Published in final edited form as:
Field methods. 2010 August 1; 22(3): 217–230. doi:10.1177/1525822X10370796.

Sampling to reduce respondent burden in personal network studies and its effect on estimates of structural measures

Daniela Golinelli, Ph.D., **Gery Ryan, Ph.D.**, **Harold D. Green Jr., Ph.D.**, **David P. Kennedy, Ph.D.**, **Joan S. Tucker, Ph.D.**, and **Suzanne L. Wenzel, Ph.D.**
RAND Corporation, 1776 Main Street, P.O. Box 2138, Santa Monica, CA, 90407, U.S.A.

Abstract

Recently, researchers have been increasingly interested in collecting personal network data. Collecting this type of data is particularly burdensome on the respondents, who need to elicit the names of alters, answer questions about each alter (network composition), and evaluate the strength of possible relationships among the named alters (network structure). In line with McCarty et al.'s (2007) research, we propose reducing respondent burden by randomly sampling a smaller set of alters from those originally elicited. Via simulation, we assess the estimation error we incur when measures of the network structure are computed on a random sample of alters and illustrate the trade-offs between reduction in respondent burden (measured with the amount of interview time saved) and total estimation error incurred. Researchers can use the provided trade-offs figure to make an informed decision regarding the number of alters to sample when in need to reduce respondent burden.

Keywords

personal network; structural measures; simulation; bias; variance and mean squared error

Introduction

Personal network studies are particularly burdensome for the respondents. In this paper, we investigate the effect of randomly sampling alters on the behavior of four structural measures as a way to reduce respondent burden

A personal network is the set of ties that an individual (ego/respondent) reports to exist among his or her network members (alters). It differs from a whole network, which represents the set of relationships among the members of a bounded population

The recent interest in personal networks stems from the belief that the characteristics of a person's social network explain the person's behaviors and or attitudes. In the study motivating this paper, the main aim was assessing how respondents' social network structural and compositional characteristics are associated with alcohol use and condom use in a sample of 445 homeless women (Wenzel 2005).

Collecting personal network data is challenging in that it places a heavy burden on the respondent. Personal network interviews are divided into three sections: questions designed to elicit the names of people in the respondent's social network (network generator), questions

Corresponding author: Daniela Golinelli, PhD, RAND Corporation, P.O. Box 2138, 1776 Main Street, Santa Monica, CA 90407, U.S.A., daniela@rand.org, phone: 310-393-0411 x6187, fax: 310-260-8152.

about each alter (network composition), and questions about the relationships between each unique pair of alters (network structure). The alter elicitation phase is usually not burdensome. Like others (McCarty et al. 2007), we have found that respondents can name between 20 and 45 people with little effort. On the other hand both the network composition and structure phases can be lengthy. Concerning network composition, if, for example, we ask the respondent 15 questions about every alter, then the composition section will contain 300 questions for a network of 20 alters. The network structure phase requires the respondent to report on whether each unique pair of named alters interacts and how often. This process creates an adjacency matrix with as many rows and columns as the number of alters elicited by the respondent. The cells in the matrix record the relationship between any two alters. The number of entries grows quadratically with the number of alters. In a network of 10 alters, the number of unique pairs of alters is 45; with 20 alters, the number of unique pairs rises to 190. Clearly respondent burden for network composition and structure can be heavy even with a modest number of alters.

McCarty et al. (2007) studied several methods for reducing respondent burden in personal network studies and argued that the best way to do so is randomly sampling a smaller set of alters. That is, respondents elicit N alters, but assess the ties for only n alters, where $n<N$.

We build on McCarty et al.'s (2007) work. In particular we describe the estimation error, in terms of bias and variance, incurred in computing the structural measures with a sample of alters (we call them *sample structural measures*). The estimation error is defined as the difference between the sample measure and the measure computed on the full set of alters N. This last measure is our target; since with the sample measures the best we can do is matching the measures computed on all N alters. This does not mean that the structural measures computed on N alters are correct, but dealing with error due to the respondent's recall bias, for example, is outside the scope of this study. We then assess the overall estimation error using both mean squared error (MSE) and root mean squared error (RMSE).

We also compute the range of RMSE for a sample of 28 networks and show how RMSE varies as function of the number of sampled alters and interview time saved.

Background

Description of the studied structural measures

In this paper we focus on four structural measures: *density*, *percentage of isolates*, *maximum degree* and *degree centralization* (Wasserman and Faust 1994). These measures are among the most commonly used in the social networks and risk behavior literature (see for example: Latkin et al 1995, Freeman 1979, and Kameda, Ohtsubo & Takezawa 1997).

Structural measures are computed from the adjacency matrix elicited during the structure phase of the interview; which for personal networks is symmetric: if the respondent reports that alter i interacts with alter j, then alter j knows alter i. We consider the case in which the adjacency matrix is dichotomous.

Let X be the $N×N$ adjacency matrix with cell (i,j), denoted by x_{ij}, equal to 1 if alter i and j interact and 0 otherwise. An alter's *degree* is the number of relationships that an alter has with the other alters in the network. The degree for alter i is given by the sum of the i-th row of X.

The degree for each alter ranges from 0, when an alter has no relationships (such an alter is called an “isolate”), to N-1, which represents the theoretical maximum degree. *Density* is a measure that varies between 0 and 1 and represents the proportion of ties that are present in a network relative to the total number of possible ties (N*(N-1)). It measures how connected a network is and is equal to the sum of all entries in X divided by (N*(N-1)).

The *percentage of isolates* is the number of alters with degree zero divided by *N*. The *maximum degree* is the largest observed degree in the network. We divide this measure by the theoretical maximum degree, *N*-1, to make it invariant to network size and call it "percent maximum degree."

Degree centralization varies between 0 and 1, and measures the degree to which the ties in the network are concentrated on few alters. Networks with centralization equal to 0 have ties evenly distributed across alters, while networks with centralization equal to 1 have a single individual with connections to all others.

Methods

This paper has two main goals. First, we want to quantify the estimation error, in terms of bias and variance, for structural measures calculated using a subsample of alters. Second, we are interested in using that information to assess whether random sampling can be an effective strategy for reducing respondent burden.

We address these aims with simulation studies. For the first aim we use a synthetic personal network of 20 alters and describe the bias and variance of the sample structural measures for samples of different sizes. To generate this personal network we assumed that the probability of a tie between any two alters is equal to 0.1 and that the presence of a tie between two given alters is independent of the presence of any other tie.

The simulation consisted of the following steps:

1. We consider 15 sample sizes ranging from 5 to 19 (since the considered networks have *N*=20 alters). Start setting *n* = 5.
2. Given *n*, randomly draw 10,000 samples without replacement of *n* alters from the given network (sampling *n* alters from a network means sub-setting the adjacency matrix to only those rows and columns corresponding to the sampled alters. Only the ties among the sampled alters are observed, since the respondent would be asked to report only about the sampled alters).
3. For every sample compute the four structural measures.
4. Use the 10,000 computed measures to assess MSE, bias, and variance. Increase *n* by one and repeat steps 2-4. The simulation stops after step 4 when *n*=19.

The MSE is the average squared difference between the sample structural measure and the "true/target" structural measure computed on the *N*=20 alters network. The bias is the difference between the average sample structural measure and the structural measure computed on the complete 20-alter network. The variance is the variance of the 10,000 sample structural measures. MSE equals the square of the bias plus the variance. We also compute RMSE since it is on the same scale as the structural measure.

For the second aim, we use 28 real personal networks of size 20. During the pilot phase of the project motivating this study, 28 homeless women were interviewed and asked to name 20 people with whom they had contact in the previous 12 months and to provide information about the relationships among them (Ryan et al. 2009). We repeat 28 times the simulation steps described above to provide a range for the total estimation error of the sample structural measures using RMSE.

Results

Behavior of the sample structural measures

The first row of Table 1 shows the value of the four structural measures computed on the full set of alters (*N*=20), the "true value". The two assumptions we made control the density value, which is close to the probability of a tie (0.1). The independence assumption enforces a "structure" in the network such that the percentage of isolates and maximum degree are inversely related to the density and the centralization is relatively low since the assumption implies a fairly uniform distribution of the degrees.

The other rows of Table 1 report the sample measures bias, variance and MSE for the 15 considered sample sizes. For example, the row corresponding to *n*=5 reports the bias of the sample density computed using a random sample of 5 alters and the column next to it reports the variance of the 10,000 sample densities. The third column reports the MSE, which in this case is equal to the variance since the sample density is unbiased. The remaining entries in the row report bias, variance and MSE for the sample percentage of isolates, sample percent maximum degree and sample degree centralization.

Unlike sample density, the sample percentage of isolates is a biased estimate; since it overestimates the true percentage of isolates. The bias is inversely related to the number of sampled alters and it decreases relatively quickly as the sample size increases. When we take a sample of alters, we are likely to "break" ties between the sampled alters and those not sampled; therefore, the sample percentage of isolates is more likely to be higher than the true percentage of isolates.

The sample maximum degree is also biased, though the bias, at least for this network, appears to be negligible. The sample centralization behaves similarly to the sample maximum degree, even though the bias is more substantial. Table 1 shows that variance and MSE for the four sample measures decline quickly as *n* increases and that the rate of decline plateaus for larger values of *n*. For the density, the zero bias implies that the MSE coincides with the variance. On the contrary for the percentage of isolates, most of the MSE is due to the bias at least for smaller values of *n*. For maximum degree and degree centralization, the MSE is mostly driven by the variance, since as we observed before the bias for these two measures is small and declines almost to zero for larger values of *n*.

This simulation study verifies what we know in theory: the sample density is an unbiased estimate of the true network density; whereas the other three sample measures are biased with the sample percentage of isolates most biased. We have also shown that the MSE declines relatively quickly as the sample size increases and that this decline plateaus for larger sample sizes.

We note that the MSE exhibits the same behavior when the size *N* of the true network is different. That is the rate of decline of the MSE slows down when at least 50 to 60% of the alters are sampled, regardless of network size.

Overall estimation error as function of time savings

The previous section gave us a better understanding of the sample structural measures behavior. However the analysis of only one network does not tell us to which degree the total estimation error depends on the underlying network and on the specific values of the structural measures.

In this section we use real data and make no assumption on the network structure. Having a sample of 28 networks provides a wide range of values and combination of values for the considered structural measures.

Table 2 reports minimum, mean and maximum values of the four structural measures for the 28 networks and indicates that these 28 networks are different in terms of the four structural measures.

We repeated the same simulation study described above on the 28 personal networks computing the RMSE for the four measures for each considered sample size *n*. We then computed the minimum, mean and maximum RMSE across the 28 networks.

Figure 1 shows the minimum, mean and maximum RMSE for the four structural measures as function of *n* (with *n* ranging from 20 to 5) and the time saved (measured in minutes) when eliciting an adjacency matrix of size *n*.

To evaluate the average time needed to elicit the adjacency matrix, we used the average amount of time it took the 445 homeless women interviewed for the study motivating this paper. Women were asked to name 20 alters. To reduce respondent burden, women were asked to elicit the adjacency matrix on 12 randomly selected alters. On average women took 5 to 6 minutes to report on 66 ties. This means that to elicit the adjacency matrix for the full set of 20 alters women would have taken approximately 16 minutes. If we measure respondent burden with the time it takes the respondent to elicit the adjacency matrix, Figure 1 provides guidance on how to make an informed decision in selecting the number of alters to sample when respondent burden is an issue. Figure 1 quantifies the likely amount of error that a researcher would incur in computing the structural measures with a sample of alters in relation to the amount of interview time saved per respondent. Studies that need to reduce respondent-burden, for feasibility or budget reasons, are faced with the decision of trading-off accuracy of the network measures for time/cost savings. For example, from Figure 1 we see that eliciting an adjacency matrix with 12 alters (vertical line in Figure 1) instead of 20 would mean a saving of slightly more than 10 minutes per respondent. However, this saving comes at a cost: the sample measures on average will be off by 10%.

This figure also shows that if we want to save a little more than one minute (corresponding to sampling 19 alters) we would incur little loss of accuracy. However, the rate of increase in RMSE seems low when saving up to 10 minutes (corresponding to sampling 12 alters) but increases sharply for larger time savings. The general behavior of RMSE is similar across the four measures; however, it is lowest for density and largest for percentage of isolates. RMSE also shows the sharpest increase for the percent of isolates as the time saving increases or the number of sampled alters decreases.

Conclusions and Discussion

Personal network studies are particularly burdensome for the respondents. In this paper, we investigated the effect of randomly sampling alters to reduce respondent burden on the behavior of four structural measures. We also assessed the range of the total amount of error we incur when computing these measures using a sample of alters and showed how this error varies as function of the number of alters sampled and the amount of time saved.

We provide researchers with a figure illustrating the amount of error they should expect to incur when sampling alters and the amount of time saved. This figure provides guidance for making an informed decision on the number of alters to sample when in need to reduce respondent burden. The only limitation to this figure is that the likely total amount of error was derived using a sample of 28 networks of homeless women. This sample might not be representative of the networks of other populations, though these 28 networks show a wide range of values for the four considered structural measures.

We think that sampling alters represents an effective way of reducing respondent burden. While the focus of this paper was on structural measures, it should be noted that sampling a smaller set of alters can also reduce the respondent burden for the network composition phase of the interview. The time savings, in terms of a shorter interview, can actually be more substantial for the composition phase than the structure phase. However, since most (if not all) of the composition measures are either means or percentages, such as the alters' mean age or the percentage of family members in the network, their statistical properties are well known. We know that composition variables computed on a sample are unbiased; hence the only source of error is their variability. In the study that motivated this paper we reduced respondent burden for both composition and structure sections. Only a small set of the composition questions (4 questions) was asked of all 20 alters, while the bulk of these questions (14 questions) was asked to the 12 sampled alters. Respondents on average took 5 seconds per composition question; this means that the second part of the composition phase took on average 14 minutes instead of 23.3 minutes. So the total (composition and structure phases combined) time saving per respondent obtained by sampling 12 of the 20 named alters amounts to about 20 minutes: a substantial reduction of the overall interview time.

In that study we went a step further in that we took a stratified sample of 12 alters. More specifically, the 20 named alters were grouped into two strata: sex partners and non-sex partners. Since one of the major goal of the study was analyzing the relationship between risky sexual behaviors (such as unprotected sex) and social network characteristics, it was important to include in the sample of 12 alters some of the sex partners the woman named.

The use of stratified sampling is one solution to the situation in which researchers need to collect information on specific types of alters and find themselves in the need of reducing respondent burden. The only drawback is that if alters from different strata are sampled at different rates, the sample measures need to be weighted (Lehtonen and Pahkinen 1994).

A stratified sample represents a potential solution to the disadvantage of randomly sampling alters noted by McCarty et al. (2007): since the selection of the alters is random, it is likely that key alters are not sampled and therefore that the sample measures might be significantly different from the true structural measures. In a study like ours with 445 cases, the case in which for few respondents the sample structural measures estimate poorly the true structural measures might have very little consequence on the analysis results. It is unlikely that, for example, the correlation between drug use and network density is affected if the network density is poorly estimated for a few respondents.

While this paper focuses on measuring the quality of estimates for personal network structural measures that is rarely the end product of the analysis. More commonly these network features act as explanatory variables in regression models. Including covariates measured with error, as would be the case with sample structural measures, results in their regression coefficients being biased toward 0. Methods for adjusting for this “attenuation bias” (Frost and Thompson, 2000) utilize the estimated variance of the covariate to project what the coefficient would be if the true structural measure had been used. Explanatory variables measured with bias are more problematic. Future work should assess the impact of estimation error in the sample structural measures on estimated regression coefficients and develop methods for adjusting regression coefficients for bias and variance in the estimated structural measures.

This paper showed that if researchers are interested in measuring network density, respondent burden can be reduced substantially since the overall estimation error is low even for relatively small samples. For the other three sample measures the overall estimation error tends to be higher. Therefore researchers need to assess the amount of error that they are willing to tolerate in order to reduce respondent burden. This is particularly true for the percentage of isolates.

The fact that three of the four considered sample structural measures are biased suggests that sample measures might not be the best estimators of the true structural measures. Future research should investigate alternative estimators that eliminate or reduce the bias and that ultimately exhibit a smaller overall estimation error. Though using the sample structural measures is advantageous since they require hardly any computation time; therefore the use of alternative estimators is warranted only if the reduction in the estimation error outweighs the increase in computing time and complexity.

References

Freeman LC. Centrality in networks: Conceptual clarification. Social Networks 1979;1:215–239.

Frost C, Thompson S. Correcting for regression dilution bias: comparison of methods for a single predictor variable. Journal of the Royal Statistical Society Series A 2000;163:173–190.

Kameda T, Ohtsubo Y, Takezawa M. Centrality in sociocognitive networks and social influence: an illustration in a group decision-making context. Journal of Personality and Social Psychology 1997;73 (2):296–309.

Latkin CA, Mandell W, Vlahov D, Knowlton AR, Oziemkowska M, Celentano DD. Personal network characteristics as antecedents to needle-sharing and shooting gallery attendance. Social Networks 1995;17(3-4):219–228.

Lehtonen, R.; Pahkinen, EJ. Practical methods for design and analysis of complex surveys. Wiley; 1994.

McCarty C, Killworth PD, Rennell J. Impact of methods for reducing respondent burden on personal network structural measures. Social Networks 2007;29(2):300–315.

Ryan GW, Stern SA, Hilton L, Tucker JS, Kennedy DP, Golinelli D, Wenzel SL. When, where, why and with whom homeless women engage in risky sexual behaviors: A framework for understanding complex and varied decision-making processes. Sex Roles. 2009 In press.

Wasserman, S.; Faust, K. Social network analysis: methods and applications. Cambridge: University Press; 1994.

Wenzel, SL. Alcohol use and HIV risk among impoverished women (R01 AA015301). Rockville, MD: National Institute on Alcohol Abuse and Alcoholism; 2005.

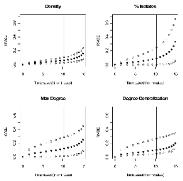

Figure 1.

Minimum (open circles), mean (filled circles) and maximum (open circles) RMSE as function of the time saved. Vertical line represents *n*=12. *n*=20 corresponds to 0 time saved and *n*=5 corresponds to a saving of 15.8 minutes.

Table 1

Bias, variance and MSE of the sample structural measures for each considered sample size *n*

	Density			**Percent Isolates**			**Maximum degree**			**Degree centralization**		
True value	**.110**			**.100**			**.263**			**.164**		
Sample size (*n*)	**Bias**	**Var**	**MSE**	**Bias**	**Var**	**MSE**	**Bias**	**Var**	**MSE**	**Bias**	**Var**	**MSE**
5	.000	.009	.009	.502	.085	.337	-.013	.037	.037	.058	.034	.038
6	.000	.006	.006	.435	.069	.259	.002	.025	.025	.062	.022	.026
7	.000	.004	.004	.370	.056	.193	.009	.017	.017	.056	.015	.018
8	.000	.003	.003	.307	.045	.140	.013	.012	.012	.050	.011	.013
9	.000	.002	.002	.259	.035	.102	.013	.009	.009	.043	.008	.010
10	.000	.002	.002	.211	.027	.071	.011	.007	.007	.033	.007	.008
11	.000	.001	.001	.173	.021	.051	.006	.005	.005	.024	.005	.006
12	.000	.001	.001	.141	.016	.036	.004	.005	.005	.017	.004	.005
13	.000	.001	.001	.112	.011	.024	.001	.004	.004	.012	.004	.004
14	.000	.001	.001	.084	.009	.016	-.002	.003	.003	.006	.003	.003
15	.000	.000	.000	.064	.006	.010	-.004	.003	.003	.002	.003	.003
16	.000	.000	.000	.045	.004	.006	-.005	.002	.002	-.001	.003	.003
17	.000	.000	.000	.031	.003	.004	-.006	.002	.002	-.003	.002	.002
18	.000	.000	.000	.018	.002	.002	-.004	.001	.001	-.003	.001	.001
19	.000	.000	.000	.008	.001	.001	-.001	.001	.001	-.001	.001	.001

Table 2

Minimum, mean and maximum values of the four structural measures for the 28 networks

	Density	Percent of isolates	Percent maximum degree	Degree centralization
Minimum	0.047	0.000	0.158	0.070
Mean	0.324	0.087	0.603	0.310
Maximum	0.868	0.650	1.000	0.643