How to Use SAS to Study Egocentric Networks

Barry Wellman
Centre for Urban & Community Studies
University of Toronto

Many network analysts study egocentric networks, that is, networks defined "Ptolemaically" from the standpoints of focal individuals (Figure 1). Egocentric network analyses are common in the study of personal, community, and social support, and they can be used for studying other matters such as corporate relations. Most egocentric analyses deal only with the direct ties that focal individuals have with the members of their network. A few researchers have studied the links that network members have among themselves, and an even smaller number have studied a focal individual's indirect ties with "friends of friends," etc. In this paper, I deal only with the basic and most common case: how to study direct ties. But even when dealing with direct ties, you still have to keep track of a lot of information:

- characteristics of focal individuals (e.g., gender, ethnicity);
- characteristics of ties (relationships) between focal individual and network members (e.g., frequency of contact), contents of relationships (e.g., providing emotional aid), and the basis of relationship (e.g., friendship);
- characteristics of the network members with whom focal individuals have ties (e.g., gender, ethnicity);
- aggregated characteristics of the network members and ties in each network, i.e., network composition (e.g., mean frequency of contact, proportion of network members providing emotional aid); network structural characteristics (e.g., density, number of clusters).

SAS's data-handling facility lets you store and link all these different kinds of data. (SAS-PC and SPSS may have similar capabilities, but I have not used those programs.) The basic procedure is to:

(a) store network member and tie data in one TIEWISE data set;
(b) store focal individual and network structure data in a separate NETWISE data set;
(c) use the same NETID variable and values in the two data sets to identify the focal individuals;
(d) use SAS's UNIVARIATE and MERGE procedures to link the data.

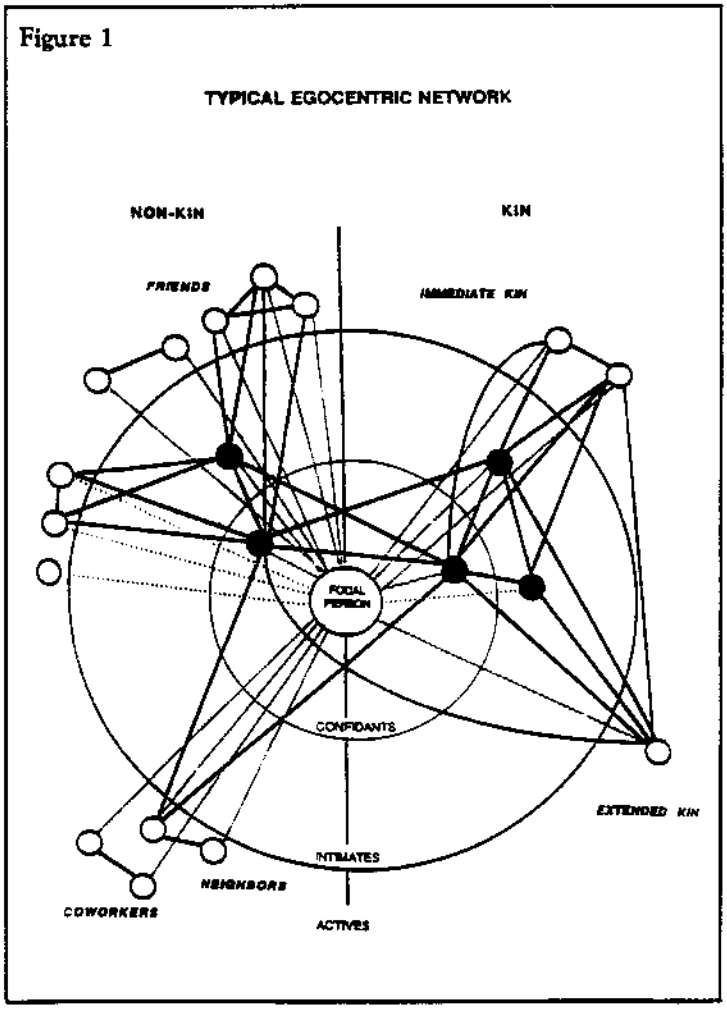
The Original Data Sets

Tiewise: In the tiewise data set, network members' characteristics not only include personal characteristics such as age and gender, they also include tie characteristics such as frequency of contact with the focal individual. By definition, a focal individual and a network member have exactly one tie, although a tie often contains different role relationships. Even though the tie is between the focal individual and the network member, you can treat network members as "possessing" the characteristics of their ties with focal individuals. This makes it possible to store both network member and tie variables in one data set (arbitrarily called *TIE* here).

Netwise: By definition, focal individuals are at the centers of their own egocentric networks (Figure 1). Thus focal individuals not only have personal characteristics such as gender and ethnicity, they have networks with different densities, numbers of clusters, etc. You can, therefore, store information about focal individuals and their networks in the same records of a netwise data set (arbitrarily called NET in this article). Note, though, that it is more useful to use the procedures described in the next section to compute network compositional data.

There is one special condition for data entry. The otherwise separate tiewise and netwise data sets must each contain the same variable that identifies the focal individual. In the tiewise data set, the NETID variable identifies the network to which each network member belongs. If several network members belong to the same network, each of the network members will have the same NETID number. In the netwise data set, the NETID variable uniquel identifies the focal individual and his or her network. In the tiewise data set, the

Figure 1



NETID variable is used to produce summary information about each network and to join this summary information with information in the netwise data set about focal individuals.

Each tie in the TIE data set should also have a unique TIEID. I also use a third identification variable to identify the members of each network, numbering them within each network from "1."

The Univariate Procedure

In addition to computing summary statistics such as means, UNIVARIATE can create a new data set containing summary statistics. This feature allows UNIVARIATE to produce network compositional data such as the mean frequency of contact for each network, the percentage of network members who provide emotional support, and the number of network members who provide emotional support.

The BY statement in the UNIVARIATE procedure specifies the variable by which records will be grouped when computing statistics. This allows analysts to produce summary statistics for each egocentric network. For example, instead of computing the mean frequency of contact for the entire sample of ties, you can compute separate means for each focal individual's network. In the current example, the statement would be: "BY NETID;".

The VAR statement in UNIVARIATE lists the variables for which summary statistics will be produced. The OUTPUT statement defines and creates a new summary data set from the statistics that have been computed from the TIE variables named in the VAR statement. In the OUTPUT statement, you specify which statistics (mean, etc.) are to be used to create the new summary variables in the TIESUM data set. (TIESUM is an arbitrary name for the data set produced by UNIVARIATE.)

Example: data on 343 network members are stored in the TIE file (the example is from Wellman 1992). The following SAS statements create the new data set TIESUM, containing summary information about 29 networks. (The numbers at the beginning of each line are for convenience in the discussion immediately following; do not use them in preparing SAS statements.)

```
1. PROC UNIVARIATE DATA = TIE
   NOPRINT;
2. VAR FTF EMAID CTAGE COUNT;
   *{these are 3 variables in the tie data set for
   residential-distance, emotional age and
   network member's age};
3. BY NETID;
4. OUTPUT OUT = TIESUM
   MEDIAN = MDFTF MDEMAID
        MDCTAGE MCOUNT
   MEAN = MFTF PEMAID MCTAGE
        MDCOUNT
   SUM = SFTF SEMAID SCTAGE
        NETSIZE;
```

Notes on this procedure:

1. This starts the UNIVARIATE procedure for the tiewise data set. The NOPRINT statement is optional. Without it, UNIVARIATE will print summary statistics for each of the 29 egocentric networks.

2. The VAR statement includes selected variables from the tiewise data set (e.g., FTF = frequency [number of days annually] of face-to-face contact between the network member and the focal individual). In this example, a tiewise emotional support variable (EMAID) is coded "0" or "1." The "0" code indicates that the network member does not provide emotional support to the focal individual, while the "1" code indicates that he or she does. This 0/1 coding is a handy tool for using variable means to calculate the proportion of network members who have a given characteristic such as "provides emotional aid." (See also #4, OUTPUT below.)

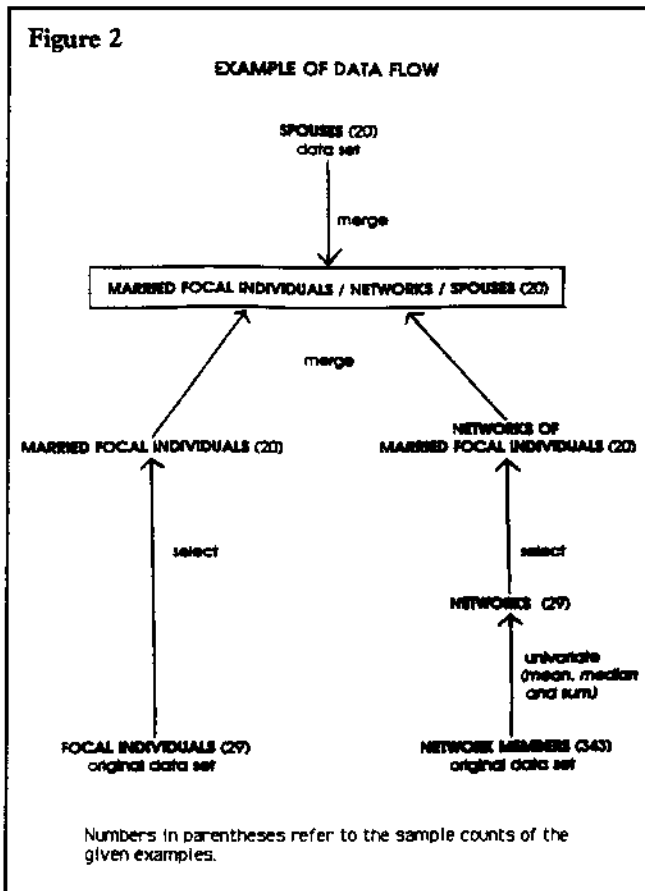
3. The BY statement specifies that statistics are to be computed separately for each value of the variable NETID. This is the crucial step. As each of the 343 network members has one of 29 NETID values, summary statistics for each of the 29 networks will be computed for the variables listed in the VAR statement. Thus, the BY statement collapses tie information into by-network summaries: one summary for each of the 29 egocentric networks.

Note that the BY statement assumes that the file is already sorted, in ascending order by the desired reference variable. Storing the data permanently in this way will save time and avoid problems. However, if the data are not already sorted, SAS's SORT procedure will sort and (can) save the data for later runs.

4. The OUTPUT statement directs SAS to create and store a new data set: TIESUM. This data set will contain the summary statistics requested in the UNIVARIATE procedure for each value of NETID (that is, each network). The keys MEAN=, MEDIAN=, and SUM=, assigned variable names to the variables contained in the TIESUM data set. For example, the keyword MEAN= is followed by the variable names that will be assigned to the mean values of the original variables in the tiewise data set.

In the example, I have adopted the convention of keeping the variable names the same in both the TIE and TIESUM data sets except that the names are preceded by an "M," "MD," or "S." I use a "P" prefix for mean variables that

Figure 2



Numbers in parentheses refer to the sample counts of the given examples.

have been calculated from 0/1 binary codes in the tiewise data set. The mean in such cases is also the proportion of ties in a network that have a particular characteristic, such as giving emotional aid. (See the description of such binary variables above.) Although you can choose any names for variables, these conventions help to keep track of things by associating the original tiewise variables with their newly created netwise summaries.

These new variables are netwise summaries of the tiewise information for each focal individual. Thus, MFTF equals the mean face-to-face contact between the focal individual and the members of his or her network. Because emotional support was coded 0/1 in TIE, the MEAN = keyword computes the proportion of ties in each network that provide it and outputs it in the PEMAID variable in TIESUM. In another example, the MEDIAN = keyword for the FTF variable creates a new MDFTF variable, the median amount of face-to-face contact between the focal individual and all of his or her network members. Similarly, the SUM= keyword creates a new SFTF variable, the total amount of face-to-face contact between each focal individual and all the members of his or her network.

Further notes on the OUTPUT statement:

(a) SAS requires that the variables in the OUTPUT statement must be entered in the same order as their tiewise counterparts in the preceding VAR statement. If EMAID is second in the VAR statement, then MDEMAID must be the second variable in the OUTPUT statement. If you scramble the order or omit a variable name, your output will be horribly (and sometimes unobtrusively) wrong.

(b) Note in the example that the same output statistics are requested for each of the three variables. The researcher cannot choose particular statistics to be computed for only some variables. For example, even if you only need MDFTF and SEMAID, your OUTPUT statement must still also include SFTF and MDEMAID. MDEMAID is meaningless in itself, but you must include it to get medians for other input variables. If you request any output summary statistic for any input variable, then you must request that output statistic (and define it as an output variable) for all input variables. You must request these output variables in the same order of the input variables for all the summary statistics you request in a single UNIVARIATE procedure.

(c) The only semicolon in the OUTPUT statement comes after all the summary statistics and new variables have been named, even though this may be many lines later.

(d) UNIVARIATE can also output other summary statistics. For example, I have used the standard deviation to measure the SES and age heterogeneity of egocentric networks.

Calculating Network Size

It is easy to use UNIVARIATE to calculate network size. First, copy NETID to a new variable (arbitrarily called COUNT here) so that you can recode it without destroying it. (Use NETID because it should never have missing values.) Recode COUNT so that all nonmissing values = "0." If you include COUNT in your PROC UNIVARIATE, the SUM of COUNT will be the size of each network. I call this variable NETSIZE in the example above; it is the one time I deviate from my strict naming rule. Using the same approach with more recoding will provide more specialized counts, such as the number of kin.

Analyzing Network Composition

The approach described above has provided information about the composition of each egocentric network. You can now use TIESUM directly to compare networks. In this example I use PROC CORR to correlate the mean and total amounts of face-to-face contact in each network with the proportion of network members who provide emotional aid.

```
PROC CORR DATA = TIESUM;
VAR MFTF SFTF PEMAID;
```

Netwise Analysis

With a MERGE statement, you can combine the newly created TIESUM data set with the NET data set that contains information about focal individuals and the structure of their networks. SAS does this by "match merging" the TIESUM and NET data sets. It combines records that have the same value for the network identification variable, NETID. That is why analysts must make sure during data entry that both TIESUM and NET contain matching NETID values. NETID is in the original NET data set. It also is carried over automatically from TIE to TIESUM when it is used in the BY option of UNIVARIATE.

```
DATA NETALL;
MERGE TIESUM NET;
BY NETID;
```

The preceding commands create a new netwise data set, NETALL, formed by the merger of the TIESUM and NET data sets. Now you can examine such matters as the relationship between a focal individual's gender (from NET) and the percentage of emotional aid in his or her network (from TIESUM).

```
PROC CORR DATA = NETALL;
VAR GENDER PEMAID;
```

Check for deletes in previous sentence. For example, the following will only do correlations for the networks of men.

```
DATA MEN;
MERGE TIESUM NET;
BY NETID;
IF GENDER = 1;
PROC CORR DATA = MEN;
VAR MFTF SFTF PEMAID;
```

Integrating Tiewise, Individual, and Network Analysis

Analysts also may want to retain the tiewise organization of the TIE data set, but supplement it with information about focal individuals and network structure. For example, our research group needed to know the gender of focal individuals and of network members to compare ties between men, between women, and between men and women (Wellman 1992). The sample size in this example is 343 ties and not the 29 networks produced through the UNIVARIATE and MERGE examples described above. In another study, we

used a similar technique to analyze the ties of married people.

A simple MERGE will do these things; there is no need to use UNIVARIATE. In the example, the NET data set is merged with the tiewise TIE data set to form a new data set (arbitrarily called TIEFOCAL). As in the preceding MERGE example, BY NETID associates the appropriate records in TIE and NET. But now, if the focal individual has 10 ties, the information from NET will be copied 10 times and merged with each network member's record. The merged TIEFOCAL data set will have 343 records, like TIE, but it may be much larger because the focal individual's information is repeated for each member of his or her network.

```
DATA TIEFOCAL;
MERGE TIE NET;
BY NETID;
```

Notes about this procedure:

1. It will work only if the original TIE data set is used and not the summary TIESUM data set that UNIVARIATE creates.

2. It will work only if similar variables in the original TIE and NET data sets have different names. Otherwise, disaster can strike as when a TIE variable named SEX is merged with a NET variable named SEX. I suggest using consistent, unique prefixes (e.g., TSEX and FSEX) in the original TIE and NET data sets.

3. You can reduce the size of the merged TIEFOCAL data set by using a KEEP or DROP statement in a preceding DATA step to limit the number of variables that will be merged. This is especially useful in reducing the size of the NET data set because its variables will be repeated for many (TIE) records when TIEFOCAL is created.

Capabilities and Limitations

1. It is easy to link summary data to information about the characteristics of focal individuals. Moreover, you need not make linkage decisions ahead of time. At any time, analysts can choose to combine different characteristics of focal individuals and networks. It is also easy to focus on the ties or networks of specific types of focal individuals.

Continued on p. 12

Text Managment Programs: Using GOfer

H. Russell Bernard

A few issues ago, I discussed the subject of text management. The big question that people have about text management (TM) programs, of course, is "Which one should I buy?" If you need an industrial strength TM program--one that lets you work on several files of thousands of pages each, done by different researchers at various sites--then I recommend ZY-Index. More on ZY in another article. This time I want to talk about GOfer, a much less powerful program than ZY-Index, but just the thing for managing a couple of thousand pages of field notes, or finding a phrase or sentence in one of the hundreds of text files lying around on your hard disk (old papers that have long since been published, for example, or drafts of grant proposals, and so on).

GOfer (as in "go fer this and go fer that") is menu driven, easy to use, and doesn't require indexing. This is a big plus. Big TM programs force you to make an index of your text files. An index is a locator file: each occurrence of each word in your text is given a unique address, in a language that the computer understands. When you do searches for words or phrases with indexing programs, the programs look through the index rather than through the raw text. This is what makes those programs so fast.

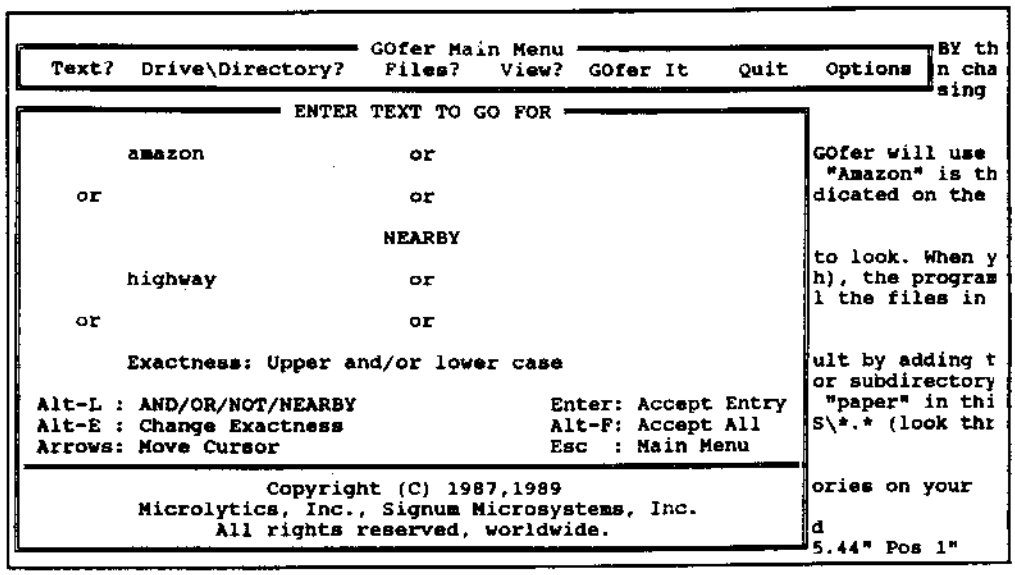
That's the good news. The bad news is that indexing takes time, and indices take up a lot of extra room on your disk drive. Also, if you change you original text, you have to run the index again. GOfer lets you operate on raw text, and it recognizes text in all the major word processor formats (WordStar, WordPerfect, Word, PC-Write, etc.).

GOfer is a terminate-and-stay-resident (TSR) program. When you install GOfer, you tell it what word processors you use and define a hot key. The default hot key is ALT-G. You run GOfer, see the program logo, and then bring up your word processor. The GOfer program log vanishes, but the program runs in the background. When you hit the hot key (like ALT-G) from inside your word processor, GOfer pops up and lets you tell it to find something for you.

You can get an idea of what GOfer looks like in the figure below. You can see some of the text from my WordPerfect document sticking out along the right side of the GOfer menu. To get out of GOfer, you just tell it to quit (hit Q).

In the example, I asked GOfer to find the word "Amazon" if it occurs NEARBY

Continued on next page



How to Use SAS, *continued from p. 9*

2. The ID option of Univariate offers a bonus by identifying focal individuals, ties, and networks that have high or low values on a variable. For example, I have used this option to identify those focal individuals whose networks provide very high or low levels of emotional support.

3. Everything discussed in this paper can be done in the same run. This is especially feasible if you are using a fast mainframe or a small data set. Rather than defining variables semi-permanently, doing everything in the same run encourages you to redefine variables for analytic purposes.

4. Keeping two separate data sets is more efficient than combining tie and network data in one set because it avoids the repetition of individual, tie, and network information. Moreover, separate tie and network data sets permit doing more efficient computer runs when only one data set is needed.

5. The general approach described here can be extended. For example, Wellman and Wellman (1992) linked TIE and NET data with a third data set that contained information about ties with spouses. We used multiple UNIVARIATEs and MERGEs to accomplish this (Figure 2), but the logic was the same.

6. The confirmatory statistics produced by SAS (such as correlation coefficients and their associated significance levels) assume that each record is an independent unit of analysis. This is not often true in egocentric network analysis. To be sure, focal individuals and networks often are independent units. Hence, analyses using NET and TIESUM rarely have this problem. However, the ties of a focal individual are inherently not independent from each other. Therefore, a sample of many focal individuals' ties--as stored in TIE or TIEFOCAL--is not a fully independent sample even if the focal individuals were sampled independently. The variance in such data sets should be lower than in a fully independent sample.

It may be possible to treat a tiewise data set as a cluster sample log-linear analysis. However, this is suitable only for analyzing a small number of variables in a large sample. Despite the uncertainty on the question of fully independent units, my comments are a caution rather than a roadblock. Until now analysts have treated each tie as an independent unit of analysis, and I do not know of any complaints from referees or misleading results. (See the studies reviewed in Wellman 1988; Campbell and Lee 1991.)

7. It is difficult to use SAS to calculate measures of network structure. Use NEGOPY, STRUCTURE, or UCINET. Analysts then can add structural measures calculated with these programs to the NET data set for further analysis using SAS.

References

Campbell, Karen and Barret Lee. 1991. "Name Generators in Surveys of Personal Networks." *Social Networks* 13 (Sept.): 203-22.

Wellman, Barry. 1988. "The Community Question Re-evaluated," p. 81-107, in *Power, Community and the City*, edited by Michael Peter Smith. New Brunswick, NJ: Transaction Books.

Wellman, Barry. 1992. "Men in Networks: Private Communities, Domestic Friendships," p. 74-114, in *Men's Friendships*, edited by Peter Nardi. Newbury Park, CA: Sage.

Wellman, Barry and Susan Gonzalez Baker. 1985. "Using SAS Software to Link Network, Tie and Individual Data." *Connections* 8 (2-3):176-87.

Wellman, Beverly and Barry Wellman. 1992. "Domestic Affairs and Network Relations." *Journal of Social and Personal Relationships* 9 (August): in press.

Acknowledgments

This is a thoroughly revised version of Wellman and Baker (1985). I am grateful for the financial support of The Social Sciences and Humanities Research Council of Canada in preparing this paper and for the assistance of Milena Gulia and for the editing of Beverly Wellman.