A Note on Missing Network Data In the General Social Survey

Fall, 1985

Ronald S. Burt

Department of Sociology and Center for the Social Sciences
Columbia University
New York, NY 10027

GSS **Technical Report No.** 64

AUTHOR NOTE -- This technical note is a by-product of support from the National Science Foundation Sociology Program (SES-8208203) and the Measurement Methods and Data Improvement Program (SES-8513327) and has been produced as part of the Research Program in Structural Analysis housed at Columbia University's Center for the Social Sciences.

A Note on Missing Network Data in the General Social Survey

There is very little network data missing on respondents to the 1985 Ceneral Social Survey. Missing data on relations between discussion partners pose the greatest problem but the unknown relations are strongly associated with relations known to be weak. The association between missing and weak relations remains strong after controls for the number, strength, and nature of a respondent's discussion relations. Moreover, the association remains strong across different kinds of respondents despite significant tendencies for certain kinds of respondents to have provided incomplete network data. The implication is that the missing network data can be replaced with quantitative data indicating a weak relation.

Missing data are doubly a curse to survey network analysis. First, network items are more complex than the usual opinion survey item and so might seem more likely to generate missing data. Second, network analysis is especially sensitive to missing data. Models of network structures assume that complete information is available on the relations to be described. Even low aggregate proportions of missing data can seriously truncate the number of respondents available for analysis. For example, if two or three relations in every respondent's network are unknown, then network models cannot be applied routinely to the data without making assumptions about the meaning of the missing data.

It is accordingly surprising to observe that so little is known about missing data in survey network analysis. Even the two pace setting studies in this genre have little to offer. Some analysis of tendencies for respondents not to know the attributes of their friends in the 1966 Detroit Area Study is provided in Laumann's (1973:30-31) book on the study, but no results are provided on failures to answer the items eliciting the names of, and relations between, friends. Similarly, Fischer's (1982) book on the 1977 Northern California Communities Study provides no discussion of incomplete network data although a useful appendix is provided on the construction of survey network items.1

1Fischer does describe an index of respondent "cooperativeness" assembled from interviewer post-interview impressions of the respondent's attitude toward the interview (Fischer, 1982:302-303, 340-341). Respondent variation in cooperativeness was held constant aa a confounding factor in much of Fischer's analysis. For example, respondents who were judged by interviewers on three multi-category scales to have been more cooperative during the interview named more contacts than respondents judged to be uncooperative (Fischer, 1982:38,302). For my purposes in this note on missing data, too many respondent and interviewer factors are combined in Fischer's cooperativeness index to be able to say with any certainty how the index reflects respondent tendencies to provide incomplete network data.

Table 1

Frequency of Missing Data

	Number of Respondents	Number of Discussion Partners
No Answer to Name Generator	3	0
No Answer to Closeness with Individual Discussion Partners	3	11
No Answer to Closeness between One or More Pairs of Discussion Partners	60	184
Complete Data on Relations with and between Discusion Partners (includes 136 isolates)	1468	4288
Total	1534	4483

NOTE -- The first column gives the number of networks incomplete on each of the three network items. The second column gives the number of discussion partners on whom network data are incomplete. The 11 discussion partners cited by the 3 respondents failing to answer the second network item are missing data on the relation with the respondent citing them. Of the 251 discussion partners cited by the 60 respondents failing to answer one or more subparts of the third network item, 184 are missing data on relations with one or more other of their respondent's discussion partners.

The 1985 General Social Survey (GSS) provides a unique opportunity to study missing data in survey network analysis because the data represent the diversity of discussion relations likely to be encountered in survey network studies of the American population. Each respondent was asked to describe relations with and among up to five important discussion partners.2 My purpose in this brief note is to indicate the magnitude of the missing network data problem, the kinds of networks in which data are likely to be missing, and the kinds of people likely to provide incomplete network data.

THE MAGNITUDE OF THE MISSING DATA PROBLEM

The first three of the GSS network items defined the structure of a respondent's network and each poses a different class of problems for data analysis.3 The frequency with which respondents failed to answer each of the first three items is given in table 1.

Little can be done to salvage interviews in which the respondent failed to answer the initial name generator. Each respondent was asked: "Looking back over the last six months, who are the people with whom you discussed matters important to you?" Fortunately, only 3 respondents were lost at this point.

Next, respondents were asked to distinguish between especially close and less close discussion partners beginning with the filter item; "Do you feel equally close to all of these people?" Here again, only 3 respondents failed to answer the question and each went on to provide complete network data on relations between their discussion partners. Respondent 283 cited five discussion partners and said that all five were especially close to one another. Similarly, respondent 990 cited two people who were especially close to one another. Given equally strong relations among the discussion partners, each might be assumed to have equally strong relations with the respondent citing them (a "yes" answer to the filter question). Less obviously, the third respondent failing to answer this item (respondent 721) cited four people, all of whom were acquainted and two of whom were especially close. Given the mixed network structures and so few data lost at this point -- 3 observations

2Burt (1984) provides a detailed discussion of the data and various issues taken into account by the GSS Board of Overseers in their deliberations over the network items.

Subsequent questions elicited data on the nature of the respondent's relation with each discussion partner and various attributes of each discussion partner.

lost if respondents are the units of analysis and 11 observations lost if discussion relations are the units of analysis -- it seems best not to make any assumptions about the missing data on relations with discussion partners.

The final item defined the strength of relations among a respondent's discussion partners. These network data were elicited by two questions about each pair of discussion partners; "Are and total strangers?", and if not; "Are they especially close?" For the respondent naming five discussion partners, this amounted to twenty questions since his five citations created ten unique pairs of discussion partners. At some point in answering these questions, 60 respondents indicated that they could not describe relation(s) between one or more of their discussion partners. These missing data affect 184 of the discussion partners cited in the survey. Fortunately, every one of the 60 respondents missing data on relations between discussion partners provided complete data on his or her relations with each discussion partner. Further, most (54 of 60) provided partial data on relations between their discussion partners and the majority (38 of 60) provided data on more than half the relations between their discussion partners.

In sum, missing data are not a serious problem with the GSS network items. Only 66 of the 1,534 respondents provided incomplete network data and only 195 of the 4,483 discussion partners cited by the respondents are involved in missing data -- making 96% of the networks complete. Moreover, most of the incomplete networks are only missing data on some of the relations between discussion partners. Thus, data on the observed relations in a network can be used to study the nature of the relations missing.

NETWORK CONDITIONS PRONE TO MISSING DATA

The empirical question is whether or not there are certain network conditions that seem to have elicited missing data. This can be studied by comparing two kinds of discussion partners; 3,605 on whom network data are complete and 155 on whom partial network data are available.4

4These discussion partners are drawn from networks containing three or more discussion partners and complete or partially missing data. The 184 discussion partners in table 1 involved in missing relations divide into 155 on whom partial data are available and 29 whose relations with other discussion partners are completely missing. The latter are deleted here because they provide no information on relations between discussion partners. Also deleted are the 228 networks containing one discussion partner and the 136 containing none because they too contain no information on relations between discussion partners. Similarly, the 235 networks composed of two discussion partners are deleted because they contain no information on relations between discussion partners if one relation is missing.

There is an association between missing and weak relations apparent from the tendency for the incomplete discussion partners to have been more involved in weak relations than the partners on whom data are complete. Among the discussion partners on whom data are complete, 35 out of 100 were not especially close to any one cited by the same respondent. Among those involved in a relation which the respondent could not describe, however, the percentage increases to 59 out of 100. The hypothesis that missing relations are independent of weak relations is strongly rejected with these data. In a two-way tabulation of discussion partners by involvement in a missing relation (yes, no) and especially close to anyone cited by the same respondent (no, yes), the z-score loglinear interaction effect indicates a strong tendency for data to be complete on anyone especially close to anyone else (5.91, $p < .001$) and, of course, independence is improbable (35.74 chi-square statistic with 1 degree of freedom, $p < .001$).[5]

This association between missing and weak relations is disaggregated across categories of network size in figure 1. Missing and weak relations remain strongly associated. The z-score loglinear tendency for data to be complete on anyone especially close to anyone else is very strong (4.42, $p < .001$) and the hypothesis that missing and weak relations are independent across categories of network size is obviously improbable (46.75 chi-square with 3 degrees of freedom, $p < .001$). Nevertheless, there is a modest interaction with network size. The interaction is illustrated by the relative magnitudes of the bars in figure 1. The contrast between complete and incomplete discussion partners is greatest in large networks.[6] Ceteris paribus, the more people in a network, the more likely by random chance that any two would be viewed as especially close. Notice in figure 1 that discussion partners in networks of five are most the likely to have been especially close

[5] Likelihood ratio chi-square statistics are presented throughout the discussion. It is worth noting that routine statistical inference is imprecise here because the respondent to discussion partner dyads are not independent observations. Dyads elicited from different respondents are independent, but the three to five elicited from a single respondent are not independent. The more interdependent the discussion partners named by a respondent, the higher the intraclass correlation within respondent networks, and the more that routine test statistics computed from dyads exaggerate statistical significance. In the absence of any systematic correction for correlation between dyads within respondent networks, I report routine statistical tests and rely on the relative magnitude of test statistics. Routine statistical significance in this case is an upper limit on the actual significance of effects.

[6] The chi-square statistic for the hypothesis that the interaction between missing and weak relations is independent of network size is faint but noticeable; 5.42 with 2 degrees of freedom (p - .07).

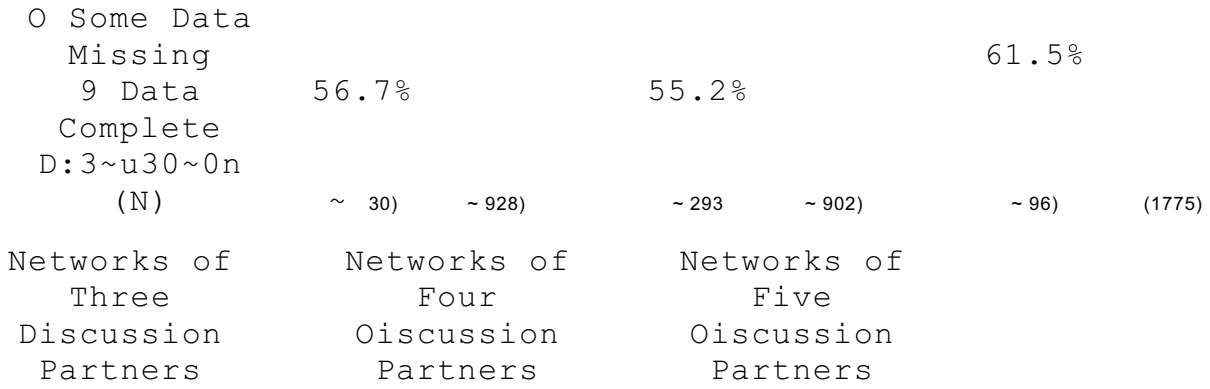

Figure 1

- Missing Relations Are Associated with Weak Relations

to one or more other persons in the network (70% versus 61% in networks of four and 55% in networks of three). All the more striking therefore, that the discussion partners involved in missing data are least likely in large networks to have been especially close to anyone. The strong overall interaction between missing and weak relations increases significantly in networks containing five discussion partners (2.32 z-score).

Given the association between missing and weak relations, is it equally true of all discussion partners or principally true for specific kinds of discussion partners? The results in table 2 are taken from tabulations of missing and weak relations across kinds of relationships between respondents and discussion partners.

Clearly, the strength of the relation between respondent and discussion partner had little effect on missing data. Judging from the first three rows of column one in table 2, there is no tendency for data to be complete on people who were especially close to the respondent, in daily contact with the respondent, or recently met by the respondent. Close and distant discussion partners were equally likely to have been involved in relations unfamiliar to the respondent. Further, missing relations are strongly associated with weak relations regardless of controls for the strength of relationship between respondent and discussion partner (second column in table 2).

Similar, but less homogeneous, results are obtained when role relations between respondent and discussion partner are compared. There is no general evidence of a direct association between role and missing data (first column in table 2). Something of an exception exits with co-members -- those discussion partners affiliated with a group containing the respondent.7 About one in five discussion partners were co-members of respondent affiliations (18.2% of all people named). There is no tendency for these people to have had strong or weak relations with other discussion partners, however, their relationships with one or more of the other discussion partners were likely to be unfamiliar to a respondent (2.31 z-score in table 2). Co-members notwithstanding, the dominant effect in each of the eleven tabulations by role relation is the association between missing relations and weak relations (second column of table 2). Moreover, there is no significant tendency for this

7The EhOw card given to the respondent for this item den~ned the co-member option aa "MEMBER OF A GROUP TO WHICH YOU BELONG -- for example, aomeone who attends your church, or who9e children attend the same School a5 your children, or belong9 to the same club, cla9smate."

Table 2
Missing Relations Across Kinds of Discussion Partners

	Z-score Loglinear Interaction with Missing Relations	Z-score Loglinear Interaction between Missing and Weak
Strength of Relationship		
Especially Close	- I.18	3.96
Daily Contact	-1.44	5.62
Known Less Than 3 Years	-1.80	3.49
Nature of Relationship		
Spouse	- 1.15	3.46
Mother or Father	0.40	1.96
Brother or Sister	0.71	3.43
Child	-0.30	3.82
Other Family	-0.52	2.17
Co-worker	-0.18	3.52
Co-member of Group	2.31	5.50
Nonkin Neighbor	1.97	4.01
Friend	1.63	4.55
Advisor	1.46	5.31

NOTE -- Results are taken from the three-way tabulation of discussion partners (see footnote 4 to the text) across involvement in a missing relation (yes, no), especially close to any other discussion partner (no, yes), and relation with respondent (yes, no; defined as indicated by row). Note that these test statistics define the upper limit of statistical significance (see footnote 5 to the text).

association to be contingent on the role relation between respondent and discussion partner.8

In sum, missing data were very much contingent on the structure of the network in which they occurred. More specifically, it appears that the relations on which data are missing are between discussion partners who were not especially close to one another. Network data tend to be complete on discussion partners who were especially close to one or more other partners. Data tend to be incomplete on discussion partners who were strangers or merely acquainted with the respondent's other discussion partners. The association between missing and weak relations remains strong after controls for the number, strength, and nature of a respondent's discussion relations.

RESPONDENTS PRONE TO MISSING DATA

The final question is whether or not the association with weak relations is contingent on the kind of respondent from whom network data were elicited. The results in table 3 are taken from tabulations of missing and weak relations across some basic distinctions among the survey respondents.9

8Chi-square statistics for the hypothesis that the association between missing and weak relations is independent of relation with respondent have 1 deRree of freedom and vary from 0.00 to 2.24 with a mean of 0.75 across twelve of the tabulations for which results are presented in table 2. The one significant three-way interaction occurs in the co-worker tabulation; a 4.03 chi-square (p - .05 under routine inference) reflects the increased tendency for co-workers to have been involved in a missing relation if they were especially close to any other discussion partner. With the number of tests being made and the fact that these test statistics define an upper limit for statistical significance (see footnote 5), it seems safe to conclude that the association between missing and weak relations is strong regardless of the relationship between respondent and discussion partner.

Some commonly discussed background attributes are not reported in table 3 because of a negligible association with missing network data; religion (Protestant, Catholic, Jewish, None), marital status (married, divorced or separated, never married), city (large central city, small city or suburb of large city, town or village or tiny city), and geographic mobility since age 16 (lives in same city, lives in same state but different city, lives in different state). Respondents from nine geographical regions were also compared. Beyond a strong tendency for New Englanders to provide incomplete network data, there are no significant regional differences. Respondent occupation was considered. Respondents were sorted into the ten broad U.S. Bureau of Census occupational categories. Given the small number of missing data available, categories were combined where respondents were similarly likely to provide incomplete network data; Professional (codes 0,1), Administrative and Clerical (codes 2,3), Craftsmen (codes 4,5), and Laborers (codes 6, 7, 9). Farmers and farm laborers were excluded from the tabulation. The one signifcant association with missing data is the tendency for incomplete network data from craftsmen, but it is contingent upon ties between discussion partners. The tendency for a craftsman's discussion partner to be involved in missing data (2.56 z-score for a .234 loglinear effect) disappears if the discussion partner was not especially close to anyone else named by the craftsman (-2.60 z-score for a -.238 loglinear effect). Therefore, in conjunction with the fact that the association between missing and weak relations is strong with respondent occupation held constant (4.56 z-score), occupation is not discussed in the text.

Table 3
Missing Relations Across Kinds of Respondents

	Z-score Loglinear Interaction with Missing Relations	Z-score Loglinear Interaction between Missing and Weak
Male	2.17	5.81
Black	3.09	3.39
Age		5.21
18-29 Years Old	-0.22	
30-37 Years Old	1.68	
38-59 Years Old	1.62	
60-87 Years Old	-2.18	
Education		2.86
Less Than High School	-2.40	
High School Graduate	1.72	
Some College	0.21	
College Graduate	2.57	

NOTE -- Results are taken from the three-way tabulation of discussion partners (see footnote 4 to the text) across involvement in a missing relation (yes, no), especially close to any other discussion partner (no, yes), and respondent attribute (defined as indicated by row). Note that these test statistics define the upper limit of statistical significance (see footnote 5 to the text).

It is clear that the respondents were differentially inclined toward incomplete network data. Discussion partners named by men are more likely to be incomplete than those named by women. Those named by Blacks are more likely to be incomplete than those named by others.10 Respondent age makes a difference, but only in the tendency for people over 60 to have provided complete network data.11 Respondent education is significant with college graduates tending to provide incomplete network data and poorly educated respondents tending to provide complete network data.12

Still, the association between missing and weak relations persists. Note the strong z-scores in the second column of table 3 despite the controls for respondent sex, race, age and education (not to mention the variables in footnote 9). Moreover, this strong association is independent of the respondent attributes. Chi-square statistics for the hypothesis of no three-way interaction effects are negligible; 0.18 and 0.47 with 1 degree of freedom for sex and race, 2.07 and 5.39 with 3 degrees of freedom for age and education.

Finally, the association between missing and weak relations is evident in the kind of respondents likely to have provided incomplete network data. This is nicely illustrated with education. Discussion partners named by poorly educated respondents tended to be complete while those named by respondents with a college education tended to be involved in missing relations. This positive association between education and missing data is just the opposite of what would be expected from research showing a tendency for education to decrease respondent use of "don't know" (e.g., Schuman and Presser, 1981:137-141). But it is exactly what would be expected if network data are missing because of weak ties in a network rather than respondent incompetence. In the networks of respondents with less than a high school education, 56.9% of the relations between discussion partners were especially

10Asiana were also more likely to provide incomplete data, but there are so few Asians in the GSS that it is difficult to make thia statement with confidence. Hispanics and whites more generally tended to report complete network data.

11This tendency ia moat pronounced among respondents past the retirement age of 65, however, there are ao few obaervationa on thia group that I have combined them with the next lower age atatua (60-65 yeara). The age groups in table 3 are bssed on an analysia of atructural equivalence in age a/ratification evident from the GSS network data (Burt, 1985, figure 4).

12Reapondenta were initially aorted into aeven education categoriea based on yeara of education and highest degree; primary achool, aome high achool, high achool graduate, aome college, Aasociate degree, Bachelor'a degree, graduate or profesaional achool. Given the amall number of misaing data available, categoriea have been combined where reapondenta were aimilarly likely to provide incomplete network data.

close.13 These respondents had fewer weak relations between their discussion partners and so had an easier task in completing the network items. Thus their tendency to provide complete network data. In the networks of college graduates, a much lower 31.3% of the relations between discussion partners were especially close (density differences between the levels of education in table 3 generate a 21.8 F statistic with 3 and 924 degrees of freedom, $p < .001$). In other words, college graduates had more weak relations between their discussion partners and so were likely to be asked in the network items to describe relationships with which they were unfamiliar. Thus their tendency to provide incomplete network data. Regressing network density over the respondent attributes in table 3 yields the following results:

Density = .53 - .OlMale + .06Black + .1301d - .06Education,

(0.4) (1.3) (4.5) (5 8)

where t-tests are given in parentheses, Old is a dichotomy between respondents over 60 and those under 60, and Education is created simply by numbering 1, 2, 3, and 4 the education categories in table 3. Education is the strongest determinant of density, but note too the tendency for older respondents to have named discussion partners especially close to one another and their tendency in table 3 to provide complete network data.

SUMMARY

There is very little network data missing on respondents to the 1985 General Social Survey. Missing data on relations between discussion partners pose the greatest problem but the unknown relations are strongly associated with relations known to be weak. The association between missing and weak relations remains strong after controls for the number, strength, and nature of a respondent's discussion relations. Moreover, the association remains strong across different kinds of respondents despite significant tendencies for certain kinds of respondents to have provided incomplete network data.

13Theae reaulta are based on the 928 reapondenta who named three or more diacussion partners and provided complete or partial data on relations between their diacussion partnera. Denaity ia measured by the following ratio for each reapondent: number of eapecially close relationa between diacusaion partnera divided by number of nonmisaing relationa between diacusaion partnera.

.

REFERENCES

Burt, Ronald S. (1984) "Network items on the General Social Survey," *Social Networks* 6:293-339.

Burt, Ronald S. (1985) "Kinds of relations in American discussion networks," Paper presented at the annual meetings of the American Sociological Association.

Fischer, Claude S. (1982) *To Dwell Among Friends.* Chicago: University of Chicago Press.

Laumann, Edward 0. (1973) *Bonds of Pluralism.* New York: Wiley Interscience.

Schuman, Howard, and Stanley Presser (1981) *Questions and Answers in Attitude Surveys.* New York: Academic Press.