Centre for Statistical & Survey Methodology
Working Paper Series

Faculty of Engineering and Information Sciences

2012

Using Social Network Information for Survey Estimation

Thomas Suesse
University of Wollongong, tsuesse@uow.edu.au

Ray Chambers
University of Wollongong, ray@uow.edu.au

Recommended Citation

Suesse, Thomas and Chambers, Ray, Using Social Network Information for Survey Estimation, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 11-12, 2012, 33.
http://ro.uow.edu.au/cssmwp/100

CSSM
Centre for Statistical & Survey Methodology

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

11-12

Using Social Network Information for Survey Estimation

Thomas Suesse and Ray Chambers

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

Using Social Network Information for Survey Estimation

Thomas Suesse
Centre for Statistical and Survey Methodology
University of Wollongong, Wollongong, Australia
E-mail: `tsuesse@uow.edu.au`

Ray Chambers
Centre for Statistical and Survey Methodology
University of Wollongong, Wollongong, Australia
E-mail: `ray@uow.edu.au`

Summary. Standard model-based and model-assisted methods of survey estimation aim to improve the precision of estimators of the population total or mean. These methods are often based on a linear regression model defined in terms of auxiliary variables whose values are assumed known for all population units. Friendships and other social relationships represent another form of auxiliary information that might increase the precision of these estimators. Such relationships are typically expressed in terms of a social network. Common linear models that use social networks as an additional source of information include autocorrelation, disturbance and contextual models. In this paper we investigate how much of the population network needs to be known for estimation methods based on these models to be useful. In particular, we use simulation to compare the performance of the best linear unbiased predictor under a model that ignores the network with model-based estimators that incorporate network information. Our results show that incorporating network information via a contextual model is the best performer overall. We also show that the full population network is not required, but that the partial network linking the sampled population units to the non-sampled population units needs to be known. Finally, we illustrate the contextual model by applying it to friendship network information collected in the British Household Panel Study.

Keywords: BLUP, social network models, linear models, model-based survey

estimation

1 Introduction

Model-based and model-assisted methods of survey estimation, see Chambers and Clark (2012) and Srndal et al. (1992), are commonly used to increase the precision of estimators of the population mean or total of a survey variable. Typically, a linear regression model linking the population values of this variable to corresponding values of a set of auxiliary variables is assumed, and sample weights are calibrated against the known population totals of these auxiliary variables. Commonly used auxiliary variables are individual level variables like age and ethnicity. Model-based methods based on the best linear unbiased predictor (BLUP) can also adjust for correlation between observations, e.g. the correlation between the values of the survey variable for members of the same household, to gain further precision, see Chapter 7 of Chambers and Clark (2012).

The British Household Panel Study (BHPS, `http://www.iser.essex.ac.uk/survey/bhps/`) is an annual longitudinal survey of British households that has been conducted from 1991. It is based on a representative sample of approximately $5,500$ households, covering more than $10,000$ individuals. The main objective of the survey is to further the understanding of social and economic change at the individual and household level in Britain.

In addition to information about the surveyed individual, the BHPS provides information about his or her three closest friends. Variables collected on the three closest friends are: age, sex, ethnicity, distance to friend (< 1 mile, $1-5$ miles, $5-50$ miles, > 50 miles) and unemployment status. This information is available in seven waves, corresponding to the even-numbered years 1992 - 2004. Because friends tend to share common characteristics, it is plausible that the friendship ties provided by the BHPS are important in modelling the survey variables. Consequently this extra information should not be excluded in the survey estimation process.

Friendship ties is a typical example of a (social) network that connects individuals. It is standard to represent a network by a matrix of zeros and ones, $\mathbf{Z} = (Z_{ij})_{i,j=1}^{N}$ with $Z_{ii} = 0$ by convention. If a relationship exists between two individuals i and j, then $Z_{ij} = 1$, otherwise $Z_{ij} = 0$. A network is said to be undirected if $\mathbf{Z} = \mathbf{Z}'$, otherwise it is a directed network.

Linear models that use a social network as additional information to model a response variable include contextual network models (Friedkin, 1990), network effects models, also known as autocorrelation (AR) models, and network disturbance models (Ord, 1975; Doreian et al., 1984; Duke, 1993; Marsden and Friedkin, 1993; Leenders, 2002). Linear versions of contextual network models assume that the individual response variable Y_i depends on individual level covariates, such as the individual's age and sex, as well as on a network covariate, e.g. the average age of the friends of individual i. AR models assume that Y_i depends on the average response Y_j of those j individuals named as friends by individual i. Network disturbance models are another class of AR models, where the AR structure holds for model residuals, i.e. the residual r_i is assumed to depend linearly on the average r_j of the residuals of the j individuals named as friends by individual i.

When the social network $\mathbf{Z}$ is known for all N individuals in the population, the contextual network, AR and network disturbance models can all be fitted. In the case of the contextual network model, this is because knowledge of the population values of the covariates and the population network $\mathbf{Z}$ implies that one can compute any function of these covariates (including their average for the j individuals connected to individual i in the network). In the case of the AR and network disturbance models, this is because knowledge of the complete network allows one to re-express these models in a marginal form where the covariance matrix of the population values of the response variable depends only on $\mathbf{Z}$.

In practice it is unlikely that $\mathbf{Z}$ will be known, and a more realistic scenario is one where only the sub-network $\mathbf{Z}_{ss}$ defined by the individuals in the sample s of size n is known, or where this sub-network as well as the sub-network $\mathbf{Z}_{sr}$ of links between the sampled individuals and the remaining $N - n$ non-sampled individuals, denoted by r, is known. That is, the sub-network defined by the links of the non-sampled individuals with the sampled individuals, $\mathbf{Z}_{rs}$, and the sub-network defined by the links within the $N - n$ non-sampled population units, $\mathbf{Z}_{rr}$, will be unknown. Clearly, under symmetry, $\mathbf{Z}_{rs}(= \mathbf{Z}'_{sr})$ will be known.

When such partial network information is observed, one needs to either use more complicated fitting methods or to impute the missing network components. For example, the BHPS provides information on a limited range of characteristics of the three best friends, but does not identify friends within or outside the sample. In this case $\mathbf{Z}$ is unknown and application of the AR and network disturbance models is not possible. However the average of the collected variables (age, sex, ethnicity etc.) for a person's three best friends can be calculated and the contextual network model can be fitted.

The main focus of this paper is on the use of network information in survey estimation. In particular, we aim to address the questions: (i) Is embedding network information useful for survey estimation? (ii) If the answer to (i) is yes, then which models are potentially useful? and (iii) How much network data needs to be collected in order to obtain potentially higher precision for survey estimation? In this context, Section 2 defines the contextual network, AR and network disturbance models. In Section 3 we then briefly discuss the EBLUP estimator of the population mean, and its application under these models. In Section 4, the exponential random graph model (ERGM) for a network is introduced and its use in imputation of missing network information is described. Section 5 then describes a small-scale simulation study that investigates the performance of the EBLUP both when the network information is ignored and when either all or part of the network is used. In Section 6 we use data from wave N (2004) of the BHPS to illustrate age by sex by region estimation of population means based on a model that includes age by sex effects and a contextual variable corresponding to the maleness proportion of an individual's three best friends. Section 7 completes the paper with a discussion of our findings as they relate to the three questions raised above.

2 Linear Network Models

In this section we describe a number of population level linear models that use network information. Throughout, we use a friendship social network structure for simplicity of exposition. In order to develop our notation, the starting point is the linear model that assumes uncorrelated errors.

2.1 The Standard Model

The classical linear model for a population of N individuals has the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_N)'$ is a vector of responses, $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_N')'$ with $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$ is the model design matrix with p rows defined by a set of covariates that depend on auxiliary population information, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)'$ is the vector of residuals with $\epsilon_i \sim N(0, \sigma^2)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the vector of regression coefficients. The population mean vector and population covariance matrix of $\mathbf{Y}$ are then $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{V} = \sigma^2\mathbf{I}_N$. Here $\mathbf{I}_N$ denotes the identity matrix of order N.

This model does not use social network information. In the rest of this section we therefore consider models where the response variable depends linearly on a set of known covariates but also uses a social network as an additional source of information.

2.2 The Contextual Network Model

Consider an educational modelling exercise where academic performance (AP) is the response variable and socio-economic status (SES) of the student is the explanatory variable. A classical contextual approach might then lead one to include the average SES of the student's school calculated over all students attending this school as another explanatory variable. This could be motivated by the argument that the extra educational resources available to a working-class student attending a school in an affluent suburb should lead to a higher AP value than a similar student attending a school in a poor neighbourhood. However Duke (1993), see Subsection 2.1, argues that such models are not appropriate, because the "socio-economic status of students in a school is a poor proxy of the actual content of the interpersonal influences to which a student is exposed". Given the friendship network of a student, this immediately suggests the alternative contextual model

$$AP_i = \beta_0 + \beta_1 SES_i + \beta_2 \overline{SES}_i + \epsilon_i, \tag{2}$$

where index i stands for the ith student and $\epsilon_i \sim N(0, \sigma^2)$. Here the academic performance of individual i depends linearly on the average level of socio-economic status of his/her friends (denoted $\overline{SES}_i$ above) and linearly on his/her socio-economic status, see Friedkin (1990). In general, this type of contextual model has the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \bar{\mathbf{X}}\bar{\boldsymbol{\beta}} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_N), \tag{3}$$

where $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ have the same meaning as for model (1). The model includes the contextual variables contained in the matrix $\bar{\mathbf{X}}$ with the vector $\bar{\boldsymbol{\beta}}$ denoting the associated contextual effects. In the above example $\mathbf{X}$ contains two columns, the first is a column of ones for the intercept and the second contains SES_i, and $\bar{\mathbf{X}} = (\overline{SES}_1, \ldots, \overline{SES}_N)'$ is the vector containing the contextual socio-economic status of all N students in the study population. The population mean vector and population covariance matrix of $\mathbf{Y}$ in this case are $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}$ and $\mathbf{V} = \sigma^2 \mathbf{I}_N$.

Let $\boldsymbol{\Xi}$ be the matrix containing the population values of the relevant contextual variables. In the above example, $\boldsymbol{\Xi}$ is the vector of values SES_i. In general, the set of covariates used in $\mathbf{X}$ can be different from those in $\boldsymbol{\Xi}$, so the contextual variables contained in $\bar{\mathbf{X}}$ are not necessarily based on $\mathbf{X}$. In particular, we then have

$$\bar{\mathbf{X}} = \mathbf{W}\boldsymbol{\Xi} \tag{4}$$

with $\mathbf{W} = \text{Diag}(\mathbf{Z}\mathbf{1}_N)^{-1}\mathbf{Z}$, where $\mathbf{1}_N$ is a column vector of ones of length N.

Remark
A contextual variable for person i often includes the value for this person, for example a household contextual effect is computed over all household members including person i. However, the contextual value for person i defined by (4) excludes person i, because $Z_{ii} = 0$ by definition.

2.3 The Autocorrelation Model

Autocorrelation (AR) models, also known as network effects models, are another class of models that incorporate network information into a linear structure. See for example Doreian et al. (1984), Duke (1993), Marsden and Friedkin (1993) and Leenders (2002), and in the context of spatial models, Ord (1975).

Continuing with the academic performance example introduced in the previous sub-section, Duke (1993) considered the model where

$$AP_i = \beta_0 + \beta_1 SES_i + \theta\overline{AP}_i + \epsilon_i,$$

depends linearly on the socio-economic status SES_i of student i and the average academic performance, denoted $\overline{AP}_i$, of his/her friends.

As before let $\mathbf{Y}$ be the vector of responses, $\mathbf{X}$ the design matrix depending on a set of covariates and $\boldsymbol{\epsilon}$ be a vector of residuals. This model is usually referred to as an AR model and can be expressed more generally as

$$\mathbf{Y} = \theta\bar{\mathbf{Y}} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_N), \tag{5}$$

where $\bar{\mathbf{Y}} = (\bar{Y}_1, \ldots, \bar{Y}'_N)'$ and $\bar{Y}_i$ stands for the average response of the friends of individual i, so $\bar{\mathbf{Y}} = \mathbf{W}\mathbf{Y}$, with $\mathbf{W}$ defined in the previous sub-section. Typically $\mathbf{W}$ is row-normalised, i.e. $\sum_{j=1}^{N} W_{ij} = 1$, implying the parameter θ is restricted to the open interval $(-1, +1)$.

The mean and variance of $\mathbf{Y}$ under (5) are $\boldsymbol{\mu} = \mathbf{U}^{-1}\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{V} = \sigma^2(\mathbf{U}'\mathbf{U})^{-1}$ with $\mathbf{U} = \mathbf{I}_N - \theta\mathbf{W}$. Note that $\mathbf{W}$ can be defined in a variety of ways, see Leenders (2002).

2.4 The Network Disturbance Model

Models of this type have been considered by Ord (1975) and Leenders (2002) among others, and correspond to imposing an AR structure on the linear model residuals. They are referred to as network disturbance models and are specified by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \theta\bar{\boldsymbol{\epsilon}} + \mathbf{v}, \mathbf{v} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_N). \tag{6}$$

where $\bar{\boldsymbol{\epsilon}} = (\bar{\epsilon}_1, \ldots, \bar{\epsilon}_N)$ and $\bar{\epsilon}_i$ is the average residual of the friends of person i.

In the context of the academic performance example considered in the preceding sub-sections, this implies that AP_i, the academic performance of the ith student, depends linearly on SES_i and that the associated error ϵ_i depends linearly on the average $\bar{\epsilon}_i$ of the errors of the friends of student i.

The model (6) is equivalent to

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2(\mathbf{U}'\mathbf{U})^{-1}) \tag{7}$$

with $|\theta| < 1$. It implies that the mean and variance of the response vector $\mathbf{Y}$ are then $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{V} = \sigma^2(\mathbf{U}'\mathbf{U})^{-1}$ respectively, where $\mathbf{U}$ is defined at the end of the previous sub-section.

Note that the mean $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ under (7) is unaffected by the social network, whereas under the AR model (5) the mean $\boldsymbol{\mu} = \mathbf{U}^{-1}\mathbf{X}\boldsymbol{\beta}$ depends on the network through $\mathbf{U}$. The parameter θ is an indicator for the strength of the correlations imposed by the network. For $\theta = 0$, these correlations are zero and $\mathbf{V}$ equals the variance under the standard model, i.e. $\mathbf{V} = \sigma^2\mathbf{I}_N$. Consequently it makes more sense to refer to the network disturbance model as a network covariance model, because the network only affects the covariance structure of the response and not its mean.

2.5 Other Network Covariance Models

Although the AR and network disturbance models are popular ways of representing the influence of a network on a response, they are not the only ways that one can achieve this aim. Other network covariance models can be constructed. For example, one can assume that the correlation between two responses Y_i and Y_j is θ if $Z_{ij} = 1$, and zero otherwise, and let the subject-level variance be σ^2. This model has the form

$$\mathbf{Y} \sim N\left(\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \mathbf{V} = \sigma^2\mathbf{I}_N + \theta\mathbf{Z}\right) \tag{8}$$

and reflects the idea that the responses of two people with a social connection ($Z_{ij} = 1$) should be more alike in terms of their responses than two people who are not connected ($Z_{ij} = 0$). To maintain the positive definiteness and symmetry of $\mathbf{V}$, $\mathbf{Z}$ needs to be symmetric and the parameter space for θ must be restricted to small values. Furthermore, in order to allow for the hypothesis that many friends have individually less influence on a person than a single friend, one could use $\mathbf{W}$ instead of $\mathbf{Z}$ in (8), reducing each individual's contribution to the correlation from θ to θ divided by the number of friends. Symmetry can be guaranteed by replacing $\mathbf{W}$ by $(\mathbf{W} + \mathbf{W}')/2$, which in turn implies that an only friend j of individual i who in turn has many friends cannot have as much influence on i as an only friend j who has only i as a friend.

3 Prediction of Population Totals Using Network Models

The models discussed in the previous section are all predictive models, i.e. when second order moments are known, they can be used to compute efficient predictions of unknown values of the response variable. Here we are specifically interested in using these models to predict the value of the population total $t = \sum_{i \in P} Y_i = \mathbf{1}'_N \mathbf{Y}$ (and hence the mean t/N) from knowledge of the sample values in $\mathbf{Y}$, the matrix of model covariates $\mathbf{X}$ and either part of or all of the network matrix $\mathbf{Z}$. Throughout we assume that inclusion in sample does not depend on $\mathbf{Z}$ and that there is non-informative sampling given $\mathbf{X}$, see Section 1.4 of Chambers and Clark (2012). Consequently, all unknown parameter values for the standard model (1) can be estimated from the sample data and predicted values of Y for the non-sampled population individuals can be computed. In this section we discuss how the more complicated network models introduced in the previous section can also be fitted to the sample data and conclusions drawn about the unknown value of t.

3.1 The Empirical Best Linear Unbiased Predictor

The best linear unbiased predictor or BLUP, see Royall (1976), is an efficient estimator of the population total t that only requires specification of the first and second order moments of $\mathbf{Y}$ given $\mathbf{X}$ and $\mathbf{Z}$. It assumes that the mean is linear, i.e. $\mathbb{E}\mathbf{Y} = \boldsymbol{\mu} = \mathbf{T}\boldsymbol{\beta}$, and that $\mathrm{Var}(\mathbf{Y}) = \mathbf{V}$ is known up to a constant of proportionality. Here $\mathbf{T}$ is a well-defined function of $\mathbf{X}$ and $\mathbf{Z}$.

Let s and r denote the n sampled and $N - n$ non-sampled population individuals respectively. Put $\mathbf{T} = (\mathbf{T}_s', \mathbf{T}_r')'$ and $\mathbf{Y} = (\mathbf{Y}_s', \mathbf{Y}_r')'$. The matrix $\mathbf{V}$ can be partitioned conformably as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{pmatrix}.$$

The BLUP is then the weighted sum $\hat{t}_{BLUP} = \mathbf{w}_s'\mathbf{y}_s$ of the sample values of the response, where

$$\mathbf{w}_s = \mathbf{1}_n + \mathbf{V}_{ss}^{-1}\left\{\mathbf{V}_{sr}\mathbf{1}_{N-n} - \mathbf{T}_s(\mathbf{T}_s'\mathbf{V}_{ss}^{-1}\mathbf{T}_s)^{-1}\left(\mathbf{T}_s'\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr} - \mathbf{T}_r'\right)\mathbf{1}_{N-n}\right\} \tag{9}$$

is the vector of BLUP weights. Here $\mathbf{1}_n$ and $\mathbf{1}_{N-n}$ are vectors of ones of size n and $N-n$ respectively. An alternative expression for the BLUP is its so-called predictive form

$$\hat{t}_{BLUP} = \sum_{i\in s} Y_i + \sum_{i\in r} \hat{Y}_i + \sum_{i\in s} \gamma_i(Y_i - \hat{Y}_i) \tag{10}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{T}_s'\mathbf{V}_{ss}^{-1}\mathbf{T}_s)^{-1}\mathbf{T}_s'\mathbf{V}_{ss}^{-1}\mathbf{Y}_s$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$, γ_i is the ith element of the vector $\mathbf{1}_{N-n}\mathbf{V}_{sr}'\mathbf{V}_{ss}^{-1}$ and $\hat{Y}_i = \mathbf{T}_i\hat{\boldsymbol{\beta}}$.

A key assumption of the BLUP is that the variance matrix $\mathbf{V}$ is known. This is usually incorrect, as the variance matrix $\mathbf{V}$ can depend on unknown parameters, which must be estimated. Substituting these estimates in $\mathbf{V}$ defines the plug-in estimator $\hat{\mathbf{V}}$, which is then used in (9) instead of $\mathbf{V}$. The resulting estimator of the population total is called the empirical BLUP (EBLUP).

In order to apply the EBLUP to the different network models defined in the previous section, we need to specify $\mathbf{T}$ and $\mathbf{V}$ as well as estimators of the unknown parameters that underpin these matrices. These are defined as follows:

Standard Model : Here $\mathbf{T} = \mathbf{X}$ and $\mathbf{V} = \sigma^2\mathbf{I}_N$. The residual mean squared error defines an unbiased estimator of σ^2.

Contextual Network Model : For the contextual model $\mathbf{T} = [\mathbf{X}, \bar{\mathbf{X}}]$, where $\bar{\mathbf{X}} = \mathbf{W}\boldsymbol{\Xi}$ and $\mathbf{W} = \text{Diag}(\mathbf{Z}\mathbf{1}_N)^{-1}\mathbf{Z}$. Also, $\mathbf{V} = \sigma^2\mathbf{I}_N$ and we can unbiasedly estimate σ^2 using the residual mean squared error.

Network Covariance Model : Here $\mathbf{T} = \mathbf{X}$ and $\mathbf{V} = \sigma^2\mathbf{I}_N + \theta(\mathbf{W} + \mathbf{W}')/2$. Estimates of σ^2 and θ can be obtained via restricted maximum likelihood estimation (REML) using the iterative generalised least squares algorithm (IGLS), see Goldstein (1989).

Autocorrelation Model : In this case the pseudo-design matrix $\mathbf{T} = \mathbf{U}^{-1}\mathbf{X}$ with $\mathbf{U} = \mathbf{I}_N - \theta\mathbf{W}$ and $\mathbf{V} = \sigma^2\mathbf{U}^{-1}(\mathbf{U}^{-1})'$. Estimates of σ^2 and θ can be obtained by maximum likelihood (ML). Note that REML cannot be used here, because the mean and variance depend on the parameter θ.

Network Disturbance Model : Here $\mathbf{T} = \mathbf{X}$ and $\mathbf{V} = \sigma^2\mathbf{U}^{-1}(\mathbf{U}^{-1})$. ML estimation of σ^2 and θ can be carried out.

ML estimation for the AR and network disturbance models is not straightforward. Both models are not reproducible, i.e. they do not share the property that the model for a subset of units of the population has the same form as the model for the whole population. To see this, note that the variance of the population response vector $\mathbf{Y}$ under both models is $\sigma^2(\mathbf{U}'\mathbf{U})^{-1}$ so that the variance for the sample response vector $\mathbf{Y}_s$ is $\sigma^2[(\mathbf{U}\mathbf{U})^{-1}]_{ss}$. In general, this will not equal $\sigma^2(\mathbf{U}_{ss}'\mathbf{U}_{ss})^{-1}$, which is the assumed variance if the model is fitted via ML at the sample level. This misspecification can lead to biased estimates of the model parameters. A modified approach that yields unbiased estimates of the fixed effects in the model is described in Suesse (2012). However this is computationally intensive. An alternative approach replaces $\mathbf{U}^{-1}$ by a 4th order Taylor series approximation. This speeds up computation considerably since it effectively replaces matrices of dimension $N \times N$ by matrices of dimension $n \times n$. See Suesse (2012) where it is shown that ML estimates based on this approximation are essentially identical to those obtained using the modified ML method.

3.2 Variance Estimation for the EBLUP

The prediction variance of the BLUP is

$$\mathrm{Var}(\hat{t}_{BLUP} - t) = \tilde{\mathbf{w}}'\mathbf{V}\tilde{\mathbf{w}} \tag{11}$$

with $\tilde{\mathbf{w}} = (\mathbf{w}_s - \mathbf{1}_n, -\mathbf{1}_{N-n})$. This formula assumes that the vector of survey weights $\mathbf{w}_s$ is fixed. We can use the same formula for the EBLUP, although from (9) it is apparent that the EBLUP weights are not fixed, because $\mathbf{V}$ is estimated, as is the design matrix for the AR model.

In general for a set of parameters $\boldsymbol{\eta}$ describing a model, the difference between the true and the ML/REML estimates of the parameter values approaches zero for large n provided the model holds, i.e. $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} = o(n^{-1/2})$. In what follows, we assume that the sample size is large, so it is reasonable to assume that the difference between the BLUP weights (9) and their corresponding EBLUP approximations can be ignored, justifying use of the formula (11) with EBLUP weights as a first order approximation to the prediction variance of the EBLUP.

Using (11) to estimate this prediction variance depends on estimation of $\mathbf{V}$. For the standard model and the contextual network model, we use a robust version of the resulting variance estimator, see Section 9.2 of Chambers and Clark (2012), given by

$$\widehat{\mathrm{Var}}(\hat{t}_{BLUP} - t) = \sum_{i \in s} (w_{is} - 1)^2 (Y_i - \hat{\mu}_i)^2 + (N - n)\hat{\sigma}^2,$$

where $\hat{\mu}_i$ is the predicted mean for $i \in s$, i.e. $\hat{\mu}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}}$ for the standard model and $\hat{\mu}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}} + \bar{\mathbf{X}}_i \hat{\hat{\boldsymbol{\beta}}}$ for the contextual network model, with $\hat{\sigma}^2$ corresponding to the usual unbiased estimator of σ^2 under each model.

For the other models we use equation (11) with a plug-in estimator $\hat{\mathbf{V}}$. In this context, we note that ML estimates of variance parameters are known to be biased, which could therefore lead to a bias in $\hat{\mathbf{V}}$ and in the resulting plug-in estimator defined by (11). The standard approach to dealing with this issue is to apply REML instead of ML. Unfortunately, the AR model does not allow the application of REML, and furthermore REML is computationally more complex when fitting these population models. Consequently a bias-corrected version of ML was applied, based on the approach set out in Goldstein (1989) which adjusts IGLS to obtain estimates that are equivalent to REML. The details of this are outlined in the Appendix.

4 Modelling and Imputation of Social Networks

The network models discussed in the previous Section all imply a population mean vector $\boldsymbol{\mu}$ or a population variance matrix $\mathbf{V}$ that depends on the population network $\mathbf{Z}$. Furthermore, our development of the EBLUP assumed that $\mathbf{Z}$ is known. However this is extremely unlikely. It is far more likely that we know either just that part of the network defined by the sampled individuals (i.e. $\mathbf{Z}_{ss}$) or that part of the network involving the sampled individuals (i.e. $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$). Consequently we now consider methods of estimation that take account of this incomplete data. In particular, we develop model-based imputation methods that can be used to construct an estimate $\hat{\mathbf{Z}}$ of the full network. We start by discussing models for networks.

4.1 Exponential Random Graph Models

The most popular class of models for a network $\mathbf{Z}$ is the class of (curved) exponential random graph models (ERGMs). See Wasserman and Faust (1994) and Carrington et al. (2005). Under an ERGM, the distribution of $\mathbf{Z}$ is characterised by

$$\Pr(\mathbf{Z} = \mathbf{z}) = \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})'\mathbf{G}(\mathbf{z}) - \kappa(\boldsymbol{\theta})\right), \tag{12}$$

where $\boldsymbol{\theta}$ is the vector of model parameters, $\boldsymbol{\eta}(\boldsymbol{\theta})$ is a mapping from p-dimensional to q-dimensional space with $p \leq q$, and

$$\kappa(\boldsymbol{\theta}) = \log\left\{\sum_{\mathbf{z}\in\Omega} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})'\mathbf{G}(\mathbf{z}))\right\}$$

is the normalising constant where the sum is over the sample space Ω of all $\exp(\binom{n}{2}\log 2)$ undirected networks. Here $\mathbf{G}(\mathbf{z})$ is a vector of network statistics which, together with $\boldsymbol{\theta}$, completely characterises the joint distribution of the network.

Fitting such models is complicated, mainly because direct calculation of $\kappa(\boldsymbol{\theta})$ is typically infeasible. One way of circumventing this problem is to sample from the network distribution (12) using a Markov-Chain-Monte-Carlo (MCMC) algorithm in order to obtain a stochastic approximation to the maximum likelihood estimate of $\boldsymbol{\theta}$. Such estimates are called MCMC ML estimates (Hunter and Handcock, 2006). Describing the network distribution via simple network statistics, such as the number of triangles (a triangle is said to exist between individuals i, j and k, if $Z_{ij} = 1$, $Z_{jk} = 1$, $Z_{ik} = 1$), then becomes problematic, because such specifications often lead to degenerate MCMC samples. Some authors (Snijders, 2002; Snijders et al., 2006) have therefore proposed the use of more complex network statistics, such as the geometrically weighted edgewise shared partner (GWESP) statistic, for which degeneracy seems less of a problem. For more details of social network modelling, see Strauss and Ikeda (1990); Hunter and Handcock (2006); Hunter (2007); Hunter et al. (2008) and Butts (2008).

4.2 Imputation of Partly Observed Networks

As mentioned earlier, an estimate $\hat{\mathbf{Z}}$ of the full network is necessary for calculation of the EBLUP under the network models considered in this paper. However, in practice only part of network will be observed, say $\mathbf{Z}^{obs}$, and another part will be missing, say $\mathbf{Z}^{mis}$. For example, the observed network

$\mathbf{Z}^{obs}$ could be $\mathbf{Z}_{ss}$, in which case the missing network $\mathbf{Z}^{mis}$ is $\mathbf{Z}_{sr} \cup \mathbf{Z}_{rs} \cup \mathbf{Z}_{rr}$. In what follows we assume an undirected network, i.e. $\mathbf{Z} = \mathbf{Z}'$, so $\mathbf{Z}_{sr} = \mathbf{Z}_{rs}$. We also focus on single-value imputation of $\mathbf{Z}^{mis}$. Our approach can be extended to multiple imputation.

A model-based approach to imputing the missing network components is based on the minimum mean square error predictor $\mathbb{E}(\mathbf{Z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs}; \boldsymbol{\theta})$. Note that the expectation here is with respect to the ERGM model (12), with unknown parameters replaced by ML estimates. To compute $\mathbb{E}(\mathbf{Z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs}; \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{ML})$ we need the distribution of $\mathbf{Z}^{mis}|\mathbf{Z}^{obs}$ under an ERGM, which is

$$\Pr(\mathbf{Z}^{mis} = \mathbf{z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs}; \boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})'\mathbf{G}((\mathbf{z}^{mis}, \mathbf{z}^{obs}))\right)}{\sum_{\mathbf{z}^{mis} \in \Omega^{mis}} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})'\mathbf{G}((\mathbf{z}^{mis}, \mathbf{z}^{obs}))\right)}. \tag{13}$$

Here Ω^{mis} is the sample space of the missing networks, which is typically extremely large. We therefore consider three methods of approximating this conditional expectation.

Method 1

The first method is via MCMC, for example the Metropolis-Hasting algorithm. Given a sample $\mathbf{z}_1^{mis}, \ldots, \mathbf{z}_m^{mis}$ from $\mathbf{Z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs}$, we can approximate $\mathbb{E}(\mathbf{Z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs})$ by $\frac{1}{m}\sum_{i=1}^{m} \mathbf{z}_i^{mis}$. Unfortunately this method is not feasible for the larger networks that occur in surveys because the computation of the network statistics $\mathbf{G}(\mathbf{z})$ is very time-consuming. As a consequence we do not consider this method further in this paper. Note however that for a single small to medium size data set (say $N \leq 2,000$), sampling from the conditional distribution (13) is recommended.

Method 2

The second method represents a simpler, more feasible, approach. Suppose conditionally on $\mathbf{z}^{obs}$ that Z_{ij}^{mis} and Z_{kl}^{mis} are conditionally independent for any two distinct pairs of individuals i, j and k, l, where both pairs are in mis and by distinct we mean that $(i, j) \neq (k, l)$ and $(i, j) \neq (l, k)$ hold. Then $\Pr(\mathbf{Z}^{mis} = \mathbf{z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs}) = \prod_{ij \in mis} \Pr(Z_{ij} = z_{ij}|\mathbf{Z}^{obs} = \mathbf{z}^{obs})$, where $\prod_{ij \in mis}$ denotes the product over which all distinct pairs is in the set mis. It follows that we can write, for a distinct pair $(i, j) \in mis$,

$$\frac{\Pr(Z_{ij} = 1|\mathbf{Z}^{obs} = \mathbf{z}^{obs})}{\Pr(Z_{ij} = 0|\mathbf{Z}^{obs} = \mathbf{z}^{obs})} = \exp(\boldsymbol{\eta}^T(\boldsymbol{\theta})\Delta\mathbf{G}_{ij}^{mis}), \tag{14}$$

where $\Delta\mathbf{G}_{ij}^{mis}$ is the change statistic, i.e. the difference in $\mathbf{G}$ between $(Z_{ij}, \mathbf{z}^{mis-(i,j)}, \mathbf{z}^{obs}) = (1, \mathbf{z}^{mis-(i,j)}, \mathbf{z}^{obs})$ and $(Z_{ij}, \mathbf{z}^{mis-(i,j)}, \mathbf{z}^{obs}) = (0, \mathbf{z}^{mis-(i,j)}, \mathbf{z}^{obs})$. Note that $mis - (i,j)$ here denotes the set mis with the distinct pair (i,j) excluded. It remains to observe that (14) implies that it is only necessary to compute $\Delta\mathbf{G}_{ij}^{mis}$ in order to obtain $\mathbb{E}(Z_{ij}|\mathbf{Z}^{obs} = \mathbf{z}^{obs}) = \Pr(Z_{ij} = 1|\mathbf{Z}^{obs} = \mathbf{z}^{obs})$ for any distinct pair $(i,j) \in mis$. Since the conditional independence assumption underpinning this method is generally unwarranted, it can only be considered as approximating $\mathbb{E}(\mathbf{Z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs})$. However, it is computationally feasible for realistic sample and population sizes.

Method 3

An even simpler approach is to calculate the proportion of $Z_{ij} = 1$ in $\mathbf{z}^{obs}$ and use this proportion to impute $\mathbf{z}^{mis}$. This corresponds to fitting an ERGM model with just the EDGES statistic to the network, which in turn is equivalent to assuming that each Z_{ij} in the network matrix $\mathbf{Z}$ is an independent Bernoulli variable with a common probability of a success.

5 Simulation Study

5.1 Study Design

The aim of the study is to investigate the effect of using networks as an additional source of information when estimating the population total t of a survey variable Y. A population size of $N = 1,000$ is assumed, balancing computation time against the number of different scenarios that are explored in the study. Sample sizes were set at $n = 100$ and $n = 200$, and simple random sampling without replacement was employed.

Network Generation

Two types of networks were investigated. In the first scenario (ERGM network), $\mathbf{Z}$ was generated via an ERGM with an EDGES statistic equal to -4.18 and a weight parameter of 1.0 for the GWESP statistic distribution. These values were chosen in order to generate a realistic network density of approximately 15 network connections for each subject.

The second scenario (Gang network) simulates a network consisting of 100 gangs each of size 10. It is assumed that each gang member only knows every other gang member of his/her gang. In this case $\mathbf{Z}$ is a block diagonal matrix with 100 blocks, with each block equal to the identity matrix of order 10 minus a 10×10 matrix of ones.

Figures 1 and 2 show a simulated ERGM network and the Gang network. Note that the ERGM network only shows the network connections between 300 randomly chosen individuals out of the population of $N = 1,000$. On average every individual has 7.541 connections for the ERGM network. By definition, every individual has 9 connections for the Gang network.

Imputation of Partly Observed Networks

In the simulation study we restricted ourselves to two realistic scenarios where part of the network is unobserved and so must be imputed. In the first scenario, denoted by SS in what follows, only $\mathbf{Z}_{ss}$ is observed and $\mathbf{Z}_{sr}$ and $\mathbf{Z}_{rr}$ are missing. This could be the case where clusters of individuals are sampled, and the components of the network defined by the individuals making up each cluster are measured (assuming that there are no network connections between individuals in different clusters). In the second scenario, denotes SS-SR in what follows, $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ are observed and $\mathbf{Z}_{rr}$ is missing. This is a more realistic situation, where network information relating to all individuals in the population is collected from the sampled individuals.

For the ERGM network both imputation Method 2 and imputation Method 3 lead to the same imputed value of $\mathbf{Z}^{mis}$ in the SS scenario. In contrast, these methods lead different imputed values in the SS-SR scenario under the ERGM network. We therefore denote the application of imputation Method 2 in the SS-RS scenario for the ERGM network by SS-SR-C, and the corresponding application of imputation Method 3 by SS-SR-S. The SS, SS-SR-C and SS-SR-S imputation scenarios were all considered in our simulation of the ERGM network.

In contrast, imputation of the Gang network can be carried out relatively effectively without using the methods described in the previous Section. This is because the gang structure implies transitivity, i.e. in many situations the true value of Z_{ij} for $(i, j) \in mis$ can be deduced. When no conclusion can be drawn about the true value of Z_{ij} then the observed proportion is used as a predictor. That is, for the Gang network we considered two missing network data scenarios, SS and SS-SR-S.

Finally, we also considered the situation where no network data are used (the standard model) and where the network is fully known.

Figure 1: Simulated ERGM network for a random sample of 300 individuals from a population of $N = 1,000$, with parameters EDGES$= -4.18$ and GWESP$= 1.0$. Note that although the isolated individual has no connections with the 299 other individuals that make up this sample, this does not mean that this individual does not have connections with the remaining 700 individuals in the population.

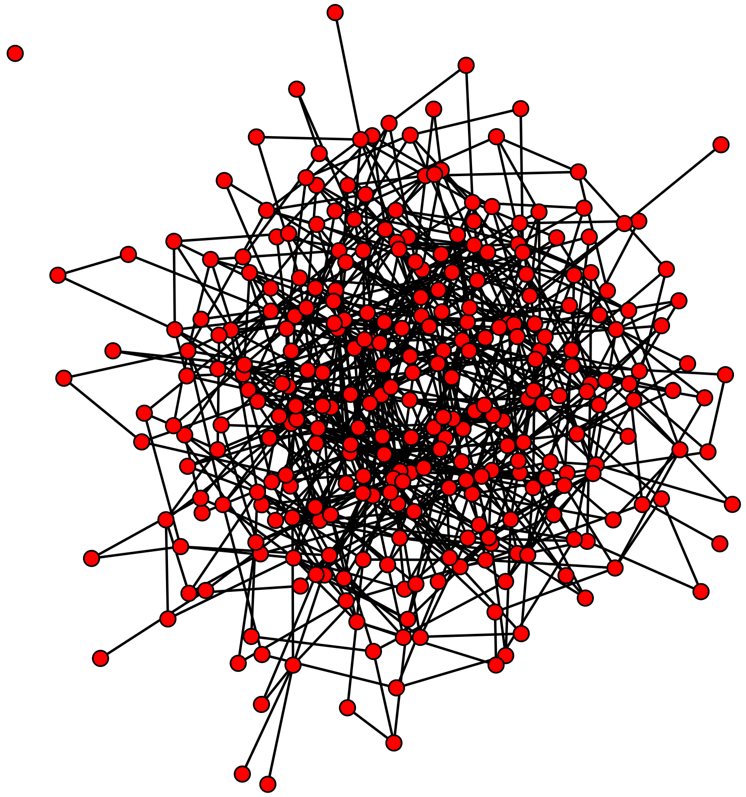
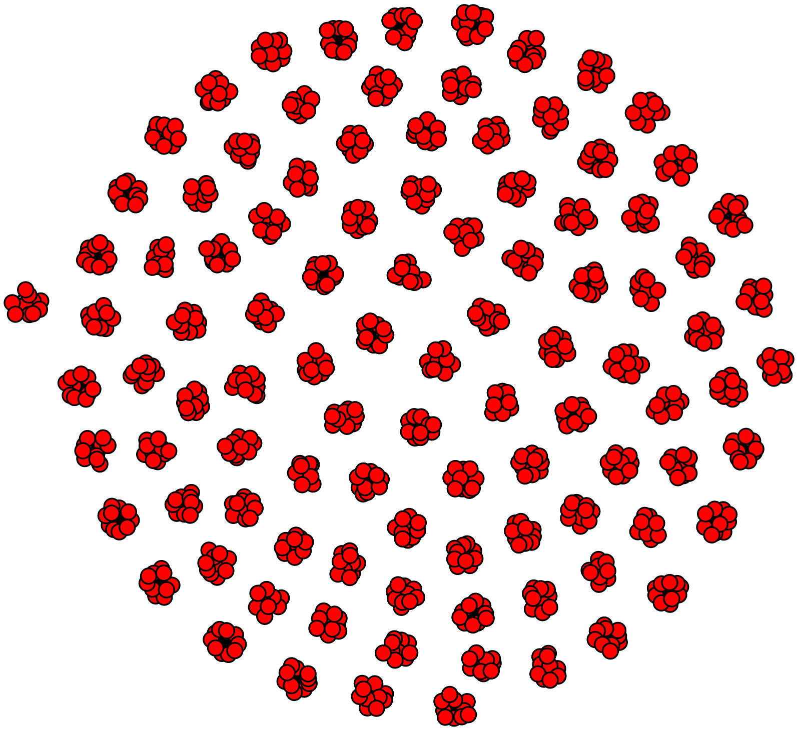


Figure 2: Gang network for $N = 1,000$ individuals with 100 gangs, each of size 10.



Parameter Specification for Linear Network Models

We generated data under all four of the linear network models discussed in Section 2. In each case all of these models were fitted to the resulting sample data, and EBLUP estimates of the population total t were then computed based on these fits (see the discussion in Section 3). In all cases the auxiliary variable took values randomly in the set $X_i = 1, \ldots, 9$, $\sigma^2 = 1$, $\beta_0 = 40$ and $\beta_1 = 5$. The network models used were

Contextual Network Model :

$$Y_i = \beta_0 + X_i\beta_1 + \bar{X}_i\bar{\beta} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

Here the contextual variable $\bar{X}_i$ is defined by the network $\mathbf{Z}$ and X_i, i.e. $\bar{\mathbf{X}} = \mathbf{W}\boldsymbol{\Xi}$, where $\boldsymbol{\Xi} = (X_1, \ldots, X_N)'$, and $\bar{\beta} = 2.0$.

Network Covariance Model :

$$\mathbf{Y} = \beta_0 + \mathbf{X}\beta_1 + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(0, \mathbf{V})$$

with $\mathbf{V} = \sigma^2\mathbf{I} + \theta(\mathbf{W} + \mathbf{W}')/2$; $\theta = 1.0$; $\sigma^2 = 1$, $\beta_0 = 40$ and $\beta_1 = 5$.

Autocorrelation Model :

$$\mathbf{Y} = \theta\mathbf{W}\mathbf{Y} + \beta_0 + \mathbf{X}\beta_1 + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_N)$$

with $\theta = 0.5$, $\sigma^2 = 1$, $\beta_0 = 40$ and $\beta_1 = 5$.

Network Disturbance Model :

$$\mathbf{Y} = \beta_0 + \mathbf{X}\beta_1 + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} = \theta\mathbf{W}\boldsymbol{\epsilon} + \mathbf{v},\ \mathbf{v} \sim N(0, \sigma^2\mathbf{I}_N)$$

with $\theta = 0.5$, $\sigma^2 = 1$, $\beta_0 = 40$ and $\beta_1 = 5$.

5.2 Simulation Results

Results for the $n = 100$ case are presented jointly for the ERGM and Gang networks. Table 1 shows the relative mean squared error (RMSE) of the estimates of t, while Table 2 shows the average lengths of nominal 95% Wald-type confidence intervals with their corresponding coverages shown as subscripts. Results for $n = 200$ are similar and are not presented here. No results are presented for the network covariance model, because these are similar to those of the network disturbance model. Also bias results are omitted, because these are effectively zero for all methods. These results include the cases where the network is ignored (the standard model) and when the network $\mathbf{Z}$ is fully known (the known network model). For partially observed network data we show results for the SS case (only $\mathbf{Z}_{ss}$ known), the SS-SR-C case ($\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ known, Method 2 imputation) and the SS-SR-S case ($\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ known, Method 3 imputation). Also, all results are shown relative to those for the BLUP, i.e. where the underlying network model and its parameter values are known. Clearly both this situation and the known network situation are unrealistic. However, they do allow one to gauge the relative benefit of putting more effort into collecting more network information and in carrying out more intensive network modelling.

Interestingly, when $\mathbf{Z}$ is known, but not the associated network and variance parameters, there is no loss in efficiency under the contextual network model. This is because knowing the variance parameter for the contextual model has no impact on the value of the EBLUP under this model. In contrast, we see an effect for the AR model, corresponding to a loss of efficiency of around $8-10\%$, mainly because the pseudo-design matrix $\mathbf{U}^{-1}(\theta)\mathbf{X}$ for this model depends on the estimated value of θ. This problem is much less of an issue for the network disturbance model because the design matrix under this model is $\mathbf{X}$.

It is clear from the results shown in Table 1 that ignoring the network (i.e. using the standard model) can lead to a large loss in efficiency if either the AR or the contextual network models are true. Interestingly, our results also seem to indicate that adopting the contextual network model when in fact the AR model is true seems as good as adopting the correct AR model specification. This is in contrast to the situation where the contextual network model is true and the EBLUP is based on either the AR model or the network disturbance model. Note that when the network disturbance model (or the network covariance model) is true, then ignoring the network information in the data only leads to a marginal loss in efficiency. Similarly, the efficiencies of the EBLUPs based on the different network models are also almost fully efficient in this case, irrespective of whether the assumed network model is true.

In order to see why the contextual network model yields similar results as the AR model when the AR model holds, we note that the mean of the AR model is $\boldsymbol{\mu} = \mathbf{U}(\theta)^{-1}\mathbf{X}$. If we approximate $\mathbf{U}(\theta)^{-1}$ by a first order Taylor series around zero, i.e. $\mathbf{U}(\theta)^{-1} = (\mathbf{I} - \theta\mathbf{W})^{-1} \approx \mathbf{I} + \theta\mathbf{W}$, then

$$\boldsymbol{\mu} \approx \mathbf{X}\boldsymbol{\beta} + \theta\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} + \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}$$

with $\bar{\boldsymbol{\beta}} = \theta\boldsymbol{\beta}$ and $\bar{\mathbf{X}} = \mathbf{W}\boldsymbol{\Xi}$, where $\boldsymbol{\Xi} = \mathbf{X}$. That is, the implied mean structure under the AR model is approximately the same as that under a contextual network model.

When $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ are observed, the EBLUP based on the contextual network model appears to perform well generally. This is because the EBLUP under this model does not depend on $\mathbf{Z}_{rr}$ and hence is unaffected by imputation of this part of the network. This is in contrast to the performance of this EBLUP when only $\mathbf{Z}_{ss}$ is observed. Here we see that the need to impute $\mathbf{Z}_{sr}$ leads to a significant loss of efficiency. Since estimation of θ in the pseudo-design matrix $\mathbf{U}(\theta)^{-1}\mathbf{X}$ under the AR model has a larger negative effect than the approximation of the AR model by the contextual network model, we conclude that the EBLUP based on the contextual network model seems a generally more robust method for estimating the population total than the EBLUP based on the AR model.

Turning now to the impact of the different network types used in the simulation, we see that for the ERGM network, our results indicate that the imputation method SS-SR-C, based on the conditional independence method (Method 2), performs worse than SS-SR-S, the simple proportion approach (Method 3). While this result is somewhat surprising, it could be explained by the roughness of the approximation implicit in the imputes generated using SS-SR-C and the robustness of the simple proportion method used in SS-RC-S. A priori, however, we would expect that a model-based approach using $\mathbb{E}(\mathbf{Z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs})$ as a predictor, say obtained by the MCMC technique, should lead to higher efficiency than using SS-SR-S. In particular, it is possible that a multiple imputation version of SS-SR-C might yield better results. However such methods still need to be developed for ERGMs, see for example Koskinen et al. (2011). In any case, it should be noted that the small observed differences between SS-RS-S and the complete network known case indicates that the possible gains from the use of these more sophisticated methods may be minimal.

Finally, for the Gang network the differences between SS and SS-SR-S for the contextual network and AR models are larger. A possible reason might be that for the Gang network some of the missing network structure can be deduced from the sample data, so that in fact part of the missing network, for example part of $\mathbf{Z}_{rr}$, is known in the SS-SR case. In contrast, in the SS case less can be inferred about the missing components of the network, explaining the larger differences compared with those obtained under the ERGM network structure. This suggests that obtaining as much additional information about the missing network components as possible is valuable. For example, if a social relationship can be ruled out for certain groups of the population, then these zero valued entries in $\mathbf{Z}$ represent valuable observed

Table 1: Estimated relative MSE of EBLUP, relative to BLUP, for ERGM network (left) and Gang network (right) and $n = 100$, $N = 1000$

		ERGM network			Gang Network		
		True Model			*True Model*		
Fitted model		Cont	AR	ARerr	Cont	AR	ARerr
BLUP - actual MSE		$9,390$	$8,739$	$8,736$	$9,421$	$12,330$	$12,315$
full network known		1.00	1.10	1.02	1.00	1.08	1.01
standard		2.30	3.50	1.00	2.87	10.5	1.04
contextual	SS	2.28	3.38	1.00	2.66	9.57	1.05
model	SS-SR-C	1.19	1.39	1.00	–	–	–
(Cont)	SS-SR-S	1.14	1.31	1.00	1.01	1.07	1.06
autocorrelation	SS	2.30	3.34	1.01	2.57	8.67	1.05
model	SS-SR-C	1.45	1.42	1.00	–	–	–
(AR)	SS-SR-S	1.31	1.30	1.00	1.24	1.10	1.06
disturbance	SS	2.30	3.52	1.02	2.54	8.79	1.01
model	SS-SR-C	2.30	3.48	1.02	–	–	–
(ARerr)	SS-SR-S	2.30	3.49	1.02	3.14	11.2	1.01

network information. In particular, it seems clear that there are significant advantages in collecting information about the complete network associated with the sampled individuals, i.e. both $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{rs}$.

6 Application to the British Household Panel Study

The British Household Panel Study (BHPS) is an annual multi-purpose household panel survey in the United Kingdom that focuses on gaining insight into the social and economic change at the individual and household level in Britain and the UK, see `http://www.iser.essex.ac.uk/survey/bhps/` for more details.

We focus on an individuals annual income in British pounds as the variable of interest. Our aim is to investigate how the use of network information available in BHPS impacts on average income estimates for the cross-classification age by gender by region, using six categories for age $15-18$ (1), $19-21$ (2), $22-30$ (3), $31-50$ (4), $51-64$ (5), 65+ (6) (in years), two for gender (1: male, 0: female) and five regions defined as: (1) 'Not London - North' consisting of East Midlands, West Midlands Conurbation, Rest of West Midlands, Greater Manchester, Merseyside, Rest of North West, South Yorkshire, West Yorkshire, Rest of Yorks & Humberside, Tyne & Wear, Rest

Table 2: Estimated relative length of nominal 95 per cent confidence interval (CI) for EBLUP, relative to BLUP, for ERGM network (left) and Gang network (right) and $n = 100$, $N = 1000$, actual coverage shown in subscript

		ERGM Network			Gang Network		
		True Model			*True Model*		
Fitted Model		Cont	AR	ARerr	Cont	AR	ARerr
BLUP - actual CI length		$370_{94.5}$	$382_{96.2}$	$382_{96.4}$	$370_{94.0}$	$418_{93.9}$	$418_{93.4}$
full network known		$1.00_{94.4}$	$0.99_{95.0}$	$0.98_{95.3}$	$1.00_{94.0}$	$1.00_{92.8}$	$1.27_{94.1}$
independence		$1.59_{94.8}$	$1.84_{96.0}$	$0.99_{95.7}$	$1.72_{93.8}$	$3.23_{92.8}$	$1.01_{93.6}$
contextual	SS	$1.54_{93.5}$	$1.79_{95.4}$	$0.99_{95.5}$	$1.65_{94.4}$	$3.07_{91.8}$	$1.00_{93.2}$
model	SS-SR-C	$1.00_{92.3}$	$1.01_{92.2}$	$0.99_{95.6}$	–	–	–
(Cont)	SS-SR-S	$1.00_{92.9}$	$1.01_{92.7}$	$0.99_{95.8}$	$1.00_{93.7}$	$1.01_{93.0}$	$1.01_{93.7}$
autocorrelation	SS	$1.54_{94.1}$	$1.77_{95.3}$	$0.99_{95.4}$	$1.62_{94.2}$	$2.89_{92.4}$	$1.00_{93.4}$
model	SS-SR-C	$1.10_{92.2}$	$1.01_{92.2}$	$0.99_{95.5}$	–	–	–
(AR)	SS-SR-S	$1.10_{93.1}$	$1.01_{92.8}$	$0.99_{95.6}$	$1.06_{93.7}$	$0.99_{92.4}$	$1.00_{93.4}$
disturbance	SS	$1.59_{94.8}$	$1.83_{95.7}$	$0.98_{95.1}$	$1.59_{93.5}$	$2.93_{92.3}$	$0.98_{93.2}$
model	SS-SR-C	$1.59_{94.8}$	$1.77_{94.4}$	$0.97_{94.4}$	–	–	–
(ARerr)	SS-SR-S	$1.60_{94.8}$	$1.80_{95.2}$	$0.98_{94.9}$	$2.01_{90.1}$	$53.6_{89.5}$	$1.27_{94.1}$

of North; (2) 'Not London - South' containing Rest of South East, South West and East Anglia; (3) 'London' includes inner and outer London; and finally (4) 'Scotland' and (5) 'Wales'. We exclude Northern Ireland from our analysis because BHPS sample sizes were too small to cross-classify by age and gender. We also exclude persons who did not report a positive income.

We start by noting that the estimates that we report are not meant to be an improvement on standard BHPS estimates. They have been calculated purely in order to illustrate how network information can be used in a realistic application, and the potential impact. In this context, we observe that BHPS collects information from a respondent on his/her three best friends, corresponding to the genders and ages of these friends, duration of friendships, frequency of contact, distances to the friends, their job/employment statuses, and their ethnicities. If we define $\mathbf{Z}$ in this case as the best friendship network for the BHPS target population then $\mathbf{Z}$ is unknown. We therefore use the contextual network model to incorporate the network information into estimation. We consider two models, a standard model based on age and gender and a contextual model that, in addition to age and gender, includes a variable equal to the arithmetic average of the gender indicators for a respondents three best friends. We use wave N of the BHPS (2004), the last available wave for which friendship information is collected. Sample size for this wave of the BHPS is $7,968$ people, of which 285 (3.6%) either report zero income or are from Northern Ireland and so are excluded, resulting in a final sample size of $n = 7,683$.

The estimated mid-2004 population for Great Britain (England, Scotland and Wales) is $58,124,700$. In the 2001 census the percentage of people aged less than 15 was 18.8% (Source: Office for National Statistics; National Assembly for Wales; General Register Office for Scotland). As a rough estimate we therefore have $N = 58,124,700 \times (1 - 0.188) \times (7713/7968) = 46,984,706$ people aged 15 years and older living in in Great Britain and with a positive income at the time of wave N of the BHPS.

Let $Age_i = 1, \ldots 6$ be the age category person $i = 1, \ldots, N$ falls in, $Reg_i = 1, \ldots, 5$ the region person i lives in and $Gender_i = 0, 1$ the gender of person i. A standard linear model for the population P that uses age by gender categories as auxiliary variables is

$$Y_{ijkl} = \alpha + \beta_{jk} + \epsilon_{ijkl},\ i = 1, \ldots, n;\ j = 1, \ldots, 6;\ k = 0, 1;\ l = 1, \ldots, 5 \quad (15)$$

where Y_{ijkl} is income of person i with $Age = j$, $Gender = k$ and $Region = l$. Prediction using the BLUP based on this model requires that population totals for each age by gender cell be known. Wave N BHPS weights were used to obtain these totals, which were then adjusted so that they summed to the (estimated) population value of N derived above.

We next define a contextual network model by adding a contextual variable to the standard model (15). This is the variable $\overline{MALE}_i = 0, 1/3, 2/3, 1$, defined as the average value of the gender indicators for an individuals three best friends. This leads to the model

$$Y_{ijkl} = \alpha + \beta_{jk} + \gamma \overline{MALE}_i + \epsilon_{ijkl},\ i = 1, \ldots, n;\ j = 1, \ldots, 6;\ k = 0, 1. \quad (16)$$

BLUP weights under both (15) and (16) can be calculated using the formula (9), and mean annual income for any domain $A \subset P$ can then be estimated via the weighted domain mean

$$\frac{\sum_{i \in A \cap s} w_i Y_i}{\sum_{i \in A \cap s} w_i}. \quad (17)$$

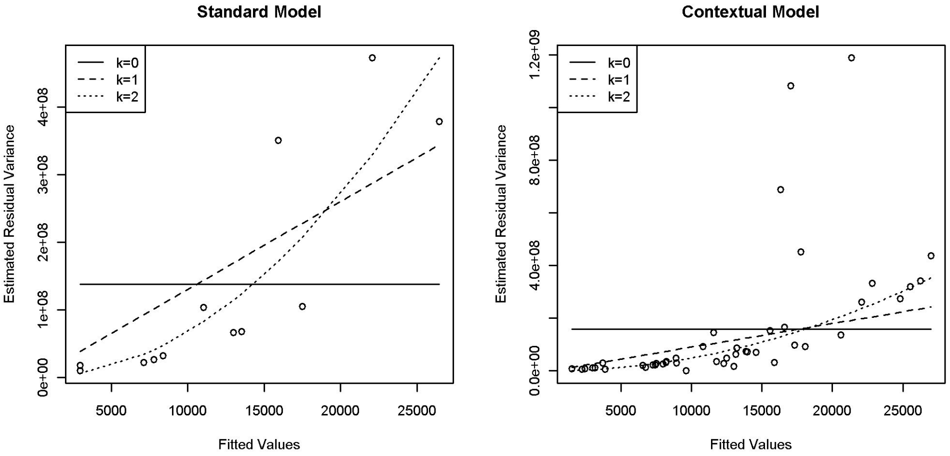
There are two issues with calculating the BLUP weights under the models (15) and (16). The first is that under the contextual model (16) these weights require that we know the population total of the contextual variable $\overline{MALE}_i$. In the results shown below, we substituted an estimate of this population total based on the wave N BHPS weights. Specifically, we estimated the non-sample mean of this variable by its corresponding (BHPS-) weighted sample mean.

Table 3: Average proportion of male friends $\overline{MALE}$

		Regions in Great Britain					
Sex	*Age Group*	London	England North	England South	Scotland	Wales	Average Proportion
	≤ 18	27.46	30.14	31.69	17.26	28.66	27.04
female	$19-21$	20.33	25.63	26.49	31.86	35.31	27.92
	$22-30$	26.23	19.80	18.27	16.08	14.22	18.92
	$31-50$	16.04	13.36	12.51	12.49	11.19	13.12
	$51-64$	11.29	10.44	11.70	8.61	10.48	10.50
	≥ 65	15.32	13.72	15.90	10.14	13.19	13.66
	≤ 18	80.87	72.54	71.98	73.52	78.80	73.7
male	$19-21$	67.95	72.23	70.48	59.23	78.80*	70.5
	$22-30$	73.06	74.70	70.48	69.17	74.06	73.2
	$31-50$	69.74	72.94	71.22	75.83	71.77	73.0
	$51-64$	59.05	65.17	63.30	68.54	67.48	64.3
	≥ 65	61.29	53.45	55.88	51.64	53.89	55.2

* sample size is one and estimated proportion from age group ≤ 18 is used instead

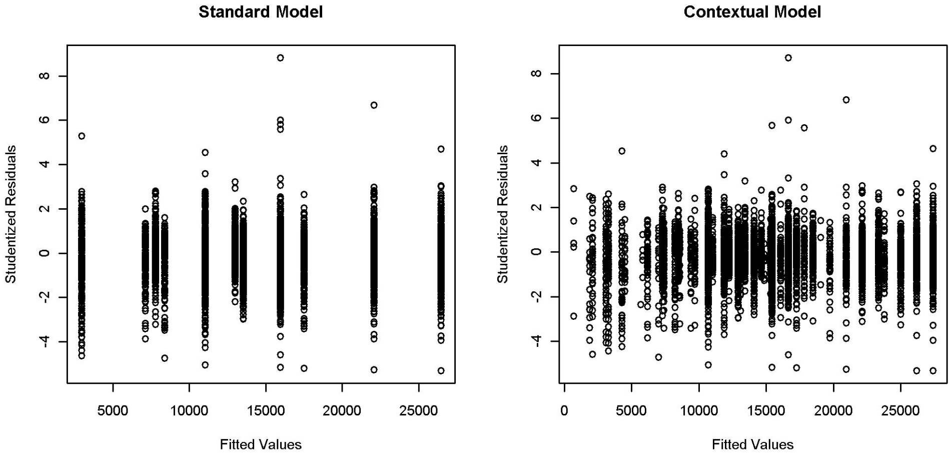
Figure 3: Estimated variance versus fitted values for standard model and contextual model



The second issue is that the covariance structure of the errors in both (15) and (16) requires specification. Figure 3 shows estimated residual variance (for each level of predicted value) versus predicted value under standard model (15) and the contextual model (16) based on an assumption of homoskedastic errors for both models. Curves corresponding to the heteroskedasticity model $\mathrm{Var}(Y_{ijkl}) = \phi\mathbb{E}(Y_{ijkl})^k$ for $k = 0, 1, 2$ were fitted to these values in order to assess whether the homoscedasticity assumption is reasonable. The four points that appear to be outliers in the contextual model plot were omitted from this fit. The fitted curves for $k = 0, 1, 2$ clearly indicate that $k = 2$ is the best choice and that the assumption of homoscedasticity ($k = 0$) is inappropriate for the BHPS income variable under both (15) and (16). This conclusion is supported by Figure 4 which shows plots of studentized residuals versus predicted values for $k = 2$.

We accommodate this heteroskedasticity when fitting (15) and (16) via ML by assuming that the model errors follow the Tweedie distribution (Jorgensen, 1997). This allows a general power link of the form $\mathbb{E}Y_{ijkl} = (\mathbf{X}_{ijkl}\boldsymbol{\beta})^p$, where $\mathbf{X}_{ijkl}$ is a vector of predictors and $\boldsymbol{\beta}$ is the vector of regression coefficients, and a general power-variance function of the form $\mathrm{Var}(Y_{ijkl}) = \phi\mathbb{E}(Y_{ijkl})^k$. The variance power coefficient k can take any real value outside the open interval $0 < k < 1$. Special cases are the Normal distribution ($k = 0$), the Poisson distribution ($k = 1$) and the Gamma dis-

Figure 4: Residuals versus fitted values for standard model and contextual model



tribution ($k = 2$). The standard assumption of normally distributed errors seems unrealistic for a strictly non-negative response variable like income. Consequently we fitted both (15) and (16) under the assumption of a linear link ($p = 1$) and Gamma errors ($k = 2$).

Table 4 shows the fitting results for the two models. The log-likelihood under (15) is $L = -80,785.75$ while that under (16) is $L = -80,750.07$, indicating that the contextual model represents a significant improvement over the standard model (the likelihood ratio statistic has a p-value of 0.00).

Average income estimates for age by gender by region under the standard model are displayed in Table 5. Corresponding estimates under the contextual model (expressed relative to those obtained under the standard model) can be found in Table 6. This also shows where the estimates generated under the two models have absolute differences larger than 1,2,3 and 4 standard errors. We see that 49 out of the 60 differences are larger than two standard errors.

The results of Table 6 show that ignoring network information, i.e. contextual effects, can lead to significantly different survey estimates in practice. Based on the results of the simulation study reported in Section 5, there is an indication that the estimates derived under the contextual network model (i.e. the model that allows for network effects) may also be more accurate.

Table 4: Fitting results for the standard model and contextual model for k=2

	Standard Model			*Contextual Model*		
Variable	estimate	Std. Error	p-value	estimate	Std. Error	p-value
Intercept	2,944	218	$< 2e\text{-}16$	2,439	237	$< 2e\text{-}16$
Age 19-21	4,181	525	$< 1.99e\text{-}15$	4,107	518	2.38e-15
Age 22-30	10589.8	536	$< 2e\text{-}16$	10,725	532	$< 2e\text{-}16$
Age 31-50	13,029	418	$< 2e\text{-}16$	13,182	415	$< 2e\text{-}16$
Age 51-64	8,098	3956	$< 2e\text{-}16$	8,391	396	$< 2e\text{-}16$
Age 65+	4,833	324	$< 2e\text{-}16$	5,019	323	$< 2e\text{-}16$
Male	-15.4	307	0.960	-925	338	0.0063
Male - Age 19-21	1,303	831	0.117	1,219	815	0.1349
Male - Age 22-30	3,939	883	8.25e-06	3,619	875	3.56e-05
Male - Age 31-50	10,526	786	$< 2e\text{-}16$	10,157	781	$< 2e\text{-}16$
Male - Age 51-64	11,043	845	$< 2e\text{-}16$	10,741	837	$< 2e\text{-}16$
Male - Age 65+	5,233	604	$< 2e\text{-}16$	5,260	595	$< 2e\text{-}16$
Contextual Male	–	–	–	2,146	409	1.55e-07
Log-likelihood	-80,785.75			-80,750.07		

Table 5: Mean annual income in British pounds for age by sex by region with weighting based on a standard linear model with covariates age by sex

		Regions in Great Britain				
Sex	*Age Group*	London	England North	England South	Scotland	Wales
female	≤ 18	2,668	3,666	2,279	1,678	4,120
	$19 - 21$	6,989	7,134	6,818	7,036	8,976
	$22 - 30$	17,068	12,759	13,403	14,147	12,774
	$31 - 50$	20,266	14,881	16,514	16,043	14,565
	$51 - 64$	12,129	10,725	10,931	11,822	11,279
	≥ 65	8,582	7,283	7,952	7,851	8,793
male	≤ 18	1,257	3,896	2,180	2,578	4,897
	$19 - 21$	10,102	7,600	9,735	6,735	16,270
	$22 - 30$	15,617	15,617	18,294	20,719	15,960
	$31 - 50$	23,884	23,884	29,216	29,216	21,947
	$51 - 64$	20,186	20,186	23,293	27,237	19,305
	≥ 65	11,775	11,775	14,540	12,055	12,613

7 Discussion

As stated at the end of Section 1, our aim in this paper was to address the questions: (i) Is embedding network information useful for survey estimation? (ii) If the answer to (i) is yes, then which models are potentially useful? and (iii) How much network data needs to be collected in order to obtain potentially higher precision for survey estimation? Given the simulation results that we present in Section 5, our answer to (i) is yes, and to (ii) is the contextual network and AR models when either model is true. When the network disturbance model or the network covariance model is true, our results suggest that ignoring the network does not result in a significant loss of efficiency. Ignoring the network under the AR and contextual network models leads to a mis-specification of the mean model, but this does not apply for the network covariance and network disturbance models. Furthermore, our answer to (iii) is that in realistic applications it will usually be impossible to collect the full network, and our simulation results show that when either the contextual network model or the AR model is true then both $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ must be collected in order to obtain efficiency gains. Knowledge of $\mathbf{Z}_{ss}$ alone is not enough.

Table 6: Change in mean annual income for the contextual model relative to standard model, positive/negative values show an increase/decrease in mean annual income

		Regions in Great Britain				
Sex	*Age Group*	London	England North	England South	Scotland	Wales
female	≤ 18	162	−10[2]	−110[4]	458[4]	391[4]
	19 − 21	324[1]	47[4]	0	−127[4]	−372[4]
	22 − 30	22[2]	2[4]	−3[4]	−10[4]	−7[4]
	31 − 50	77[2]	−2[4]	−6[4]	−3[4]	−29[4]
	51 − 64	−8	−1[4]	5[4]	−6[4]	−2[4]
	≥ 65	−7	5[4]	−1	4[4]	4[4]
male	≤ 18	−176[3]	−20[4]	99[4]	106[4]	−771[4]
	19 − 21	139[1]	−5[4]	−15[4]	71[4]	−710[4]
	22 − 30	−23	41[4]	−47[4]	−89[4]	20[4]
	31 − 50	−134[1]	30[4]	−60[4]	114[4]	−14[4]
	51 − 64	−214[1]	40[4]	−50[4]	209[4]	56[4]
	≥ 65	119[1]	−5[4]	−2[3]	−74[4]	12[4]

[1],

[2], [3], [4] Difference larger than 1,2,3,4 standard errors

In practice, we suggest a careful model fitting exercise be carried out before attempting to use either the contextual network model or the AR model for survey estimation. Given the numerical difficulties with fitting the AR model, see Suesse (2012), we recommend that the contextual network model be used if it is a good fit to the data, otherwise caution is warranted and ignoring the network might be the best option.

Clearly, more extensive information on networks needs to be collected in conjunction with standard survey data to gain further insight into the usefulness of network models for survey estimation. In this paper we have focused on undirected networks, so knowing $\mathbf{Z}_{sr}$ is equivalent to knowing $\mathbf{Z}_{rs}$. For directed networks, this equivalence does not apply and conclusions, particularly for the case when $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ are known, are likely to be different. The issue of imputation methods for the missing network information has been addressed in this paper, but many questions remain. Is an appropriate single value imputation (let alone multiple imputation) method using $\mathbb{E}(\mathbf{Z}^{mis}|\mathbf{Z}^{obs} = \mathbf{z}^{obs})$ (Method 2) better than the simple proportion approach (Method 3)? The numerical intensity of the MCMC methods used to fit network models like the ERGM when population sizes are large meant that we could not fully explore this issue. There is current research that tries to address some of these issues, see (Koskinen et al., 2011), but more is required. However, given that we observed only small differences in efficiency between the SS-RS-S case and the full network known case, we anticipate that more sophisticated imputation methods will not lead to substantial efficiency gains.

Finally, we note that all network models considered in this paper assume that the value of the response variable Y for an individual in the study population depends on a linear combination of the values of this variable for the other individuals in the population that are linked to this person in the network. If there is an implicit ordering in the strength of these links, then this can be allowed for in the network model for Y. For example, in the case of a best friend network, where the friendships are ordered by their strength, one can modify the contextual network model so that there is a separate parameter for each level of best friend. To illustrate, in the BHPS application reported in the previous Section, the effect of the first male friend is 948.24, the effect of the second male friend 648.23 and that of the third male friend 513.66. The corresponding coefficient for the contextual network model with a common effect is $2,146/3 = 715.33$. A Wald test for equality of these effects (based on the model (16) with heteroskedasticity parameter $k = 2$) returns a p-value of 0.43, supporting the assumption of a common effect.

References

Butts, C. (2008). network: A package for managing relational data in r. *Journal of Statistical Software 24*(2), 1–36.

Carrington, P., J. Scott, and S. Wasserman (2005). *Models and methods in social network analysis.* New York: Cambridge University Press.

Chambers, R. L. and R. G. Clark (2012). *An Introduction to Model-Based Survey Sampling with Applications.* Oxford: Oxford University Press.

Doreian, P., K. Teuter, and C. H. Wang (1984). Network auto-correlation models - some monte-carlo results. *Sociological Methods & Research 13*(2), 155–200.

Duke, J. B. (1993). Estimation of the network effects model in a large data set. *Sociological Methods & Research 21*(4), 465–481.

Friedkin, N. E. (1990). Social networks in structural equation models. *Social Psychology Quarterly 53*(4), 316–328.

Goldstein, H. (1986). Multilevel mixed linear-model analysis using iterative generalized least-squares. *Biometrika 73*(1), 43–56.

Goldstein, H. (1989). Restricted unbiased iterative generalized least-squares estimation. *Biometrika 76*(3), 622–623.

Hunter, D. R. (2007). Curved exponential family models for social networks. *Social Networks 29*(2), 216–230.

Hunter, D. R., S. M. Goodreau, and M. S. Handcock (2008). Goodness of fit of social network models. *Journal of the American Statistical Association 103*(481), 248–258.

Hunter, D. R. and M. S. Handcock (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics 15*(3), 565–583.

Jorgensen, B. (1997). *Theory of Dispersion Models.* London: Chapman and Hall.

Koskinen, J., G. Robins, and P. Pattison (2011, November). Missing data in social networks: Problems and prospects for model-based inference. Working Paper No. 09-01, MelNet Social Networks Laboratory.

Leenders, R. (2002). Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks 24*(1), 21–47.

Marsden, P. V. and N. E. Friedkin (1993). Network studies of social-influence. *Sociological Methods & Research 22*(1), 127–151.

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association 70*(349), 120–126.

Royall, R. M. (1976). Linear least-squares prediction approach to 2-stage sampling. *Journal of the American Statistical Association 71*(355), 657–664.

Snijders, T. (2002). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 1–40.

Snijders, T., P. Pattison, G. Robins, and M. Handcock (2006). New specifications for exponential random graph models. *Sociological Methodology 36*, 99–153.

Srndal, C., B. Swensson, and J. Wretman (1992). *Model assisted survey sampling*. Springer series in statistics. New York: Springer-Verlag.

Strauss, D. and M. Ikeda (1990). Pseudolikelihood eestimation for social networks. *Journal of the American Statistical Association 85*(409), 204–212.

Suesse, T. (2012). Estimation in autoregressive population models. In *Proceedings of Fifth Annual ASEARC Research Conference*, University of Wollongong. ASEARC. 2-3 February 2012.

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

A Bias-Correction for Variance Estimator

For any linear model with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and variance $\mathbf{V}$, the ML estimator for $\boldsymbol{\beta}$ as a function of the variance $\mathbf{V}$ is $\hat{\boldsymbol{\beta}}_{ML} = \mathbf{M}\mathbf{Y}$ with $\mathbf{M} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}^{-1}$. One can then show that

$$\mathbb{E}\mathbf{r}_d\mathbf{r}_d' = \mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}', \tag{18}$$

where $\mathbf{r}_d := \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{X}\mathbf{M})\mathbf{Y}$, see Goldstein (1989). The iterative generalised least squares method (IGLS) for fitting a multi-level model is equivalent to ML under normality (Goldstein, 1986). This is true for any variance function $\mathbf{V}(\boldsymbol{\alpha})$ depending on $\boldsymbol{\alpha}$ and not just for a multi-level model where the variance model is in fact also a linear model. However, ML estimates for the variance components θ are biased, because $\mathbb{E}\mathbf{r}_d\mathbf{r}_d' \neq \mathbf{V}$, see (18). However, adjusting the pseudo-observations $\mathbf{S} = \mathbf{r}_d\mathbf{r}_d'$ to $\mathbf{S}^* = \mathbf{S} + \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'$ and applying least squares to $\mathbf{S}^*$ instead of $\mathbf{S}$ to obtain estimates for the parameter $\boldsymbol{\alpha}$ of the variance model is equivalent to applying REML. In this context, the added term can be seen as a bias correction (Goldstein, 1989).

For the population disturbance and AR models $\mathbf{V}(\theta, \sigma^2) = \sigma^2\mathbf{U}(\theta)^{-1}$ $(\mathbf{U}(\theta)^{-1})'$. Instead of applying this bias correction within each iteration of the fitting algorithm, we use the ML estimates $\hat{\theta}_{ML}$ and $\hat{\sigma}^2_{ML}$ to define

$$\hat{\mathbf{V}} = \mathbf{V}(\hat{\theta}_{ML}, \hat{\sigma}^2_{ML}) + \mathbf{X}(\mathbf{X}'\mathbf{V}(\hat{\theta}_{ML}, \hat{\sigma}^2_{ML})^{-1}\mathbf{X})^{-1}\mathbf{X}'$$

as an estimate for $\mathbf{V}$ that retrospectively accounts for the bias of the ML estimates of the variance components and consequently that of $\mathbf{V}(\hat{\theta}_{ML}, \hat{\sigma}^2_{ML})$.

For the AR model the residuals are defined as $\mathbf{r}_{AR} := \mathbf{Y} - \mathbf{U}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}$. We can show that

$$\mathbb{E}\mathbf{r}_{AR}\mathbf{r}_{AR}' = \mathbf{V} - \sigma^2\mathbf{U}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{U}^{-1})'. \tag{19}$$

This formula can also be obtained from (18) by replacing $\mathbf{X}$ by $\mathbf{U}^{-1}\mathbf{X}$ and using $\mathbf{V}(\theta, \sigma^2) = \sigma^2\mathbf{U}(\theta)^{-1}(\mathbf{U}(\theta)^{-1})'$. Assuming ρ is known, and again adjusting $\mathbf{S}$ to $\mathbf{S}^* = \mathbf{S} + \sigma^2\mathbf{U}^{-1}\mathbf{X}\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{U}^{-1})'$ makes IGLS equivalent to REML. However this only holds when θ is known, since otherwise $\partial\mathbf{S}(\rho)/\partial\rho \neq \mathbf{0}$, and IGLS and REML are equivalent only if $\partial\mathbf{S}(\theta)/\partial\theta = \mathbf{0}$, i.e. the residuals forming $\mathbf{S}$ do not depend on unknown variance parameters. For the AR model the mean and the variance depend on θ, so this equivalence does not hold for unknown θ. However the assumption of known θ still allows us to apply a partial bias correction, in that it does not account for the variation in $\hat{\theta}_{ML}$. This bias corrected estimate for $\mathbf{V}$ is

$$\hat{\mathbf{V}} = \mathbf{V}(\hat{\theta}_{ML}, \hat{\sigma}^2_{ML}) + \hat{\sigma}^2_{ML}\mathbf{U}(\hat{\theta}_{ML})^{-1}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{U}(\hat{\theta}_{ML})^{-1})'.$$

Note that simulation results for the AR model reported in Suesse (2012) indicate that $\hat{\theta}_{ML}$ is unbiased.