The Reliability of Network Density and Composition Measures

Peter V. Marsden

Harvard University

May 12, 1992

GSS Methodological Report No. 72

A previous version of this paper was presented at the 1992 Sunbelt Social Network Conference, San Diego, CA. I am indebted to James A. Davis for helpful comments and to Nancy P. Williamson for computational assistance.

Abstract

Many measures based on egocentric network data, such as age composition or (local) network density, can be viewed as "aggregate" measures: they are mean values of the alter attributes or the dyadic attributes that fall within a given respondent's egocentric network. Internal consistency methods of classical test theory are not suitable for assessing the reliability of such measures: such methods presume a "crossed" design for data collection in which each respondent is scored on the same set of indicators. In designs for gathering egocentric network data, alters are instead "nested" within respondents; moreover the number of alters may differ across respondents. This paper evaluates the reliability of composition and density measures via analysis-of-variance approaches to reliability known as generalizability theory. Reliability estimates are presented for egocentric measures based on the 1985, 1987, and 1988 General Social Surveys and for the 1977-78 Northern California Community Study. Ethnoreligious composition, political composition, density, and composition of a network by "friends" or co-members of organizations are measured with relatively high reliability, even for a relatively small number of alters. Other measures require more alters to attain adequate reliability, and some, such as sex composition, remain problematic even when the number of alters grows quite large. The sensitivity of reliability estimates to differences in instrument design is examined using design variations in the surveys studied.

Many measures based on egocentric network data, such as age composition or (local) network density, can be viewed as "aggregate" measures: they are mean values of the alter or dyadic attributes that fall within a given respondent's egocentric network. Due to the design used to collect such data, the reliability of these measures cannot be evaluated with internal consistency methods based on classical test theory. This paper suggests that methods associated with generalizability theory (Shavelson and Webb, 1991; O'Brien, 1990) can be used for this purpose, and provides reliability estimates for measures from the 1985, 1987, and 1988 General Social Surveys (GSS) and from the 1978-79 Northern California Community Study (NCCS).

The results suggest that certain aspects of networks--ethnoreligious composition, political composition, density, and composition by persons bearing certain "role relations" to a respondent--can be measured with adequate reliability for a small number of alters per respondent. Other features require larger number of alters, and some, such as sex composition, may be problematic even with a relatively large number of alters. Several comparisons are presented to indicate the effects of variations in data collection designs on reliability; these suggest that core segments of networks are more reliably measured than extensive ones.

Egocentric Network Data

Also known as "personal" or "survey" network data, egocentric network data provide information on the nature of the local social environment surrounding an actor (usually, though not exclusively, a respondent to a social survey). The typical design for collecting such data treats the actor or respondent as an informant about his or her egocentric network. Instruments for gathering egocentric network data begin (see Burt, 1984) with one or more name generators; these place boundaries on an egocentric network by identifying a set of alters included in it. Delimitation of the network is followed by a series of name interpreter items that ask respondents for information about (1) the attributes of alters (e.g. age, race/ethnicity); (2) features of the relationships linking the respondent to the alters (e.g. intensity, duration); and (3) features of the relationships linking alters to one another. The data collection design may gather name interpreter data on all alters cited in response to the name generator, or on some subset of those alters (e.g. the first three or five alters named, or a sample of the alters named).

Composition and Density Measures

Most measures based on egocentric network data involve some aggregation of the responses to name interpreter items. In this paper I am concerned with measures of properties of a respondent's egocentric network that can be written as within-respondent means of name interpreter items. Composition measures reflect the central tendency of the distribution of an attribute across the alters or relationships included in an egocentric network; "age composition" is thus the mean age of alters, while "kin composition" reflects the extent to which respondents have kinship relations to alters. If x_{ik} is the i^{th} respondent's value on a name interpreter item for the k^{th} of n_i alters, a composition measure c_i for the i^{th} respondent's network is defined as follows:

$$c_i = n_i^{-1} \sum_{k=1}^{n_i} x_{ik} \, . \qquad (1)$$

It is not as well recognized that a familiar measure of network structure, local network density, takes a similar form. Density is conventionally defined (under the implicit assumption that relationships are either present or absent) as "the extent to which links which could possibly exist among persons do in fact exist" (Mitchell, 1969: p. 18). If a_i is the number of links that exist among the alters in the egocentric network of the i^{th} respondent, the local network density d_i can be written

$$d_i = n^{*}{}_i^{-1} a_i,$$

where n^{*}_i, the number of possible relationships among alters in respondent i's egocentric network, is defined as $n_i(n_i-1)/2$.

The number of links present, a_i, can be written as the sum of n^{*}_i dichotomous indicators x_{ijk} which tell whether or not a relationship is present for each distinct pair of alters j and k in respondent i's egocentric network. Rewriting a_i in this way illustrates the parallelism of the density measure to the composition measures defined above:

$$d_i = n^{*}{}_{i}^{-1} \sum_{j=2}^{n_i} \sum_{k=1}^{n_i-1} x_{ijk} \, . \qquad (2)$$

Moreover, when the density measure is written as in (2), it can be extended straightforwardly to the case in which indicator x_{ijk} is not dichotomous; a density measure is thus more generally understood as the mean tie strength linking pairs of alters in respondent i's egocentric network.

Sources of Unreliability

The reliability of a measure is defined as the extent to which it yields the same results on repeated trials (Carmines and Zeller, 1979: p. 11). Measures like (1) and (2) are subject to two distinct sources of variation across repetitions of a survey: (a) fluctuation in the specific alters or dyads that are included in a respondent's egocentric network on the basis of responses to a name generator, and (b) fluctuation in the scores x_{ik} or x_{ijk} obtained for a given alter or dyad on the basis of responses to a name interpreter. Response errors like (b) are typical for survey items; source (a) is particular to aggregate measures like those studied here.

One perspective on egocentric network data would minimize the importance of over-time fluctuation in alters, viewing the alters enumerated by a name generator as "the network." Those taking this view would, in the parlance of generalizability theory, treat alters as a "fixed" facet of measurement that does not vary across replications. The name generators used to elicit alters rarely ask respondents to explicitly select a subset of alters from some larger set that constitutes the "population" of alters in the egocentric network. Instead they request that respondents name, for example, "the three men who are your closest friends and whom you see most often" (Laumann, 1973: p. 264) or "the people with whom you discussed matters important to you [over the last six months]" (Burt, 1985: p. 119). If one regards alters as a fixed facet, the problem of unreliability for egocentric network measures dissolves, in large part; for a fixed set of alters, the sole source of unreliability in composition or density measures would be response variability in reports about the alters or dyads themselves--that is, only source (a) of unreliability.[1]

There are several reasons to be wary of treating alters as a fixed facet, despite the wording of many name generator items. First, in some data collection designs name interpreter data are gathered for only a subset of alters that is drawn from a respondent-specific population of alters enumerated by name generators. In the NCCS (Fischer, 1982a), each respondent's egocentric network was elicited by a set of nine name generators varying considerably in intensity, from spending leisure time and providing minor household aid to discussing personal problems and lending substantial sums of money. Many name interpreter items, including network density items, were asked only about a subset of the alters elicited (see below for details).

Second, most name generators of necessity ask respondents to recall the alters in their networks rather to recognize them from a prior enumeration. Sudman's (1985) results suggest that recall methods tend to underenumerate network membership, sometimes substantially.

Third, in practice there is variability across occasions in the alters elicited when an instrument for gathering egocentric network data is readministered to a sample of respondents after a short time interval. Several panel studies using network items demonstrate such short-term variability. Using a four-week interval and an "affective" name generator, Broese van Groenau, van Sonderen and Ormel (1990) find that 78% of the names obtained on the first occasion were also obtained on the second. Comparable figures for a "role relation" name generator and a composite 20-item "exchange" name

[1] Note, however, that unless there are multiple measures of a given property of a given alter or dyad, it is impossible to separate such sources of variation (ϵ_{ik} in equation (3) below) from alter effects ($\mu_{ik} - \mu_i$ in equation (3)). If alters are fixed, then, one is unable to assess the reliability of a density or composition measure.

generator are 88% and 74%, respectively.[2] For a four-week interval, Hoffmeyer-Zlotnik (1990) reports an overlap of 63% for a composite 8-item exchange name generator and 45% for the GSS "important matters" name generator. Bien, Marbach and Neyer (1991) report 76% overlap for a composite 11-item name generator and a three-month interval.

Other aspects of the data collection process may also result in variations in the alters included in a respondent's egocentric network. Many name generators ask for those persons with whom a respondent has had a given kind of contact within a specified time period; variations in the length of this period or respondent errors in recalling contacts that fall within it can lead to variations in the inclusion of alters. Indeed, all forms of error that affect responses to name generators may lead to short-term fluctuations in the alters elicited and subsequently described via name interpreters, and these in turn may result in unreliability in measures of composition and density.

Evaluating Reliability for Aggregate Measures

The density and composition measures of eqs. (1) and (2) are multiple-item measures. They differ, however, from familiar indices based on multiple indicators because the data used in constructing them are gathered by way of a nested rather than a crossed design. Bryk and Raudenbush (1992) call such designs as "hierarchical" because alters are studied only within respondents--a given alter or dyad is used to measure the network of a single respondent. This feature of the study design has important implications for the way in which the reliability of measures like (1) and (2) is to be assessed.

Consider, by way of contrast, the multiple-item measures of tie strength studied by Marsden and Campbell (1984) or Friedkin (1990). In these studies, the unit of analysis is the dyad; Marsden and Campbell examine a measure of tie strength constructed from dyadic indicators of intensity, frequency, duration, and confiding, while Friedkin constructs

[2] For the individual exchanges in the composite name generator, the percent overlap across occasions varies from 44% for giving (unpaid) help with household tasks to 77% for receiving aid in the care of one's home while away (Broese van Groenau et al., 1990: p. 128).

a Guttman scale of tie strength from items measuring frequency, help-seeking, and friendship. In each case, items are a facet of measurement and the data collection design "crosses" items with dyads: every dyad is scored on the same items. The number of items is necessarily the same for all dyads, and the scores assigned to dyads are comparable within items (e.g., the measurement of intensity for one dyad is comparable to the measurement of intensity for another dyad--but not comparable to measurements of frequency).

Measures like c_i and d_i for egocentric networks, however, are based on a design that "nests" alters or dyads within respondents. Different alters or dyads appear in the networks of different respondents. There is no one-to-one correspondence between the alters or dyads for one respondent and those for another; that is, there is no sense in which the age of alter 1 cited by respondent a is to be compared only with the age of alter 1 cited by respondent b, rather than with the ages of other alters cited by b. Density and composition measures may be based on different numbers of alters or dyads for different respondents; that is, n_i and n^*_i may (indeed should) differ across respondents to the degree that respondents' networks differ in size.

Methods for studying reliability based on classical test theory assume that data come from a crossed design. To estimate reliability for aggregate measures based on egocentric network data gathered via a nested design, I draw on generalizability theory (Shavelson and Webb, 1991; O'Brien, 1990), which suggests that the reliability of measurements gathered via a wide variety of designs can be evaluated using the analysis of variance. Generalizability theory allows one to examine the way in which the reliability of measurement for a given "object of measurement" (here respondents) is affected by any measurement facet; beyond items, facets of measurement could include interviewers, raters, or occasions of measurement. Here, alters or dyads are the facet of interest. Moreover, the approach is applicable to data from crossed, nested, or mixed designs.

With this approach, reliability (or "generalizability") coefficients are estimated using intraclass correlations based on variance components (Shrout and Fleiss, 1979; Mitchell, 1979). The analyses are based on the following linear model for a respondent's report about a single alter or dyadic characteristic:

$$X_{ik} = \mu + (\mu_i - \mu) + (\mu_{ik} - \mu_i) + \epsilon_{ik}, \qquad (3)$$

where X_{ik} is the report about the kth alter or dyad by the ith respondent, μ is the expected value of X_{ik} over all respondents and all alters/dyads, μ_i is the expected value of X_{ik} over all alters or dyads in the egocentric network of respondent i (the "universe score"), μ_{ik} is the expected value of X_{ik}, and ϵ_{ik} is a random error component capturing sources of unreliability in a respondent's report of an alter or dyadic characteristic for a specific alter or dyad--that is, source of unreliability (a) discussed above. The quantity $(\mu_i - \mu)$ represents the respondent effect on the measure, while $(\mu_{ik} - \mu_i)$ is the alter effect.

In principle, variance in X_{ik} can be separated into four components associated with respondents (σ^2(r)), alters (σ^2(a)), the interaction of respondents and alters (σ^2(ra)) and error (σ^2(e)). With a nested design, however, only one observation per alter is available, and it is not possible to separate the last three of these variance components; they are hence pooled into a term σ^2(a:r,e) for variance attributable to variability of alters within respondents and error.

Using a one-way analysis of variance (random effects model) with respondents as the factor, one can estimate the variance components σ^2(r) and σ^2(a:r,e); see O'Brien (1990: p. 483). With these, one can assess the reliability of the mean within-respondent scores obtained with a given number of alters or dyads per respondent. In doing this, it is assumed that alters or dyads are drawn at random from a larger population, and that as a result of this, the reports on which the mean scores are based would differ if the process of measurement were to be repeated (issues of the applicability of this assumption to egocentric network data are discussed further below).

The reliability of a measure of property X of an egocentric network based on a single alter or dyad can be estimated as[3]

$$\hat{\rho}_X^{(1)} = \hat{\sigma}^2(r)/[\hat{\sigma}^2(r) + \hat{\sigma}^2(a{:}r,e)]. \qquad (4)$$

Measure (4) assesses the reliability of single name interpreter items as measures of composition or density. We can also ask about the reliability of the mean scores for respondents obtained using reports on a larger number of alters or dyads. This will exceed $\hat{\rho}_X^{(1)}$, since random errors and deviations involving particular alters or dyads within the network of a given respondent will tend to cancel one another out (under the assumption that errors and alter/dyad effects are uncorrelated with respondent effects), leaving a mean that more closely approximates the "true" character of the social environment surrounding respondent i than does a report on a single alter or dyad. For m alters per respondent, we have

$$\hat{\rho}_X^{(m)} = \hat{\sigma}^2(r)/[\hat{\sigma}^2(r) + \hat{\sigma}^2(a{:}r,e)/m]. \qquad (5)$$

The estimates reported below give the estimated reliability of aggregate measures assuming the number of alters per respondent actually obtained in the surveys (GSS and NCCS) studied. Because the number of alters differs across respondents in these studies, we must replace m in formula (5) by (see O'Brien, 1990: 501n)

$$k_0 = [1/(n-1)][\Sigma_i n_i - (\Sigma_i n_i^2/\Sigma_i n_i)]$$

where n is the number of respondents and we replace n_i by n^*_i when evaluating the reliability of a density measure (which is based on dyads, rather than alters, nested within respondents). When the numbers of respondents and alters are large, the value of k_0 is close to the mean number of alters/dyads per respondent (Feldt and Brennan, 1989: 127); due to missing data and other differences in study design, k_0 varies for the measures evaluated here (see Tables 1 and 2 below).[4]

[3] One can estimate the reliability of a measure based on a single alter as shown here, but it should be noted that at least some respondents must cite multiple alters if the variance components involved are to be estimated.

[4] Below, I present "unpooled" reliability estimates, which estimate the error variance component $\sigma^2(a{:}r,e)$ under the assumption that the variance of alter or dyadic

The Assumption of Random Selection of Alters. Key to the use of this approach to evaluating the reliability of network measures is the postulate that the set of alters (dyads) described by a given respondent would differ if the data collection process were to be repeated. The variance components used in calculating the reliability coefficients in (4) and (5) are derived under the assumption that the alters/dyads described by a respondent are drawn at random from some population of alters or dyads that constitutes the local social environment surrounding the respondent.

As mentioned, some would regard alters instead as a fixed facet of measurement. Several aspects of the measurement process discussed previously suggest that this view is inaccurate, however: the researcher is not in a position to fix the set of alters described by a respondent, and fluctuations in inclusion of alters within this set over short time periods are appreciable.

A different sense in which the set of alters can be seen to be random is to regard them as representative of a hypothetical "superpopulation" of alters surrounding each respondent. Taking this perspective would treat the set of alters with whom a respondent currently maintains relationships to be a random sample from the opportunity structure within which that respondent's egocentric network forms. The superpopulation view would imply that if a given alter were to leave an egocentric network, he or she would be replaced by a similar alter.

These considerations suggest that it is worthwhile to examine the reliability coefficients given above for some important sources of egocentric network data. Reliability coefficients (4) and (5) treat alters or dyads as if they were randomly drawn, however, and in fact it is likely that that while alters are not fixed, neither are they strictly random. It is

characteristics may vary across respondents, because conventional handling of egocentric network data assumes this when calculating measures of network heterogeneity (see, e.g., Marsden, 1987). "Pooled" estimates instead assume that variances are the same for all respondents. In most cases, pooled and unpooled estimates of reliability are very similar (see O'Brien, 1990: 484).

therefore of interest to ask how departures from random selection would affect the reliability coefficients.

To the extent that alters are fixed rather than random--perhaps because there is more stability in the inclusion of "close" alters (such as spouses or "best friends") than more distant ones--alter effects (μ_{ik} - μ_i) in equation (3) will be stable across occasions of measurement. This would mean that a portion of the variance component σ^2(a:r,e) attributable to sampling of alters should be removed from the denominators of the ratios (4) and (5) used to estimate reliability coefficients. If this is done, the coefficients will become larger; in this sense, the estimates provide conservative indices of reliability.[5] A carefully conducted panel study might permit the isolation of components of alter effects attributable to stability across occasions; with the information available in cross-sectional designs like those analyzed below, though, all alter effects must be pooled with other sources of "error." Variations in the designs of the GSS and NCCS will allow us to gain some sense of the extent to which reliability estimates are affected by differences in research methods used to select alters for collection of name interpreter data.

Reliability Estimates for GSS Egocentric Network Measures

The GSS is an almost-annual study based on an area probability sample of roughly 1500 noninstitutionalized, English-speaking U.S. adults. The 1985 (n = 1534) and 1987 GSS (n = 1817)[6] network instruments enumerated network alters using the "important matters" name generator quoted above.[7] In 1985, name interpreter data were gathered on up to the first five alters cited; most respondents cited fewer than five. The name interpreter data included the age, education, race, religion, and sex of each alter; the

[5] Of course, conservative indices are not necessarily benign; a conservative estimate of reliability might lead an investigator needlessly to increase the number of alters for which name interpreter information is collected, in an effort to raise the reliability of density or composition measures. If the estimate is unduly conservative, this could involve a considerable excess expenditure of data collection resources.

[6] The 1987 study included a black oversample of size 353.

[7] For exact question wordings used for all GSS network items, see Davis and Smith (1991: pp. 328-360, 410-415).

frequency and duration of the respondent's relationship to each alter; the role labels (kin, friend, coworker, etc.) applicable to each relationship; and the closeness of the relationship between each pair of alters, which yields a measure of network density. Marsden (1987) gives an overview of the networks measured.

More limited name interpreter data were collected in 1987 for up to the first three alters cited. They included information on role labels, in the same format used in 1985, together with each alter's political party identification and the frequency of political discussions between respondent and alter. Information needed to measure network density among alters was not gathered. See Knoke (1990) for a discussion of these data.

In the 1988 GSS, limited data having to do with religion were collected for up to three alters. These alters were enumerated with a name generator asking respondents to name "good friends (other than your spouse)." Respondents were then asked about the religious preferences of these friends and whether each friend is a member of the respondent's congregation. Detailed denominations were recorded in 1988 (but not in 1985); here, we examine a dichotomous measure of whether each friend is Protestant and a coding of the friend's denomination along a fundamentalist-liberal continuum (Smith, 1990).

Table 1 presents reliability estimates for thirteen measures calculated for the 1985 data, seven based on the 1987 data, and three obtained using the 1988 data. The first column gives the one-alter reliabilities estimated via equation (4). The remaining columns present aggregate reliability estimates [equation (5)] for three and five alters, as well as for the average number of alters cited in these studies (k_0). Three and five are common upper limits placed on the number of alters for the purpose of collecting name interpreter data.

Insert Table 1 about here

Table 1 shows wide variations in the reliability of measures and in the number of alters necessary to obtain measures with adequate reliability. Race composition is quite

Table 1

Reliability of General Social Survey Egocentric Network Measures*

A. 1985 GSS	Reliability of Measure Based on ...				
Measure	1 alter	3 alters	5 alters	k_0 alters	(k_0)
A. 1985 GSS					
Race (White) Composition	.792	.919	.950	.924	(3.21)
Religious (Protestant) Composition	.540	.779	.854	.782	(3.05)
Friend Composition	.497	.748	.832	.760	(3.21)
Co-member Composition	.455	.714	.807	.728	(3.21)
Education Composition	.378	.646	.752	.656	(3.14)
Density	.348**	.615	.842	.740	(5.34)
Age Composition	.343	.610	.723	.626	(3.20)
Co-worker Composition	.299	.561	.680	.578	(3.21)
Neighbor Composition	.258	.510	.635	.527	(3.21)
Kin Composition	.245	.494	.619	.511	(3.21)
Mean Duration	.239	.485	.611	.502	(3.21)
Mean Frequency	.227	.468	.595	.485	(3.21)
Sex Composition	.020	.059	.095	.063	(3.21)
B. 1987 GSS					
Mean Frequency Political Discussions	.618	.829	.890	.798	(2.45)
Co-member Composition	.518	.763	.843	.725	(2.45)
Party Composition	.517	.762	.842	.698	(2.16)
Friend Composition	.496	.747	.831	.707	(2.45)
Neighbor Composition	.309	.573	.691	.523	(2.45)
Co-worker Composition	.256	.507	.632	.457	(2.45)
Kin Composition	.240	.487	.613	.437	(2.45)
C. 1988 GSS					
Religious (Protestant) Composition	.559	.792	.864	.760	(2.50)
Proportion Same Congregation	.520	.765	.844	.747	(2.69)
Mean Denominational Fundamentalism	.456	.716	.807	.668	(2.40)

* Unpooled estimates

** Based on 2 alters (1 dyad)

reliably measured in the 1985 data; a measure based on a single alter has a reliability of nearly 0.8. With three alters, the reliability of the percent white among the alters in a respondent's network increases to over 0.9. Religious composition too can be measured rather reliably with data on a small number of alters, as indicated by the results for 1985 and 1988. If we take the conventional level of 0.7 as the threshhold for an "adequately reliable" measure, three alters seems sufficient to obtain measures of the percentages of alters in a respondent's network who are Protestant or in the same congregation, and of the mean level of fundamentalism of those denominations.[8] Similarly, the political composition variables in the 1987 data appear to be adequately reliable; with three alters, a measure of party composition (measured along a three-point scale from Republican to Democrat) has a reliability of over 0.75, while the mean frequency of political discussions has a reliability over 0.8.[9]

At the other extreme, sex composition appears to be quite unreliably measured. The alters cited by most respondents include a mixture of men and women; this is in part the product of the relatively high number of alters who are tied to respondents through kinship (Marsden, 1987, 1988). The within-respondent variability in the sex of alters is large enough to suggest that estimates of the percentage female made at two occasions would be quite weakly correlated if based on a small sample of alters surrounding a respondent.

Between these extremes, we find a number of measures that evidently can be measured with adequate reliability provided that data are collected on a sufficiently large number of alters. Reliabilities of the mean age and education levels of alters rise above 0.7 only once five alters are sampled. Measures of the mean frequency and duration of

[8] Differences between the name generators used in obtaining the 1985 and 1988 data on religious composition do not seem to have had an appreciable effect on the reliability of these measures.

[9] Note that this result refers to the reliability of the measure only; validity or accuracy is a different matter. For party composition, Laumann's (1973: Chapter 2) results from alter interview studies suggest that respondents are likely to project their own party preferences onto alters.

relations to alters are more problematic; for three alters the reliabilities of these measures remain beneath 0.5, while for five alters they are close to 0.6. The results suggest that one would need data on seven or eight alters per respondent in order to attain adequate reliability for indicators of between-respondent differences in the average strength of relations.

Measures telling the degree to which egocentric networks are composed of people having a given role relation to respondents vary in their reliability. The absolute and relative reliabilities of these measures are quite similar for the 1985 and 1987 GSS data. Most reliable are the measures of the proportions of "friends" and "group members" in a respondent's network; their aggregate reliabilities exceed 0.7 when based on three alters and 0.8 when based on five. These results may reflect inter-respondent differences in levels of affiliation with groups, or in the likelihood of applying the label "friend" to a relationship (Fischer, 1982b); there is relatively little within-respondent variation in designation of alters as group members or friends.

Estimated reliabilities are lower for the remaining three role relation composition measures considered, the proportions of a network made up of kin, coworkers, and neighbors. For the numbers of alters measured in the 1985 and 1987 studies, these measures have reliabilities between .4 and .6; the results suggest that reliable indicators of between-respondent differences in these respects could be obtained with measurements based on 7 or 8 alters.

Network density is measured with acceptable reliability in the 1985 data. On average, respondents reported on roughly 5.3 dyads connecting pairs of alters, and the estimated reliability of this density measure is 0.74. If all respondents were to report on networks of size 5, including 10 dyads, reliability would be forecast to be nearly 0.85.

The GSS composition and density measures are based on the "core" networks elicited by the name generators used in obtaining these network data, and the reliabilities reported in Table 1 refer to such networks only. We next examine results for the NCCS network data, which describe larger segments of the interpersonal environments that surround respondents.

Reliability Estimates for the NCCS

The NCCS drew a multistage cluster sample of 1050 noninstitutionalized, English-speaking adults from predominantly white neighborhoods in northern California, principally in the San Francisco/Oakland and Sacramento SMSAs. As noted above, the alters sampled in the networks of NCCS respondents were enumerated by nine different name generators; see Fischer (1982a: pp. 36, 288) for specific details.

Name interpreter data were recorded in two different ways. Many questions were posed about any alter enumerated; these name interpreters included sex, role relations and closeness, among others. The networks of NCCS respondents ranged in size between 2 and 67, with a mean size of 18.5, so these data are quite extensive.

Due to pressures on interview time, other name interpreter data were gathered only for a subset of alters consisting of those named first in response to several of the name generators.[10] Though this was not a random sample of the alters cited by a respondent, it was "intended to be representative of the core network" (see Fischer, 1982a: pp. 290-291), and it is easy to imagine that there would be variability in the alters in such a subset across occasions of measurement. About four alters fell into the subsample for the typical respondent. Most data about alters in the subsample were obtained on a self-administered questionnaire (SAQ). Items appearing here included the duration and frequency of respondent-alter contact and the alter's age, marital/family status and labor force status. Data needed to construct a measure of network density were obtained by asking respondents, for each pair of alters in the subsample, whether the alters "know each other well."

[10] The name generators were: caring for the home, visiting or going out socially, discussing hobbies, discussing personal matters, giving advice, and lending money. Interviewers were instructed to include the first name given in response to each of these in the subsample, until five names were obtained (see Fischer, 1982a: p. 145).

Estimated reliabilities for fifteen measures calculated from the NCCS data appear in Table 2. While not all of the measures are directly comparable to the corresponding indices for the GSS, the relative level of reliability for substantively related measures is much the same across the studies, as we see by comparing the figures in Table 2 with those in Table 1.

Insert Table 2 about here

As in the GSS, measures of ethnoreligious composition and the proportions of alters who are friends or members of groups with which the respondent is affiliated are among the most highly reliable measures. Of these, however, only the proportion of alters who are the "same religion" as the respondent has an estimated reliability of over 0.7 for five alters. Sex composition again is estimated to have very limited reliability (0.37, even when based on 18.5 alters per respondent); the proportion of alters in the labor force is similarly unreliable.

Estimates of the reliability of indicators of the average strength (frequency, duration, closeness) of respondent-alter ties have modest reliability at best. Kin and coworker composition have unacceptable reliabilities when based on small numbers of alters, but like many of the other measures, these increase above the 0.7 threshhold if based on the relatively large number of alters included in the personal networks of NCCS respondents.

Network density is measured with slightly lower reliability in the NCCS data, but for five alters its estimated reliability of 0.771 is still acceptable. When based on the average number of ties reported by respondents to the NCCS, density has an estimated reliability of 0.69. Together with the GSS results for density, this suggests that adequate measures of density can be obtained with the relatively efficient designs used by these studies; the quality of the measure is not greatly affected by whether it is based on a core network, as in the GSS, or on a sample from a more extensive enumeration, as in the NCCS.

Table 2

Reliability of Northern California Community Study Egocentric Network Measures*

	Reliability of Measure Based on ...				
Measure	1 alter	3 alters	5 alters	k_0 alters	(k_0)
Proportion Same Religion[++]	.331	.597	.712	.903	(18.73)
Co-member Composition	.316	.581	.698	.895	(18.49)
Proportion Same Ethnicity[+]	.300	.562	.682	.833	(17.70)
Age Composition	.263	.518	.641	.586	(3.96)
Density	.252**	.503	.771	.689	(6.57)
Friend Composition	.220	.459	.586	.839	(18.49)
Mean Duration	.218	.455	.582	.526	(3.98)
Mean Closeness	.150	.346	.469	.765	(18.45)
Kin Composition	.139	.326	.446	.749	(18.49)
Co-worker Composition	.128	.306	.424	.731	(18.49)
Mean Frequency	.119	.289	.404	.349	(3.95)
Proportion Same Work[+++]	.090	.232	.334	.651	(18.57)
Proportion in Labor Force	.084	.216	.315	.264	(3.90)
Neighbor Composition	.082	.212	.310	.624	(18.49)
Sex Composition	.030	.086	.135	.366	(18.48)

* Unpooled estimates

** Based on 2 alters (1 dyad)

[+] Measured only for those identifying with a particular race, ethnicity or nationality ("Just Americans" excluded)

[++] Measured only for those identifying with a particular religion ("None" excluded)

[+++] Measured only for those in or retired from the labor force

Effects of Study Variations on Reliability

This section presents several methodological comparisons of the reliability levels of measures. These provide some indication of how measure quality is affected by differences in methods used to select the alters for whom name interpreter data are gathered. They also suggest that high quality measures are more readily obtained for the "core" egocentric network surrounding a respondent.

In general, the absolute level of reliability appears to be slightly lower, for a given number of alters, in the NCCS than in the GSS (compare Table 2 to Table 1). To gain some indication of whether this reflects differences in the range of alters included in egocentric networks by the two studies, or other study differences, I reestimated the reliability of the NCCS measures using a subset of "core" alters. These estimates were obtained by considering only those alters with whom the respondent said that he or she discussed "personal matters"; this is the NCCS name generator most comparable to that used in the GSS. Five-alter reliabilities estimated using these data appear in the first column of Table 3.[11]

Insert Table 3 about here

Two comparisons involving the "personal matters" reliability estimates are of interest. First, they are higher than the corresponding estimates in Table 2. Thus, it appears that these aspects of networks are more reliably measured for core networks than for extended ones, since core networks include alters who are more homogeneous in most respects. Second, most reliability estimates in Table 3 are slightly lower than the estimates for the corresponding measures in the GSS shown in Table 1 (kin composition is an exception). These differences could be due to the different "core" name generators used in the studies, the wording of name interpreter items, or other study differences.

[11] The reliability of measures based on items on the subsample SAQ could not be reestimated, since the subsample included only one "personal matters" alter per respondent.

Table 3

Variation in Reliability of NCCS Network Measures under Differences in Methods of Selecting Alters*

	5-Alter Reliability of Measure Based on ...		
	"Personal Matters" alters	SAQ alters	Randomly sampled alters
Co-Member Composition	.790	.697	.671
Friend Composition	.784	.576	.626
Proportion Same Ethnicity+	.753	.627	.713
Proportion Same Religion++	.735	.669	.720
Kin Composition	.660	.318	.478
Mean Closeness	.546	.383	.425
Co-Worker Composition	.512	.434	.439
Proportion Same Work+++	.428	.361	.329
Neighbor Composition	.434	.196	.359
Sex Composition	a	.199	.164

* Unpooled estimates
+ Measured only for those identifying with a particular race, ethnicity, or nationality ("Just Americans" excluded)
++ Measured only for those identifying with a particular religion ("None" excluded)
+++ Measured only for those in or retired from the labor force
a Not reported because variance component for respondents is negative

A second methodological query is how reliability estimates are affected by sampling from the list of alters enumerated by one or more name generators. To explore this, I recalculated reliability estimates for those NCCS measures based on name interpreter items administered for all alters after sampling alters in two different ways. First, I considered the sampling procedure actually used in the NCCS by using data for only the subsample of alters on which SAQ data were collected. Second, I randomly sampled a set of five alters from the distribution elicited for each respondent, to simulate results obtained under a random selection procedure. The five-alter reliabilities obtained from these analyses are presented in the middle and right columns of Table 3. Comparing these results to those based on all of the NCCS data (Table 2) allows us to assess the manner in which these sampling schemes for alters affect reliability.

For most measures, the estimates of reliability based on the SAQ subsample are slightly smaller than the complete-data estimates. Differences are especially notable for kin and neighbor composition. Evidently the procedure used to obtain the SAQ subsample alters actually overestimated the amount of diversity in the personal networks of the NCCS respondents, though it is not immediately plain how the procedure would have had such an effect. As expected, estimates in the final column, based on a random sample of five NCCS alters for each respondent, are sometimes greater and sometimes less than the complete-data estimates displayed in Table 2.

A final set of methodological comparisons demonstrates the effect on reliability of raising the number of alters for whom name interpreter data are collected. From formula (5) it is plain that reliability should increase with the number of alters, as long as the error variance component $\sigma^2(a{:}r,e)$ does not rise rapidly as the number is increased. When data are recorded for the "first k" alters in studies like the GSS, however, it is by no means plain that this will not occur.

The 1985 GSS contains good information on citation order, so we are in a position to simulate the data that would have been obtained there had name interpreter data been gathered on only the first three or four alters cited, rather than the first five as was actually done. The way in which this affects the estimated reliability of the measures in the 1985 GSS is displayed in Table 4.

Insert Table 4 about here

If alters are drawn at random from the social environment surrounding a respondent, then the reliability of an aggregate measure based on the data should be affected only by the number of alters selected. This would mean that reliability estimates should increase as we move from left to right within a given row in Table 4, but should not change as we move down columns for a given measure. If, on the other hand, alters cited later (and therefore included only when the limit on alters is raised) are more diverse, then σ^2(a:r,e) will rise as the limit on alters does, with a corresponding decline in estimated reliability for a given number of alters.

In Table 3, I present results for seven measures from the 1985 GSS. For some measures, we can observe a slight tendency for reliability (assuming a given number of alters) to decrease as the number of alters rises. For example, one-alter reliability for race composition shrinks from .815 under the three-alter limit to .792 for the five-alter limit. A similar decline is observed for kin composition.

In other cases, however, a clear trend is less apparent. One-alter reliability of Protestant composition actually rises as we move from three to four alters, while the reliability of the density measure fluctuates. The general conclusion, as prior research (Burt, 1986) leads us to expect, is that the selection procedure implemented through the use of the "important matters" name generator is not random--more homogeneous alters are cited early. The effect of this on the estimated reliability of measures seems to be rather modest, however; certainly the greater heterogeneity of later-cited alters (reflected in increases in σ^2(a:r,e)) is more than counterbalanced by the reduction in the error variance of aggregate measures (achieved by increasing m in the denominator of (5)) that

Table 4

Reliability of Selected Egocentric Measures, 1985 GSS
Under Different Limits on the Number of Alters
for Whom Name Interpreter Data are Collected

		Reliability of Measure Based on ...				
Measure	Limit on Alters	1 alter	3 alters	5 alters	k_0 alters	(k_0)
Race (White)	3	.815	.930	.957	.917	(2.51)
Composition	4	.800	.923	.952	.922	(2.94)
	5	.792	.919	.950	.924	(3.21)
Religious (Protestant)	3	.531	.773	.850	.730	(2.38)
Composition	4	.542	.780	.855	.768	(2.79)
	5	.540	.779	.854	.782	(3.05)
Co-member Composition	3	.465	.723	.813	.685	(2.50)
	4	.459	.718	.809	.714	(2.94)
	5	.455	.714	.807	.728	(3.21)
Education Composition	3	.388	.656	.761	.609	(2.45)
	4	.385	.653	.758	.643	(2.88)
	5	.378	.646	.752	.656	(3.14)
Density	3	.357**	.625	.847	.588	(2.56)
	4	.342**	.610	.839	.681	(4.09)
	5	.348**	.615	.842	.740	(5.34)
Kin Composition	3	.269	.524	.647	.479	(2.50)
	4	.251	.502	.627	.497	(2.94)
	5	.245	.494	.619	.511	(3.21)
Mean Frequency	3	.230	.473	.599	.428	(2.50)
	4	.227	.468	.595	.463	(2.94)
	5	.227	.468	.595	.485	(3.21)

* Unpooled estimates
** Based on 2 alters (1 dyad)

comes with the introduction of additional alters. We can see this by making comparisons along the main diagonal in each panel of Table 4. The modest magnitude of the heterogeneity effect on reliability estimates suggests that these indices provide useful indications of the extent to which scores on a measure would be reproduced if data collection were to be repeated after a short time interval.

Discussion

The nested design used to collect egocentric network data prevents researchers from evaluating the reliability of aggregate measures or indices based on such data with the usual techniques. This paper has outlined techniques for obtaining reliability estimates, provided such estimates for several important sources of egocentric network data, and performed a number of methodological comparisons which allow us to assess the proposed approach.

The results clearly indicate differences in the capacity to measure different network concepts reliably. Ethnoreligious composition, political composition, and the tendency to cite "friends" and co-members of organizations are relatively reliable; sex composition is quite unreliable. Network density appears to be measured with adequate reliability. Other composition measures examined have more modest reliability for the limits on numbers of alters typically imposed in survey designs. In general, it appears easier to obtain highly reliable measures for the core segments of egocentric networks than for more extensive portions.

It will be observed that reliability of aggregate measures is linked closely to the extent of homogeneity among the alters in a respondent's network. This is as it should be: if alters are homogeneous, then the score assigned to a respondent's network on a given property depends relatively little on which alters are sampled, or how. If a respondent's alters are diverse, however, then scores on a measure can be quite volatile, changing substantially due to happenstance inclusion of some alters and exclusion of others. The advantage of the measures used here is that they yield an indication of the expected extent of these variations in scores.

Several caveats and implications should be appended. The results suggest that to obtain adequately reliable indicators for many measures, name interpreter data should be gathered for five or more alters. This implication, however, should be tempered by the recognition that the reliability estimation technique assumes random selection of alters. It is clear that name generators introduce some variability into the selection of alters, but also clear that alter selection is not strictly random. Since the likely effect of nonrandom selection is to understate the reliability of composition and density measures, researchers should not move too quickly to increase the limit on alters in an instrument for data collection on the basis of these results.

The analysis of variance approaches to reliability advocated here can be extended by incorporating occasions as a measurement facet, as in panel or test-retest studies. Such studies would be valuable in efforts to separate the stable and random components of alter effects and hence in obtaining more precise indications of reliability than can be extracted from cross-sectional studies like the GSS and the NCCS. Ideally, such studies would include at least three waves, to facilitate the separation of stability and unreliability. They should be careful to track the identities of alters from wave to wave, so that effects due to turnover in alters can be separated from those due to fluctuation in reports about specific alters.

Density and composition measures, of course, by no means exhaust the measures of interest to researchers who make use of egocentric network data. In particular, measures of range or variability among the set of alters are common. Because such measures are sensitive to the within-respondent variability of a set of alters, as distinct from the within-respondent central tendency examined here, evaluating their quality would require techniques that go beyond those used above.

References

Bien, W., J. Marbach and F. Neyer. 1991. "Using Egocentered Networks in Survey Research: A Methodological Preview on an Application of Social Network Analysis in the Area of Family Research." Social Networks 13: 75-90.

Broese van Groenau, Marjolein, Eric van Sonderen, and Johan Ormel. 1990. "Test-Retest Reliability of Personal Network Delineation." Pp. 121-136 in C.P.M. Knipscheer and T.C. Antonucci (eds.) Social Network Research: Substantive Issues and Methodological Questions. Lisse, The Netherlands: Swets and Zeitlinger.

Bryk, Anthony S. and Stephen W. Raudenbush. 1992. Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park, CA: Sage.

Burt, Ronald S. 1984. "Network Items and the General Social Survey." Social Networks 6: 293-339.

Burt, Ronald S. 1985. "General Social Survey Network Items." Connections 8: 119-123.

Burt, Ronald S. 1986. "A Note on Sociometric Order in the General Social Survey Network Data." Social Networks 8: 149-174.

Carmines, Edward G. and Richard A. Zeller. 1979. Reliability and Validity Assessment. Beverly Hills: Sage.

Davis, James A. and Tom W. Smith. 1991. General Social Surveys, 1972-1991: Cumulative Codebook. Chicago: National Opinion Research Center.

Feldt, Leonard S. and Robert L. Brennan. 1989. "Reliability." Pp. 105-146 in Robert L. Linn (ed.) Educational Measurement. New York: Macmillan.

Fischer, Claude S. 1982a. To Dwell Among Friends: Personal Networks in Town and City. Chicago: University of Chicago Press.

Fischer, Claude S. 1982b. "What Do We Mean By 'Friend'?: An Inductive Study." Social Networks 3: 287-306.

Friedkin, Noah E. 1990. "A Guttman Scale for the Strength of an Interpersonal Tie." Social Networks 12: 239-252.

Hoffmeyer-Zlotnik, Juergen H.P. 1990. "The Mannheim Comparative Network Research." Pp. 265-279 in Jeroen Weesie and Henk Flap (eds.) Social Networks Through Time. Utrecht, The Netherlands: ISOR/ETS, Rijksuniversiteit Utrecht.

Knoke, David. 1990. "Networks of Political Action: Toward Theory Construction." Social Forces 68: 1041-1063.

Laumann, Edward O. 1973. Bonds of Pluralism: The Form and Substance of Urban Social Networks. New York: Wiley Interscience.

Marsden, Peter V. 1987. "Core Discussion Networks of Americans." American Sociological Review 52: 122-131.

Marsden, Peter V. 1988. "Homogeneity in Confiding Relations." Social Networks 10: 57-76.

Marsden, Peter V. and Karen E. Campbell. 1984. "Measuring Tie Strength." Social Forces 63: 482-501.

Mitchell, J. Clyde. 1969. "The Concept and Use of Social Networks." Pp. 1-50 in J. Clyde Mitchell (ed.) Social Networks in Urban Situations. Manchester, UK: Manchester University Press.

Mitchell, Sandra K. 1979. "Interobserver Agreement, Reliability, and Generalizability of Data Collected in Observational Studies." Psychological Bulletin 86: 376-390.

O'Brien, Robert M. 1990. "Estimating the Reliability of Aggregate-Level Variables Based on Individual-Level Characteristics." Sociological Methods and Research 18: 473-504.

Shavelson, Richard J. and Noreen M. Webb. 1991. Generalizability Theory: A Primer. Newbury Park, CA: Sage.

Shrout, Patrick E. and Joseph L. Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." Psychological Bulletin 86: 420-428.

Smith, Tom W. 1990. "Classifying Protestant Denominations." Review of Religious Research 31: 225-245.

Sudman, Seymour. 1985. Experiments in the Measurement of the Size of Social Networks. Social Networks 7: 127-151.