2012

Multiple membership models for social network and group dependencies

Mark Tranmer
University of Manchester

David Steel
University of Wollongong, dsteel@uow.edu.au

William J. Browne
University of Bristol

Recommended Citation

Tranmer, Mark; Steel, David; and Browne, William J., Multiple membership models for social network and group dependencies, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 1-12, 2012, 17.
http://ro.uow.edu.au/cssmwp/88

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

01-12

Multiple Membership Models for Social Network and Group Dependencies

Mark Tranmer, David Steel and William J Browne

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Multiple Membership Models for Social Network and Group Dependencies.

Mark Tranmer
CCSR & Mitchell Centre for Social Network Analysis
Social Statistics
University of Manchester
Manchester, M13 9PL, UK.

David Steel
Centre for Statistical and Survey Methodology
School of Mathematics and Applied Statistics
University of Wollongong
Wollongong, NSW 2522, Australia.

William J Browne
School of Veterinary Science
University of Bristol
Bristol, BS40 5DU, UK.

Summary. Multilevel models have been developed and applied for individuals in groups, such as schools or areas, but these models tend to not consider the networks of an individual's connections within and between groups; the social network has largely been ignored as an *additional* source of dependence in multilevel modelling that is carried out in social statistics. Typical models for network dependencies in the social networks literature, such as network autocorrelation models, have largely ignored other sources of dependence, such as the school or area in which an individual lives. To bridge this divide, a multiple membership modelling approach for jointly investigating social network and group dependencies is presented that allows social network and group dependencies on individual responses to be investigated and compared, and which can be analysed using MCMC estimation in standard statistical software for multilevel modelling. This approach is used to analyse a subsample of the Adolescent Health dataset from the US, where the two response variables of interest are individual level educational attainment and self-assessed health status, and the three individual level covariates are sex, ethnic group and age. Individual, network, school and area levels are included in the analysis. The network level can be represented with various configurations. The results suggest that the network should not be ignored from a statistical perspective when studying variations in educational attainment, as ignoring this level impacts on the estimates of variation at the other levels (school, area, individual), as well as having some impact on the point estimates and standard errors of the estimates of regression coefficients for covariates in the fixed part of the model. From a substantive perspective, this approach provides a flexible and practical way of investigating the network level, and comparing its relative importance to other group levels such as areas or schools.

Keywords: Multilevel models ; Linear regression ; Logistic regression ; Autocorrelation.

E-mail: mark.tranmer@manchester.ac.uk; CSSM Working Paper, January 2012.

1. Introduction

Many social populations have a multilevel structure, such as individuals in households and areas, students in schools, workers in organisations, or observations made on individuals at different time points. Multilevel models have been developed to analyse data collected from these populations (see, for example Snijders & Bosker (1999), Goldstein (2003)), and to allow relationships between variables to be investigated at a particular level, often the individual level, whilst taking into account effects operating at other levels of the population structure. In some situations, the aim of the analysis is to assess whether an individual level relationship differs from one group to another. For example, is the relationship between prior and current examination performance stronger in some schools than others? Multilevel modelling can also be used as a model-based approach for analysing data that have been collected from a multi-stage sample, when the target of inference is the individual level relationship between a set of variables. In this approach the primary sampling unit is included as a level, and the population structure may be regarded as a nuisance to be taken into account in the analysis of an individual level relationship, so that regression coefficients are estimated efficiently and appropriate standard error estimates are calculated. However, more often in multilevel analysis the population structure is of direct substantive interest, and the aim of the analysis may be twofold: to estimate an individual level relationship given the complex population structure, and to measure the nature and extent of variations in variables of interest at the various levels of this population structure.

Numerous studies have been carried out to investigate household level, school level, or area level variations in individual level responses for social, educational and health outcome variables using multilevel models. However, the social network has been largely ignored in the multilevel modelling literature as an *additional* source of dependence in the context of these other levels. Network autocorrelation models have been developed for social network dependencies in the social network literature, adapted from models originally developed for spatial dependencies (Cliff & Ord 1975, Ord 1975). Whilst network autocorrelation models take the social network dependencies into account, they essentially ignore other levels of the population structure unless these are added to the model as fixed effects, which can be problematic when there are many groups. For network autocorrelation models it is also assumed that the response variable is continuous. Moreover, network autocorrelation models do not allow an easy assessment of the relative share of variation in an individual response at the individual, network and group levels.

This paper illustrates some new approaches for assessing social network dependencies in individual response variables - academic performance and self-assessed quality of health, while including other group dependencies such as the school or area to which an individual belongs, by applying multiple membership models (Hill & Goldstein 1998) to the Adolescent Health (hereafter, ‘Add Health’) dataset (Harris et al. 2009). The main aim of the analysis is twofold: to estimate an overall individual level relationship between a set of variables, and to measure the nature and extent of variations in individual level responses at various levels of the population structure. There are two other aims of the analysis: an assessment of the consequences, in terms of model parameter estimation and subsequent statistical inference, of ignoring different features of the population structure in the analysis, and a comparison of multiple membership models with existing models for network autocorrelation.

An empirical study has been carried out using a subsample of the Add Health data for one state in the US, focusing on two individual level response variables: academic performance and self-assessed health status, each of which is related to three individual level covariates: gender, ethnic group and age. Friendship networks are available for the individuals in the Add Health data. The levels of the population structure that are studied alongside the social network are: area (defined by US county) and school, as well as the individual level.

In the multiple membership modelling approach for social networks, individuals are 'members' of cliques (sub-networks of a particular size where every individual is directly connected to one another), or ego-nets (a particular individual in the network, and all the other individuals in the network to which they are directly connected). Each individual may 'belong' to more than one of these groups, or in some cases not be a member of any groups - for example population network isolates, or members of the sampled network with very few connections to other sampled individuals. Members of the network with few connections are more likely to occur in sampled networks, because the friends that sampled individuals nominate may be out-of-sample. Other group dependencies, such as those at the school or area level are added as additional crossed levels in multiple membership models, which are a type of multilevel model.

Where possible in this paper, the results of the multiple membership modelling approaches are compared with existing approaches for taking into account social network dependencies: in particular, network autocorrelation models. In addition, comparisons are made with a single level regression model, which ignores all features of the population structure. It is of interest to assess whether such a naive statistical modelling approach would lead to different substantive conclusions when compared with those models that account for population structure, as well as assessing the consequences for statistical estimation of ignoring levels of the population structure above the individual.

The remainder of the paper is organised as follows. In Section 2, a description of the data structure and preparation is given, together with some descriptive results. In Section 3 a discussion of the modelling approaches is given, including a brief review of network autocorrelation models, and a definition of the multiple membership models for social network and group dependencies. In Section 4, the results of analysing various models are given. Finally, a discussion is given and conclusions are drawn in Section 5.

2. Data description and preparation.

The data are from a restricted access Add Health dataset for Wave I (Harris et al. 2009), and were analysed in a secure data environment. Add Health is a longitudinal school-based sample of school students in grades 7-12 (typically aged 11-16) in the US, which commenced around 1995. Information collected includes a range of health outcome variables, academic performance, and individual (i.e. student level) characteristics, as well as school and geographical details, and friendship networks.

Two dependent variables are analysed, and both are individual level outcome variables. The first is a measure of academic performance, which was generated by converting the grades from four subjects (graded A-D) into four five-point scales: (0: did not take subject, 1=D or less, 2=C, 3=B, 4=A) and then summing them, so that the resulting academic performance score ranges from 0-16. 15 students achieved scores of zero for this measure, so there were very few students that did not take any subjects. We felt it was better to include all sampled students in the analysis, to retain as much of the sampled network as possible and because all students answered the self-assessed health status question; the small percentage of students with a score of zero (1.5%) will have minimal effect on the overall parameter estimates in the models for academic performance. These scores were then standardised to have mean zero and standard deviation 1 and are called 'ztotscore' in the tables. The second dependent variable is a self-assessed measure of quality of health. Survey participants were able to grade their quality of health into 'poor','fair','good' and 'excellent'. This dependent variable has then been dichotomised into 'excellent' or 'good' health = 1 and 'fair' or 'poor' health = 0.

There are three covariates: ethnicity, which is coded as a binary indicator as to whether the student is black American or not, gender and age. The first two covariates are categorical, and the third is continuous, and is centred in the modelling results that appear in Section 4. A subset of the Add Health data was used for 'State 7', a state of the US which comprises 968 valid sampled cases for individuals in 10 schools, and in 13 areas at the county scale. Whilst the sample is evenly distributed in the 10 schools, the individuals are not evenly distributed in the 13 areas. Instead, some areas have very large sample sizes, and others are very small, because Add Health is a school-based rather than area-based sample. There are also two other finer scales of geography in the Add Health dataset: 'neighbourhood 1' and 'neighbourhood 2', where the latter definition is the finest geographical scale, though both these neighbourhood definitions often include only 1 sample unit per-area. Whilst substantively these geographical scales may be interesting to study, there was insufficient data to consider these areas further as levels in the models, as it would be very difficult to disentangle individual and area level effects at these scales when there is often only one sample unit per-area.

The friendship networks are potentially available for all pupils in the sample. Each individual is asked to name up to five male and five female friends. These friends may also be in the sample, or people can be named that are out-of sample. The networks are available in the Add Health data as adjacency (node) lists. Directed binary adjacency matrices were extracted from these lists using R, via a routine for converting adjacency lists to adjacency matrices. For a sample of size n, the binary adjacency matrix, $\mathbf{D}$ has dimensions $n \times n$. An element d_{ij} in the matrix $\mathbf{D}$ takes the value 1 if i nominates j as a friend and 0 otherwise. Individual j may reciprocate by nominating i as a friend, or may not, so that the directed adjacency matrix is not usually symmetric. Diagonal elements d_{ii}, which would indicate self-friendship nominations, are assumed to be zero. These adjacency matrices were restricted to other individuals in the sample, and also in State 7. Because of these restrictions, around half the sample used here are social network isolates. This does not present problems in the modelling approaches discussed later, but it is worth noting that the structure of a sampled network is different from a population network; a sample network is likely to be a lot sparser and less structured than the corresponding population network. As well as obtaining the directed adjacency matrix, $\mathbf{D}$, a symmetrized (undirected) version of this matrix, $\mathbf{D}^s$, was obtained, and was used in the clique set analyses below. For a symmetrized matrix, $\mathbf{D}^s$, if i nominates j as a friend, if j nominates i as a friend, or if both nominate each other as friends, the element d_{ij} is given a value of 1. Thus, the symmetrized matrix is, by definition, symmetric, and tends to have much higher 'density' (have more of the possible network connections present) than the directed adjacency matrix, $\mathbf{D}$, from which $\mathbf{D}^s$ was derived.

Where the network level was considered, the multiple membership models applied in Section 4 involved clique memberships for undirected cliques of minimum size 2 (clique-2) and of minimum size 3 (clique-3), as well as ego-net membership. When clique-2 and clique-3 definitions are used together in the models, the clique-2 level represents dyads only, and the clique-3 level represents cliques of size at least 3. In theory, cliques of minimum size 4, or other network configurations could be specified in this approach, but there are too few of these to include as a level in the models for these sample data. The clique sets for these two thresholds were obtained using UCINET 6 (Borgatti et al. 2002). The program implements the Bron & Kerbosch (1973) algorithm to find all Luce and

Perry (1949) cliques of a specified minimum size. In addition, the ego nets for each individual were extracted from the adjacency matrix; the ego net for each individual is the corresponding row of the adjacency matrix. Once the ego nets and clique sets had been obtained, weight matrices were produced for each individual. These weights work as follows. If an individual is a member of 4 cliques of minimum size 3, their weights for each of these clique-3s will be $\frac{1}{4}$. Where ego-nets are used in the models a similar approach is used: if an individual (ego) had three friends in the sample (alters), each of these alters would each be given a weight of $\frac{1}{3}$ in the weight matrix.

2.1. Descriptive Statistics

Attributes.

The total sample size is 968, comprising 510 (52%) females and 161 (17%) black Americans. The average age in the sample is 14.76 with a minimum of 10 and a maximum of 19. The response variable academic performance (ztotscore) has been standardised to have mean zero and standard deviation 1. For the dichotomous health response variable, 703 (72.6%) self-assessed their health as ‘excellent’ or ’good’, the remainder self-assessing their health as ‘fair’ or ‘poor’.

Social Networks.

Each row of the directed network represents the ego net of an individual. Thus if individual i, represented by row i of the matrix $\mathbf{D}$, nominates four friends, then the four off-diagonal elements in row i of $\mathbf{D}$ that represent these friends will take the value 1 and the remaining n - 4 will take the value zero. Thus it is easy to identify the ego net of each individual. In the models that follow the ego net of each individual is only represented by their friends, so that ego is not included in their own ego-net, though alternative weighting schemes could be used. In the weighting schemes used here, if an individual (ego) has 3 friends in the sample, each of these is given a weight of $\frac{1}{3}$ in the weight matrix that is used part of the information required to set up the multiple membership model in multilevel modelling software, such as MLwiN.

A triad census was carried out in which every trio of actors is selected in turn and the nature of their connections, if any exist, is tabulated. The triad census results for the directed and undirected adjacency matrices are given in Table 1 in the Appendix. Definitions of these different types of triads can be found in Holland & Leinhardt (1976). For both directed and undirected networks, there are many empty triads (type 003) and there are 138 complete triads (type 300). In the undirected network a complete triad is equivalent to a clique of exactly size 3. These results suggest that identifying cliques in the network and using cliques as network definitions in the modelling approaches makes sense.

Groups.

The distribution of people in areas, where the areas here are counties, is uneven. The distribution is shown in Table 2 in the Appendix. Some areas, e.g. county 38, have a large sample size, whilst some other areas contain only one observation. There are six areas with sample sizes greater than 50. The sample is much more evenly distributed in the 10 schools included in State 7, because the sample design is a school based study (See Table 3).

3. Models

3.1. Multiple membership models

Social network and group dependencies can be taken into account through a random effects modelling approach via a multilevel model, known as the 'multiple membership' model . We adapt the notation of Browne et al. (2001, eqn. 5), for a multiple membership model.

The multiple membership model for social network dependencies may be defined by one level of grouping (e.g. membership of ego-nets) as:

$$\begin{aligned} y_i &= \mathbf{x}_i'\beta + \sum_{j \in group1(i)} w_{1i,j} u_j^{(2)} + e_i \\ i &= 1, ..., n \quad group1(i) \subset (i, ..., J_1) \\ u_j^{(2)} &\sim N(0, \sigma^2_{u^{(2)}}), \quad e_i \sim N(0, \sigma^2_e) \end{aligned} \tag{1}$$

In Model (1), y_i is an individual level response. In this model formulation this response is assumed to be continuous, $\mathbf{x}_i$ is a vector of fixed covariates, β is the vector of their regression coefficients. Here, $group1(i)$ is the set of network subgroups to which i is a member. The term $\sum_{j \in group1(i)} w_{1i,j} u_j^{(2)}$ involves a set of J_1 random effects $u_j^{(2)}$, where J_1 is the total number of network subgroups for the network definition included in the model e.g. the total number of cliques of minimum size 3. The weight given to each individual for their network subgroup membership is $w_{1i,j}$. These weights sum to 1 for each individual. The random effects at the individual and network levels are assumed to be uncorrelated.

The multiple membership model for social network dependencies may be defined by two levels of network subgrouping (e.g. membership of cliques of minimum size 2, and membership cliques of minimum size 3) as:

$$\begin{aligned} y_i &= \mathbf{x}_i'\beta + \sum_{j \in group2(i)} w_{2i,j} u_j^{(3)} + \sum_{j \in group1(i)} w_{1i,j} u_j^{(2)} + e_i \\ i &= 1, ..., n \quad group1(i) \subset (i, ..., J_1), \quad group2(i) \subset (i, ..., J_2) \\ u_j^{(3)} &\sim N(0, \sigma^2_{u^{(3)}}), \quad u_j^{(2)} \sim N(0, \sigma^2_{u^{(2)}}), \quad e_i \sim N(0, \sigma^2_e) \end{aligned} \tag{2}$$

The term $\sum_{j \in group1(i)} w_{2i,j} u_j^{(3)}$ involves a set of J_2 random effects $u_j^{(3)}$, where J_2 is the total number of network subgroups for the second network definition included in the model. The weight given to each individual for their membership of this second set of network subgroups is is $w_{2i,j}$, and these weights sum to 1 for each individual.

Finally, the multiple membership model may be extended to allow for both social network and group dependencies. As an example, the model for social network dependencies defined at two levels, Model (2) is extended to include two additional levels: a school level, with random effects $u^{(5)}_{school(i)}$ and a geographical level, as defined by 'county', with random effects $u^{(4)}_{area(i)}$. The area and the school are not assumed to be nested in this model formulation:

$$\begin{aligned} y_i = \mathbf{x}_i'\beta + u^{(5)}_{school(i)} + u^{(4)}_{area(i)} + \sum_{j \in group2(i)} w_{2i,j} u^{(3)}_j + \sum_{j \in group1(i)} w_{1i,j} u^{(2)}_j + e_i \\ i = 1, ..., n \quad group1(i) \subset (i, ..., J_1), \quad group2(i) \subset (i, ..., J_2) \\ u^{(5)}_{school(i)} \sim N(0, \sigma^2_{u^{(5)}}), \quad u^{(4)}_{area(i)} \sim N(0, \sigma^2_{u^{(4)}}) \\ u^{(3)}_j \sim N(0, \sigma^2_{u^{(3)}}), \quad u^{(2)}_j \sim N(0, \sigma^2_{u^{(2)}}), \quad e_i \sim N(0, \sigma^2_e) \end{aligned} \tag{3}$$

Here $school(i)$ is the school to which the student i belongs, and $area(i)$ is the area in which student i lives.

An alternative way of accounting for the group dependencies is to extend model (1) or (2) to include fixed indicator variables for the groups and this may be appropriate where the number of groups is small and all groups are included in the sample. An example of this approach, where there are fixed indicator variables for schools, rather than adding school as a level (and areas are not included in the model), can be found in Table 8.

Models (1), (2) and (3) can all be estimated in the software MLwiN (Rasbash et al. 2009). The model parameters are estimated via Monte Carlo Markov Chain (MCMC) (Browne 2009). Prior to fitting the models, data preparation to identify the groups in the networks, such as ego-nets, or cliques, and the corresponding membership weights, as discussed above in Section 2, must be undertaken. Model (3) was fitted to the Add health data in the form as specified above with area, school, network and individual levels all defined, and such a model was run as a null model (no covariates) and as a model with fixed covariates included. In this model, the two levels of network dependencies were clique-2s and clique-3s, based on the undirected adjacency matrix. A slightly reduced form of Model (3) was also fitted, with just one level of network dependencies: the ego-nets of each individual, as derived from the directed adjacency matrix. Linear formulations of these models (formulation as shown above) were estimated for the academic performance outcome and logistic versions of these models (formulation not shown) were run for the dichotomous health outcome. The results are discussed in the next section.

Fitting these models allows us to fulfil the two main aims of our analysis. Firstly, to estimate the individual level relationship between the two individual level outcome variables academic performance and self assessed health status with the three individual level covariates: gender, ethnic group and age, having taken the clustering of academic performance in networks, schools and areas into account. Secondly, to investigate the extent of variation in academic performance at the individual, network, school and area levels, both before and after the inclusion of individual-level covariates. In addition, reduced versions of this model were also fitted, including a single level regression model. That is, a model that ignores school, area and network dependencies, as well as models that ignore some of the levels above the individual. This enables us to fulfil one of the other aims of this paper; to investigate the consequences of ignoring particular features of the population structure in statistical analysis. The results are given in Section 4.

3.2. Network Autocorrelation models

Other models exist for social network dependencies, and these have been applied in the social networks literature; in particular, *network autocorrelation models*. See Leenders (2002) for a review. One such network autocorrelation model is the *network effects* model, also known in the geographical literature as the *spatial effects* model (Doreian 1980), and defined as:

$$\begin{aligned}\mathbf{Y} &= \rho\mathbf{A}_1\mathbf{Y} + \mathbf{X}\beta + \epsilon \\ \epsilon &\sim N(0, \sigma^2_\epsilon \mathbf{I}_n) \end{aligned} \tag{4}$$

where $\mathbf{Y}$ is a $n \times 1$ vector of response variables and $\mathbf{A}_1$ is an $n \times n$ matrix of weights to reflect the connections for the n individuals. The definition of $\mathbf{A}_1$ will affect the estimated model parameters, as Leenders (2002) noted. One such definition, as used here, is to derive $\mathbf{A}_1$ from the row-standardised directed adjacency matrix, so that this is the same weight information as used in the multiple membership models based on ego-nets, as defined above. The autocorrelation parameter, ρ, measures the strength of association between an individual's response and the responses to which the individual is connected. The same value of ρ is assumed for all of these connections. $\mathbf{X}$ is a $n \times p$ matrix of fixed covariates, with associated regression coefficients β, and ϵ is a vector of error terms. If there are no connections between individuals, or if ρ is zero, Model (4) reduces to an ordinary least squares regression equation.

A variation of this model, the *network disturbances* model, also known in the geographical literature as the *spatial disturbances* model (Doreian 1980), is defined as

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\beta + \epsilon \\ \epsilon &= \rho\mathbf{A}_2\epsilon + \xi \\ \xi &\sim N(0, \sigma^2_\xi \mathbf{I}_n) \end{aligned} \tag{5}$$

In Model (5) the parameters in common with Model (4) are defined identically. The matrix $\mathbf{A}_2$ is an $n \times n$ weight matrix to reflect the connections for the n individuals, which could be identical to $\mathbf{A}_1$. Again, if there are no connections between individuals, or if ρ is zero, Model (5) reduces to an ordinary least squares regression equation. Finally, it is possible to formulate a combined model based on (4) and (5), so that network autocorrelation is jointly modelled through the response and error terms, and where the matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ may be the same, or different. Models (4) and (5), and the combined model, may be estimated using the `sna` and `spdep` packages in `R` (Butts 2008, Bivand 2010). In the literature, the models tend to perform similarly in terms of statistical measures (Doreian 1980, Leenders 2002). The decision to fit the network effects model, network disturbances model, or some combination of the two, should therefore be made on substantive grounds.

3.3. Different modelling approaches

As well as the social network dependencies, other group dependencies may also be present. For example, in the Add Health data, there are geographical groups at various scales (neighbourhoods, and counties), as well as schools. In the fixed effects approaches of Models (4) and (5), it is possible to also add the groups as fixed effects via a set of indicator (dummy) variables. This approach is often appropriate when there is a relatively small number of groups. Some of these models were estimated below with schools as groups. This enabled us to compare the network disturbance model and multiple membership models for an example that included network and group dependencies (see Section 4 for a discussion). When modelling the cliques in the multiple membership models, it would be possible to just include one term e.g. just for clique-3s without clique 2s, and this may be appropriate if we had a particular substantive theory about cliques of size 3 or more with respect to the individual level responses.

The decision to fit network autocorrelation models or multiple membership models when other groups are present is likely to depend on two things: one substantive, one practical. Firstly, from a substantive perspective, if interest focuses in estimating the nature and extent of variation in individual responses at the network 'level(s)' compared with other group levels, then the multiple membership model may be preferable because variance components are explicitly included as model parameters and it is possible to model other group levels alongside the networks as random effects and compare the extent of variation at the different levels. However, if the substantive focus of the analysis is on estimating the individual level relationship and regarding the networks as a nuisance then the network autocorrelation modelling may be preferable. Moreover, the feedback correlations obtained from these models may have some substantive interpretation in some contexts. Secondly, from a practical perspective, if the number of groups such as schools or areas is large, and/or if the sample is unevenly distributed within the groups, the multiple membership modelling approach may be more statistically efficient. Also, it is relatively straightforward to define and fit the multiple membership model for a binary response, as we demonstrate in the results section, whereas it is not straightforward to fit the network autocorrelation models for binary responses. We compare the two approaches in the context of the Add Health data in the next section for the academic performance response variable, which is continuous.

4. Results.

We began by fitting multiple membership models for network and group dependencies with standardised academic score (ztotscore) as the dependent variable. In Tables 4 and 5 we defined the network through the clique-2 and clique-3 membership based on the undirected adjacency matrices. All multiple membership models presented in this paper were estimated via the Monte Carlo Markov Chain (MCMC) algorithm using default flat priors for the fixed effects and a chain of 20,000 samples, implemented in MLwiN. In all models, standard (gamma) priors were assumed for the variance parameters.

Table 4 shows null models with only a constant term in the fixed part, and Table 5 adds the three individual level fixed covariates. Because both clique membership terms are included in the models, the clique-2 represents dyad membership only, and clique-3 represents individuals that are members of cliques of minimum size 3. Because the networks are based on sample data, and only a limited number of friend nominations per-individual is allowed in the Add Health survey, there are very few cliques of size 4 or more.

Network, school and area levels were included in the 'complete' models (penultimate columns of Tables 4 and 5) to meet our two aims, of estimating the individual level relationship between academic performance and the individual level covariates, having taken into account the population structure above the individual level, and to investigate the nature and extent of variations in individual level responses at the individual and group levels. In Table 4 for the null models, we found that the best-fitting models according to the Deviance Information Criterion (DIC) involved the network and the school level above the individual level (DIC=2700). For the model that included schools areas and networks there was little difference in the goodness of fit (DIC=2701) when compared with the model that only includes network and school levels, probably reflecting the overlap of networks and areas in this example. The areas may not perform quite as well as a level when schools are not included because of the uneven sample distribution in areas (DIC=2716). The final column of Table 4 shows the results for the null model with a DIC of 2750 suggesting a much worse model fit when no population structure above the individual level is accounted for in the null model. These goodness of fit results suggest that the population structure above the individual level, including the network level, should not be ignored in statistical analysis when studying variations in academic performance.

The penultimate column of Table 4 shows the results for the model that includes networks, schools and areas. It can be seen that most of the variation is between individuals but there is also considerable variation at the network level for both clique definitions, suggesting clustering within dyads (clique-2s) and within more clustered parts of the network (clique-3s). There is also more variation between schools than between areas (variance components of .052 and .036 respectively).

In Table 4 we also removed features of the population structure from the models to assess the consequences of ignoring one or more levels of population structure in this example. If just schools or just areas are modelled above the individual we see that the individual variance component increases compared with the cases where the network level is also included, but the school or area level components do not change much. This is in contrast to the findings of Tranmer & Steel (2001) who found that ignoring a geographical level between the individual level and a more aggregated area level affected both estimates in an example based on UK census data. However this may reflect the fact that the levels were hierarchical in the case of census data. In the case of the Add Health data, the levels are cross-classified.

Looking now at the model that adds the three individual level covariates in Table 5, we see a pattern fairly similar to that described for Table 4 in terms of model goodness of fit. The penultimate column of Table 5 gives the model results with all three levels above the individual: school, area and network, included and we find that for the individual level relationship between academic score and the three covariates (Black, Female, Age) there is little change in the coefficients and only a minor reduction in their standard errors, although the coefficient for 'black' changes in its point estimate (though it is non-significant in both cases) and the standard error increases a little for the model that includes population structure above the individual level.

In terms of the share of the variation at the network, area and school level (penultimate column of Table 5) the pattern is similar to that described above for the null model (Table 4). In all models, the coefficient estimates for female and age are found to be statistically significant, and conditional on the addition of the covariates to the model, the school and area level variance components reduce. The clique-3 variance components are also reduced by the inclusion of covariates and, interestingly, in the model with all features of the population structure included (penultimate column) there is now a slightly larger variance for the clique-2 level than the clique-3 level, though both component estimates are reduced, compared with the corresponding results from the null model. At these levels, the covariates may be partly explaining homophily (people with similar characteristics tend to be more likely to be connected (McPherson et al. 2001)) in the networks with respect to age, sex and ethnic group.

In terms of model fit, the model with all features (penultimate column) shares the lowest DIC with the model that includes areas and social networks, but not a school level (DIC=2674); this may suggest that the individual characteristics added to these models are to some extent clustered within schools and are thus accounting for some of the school differences. Interestingly, in this example a similar conclusion would have been reached regarding the individual level relationship between academic performance and the three covariates: black, female and age, if a single-level regression model had been estimated with the individual level data (final column), as was found with the more sophisticated models in Table 5 that take into account population structure. However, the single level regression would not allow any potentially important substantive inferences to be made about variation in academic performance at the levels of the population structure above the individual. The fact that models that allow for population structure above the individual level have a statistically better fit, suggests that these additional levels should not be ignored. One reason why the individual level relationship between academic performance and the covariates might be similar could be the large number of social network isolates in this dataset.

In Table 6, we carried out an analysis similar to that of Tables 4 and 5, but this time we defined the network level as ego-net membership based on the directed adjacency matrix, $\mathbf{D}$, with the weight matrix calculated as a row-standardised version of $\mathbf{D}$. We found that defining the network by ego-net as opposed to cliques results in slightly better goodness of fit and similar coefficients for the individual level covariates and their standard errors for the model as found, for example, in Table 5 for the corresponding models where networks were defined via cliques. The individual level variance component estimate reduces more in the ego-net than for the corresponding clique model, however, and the network variance is higher for the single ego-net component than for either of the clique estimates for the corresponding model in Table 5. The school and area level variance components are similar for the full model in Table 6 as for the comparable model where the network is defined via cliques (penultimate column of Table 5).

In Table 7, we applied the multiple membership models to the dichotomous individual level health response 'health good/excellent' = 1 vs. 'health fair/poor' = 0. Table 7 summarises the results of a multilevel logistic regression model, where the response is now dichotomous. There is evidence that the social network and group levels are important with respect to variations in health in this example because the DIC statistics are smaller when compared with a model that ignores population structure above the individual level; we note that in this case it appears that most of the variation appears to be at the area and school levels rather than at the network level for the health response. In the single-level model shown in the final column of Table 7 in which population structure above the individual is ignored, we note that the coefficient of female would be found to be statistically non-significant at the 5% significance level, whereas this coefficient would be statistically significant at the 5% level in the corresponding models that include networks, schools and areas, regardless of whether the network is defined via cliques or ego-nets.

In Table 8, we compared the multiple membership and network autocorrelation model as closely as possible for this data-set for modelling social networks and groups by fitting a model with just networks and schools as levels above the individual. We did this as follows. Because the sample is evenly distributed across the schools, we created nine fixed indicator variables for the ten schools, choosing School 17 as a reference group. We then fitted a Multiple Membership (MM) model with networks defined by ego-nets, and estimated a Network Disturbance (ND) model using the same weight information in both cases: the row standardised version of the adjacency matrix. The network disturbance models were first run using the `lnam` command in the `sna` package in `R`, and the results are estimated via maximum likelihood, so that their goodness of fit statistic is based on the AIC rather than the DIC for the MM models that were estimated via MCMC, though the two measures are comparable. We chose to compare the ND model with the MM model because in the ND model the network effects are going through the error term, rather than directly through the response as would be the case for the network effects model In Table 8 we fitted a model with just a constant and the school indicators and also a 'full model that also includes the three individual level covariates. We find that the two models have comparable results in terms of the estimated coefficients, their standard errors, and the model goodness of fit.

5. Discussion and conclusion

An approach for estimating the nature and extent of social network dependencies on individual responses in the context of other group dependencies has been demonstrated, using multiple membership models. This approach takes into account social network and group dependencies when estimating an individual level relationship and allows the variations in individual level responses at levels of the population structure above the individual to be compared. The model can be fitted in software for multilevel modelling, such as MLwiN. The model results for the academic performance score in particular suggest that network dependencies exist, even when other levels of the population structure such as school and areas are accounted for. In many cases the variance component estimates for network cliques and ego nets are larger than the corresponding model estimates for school and area level variance components, this is because the ego-nets and cliques are smaller and compact groups than schools or areas ; thus we would expect relatively high variance components for cliques and ego-nets when compared with schools and areas.

These new approaches compare well in terms of goodness of fit with existing models used for social network dependencies, such as network autocorrelation models. Substantively, there are implications for assuming these models: in particular it is assumed for these models that the social network is exogenous to the individual response. This is a different assumption from those models that allow the co-evolution of social networks and behaviour, and hence selection and influence, to be assessed with longitudinal network data: Stochastic Actor Based Models (SABMs) (Snijders et al. 2010). The models presented here are not an alternative to SABMs; they have a different role. In a descriptive sense, the models presented here allow the variation in individual responses to be assessed at different levels of the population structure, including the network level. Hence, using the multiple membership modelling approaches it is possible to ask "Is there more variation in academic performance or health between networks, between schools, or between areas?", and to obtain a measure before and after the addition of other variables to the model, such as individual characteristics, that might capture homophily at the network level. This approach could be applied to other situations given data availability. For example: individuals, households, social networks and geographical groups.

In terms of the multiple membership model, various extensions are possible via the multilevel approach. For example: given the multiple waves of the Add Health data, a time level could be included to allow further inferences about variations in individual responses through time in the context of other population structural features. Also, covariates could be given random coefficients at different levels. For example, it is possible to make 'female' random at the clique or ego-net level to assess whether social network structure has a different association with the response variable for girls than for boys. It has already been demonstrated here that the multiple membership model can be applied to a dichotomous response and other categorical responses could also be modelled (for example, an ordinal response for self-assessed health status (excellent, good, fair, poor)). Another application of these methods is to multiplex networks, where the different types of connection are specified as a series of weight matrices in the multiple membership model specification. Moreover, a bivariate response could be considered within the multilevel model framework. Finally, other data could be combined with the network data in the multiple membership model framework; for example Census information could be combined with the survey data at area level, if the areas were identifiable in the survey data.

The goodness of fit, measured by DIC, of the models that define network by clique is almost as low as the corresponding DICs for models that define networks by ego net, and this may have implications for network data release. For example, in some cases the network may be accounted for using anonymised clique-set information rather than full network information, rather than using the full network. We note that the full network information would usually be preferable, but if this could not be released then accounting for cliques would be much better than ignoring the network dependencies altogether.

Models based on ego-net connections capture direct relationships between people, whereas those that involve cliques explicitly capture higher-order substructures in the network. In our examples both models performed similarly in terms of goodness of fit. The decision as to how to model the network may depend on the substantive theories to be tested; whether these relate to clusters in the network or simply to connected individuals. The MM model framework would allow other network configurations to be included - for example the kinds of substructures that are often included in Exponential Random Graph Models (see, for example, Robins, Pattison, Kalish & Lusher (2007), Robins, Snijders, Wang, Handcock & Pattison (2007) for an introduction), provided such substructures could be identified and the weight matrices for membership of these substructures by each individual could be obtained.

6. Acknowledgements

This research was funded by ESRC/ARC Bilateral Australia grant: “The Role of Households, Social Networks and Geographical Groups in Social Statistics”, and ESRC e-stat node.

The authors wish to thank Sam Smith (Manchester), Dai Shizuka (Chicago) and David McDonald (Wyoming) for their helpful advice on network data preparation. We also thank Harvey Goldstein (Bristol), Johan Koskinen (Manchester), Vernon Gayle and Paul Lambert (Stirling) for general comments. Finally, we wish to thank Thomas Grund (Zürich) for general comments when he was discussant for an early version of this paper, presented at the Applications of Social Network Analysis (ASNA) conference in Zurich, September 2011.

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for the analysis in this paper.

References

Bivand, R. (2010), ‘Package spdep in R’.

Borgatti, S. P., Everett, M. G. & Freeman, L. C. (2002), ‘UCINET 6 For Windows: Software for Social Network Analysis’.
URL: *www.analytictech.com*

Bron, C. & Kerbosch, J. (1973), ‘Algorithm 457: finding all cliques of an undirected graph’, *Commun. ACM* **16**(9), 575–577.

Browne, W. J. (2009), 'MCMC estimation in MLwiN', *Centre for Multilevel Modelling, University of Bristol.* .

Browne, W. J., Goldstein, H. & Rasbash, J. (2001), 'Multiple membership multiple classification (MMMC) models', *Statistical Modelling* **1**, 103–124.

Butts, C. T. (2008), 'Social Network Analysis with sna', *Journal Of Statistical Software* **24**(6).

Cliff, A. D. & Ord, J. K. (1975), 'Model Building and the Analysis of Spatial Pattern in Human Geography', *Journal of the Royal Statistical Society, Series B* **37**(3), 297–348.

Doreian, P. (1980), 'Linear Models with Spatially Distributed Data: Spatial Disturbances or Spatial Effects?', *Sociological Methods & Research* **9**(1), 29–60.

Goldstein, H. (2003), *Multilevel Statistical Models [Third Edition]*, Edward Arnold.

Harris, K., Halpern, C., Whitsel, E., Hussey, J., Tabor, J., Entzel, P. & Udry, J. (2009), 'The National Longitudinal Study of Adolescent Health: Research Design'.
URL: *http://www.cpc.unc.edu/projects/addhealth/design*

Hill, P. W. & Goldstein, H. (1998), 'Multilevel Modeling of Educational Data With Cross-Classification and Missing Identification for Units', *Journal of Educational and Behavioral Statistics* **23**(2), 117–128.

Holland, P. W. & Leinhardt, S. (1976), 'Local Structure in Social Networks', *Sociological Methodology* **7**(1976), 1.

Leenders, R. (2002), 'Modeling social influence through network autocorrelation: constructing the weight matrix', *Social Networks* **24**, 21–47.

McPherson, M., Smith-Lovin, L. & Cook, J. M. (2001), 'Birds Of A Feather: Homophily In Social Networks', *Annual Review of Sociology* **27**, 415–444.

Ord, K. (1975), 'Estimation Methods for Models of Spatial Interaction', *Journal of the American Statistical Association* **70**(349), 120– 126.

Rasbash, J., Steele, F., Browne, W. J. & Goldstein, H. (2009), 'A Users Guide to MLwiN'.

Robins, G., Pattison, P., Kalish, Y. & Lusher, D. (2007), 'An introduction to exponential random graph (p) models for social networks', *Social Networks* **29**(2), 173–191.

Robins, G., Snijders, T., Wang, P., Handcock, M. & Pattison, P. (2007), 'Recent developments in exponential random graph (p*) models for social networks', *Social Networks* **29**, 192–215.

Snijders, T. A. B. & Bosker, R. (1999), *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling*, Sage Publications: London.

Snijders, T. A. B., van de Bunt, G. G. & Steglich, C. E. G. (2010), 'Introduction to stochastic actor-based models for network dynamics', *Social Networks* **Volume 32**(Issue 1, January 2010), Pages 44–60.

Tranmer, M. & Steel, D. G. (2001), 'Ignoring a level in a multilevel model: evidence from UK census data', *Environment and Planning* **33**, 941–948.

Table 1. Triad census results for directed and undirected friendship networks. For triad definitions see: Holland & Leinhardt (1976).

Type	003	012	102	021D	021U	021C	111D	111U
Directed	149955404	572001	175813	246	311	411	309	276
Undirected	149955404	0	747814	0	0	0	0	0
type	030T	030C	201	120D	120U	120C	210	300
Directed	28	2	107	19	34	9	34	12
Undirected	0	0	1660	0	0	0	0	138

Table 2. Area (county) sample distribution.

County ID	20	22	23	25	28	29	30	31	34	35	36	37	38
Frequency	1	272	2	83	4	1	1	123	1	167	1	59	253

Table 3. School sample distribution.

School ID	17	28	44	49	85	117	144	146	149	185
Frequency	102	89	179	191	74	66	74	59	83	51

Table 4. Null Multiple Membership Models. Dependent variable: academic performance (ztotscore). Social Networks defined by undirected clique-2 and clique-3 membership.

							School	
	School	Area	Network	School	School	Area	+Area	Ind.
Param.	only	only	only	+Area	+Network	+Network	+Network	only
Constant	.04 (.10)	.03 (.10)	-.02 (.03)	.03 (.12)	.02 (.09)	.02 (.11)	.03 (.10)	.00 (.03)
School var.	.075			.059	.073		.052	
Area var.		.070		.036		.070	.036	
Network var.:								
Clique-3			.188		.148	.218	.153	
Clique-2			.139		.139	.040	.139	
Individual var.	.958	.973	.924	.958	.884	.930	.885	1.00
DIC	2714	2727	2736	2715	2700	2716	2701	2750

Table 5. Main Effects Multiple Membership Models with individual covariates. Dependent variable: academic performance (ztotscore). Social Networks defined by undirected clique-2 and clique-3 membership.

							School	
	School	Area	Network	School	School	Area	+Area	Ind.
Param.	only	only	only	+Area	+Network	+Network	+Network	only
Constant	-.09 (.07)	-.07 (.09)	-.09 (.05)	-.08 (.09)	-.10 (.07)	-09 (.09)	-.09 (.08)	-.08 (.05)
Black	-.03 (.09)	-.04 (.09)	-.07 (.09)	-.03 (.09)	-.03 (.09)	-.04 (.09)	-.03 (.09)	-.06 (.08)
Female	.17 (.06)	.17 (.06)	.17 (.06)	.17 (.06)	.17 (.06)	.17 (.06)	.17 (.06)	.17 (.06)
Age	-.13 (.02)	-.12 (.02)	-.13 (.02)	-.12 (.02)	-.12 (.02)	-.12 (.02)	-.12 (.02)	-.13 (.02)
School var.	.023			.010	.027		.012	
Area var.		.034		.029		.036	.026	
Network var.:								
Clique-3			.106		.110	.106	.102	
Clique-2			.016		.103	.099	.108	
Individual var.	.930	.927	.922	.927	.875	.875	.871	.941
DIC	2687	2682	2689	2683	2678	2674	2674	2692

Table 6. Main Effects Multiple Membership Models. Dependent variable: ztotscore. Social Networks defined by ego-net membership.

				School				School
	Network	School	Area	+Area	Network	School	Area	+Area
Param.	only	+Network	+Network	+Network	only	+Network	+Network	+Network
Constant	-.01 (.03)	.03 (.09)	.07 (.06)	.04 (.1)	-.09 (.04)	-.09 (.07)	-.07 (.09)	-.08 (.09)
Black					-.06 (.09)	-.03 (.09)	-.04 (.09)	-.03 (.08)
Female					.17 (.06)	.16 (.06)	.17 (.06)	.17 (.06)
Age					-.13 (.02)	-.12 (.02)	-.12 (.02)	-.12 (.02)
School var.		.08		.06		.03		.01
Area var.			.07	.03			.04	.03
Network var.:								
Ego	.17	.20	.15	.19	.15	.15	.15	.15
Individual level:	.926	.870	.902	.873	.873	.861	.856	.862
DIC	2737	2697	2713	2697	2681	2674	2669	2672

Table 7. Multiple Membership Models: null and with individual covariates. Dependent variable: health good/excellent.

	School	School		School	School	
	+Area	+Area	Ind.	+Area	+Area	Ind.
Param.	+Network	+Network	only	+Network	+Network	only
Constant	1.20 (.29)	1.157 (.243)	.98 (.07)	1.28 (.26)	1.27 (.29)	1.1 (.11)
Black				.07 (.22)	.09 (.22)	.14 (.20)
Female				-.31 (.14)	-.31 (.15)	-.28 (.15)
Age				-.05 (.05)	-.05 (.05)	-.07 (.04)
School var.	.109	.101		.152	.129	
Area var.	.301	.280		.198	.258	
Network var.:						
Clique-3	.095			.090		
Clique-2	.099			.018		
Ego		.000			.008	
DIC	1120	1118	1138	1120	1119	1137

Table 8. Multiple Membership (MM) and Network Disturbance (ND) models with school indicators as fixed covariates ; Full model also includes all individual level covariates. Dependent variable: academic performance (ztotscore). reference school = School 17. Dependent variable = ztotscore

	Constant + School Indicators		Full Model	
	MM	ND	MM	ND
Parameter				
Constant	-.004 (.110)	.001 (.104)	.030 (.115)	.034 (.110)
School 28	.259 (.153)	.235 (.152)	.076 (.160)	.043 (.159)
School 44	-.196 (.128)	-.205 (.129)	-.174 (.134)	-.180 (.128)
School 49	-.107 (.131)	-.136 (.128)	-.112 (.130)	-.131 (.126)
School 85	-.327 (.151)	-.342 (.156)	-.296 (.156)	-.307 (.154)
School 117	.344 (.165)	.322 (.164)	-.012 (.178)	-.044 (.178)
School 144	.138 (.156)	.123 (.158)	-.185 (.173)	-.207 (.170)
School 146	.407 (.167)	.388 (.167)	.050 (.185)	.022 (.180)
School 149	.258 (.151)	.251 (.152)	-.083 (.167)	-.099 (.166)
School 185	-.327 (.173)	-.327 (.173)	-.612 (.189)	-.626 (.186)
Black			.006 (.092)	.015 (.093)
Female			.165 (.062)	.173 (.062)
Age			-.119 (.025)	-.125 (.025)
Ego-net var.	.193		.165	
Individual var.	.872		.855	
DIC	2698		2674	
$\hat{\rho}$		.107 (.041)		.111 (.041)
AIC		2711		2675
$\hat{\sigma}^2$		.917		.883