, so the probability of forming an in-group tie is the product of the probabilities of these two events, i.e., if $H_a < 0$, then $S_{aa} = (1 + H_a)P_a$. In sum, homophily's absolute value $|H_a|$ is the probability that homophily governs tie formation. When homophily is positive, it is the probability that a tie is formed from within rather than being formed irrespective of group affiliation, and when homophily is negative, it is the probability that a tie is formed from the out-group rather than being formed irrespective of group affiliation. With this set of equations, the model specifies the relationship between the population size (P), and selection probability (S).

Based on the combination of this model for homophily and the Markov model, a theorem was derived demonstrating that the equilibrium distribution (E) equals the population distribution (P) if homophily is equal. In that case, the equilibrium provides a nonbiased population estimate. In essence, what was shown was that groups with higher homophily are over sampled, but over sampling cancels out if all groups have equal homophily. The equilibrium

thereby converges with the population distribution, and hence the former becomes a nonbiased estimator for the latter. This conclusion was stated in a third theorem:

> *Theorem 3:* A respondent-driven sample is unbiased by homophily (i.e., E = P) if the homophily of each group is equal (i.e., for each group x and y, $H_x = H_y$).

Theoretic arguments were then offered suggesting that high homophily in one group tends to produce high homophily in other groups, thereby reducing differentials in homophily. Then a sensitivity analysis was conducted in which, assuming varying degrees of association in homophily, the results indicated that the correlation between equilibrium and population composition was high even when homophily was weakly associated, and it remained substantial even when homophily across groups were independent. Therefore, the sensitivity analysis suggested that the biasing effects of unequal homophily would be modest. This analysis had two limitations. First, no means for measuring homophily were offered, because a crucial term, population size (i.e., P in equations 4 and 5), is not known when sampling a hidden population. Second, no means were provided for controlling for any bias resulting from unequal homophily. The limitations are overcome in Section II below.

By way of closing this section, it should be noted that the initial presentation of RDS had an unacknowledged strength, providing a means for controlling bias due to differential recruitment. This bias occurs when one group recruits especially effectively and its distinctive recruitment pattern is thereby overrepresented in the sample. Though recruitment quotas reduce this form of potential bias, given that not all subjects fulfill their quotas, variation remains. The Tables provide evidence for such recruitment differentials. For example, in the gender analysis, male IDUs made up 69% of the sample (132/190), but they recruited 74% of respondents (140/190), so they recruited more on average than did female IDUs. Differentials also exist in the ethnicity analysis, with Hispanics and others recruiting more, and blacks and whites recruiting less. The observed differences between sample composition and the equilibrium in Table 1 results, in part, from these differentials in recruiting.

This source of bias does not affect the Markov model's equilibrium, because it depends not on the absolute number of recruits from each group, but rather on the *proportional* distribution of recruits (i.e., the S terms). These are what drive the model. That is, because transition probabilities are based on the proportional distribution of each group's recruits, the probabilities remain the same whether all groups recruit equally or some groups recruit more or less than others. To see why this is the case, consider the effect of *demographically adjusting* the recruitment counts to compensate for differences in recruiting success. This was done in the third entry in Table 1's recruitment cells, in which recruitment counts were adjusted based on the assumption of equal recruitment success such that the number of recruits from each group (the row sum) equals the number of recruitments by each group (the column sum), without any change in recruitment pattern or sample size. That is, the adjusted recruitment count is the selection proportion multiplied by the equilibrium proportion of recruits from that category and the total number of recruitments for all categories. For example, females recruited a proportion of .42 other females, the proportion of females in the equilibrium sample is .313, and total recruitment was 190, so the adjusted recruitment count for females is $.42 \times .313 \times 190 = 24.98$. This is the *expected* number of recruits of female IDUs by female IDUs had both genders recruited with equal success. As thus adjusted, the number of recruitments of and by female IDUs (i.e., the column and row sums) is the same, 59.475. Because the recruitment proportions (i.e., the second entry in each cell, the S terms) are the same whether the calculation is based on the actual or the adjusted counts, a term computed based on these proportions, such as the equilibrium, E, is independent of differential recruitment. In essence, the equilibrium is computed as though all groups recruited equally.

II. Extending The Respondent-Driven Sampling Method

Given the limitations of the above-described models, the need for additional theoretic development is evident. Such modeling can begin by analyzing in more detail the structure of recruitment networks. An important feature of recruitment networks generated by RDS is that they reflect preexisting social relationships among subjects. For example, recruiters were usually friends or acquaintances (90%), and most other recruiters had even closer relationships, e.g., sex partner, spouse, or other relative (7%). Only a small proportion of recruiters were identified as strangers (1%), and a few others had a relationship identified as "other" (2%).

The recognition that a preexisting social relationship linked recruiters and recruits is significant theoretically because such relationships generally are *reciprocal* (e.g., if A has a link to B, then B has a link to A). This provides the basis for a richer model of recruitment structure. When relationships are reciprocal, for any two groups A and B, the number of ties from A to B, T_{ab}, equals those from B to A, T_{ba}, i.e.,

$$T_{ab} = T_{ba} \tag{6}$$

The number of such crosscutting ties depends on three factors. One is the mean personal network size for members of the group, where each individual's personal network size is the number of other population members he or she knows. This is therefore also the number of potential recruits known to the person. A second factor is the proportional size of the group, P_a. The third term is the proportion of crosscutting ties. This proportion is the second line in each recruitment cell in Table I, and in the model it is also treated as reflecting the probability of forming a crosscutting tie; e.g., S_{ab} is the probability that a member of group A will form a tie with a member of group B. Therefore, the number of ties from group A to B is the product of these three terms,

$$T_{ab} = P_a N_a S_{ab} \tag{7}$$

Equation 6 can be expanded based on equation 7 as follows:

$$P_a N_a S_{ab} = P_b N_b S_{ba} \tag{8}$$

After substituting $1 - P_a$ for P_b, this expression can be solved for P_a to derive an estimate of population size,

$$P_a = \frac{S_{ba} N_b}{S_{ba} N_b + S_{ab} N_a} \tag{9}$$

This is the estimate of population size based on the *reciprocity model.* It provides an estimate of the proportional size of the hidden population based on two sources of data: the transition probabilities derived from the analysis of recruitment patterns, and self-reported personal network size. The latter is gathered routinely in health-related research (Killworth, et al. 1990) because it is used to measure social integration and the risk of a number of other conditions, including depression (Marsden 1990). The current study was no exception, so data on network sizes were available. For example, female IDUs had substantially smaller networks ($N_f = 37.5$) than did male IDUs ($N_m = 57.1$). When the above expression is used to estimate the proportion of IDUs by gender, the estimated proportion of female IDUs is $P_f = .41$ (i.e., $P_f = (.264 \times 57.1)/(.264 \times 57.1 + .58 \times 37.5)$). Similarly, the estimated proportion of male IDUs is $P_m = .59$.

When the population estimate derived from the reciprocity model is compared with the equilibrium or sample distributions, the results are consistent with the commonsense notion that groups with larger networks will be over sampled. For example, male IDUs had personal networks that were on average one half larger than female IDUs. Though male IDUs made up an estimated 59% of the population based on the reciprocity model, they were 68.7% of the

equilibrium and 69.5% of the actual sample. Similarly, groups with smaller networks are under sampled. Though female IDUs made up an estimated 41% of the population based on the reciprocity model, they made up only 31.3% of the equilibrium and 30.5% of the actual sample.

Just as the analysis of the equilibrium extends straightforwardly to systems with more than two groups, so too does the reciprocity model. In both cases, analyzing a system requires solving a system of linear equations. The reciprocity model in a system with N groups can be represented by a system of equations, in which the first states that proportional population sizes must sum to one, and the others express the reciprocity principle for each of the pair of groups, where the number of pairs is $(N(N-1))/2$. Therefore, a system with four groups is described by seven equations, as follows:

$$\begin{aligned} 1 &= P_a + P_b + P_c + P_d \\ P_a N_a S_{ab} &= P_b N_b S_{ba} \\ P_a N_a S_{ac} &= P_c N_c S_{ca} \\ P_a N_a S_{ad} &= P_d N_d S_{da} \\ P_b N_b S_{bc} &= P_c N_c S_{cb} \\ P_b N_b S_{bd} &= P_d N_d S_{db} \\ P_c N_c S_{cd} &= P_d N_d S_{dc} \end{aligned} \tag{10}$$

Here the N terms refer to network sizes, the S terms derive from the transition matrix, and the P terms are the population estimates to be derived from the reciprocity model. Solving this system of equations requires more than standard algebra because of a superabundance of equations. To be determinate, a system of linear equations must have the same number of equations and unknowns, yet here there are seven equations and only four unknowns (i.e., the P terms), so the system is *over-determined.*

If the fit between the reciprocity model and the data were perfect, it would suffice merely to choose four equations arbitrarily and solve for the four P terms. However, since fit with real data is never perfect, that choice matters. For example, if one focuses only on recruitments involving non-Hispanic whites, and therefore considering only the first four equations of equation system 10, solving them using standard algebra yields a population estimate of .681, .181, .075, and .062 for whites, Hispanics, blacks, and others, respectively. However, if one focuses only on recruitments involving Hispanics, therefore considering equations 1, 2, 5, and 6 of equation system 10, the population estimate is .725, .193, .078, and .004. That these two estimates differ by only an average of 2.9% indicates that the fit with the reciprocity model is substantial.

The standard means for solving over-determined systems is linear least squares (Farebrother 1988), a procedure that employs the same logic as linear regression.[2] When the ordinary least squares (OLS) version of this method is applied to the race/ethnicity analysis, the estimated sizes are .702, .198, .08, and .02, for whites, Hispanics, blacks, and others, respectively. This corresponds fairly closely to the above two partial estimates, differing by an average of 2.1% and 1.2%, respectively. The advantage of the linear least squares approach is that it relies on a standard statistical method to resolve conflicts among the equations.

Data Smoothing

An alternative method for deriving the population estimate draws on the logic of the reciprocity model to solve the problem of over-determination. Recall that in a system where ties are reciprocal, the number of directed ties between any pair of groups will be equal. Therefore, if recruitment patterns reflect the distribution of ties in the system, as would occur in equilibrium if all groups recruited equally effectively, the number of recruitments across groups would also tend to be equal. Therefore, what were above termed *demographically adjusted* recruitment counts can be theoretically expected to approach equality across groups.

2. I am grateful to Mathew Salganik for suggesting the use of linear least squares in this context.

Table 2 • ***Recruitment by Race/Ethnicity and Gender, with Data Smoothing***

	Race and Ethnicity of Recruit				
	White	*Hispanic*	*Black*	*Other*	*Total*
Race and ethnicity of person who recruited					
Non-Hispanic white					
Adjusted and smoothed count	107.657	12.142	8.692	4.497	132.988
Selection proportion	.810	.091	.065	.034	1
Hispanic					
Adjusted and smoothed count	12.142	14.493	2.803	2.599	32.036
Selection proportion	.379	.452	.087	.081	1
Non-Hispanic black					
Adjusted and smoothed count	8.692	2.803	6.386	0	17.881
Selection proportion	.486	.157	.357	0	1
Other					
Adjusted and smoothed count	4.497	2.599	0	0	7.096
Selection proportion	.632	.366	0	0	1
Total distribution of recruits	132.988	32.036	17.881	7.096	190
Sample distribution, SD	.7	.169	.094	.037	1
Equilibrium, E	.7	.169	.094	.037	
Homophily, H	.430	.288	.303	−1	
Population estimate, P	.666	.231	.078	.026	
Standard error of P	.05	.042	.028	.013	

Consistent with this proposition, the tables show that these cross-recruitment counts are positively associated. For example, in Table 1A, the correlation between raw cross-group counts is .78, and this increases to .85 when the counts are demographically adjusted.

From the standpoint of the reciprocity model, differences across groups in demographically adjusted recruitment counts reflect sampling error. Therefore, the *best estimate* for these counts is the mean count across groups, which may be termed the *smoothed* count. These means are reported in the first line of each recruitment cell of Table 2. For example, the smoothed cross-recruitment count between Hispanics and non-Hispanic whites reported in Table 2 is the mean of their demographically adjusted counts from Table 1A, (8.444 + 8.94)/2 = 8.692. Thus, by averaging demographically adjusted recruitment counts in reciprocal cells, reciprocal counts become identical, and the recruitment matrix is rendered exactly compatible with the reciprocity model.

Table 2 shows the effect of data smoothing on the ethnicity analysis. Smoothing brings the recruitment counts into precise alignment with the reciprocity model's assumptions and thereby provides an alternative solution to the problem of over-determination. For example, in a four-category case such as the race/ethnicity analysis, the population estimate is the same whether based on all recruitment information (i.e., all seven equations in equation system 10 solved by linear least squares), or on only a partial model (e.g., only the first four equations in equation system10). The smoothed estimate is P = (.666, .231, .078, .026), for whites, Hispanics, blacks, and others, respectively. This differs by an average of 1.9% from the population estimate derived by linear least squares without smoothing. This relatively small discrepancy reflects the strong, but not perfect, correspondence of the original recruitment matrix to the reciprocity model. Of course, data smoothing has no effect on two-category systems because the over-determination problem does not arise. Unless otherwise specified subsequently in this paper, population estimates for systems with more than two categories will employ data smoothing.

Controlling Bias Due to Differential Network Size

A commonsense notion holds that when network sizes are equal, the group with the larger network will be over sampled. The implication is that when network sizes are equal, this source of bias will not exist. This intuition can be tested formally by checking to see whether the population estimate, P, equals the equilibrium distribution, E, when network sizes are equal. In that case, in a two-group system, network sizes for each group will be equal so N_a can be substituted for N_b in equation 9's expression for P. Furthermore, $1 - S_{aa}$ can be substituted for S_{ab}. Making these substitutions into equation 9 yields,

$$P_a = \frac{S_{ba}N_a}{S_{ba}N_a + (1 - S_{aa})N_a} \tag{11}$$

With algebraic manipulation, network size cancels out and the expression can be simplified and rearranged as follows:

$$P_a = \frac{S_{ba}}{1 - S_{aa} + S_{ba}} \tag{12}$$

Note that this is the same as the equation for the equilibrium (see equation 3 above). Thus, if network sizes are equal, the equilibrium distribution is equivalent to the distribution from which the sample was drawn (i.e., E = P), and in this sense the sample is unbiased. The converse is also the case, that is, if the sample is unbiased (i.e., if E = P), then network sizes are equal. That is, in an unbiased system,

$$E_a = P_a \tag{13}$$

From equations 3 and 9 above, this expands to

$$\frac{S_{ba}}{1 - S_{aa} + S_{ba}} = \frac{S_{ba}N_b}{S_{ba}N_b + S_{ab}N_a} \tag{14}$$

Given that $S_{aa} + S_{ab} = 1$, $1 - S_{ab}$ can be substituted for S_{aa}. With this substitution, solving for network size, N, yields

$$N_a = N_b \tag{15}$$

Furthermore, given that the reciprocity model treats systems in essence as pairs of dyads, this result generalizes to systems with more than two categories. This leads to a fourth theorem:

> *Theorem 4:* A respondent-driven sample is unbiased by network size (i.e., E = P) if and only if the network sizes of each group are equal (i.e., for each group x and y, $N_x = N_y$).

This theorem demonstrates the consistency of the reciprocity model with the assumption that differential network sizes are a source of bias in chain-referral sampling. However, the prime significance of the reciprocity model on which it is based is that it provides a means for computing a population estimate that is not biased by unequal network sizes.

Network Structural Constraints on Homophily

Theorems 3 and 4 are parallel in that each holds that in the presence of a condition—equal homophily and network size, respectively—the equilibrium will be unbiased (i.e., E = P). Furthermore, theorem 4 holds that the reverse is also the case; if the equilibrium is unbiased, network sizes are equal. The conditions imply a connection between homophily and network size, even though these two terms have been treated as independent in the liter-

ature, because the absence of a connection would introduce a contradiction. For example, consider a hypothetical case in which homophily is equal but network sizes are not. According to theorem 3, E = P, but according to theorem 4, E ≠ P. Obviously, if both theorems are valid, this hypothetical case is impossible, and equal homophily implies equal network size. An examination of the structure of networks of reciprocal ties shows that this nonintuitive conclusion is correct.

Consider first, a system contrary to the assumptions of the reciprocity model, where ties are *nonreciprocal.* Examples of nonreciprocal ties are admiration and knowing a person's name. For example, celebrities are known and perhaps admired by many people they do not know. In such cases, any homophily combination is possible. For example, if individuals admire only members of their own group, there is strict homophily; if individuals admire only members of the other group, there is strict heterophily; if everyone admires only members of one group, that group is homophilous and the other group is heterophilous.

If ties are reciprocal, as in the cases of marriage and friendship, this range of possibilities does not exist, because establishing a tie requires mutual consent and, thus, the combination of strict homophily and heterophily becomes impossible. If one group is strictly homophilous, the other has no out-group members with which to establish heterophilous ties, so it, too, must be homophilous. The network-structural constraints on homophily are especially clear when groups of differential status interact. For example, according to Eder's (1985) study of U.S. high schools, athletes and cheerleaders have the highest status. Members of these groups tend to associate with one another, thereby limiting others' opportunity to gain status by associating with them. This may be termed *power homophily* because it reflects the elite's control over tie formation. The homophily of the high school elite thereby induces homophily among lower-status students. A similar process occurs in career-based hierarchies, as among musicians where the opportunity to play with a luminary represents an important advance in one's career (Heckathorn and Jeffri 2001). This may be termed *exclusion homophily* because it reflects their exclusion from elite circles.

When a group replaces out-group ties with in-group ties and its network size remains unchanged, it becomes more homophilous. This loss of out-group ties also increases the homophily of other groups. Thus, homophily levels are *positively related,* and this occurs not because of the psychodynamic processes addressed by Simmel (1955), but because of the structural properties of networks of reciprocal relationships. In contrast, when a group adds in-group ties without altering out-group ties, it becomes more homophilous and network size increases. The homophily of other groups is unchanged, so whether a change in a group's homophily affects other groups depends on network size. Thus, changes in homophily and network size are interrelated.

The association between homophily and network size can be specified by combining the homophily and reciprocity models. To simplify the analysis, again consider a two-group system. Equation 5 specified the relationship between population size and homophily. If it is solved for population size, P_a, the result is,

$$P_a = \frac{S_{ba}}{1 - H_b} \tag{16}$$

Similarly, for group B, $P_b = S_{ab}/(1 - H_a)$. If these expressions for P_a and P_b are substituted into equation 8 above, the result is,

$$\frac{S_{ba}}{1 - H_b} N_a S_{ab} = \frac{S_{ab}}{1 - H_a} N_b S_{ba} \tag{17}$$

This equation provides the means for identifying the implications of equal homophily for the reciprocity model. This can be done by setting the homophily equal (i.e., substitute H_b for H_a), and solving for N_a. Algebraic manipulation produces,

$$N_a = N_b \tag{18}$$

Thus, if homophily is equal, so too is network size. This demonstrates the consistency between theorems 3 and 4. In contrast, equal network sizes do not imply equal homophily, except in a special set of systems, that is, two-category systems with positive homophily. For example, in the ethnicity analysis, if network sizes are made equal, homophily remains unequal (i.e., H = (.365, .341, .29, −1) for whites, Hispanics, blacks, and others, respectively). Nonetheless, changes in network size remain linked to changes in homophily. For example, in a system with any number of categories, if a cross-cutting tie is severed with no other changes in the system, the affected groups have both smaller mean network sizes and higher homophily; and if a cross-cutting tie is created, the affected groups have both larger mean network sizes and lower homophily, so homophily levels remain positively related. More generally, the creation or abandonment of ties have implications for both network size and homophily whose effect is to establish a conceptual link between these two concepts, a link specified by the above model. Therefore, *by controlling for the effects of differential network size, the reciprocity model thereby also controls for the effects of differential homophily.*

The combination of the homophily and reciprocity models provides the basis for point estimates of homophily that were not possible using only the Markov and homophily models. Homophily (i.e., see equation 5) is a function of both transition probability, which is provided by the chain-referral data, and population size, which can be computed using the reciprocity model based on both transition probability and network size. For example, homophily by gender (see Table 1B) can be computed as follows: From equation 5, above, $S_{ba} = (1 - H_b) \times P_a$, so by substitution where females are group A and males are group B, $.264 = (1 - H_b) \times .41$, which yields $H_b = .355$. Therefore, the male IDUs exhibited substantial homophily, forming networks as though 35.5% of the time they formed a tie to another male IDU, and the rest of the time they formed a tie randomly irrespective of gender. Solving for the homophily of female IDUs yields $H_a = .018$, a near zero (1.8%) level of homophily. Therefore, only the male IDUs were meaningfully homophilous. The relationship between homophily and ethnicity is similarly variable (see Table 2). For example, non-Hispanic whites were the most homophilous (.43). Non-Hispanic blacks and Hispanics had substantial and nearly equal homophily levels (.303 and .288, respectively). These differences in homophily point to the need to take homophily into account as a potential biasing factor in chain-referral samples. That is, despite the positive relationship among homophily levels, differences can be great enough to very significantly affect the sampling process, as is illustrated in particular by the gender analysis.

III. Reliability of Indictors Drawn from Respondent-Driven Samples

Respondent-driven sampling, like other sampling methods, yields indicator subject to both systematic and nonsystematic error. Systematic error is of special concern because increasing sample size does not reduce it. The above analyses showed how to reduce several such sources of systematic error. However, because sample sizes are always limited, an equally important focus in sampling is to quantify the relationship between sample size and the variability of indicators. The question is, if samples of a given size were drawn not once, but many times, how would the results vary? More precisely, what would be the standard deviation across these samples, that is, the standard error? The standard error of the reciprocity-based population estimate can, therefore, be computed through a procedure somewhat like *bootstrapping* (Berkowitz and Diebold 1998), in which the sampling process is simulated. The simulations of the sampling process employ the following steps. First, a seed from which recruitment begins is arbitrarily selected. Second, the first recruit is chosen randomly based on the matrix of transition probabilities. The second recruit then chooses a recruit in the same manner. This process is continued until the number of recruits equals the sample size. The reciprocity-based popula-

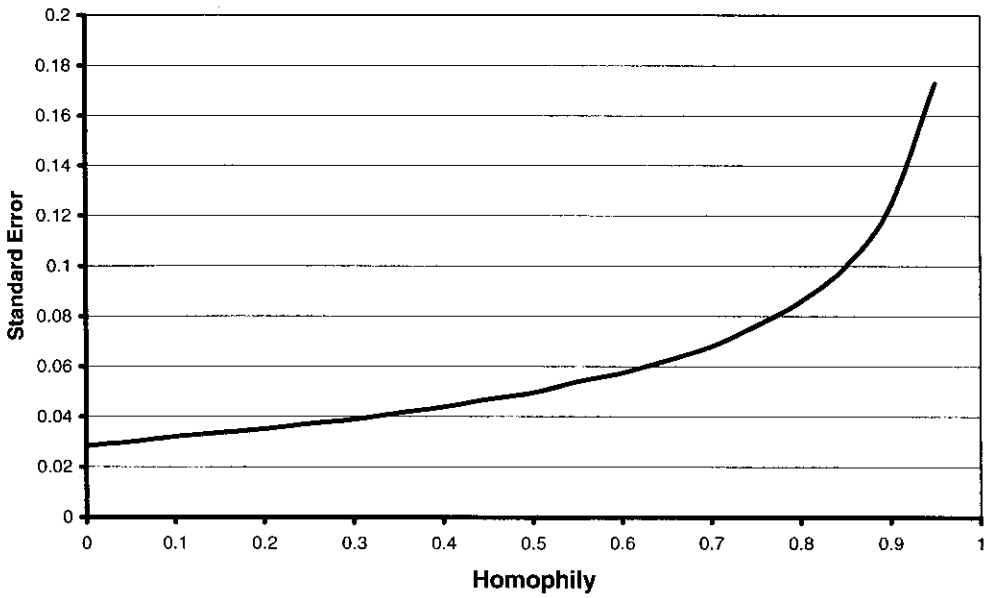

Figure 2 • *Homophily and Standard Error of the Population Estimate, P*

tion estimator, P, is then computed based on that simulated recruitment data. Ten thousand such simulations of the sampling process were conducted, and the standard deviation of these population estimates was then computed, to yield the estimate of that estimate's standard error.

Incidentally, the simulation process could have been made more realistic, by taking into account the number and composition of the seeds with which sampling began and that subjects could recruit more than one other subject. However, in a series of simulations, these factors had no significant effect on samples above a minimum size (i.e., about 20). This is expected because of theorem 2, which states that equilibrium is approached quickly. Therefore, the simulations employed the simplest structure—a linear recruitment chain that begins with an arbitrarily chosen seed and whose length corresponds to the sample size.

Consistent with usual expectations, the standard error of the population estimate depends on sample size. Less obviously, it also depends on homophily. As Figure 2 shows, error is an increasing function of homophily. This occurs because when homophily governs an act of recruitment, the recruiter does not reach outward and provide information about the groups composing the system. The greater the homophily, the less information is generated by each act of recruitment. In the limiting case of perfect homophily, recruitment generates no information about group composition. Subjects merely recruit others like themselves, so the sample's composition remains the same as the seeds from which it began, thereby revealing no information about the larger population.

Because of the association between homophily and standard error of the population estimate, RDS loses efficiency as homophily increases. That is, the greater the homophily, the larger the sample size must be to attain any given level of standard error. Furthermore, as is apparent from the sharply accelerating curve in Figure 2, this relationship is nonlinear. When homophily levels are moderate, a given increase has a smaller effect on standard error. For example, when compared with a baseline of zero homophily, a homophily of 58% doubles

the standard error, 79% triples it, 88% quadruples it, and 93% quintuples it. These exact figures depend on the recruitment matrix used for this analysis, but the pattern is characteristic of RDS: Standard error is an accelerating function of homophily. This relationship between homophily and standard error reflects variations in the independence of observations. The higher the homophily, the greater the dependence of each observation upon the previous observation (i.e., the dependence of the recruit's characteristics upon the recruiter's characteristics), and hence the less information is contained in each additional observation. An implication is that RDS has greater efficiency when sampling systems that have moderate to low homophily. In the study reported here, the observed homophily levels were moderate (maximum = 44%), producing slightly less than a doubling of standard error.

Conclusion

In this paper, the RDS method was expanded to include procedures for computing population estimates which compensate for bias resulting from differences in respondents' homophily, personal network size, and recruitment effort. In addition, means for computing the population estimate's standard error were introduced. The analyses show that nonprobability samples need not be dismissed as mere convenience samples. If the analysis of bias is detailed enough to permit the construction of a statistical theory, if the sampling process is redesigned based on that theory to reduce preventable biases, and if information required by the theory with which to quantitatively estimate and control for other biases is gathered during the sampling process, such as information regarding recruitment patterns and network sizes, valid inferences are possible. Though some of the procedures are computationally intensive, because they involve solving systems of simultaneous equations and simulations of the sampling process, they are well suited to computer implementation.[3]

The analyses in this paper focused on deriving simple population estimates, such as the proportional distribution of the population by race and ethnicity. These population estimates could be employed to weight the sample in further statistical analyses. This use of weights, it should be noted, is unusual. Normally, weights are designed into a study, as a means for increasing the representation of small groups or groups of special interest. In contrast, in RDS, the weights are computed after the data has been collected, based on such factors as differences in network size and homophily that determine which groups are under or over-sampled. For example, in a previous study of injection drug users (Heckathorn, et al. 1999), it was found that both HIV positive injectors, and those whose frequency of injection was especially high, tended to be over-sampled because they had especially large networks.

The formal modeling on which the RDS method is based points to additional sources of sampling bias that can be investigated in further studies, areas in which applications of the method can be expected to be most and least fruitful, and potential areas of further theoretic elaboration. As noted above, a prerequisite for the use of any chain-referral method is that the target population be linked by a contact pattern. That is, members must know one another *as members of the population*. A further limitation is that the sample should be small relative to the population. Respondents can be recruited only once, so recruitment in effect depletes the population. Hence, the sampling-with-replacement model implicit in the Markov and reciprocity models is a reasonable approximation for large and densely connected populations, but not for populations that are either small or sparsely connected, for each recruitment then significantly

3. Custom software for Windows 95/98 written by the author implements the analyses described in this paper, including the Markov, homophily and reciprocity models, and computation of standard errors. The program analyzes systems with up to eight groups, and includes sample data on injection drug users for four variables (race/ethnicity, gender, HIV status, and homelessness) from each of three sites (New London, Middletown, and Meridien, Connecticut). This software is available free from the author.

depletes the further recruitment opportunities for both the recruiter and those with whom the recruiter is connected. A subsequent paper will examine depletion effects and thereby extend the sampling method for use in smaller and sparser populations. A related limitation derives from the assumption that transition probabilities are stable. Ideally, sampling should proceed quickly enough for this assumption to be plausible, and if not, analyses comparing early and late recruits should test for instabilities in transition probabilities due to secular trends or other factors. Further analysis will also be required to determine whether data smoothing or a statistically based approach such as linear least squares is most appropriate, and if the latter is used, whether it should employ ordinary least squares, partial least squares, or some other method. Given the crucial role of self-reported network size for the reciprocity model, improved means for measuring network size and determining its level of reliability are also important, as are means for assessing the effects of within-group variations in network size. Finally, the standard error of the population estimates should be derived analytically rather than through simulation.

Because of its reliance on behavioral contact tracing, RDS represents a compromise between the large-scale community studies that once played a focal role in the discipline (Lynd and Lynd 1929), and the survey-based probability samples that are now dominant. The former analyzed individuals in the social structures in which they were embedded, including the groups and associations that served as the focal points for affiliation. In contrast, probability sampling pulls individuals out of these structures, and the information that is retrievable regarding relationships with others is limited to self-reports. Hence, these latter methods are limited to the information respondents possess regarding those with whom they interact.

The advantage of tracing network linkages behaviorally, as in RDS, derives not merely from concerns about the validity and reliability of self-reports. It also permits the investigation of affiliation patterns based on types of information that are not shared among respondents. For example, information regarding HIV status is frequently not shared among injectors, but RDS can be used to study affiliation patterns based on HIV status (see Heckathorn, et al. 1999, in which HIV status was determined for each respondent, and the social ties linking them were established behaviorally, through recruitment relationships). Such studies can investigate affiliation patterns of which the respondents themselves are unaware. Similarly, affiliation patterns can be studied regarding concepts, such as self-efficacy, that are not meaningful to respondents.

A final advantage of RDS is that it provides not merely a sample, but also data about the social structure in which respondents are embedded, where social structure is defined, consistent with Simmel (1955) and with Blau's (1977) macrostructural theory, in terms of patterns of affiliation. According to that definition, in an unstructured system affiliations are formed randomly, that is, homophily is zero. Structure emerges when affiliations are formed nonrandomly. When ties are based on similarity (e.g., friendships among persons similar in education or ethnicity), the result is homophily. When ties are formed based on complementarity or difference (e.g., exogamous marriage norms or heterosexual relationships), the result is heterophily. Thus, according to this conceptualization of social structure, homophily and heterophily are the elements out of which social structures are built. The concept of affiliation can be formalized through an extension of the homophily model, if homophily is conceived as affiliation to one's own group. The affiliation between any two groups, X and Y, A_{xy}, therefore, can be defined as follows:

$$A_{xy} = \begin{cases} \dfrac{P_y - S_{xy}}{P_y - 1} & \text{if } P_y < S_{xy} \\ \dfrac{S_{xy} - P_y}{P_y} & \text{if } P_y > S_{xy} \end{cases} \tag{19}$$

Affiliation is positive if $P_y < S_{xy}$, indicating that the proportion of ties to the group is greater than that group's proportional size. Affiliation is negative if $P_y > S_{xy}$, indicating that the pro-

Table 3 • *Affiliation by Race/Ethnicity and Gender*

A: Affiliation Index by Race and Ethnicity	*Target of Affiliation*			
Source of affiliation	*White*	*Hispanic*	*Black*	*Other*
Non-Hispanic white	0.430	−0.604	−0.163	0.008
Hispanic	−0.431	0.288	0.01	0.057
Non-Hispanic black	−0.27	−0.32	0.303	−1
Other	−0.048	0.176	−1	−1

B: Affiliation Index by Gender	*Target of Affiliation*	
Source of Affiliation	*Female*	*Male*
Female	0.018	−0.018
Male	−0.355	0.355

portion of ties to the group is less than that group's proportional size. By this definition, if a group forms ties only with another group, the former group's affiliation to the latter is perfect (i.e., if $S_{xy} = 1$, then $A_{xy} = 1$), if a group never forms ties with another group, affiliation is minimal (i.e., if $S_{xy} = 0$, $A_{xy} = -1$), and if a group forms ties only in proportion to the group's representation in the population, affiliation is zero (i.e., if $S_{xy} = P_y$, $A_{xy} = 0$).

Were affiliations formed randomly, irrespective of group membership, affiliation indexes would be zero. Alternatively, were affiliations formed in the manner assumed by the Fararo and Sunshine (1964), in which ties within the group reflect homophily, and ties with out-group members are formed in exact proportion to group size, levels of affiliation would be uniform across out-groups, and that level of out-group affiliation would equal the inverse of the group's homophily. Examination of affiliation data from the race/ethnicity analysis (see Table 3A) shows that this assumption does not hold. Instead, affiliation levels with out-groups are variable. For example, Hispanics are socially distant from whites ($A_{hw} = -.431$), but have a near neutral affiliation toward blacks ($A_{hb} = .01$). In contrast, blacks have rather uniformly negative affiliations toward Hispanics ($A_{bh} = -.32$) and whites ($A_{bw} = -.27$). Note also, that affiliations need not be the same in both directions (e.g., the affiliation from Hispanics to blacks is neutral, yet affiliation from blacks to Hispanics is negative). As a result, the social structure of race/ethnicity is rather complex. Only in the simplest case, where there are only two categories, does the affiliation with the out-group equal the inverse of homophily (see Table 3B, where the self-affiliation of each gender is the inverse of affiliation to the other gender).

By providing a measure for affiliation, RDS offers a means for studying social structure. When RDS is viewed exclusively as a sampling method, affiliation is taken into account only so its potential biasing effects can be quantified and controlled. In contrast, when RDS is used as a means to study social structure, the ability to measure affiliation and to thereby quantitatively specify social structure becomes the essential focus. A sampling method that must take into account the potentially biasing effects of social structure thereby becomes also a method for studying that structure.

Appendix: Extending Theorem 4 to Systems with More Than Two Groups

Consider first a three-category system. Let x, y, and z be the smoothed number of recruitments across categories A and B, A and C, and B and C, respectively. For example, As recruited x

Bs, and Bs also recruited x As. Similarly, let i, j, and k be the total number of recruits by members of A, B, and C, respectively. This suffices to specify the recruitment matrix. The number of As recruited by As is i − x − y, so S_{aa} = (i − x − y)/i. Similarly, S_{ab} = x/i, S_{ac} = y/i, and the rest of the recruitment matrix is derived in the same manner. Consistent with equation 1, the equilibrium sampling distribution can be derived by solving the following system of equations:

$$\begin{aligned} 1 &= E_a + E_b + E_c \\ E_a &= \frac{i-x-y}{i}E_a + \frac{x}{j}E_b + \frac{y}{k}E_c \\ E_b &= \frac{x}{i}E_a + \frac{j-x-z}{j}E_b + \frac{z}{k}E_c \end{aligned} \tag{a}$$

The solution to this system of equations is,

$$\begin{aligned} E_a &= \frac{i}{i+j+k} \\ E_b &= \frac{j}{i+j+k} \\ E_c &= \frac{k}{i+j+k} \end{aligned} \tag{b}$$

The reciprocity-model-based population estimate can be derived in a similar manner. Given that network sizes are assumed to be equal, let N_b and N_c equal N_a. This, plus the above specified recruitment matrix for reciprocity-compatible systems, suffices to specify the model. Consistent with equation 10, the population estimate can be derived by solving the following system of equations:

$$\begin{aligned} 1 &= P_a + P_b + P_c \\ P_a N_a \frac{x}{i} &= P_b N_a \frac{x}{j} \\ P_a N_a \frac{y}{i} &= P_c N_a \frac{x}{k} \end{aligned} \tag{c}$$

The solution to this system of equation is,

$$\begin{aligned} P_a &= \frac{i}{i+j+k} \\ P_b &= \frac{j}{i+j+k} \\ P_c &= \frac{k}{i+j+k} \end{aligned} \tag{d}$$

Note that this solution is identical to the equilibrium distribution. Therefore, if network sizes are equal, E = P. Furthermore, the structure of the solution suggests, and analyses confirm, that the conclusion extends to the general case. For example, in a system with M groups and when m is the total number of group M's recruits, E_a = i/(i + j + k + . . . + m), and P_a = i/(i + j + k + . . . + m), so the equilibrium and population distributions remain the same.

References

Berkowitz, Jeremy, and Francis X. Diebold
1998 "Bootstrapping multivariate spectra." *The Review of Economics and Statistics* 80:664–666.

Blau, Peter M.
1977 *Inequality and Heterogeneity.* New York: Free Press.
1994 *Structural Contexts of Opportunities.* Chicago: University of Chicago Press.
Broadhead, Robert S., and Douglas D. Heckathorn
1994 "AIDS prevention outreach among injection drug users: Agency problems and new approaches." *Social Problems* 41:473–495.
Brown, Lawrence D., Morris L. Eaton, David A. Freedman, Stephen P. Klein, Richard A. Olshen, Kenneth W. Wachter, Martin T. Wells, and Donald Ylvisaker
1999 "Statistical controversies in Census 2000." Technical Report 537, Department of Statistics, University of California, Berkeley.
Burt, Ronald S.
1986 "A note on sociometric order in the general social survey network data." *Social Networks* 8:146–174.
Eder, Donna
1985 "The cycle of popularity: Interpersonal relations among female adolescents." *Sociology of Education* 58:154–165.
Erickson, Bonnie H.
1979 "Some problems of inference from chain data." *Sociological Methodology* 10:276–302.
Fararo, Thomas J., and Morris H. Sunshine
1964 *A Study of a Biased Friendship Net.* New York: Syracuse University Press.
Farebrother, Richard William
1988 *Linear Least Squares Computations.* New York: Marcel Dekker.
Frank, Ove, and Tom Snijders
1994 "Estimating the size of hidden populations using snowball sampling." *Journal of Official Statistics* 10:53–67.
Goodman, Leo A.
1961 "Snowball sampling." *Annals of Mathematical Statistics* 32:148–170.
Heckathorn, Douglas D.
1990 "Collective sanctions and compliance norms: A formal theory of group-mediated social control." *American Sociological Review* 55, June:366–384.
1997 "Respondent driven sampling: A new approach to the study of hidden populations." *Social Problems* 44:174–199.
Heckathorn, Douglas D., Robert S. Broadhead, Denise L. Anthony, and David L. Weakliem
1999 "AIDS and social networks: Prevention through network mobilization." *Sociological Focus* 32:159–179.
Heckathorn, Douglas D., Robert S. Broadhead, and Boris Sergeyev
2001 "A methodology for reducing respondent duplication and impersonation in samples of hidden populations." *Journal of Drug Issues* 31:543–564.
Heckathorn, Douglas D., and Joan Jeffri
2001 "Finding the beat: Using respondent-driven sampling to study jazz musicians." *Poetics* 28:307–329.
Heckathorn, Douglas D., Salaam Semaan, Robert S. Broadhead, and James J. Hughes
2002 "Extensions of respondent-driven sampling: A new approach to the study of injection drug users aged 18–25." *AIDS and Behavior.*
Kalton, Graham
1983 *Introduction to Survey Sampling.* Newbury Park, CA: Sage.
Kemeny, John G., and J. Laurie Snell
1960 *Finite Markov Chains.* Princeton, NJ: Van Nostrand.
Killworth, Peter D., and H. Russell Bernard
1978–1979 "The reversal small world experiment." *Social Networks* 1:159–192.
Killworth, Peter D., Eugene C. Johnsen, H. Russell Bernard, Gene Ann Shelly, and Christopher McCarthy
1990 "Estimating the size of personal networks." *Social Networks* 12:289–312.
Klovdahl, Alden. S.
1989 "Urban social networks: Some methodological problems and possibilities." In *The Small World,* M. Kochen, ed. Norwood, NJ: Ablex: 176–210.

Laumann, Edward O., John H. Gagnon, Stuard Michaels, Robert T. Michael, and James S. Coleman
1989 "Monitoring the AIDS epidemic in the United States: A network approach." *Science* 244:1186–1189.

Lynd, Robert S., and Helen Merrell Lynd
1929 Middletown, A Study in Contemporary American Culture. Middletown, NY: Harcourt Brace and Co.

MacKellar, Duncan A., Linda Valleroy, John M. Karon, George F. Lemp, and Robert S. Janssen
1996 "The young men's survey: Methods for estimating HIV seroprevalence and risk factors among young men who have sex with men." *Public Health Reports* 111, Supplement: 138–144.

Marsden, Peter V.
1990 "Network data and measurement." *Annual Review of Sociology* 16:435–463.

McPherson, J. Miller, and Lynn Smith-Lovin
1987 "Homophily in voluntary organizations: Status distance and the composition of face-to-face groups." *American Sociological Review* 52:370–379.

Rapoport, Anatol
1979 "A probabilistic approach to networks." *Social Networks* 2:1–18.

Simmel, Georg
1955 *Conflict: The Web of Group Affiliations*, Kurt H. Wolff and Reinhard Bendix, trans. New York: Free Press.

Spreen, Marius
1992 "Rare populations, hidden populations, and link-tracing designs: What and why?" *Bulletin de Méthodologie Sociologique* 6:34–58.

Sudman, Seymour, and Graham Kalton
1986 "New developments in the sampling of special populations." *Annual Review of Sociology* 12:401–429.

Watters, John K. and Patrick Biernacki
1989 "Targeted sampling: Options for the study of hidden populations." *Social Problems* 36, 4:416–430.