The Use of Multilevel Modeling for Analysing Personal Networks: Networks of Cocaine Users in an Urban Area*

TOM SNIJDERS, MARINUS SPREEN and RONALD ZWAAGSTRA

Department of Statistics and Measurement Theory
University of Groningen
The Netherlands

ABSTRACT: This paper explains how multilevel methods can be employed to analyse personal network data, when the dependent variable under consideration is a function of the relations contained in the personal networks. These methods take into account the mutual dependence of relations of the same respondent, and allow us to study the variability between respondents as well as the variability between different relations within respondents. As an illustration, multilevel models are applied to an analysis of personal networks of cocaine users, focusing on the significance of cocaine in their personal relations with other cocaine users.

KEY WORDS: cocaine, hierarchical linear model, multilevel analysis, personal network, random effects, snowball sample

1. INTRODUCTION: THE NESTING STRUCTURE OF PERSONAL NETWORK DATA

The statistical analysis of relational data, which is essential to social network research, poses special problems because the independence assumptions that are fundamental to many statistical methods can be vitiated by the relational nature of the data. In this article, a relation is understood always to refer to a *dyad*, i.e., a pair of individuals (or other kinds of actors). The two most common forms of relational data are complete networks and personal, or egocentric, networks. A *complete network* refers to a group of individuals and one or more types of relation, and the data indicates, for every pair of individuals, whether or not the relation is present between them. A *personal* or *egocentric network* refers to one central individual, indicated as *ego*, the respondent, or the focal individual; the other individuals with whom ego has a relation of the considered type together constitute ego's personal network. These other individuals are referred to as *alters* and also, in their relation to the respondent, as *nominees*. The usual type of personal network data consists of a sample of respondents, wherein for each respondent, all relations in his personal network are identified.

Many statistical methods are available for the analysis of complete network data; see, for example, the textbook by Wasserman and Faust (1994). This is not the case for the analysis of personal networks. Personal

Journal of Quantitative Anthropology 5, 85–105, 1995.

network data distinguish themselves from other types of data in two aspects: first, the mutual dependence of the various relations of each single respondent; and second, the fact that diverse kinds of variables may be present. Variables can be attributes of individual respondents, of individual nominees, or of dyadic respondent-nominee relationships. Variables that refer, through an aggregation procedure, to the entire personal network, e.g., the number of relations mentioned by ego, are considered here as attributes of ego. For questions where the dependent variable as well as the explanatory (independent) variables are attributes of the focal individual, traditional statistical methods, such as OLS regression analysis, can be used. Examples are studies of the size and composition of social networks, and studies of the effect of social support on health. For such questions, the unit of analysis is ego together with his personal network; characteristics of the personal network can be included in the analysis by aggregating characteristics of relations to the level of the personal network. This procedure is often followed; an example is Campbell and Lee (1992; see p. 1085). However, aggregation of characteristics of relations to the level of the personal network is not always satisfactory, especially when the dependent variable is itself a function of the personal network.

This article is concerned with the statistical analysis of questions where the unit of analysis is the dyad or the relation, and where a dependent variable is analyzed that is a function of the relation, for example an indicator of the relation's intensity or its content, while the available data is a sample of personal networks. Since each respondent, or ego, can report several relations with alters, the various relations in the data set may not be treated as independent. We assume that respondents have been sampled in a sufficiently large social environment for the overlap between different respondents' personal networks to be negligible. This implies that it is legitimate to treat the different personal networks in the data set as independent. The structure of such a data set is nested or clustered, i.e., relations are nested within respondents.

One might object that relations are dyadic in nature, and therefore not really nested within individuals. We argue that the nesting structure holds for the data consisting of the non-overlapping personal networks of a sample of respondents. However, we do not imply that it holds for the (unobserved) global network in which these personal networks are embedded. A complementary point of view is that the personal networks in its entirety is a complex attribute of the individual. Regarded from this viewpoint, relations are clustered within egos (but not independent between egos for a population in which the personal networks overlap).

An interesting article about the statistical analysis of personal networks where data for relations between alters are available is Wellman, Frank, Espinoza, Lundquist, and Wilson (1991). In their model, alters are characterized by a categorical variable while the composition of the personal

network in terms of these categories is modeled as well as the frequencies of ties between the various types of alters. The present paper addresses a different question: it takes the personal network as given, and presents a method for studying a characteristic of the relations between ego and alters.

If the dependent variable is an attribute of the relations, then one possibility to simplify the question of how to analyse the data is to aggregate the characteristics of the personal network to the level of the focal individual and use this aggregate variable as the dependent variable in an OLS regression analysis. However, such an aggregation implies an important loss of information. Furthermore, the variable number of relations per individual can make it difficult to define an aggregate variable (e.g., a mean) that is well comparable across individuals. It is often desirable to retain the two conceptual levels in the analysis, and analyse the personal network data as *relationships nested within individuals*. The purpose of this paper is to show how this can be achieved by the use of the hierarchical linear model of multilevel analysis.

Technically as well as conceptually, it will be useful to distinguish variables according to the level at which they are defined. The level of the relation is called the first, or lowest level, while the level of the individual respondent is called the second, or higher, level. Variables that are functions of relations, alters, or dyads of related individual are called level-1 variables whereas variables that are functions of individual respondents are called level-2 variables. An analysis that jointly considers these two levels needs to take into account differences between individuals as well as differences between relationships with individuals.

The *hierarchical linear model* is a statistical model that is suitable for a nested data structure, in the case where the dependent variable is a level-1 variable. This model is treated in the textbook by Bryk and Raudenbush (1992). It has been used in many studies with a multilevel character. Bryk and Raudenbush (1992) give many examples of such multilevel studies, in fields such as education (pupils within classes within schools) and organizational sociology (individuals within organizations). The hierarchical linear model can be used as a kind of regression analysis for personal network data if the following conditions apply:

- the dependent variable is at the lowest level, i.e., the level of the relation or the alter;
- the data contains no overlap of personal networks of different egos, or at least this overlap is negligible;
- the data obtained from different egos (respondents) is mutually independent.

In contrast to the dependent variable, the explanatory variables can be variables at the first as well as at the second level. One of the attractive features of the hierarchical linear model is that the effect of first-level variables on the dependent variable may be variable across respondents. Part

of this variability may be explained by second-level variables and part may remain unexplained (modeled as random variability). An example will be given below.

Some basic characteristics of the hierarchical linear model are treated in Section 3. The use of this model for analysing personal networks is illustrated with an investigation into relationships of cocaine users in Rotterdam, using data collected by means of a snowball sample (Intraval 1992; Bieleman *et al.* 1993).

2. PERSONAL NETWORKS

Network data must, by their nature, refer to a *population of individuals* jointly with a *relationship* defined in this population. We denote that population by $\mathcal{P}$ and the relationship by $\mathcal{R}$. This relationship may be directed or undirected (symmetric). The variables in the data can refer to individuals (e.g., age) but some of the variables may refer to relations; e.g., the length of time that the relationship has been in existence. For the data collection, it is necessary to have an inclusion criterion for being a member of population $\mathcal{P}$ as well as an inclusion criterion for being related according to $\mathcal{R}$. These two inclusion criteria define the total network of interest. The inclusion criterion for $\mathcal{R}$ may be defined as just knowing the nominee (e.g., using some practical inclusion criterion for $\mathcal{P}$, "mention all individuals whom you know and who are members of $\mathcal{P}$") but also as a more specific characteristic (e.g., "mention all individuals who are a member of $\mathcal{P}$ and whom you know on a first-name basis") or by a so-called name generator (see, e.g., Fischer 1982). In the research presented below, $\mathcal{P}$ is the population of all adult inhabitants of Rotterdam who used cocaine at least 25 times in their life and/or at least 5 times in the last half year; $\mathcal{R}$ is the directed (i.e., not necessarily symmetric) relationship defined by the inclusion criterion that the respondent knows the nominee by name, knows that the nominee satisfies the inclusion criterion for population $\mathcal{P}$, and believes that the nominees also knows her/him by name.

This paper is concerned with methods for the analysis of some aspect of the relations, expressed by a *dependent variable*, denoted Y, that is a function of related pairs: for every ordered pair (j, i) of individuals in $\mathcal{P}$ who are related according to $\mathcal{R}$, the value Y_{ij} must be defined. The explanatory variables may be functions of the relations but also of the two individuals involved.

A sample of personal networks may be regarded as a particular kind of sample from the network, where the sampling acts on the individuals in $\mathcal{P}$. Each sampled person, referred to as a *respondent*, is asked to mention all other person, referred to as *nominees*, who belong to population $\mathcal{P}$ and are related to the respondent according to $\mathcal{R}$. Furthermore, each sampled

person can give information about his nominees, and about his relationships with them.

Nominees mentioned by the same respondent will usually be more alike than nominees mentioned by different respondents. Their likeness can be caused by a direct interaction among the nominees or by their interaction with the same respondent. Also, their social environment will usually show a greater degree of overlap than the environment of nominees mentioned by different respondents. Social environment and social interactions can have an effect on the dependent variable Y. This implies that variables Y_{ij} and Y_{kj}, referring to two nominees i and k of the same respondent j, will usually be correlated, whereas variables Y_{ij} and $Y_{kj'}$ for $j \neq j'$, referring to two nominees of different respondents, are uncorrelated.

Since the respondents may mention several relations according to the defined inclusion criterion, these relations are *nested* within respondents. In accordance with the terminology of multilevel research (Bryk and Raudenbush 1992), we denote the respondents as *units at level two* and the relations as *units at level one*. We assume that the population is so large that it will rarely happen that a respondent is mentioned as a nominee by another respondent, or that two respondents mention the same person as a nominee; such events will be neglected in the analysis. This assumption implies that it is permissible to treat nominees interchangeably with respondent-nominee relations as the units at level one.

3. MULTILEVEL MODELS FOR PERSONAL NETWORKS

The multilevel model presented in this paper is the *hierarchical linear model (HLM)* treated in detail by Bryk and Raudenbush (1992). This model is a variant of multiple linear regression analysis for data with a hierarchical nesting structure where the dependent variable is defined at the lowest level of the hierarchy. We discuss briefly the essential aspects of this model in the context of personal network data. Respondents (level 2 units) are indicated by j and nominees, or relations with nominees (level 1 units), by i. A crucial aspect of the HLM is that the magnitude of the effect of explanatory ("independent") variables X on the dependent variable Y may differ between respondents. A simple two level model for the effect of X on Y can be formulated as a regression model with coefficients that differ between respondents:

$$(1) \qquad Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + R_{ij},$$

where

Y_{ij} is the value of the dependent variable,

β_{0j} is the respondent-specific intercept,

β_{1j} is the respondent-specific regression slope,

x_{ij} is the value of the explanatory variable,
R_{ij} is the unexplained part ("residual") of the dependent variable Y_{ij}.

In equation (1), Y and R are indicated by capital letters and x by a small letter because the former two variables are considered to be random variables, while the model is conditional on the value of X, i.e., the observed value x is treated as a given non-random value. In the research presented below where $\mathcal{P}$ is a population of cocaine users, the dependent variable Y is a measure for the significance of cocaine use in the relationship; a possible explanatory variable X is the duration of the relationship. For each respondent, the regression equation of Y on X may be different. This is reflected by the fact that in (1), the intercept β_{0j} as well as the regression coefficient β_{1j} depend on the respondent j. If the effect of relationship duration on the significance of cocaine use is positive, but slight for some individuals and strong for others, then the former individuals have a small positive value for β_{1j} and the latter a high positive value.

A characteristic feature of the HLM is the assumption that β_{0j} and β_{1j} are random variables: the HLM is a *random effects model*. It differs from the random effects models that have been extensively used by econometricians. The main difference is that in the HLM not only regression intercepts, but also regression slopes can be random. The HLM being a random effects model means that the analysis does not focus on individual values of the coefficients β_{0j} and β_{1j}. Instead, it focuses on the *population* of these coefficients that tells something about the *population* $\mathcal{P}$. It is convenient to split the coefficients β_{0j} and β_{1j} in (1) into a fixed part (the mean) and a random part (with mean 0):

$$(2) \qquad \begin{aligned} \beta_{0j} &= \gamma_{00} + U_{0j}, \\ \beta_{1j} &= \gamma_{10} + U_{1j}, \end{aligned}$$

where

γ_{00} is the population mean of the intercepts,
γ_{10} is the population mean of the regression coefficients,
U_{0j} is the respondent specific part of the intercept,
U_{1j} is the respondent specific part of the regression coefficient.

Since the population means are split off, the means of U_{0j} and U_{1j} are 0. The residual at level 1, R_{ij}, is assumed to be statistically independent of the random effects at level 2, U_{0j} and U_{1j}. It is usually assumed that the latter variables have a bivariate normal distribution. This implies that the distribution is characterized by the variances $\mathrm{var}(U_{0j})$, $\mathrm{var}(U_{1j})$ and the covariance $\mathrm{cov}(U_{0j}, U_{1j})$. If $\mathrm{var}(U_{0j}) = \mathrm{var}(U_{1j}) = 0$, the coefficients are the same over all respondents and an OLS regression analysis could be performed, without taking into effect the clustering of relations within respondents. If $\mathrm{var}(U_{1j}) = 0$, then the regression coefficients do not vary, but the intercepts may be variable. In the example, the mean level of frequency of

joint cocaine use then does vary over respondents, but the effects of duration of the relationship on frequency of joint use is the same for all respondents. In the general case, $\text{var}(U_{0j})$, $\text{var}(U_{1j})$ and $\text{cov}(U_{0j}, U_{1j})$ are free parameters that are estimated from the data.

The three variances $\text{var}(R_{ij})$, $\text{var}(U_{0j})$, and $\text{var}(U_{1j})$ measure different kinds of unexplained variation in the model. The parameter $\text{var}(R_{ij})$ measures unexplained variation between nominees within respondents and $\text{var}(U_{0j})$ measures unexplained variation between respondents, more specifically, respondents' main effects. The parameter $\text{var}(U_{1j})$ measures unexplained interaction between respondents and variable X. In OLS regression analysis there is only one variance parameter and the analyst tries, by using explanatory variables, to obtain a small value for this residual variance. In the multilevel model, one tries in a similar way to obtain small values for all variance parameters. Typically, inclusion of explanatory variables that are defined at the lowest level (functions of relations of nominees) will diminish $\text{var}(R_{ij})$ and $\text{var}(U_{0j})$; inclusion of explanatory variables at the higher level (functions of respondents) will diminish $\text{var}(U_{0j})$; inclusion of interactions between X and higher-level variables will diminish $\text{var}(U_{1j})$. However, this holds not exclusively or necessarily. An extensive discussion of this point is given in Snijders and Bosker (1994). Note that the last-mentioned type of interactions are interactions between relationship-level and respondent-level variables. Effects explainable by such interactions are called cross-level interaction effects.

The following formulae express the use of respondent-level variables to diminish the unexplained variances $\text{var}(U_{0j})$ and $\text{var}(U_{1j})$. A respondent-level variable W_j is used to explain part of the differences between the intercepts β_{0j} as well as between the regression coefficients β_{1j}. This means that W_j is used as an explanatory variable for β_{0j} and β_{1j}, which leads to an expansion of equations (2). Since the value of variable W_j is regarded as a given value and not as a random variable, it is indicated by a small letter w_j.

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} w_j + U_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} w_j + U_{1j}. \end{aligned} \tag{3}$$

Parameters γ_{00}, γ_{01}, γ_{10} and γ_{11} are called fixed coefficients. Combining equations (1) and (3) yields

$$Y_{ij} = \gamma_{00} + \gamma_{01} w_j + \gamma_{10} x_{ij} + \gamma_{11} w_j x_{ij} + U_{0j} + U_{1j} x_{ij} + R_{ij}. \tag{4}$$

The first four terms in the right-hand side of (4) are the fixed part of the model, which is just like the fixed part in an OLS regression model. The last three terms constitute the random part of the model. The fixed part contains a main effect for w_j due to the use of w_j to explain the intercepts β_{0j}. It also contains the cross-level interaction effect for $w_j x_{ij}$ due to the use of w_j to explain the regression coefficients β_{1j}.

The random part of the model, $U_{0j} + U_{1j}x_{ij} + R_{ij}$, makes it possible to separate distinct components of variance. The first two components are attributable to the respondents and the last one to the relations. The presence of the respondent-level random effects U_{0j} and U_{1j} results in correlations between nominees of the same respondent; if $\text{var}(U_{1j}) > 0$ these correlations depend on the values for x_{ij}.

Model (4) can be extended by using more than one relationship-level and more than one respondent-level explanatory variable. It is not necessary that all relationship-level variables with a fixed effect also have a random effect. However, it is a questionable practice to include variables with a random effect that do not also have a fixed effect. (If some variable X were to be included with a random but without a fixed effect, the researcher would have to explain why a model is plausible in which the effects of X on Y, non-zero for practically all respondents, yet average out to a precisely zero effect over the whole population.) Furthermore, one may use different respondent-level variables to explain different regression coefficients.

To test whether parameters are zero, the fixed coefficients must be treated differently from the variance or covariance parameters. Testing the fixed coefficients is straightforward. To test the null hypothesis that some coefficient γ equals 0, a t-statistic

$$t = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

can be used, where $\hat{\gamma}$ is the estimated coefficient and $SE(\hat{\gamma})$ its standard error. This statistic does not have an exact t-distribution under the null hypothesis, so the number of its degrees of freedom is not clearly defined. Some authors advise to use a liberal standard normal distribution. This is satisfactory for testing effects of level-one variables. For testing main effects of level-two variables, one may use a t-distribution with $J - k - p - 1$ degrees of freedom, where J is the number of level-two units, k the number of fixed coefficients, and p the number of parameters for the random part of the model at level 2. We will describe the process of testing random effects, that is, testing variance parameters, in section 5.

Several software packages for estimation of multilevel models are available, each with its own advantages and disadvantages. ML3 (Prosser, Rasbash, and Goldstein 1991), VARCL (Longford 1988), and HLM (Bryk, Raudenbush, Seltzer, and Congdon 1988) can all be used for estimating and testing.

The described model can be further extended in several ways. Three-level models are discussed by Goldstein (1987) and Bryk and Baudenbush (1992). For the analysis of personal networks, three-level models can be relevant when respondents are arranged in groups (e.g., schools, companies, neighbourhoods) between which unexplained variation may exist. The groups will

then constitute a higher (third) level. In other cases, it may be relevant to add a lower level under that of the relationship, e.g. time, when repeated measurements are made on the dependent variable. Another extension, discussed in Goldstein (1987), is to allow for heteroscedasticity of the random effects.

4. DESIGN OF THE COCAINE USERS RESEARCH PROJECT

Multilevel analysis of personal networks is illustrated here with data obtained in a research project on cocaine use in Rotterdam (Intraval 1992; Bieleman *et al.* 1993). Cocaine users form a *hidden population* (cf. Spreen 1992) because a sampling frame is missing, respondents are difficult to locate and to identify, and respondents are reluctant to cooperate. A frequently applied data collection approach in studies of hidden populations is snowball sampling: interviewees are asked to mention other population members known to them, who then are interviewed in their turn. This procedure provides the researcher with additional population members as well as with data about the personal networks of the interviewees. The reason for using a snowball approach in these studies is that it often seems the only possible way to collect a reasonable amount of respondents. Snowball samples have proved to be rather successful in studies of populations of drug users (Biernacki and Waldorf 1981; Morrison 1988). However, subsequent analyses often are restricted to non-structural descriptions of the respondents and their environment (an exception is Johnson 1990). An important reason why snowball samples are successful for locating potential respondents in studies of, for example, urban drug users populations is the fact that such populations can be considered to be large social networks. Their members are often tied to one another through direct or indirect social links. Coleman (1958) had this in mind when introducing snowball sampling. He intended that this sampling method be used for sampling with reference to the social structure, so that different aspects of network structure could be analysed in large networks.

In the Rotterdam cocaine research project, a snowball sampling procedure was used for two purposes. First, to get a reasonable number of users in a relatively economic way, and second to study relationships in which cocaine plays a role. The respondents were asked to mention other cocaine users, to a maximum of 50. The inclusion criteria for the population $\mathcal{P}$ and for the relation $\mathcal{R}$ were mentioned above in section 2.

Respondents had to assign users known to them to five pre-defined circuits, in which the nominee used cocaine most frequently. These circuits were: the world of entertainment, work, home or private parties, sport or hobbies clubs, and the hard drugs scene. The circuits had been defined on

the basis of a pilot field study. Each respondent could mention a maximum of 10 other users per circuit. There were two purposes to this use of circuits. First, to "force" the respondents to nominate some users in the periphery of their network, and second to get some insight into the relations. Furthermore, from each respondent two nominees per circuit were selected at random and more detailed questions about these relationships were asked. In this way, each respondent provided information about at most ten relationships.

Due to the nature of this specific hidden population, an initial random sample was not possible. We refer to Intraval (1992) and Bieleman *et al.* (1993) for the specification of how initial respondents were obtained, and the methods used to make this initial sample "well spread" over the population of cocaine users in Rotterdam. The size of the population (number of cocaine users according to these inclusion criteria) was estimated by a method developed by Frank and Snijders (1994) at about 12,000.

The use of multilevel models for the analysis of personal networks is illustrated by exploring which factors contribute to the significance of cocaine in relationships for the respondents. For other multilevel analyses in the same study we refer to Spreen (1993). The degree of significance of cocaine in relationships is important. Relations where this degree of significance is high will be particularly important for determining the extent to which the individuals's social personality is dominated by his/her cocaine use, and the increase or decrease in the frequency of his/her cocaine use. A more extensive discussion is presented in Bieleman *et al.* (1993). The significance of cocaine use in relationships, as evaluated by the respondents, is indicated by Y_{ij}. This variable is measured by an index which takes two aspects of a cocaine relationship into account These are the role played in the relationship by the procurement of cocaine and, second, the frequency of the respondent meeting the nominee in a context of cocaine use. The significance of cocaine in a relationship of respondent j with nominee i is considered to be high if respondent j meets nominee i always or usually in a context of cocaine, and if respondent j buys or obtains cocaine from nominee i or vice versa. The next diagram shows the different combinations of these two aspects, and their value on the index Y_{ij}. The index is assumed to form an ordinal measurement scale.

The total number of relationships in the sample is 269, nested within 60 respondents: 39 initial respondents and 21 extensions (nominees of initial respondents). Only 3 of these extensions were found in the second wave of the snowball sample. The maximum number of 10 nominees per circuit, which the respondents were allowed to mention, was found to be a restriction for 17 respondents in 1 circuit and for 4 respondents in 2 or 3 circuits. It was a restriction for the entertainment circuit 14 times, 9 times for the home circuit and 3 times for the work circuit.

The distribution of the dependent variable Y_{ij} is given in Figure 1.

Value	Description
0	The contact between respondent *j* and nominee *i* never takes place in a context of cocaine use.
1	The contact between respondent *j* and nominee *i* sometimes takes place in a context of cocaine use. There has not been any relation between respondent *j* and nominee *i* with respect to the procurement of cocaine in the last six months.
2	The contact between respondent *j* and nominee *i* usually takes place in a context of cocaine use. There has not been any relation between respondent *j* and nominee *i* with respect to the procurement of cocaine in the last six months.
3.	The contact between respondent *j* and nominee *i* sometimes takes place in a context of cocaine use. Respondent and nominee take turns to obtain cocaine, or they buy it jointly from a third person.
4	The contact between respondent *j* and nominee *i* sometimes takes place in a context of cocaine use. Either the respondent or the nominee usually procures cocaine for the other.
5	The contact between respondent *j* and nominee *i* usually takes place in the context of cocaine use. Respondent and nominee take turns to obtain cocaine or buy it jointly from a third person.
6	The contact between respondent *j* and nominee *i* usually takes place in the context of cocaine use. Either the respondent or the nominee usually procures cocaine for the other.

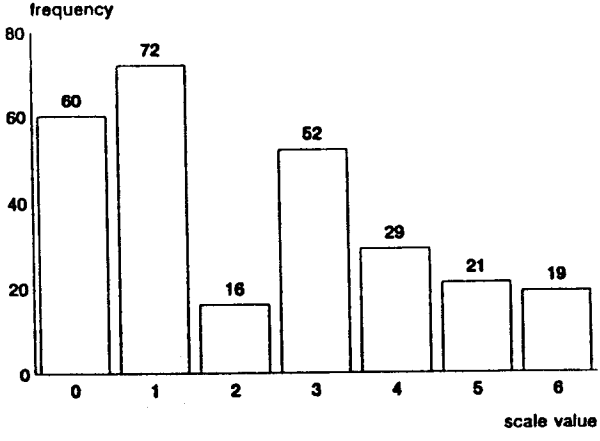

Fig. 1. Frequencies of Y_{ij}.

The explanatory variables used in the analysis are:
- the main circuit of use of the respondent (5 categories),
- the main circuit of use of the nominee (5 categories),
- the total number of people the respondent provided information on,
- the duration of the relationship,
- the gender of the respondent,
- the gender of the nominee,
- the way of use of the respondent (3 categories; see below), and
- extension: whether or not the respondent was obtained as a nominee of another respondent.

The way of using cocaine is coded into 3 categories: (1) snorting, (2) a single other way of use like injecting, basing, smoking etc., and (3) two or more of these methods. Since some of the respondents were obtained as initial respondents in a sample resembling a random sample, whereas other respondents were obtained as nominees by initial respondents, the assumption that the respondents are obtained as a random sample is not quite satisfied. The last variable in the list above is used in order to see whether there is a systematic difference between the first and the second group of respondents. It is scored as 1 (obtained as nominee) or 0 (obtained as initial respondent).

The models mentioned in the next section were estimated using ML3 (Prosser, Rasbash, and Goldstein 1991).

5. MODEL SELECTION

Several procedures to model selection are possible. The approach of forward modeling consists of starting with an empty model and adding effects. Backward modeling starts with a full model (all variables included) and successively removes unimportant effects step by step.

Both methods have their disadvantages. The forward approach has the risk of missing important effects because of masking effects. The backward approach suffers from the drawback that it can be extremely time consuming. We started with a model without any explanatory variables in order to obtain the ratio of between-respondent to within-respondent variability. After that we added main effects of all available theoretically important variables. The model was further refined in an exploratory way by excluding non-significant effects and testing some plausible interaction effects. We were not quick to exclude variables with non-significant effects but preferred to retain them in later model fits, because their exclusion might mask effects of other variables.

A question about the form of the model is, whether or not it should be required that attributes of the respondent and the same attributes of the nominee (e.g., gender) have identical effects. If the respondent and the

nominee would have occupied identical positions with regard to how the dependent and explanatory variables are defined, this requirement would have been natural. In this research, however, asymmetry was present because the dependent variable and the information about the nominee's main circuit of use were obtained from interviewing the respondent. Therefore, such a requirement of respondent-nominee symmetry was not made.

In the following, the search for a satisfactorily fitting model is presented in a few steps. The interpretation of the resulting estimates is deferred till after the presentation of the final model.

The first model estimated is the model with only a random intercept,

$$(5) \qquad Y_{ij} = \gamma_{00} + U_{0j} + R_{ij}.$$

Denote $\text{var}(U_{0j}) = \sigma_0^2$ and $\text{var}(R_{ij}) = \sigma^2$. Expression (5) is identical to a one-way random effects ANOVA; it is also called "empty model" because it contains no explanatory variables and doesn't explain anything. This model is useful because it shows how much variability exists at each level. A comparison of more complex estimated models with this model makes it possible to gain an insight in the explanatory power (explained proportion of variance) of estimated models (Snijders and Bosker 1994).

The proportion of variance due to differences between respondents can be expressed by

$$(6) \qquad \rho = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2},$$

the intraclass correlation coefficient. The results of the first model fit are presented in Table 1. This table also includes the deviance, defined as minus twice the log-likelihood, which is a measure of badness of fit. The deviance will be used below for model testing.

The estimates $\hat{\sigma}^2 = 2.71$ and $\hat{\sigma}_0^2 = 0.84$ yield an intraclass correlation coefficient $\hat{\rho} = 0.24$, which is substantial. This confirms the expectation that there exists a correlation between relationships of the same respondent;

TABLE 1
Estimates for one-way random effects ANOVA.

Fixed effect	Coefficient	S.E.
Intercept	2.14	0.16
Random effect	**Variance component**	**S.E.**
$\sigma_0^2 = \text{var}(U_{0j})$	0.84	0.29
$\sigma^2 = \text{var}(R_{ij})$	2.71	0.27
Deviance	1049.02	

this correlation implies that using OLS rather than a multilevel approach could result in erroneous conclusions.

Since the data were obtained by a snowball sample, the next variable we included in the model was the extension variable in order to determine if a selection effect, due to the different waves in the sample, was observable. The addition of this explanatory variable to the first model did not result in a significant effect. However, in analyses of other dependent variables (reported in Spreen 1993) the effect of the extension variable appeared to be masked due to interaction effects, so this variable was included in some of the later models.

The next variables added to the model were the main circuits in which the respondents and their nominees used cocaine, the gender of the respondent and of the nominee, and the way of use. All these variables were coded as dummies. The reference category for the circuits was the entertainment circuit. Since the number of respondents from the sport circuit was very low, while the estimated effect differed little from the effect for the entertainment circuit, the dummy for respondents from the sports circuit was omitted from the model. Furthermore, it turned out that the effect of "way of use" differentiated only between the categories 1 and 2 (one way of use) on one hand, and 3 (multiple ways of use) on the other hand. Therefore only the dummy for multiple ways of use was retained. The gender of the respondent did not have a significant effect, and therefore also was omitted from the model.

In a second step, the composition of the personal network was taken into account by adding, as explanatory variables, the fractions of nominees mentioned by the respondent in the different circuits. It turned out that of these composition variables, only the fraction of nominees mentioned in the sport circuit had a significant effect. This may be related to the fact that hardly any respondents mentioned the sport circuit as their own main circuit of use. This supports a conclusion that was drawn from qualitative analyses of the field work, viz., that the importance of the sport circuit consists especially in the fact that it is a secondary circuit of use for certain cocaine users. The estimates are given as model 2 in Table II.

As a next step, it was investigated whether extra random effects give an improvement of the model fit. Recall that a random effect for a variable X means that the effect of X on Y varies between respondents. It turned out that including a random effect for the dummy variable that indicates whether nominees are in the hard drugs circuit, results in a better model fit. This means that the difference between Y values for relationships with nominees in the hard drugs circuit on one hand and with nominees in other circuits on the other hand, differs between respondents.

The statistical significance of the improvement in fit due to the inclusion of extra parameters in the random part of the model can be tested as follows. Suppose a model M_0 has been estimated with N parameters (vari-

TABLE II
Models 2 and 3.

Fixed effects	Model 2		Model 3	
	Coefficient	S.E.	Coefficient	S.E.
Intercept	2.13	0.24	1.96	0.23
Main respondent circuit:				
Entertainment/sport	0	0	0	0
Work	1.67	0.52	1.85	0.38
Home	–0.80	0.29	–0.49	0.26
Hard drugs	0.94	0.54	1.52	0.37
Main nominee circuit:				
Entertainment	0	0	0	0
Sport	–1.34	0.43	–1.45	0.41
Work	–0.09	0.29	–0.14	0.28
Home	0.47	0.26	0.45	0.25
Hard drugs	–1.33	0.39	–1.51	0.46
Other variables:				
Female nominee	–0.55	0.25	–0.53	0.24
Multiple ways of use	1.15	0.59	1.40	0.38
Fraction of nominees in sport circuit	2.45	0.95	3.06	0.91
Random effects	Variance component	S.E.	Variance component	S.E.
Level two random effects:				
Random intercept var(U_{0j})	0.33	0.17	0.53	0.21
Hard drugs nominee var(U_{1j})			3.71	1.49
cov(U_{0j}, U_{1j})			–1.55	0.50
Level one variance:				
Residual R_{ij}	2.41	0.24	2.19	0.22
Deviance	996.67	984.71		

ances and covariances) for the random effects and certain fixed effects. Now a second model M_1 is estimated with $N + k$ parameters for the random effects, and the same set of fixed effects as model M_0. The null hypothesis is that model M_0 is valid, i.e., the additional parameters for random effects in model M_1 are 0. Denote the deviance of model M_0 by D_0 and the deviance of model M_1 by D_1. Under the null hypothesis that model M_0 holds, the difference $D_0 - D_1$ has (approximately, for large sample sizes) a χ^2-distinction with k df. For instance, model 2 has 10 fixed coefficients, 2 variance parameters, and a deviance of 996.67. Model 3 has fixed coefficients for the same variables, one variance and one covariance parameter extra, and a deviance of 984.71. This results in model M_0 having

2 parameters more than model 2 while the difference in deviance is 11.96. Testing this difference in the χ^2-distribution with 2 degrees of freedom shows that it is a highly significant ($p < 0.005$) value.

Some of the estimated regression coefficients differ considerably between models 2 and 3. In model 3, almost all standard errors are smaller than in model 2. Since model 3 fits significantly better than model 2, the results of model 2 are less trustworthy than those of model 3. The attentive reader may have noticed that in the estimates for model 3, the correlation $\rho(U_{0j}, U_{1j})$ is less than -1. This is possible in estimates obtained from ML3. Although a correlation of -1 is incompatible with the model expressed by (4), ML3 allows such estimates when the corresponding covariance matrix for the observations Y_{ij} is positive semi-definite. We return to this when interpreting the parameter estimates of the final model.

To improve model 3, the procedure followed was explorative in nature. Several models with more fixed and random effects were estimated. Some of the fixed effects were interaction effects, for example, the extension variable was re-included as a main effect and as a part of interactions. Since the extension variable was indeed significant in interaction with other variables but the main effect of the extension variable was very close to 0, this main effect was omitted in the final model. The resulting model is presented in Table III. The variable "extension respondents" is a dummy variable indicating that the respondent was obtained as a nominee from an earlier respondent. "Home $\Rightarrow$ home" is a dummy variable which indicates the interaction effect of the combination of a respondent in the home circuit with a nominee also in the home circuit.

6. INTERPRETATION OF THE FINAL MODEL

6.1. *Fixed coefficients*

It is apparent from the estimates of the final model that, contrary to what might have been expected, attributes of the respondent and the nominee do not have similar effects. This asymmetry could not be removed by taking into account interactions between respondent attributes and nominee attributes. As an example, the respondent's gender has no significant effect whereas in relationships with female nominees, cocaine is reported (by male as well as female respondents) as less important than in relationships with male nominees.

As to the effect of respondent and nominee circuits, note that only the relative and not the absolute values of the effects should be interpreted. Cocaine has the highest significance in relations for those respondents who primarily use cocaine in the work and hard drugs circuits, while it has a lower significance for respondents who mainly use cocaine in the

TABLE III
Final model.

Fixed effects	Coefficient	S.E.
Intercept	1.89	0.22
Main respondent circuit:		
Entertainment/sport	0	0
Work	1.93	0.29
Home	0.10	0.26
Hard drugs	1.79	0.29
Main nominee circuit:		
Entertainment	0	0
Sport	−1.46	0.41
Work	−0.15	0.27
Home	0.59	0.29
Hard drugs	−1.68	0.43
Other variables:		
Female nominee	−0.59	0.24
Multiple ways of use	1.24	0.30
Fraction of nominees sport circuit	3.51	0.85
Interactions:		
Home ⇒ home	−1.01	0.53
Extension respondent in home circuit	−1.49	0.54
Extension respondent in home circuit ⇒ nominee in home circuit	1.71	0.74
Random effects	**Variance component**	**S.E.**
Level two random effects:		
Random intercept var(U_{0j})	0.56	0.21
Hard drugs nominee var(U_{1j})	4.08	1.43
Cov(U_{0j}, U_{1j})	−1.77	0.50
Level one variance:		
Residual R_{ij}	2.14	0.22
Deviance	975.64	

entertainment, sport, and home circuits. On the other hand, for all relations (regardless of the circuit of the respondent) cocaine has the highest significance in relations with nominees in the home circuit and the lowest in relations with nominees in the sport and hard drugs circuits. These differences between respondent and nominee circuits may be caused by the sample design. Because each respondent is "forced" to mention some users in the periphery of his personal network, he is likely to mention well-known, "public" users although he has no personal relations with them. This situation corresponds with $Y_{ij} = 0$. The negative effect of relations being

with hard drugs circuit nominees is probably due to a combination of generally known "open" cocaine users and superficial relationships in the environment of non hard drugs respondents.

Another significant fixed effect in the model is the way respondents take cocaine. Respondents who use cocaine in multiple ways, report a higher significance of cocaine in their relations than respondents who have only one single way of intake of cocaine.

Concerning the fraction of nominees in the sports circuit, it has already been noted that this circuit can be considered a secondary circuit for certain cocaine users. The high coefficient shows that users for whom the sports circuit is a primary or secondary circuit report a high importance of cocaine in their relations.

The three observed interaction effects all involve the home circuit. The importance of cocaine in relations involving the home circuit seems to be determined in different ways than in relations outside the home circuit. For relations of respondents who use cocaine mainly at home or at private parties and mention other users within their own circuit, we found a less significant role of cocaine. We also found a less significant role of cocaine for home circuit respondents who were nominees of an earlier respondent. However, the relations of these extension respondents with others within the home circuit showed a more significant role of cocaine. At first sight, these effects seem contradictory. Recalling that the extension variable was included to take the sampling design into account, they may be due to "snowball effects" (Spreen 1993). In the initial stage of the snowball sample most initial home circuit respondents were found in other circuits, because the fieldworkers had no access to this circuit. Respondents in the initial sample classified in the home circuit were later assigned to this circuit because during the analysis of the interviews, it appeared to be their main circuit of use. In other words these initial "home" respondents were found in other, "better accessible" circuits in the initial stage of the snowball sample. Persons who use cocaine mainly in the home circuit are under-represented, because this circuit is difficult to access. However, the extension respondents in the home circuit are probably the more typical users at home or at private parties. As a group, they assign to cocaine a relatively low significance in their general relations with other cocaine users, but in their relations within their own circuit they report a higher significance of cocaine. A qualitative analysis of the interviews suggests that these persons are users who primarily use cocaine in their own circuit although they do know other users outside the home circuit. These effects show that by using a snowball sample to get access to this hidden population, individuals are indeed reached who are less likely to be identified in most other practical ways of obtaining respondents.

with hard drugs circuit nominees is probably due to a combination of generally known "open" cocaine users and superficial relationships in the environment of non hard drugs respondents.

Another significant fixed effect in the model is the way respondents take cocaine. Respondents who use cocaine in multiple ways, report a higher significance of cocaine in their relations than respondents who have only one single way of intake of cocaine.

Concerning the fraction of nominees in the sports circuit, it has already been noted that this circuit can be considered a secondary circuit for certain cocaine users. The high coefficient shows that users for whom the sports circuit is a primary or secondary circuit report a high importance of cocaine in their relations.

The three observed interaction effects all involve the home circuit. The importance of cocaine in relations involving the home circuit seems to be determined in different ways than in relations outside the home circuit. For relations of respondents who use cocaine mainly at home or at private parties and mention other users within their own circuit, we found a less significant role of cocaine. We also found a less significant role of cocaine for home circuit respondents who were nominees of an earlier respondent. However, the relations of these extension respondents with others within the home circuit showed a more significant role of cocaine. At first sight, these effects seem contradictory. Recalling that the extension variable was included to take the sampling design into account, they may be due to "snowball effects" (Spreen 1993). In the initial stage of the snowball sample most initial home circuit respondents were found in other circuits, because the fieldworkers had no access to this circuit. Respondents in the initial sample classified in the home circuit were later assigned to this circuit because during the analysis of the interviews, it appeared to be their main circuit of use. In other words these initial "home" respondents were found in other, "better accessible" circuits in the initial stage of the snowball sample. Persons who use cocaine mainly in the home circuit are under-represented, because this circuit is difficult to access. However, the extension respondents in the home circuit are probably the more typical users at home or at private parties. As a group, they assign to cocaine a relatively low significance in their general relations with other cocaine users, but in their relations within their own circuit they report a higher significance of cocaine. A qualitative analysis of the interviews suggests that these persons are users who primarily use cocaine in their own circuit although they do know other users outside the home circuit. These effects show that by using a snowball sample to get access to this hidden population, individuals are indeed reached who are less likely to be identified in most other practical ways of obtaining respondents.

6.2. *Random effects*

In order to explain the random part of the model, we calculate the consequences of the parameter estimates for the residual (i.e., unexplained) variances and correlations of the dependent variable Y_{ij}. The unexplained part of Y_{ij} for a relationship with a nominee outside the hard drugs circuit is $U_{oj} + R_{ij}$ while the unexplained part for a relationship with a nominee in the hard drugs circuit is $U_{oj} + U_{1j} + R_{ij}$.

The estimated variances and covariance in Table III imply (using the usual rules for computing the variance of a sum of random variables) that these residuals have quite similar variances of 2.7 and 3.2, respectively. Correlations between residuals for two different nominees of the same respondent are as follows:

- two nominees outside the hard drugs circuit: 0.21;
- two nominees in the hard drugs circuit: 0.34;
- one nominee outside and one nominee in the hard drugs circuit: −0.41.

For intra-class correlations, these are substantial. (Note that, although the correlation between U_{0j} and U_{1j} is less than −1, this nevertheless produces residual correlations for the dependent variable that have an acceptable value.) The interpretation is that, in addition to the fixed effects already taken into account in the model of Table III, there are influences on Y that lead to an unexplained standard deviation of about 1.7. This points to an increased importance of cocaine in relations either with individuals outside the hard drugs circuit, or with individuals in the hard drugs circuit, but usually not both. In other words, in the explanatory variables we still miss a variable that represents an influence differentiating between a higher importance of cocaine in relations with hard drug users versus relations with non-hard drug users.

7. CONCLUDING REMARKS

By applying multilevel analysis to the analysis of personal networks, the effects of the two conceptual levels inherent in each personal network can be simultaneously analysed. This can lead to better insight into relational structures within the personal networks. The regression coefficients at the ego-level can be considered to characterize the relational patterns of the egos. For example, in the Rotterdam cocaine research project, the respondents of the hard drugs circuit assigned in their relationships a higher significance to cocaine than the other respondents. The regression coefficients at the relation-level can be considered to characterize relational structures of all relations. For instance, in the total sample of relations, relations with female users showed a lower significance of cocaine. An advantage of the multilevel technique is the possibility of defining cross-

level interactions. In this way, more precise influences of the ego-level on the relationship-level can be analysed. For instance, the cross-level interaction effect of respondents using cocaine mainly at home or at private parties with nominees using mainly in the same circuit showed that relations between such cocaine users tend to have a lower significance of cocaine. By using random effects the unexplained variance can be analysed more precisely. In this way, we were able to give an indication for an important explanatory variable that is missing in the present data set. Finally, by using the extension variable we were also able to model some of the effects of the snowball sampling design.

It should be noticed that the presented research could also be considered as a three level design. Some of the respondents were not found independently of each other. This happened mainly because they were jointly referred to by other respondents as people willing to cooperate with the research. These clusters of respondents may be seen as level three units. We neglected this overlap between personal networks and preferred to model this design with two levels. A three level model would have been harder to interpret because of its more complicated random effects structure.

Introducing a network approach and subsequent multilevel modeling of the personal networks in studies of hidden populations like cocaine users provides a useful way of investigating these difficult populations. More extensive analyses of the cocaine research project in Rotterdam are given in Bieleman *et al.* (1993).

By using the network concept in this specific area of research, not only the usual individual-oriented qualitative analysis techniques, but also relation-oriented quantitative analyses can be used. In this way, besides using multilevel analysis, it was also possible to investigate the spread of cocaine use by means of an outdegree analysis (Spreen 1993) and to elaborate estimators for the size of the network (Intraval 1992; Frank and Snijders 1994).

NOTE

* This research is a sequel to an earlier cocaine research project commissioned by the municipality of Rotterdam, conducted and reported by research bureau Intraval.

REFERENCES

Bieleman, B., A. Diaz, G. Merlo, and Ch. D. Kaplan 1993 (eds.) Lines across Europe: Nature and Extent of Cocaine Use in Barcelona, Rotterdam and Turin. Swets en Zeitlinger, Amsterdam/Lisse.

Biernacki, P. and D. Waldorf 1981 Snowball Sampling: Problems and Techniques of Chain Referral Sampling. Sociological Methods and Research 10: 141–163.

Bryk, A. S., S. W. Raudenbush, M. Seltzer, and R. T. Congdon 1988 HLM. Chicago: University of Chicago.

Bryk, A. S. and S. W. Raudenbush 1992 Hierarchical Linear Models, Applications and Data Analysis Methods. London: Sage Publications.

Campbell, K. E. and B. A. Lee 1992 Sources of Personal Neighbor Networks: Social Integration, Need, or Time? Social Forces 70: 1077–1100.

Coleman, J. S. 1958 Relational Analysis: The Study of Social Organizations with Survey Methods. Human Organization 17: 28–36.

Fischer, C. S. 1982 To Dwell among Friends: Personal Networks in Town and City. Chicago: University of Chicago Press.

Frank, O. and T. A. B. Snijders 1994 Estimating the Size of a Hidden Population by Use of Snowball Samples. Journal of Official Statistics 10: 53–67.

Goldstein H. 1987 Multilevel Models in Educational and Social Research. London: Charles Griffin & Co.

Intraval 1992 Between the Lines: A Study of the Nature and Extent of Cocaine Use in Rotterdam. Groningen and Rotterdam: Intraval.

Johnson, J. C. 1990 Selecting Etnographic Informants. Newbury Park, CA: Sage.

Longford, N. T. 1988 VARCL: Software for Variance Component Analysis of Data with Hierarchically Nested Random Effects (Maximum Likelihood). Manual. Princeton, N.J.: Educational Testing Service.

Morrison, V. L. 1988 Observation and Snowballing: Useful Tools for Research into Illicit Drug Use? Social Pharmacology 2: 247–271.

Prosser, R., J. Rasbash, and H. Goldstein 1991 ML3: Software for Three-Level Analysis. London: Institute of Education, University of London.

Snijders, T. A. B. and R. J. Bosker 1994 Modeled Variance in Twolevel Models. Sociological Methods and Research 22: 342–363.

Spreen, M. 1992 Rare Populations, Hidden Populations and Link-tracing Designs: What and Why? Bulletin de Methodologie Sociologique 36: 34–59.

Spreen, M. 1993 Sampling and Analysing Structure in Hidden Populations. In: B. Bieleman *et al.* (eds.) Lines across Europe: Nature and Extent of Cocaine Use in Barcelona, Rotterdam and Turin. Swets en Zeitlinger, Amsterdam/Lisse, 185–205.

Wasserman, S. and K. Faust 1994 Social Network Analysis: Methods and Applications. New York and Cambridge: Cambridge University Press.

Wellman, B., O. Frank, V. Espinoza, S. Lundquist, and C. Wilson 1991 Integrating Individual, Relational and Structural Analysis. Social Networks 13: 223–249.