# Social Networks with Mismeasured Links

Arthur Lewbel
Boston College

Xi Qu
Shanghai Jiao Tong University

Xun Tang
Rice University

April 2021

Abstract

We consider estimation of peer effects in social network models where some network links are incorrectly measured. We show that if the number of mismeasured links does not grow too quickly with the sample size, then standard instrumental variables estimators that ignore the measurement error remain consistent, and standard asymptotic inference methods remain valid. These results hold even when measurement errors in the links are correlated with regressors, or with the model errors. Monte Carlo simulations and real data experiments confirm our results in finite samples. These findings imply that researchers can ignore small amounts of measurement errors in networks.

JEL classification: C31, C51

Keywords: Social networks, Peer effects, Misclassified links, Missing links, Mismeasured network.

## 1 Introduction

In many social and economic environments, an individual's behavior or outcome (such as a consumption choice or a test score) depends not only on his or her own characteristics, but also on the behavior and characteristics of other individuals. Call such dependence between two individuals a link, and call individuals with such links friends. A social network consists of a group of linked individuals. Each individual may have a different set of friends in the network, and each individual may assign heterogenous weights to his or her links. The structure of a social network is fully characterized by a square adjacency matrix, which lists all links (with possibly heterogenous weights) among the individuals in the network.

Much of the econometric literature on social networks focuses on disentangling and estimating various social or network effects, based on observed outcomes and characteristics of network members. These structural parameters include the effects on each individual's outcome by (i) the individual's own characteristics (direct effects) and possibly group characteristics (correlated effects), (ii) the characteristics of the individual's friends (contextual effects) and (iii) the outcomes of the individual's friends (peer effects). Standard methods of identifying and estimating these structural network effect parameters assume that the adjacency matrix of links among individuals in the sample is perfectly observed.

**1.1. Our contribution.** We consider the case where network links are misclassified, or generally measured with errors. Here we provide good news for empirical researchers, by showing that relatively small amounts of measurement error in the network can be safely ignored in estimation. More precisely, we show that instrumental variable estimators like Bramoullé, Djebbari and Fortin (2009), and their standard errors, remain consistent and valid, despite the presence of misclassified, unreported, or mismeasured links, as long as the number and size of these measurement errors grows sufficiently slowly with the sample size. Moreover, these results hold even when the measurement errors are correlated with the regressors, or with the model errors. Below in subsection 1.3 we give examples of applications where measurement errors grow at these slow rates.

one in which $n$ grows to infinity. Our main results focus on situations where the sum of network measurement errors (the differences between $H_n$ and $G_n$) grows at a rate less than $\sqrt{n}$. Here we list a range of empirical situations in which network measurement errors would be expected to grow at these slow rates.

Consider first the common modeling environment in which data are collected from many groups of individuals, like villages or schools. Data are often collected on links within these groups, such as friendships within class rooms, or kinship relationships within villages. Models using such data often assume no links between individuals in different groups, either for theoretical convenience, or because data are not collected on links between groups. This is equivalent to misclassifying as zero all links that exist outside of diagonal blocks of $G_n$. In other words, this means using a block-diagonal $H_n$ in place of the actual $G_n$ in the data-generating process. The measurement errors will grow at a rate slower than $\sqrt{n}$ if the number of sampled groups grows at rate slower than $\sqrt{n}$, and the number of links between groups is relatively small.

Another example comes from panel data. Suppose the sample consists of $L$ individuals, each of which is observed for $T$ time periods, so the sample size is $n = LT$. For example, the data couldthethe hindiv

62.viduals,1734-16Rgrows to423sa355(oTf10(c

[illegible]

=

and contextual effects respectively.[2] Let $G_{ij}$ ($C_{ij}$) denote the element in row $i$ and column $j$ of $G_n$ ($C_n$). We have $G_{ij} > 0$ if $i$ and $j$ are linked for peer effects and $G_{ij} = 0$ otherwise. Similarly, $C_{ij} > 0$ if $i$ and $j$ are linked for contextual effects and $C_{ij} = 0$ otherwise. For each individual $i$, let $G_{ii} = 0$ and $C_{ii} = 0$ by convention in the literature. Note that $G_{ij}$ can be binary (with $G_{ij} \in \{0; 1\}$ indicating the absence or presence of a link), or continuous and non-negative with $G_{ij} \in {}_{+}$ signifying the strength of the link. The same applies for $C_n$. Throughout the paper, we maintain that $\min_i \sum_{j=1}^{n} G_{ij} > 0$ and $\min_i \sum_{j=1}^{n} C_{ij} > 0$ with probability one. This means there are no isolated individuals in the network, or equivalently no rows of zeros in $G_n$ or $C_n$, almost surely. This condition is standard in the literature.

We assume a linear social network model:

$$Y_n = {}_0 + {}_0 G_n Y_n + X_n {}_0 + C_n X_n {}_0 + {}_n, \tag{1}$$

where $G_n$ and $C_n$ can be either the original adjacency matrices $G_n$ and $C_n$, or normalized versions of $G_n$ and $C_n$. For example, a row-normalized $G_n$ is defined by $G_{ij} = G_{ij} / \sum_{j'=1}^{n} G_{ij'}$. Row normalization is common; we will show our results hold with or without such normalization.

The parameters in equation (1) are as follows: ${}_0 \in$ is a scalar peer effect, ${}_0 \in {}^K$ is a vector of direct effects, ${}_0 \in {}^K$ is a vector of contextual effects, and ${}_0 \in$ is the structural intercept. If individuals are divided into groups (such as villages or classrooms), then what are known as correlated effects are group-level fixed effects, i.e., elements of ${}_0$ where the corresponding element of $X_n$ is a group membership indicator.

Our goal is estimation of ${}_0 \equiv ( {}_0; {}_0; {}^0_0;$ n to be completely different, by assuming that two different adjacency matrices are observed, one a mismeasured version of $G_n$ and the other a mismeasure of $C_n$. We do not do so to save on notation, and because it is extremely rare in practice to observe two different adjacency matrices, where one is known to measure peer effects and the other is known to measure contextual effects.

Like $G_n$ and $C_n$, the matrix $H_n$ by convention has zeros on the diagonal. When $G_{n;ij}$ equals zero or one, misclassification of that link corresponds to $H_{ij} = 1 - G_{ij}$, and similarly for $C_{ij}$. More pointed out by Manski (1993). This identification problem can be overcome in models with more complicated social interaction structures. Lee (2007) uses conditional maximum likelihood and instrumental variable methods to estimate peer and contextual effects in a spatial autoregressive social interaction model, assuming links are perfectly observed in the data. Bramoullé, Djebbari and Fortin (2009) and Lin (2010) provide specific conditions on observed network structure in order to identify peer effects in social interaction models, using characteristics of friends of friends as instruments.

Given results like these, the model described in the introduction has been widely used to estimate peer effects in a variety of settings (usually assuming either $C_n = G_n$ or $C_n = 0$, though see Blume et. al. 2015). Examples are studies of peer influence on students' academic performance, sport and club activities, and delinquent behaviors (Hauser et al., 2009; Calvó-Armengol et al., 2009; Lin, 2010; Lee et al., 2010; Liu et al., 2014; Boucher et al., 2014; Patacchini and Zenou, 2012). These models all assume that the network structure is correctly measured in the data.

Regarding selection and comparison of adjacency matrices, LeSage and Pace (2009) use the Bayesian posterior distribution to choose among models with different adjacency matrices. Empirical research may also report estimates using different link weights as robustness checks. These practices are feasible in, e.g., spatial econometric models, where link weights are assumed to be a function of observable geographic information, as in gravity models of trade. Errors in constructing such links would fit in our framework. There is also a small literature on identification and estimation of peer effects when networks are unobserved. Examples include de Paula et al. (2018) and Lewbel et al. (2021).

The issue of potentially misclassified links is acknowledged and discussed in Patacchini and Venanzoni (2014), Liu et al. (2014), and Lin (2015) among others. But these papers do not provide a formal analysis of the asymptotic impact of mismeasured links on the performance of standard estimators. Griffith (2021) studied the impact on inference when misclassification in the adjacency matrix occurs because of binding caps on the number of self-reported links. Our results fill a void in the literature by analyzing how ignoring small amounts of general measurement errors in the adjacency matrix affects the consistency of standard estimators and the validity of inference.[4]

[4] Referring to potential omission of friends, Patacchini and Venanzoni (2014) say that, "in the large majority of cases (more than 94%), students tend to nominate best friends who are students in the same school and thus are systematically included in the network (and in the neighborhood patterns of social interactions)". Liu et al. (2014) report that "less than 1% of the students in our sample show a list of ten best friends, less than 3% a list of five males and roughly 4% a list of five females. On average, they declare that they have 4.35 friends with a small dispersion around this mean value (standard deviation equal to 1.41), and in the large majority of cases (more than 90%) the nominated best friends are in the same school." Lin (2015) says, "this nomination constraint only affects a small portion of our sample, as less than 10% of the sample have listed five male or female friends. Therefore, this restriction should not have a significant impact on the results." This last speculation is precisely what our first set of results establishes: that consistency of estimates will not be effected if the number of omitted (and hence misclassified) links is sufficiently small.

## 3 2SLS Estimation With Mismeasured Links

In this section we derive asymptotic properties of the 2SLS estimator for the model in (1) when the mismeasured adjacency matrix

Part (i) of Assumption 2 states that $X_n$

As noted in the introduction, even slowly growing measurement errors could asymptotically corrupt b if the stochastic order of quadratic terms in $b - {}_{0}$ isn't bounded. The closed form of the

3. The average standard errors do a good job of estimating the standard deviations for all values of s. This is as expected, because the problem with inference for larger values of s is that the bias in the estimator shrinks at rate $n^{s \ 1}$. Similarly, with $s \geq 0.5$, the parameter estimates deteriorate primarily due to bias rather than variance.

4. With both the true and mismeasured adjacency matrices, the mean-squared errors are much smaller for the direct effects than for the peer and contextual effects and , and the mean squared errors are much lower for the discrete regressor effects $_1$ and $_1$ than for the continuous regressor effects $_2$ and $_2$

Table 1. 2SLS Estimators with Misclassi..ed Links

| | $n$ = 200 | | | | $n$ = 500 | | | | $n$ = 1000 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | m.s.e. | bias | std | a.s.e. | m.s.e. | bias | std | a.s.e. | m.s.e. | bias | std | a.s.e. |
| True | | Mis.# | 0 | | | | 0 | | | | 0 | |
| $\alpha$ | 3.880 | -0.114 | 1.971 | 2.197 | 1.519 | 0.031 | 1.235 | 1.310 | 0.762 | 0.065 | 0.873 | 0.887 |
| $\lambda$ | 0.336 | 0.025 | 0.581 | 0.654 | 0.131 | -0.010 | 0.362 | 0.386 | 0.068 | -0.019 | 0.260 | 0.264 |
| $\beta_1$ | 0.003 | 0.005 | 0.058 | 0.058 | 0.001 | -0.003 | 0.036 | 0.036 | 0.001 | -0.000 | 0.027 | 0.026 |
| $\beta_2$ | 0.005 | 0.008 | 0.072 | 0.073 | 0.002 | 0.001 | 0.048 | 0.045 | 0.001 | -0.000 | 0.032 | 0.032 |
| $\gamma_1$ | 0.802 | -0.029 | 0.898 | 1.006 | 0.301 | 0.019 | 0.549 | 0.597 | 0.165 | 0.030 | 0.406 | 0.410 |
| $\gamma_2$ | 1.571 | -0.040 | 1.256 | 1.348 | 0.561 | 0.020 | 0.750 | 0.796 | 0.278 | 0.032 | 0.528 | 0.545 |
| $s$ = 0.1 | | Mis.# | 66 | | | | 81 | | | | 88 | |
| $\alpha$ | 4.100 | -0.058 | 2.029 | 2.254 | 1.576 | 0.033 | 1.258 | 1.325 | 0.780 | 0.070 | 0.883 | 0.894 |
| $\lambda$ | 0.365 | 0.008 | 0.605 | 0.672 | 0.135 | 0.010 | 0.368 | 0.391 | 0.070 | -0.020 | 0.263 | 0.266 |
| $\beta_1$ | 0.003 | 0.004 | 0.058 | 0.058 | 0.001 | -0.003 | 0.036 | 0.036 | 0.001 | -0.000 | 0.027 | 0.026 |
| $\beta_2$ | | | | | | | | | | | | |

## 5 Application

Lin and Lee (2010) model teenage pregnancy rates, using the model

$$Teen_i = \quad \sum_{j=1}^{n} G_{ij} Teen_j + \quad + Edu_i \; {}_1 + Inco_i \; {}_2 + FHH_i \; {}_3 + Black_i \; {}_4 + Phy_i \; {}_5 + "_i;$$

where $Teen_i$ is the teenage pregnancy rate in county i, which is the percentage of pregnancies occurring to females 12-17 years old, and $G_{ij}$ is the row-normalized entry of the original link matrix $G_n$, where $G_{ij} = 1$ if counties i and j are neighboring counties. $Edu_i$ is the education service expenditure (in units of $100), $Inco_i$ is median household income (divided by 1000), $FHH_i$ is the percentage of female-headed households, $Black_i$ is the proportion of black population and $Phy_i$ is the number of physicians per 1000 population, all in county i.[5]

The sample size is $n = 761$. Among all the $761 \times 760 = 578;360$ entries (diagonal are zero) in the original network $G_n$, there are 4;606 non-zero links. We treat the adjacency matrix they report as the true network, artificially introduce misclassified links, and then evaluate how this affect the 2SLS estimates. We generate misclassified links using $H_{ij} = G_{ij} \cdot e_{1i} + (1 - G_{ij}) \cdot e_{2i}$, where $e_{1i}$ and $e_{2i}$ are binary variables with probabilities $\;_{1i} = \;_i n^{s \; 1}$ and $\;_{2i} = 100 \;_i n^{s \; 2}$ of equaling 1. We set $\;_i = (y_i = \bar{y})^2$; so for each individual i misclassification is more likely to happen the larger is the magnitude of the observed outcome $y_i$.

We report 2SLS estimates using $H_n X_n$ and $H_n^2 X_n$ as instruments. Unlike our structural model, Lin and Lee (2010) assume contextual effects (the coefficients) are zero, so $G_n X_n$ does not appear as regressors. It would therefore have been possible to just use $H_n X_n$ as instruments for estimation. Nonetheless, to illustrate our proposition, we use both $H_n X_n$ and $H_n^2 X_n$ as instruments here.

Table 2 reports results based on 1000 Monte Carlo replications for each value of s. Results are reported where the model is estimated both with and without row normalization.

Consistent with our propositions, when the misclassification rate is low ($s < 0:5$), the 2SLS

Table 2. Estimation Results with Different Misclassification Rates

| | | | 100 $_1$ | $_2$ | $_3$ | $_4$ | $_5$ | Mis. # |
|---|---|---|---|---|---|---|---|---|
| Row-normalized adjacency matrices $G_{ij} = G_{ij}/\sum_j G_{ij}$ and $H_{ij} = H_{ij}/\sum_j H_{ij}$ | | | | | | | | |
| True | 0.4813 | 6.1911 | -0.9824 | -0.1871 | 0.7347 | 0.1267 | -0.4956 | 0 |
| | (0.079) | (1.469) | (0.651) | (0.040) | (0.063) | (0.057) | (0.188) | |
| s = 0:1 | 0.4897 | 6.1085 | -0.9910 | -0.1878 | 0.7355 | 0.1289 | -0.4980 | 111 |
| | (0.081) | (1.480) | (0.651) | (0.040) | (0.063) | (0.057) | (0.188) | |
| s = 0:3 | 0.5132 | 5.8759 | -1.0086 | -0.1895 | 0.7375 | 0.1341 | -0.5049 | 418 |
| | (0.085) | (1.512) | (0.652) | (0.040) | (0.063) | (0.057) | (0.188) | |
| s = 0:5 | 0.6017 | 4.9578 | -1.0542 | -0.1943 | 0.7422 | 0.1465 | -0.5227 | 1578 |
| | (0.099) | (1.626) | (0.654) | (0.040) | (0.063) | (0.057) | (0.189) | |
| s = 0:7 | 0.8138 | 2.7629 | -1.1726 | -0.2092 | 0.7589 | 0.1683 | -0.5535 | 5948 |
| | (0.139) | (1.985) | (0.660) | (0.040) | (0.064) | (0.057) | (0.191) | |
| Original adjacency matrices $G_{ij}$ and $H_{ij}$ without normalization | | | | | | | | |
| True | 0.0239 | 10.840 | -1.5244 | -0.2348 | 0.8151 | 0.2061 | -0.5731 | 0 |
| | (0.009) | (1.261) | (0.669) | (0.041) | (0.064) | (0.058) | (0.194) | |
| s = 0:1 | 0.0275 | 10.491 | -1.5290 | -0.2317 | 0.8087 | 0.2069 | -0.5658 | 111 |
| | (0.009) | (1.248) | (0.666) | (0.040) | (0.064) | (0.057) | (0.193) | |
| s = 0:3 | 0.0356 | 9.6492 | -1.5361 | -0.2239 | 0.7916 | 0.2079 | -0.5463 | 418 |
| | (0.008) | (1.216) | (0.659) | (0.040) | (0.063) | (0.057) | (0.191) | |
| s = 0:5 | 0.0486 | 7.5887 | -1.5473 | -0.2039 | 0.7351 | 0.2058 | -0.4813 | 1578 |
| | (0.005) | (1.130) | (0.633) | (0.038) | (0.061) | (0.055) | (0.184) | |
| s = 0:7 | 0.0442 | 4.9575 | -1.5211 | -0.1749 | 0.6170 | 0.1858 | -0.3396 | 5948 |
| | (0.003) | (0.984) | (0.571) | (0.034) | (0.055) | (0.049) | (0.166) | |

Note: The table reports average estimates and average standard errors (in parentheses) from 1000 simulated samples.

# 6 Conclusions

We show that in 2SLS estimation of linear social network models, measurement errors in the network can be safely ignored by the researcher if the number and magnitude of measurement errors in the adjacency matrix grows sufficiently slowly with the sample size. Moreover, these results hold even if the measurement errors are correlated with model errors, covariates, and outcomes. A useful agenda for future work would be to see if similar results can be obtained for more general network models.

# Appendix

For a generic matrix A, let $A_{(i)}$, $A_{[k]}$ denote its i-th row and k-th column respectively; and $A_{ij}$ denote its $(i; j)$-th component, so that $A_{(i)}$ is the sum of the i-th row in A. Let $_{1n} \equiv H_n - G_n$ and $_{2n} \equiv H_n - C_n$ with $H_{ii} = 0$ by construction. With row normalization,

$$_{1n} \equiv H_n - G_n = \mathrm{diag} \quad \frac{1}{G_{(1)}}; ::::; \frac{1}{G_{(n)}} \quad _{1n} + \mathrm{diag}$$

Hence, there exists some constant $M < \infty$ with $\Pr\{\sup_i E(|y_i|| \;\; _n) \leq M \} = 1$. □

Proof of Proposition 1 . Recall

$$\hat{b} - {}_0 = \left[ \frac{\tilde{R}_n'\tilde{V}_n}{n} \left( \frac{\tilde{V}_n'\tilde{V}_n}{n} \right)^{1} \frac{\tilde{V}_n'\tilde{R}_n}{n} \right]^{1} \frac{\tilde{R}_n'\tilde{V}_n}{n} \left( \frac{\tilde{V}_n'\tilde{V}_n}{n} \right)^{1} \frac{\tilde{V}_n'\tilde{e}_n}{n} \tag{4}$$

where

$$\begin{aligned}
\frac{1}{n}\tilde{V}_n'\tilde{R}_n &= \frac{1}{n}V_n'R_n + \frac{1}{n}V_n'(0; \;\; {}_{1n}Y_n; 0; \;\; {}_{2n}X_n) \\
&+ \frac{1}{n}(0; (G_n \;\; {}_{1n} + \;\; {}_{1n}G_n + \;\; {}^2_{1n})X_n; 0; \;\; {}_{2n}X_n)'R_n \\
&+ \frac{1}{n}(0; (G_n \;\; {}_{1n} + \;\; {}_{1n}G_n + \;\; {}^2_{1n})X_n; 0; \;\; {}_{2n}X_n)'(0; \;\; {}_{1n}Y_n; 0; \;\; {}_{2n}X_n):
\end{aligned}$$

$$\frac{1}{n}\tilde{V}_n'\tilde{V}_n = \frac{1}{n}V_n'V_n + \frac{1}{n}V_n'(0; (G_n \;\; {}_{1n} + \;\; {}_{1n}G_n + \;\; {}^2_{1n})n + \;\; {}^2_1$$

Using Lemma A1, we can show that

$$\widehat{A} = A + O_p(n^{s-1})$$

and

$$\widehat{B} = B + \frac{R_n'V_n}{n}\left(\frac{V_n'V_n}{n}\right)^{-1}\left(\frac{1}{n}\widetilde{V}_n'\widehat{\Sigma}_n\widetilde{V}_n - \frac{1}{n}V_n'\Sigma_n V_n\right)\left(\frac{V_n'V_n}{n}\right)^{-1}\frac{V_n'R_n}{n} + O_p(n^{s-1}).$$

Then, what left is to show that from the fact that $\frac{1}{n}\widetilde{V}_n'\widehat{\Sigma}_n\widetilde{V}_n - \frac{1}{n}V_n'\Sigma_n V_n$ is $o_p(1)$. As

$$\frac{1}{n}\widetilde{V}_n'\widehat{\Sigma}_n\widetilde{V}_n - \frac{1}{n}V_n'\Sigma_n V_n = \frac{1}{n}V_n'\left(\widehat{\Sigma}_n - \Sigma_n\right)V_n + O_p(n^{s-1});$$

and the first term on the RHS is $O_p(n^{-1/2} \vee n^{s-1})$ because

$$\begin{aligned}
\frac{1}{n}V_n'\left(\widehat{\Sigma}_n - \Sigma_n\right)V_n &= \frac{1}{n}\sum_{i=1}^{n}\left\{[(Y_n - \widetilde{R}_n\widehat{\beta})_{(i)}]^2 - E(\varepsilon_i^2)\right\}v_iv_i' \\
&= \frac{1}{n}\sum_{i=1}^{n} v_iv_i'[\varepsilon_i^2 - E(\varepsilon_i^2)] + \frac{1}{n}\sum_{i=1}^{n} v_iv_i'\left\{[\widetilde{R}_i(\beta_0 - \widehat{\beta})]^2 + [(\delta_0\,\delta_{1n}Y_n + \delta_{2n}X_n\beta_0)_{(i)}]^2\right\} \\
&\quad + \frac{2}{n}\sum_{i=1}^{n} v_iv_i'\widetilde{R}_i(\beta_0 - \widehat{\beta})\varepsilon_i - \frac{2}{n}\sum_{i=1}^{n} v_iv_i'[\widetilde{R}_i(\beta_0 - \widehat{\beta}) + \varepsilon_i](\delta_0\,\delta_{1n}Y_n + \delta_{2n}X_n\beta_0)_{(i)} \\
&= O_p(n^{-1/2}) + O_p(\beta_0 - \widehat{\beta}) + O_p(n^{s-1}) = O_p(n^{-1/2} \vee n^{s-1})
\end{aligned}$$

Together, we have $\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1} - A^{-1}BA^{-1} = O_p(n^{-1/2} \vee n^{s-1}) = o_p(1)$.

# References


Blume, L., W. Brock, S. Durlauf, R. Tayaraman, 2015. Linear so65t.879Td[TJ/F227.97Tf6.5870T3648(20

Lee, L., X. Liu, X. Lin, 2010. Specification and estimation of social interaction models with network structures. Econometrics Journal 13: 145–176.