DISCUSSION PAPER SERIES

No. 10575

THE MISSING TRANSFERS: ESTIMATING MIS-REPORTING IN DYADIC DATA

Margherita Comola and Marcel Fafchamps

DEVELOPMENT ECONOMICS

Centre for Economic Policy Research

ISSN 0265-8003

THE MISSING TRANSFERS: ESTIMATING MIS-REPORTING IN DYADIC DATA

Margherita Comola and Marcel Fafchamps

Discussion Paper No. 10575
May 2015
Submitted 22 April 2015

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801
www.cepr.org

THE MISSING TRANSFERS: ESTIMATING MIS-REPORTING IN DYADIC DATA[†]

Abstract

Many studies have used self-reported dyadic data without exploiting the pattern of discordant answers. In this paper we propose a maximum likelihood estimator that deals with mis-reporting in a systematic way. We illustrate the methodology using dyadic data on inter-household transfers from the village of Nyakatoke in Tanzania, investigating the role of wealth in link formation. Our results suggest that observed transfers are grounded in mutual self-interest, and we show that not taking reporting bias into account leads to incorrect inference and serious underestimation of the total amount of transfers between villagers. The method introduced here is applicable whenever the researcher has two discordant measurements of the same dependent variable.

JEL Classification: C13, C51 and D85
Keywords: dyadic data, informal transfer, reporting bias and social networks

Margherita Comola comola@pse.ens.fr
Paris School of Economics

Marcel Fafchamps fafchamp@stanford.edu
Stanford University and CEPR

[†] We are indebted to Joachim De Weerdt for sharing his data and answering our questions. We benefitted from useful comments from seminar participants at the Paris School of Economics, Oxford University, Norwegian School of Economics and Business Administration, Stanford and Yale. All remaining errors are our own.

1 Introduction

It is increasingly common for surveys to collect information on social links and interpersonal flows – *e.g.*, friendship, loans and gifts, advice, referral. In particular, much social network analysis is based on dyadic data reported by survey respondents – *e.g.*, answers to questions such as 'to whom did you lend money', 'who are your friends', 'with whom do you exchange information', or 'are you related to X' (*e.g.*, Fafchamps and Lund 2003; Christakis and Fowler 2009; Steglich, Snijders, and Pearson 2010; Banerjee *et al.* 2013). In principle answers to these questions should agree: if for instance i reports lending money to j, then j should report receiving money from i. Yet it is common for such data to be discordant, *i.e.*, there often are considerable discrepancies between answers given by i and j (Ball and Newman 2013). Until now, mis-reporting has typically been ignored and estimation has proceeded using information reported by i, j, or a combination of the two (*e.g.*, Glaeser, Sacerdote and Scheinkman 1996; Snijders, Koskinen and Schweinberger 2010; Liu *et al.* 2011). However, failing to properly account for mis-reporting may bias the estimation results. This paper investigates a case in which mis-reporting affects estimation and inference in self-reported dyadic data, and proposes an estimator that deals with the problem.

We illustrate our methodology using data on informal transfers from rural Tanzania. Informal transfers have been recognized a great importance for development, since they represent a source of favour exchange and insurance against idiosyncratic shocks. In particular, many studies have investigated informal transfers by using self-reported dyadic transfer data. For instance, Fafchamps and Lund (2003) and De Weerdt and Fafchamps (2011) use transfers information obtained from one of the two households only, while Fafchamps and Gubert (2007) combine the two answers to construct a unique measure of transfers (see also Attanasio *et al.* 2012). All these studies neither exploit the systematic pattern of discordant answers in the data, nor investigate the consequences of mis-reporting. In this paper we show that incorrect inference may be drawn about the coefficients of interest in such type of data, and that the amount of informal transfers occurring at the village level may be largely underestimated.

Faced with discordant transfer data, researchers typically rely on *ad hoc* assumptions. They often assume that if either i or j report a transfer, then a transfer between i and j took place; this is equivalent to assuming that all observed discordances are due to under-reporting. Alternatively, they may assume that a transfer between i and j took place only if both i and j reported it; this is equivalent to assuming that all observed discordances are due to over-reporting. Both assumptions rule out the pos-

sibility that a transfer occurred but was not declared by anyone, or that some transfer declared by both individuals in reality did not occur.

We propose a maximum likelihood estimator that deals with discordant answers in a systematic way. Our estimator accounts separately for the propensity of i and j to report a transfer, which may depend on observables. It forces the researcher to assume either under- or over-reporting in the underlying data generation process. But it also allows to investigate the sensitivity of the findings to assuming one or the other.[1] While there is an established literature on measurement error in binary variables (*e.g.* Hong and Tamer 2003; Schennach 2004), to the best of our knowledge this paper offers the first specific solution for dyadic data. The method we propose to correct for mis-reporting is of particular interest to researchers studying social networks but it is also suitable for any pairwise data with two discordant self-reported measures of the same objective phenomenon, *e.g.*, multiple measurements of schooling levels in twins (Ashenfelter and Krueger, 1994), discrepancies over earnings reported by workers and companies (Duncan and Hill, 1985), estimates of time spent on housework by the spouse (Lee and Waite 2005), bilateral trade flows reported by exporters and importers which need to be reconciled (Gaulier and Zignago 2010).

Simulations suggest that more accurate inference is obtained with our mis-reporting correction. In particular, we show that estimation results are sensitive to mis-reporting if the propensity to report is correlated with the regressors of interest. To understand why, consider the following example. Imagine we have data on households' ethnicity (two groups: A and B) and transfers between them. Assume that households from group A tend to give and receive more transfers, but they are also less likely to subsequently report transfers. If the researcher assumes that a transfer took place only if at least one side i or j reported it, the estimated coefficient of belonging to ethnic group A on the probability of a transfer from i to j will be biased downwards. This is because the researcher observes transfers less frequently when giver and receiver households are from this group – but this is entirely due to differential mis-reporting.

We illustrate our method using dyadic data from the village of Nyakatoke in Tanzania. The data contain detailed information on all monetary transfers (loans and gifts) between all households in the village, and display large discrepancies in survey

[1]The appropriateness of assuming under- versus over-reporting depends on the context. In many cases it is reasonable to assume that the main reason for discrepancies is under-reporting: a transfer took place but one of the parties involved forgot to report it to enumerators. It can also happen that links or flows are suspected to be over-reported, as when individuals inflate the number of their sexual partners.

responses about transfers given and received. Specifically, we investigate whether observed transfers are grounded in mutual self-interest. Our identifying assumption is that wealthy villagers are more desirable partners because they may be a source of material favours. We find reasonably strong evidence that monetary flows take place mostly between households who are desirable partners for each other, as defined by their wealth level. This finding is in line with much of the economic literature on risk sharing which emphasizes self-interest as basis for mutual support (Coate and Ravallion 1993, Ligon, Thomas and Worrall 2001, Attanasio *et al.* 2012). This finding, however, only emerges when we correct for mis-reporting. Not taking mis-reporting into account leads to erroneous conclusions in terms of the statistical significance of estimated coefficient, and seriously underestimates the total amount of transfers between Nyakatoke villagers. These results cast some doubt on the reliability of previous results that rely on transfers reported in household surveys. In particular, many studies have found that reported gifts and loans are insufficient to insulate households against shocks. But if actual gifts and loans are much larger, these findings may be called into question. For instance, Rosenzweig (1988) reports that loans between households represent only 2% of the value of the shocks they face. If there is as much under-reporting in his data as in ours, the correct figure is probably closer to 5%.

The paper is organized as follows. In Section 2 we describe the estimation strategy and simulation analysis. The data are illustrated in Section 3, and results are discussed in Section 4. Section 5 focuses on the estimates of under-reporting, while Section 6 concludes. Additional figures and tables are reported in Appendix A. Appendix B illustrates how to implement our estimator under the assumption of over-reporting. Appendix C demonstrates how alternative estimates can be generated under different identification assumptions about the correlation of reporting errors.

2 Estimation strategy

2.1 The estimator

In our empirical analysis, τ_{ij} refers to a binary transfer from i to j over a given time interval. More generally, we think of τ_{ij} as capturing any manifestation of a social link, typically a flow of money, goods, or favors. Our objective is to estimate a regression model of the form:

$$\Pr(\tau_{ij} = 1) = \lambda(\beta_\tau X_\tau^{ij}) \tag{1}$$

where X_τ^{ij} is a vector of controls for dyad ij, β_τ is a coefficient vector of interest, and λ is the logit function. We focus on the case where the data contain two reports on τ_{ij}, *i.e.* both i and j were (separately) asked to report τ_{ij}. Let G_{ij} be the report that the giver i made on the true transfer τ_{ij} and let R_{ij} be the report that the receiver j made on the same transfer τ_{ij}. In principle, i and j should report the same thing, *i.e.*, we should observe $G_{ij} = R_{ij}$. This is not typicaly the case, however. For instance, in the dataset that we use for illustration purposes, when respondent i reports $G_{ij} = 1$, in the majority of cases respondent j reports $R_{ij} = 0$.

In what follows we assume that the source of mis-reporting in data is under-reporting, for instance driven by poor recall. With under-reporting, if a flow is reported by either i or j, then it must have taken place. But a flow may also have taken place even if it was not reported by either i or j. We propose a maximum likelihood estimator that corrects for such mis-reporting pattern. Whether under-reporting is a reasonable assumption or not depends on the context. It seems to us the most reasonable for our application on transfers data in Tanzania. Appendix B illustrates how the methodology can be amended to deal with the polar assumption of over-reporting, and confirms that under-reporting is most appropriate for the data at hand.

Dropping the ij subscripts to improve readability, let τ denote the true binary flow or transfer from i to j, *i.e.*, $\tau = 1$ if i made a transfer to j. We have $G = 1$ if i reported making a transfer and 0 otherwise. Similarly, $R = 1$ if j reported receiving a transfer, and 0 otherwise. We do not observe τ, only G and R. Under-reporting implies that $G = 1$ only if $\tau = 1$, and that $R = 1$ only if $\tau = 1$. However, it could be the case that $G = 0$, $R = 0$ and still $\tau = 1$. Given these assumptions, the data generation process takes the following form:

$$\begin{aligned}
\Pr(G = 1,\, R = 0) &= \Pr(\tau = 1, G = 1,\, R = 0) \\
&= \Pr(\tau = 1) * \Pr(G = 1|\tau = 1) * \Pr(R = 0|G = 1,\, \tau = 1) \\
\Pr(G = 0,\, R = 1) &= \Pr(\tau = 1, G = 0,\, R = 1) \\
&= \Pr(\tau = 1) * \Pr(G = 0|\tau = 1) * \Pr(R = 1|G = 0,\, \tau = 1) \\
\Pr(G = 1,\, R = 1) &= \Pr(\tau = 1, G = 1,\, R = 1) \\
&= \Pr(\tau = 1) * \Pr(G = 1|\tau = 1) * \Pr(R = 1|G = 1,\, \tau = 1) \\
\Pr(G = 0,\, R = 0) &= 1 - \Pr(G = 1,\, R = 0) - \Pr(G = 0,\, R = 1) - \Pr(G = 1,\, R = 1)
\end{aligned}$$

As is formally shown in Appendix C, it is not possible to estimate the above probabilities from observed moments of the data: the above model is unidentified. We need to make one maintained assumption in order to achieve identification. Here we opt for what we think is the least problematic assumption, namely that under-reporting by i is independent of under-reporting by j, then $\Pr(R|G,\tau) = \Pr(R|\tau)$. This assumption is reasonable if under-reporting results primarily from mistakes and omissions.[2] With this assumption, we can rewrite the system as:

$$\begin{aligned}
\Pr(G=1,R=0) &= \Pr(\tau=1)*\Pr(G=1|\tau=1)*\Pr(R=0|\tau=1) && (2)\\
\Pr(G=0,R=1) &= \Pr(\tau=1)*\Pr(G=0|\tau=1)*\Pr(R=1|\tau=1) && (3)\\
\Pr(G=1,R=1) &= \Pr(\tau=1)*\Pr(G=1|\tau=1)*\Pr(R=1|\tau=1) && (4)\\
\Pr(G=0,R=0) &= 1-\Pr(G=1,R=0)-\Pr(G=0,R=1)-\Pr(G=1,R=1) && (5)
\end{aligned}$$

Equations (2) to (5) express the data generating process in terms of three probabilities: $P(\tau=1)$, $P(G=1|\tau=1)$ and $P(R=1|\tau=1)$. To obtain the likelihood function, we assume that these three probabilities can be represented by three distinct logit functions $\lambda(.)$ as follows:

$$\begin{aligned}
\Pr(\tau=1) &= \lambda(\beta_\tau X_\tau) && (6)\\
\Pr(G=1|\tau=1) &= \lambda_G(\beta_G X_G) && (7)\\
\Pr(R=1|\tau=1) &= \lambda_R(\beta_R X_R) && (8)
\end{aligned}$$

Together with (2) to (5), equations (6) to (8) fully characterize the likelihood of observing the data. The main equation of interest is $\Pr(\tau=1) = \lambda(\beta_\tau X_\tau)$: it is on this equation that we wish to test restrictions on the true parameter vector β_τ. Equations (7) and (8) condition on individual observables X_G and X_R, respectively. As we illustrate below, conditioning $\Pr(G=1|\tau=1)$ and $\Pr(R=1|\tau=1)$ in this manner is

[2]Setting $\tau_{ij} = \max\{G_{ij}, R_{ij}\}$ as it is common in the social network literature is equivalent to assuming perfect negative correlation between $G|\tau$ and $R|\tau$ – *i.e.*, i remembers when j does not and *vice versa*. This is an unreasonable assumption in most cases. Assuming perfect positive correlation between $G|\tau$ and $R|\tau$ rules out discordant answers, a feature that is trivially rejected in most datasets, including the one we use in our empirical illustration. With only two reports R and G, it is not possible to estimate a model that allows for arbitrary correlation between $G|\tau$ and $R|\tau$ (see Appendix C). This leaves independence as the most realistic option. As explained below, by conditioning $\Pr(G|\tau)$ and $\Pr(R|\tau)$ on individual observables X_G and X_R, we nonetheless correct for correlation in reporting G and R that is predicted by correlation between X_G and X_R.

often essential to obtain correct inference. We also note that by conditioning $\Pr(G|\tau)$ and $\Pr(R|\tau)$ on X_G and X_R, we correct for correlation in reporting G and R that is predicted by correlation between X_G and X_R.[3]

To illustrate how our correction for mis-reporting affects inference, we will compare the estimated results from $\Pr(\tau = 1)$ with two standard logit regressions which are commonly used in the network literature. In the first of them, the dependent variable equals one if at least one side has declared a transfer, which is equivalent to defining $\tau_{ij}^{max} \equiv \max\{G_{ij}, R_{ij}\}$. This assumes that when both reports agree they are true statements and all discordances are due to under-reporting. In the second regression the dependent variable equals one if both the giver and the receiver have declared a transfer, *i.e.*, it is $\tau_{ij}^{min} \equiv \min\{G_{ij}, R_{ij}\}$. This is equivalent to assuming that when both reports agree they are true statements and all discordances are due to over-reporting. These scenarios by construction rule out the possibility either that a transfer occurred but was not declared by anyone, or that a transfer declared by both parties did not occur.

Dyadic observations are typically not independent. This does not invalidate the application of standard maximum likelihood techniques to estimate β_τ, β_G and β_R in equations (6) to (8). But standard errors must be adjusted to correct for dyadic dependence across observations, otherwise inference will be inconsistent. Since we only have data from a single population,[4] we apply the formula developed by Fafchamps and Gubert (2007) which corrects for arbitrary correlation across all τ_{ij} and τ_{ji} observations involving either i or j.

2.2 Simulation analysis

Whether or not mis-reporting affects inference depends on the hypothesis being tested, that is, on the regressors in equation (6). To illustrate this point, we conduct an extensive simulation analysis to investigate how our estimator and the standard logit regressions behave when reporting propensities $\lambda_G(\beta_G X_G)$ and $\lambda_R(\beta_R X_R)$ vary systematically with the regressors of interest. Results discussed below show that our estimator always delivers satisfactory coefficients, while the results from the standard logit estim-

[3]For instance, if wealthy households are less likely to report making a transfer than poor households and wealth is correlated across giving and receiving households, this can be controlled for by including the wealth of the giver in X_G and the wealth of the receiver in X_R.

[4]If we had data from a sufficient number of distinct sub-populations we could cluster the standard errors to correct for correlation across observations from the same sub-population (Arcand and Fafchamps 2012).

ates can be severely biased. They also clarify the conditions under which the inclusion of certain regressors in the reporting equations affect inference.

We posit a data generating process of the form

$$\Pr(\tau_{ij} = 1) = \lambda(\beta_{\tau 0} + \beta_{\tau 1} x_i + \beta_{\tau 2} x_j + \beta_{\tau 3} d_{ij} + \varepsilon_{\tau ij}) \tag{9}$$

where τ_{ij} is the real transfer from i to j, x_i and x_j are two uniformly distributed individual attributes (for instance wealth), d_{ij} is a uniformly distributed relational attribute (for instance geographic distance), the error term $\varepsilon_{\tau ij} \backsim N(0,1)$ and λ is the logit function. While τ_{ij} stays unobserved, we generate the two individual binary reports G_{ij}, R_{ij} under different mis-reporting scenarios as follows:

- Under Scenario 1 we impose that mis-reporting is purely random, *i.e.,* $\Pr(G_{ij} = 1) = \lambda(\beta_{G0} + \varepsilon_{Gij})$ and $\Pr(R_{ij} = 1) = \lambda(\beta_{R0} + \varepsilon_{Rij})$ with ε_{Gij}, $\varepsilon_{Rij} \backsim N(0,1)$ and $E[\varepsilon_{Gij}\, \varepsilon_{Rij}] = 0$.

- Under Scenario 2 we generate mis-reporting on the basis of individual attributes, *i.e.,* $\Pr(G_{ij} = 1) = \lambda(\beta_{G0} + \beta_{G1} x_i + \varepsilon_{Gij})$ and $\Pr(R_{ij} = 1) = \lambda(\beta_{R0} + \beta_{R2} x_j + \varepsilon_{Rij})$. This corresponds to the case where respondents with a high x (*e.g.*, a high wealth in our empirical analysis below) are more likely to report transfers given and received. We maintain ε_{Gij}, $\varepsilon_{Rij} \backsim N(0,1)$ and $E[\varepsilon_{Gij}\, \varepsilon_{Rij}] = 0$.

- Under Scenario 3 we generate mis-reporting on the basis of the relational attribute, *i.e.,* $\Pr(G_{ij} = 1) = \lambda(\beta_{G0} + \beta_{G3} d_{ij} + \varepsilon_{Gij})$ and $\Pr(R_{ij} = 1) = \lambda(\beta_{R0} + \beta_{R3} d_{ij} + \varepsilon_{Rij})$. This corresponds to the case where transfers to (geographically or socially) proximate households are easier to recall.

- Under Scenario 4 we generate mis-reporting on the basis of both individual and relational attributes, *i.e.,* $\Pr(G_{ij} = 1) = \lambda(\beta_{G0} + \beta_{G1} x_i + \beta_{G3} d_{ij} + \varepsilon_{Gij})$ and $\Pr(R_{ij} = 1) = \lambda(\beta_{R0} + \beta_{R2} x_j + \beta_{R3} d_{ij} + \varepsilon_{Rij})$.

Under all four scenarios we calibrate the reporting propensity of givers and receivers to be 60% and 40%, respectively, conditional on $\tau_{ij} = 1$. This approximately matches the relative proportions in our observational data. For each of these scenarios we draw 250 random matrices of transfer flows and we compare the performance of our estimator with standard logit regressions. The simulation results are summarized in Table 1. Additionally, for the most complete misreporting Scenario 4, we plot kernel

densities of the Monte Carlo estimates for $\beta_{\tau 1}$ (Figure 1, appendix A), $\beta_{\tau 2}$ (Figure 2, appendix A) and $\beta_{\tau 3}$ (Figure 3, appendix A).

Table 1. Simulation results

	(1)	(2)	(3)	(4)	(5)
	true model	our estimator	our estimator	standard logit	standard logit
	τ_{ij}	intercept only	with covariates	τ_{ij}^{max}	τ_{ij}^{min}
	Scenario 1:				
$\beta_{\tau 1}$	1.73	1.75	1.76	1.48	1.13
$\beta_{\tau 2}$	1.73	1.75	1.75	1.48	1.14
$\beta_{\tau 3}$	-1.73	-1.74	-1.75	-1.45	-1.09
	Scenario 2:				
$\beta_{\tau 1}$	1.73	2.3	1.72	1.92	1.83
$\beta_{\tau 2}$	1.74	2.12	1.72	1.77	2.21
$\beta_{\tau 3}$	-1.74	-1.83	-1.73	-1.51	-0.97
	Scenario 3:				
$\beta_{\tau 1}$	1.73	1.72	1.76	1.48	1.18
$\beta_{\tau 2}$	1.73	1.73	1.76	1.48	1.19
$\beta_{\tau 3}$	-1.74	-1	-1.75	-0.8	0.52
	Scenario 4:				
$\beta_{\tau 1}$	1.74	2.26	1.73	1.92	1.85
$\beta_{\tau 2}$	1.73	2.07	1.72	1.75	2.23
$\beta_{\tau 3}$	-1.73	-1.04	-1.72	-0.86	0.64

Column (1) of Table 1 reports the average logit coefficients over the 250 replications when we estimate equation (9) using the actual transfer τ_{ij} as dependent variable. Column (2) reports the average estimated coefficients from the misreporting-corrected model of equation (6) when we only include the intercept term in the reporting equations. Column (3) reports average estimated coefficients when we include x_i and d_{ij} in X_G and we include x_j and d_{ij} in X_R for the reporting regressions. Column (4) reports average estimated coefficients when we posit $\tau_{ij}^{max} \equiv \max\{G_{ij}, R_{ij}\}$ and estimate equation (9) applying standard logit methods to τ_{ij}^{max}. Column (5) reports average estimated coefficients if we instead let $\tau_{ij}^{min} \equiv \min\{G_{ij}, R_{ij}\}$ and apply standard logit methods to τ_{ij}^{min}.

Results show that our estimator outperforms the standard logit regressions of columns (4) and (5) in all cases. Under Scenario 1 our estimator does equally well whether or not we condition the reporting equations on observables. When we do not correct for mis-reporting, the magnitude of the estimated coefficients is biased downwards – more severely in column (5) than in column (4). Under Scenarios 2, 3 and 4 where reporting propensities depend on observables, our estimator delivers consistent results only if we include the controls in the reporting equations. In particular, our estimator with covariates (column 3) always delivers satisfactory coefficients. This is not the case for our estimator with intercept only (column 2) or for the standard logit regressions (column 4 and 5). The bias in estimated coefficients is particularly severe for the variable(s) that affects reporting in the data generating process: in Scenario 2 and 4 β_1 and β_2 are upward biased in all columns except column (3), and similarly in Scenario 3 and 4 β_3 is always upward biased with the exception of column (3). Our estimator seems to perform better that the standard logit regressions even when we only include an intercept in the reporting equations, as in column (2). Indeed, for columns (4) and (5) the coefficients of regressors that do not enter the reporting equations (*i.e.,* $\beta_{\tau 3}$ for Scenario 2 and $\beta_{\tau 1}$,$\beta_{\tau 2}$ for Scenario 3) are more severely biased than in column (2). The Kernel plots from the 250 simulated networks of Scenario 4 reported in Figures 1 to 3 (Appendix A) confirm the results discussed above. They clearly show that our our estimator with covariates brings important gains in terms of inference, with no major loss in efficiency.

Overall, the simulation exercise suggests that, if the self-reporting of transfer data has the general properties sketched above, our estimator perform well in estimating equation (1) while standard logit regressions yields incorrect inference. If certain regressors are omitted from the reporting equations, they can get biased coefficients in the main regression but only if they are correlated with reporting propensities. This also coincides with the situation in which our method delivers the biggest improvement relative to alternative logit estimators. Results also indicate that identification does not require that the regressor sets X_G and X_R contain a variable absent from X_τ. In other words, correct inference does not necessarily depend on the availability of 'instruments', *i.e.*, excluded variables, to identify the reporting equations.

We also investigate the behaviour of the estimator when we relax the assumption of independence in ε_{Gij} and ε_{Rij}. In order to do that we recompute the simulation results with different nonzero values of the correlation coefficient ρ between ε_{Gij} and ε_{Rij}. Results are presented in Tables A1 to A6, Appendix A. We first focus on positive

correlation in reporting propensities, for a set of plausible correlation values. If $\rho = 1$ there is no misreporting, in which case there are no discordant answers, $G = R$ always, and $\tau_{ij}^{max} = \tau_{ij}^{min}$. In this case all three estimators are identical. To account for discordant answers, it is necessary that $\rho < 1$. The lower ρ, the more discordant answers there are. For values of ρ above 0.5, the proportion of discordant answers among all reported transfers is fairly small.[5] For this reason we focus on values of $\rho \leq 0.5$, reporting simulation results for $\rho = 0.1$ (Table A1), $\rho = 0.3$ (Table A2) and $\rho = 0.5$ (Table A3). Our estimator performs well in all three scenarios. When correlation is low (Table A1) results are very similar to those of Table 1, and our estimator with covariates (column 2) proves unambiguously superior to the standard logit regressions of column 4 and 5. When the correlation in misreporting increases (Table A2 and A3), there is a bias in the estimated coefficients, but our estimator still delivers the most accurate results. In particular, our estimator with covariates performs always better than the standard logit regressions, and it performs much better when the relational attribute enters the reporting equations: in scenarios 3 and 4 the coefficient $\beta_{\tau 3}$ is mildly biased in column (2), but severely biased in all other columns.

Finally we also recompute all simulation results under the hypothesis of negative correlation in reporting propensities for $\rho = -0.1$ (Table A4), $\rho = -0.3$ (Table A5) and $\rho = -0.5$ (Table A6). When correlation is low (Table A4) and intermediate (Table A5) we observe more noise that in the positive correlation case, but all previous results stand: under all four scenarios our estimator with covariates (column 3) displays the best performance, and this is especially true for the situations where the relational attribute enters the reporting equations. In the high correlation case (Table A6) results are less conclusive, and we observe situations in which a standard logit on τ_{ij}^{max} outperforms our estimator by a small extent. In conclusion, based on the overall evidence from the simulations above, our method still provides a safe choice even when the researcher suspect a reasonable level of correlation of unknown sign between reporting propensities. For more details on the identification of the model with correlated reports and on plausible correlation values that can be reconciled with our data we remand to Appendix C.

[5] For instance, if $\Pr(G|\tau) = \Pr(R|\tau) = 0.5$ and $\rho = 0.5$, the proportion of discordant answers among all reported tranfers is 28.5% – much lower than in our data. This proportion rises to 67% for $\rho = 0$. In our data the proportion of discordant answers is 73%.

3 Informal Transfers in Tanzania

3.1 Nyakatoke household survey

We illustrate our methodology using a unique census dataset on transfers between all the households in an African village, Nyakatoke. The village is located in the Buboka Rural District of Tanzania, at the west of Lake Victoria. The data have been the object of numerous articles (*e.g.* De Weerdt and Dercon 2006; De Weerdt and Fafchamps 2011; Vandenbossche and Demyunck 2013; Comola 2012; Comola and Fafchamps 2014).

The community is composed by 600 inhabitants, 307 of which are adults.[6] A total of 119 households were interviewed in five rounds at regular intervals from February to December 2000. In the first survey round (February 2000), each adult was asked to whom he would ask and/or provide help in case of need.[7] During each of the subsequent interview rounds, each adult was asked whether they had received or given transfers (loans or gifts). If they said yes, information was collected together with the name of the partner and the value of what was given or received, whether in cash or kind.[8] This provides us with a detailed picture of all transfers occurring within the village over one year. These transfers have been shown to serve an insurance purpose against health shocks (De Weerdt and Fafchamps 2011).[9] This is in line with the literature on informal risk sharing which has shown how informal transfers can be a way of smoothing consumption against shocks while satisfying self-enforcement constraints (Udry 1994; Kocherlakota 1996; Foster and Rosenzweig 2001; Ligon Thomas and Worrall 2001).

3.2 Transfer data

In order to map the transfers between Nyakatoke households we aggregate the individual-level information on transfers at the household level, across rounds, and across types of transfer. We aggregate at the household level to reduce the discrepancies that could arise if i mentioned giving to member a of household j but member b of household j is the one who mentions receiving a transfer from i.[10] We aggregate across rounds to

[6] Individuals aged 16 and above are considered adult.

[7] *"Can you give a list of people from inside or outside of Nyakatoke, who you can personally rely on for help and/or that can rely on you for help in cash, kind or labor?"*

[8] Loan repayment and gifts in labor are not included.

[9] This is consistent with findings reported by Fafchamps and Lund (2003) for the Philippines.

[10] When aggregating at the household level, questionnaires were carefully checked by survey supervisors to avoid any double-counting of identical gifts reported by two different members of the same household.

reduce discrepancies that could arise if household i declares a transfer in round t while household j declares that same transfer in round $t+1$. Finally, we aggregate loans and gifts into a unique transfer measure in order to avoid discrepancies due to the fact that household i declares a loan while household j reports that same transfer as a gift.[11]

Our unit of observation is the dyad: in Nyakatoke there are 119 households, which gives $119 * 118 = 14042$ dyads. For each household dyad ij we have four measurement of the transfer among them: transfer G_{ij} that i stated giving to j; transfer R_{ij} that j stated receiving from i; transfer G_{ji} that j declared giving to i; and transfer R_{ji} that i stated receiving from j. These four measurements correspond to two actual unobserved gross flows: the flow from i to j, denoted τ_{ij}, and the flow from j to i, denoted τ_{ji}. Since we focus on gross flows, the two are not the same. Hence $\{\tau_{ij}\}$ defines a directed graph.

There are major discrepancies between G_{ij} and R_{ij}. In fact, $G_{ij} \neq R_{ij}$ in nearly all cases. There are 1721 dyads (*i.e.*, 12.26% of the household dyads) for which either G_{ij} or R_{ij} is not zero. In 769 cases the report comes from the giver only (5.48% of the dyads), in 481 cases from the receiver only (3.43% of the dyads), and in 471 from both (3.35% of the dyads). Out of the 471 dyads in which both i and j report a transfer from i to j, only 23 report the exact same amount, and the amounts declared tend to differ by a large margin (*i.e.* the highest of the two declared amounts is on average double the smallest one). Amounts reported by both sides are on average larger than amounts reported by one side only.[12] The frequency distribution of transfers is given in Table A.7, Appendix A.

In summary, there are massive discrepancies between the responses given by i and j about the same transfers. These discrepancies are mostly due to the fact that one side reports something while the other reports nothing. Under-reporting by those who receive transfers may not be too surprising: they may have a strategic motive in 'forgetting' the favors that they probably have a moral obligation to reciprocate. But we also sense massive under-reporting by those who give. Consequently there may be many transfers which took place but are not observed in the data because they were not mentioned by either sides. When estimating model (1), our main challenge is to

[11] As a robustness check we have also conducted separate analysis for loans and gifts. These lead to similar conclusions.

[12] For instance, the average value declared by the receiver is 2440 Tanzanian shillings (*tzs*) when the giver also declares a non-zero amount, and 1468 *tzs* when the giver does not declare any transfer. This is consistent with the idea that respondents are more likely to recall large transfers than small transfers.

address this source of bias.[13]

3.3 Mutual self-interest in link formation

A favor exchange relationship can arise when both households see it in their mutual self-interest. This may happen, for instance, because they share the same status and consequently need not fear that exchanging favors and sharing risk will be to the sole advantage of one of them. Favor exchange may also arise when one of the two households benefits disproportionately from the relationship. This can occur for a variety of reasons, such as altruism or sharing norms.

Formally, let d_{ij} proxy for household i's material interest in exchanging favors with household j. This material interest could be, for instance, j's wealth: a richer household is a better source of material assistance. Similarly, let d_{ji} proxy for j's material interest in exchanging transfers with i. In order to investigate whether observed transfers are mutually beneficial, we estimate an equation (6) of the form:

$$\Pr(\tau_{ij} > 0) = \lambda(\alpha d_{ij} + \beta d_{ji} + \gamma d_{ij} d_{ji} + \theta Z_{ij\tau}) \tag{10}$$

Here $X_{ij\tau} \equiv [d_{ij}, d_{ji}, d_{ij}d_{ji}, Z_{ij\tau}]$. If transfers only flow between two households when both have a material interest in a favor exchange relationship, transfers between i and j should only be observed if both i and j benefit from the link, that is, when both d_{ij} and d_{ji} are large. This means that, once we control for $d_{ij}d_{ji}$, variables d_{ij} and d_{ji} should have little or no additional predictive power on the probability of observing $\tau_{ij} > 0$. If favor exchange between i and j can arise even when it is only in the material interest of i or j, both d_{ij} and d_{ji} should have a positive coefficient, and $d_{ij}d_{ji}$ should have a negative coefficient to avoid double-counting. If all three coefficients are positive, it means that both types of links coexist in the data.[14]

[13] We have no reason to suspect that respondents report flows that did not take place, since reporting a transfer to an enumerator takes time and effort. There is some evidence of this in the data itself. The fact that transfers reported by both sides are on average larger than transfers reported by one side only is in line with the hypothesis of recall mistakes that decrease in the amount transferred. See also Akee and Kapur (2012) for evidence on reporting bias about transfers.

[14] This logic is best illustrated with a stylized example. Suppose that d_{ij} is a dichotomous variable equal to 1 if i wishes to link with j, and 0 otherwise – and similarly for d_{ji}. In the mutual self-interest case, transfers between i and j should only take place when $d_{ij}d_{ji} = 1$, and variables d_{ij} and d_{ji} should have no additional effect on the probability of observing $\tau_{ij} > 0$. This means that we should observe $\alpha = \beta = 0$ and $\gamma > 0$. In contrast, consider the case when favor exchange can arise if either i or j benefits from it. We should observe $\tau_{ij} > 0$ either when $d_{ij} = 1$ or when $d_{ji} = 1$ or both. We

This logic is the basis for our empirical illustration. For the purpose of this test, the mis-reporting correction may be of great value because we suspect that d_{ij} and d_{ji} may affect not only equation (6), but also the reporting equations (7) and (8). If this is the case, only by correcting for mis-reporting can we draw correct inference about whether observed links are mutually beneficial.

To perform the test we need a proxy for the material interest of household i in establishing a favor exchange relationship with household j. To serve as proxy, we use the potential partner's wealth. The usual caveat applies since this variable is selected by us, based on *a priori* considerations regarding factors likely to affect material self-interest in a link.[15]

3.4 Variable definitions

The regressors used in our analysis are illustrative of the variables typically included in an analysis of this kind. The main equation of interest is $\Pr(\tau = 1) = \lambda(\beta_\tau X_\tau)$. The regressors entering X_τ are control variables expected to influence the actual flows of funds between households. Since τ_{ij} is directional, regressors for observation ij can differ from regressors for observation ji.[16] The regressors of interest for our testing strategy are the wealth of i and j, as well as the interaction term $wealth_i * wealth_j$.[17] From the work of Fafchamps and Lund (2003), De Weerdt and Dercon (2006) and De Weerdt and Fafchamps (2011), we know that informal arrangements are more frequent among households that are socially and geographically proximate. To capture this, we include four relational dummies for whether i and j have the same educational level, share the same religion, are blood related, and are neighbours.[18]

therefore should observe $\alpha = \beta = \alpha + \beta + \gamma > 0$ which implies that $\gamma = -\beta = -\alpha$.

[15]It would have been better if data had been collected on desire to link. However, self-reported desire to link is subject to self-censoring: people often refrain from listing people they truly wish to link with but fear being rejected by (Hitsch, Hortacsu, and Ariely 2010, Belot and Francesconi 2012). It should be possible to design a controlled experiment in which truth-telling is incentivized, or in which the true payoffs are known to the researcher, but experimental data of this kind at the moment do not exist. Given this, the results presented here should be taken as the best suggestive evidence available at this point.

[16]This stands in contrast with undirected network data where $\tau_{ij} \equiv \tau_{ji}$ and regressors by construction have to be identical such that $X_\tau^{ij} = X_\tau^{ji}$.

[17]Wealth is computed as the total value of land assets in Tanzanian shilling (1 unit = 100000 tzs).

[18]Out of 119 households in Nyalatoke, 24 are Muslim (20%), 46 are Protestant (39%) and 49 are Catholic (41%). An household is considered educated if at least one adult member finished primary education, and households i and j are said to have the same educational level if they are both educated, or both not educated. We consider households i and j blood-related if an adult member of i is the

Next we discuss the variables that enter the reporting equations of the giver $\Pr(G = 1|\tau = 1) = \lambda_G(\beta_G X_G)$ and receiver $\Pr(R = 1|\tau = 1) = \lambda_R(\beta_R X_R)$. We include wealth (wealth of i, wealth of j and interaction term) as regressor in both reporting equations since wealthy people may be more or less likely to forget a transfer given or received (Akee and Kapur 2012). All social and geographical proximity variables are included to allow for the possibility that respondents better remember transfers to and from proximate households. We also include regressors that are *a priori* expected to affect mis-reporting but not transfers themselves. For this purpose, we use the total number of declared friends, defined as the individuals living inside or outside the village which were listed in response to the first-round question on who respondents would turn to for help and to whom they would provide help. The logic underlying this choice is that households that intend to seek help from (or provide help to) many other households are probably more sensitive to the issue of inter-household transfers, and therefore recall transfers better. Following this logic we include $declared\ friends_i$ in X_G and $declared\ friends_j$ in X_R.

In Table 2 we present descriptive statistics for all variables used in the analysis. The upper section of the table reports different versions of the dependent variable. The first two rows focus on the transfers from i to j, as reported by i and j respectively. In the next two rows we report the variables $\tau_{ij}^{max} \equiv \max\{G_{ij}, R_{ij}\}$ and $\tau_{ij}^{min} \equiv \min\{G_{ij}, R_{ij}\}$ that are used as dependent variables in the standard logit regressions. These data demonstrate the extent of the divergence between the information given by households i and j on the same τ_{ij}. We also note that givers are more likely to report a transfers than receivers. The rest of Table 2 focuses on regressors. There is considerable variation in wealth levels across Nyakatoke households. 65% of household dyads have the same level of education (*i.e.* are both educated or both non-educated). There is significant diversity in religion: only 35% of households head pairs share the same religion. Around 2% of household pairs are related by blood, and 40% of households are neighbours. The average number of friends declared in the first-round question is 5.29.

parent/sibling/child of an adult member of j. We consider households i and j as neighbours if they live within 400 meters of each other (for 3 households the distance is missing, so we have imputed the sample average).

Table 2. Descriptive statistics (N=14042)

variable	dummy	mean	min	max	sd
τ_{ij}^{i}	yes	0.09			
τ_{ij}^{j}	yes	0.07			
τ_{ij}^{max}	yes	0.12			
τ_{ij}^{min}	yes	0.03			
wealth (i and j)	no	4.01	0	23.09	3.75
$wealth_i$*$wealth_j$	no	15.98	0	378.59	24.89
same education	yes	0.65			
same religion	yes	0.35			
blood link	yes	0.02			
neighbors	yes	0.40			
declared friends (i and j)	no	5.29	0	19	3.06

4 Estimation results

4.1 Main results

Table 3 presents the main results. Columns (1) and (2) report the results from standard logit regressions where the dependent variable is $\tau_{ij}^{max} \equiv \max\{G_{ij}, R_{ij}\}$ and $\tau_{ij}^{min} \equiv \min\{G_{ij}, R_{ij}\}$ respectively. Columns (3) to (5) report jointly estimated coefficients from maximizing the likelihood function defined by equations (2) to (8). Column (3) corresponds to the equation of interest (1), while columns (4) and (5) correspond to the reporting equations of the giver and receiver respectively.

The results from the two approaches are different. In columns (1) and (2) $wealth_i$ and $wealth_j$ are both significantly positive, but the interaction term in non-significant. In contrast, when we correct for mis-reporting in column (3), $wealth_i$ and $wealth_j$ loose significance while their interaction becomes significant. In all cases wealth predicts transfers, but the pattern is different: in column (3), two households are more likely to exchange a transfer only if they are both wealthy; in columns (1) and (2), it only suffices that one of them be wealthy. This means that correcting for mis-reporting affects inference: when we correct for it in (3), the evidence is consistent with mutual self-interest in link formation, but when we do not correct in columns (1) and (2),

the evidence suggests instead that mutual self-interest is not essential. As for the other covariates, people appear more likely to give to relatives and neighbours in all specifications, and to members of the same religion in columns (1) and (2).

Results for the two under-reporting regressions – columns (4) and (5) – show that respondents are more likely to recall a transfer from/to neighbours and relatives, which is not surprising. In the $\Pr(G = 1|\tau = 1)$ regression, $wealth_i$ is negatively significant, suggesting that wealthy respondents are more likely to forget reporting the transfers they have made. In the $Pr(R = 1|\tau = 1)$ regression, the coefficient of $wealth_j$ is negative as well, although not statistically significant. The excluded variables *declared* $friends$ are significantly positive in both equations as expected, indicating that households who cite more partners at baseline also better recall the transfers made and received in subsequent survey rounds.

Table 3. Main results

	(1) τ_{ij}^{max}	(2) τ_{ij}^{min}	(3) $\Pr(\tau = 1)$	(4) $\Pr(G = 1\|\tau = 1)$	(5) $Pr(R = 1\|\tau = 1)$
$wealth_i$	0.062***	0.057***	0.045	-0.053*	0.055
	(0.021)	(0.019)	(0.051)	(0.028)	(0.079)
$wealth_j$	0.096***	0.051**	0.062	0.084	-0.058
	(0.030)	(0.026)	(0.041)	(0.060)	(0.045)
$wealth_i * wealth_j$	0.004	0.002	0.013**	-0.001	-0.003
	(0.003)	(0.003)	(0.006)	(0.003)	(0.006)
$same\, education$	-0.012	0.060	-0.052	0.173	-0.143
	(0.118)	(0.177)	(0.306)	(0.359)	(0.282)
$same\, religion$	0.434***	0.464***	0.367	0.212	0.216
	(0.099)	(0.145)	(0.282)	(0.296)	(0.273)
$blood\, link$	2.718***	2.627***	2.631***	1.003**	1.321***
	(0.252)	(0.246)	(0.601)	(0.459)	(0.354)
$neighbors$	1.063***	1.503***	0.683*	0.891***	0.674**
	(0.111)	(0.157)	(0.350)	(0.283)	(0.264)
$declared\ friends_i$				0.086***	
				(0.026)	
$declared\ friends_j$					0.052*
					(0.029)
$constant$	-3.510***	-5.120***	-2.541***	-1.656**	-1.389***
	(0.210)	(0.213)	(0.647)	(0.647)	(0.518)

*** p<0.01, ** p<0.05, * p<0.1. Dyadic-robust standard errors in parentheses.

To check the robustness of these findings, we re-estimate the model with different sets of regressors. For columns (3) to (5), convergence to a stable set of coefficient estimates is smooth for a reasonably sized set of regressors. Estimated coefficients are very similar across specifications. Including significant regressors in the mis-reporting equations increases the difference between standard logit results in columns (1) and (2) and the maximum likelihood results in column (3). These findings are consistent with our simulation results and reconfirm that our estimator represents an improvement over logit if we include relevant variables in the mis-reporting equations.

5 Estimates of under-reporting

From the raw figures reported in Subsection 3.2 is it possible to compute method-of-moments (MM) estimates of under-reporting, before introducing covariates. Assuming independence in reporting probability between i and j, we use the following three equations to fit the three unconditional probabilities $\Pr(\tau = 1)$, $\Pr(G = 1|\tau = 1)$, and $\Pr(R = 0|\tau = 1)$ to the data frequencies reported in Subsection 3.2:

$$\Pr(G = 1, R = 0) = \Pr(\tau = 1) * \Pr(G = 1|\tau = 1) * \Pr(R = 0|\tau = 1) = 0.0548 \quad (11)$$

$$\Pr(G = 0, R = 1) = \Pr(\tau = 1) * \Pr(G = 0|\tau = 1) * \Pr(R = 1|\tau = 1) = 0.0343 \quad (12)$$

$$\Pr(G = 1, R = 1) = \Pr(\tau = 1) * \Pr(G = 1|\tau = 1) * \Pr(R = 1|\tau = 1) = 0.0335 \quad (13)$$

Straightforward algebra yields the solutions reported in Table 4 below:

Table 4. MM estimates of under-reporting

in data: declared by i	0.09
in data: declared by j	0.07
in data: declared by i or j $\left(\tau_{ij}^{max}\right)$	0.12
in data: declared by i and j $\left(\tau_{ij}^{min}\right)$	0.03
$\Pr(\tau_{ij} = 1)$	0.18
$\Pr(G = 1|\tau = 1)$	0.49
$\Pr(R = 1|\tau = 1)$	0.38

If we compare these estimates to the reported transfers presented in the upper part of the table, we see that not taking mis-reporting into consideration leads to serious underestimation of transfers between villagers. The simple calculation above suggests that $\tau_{ij}^{max} = 12\%$ only captures two thirds of the transfers estimated to be made.

We can obtain similar estimates from the maximum likelihood model formed by equations (2) to (8). The only difference is that these estimates are conditional on covariates, a feature that allows for correlation in reporting propensities based on observables. The result of these calculations is reported in Table 5.

Table 5. Estimates of under-reporting with covariates

	gifts
average fitted $\Pr(\tau_{ij} = 1)$	0.20
average fitted $\Pr(G = 1 \mid \tau = 1)$	0.38
average fitted $\Pr(R = 1 \mid \tau = 1)$	0.30

The average fitted propensity to give from Table 5 is 20%, very close to the figure of 18% obtained without conditioning on covariates. The average fitted propensity to report a gift is 38% for the giver and 30% for the receiver, smaller than the figures of Table 4. If anything, estimated propensities to report gifts and loans fall when we allow them to depend on household observables.

The Nyakatoke data were collected with an unusually high level of care, using multiple survey rounds and interviewing each household member separately. Yet results suggests massive under-reporting. This casts some doubt on the general reliability of self-reported data on transfers of money, goods, and favors. This matters for our understanding of the importance of favor exchange. Many studies have found that reported gifts and loans are insufficient to insulate households against shocks. But if actual gifts and loans are much larger, these findings might be called into question. For instance, Rosenzweig (1988) reports that loans between households represent only 2% of the value of the shocks they face. If there is as much loan under-reporting in his data as in ours, the corrected figure is closer to 5%.[19]

In Appendix C we investigate in detail the robustness of our estimates to the assumption that unexplained variation in under-reporting by i is independent of that in under-reporting by j. We calculate estimates of $\Pr(\tau_{ij} = 1)$ for different possible values of the correlation in under-reporting between i and j. We show that extremely high or low correlation values are irreconciliable with the data: high positive correlation would imply little discordance, which is not what the data show; and high negative correlation would imply even more discordance than what is in the data. There is a range of intermediate correlation values which are potentially consistent with the data. This range is within the range of correlation values for which simulation analysis has shown that correcting for mis-reporting improves inference. To each of the feasible correlation

[19]This estimate is obtained by multiplying the loans reported in the Rosenzweig data (2%) by a correction factor equal to (predicted transfers estimated in our model)/(transfers declared by the giver in our model), that is, by (20.2%) / (8.8%).

values corresponds an estimated value of $\Pr(\tau_{ij} = 1)$. This is summarized in Figure 4 (Appendix C), which shows that feasible estimates of $\Pr(\tau_{ij} = 1)$ vary between 13% and 27%.

6 Conclusions

Self-reported transfer data are typically discordant: i may report a transfer to j while j reports no such transfer from i. In this paper we propose a maximum likelihood estimator to deal with mis-reporting of this kind. Using simulations, we show that the consequences of neglecting mis-reporting may be severe when determinants of transfers are correlated with the propensity to report a transfer given or received. Our estimator corrects for this bias by conditioning reporting on such determinants.

We illustrate the methodology using dyadic data on inter-household transfers from the village of Nyakatoke in Tanzania, where we observe substantial discrepancies between amounts reported by givers and receivers. In particular, we combine data about flows and a proxy of desire to link to investigate whether observed transfers are in the mutual self-interest of both households involved. We find reasonably convincing evidence that the exchange of transfers is best predicted by mutual self-interest.

We provide evidence of sizable under-reporting of transfers, in spite of the care that was applied in collecting the data. This finding is hardly surprising given that reports of transfers between households are often discordant, with one household reporting it while the other does not. We also provide alternative estimates of under-reporting that allow for correlation (positive or negative) in reporting probabilities across household pairs.

The methodology presented here has potential applications in other fields as well. Gravity models are a good example of a possible application. They have long been estimated in the trade literature. Our methodology could prove useful when there are discrepancies in trade flow data reported by different countries or different sources. The model presented in the body of this paper applies if researchers suspect data may be under-reported, while the variation presented in Appendix B is applicable when over-reporting is suspected instead.[20]

The method is also applicable to non-dyadic data when the researcher has conflict-

[20]The standard error correction from Fafchamps and Gubert (2007) should also be used for gravity models, with or without correction for mis-reporting, in order to compensate for the downward bias in reported standard errors that is common to all dyadic regressions.

ing measurements of the same dependent variable from different sources. For instance, the method could be useful to deal with answers to questions about household expenditures answered by both husband and wife, or to questions about worker performance questions answered by both employer and employee, etc.[21]

References

[1] Akee, Randall and Devesh Kapur (2012), 'Remittances and Rashomon', Center for Global Development, Working Paper 285, January

[2] Arcand, Jean-Louis and Marcel Fafchamps (2012), 'Matching in Community-Based Organizations', *Journal of Development Economics*, 98 (2): 203 - 219

[3] Ashenfelter, Orley and Alan Krueger (1994), 'Estimates of the Economic Return to Schooling From a New Sample of Twins', *American Economic Review*, 84: 1157-73

[4] Attanasio, Orazio, Abigail Barr, Juan Camilo Cárdenas, Garance Genicot and Costas Meghir (2012) 'Risk Pooling, Risk Preferences and Social Networks' *American Economic Journal. Applied Economics*, 4(2): 134–67.

[5] Ball, Brian and M.E.J. Newman (2013), 'Friendship networks and social status', *Network Science*, 1(01): 16-30.

[6] Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo and Matthew O. Jackson (2013), 'The Diffusion of Microfinance', *Science*, Vol. 341.

[7] Belot, Michele and Marco Francesconi (2012), 'Dating Preferences and Meeting Opportunities in Mate Choice Decisions', *Journal of Human Resources*, 48(2), 474-507

[8] Christakis, Nicholas A. and James H. Fowler (2009), *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*, Little, Brown and Company, London

[9] Coate, Stephen and Martin Ravallion (1993), 'Reciprocity Without Commitment: Characterization and Performance of Informal Insurance Arrangements', *Journal of Development Economics*, 40: 1-24

[21] In these cases, the dyadic correction of standard errors would not be necessary.

[10] Comola, Margherita (2012), 'Estimating Local Network Externalities', SSRN Working Paper n. 946093

[11] Comola, Margherita and Marcel Fafchamps (2014), 'Testing Unilateral and Bilateral Link Formation', *The Economic Journal,* 124: 954-976.

[12] De Weerdt, Joachim and Stefan Dercon (2006), 'Risk-Sharing Networks and Insurance Against Illness', *Journal of Development Economics*, 81(2): 337-56

[13] De Weerdt, Joachim and Marcel Fafchamps (2011), 'Social Identity and The Formation of Health Insurance Networks', *Journal of Development Studies*, 47(8): 1152-1177.

[14] Dai, Bin, Shilin Ding and Grace Wahba (2012). "Multivariate Bernoulli Distribution", Department of Statistics, University of Wisconsin at Madison, Technical Report No. 1170

[15] Duncan, Greg and Daniel Hill (1985), 'An Investigation of the Extent and Consequences of Measurement Error in Labor Economic Survey Data', *Journal of Labor Economics* 3: 508-522

[16] Fafchamps, Marcel and Susan Lund (2003), 'Risk Sharing Networks in Rural Philippines', *Journal of Development Economics*, 71: 261-87

[17] Fafchamps, Marcel and Flore Gubert (2007), 'The Formation of Risk Sharing Networks', *Journal of Development Economics*, 83(2): 326-50

[18] Foster, Andrew D. and Mark R. Rosenzweig (2001), 'Imperfect Commitment, Altruism and the Family: Evidence from Transfer Behavior in Low-Income Rural Areas', *Review of Economics and Statistics*, 83(3): 389-407)

[19] Gaulier Guillaume and Soledad Zignago (2010) 'BACI: International Trade Database at the Product-Level. The 1994-2007 Version,' CEPII Working Paper 2010-23 , CEPII.

[20] Glaeser, Edward, Bruce Sacerdote,and Jose Scheinkman (1996), "Crime and Social Interactions", *Quarterly Journal of Economics*, 111: 507-48, 1996

[21] Hitsch, Gunter J., Ali Hortacsu, Dan Ariely (2010), 'Matching and Sorting in Online Dating', *American Economic Review*, 100(1): 130-63.

[22] Hong, Han and Elie Tamer (2003), "A simple estimator for nonlinear error in variable models," *Journal of Econometrics*, 117(1): 1-19

[23] Kocherlakota, Narayana R. (1996), 'Implications of Efficient Risk Sharing Without Commitment', *Review of Economic Studies*, 63(4): 595-609

[24] Lee, Yun-Suk and Linda J. Waite (2005), 'Husbands and Wives Time Spent on Housework: A Comparison of Measures', *Journal of Marriage and Family*, 67: 328-336

[25] Ligon, Ethan, Jonathan P. Thomas, and Tim Worrall (2001), 'Informal Insurance Arrangements in Village Economies', *Review of Economic Studies*, 69(1): 209-44

[26] Liu, Xiaodong, Eleonora Patacchini, Yves Zenou, and Lung-Fei Lee (2011), "Criminal Networks: Who is the Key Player?", Research Papers in Economics 2011:7, Stockholm University, Department of Economics

[27] Rosenzweig, Mark R. (1988), "Risk, Implicit Contracts and the Family in Rural Areas of Low-Income Countries," *Economic Journal*, 98: 1148-1170, December

[28] Roth, Alvin and Marilda Sotomayor (1990), *Two-Sided Matching*, Cambridge University Press, Cambridge

[29] Schennach, Susanne (2004), "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72 (1): 33-75

[30] Snijders, Tom A.B., Johan Koskinen, and Michael Schweinberger (2010), "Maximum Likelihood Estimation for Social Network Dynamics", *Annals of Applied Statistics*, 4 (2): 567-588

[31] Steglich, Christian E.G., Tom A.B. Snijders, and Michael Pearson (2010), 'Dynamic Networks and Behavior: Separating Selection from Influence', *Sociological Methodology*, 40 (1): 329-393

[32] Udry, Christopher (1994), 'Risk and Insurance in a Rural Credit Market: An Empirical Investigation in Northern Nigeria', *Review of Economic Studies*, 61(3): 495-526

[33] Vandenbossche, Joost and Thomas Demuynck (2013), 'Network Formation with Heterogeneous Agents and Absolute Friction', *Computational Economics*, 42 (1): 23-45.

Appendix A

Figure 1: Kernel densities for $\beta_{\tau 1}$ under scenario 4

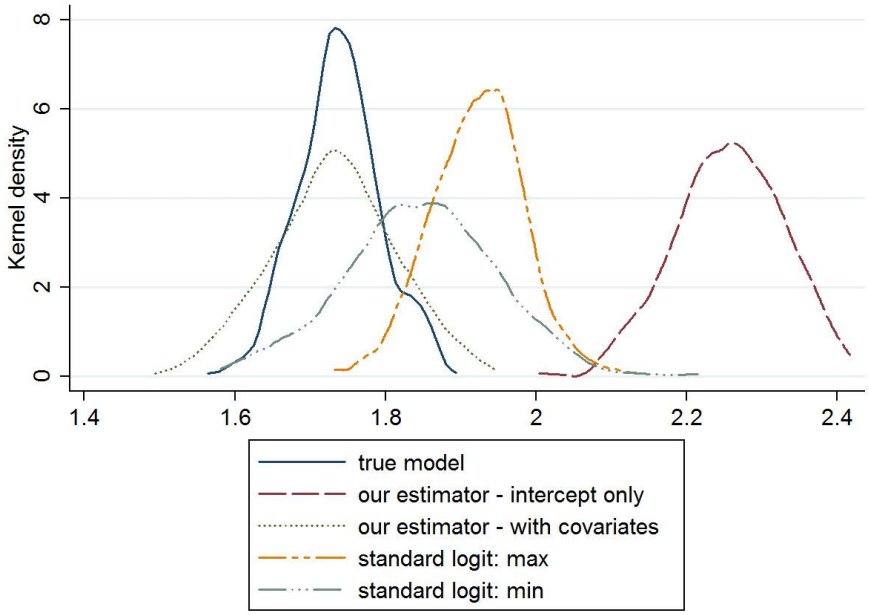


Figure 2: Kernel densities for $\beta_{\tau 2}$ under scenario 4

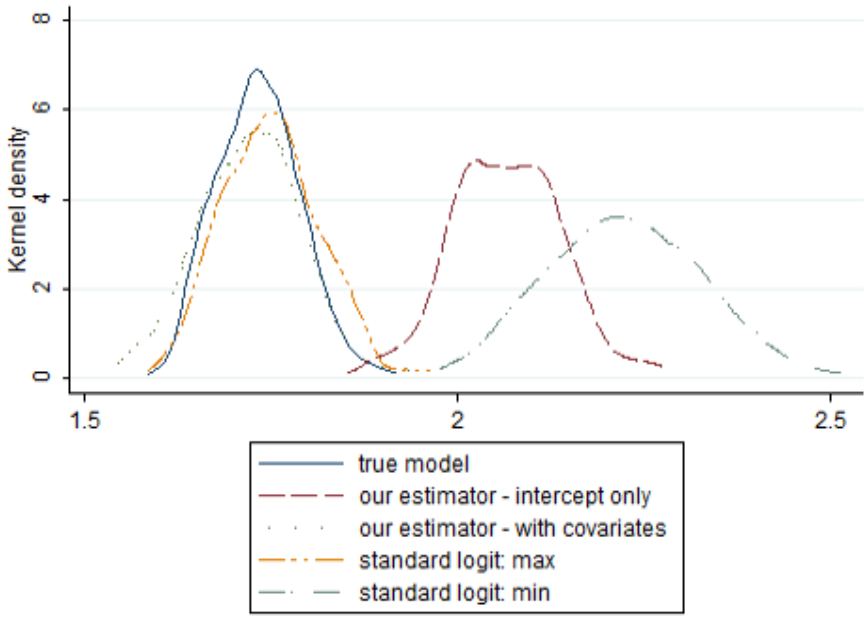


Figure 3: Kernel densities for $\beta_{\tau 3}$ under scenario 4

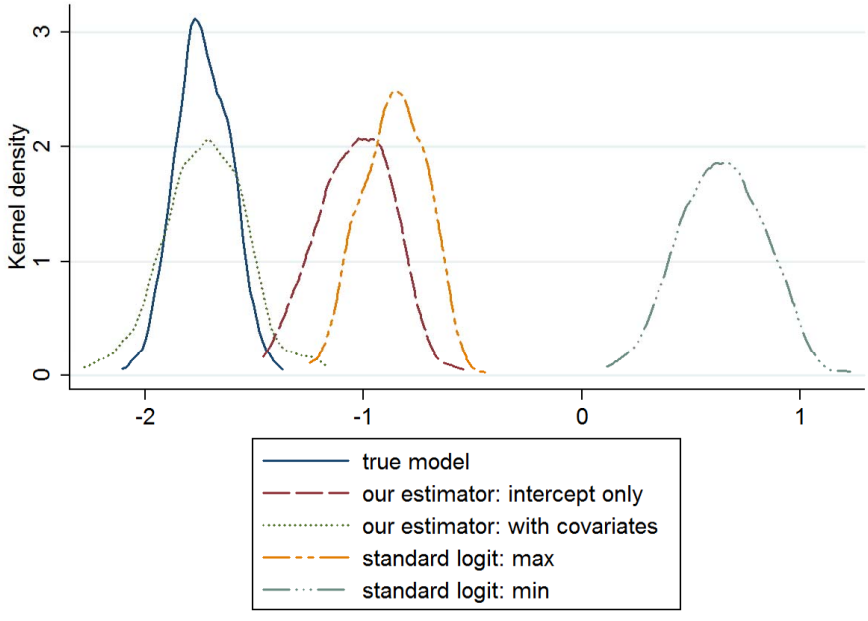


Table A1. Simulation results under correlation: $\rho = 0.1$

	(1)	(2)	(3)	(4)	(5)
	true model	our estimator	our estimator	standard logit	standard logit
	τ_{ij}	intercept only	with covariates	τ_{ij}^{max}	τ_{ij}^{min}
	Scenario 1:				
$\beta_{\tau 1}$	1.73	1.69	1.69	1.47	1.14
$\beta_{\tau 2}$	1.73	1.69	1.7	1.47	1.16
$\beta_{\tau 3}$	-1.74	-1.7	-1.72	-1.47	-1.12
	Scenario 2:				
$\beta_{\tau 1}$	1.73	2.24	1.68	1.91	1.78
$\beta_{\tau 2}$	1.73	2.05	1.67	1.74	2.19
$\beta_{\tau 3}$	-1.75	-1.76	-1.66	-1.48	-0.95
	Scenario 3:				
$\beta_{\tau 1}$	1.73	1.67	1.71	1.46	1.19
$\beta_{\tau 2}$	1.74	1.67	1.71	1.47	1.2
$\beta_{\tau 3}$	-1.74	-0.96	-1.64	-0.8	0.41
	Scenario 4:				
$\beta_{\tau 1}$	1.73	2.2	1.69	1.9	1.8
$\beta_{\tau 2}$	1.73	2.01	1.68	1.73	2.18
$\beta_{\tau 3}$	-1.74	-1	-1.58	-0.84	0.58

Table A2. Simulation results under correlation: $\rho = 0.3$

	(1)	(2)	(3)	(4)	(5)
	true model	our estimator	our estimator	standard logit	standard logit
	τ_{ij}	intercept only	with covariates	τ_{ij}^{max}	τ_{ij}^{min}
	Scenario 1:				
$\beta_{\tau 1}$	1.73	1.57	1.58	1.44	1.17
$\beta_{\tau 2}$	1.73	1.58	1.58	1.44	1.17
$\beta_{\tau 3}$	-1.75	-1.58	-1.59	-1.44	-1.16
	Scenario 2:				
$\beta_{\tau 1}$	1.73	2.14	1.66	1.9	1.74
$\beta_{\tau 2}$	1.74	1.93	1.61	1.71	2.16
$\beta_{\tau 3}$	-1.74	-1.67	-1.54	-1.46	-1.02
	Scenario 3:				
$\beta_{\tau 1}$	1.73	1.56	1.6	1.44	1.19
$\beta_{\tau 2}$	1.73	1.56	1.59	1.44	1.2
$\beta_{\tau 3}$	-1.73	-0.84	-1.3	-0.74	0.28
	Scenario 4:				
$\beta_{\tau 1}$	1.73	2.11	1.67	1.9	1.75
$\beta_{\tau 2}$	1.73	1.89	1.61	1.69	2.17
$\beta_{\tau 3}$	-1.73	-0.91	-1.29	-0.8	0.43

Table A3. Simulation results under correlation: $\rho = 0.5$

	(1)	(2)	(3)	(4)	(5)
	true model	our estimator	our estimator	standard logit	standard logit
	τ_{ij}	intercept only	with covariates	τ_{ij}^{max}	τ_{ij}^{min}
	Scenario 1:				
$\beta_{\tau 1}$	1.73	1.49	1.5	1.42	1.17
$\beta_{\tau 2}$	1.73	1.49	1.49	1.41	1.18
$\beta_{\tau 3}$	-1.73	-1.47	-1.49	-1.39	-1.13
$\beta_{\tau 1}$	1.73	2.05	1.65	1.89	1.68
$\beta_{\tau 2}$	1.73	1.8	1.54	1.65	2.14
$\beta_{\tau 3}$	-1.74	-1.57	-1.46	-1.42	-1.02
$\beta_{\tau 1}$	1.74	1.5	1.52	1.43	1.22
$\beta_{\tau 2}$	1.73	1.5	1.51	1.43	1.22
$\beta_{\tau 3}$	-1.74	-0.75	-1.03	-0.7	0.14
$\beta_{\tau 1}$	1.72	2.04	1.66	1.89	1.69
$\beta_{\tau 2}$	1.73	1.78	1.55	1.64	2.16
$\beta_{\tau 3}$	-1.73	-0.84	-1.07	-0.76	0.3

Table A4. Simulation results under correlation: $\rho = -0.1$

	(1)	(2)	(3)	(4)	(5)
	true model	our estimator	our estimator	standard logit	standard logit
	τ_{ij}	intercept only	with covariates	τ_{ij}^{max}	τ_{ij}^{min}
	Scenario 1:				
$\beta_{\tau 1}$	1.73	1.82	1.83	1.49	1.13
$\beta_{\tau 2}$	1.73	1.83	1.83	1.49	1.13
$\beta_{\tau 3}$	-1.74	-1.84	-1.83	-1.49	-1.12
	Scenario 2:				
$\beta_{\tau 1}$	1.73	2.35	1.75	1.93	1.88
$\beta_{\tau 2}$	1.73	2.17	1.76	1.77	2.23
$\beta_{\tau 3}$	-1.73	-1.85	-1.78	-1.5	-0.93
	Scenario 3:				
$\beta_{\tau 1}$	1.73	1.79	1.82	1.49	1.19
$\beta_{\tau 2}$	1.73	1.79	1.82	1.49	1.17
$\beta_{\tau 3}$	-1.74	-1.08	-1.97	-0.83	0.62
	Scenario 4:				
$\beta_{\tau 1}$	1.73	2.29	1.74	1.92	1.88
$\beta_{\tau 2}$	1.73	2.12	1.76	1.76	2.25
$\beta_{\tau 3}$	-1.75	-1.1	-1.89	-0.9	0.7

Table A5. Simulation results under correlation: $\rho = -0.3$

	(1)	(2)	(3)	(4)	(5)
	true model	our estimator	our estimator	standard logit	standard logit
	τ_{ij}	intercept only	with covariates	τ_{ij}^{max}	τ_{ij}^{min}
	Scenario 1:				
$\beta_{\tau 1}$	1.73	2	1.98	1.52	1.11
$\beta_{\tau 2}$	1.74	2	2	1.53	1.12
$\beta_{\tau 3}$	-1.74	-2.01	-1.94	-1.52	-1.09
	Scenario 2:				
$\beta_{\tau 1}$	1.73	2.47	1.84	1.94	1.96
$\beta_{\tau 2}$	1.73	2.32	1.89	1.81	2.31
$\beta_{\tau 3}$	-1.73	-1.97	-1.93	-1.54	-0.91
	Scenario 3:				
$\beta_{\tau 1}$	1.73	1.94	1.94	1.51	1.17
$\beta_{\tau 2}$	1.73	1.94	1.96	1.51	1.17
$\beta_{\tau 3}$	-1.73	-1.22	-2.29	-0.87	0.85
	Scenario 4:				
$\beta_{\tau 1}$	1.73	2.42	1.81	1.93	1.97
$\beta_{\tau 2}$	1.73	2.26	1.87	1.79	2.29
$\beta_{\tau 3}$	-1.73	-1.16	-2.22	-0.91	0.97

Table A6. Simulation results under correlation: $\rho = -0.5$

	(1)	(2)	(3)	(4)	(5)
	true model	our estimator	our estimator	standard logit	standard logit
	τ_{ij}	intercept only	with covariates	τ_{ij}^{max}	τ_{ij}^{min}
	Scenario 1:				
$\beta_{\tau 1}$	1.74	2.19	2.15	1.55	1.1
$\beta_{\tau 2}$	1.73	2.18	2.18	1.55	1.1
$\beta_{\tau 3}$	-1.74	-2.19	-2	-1.55	-1.06
	Scenario 2:				
$\beta_{\tau 1}$	1.73	2.62	1.98	1.96	2.09
$\beta_{\tau 2}$	1.73	2.48	2.08	1.85	2.42
$\beta_{\tau 3}$	-1.74	-2.09	-2.03	-1.57	-0.88
	Scenario 3:				
$\beta_{\tau 1}$	1.73	2.14	2.08	1.54	1.17
$\beta_{\tau 2}$	1.73	2.13	2.11	1.54	1.18
$\beta_{\tau 3}$	-1.74	-1.44	-2.68	-0.94	1.16
	Scenario 4:				
$\beta_{\tau 1}$	1.73	2.55	1.9	1.96	2.11
$\beta_{\tau 2}$	1.73	2.4	1.99	1.84	2.43
$\beta_{\tau 3}$	-1.73	-1.26	-2.54	-0.97	1.25

Table A7. Quintiles of declared transfers

Information given by:	giver	receiver
nonzero obs.	1240	952
cut-off values:		
0-20%	272	250
20-40%	600	500
40-60%	1100	1000
60-80%	2150	2370
80-100%	60150	47750

Note: the total sample size is 14042 dyads. Cut-off values computed on nonzero observations only. Values expressed in *tzs*.

Appendix B

In this appendix we explain how our estimator can be implemented when researchers suspect that transfers are over-estimated instead of under-estimated, *i.e.*, when respondents may report transfers that did not actually take place. In the context of our data, this could arise if people wish they had made these transfers but were ashamed to admit to enumerators that they did not, and so made up some numbers. Whether or not this is a reasonable assumption depends on the context - for our data, it is rather unlikely. It should be noted that in our data few household pairs have declared a transfer from both sides (3.4% of dyads). This means that, under the assumption of over-reporting, the number of observations for which $\tau = 1$ is small. It is nevertheless instructive to illustrate the procedure.

We now assume that unless both i and j declare a transfer, it did not take place. As long as recall errors are not perfectly negatively correlated, it is also possible that a transfer did not take place even if both i and j declare it. As before, let us assume that recall errors are independent between i and j. We can write:

$$\begin{aligned}
\Pr(G=1, R=0) &= \Pr(\tau=0) * \Pr(G=1|\tau=0) * \Pr(R=0|\tau=0) && (14)\\
\Pr(G=0, R=1) &= \Pr(\tau=0) * \Pr(G=0|\tau=0) * \Pr(R=1|\tau=0) && (15)\\
\Pr(G=0, R=0) &= \Pr(\tau=0) * \Pr(G=0|\tau=0) * \Pr(R=0|\tau=0) && (16)\\
\Pr(G=1, R=1) &= 1 - \Pr(G=1, R=0) - \Pr(G=0, R=1) - \Pr(G=0, R=0) && (17)
\end{aligned}$$

Equations (14) to (17) express the data generating process in terms of three probabilities: $P(\tau = 0)$, $P(G = 1|\tau = 0)$ and $P(R = 1|\tau = 0)$. As before, we assume that these three probabilities can be represented by three distinct logit functions $\lambda(.)$ as follows:

$$\begin{aligned}
\Pr(\tau=0) &= \lambda(\beta'_\tau X_\tau) && (18)\\
\Pr(G=1|\tau=0) &= \lambda_G(\beta'_G X_G) && (19)\\
\Pr(R=1|\tau=0) &= \lambda_R(\beta'_R X_R) && (20)
\end{aligned}$$

The main equation of interest now is $\Pr(\tau = 0)$. Define $h_{ij} = 1$ if and only if $\tau_{ij} = 0$, *i.e.*, h_{ij} is an indicator variable that takes value 1 is i does *not* give to j. We estimate

a model of the form:

$$Pr(h_{ij} = 1) = \lambda(\theta'_\tau X^{ij}_\tau) \tag{21}$$

Equations (19) and (20) can be similarly transformed. The resulting likelihood function is equivalent to equations (6) to (8), but expressed in terms of h_{ij} instead of τ_{ij}.

Table B1 reports the estimated frequency of giving and lending under the assumption of over-reporting. The estimated probabilities of reporting a transfer which did not take place range from 6.4% (for the giver) to 4.3% (for the receiver). These probabilities are very low (especially when compared to the figures of 38% and 30% reported in Table 5 under the alternative assumption of under-reporting) and accordingly the average fitted $\Pr(\tau_{ij} = 1)$ is close to the share of transfers declared by both i and j from Table 4. Since over-reporting is estimated to be small, unsurprisingly the estimated coefficients for $Pr(h_{ij} = 1)$ are in this case close to the coefficients of the standard logit regression where the dependent variable is τ^{min}_{ij} and all discordances are imputed to over-reporting (column (2) in Table 3), and are not reported here to save on space.

Table B1. Estimates of over-reporting

	transfer
average fitted $\Pr(\tau_{ij} = 1) = \Pr(h_{ij} = 0)$	0.028
average fitted $\Pr(G = 1 \mid \tau = 0)$	0.064
average fitted $\Pr(R = 1 \mid \tau = 0)$	0.043

Appendix C

In this appendix, we illustrate how it is possible to check the robustness of the mis-reporting estimates to alternative assumptions regarding the correlation between R and G. We focus on the unconditional case, which is sufficient for our purpose. We start by introducing a simplified notation:

$$\begin{aligned}
\Pr(R=0,G=0|\tau=1) &\equiv p_{00} && (22)\\
\Pr(R=1,G=0|\tau=1) &\equiv p_{10} && (23)\\
\Pr(R=0,G=1|\tau=1) &\equiv p_{01} && (24)\\
\Pr(R=1,G=1|\tau=1) &\equiv p_{11} && (25)\\
\Pr(\tau=1) &\equiv \lambda && (26)\\
\Pr(R=1|\tau=1) &\equiv \rho = p_{10}+p_{11} && (27)\\
\Pr(G=1|\tau=1) &\equiv \gamma = p_{01}+p_{11} && (28)
\end{aligned}$$

The joint distribution of R and G conditional on $\tau = 1$ is a Bernoulli distribution of the form (Dai, Ding and Wahba 2012):

$$\Pr(R=r,G=g|\tau=1) = \exp\{\log(p_{00}) + rf_1 + gf_2 + rgf_{12}\}$$

where

$$\begin{aligned}
f_1 &= \log(p_{10}) - \log(p_{00})\\
f_2 &= \log(p_{01}) - \log(p_{00})\\
f_{12} &= \log(p_{11}p_{00}) - \log(p_{10}p_{01})
\end{aligned}$$

The covariance between R and G is given by:

$$cov(R,G|\tau=1) = p_{11}p_{00} - p_{10}p_{01} \quad (29)$$

Note that $cov(R,G|\tau=1) \subset (-1,1)$ since all the elements are probabilities. The correlation coefficient between R and G is:

$$corr(R,G|\tau=1) = \frac{p_{11}p_{00} - p_{10}p_{01}}{\sqrt{\rho(1-\rho)\gamma(1-\gamma)}} \quad (30)$$

What we observe are sample moments of the following probabilities:

$$\begin{aligned}
\Pr(R &= 1, G = 0) \equiv m_{10} = \lambda p_{10} \\
\Pr(R &= 0, G = 1) \equiv m_{01} = \lambda p_{01} \\
\Pr(R &= 1, G = 1) \equiv m_{11} = \lambda p_{11} \\
\Pr(R &= 0, G = 0) \equiv m_{00} = \lambda p_{00} + (1 - \lambda)
\end{aligned}$$

which implies:

$$p_{10} = \frac{m_{10}}{\lambda} \tag{31}$$

$$p_{01} = \frac{m_{01}}{\lambda} \tag{32}$$

$$p_{11} = \frac{m_{11}}{\lambda} \tag{33}$$

$$p_{00} = \frac{m_{00} - (1 - \lambda)}{\lambda} \tag{34}$$

Equations (31) to (34) contain four quantities that are potentially observable – m_{10}, m_{01}, m_{11}, and m_{00} – and five unknown parameters – p_{10}, p_{01}, p_{11}, p_{00} and λ. It is immediately apparent that it is impossible to estimate all five unknown parameters from the four observable moments. To circumvent this difficulty, we have so far assumed that R and G are independent and thus that $cov(R, G|\tau = 1) = 0$. We now generalize this approach and assume that $cov(R, G|\tau = 1)$ takes some arbitrary value C between -1 and 1. Given this value, it is possible to obtain estimates of λ, p_{10}, p_{01}, p_{11}, p_{00}, ρ and γ as follows:

$$\begin{aligned}
p_{11}p_{00} - p_{10}p_{01} &= C \text{ from (29)} \\
p_{00} &= \frac{p_{10}p_{01} + C}{p_{11}} \\
\frac{m_{00} - (1 - \lambda)}{\lambda} &= \frac{\frac{m_{10}}{\lambda}\frac{m_{01}}{\lambda} + C}{\frac{m_{11}}{\lambda}} \text{ using (31-34)} \\
m_{00} - (1 - \lambda) &= \frac{m_{10}m_{01} + \lambda^2 C}{m_{11}} \\
\lambda^2 C - \lambda m_{11} + m_{10}m_{01} - m_{00}m_{11} + m_{11} &= 0
\end{aligned}$$

which yields a second order polynomial in λ. Solving this polynomial for λ using sample moments for m_{10} etc yields a method-of-moments estimator of λ. We thus have two

roots:

$$\begin{aligned} \widehat{\lambda} &= \frac{m_{11} \pm \sqrt{m_{11}^2 - 4aC}}{2C} \text{ with} \\ a &\equiv m_{10}m_{01} - m_{00}m_{11} + m_{11} \end{aligned}$$

Experimentation reveals that the meaningful root is the negative one. In the special case where $C = 0$ the polynomial simplifies to a linear equation, the solution of which is:

$$\widehat{\lambda} = \frac{m_{10}m_{01} - m_{00}m_{11} + m_{11}}{m_{11}}$$

Once we have an estimate of $\widehat{\lambda}$, we can derive estimates of $\widehat{p}_{10}, \widehat{p}_{01}, \widehat{p}_{11}, \widehat{p}_{00}$ using using (31-34). We can also estimate:

$$\begin{aligned} \widehat{\rho} &= \widehat{p}_{10} + \widehat{p}_{11} \\ \widehat{\gamma} &= \widehat{p}_{01} + \widehat{p}_{11} \end{aligned}$$

and use (30) to estimate the correlation between R and G that is implied by the value of C, given the data. This correlation should be between -1 and 1.

It is important to recognize that the choice of a value for C may yield non-sensical estimates, that is, estimates that are not within normal bounds: for the implied value of $\widehat{\lambda}$ some probabilities $\widehat{p}_{10}$, $\widehat{p}_{01}$, $\widehat{p}_{11}$, $\widehat{p}_{00}$, $\widehat{\rho}$ or $\widehat{\gamma}$ may turn out to be negative or above 1.[22] This occurence implies that there is no way of reconciling the assumed covariance C with the data.

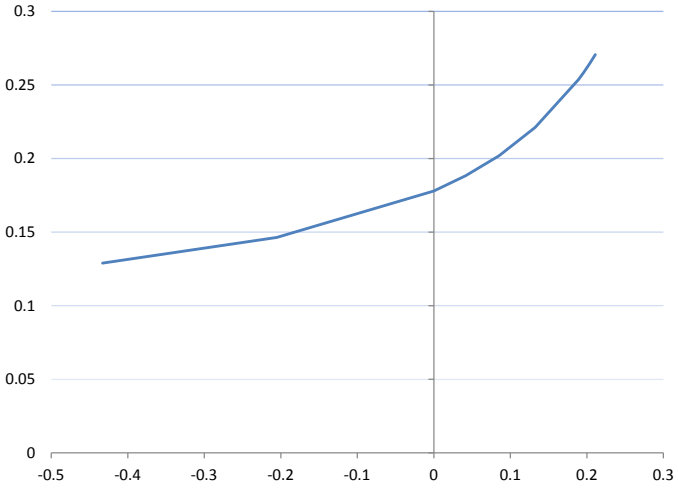
To see how this can arise, suppose we assume that R and G are highly correlated. If this is true, we should observe very small values of m_{10} and m_{01} relative to m_{11}. Suppose this is not true in the data. In this case, forcing the data to fit a data generating process that posits a high covariance can only result in non-sensical results. Similarly, suppose that we posit a low negative value for C, implying that R and G are negative correlated. In this case, we expect m_{10} and m_{01} to be large relative to m_{11}: here, if the recipient reports the transfer, the giver does not report it, and vice versa. Suppose that in fact m_{11} is not small relative to m_{10} and m_{01}. Again, forcing the data into this data generating process will result in contradiction, that is, non-sensical probability estimates and the like.

Using this approach, it is therefore possible to bracket the values of λ that can be

[22]Similarly, the correlation coefficient between R and G may be below -1 or above $+1$.

reconciled with the data. We report in Figure 4 below the different values of λ, that is, the estimates of $\Pr(\tau = 1)$ that correspond to different values of the correlation between R and G for our data.[23] We note that correlation values less than -0.44 and large than 0.21 result in non-sensical probabilities. The range of possible values of $corr(R, G|\tau = 1)$ is thus bounded by the data. We also note that estimates of $\widehat{\lambda}$ do not vary too much over the range of possible correlation values (between 13 and 27%).

Figure 4: $\Pr(\tau = 1)$



[23] In Figure 4 we opted to show $corr(R, G|\tau = 1)$ on the x-axis instead of C to improve readability.