# A Pairwise Differencing Distribution Regression Approach for Network Models

Gabriela M. Miyazato Szini

*Tilburg University*

*Preliminary, and updated frequently. Click here for the most recent version.*

**Abstract**

A novel estimation method for distribution regressions in a network setting is proposed. It considers the effects of covariates on the entire outcome distribution rather than solely on the mean. I adopt a semiparametric approach by considering two-way unit-specific effects. Thus, I extend the standard distribution regression approach to a network setting by estimating multiple binary choice models with two-way fixed effects for different thresholds of the distribution. I employ a conditional maximum-likelihood approach that differences out the unit-specific effects, avoiding the incidental parameter problem. This method yields consistent point estimates that converge at a parametric rate and remain asymptotically unbiased in the tails of the outcome distribution, where the underlying network can be seen as sparse. Monte Carlo simulations validate these findings for single cut-off points and the overall outcome distribution. The empirical application focuses on gravity equations for bilateral trade, demonstrating the effectiveness of the proposed approach in cases where the outcome variable is bounded below at zero.

⋆I would like to thank my advisors Frank Kleibergen and Artūras Juodis for all comments and suggestions. Moreover, I am grateful for comments from Bo Honoré as well as Martin Weidner, Chris Muris, Michal Kolesár, Cavit Pakel, Mikkel Plagborg-Møller, Elena Manresa, Bryan Graham, Sebastian Roelsgaard, Timo Schenk, Pavel Cizek, Otilia Boldea, Geert Dhaene, Áureo de Paula, Stephane Bonhomme, Jonas Meier, Mark Watson, Laura Liu, Wendun Wang, Stephen Redding, Luther Yap, Andrea Titton, Lina Zhang, Andrei Zeleneev, Sami Stouli and seminar participants at Princeton University, TI PhD Job Market Jamboree, Netherlands Econometric Study Group, the International Panel Data Conference, and the Bristol Econometric Study Group.

*Email address:* g.m.miyazatoszini@tilburguniversity.edu. (Gabriela M. Miyazato Szini)

## 1. Introduction

The vast majority of studies, especially for network models, propose estimates for the effects of covariates on the mean of an outcome variable. However, in many cases, the effects on the entire distribution of the outcomes are also an object of interest. For instance, in applications for international trade models, one might be interested not only in the effects of tariffs on the mean level of exports from one country to another but also in understanding how (and whether) this effect may vary for different quantiles of the distribution of trade flows. In a more straightforward cross-sectional case, such varying effects can be estimated via the distribution regression (DR) approach initially proposed by Foresi and Peracchi (1995).

Motivated by the current abundance of network datasets and the estimation of international trade flows (which naturally constitutes a network setting where countries form bilateral ties), an estimation method for the DR in a network framework is provided. The contributions of this paper are threefold: (i) I propose an estimation method that is free of the incidental parameter problem, being valid also for estimation in the tails of the distribution of the outcomes (where, as later discussed, the underlying network structure becomes sparse, imposing additional estimation challenges); (ii) I provide the asymptotic properties of the estimator, and I show in Monte Carlo simulation exercises the performance of the estimator in finite samples; (iii) I illustrate the method with an application to the estimation of gravity models for international trade flows.

A broad range of economic relationships can be modeled through a network perspective, particularly through bilateral ties of agents. Relevant examples include models for international trade flows (Helpman et al. 2008), as mentioned before, as well as models for firm-level trade (Alfaro-Urena et al. 2023), for risk sharing (Fafchamps and Gubert 2007), for the diffusion of microfinance loans (Banerjee et al. 2013), and for earnings in employee-employer data (Bonhomme et al. 2019). I, therefore, consider a directed network structure through a dyadic model, in which the outcomes reflect pairwise interactions among the sampled units (Graham 2020). A key aspect of dyadic regressions is the inclusion of observed dyad-level (pair-level) characteristics and unobserved unit-specific effects for each unit in the dyad (both senders and receivers in a

directed framework), which captures the unobserved individual heterogeneity of units. I treat the unit-specific effects as fixed parameters to be estimated, such that their distribution, conditional on the covariates, is left unrestricted. Next, because the estimated effects can also vary with the level of the outcome, the considered model is semiparametric.

The DR approach was initially proposed by Foresi and Peracchi (1995) for the cross-sectional case with independent observations. Its central idea is to directly model the conditional cumulative distribution function of the outcomes, which describes the likelihood of a random variable taking on a value less or equal to a particular value in its support. In the international trade example, it translates to the probability of trade flows being smaller or equal to a specific threshold value. Therefore, the DR approach boils down to estimating the conditional distribution of the outcome of interest with a finite sequence of binary response variables (and thus, a sequence of discrete choice models). More specifically, each binary response outcome is given by an indicator function of the observed dependent variable of interest being below some threshold (for instance, a corresponding quantile). By varying the value of these thresholds, an entire characterization of the conditional distribution of the outcome is obtained. Chernozhukov et al. (2013) extended this approach to a continuum of binary response estimators. However, in practice, the method encompasses, in both cases, the estimation of a sequence of discrete choice models over a grid of values of the dependent variable's support.

The model for each threshold is inherently non-linear since it is a discrete choice model. Due to such non-linearity, including the two-way fixed effects to accommodate the dyadic structure leads to the incidental parameter problem (Neyman and Scott 1948) when jointly estimating all the parameters for a given threshold. In the specific setting considered in this paper, the incidental parameter problem manifests itself by asymptotic bias(es) in the limit distribution of the scaled parameter(s) of interest, resulting in invalid inference.

To deal with the incidental parameter problem, I propose to extend the conditional maximum-likelihood approach of Charbonneau (2017) to estimate single binary choice network models under a logistic specification to multiple (possibly a continuum of) binary choice models for the

thresholds. Note that the estimator of Charbonneau (2017) was initially proposed for a directed network formation model; however, since the structure of those is that of a dyadic discrete choice model, it is also suitable for each of the thresholds of the DR framework. The approach mentioned above relies on conditioning the likelihood function on a specific set of conditions for quadruples of nodes of the network, such that, when the underlying distribution of the outcomes is logistic, it differences out the fixed effects from the likelihood. To show the pointwise asymptotic properties for this estimator for the DR, I use results from Jochmans (2018). This estimator is consistent and converges asymptotically (pointwise) to a Gaussian limit distribution centered around the true parameter value at a parametric rate, delivering valid inference for each of the considered thresholds.

To my knowledge, one of the few papers in the literature that proposes an estimator for the DR model in the framework of a network is Chernozhukov et al. (2020). The key difference in my approach relates to the estimation method employed for each threshold. They propose to deal with the incidental parameter problem by independently employing analytical bias correction estimates for each threshold. However, in the context of a network formation model, an essential assumption for the consistency of this estimator is that the underlying network is dense (Dzemski 2019). In the DR setting, this translates to the conditional probability of the outcomes being smaller than a given threshold to be bounded away from zero or one (Chernozhukov et al. 2020). That is, the estimates are not guaranteed to be consistent and lead to valid inference in the extremum quantiles of the conditional distribution of the outcomes. This problem is attenuated when the outcomes of interest are bounded below at zero and contain many zeros (or another value) since the thresholds of interest lie mostly in the extremum quantiles, a region in the support of the outcome of interest where for a given threshold, the generated binary variable will have little or no variation for some units in the dyad rendering the fixed effects to be not identified.

On the other hand, the conditional maximum likelihood approach proposed here allows for a higher degree of sparsity in the underlying network, being also valid in the extremum quantiles

of the conditional distribution and in situations of outcomes with many zeros. These results are also observed in finite samples through Monte Carlo studies for both the estimates of a given threshold of the distribution (which is essentially a network formation model, as seen in the following Sections), and for the estimates of the entire distribution.

I consider an empirical application to gravity equations for bilateral trade between countries, a natural application of bilateral network models of great economic relevance. In this case, the relevant outcome variable is bounded below at zero, indicating the presence of a heavy upper tail in the distribution, therefore, the DR approach is well-suited for this application (Chernozhukov et al. 2020). I show that the estimated coefficients of the distribution regression vary substantially across the different quantiles of the level of trade flows and are also substantially different from the estimates obtained via analytical bias corrections. Although the estimation procedure proposed does not deliver the average effects of the trade barriers on trade flows (since the individual fixed effects are differenced out and, hence, not estimable), the estimated coefficients have a clear relation to the marginal effects of the quantile function, providing a further interpretation of the estimates. Moreover, in this particular application, joint confidence bands on the estimates allow for testing whether the elasticities of gravity models of trade are heterogeneous, a property that ultimately affects welfare analysis in the international trade literature (Arkolakis et al. (2012), Melitz and Redding (2015), Chen and Novy (2022)).

**Plan of the paper.** Section 2 outlines the main model to be estimated; Section 3 provides the estimation method; Section 4 shows the asymptotic properties of the proposed estimator; Section 5 provides the Monte Carlo simulation results for both a single threshold of the distribution and for the entire distribution; Section 6 outlines the application for gravity models of international trade; and Section 7 concludes.

## 2. A Distribution Regression Model for Networks

This Section introduces a model for the DR approach that considers a directed network structure formed through bilateral ties of units. As initially proposed by Foresi and Peracchi

(1995), the DR provides a modeling approach for the conditional distribution of the outcomes. To accommodate the network framework, a general dyadic setting is considered. Therefore, the conditional distribution function is parametrized as a function of dyad-specific characteristics and fixed effects for each unit in the observed pair of nodes.

Let $\{(y_{ij}, \boldsymbol{x}_{ij}) : (i,j) \in \mathcal{D}\}$ be the observed dataset, where $y_{ij}$ is a scalar outcome variable that can be discrete, continuous or mixed for a dyad $(i,j)$, and $\boldsymbol{x}_{ij}$ is a vector of covariates. I assume that there is a specific region of interest $\mathcal{Y}$ contained in the support of the outcome of interest and that the vector of covariates has support $\mathcal{X} \subseteq \mathbb{R}^{d_x}$. The set of nodes[1] in the network is given by $\mathcal{N} = \{1, 2, \ldots, N\}$, and the total number of observed dyads is given by $n = |\mathcal{D}| = N(N-1)$. The set $\mathcal{D}$ contains, without loss of generality, the indices of the pairs $(i,j)$ that are observed in a directed network without self-links, i.e., $\mathcal{D} = \{(i,j) : i = 1, \ldots, N, j = 1, \ldots, N\} \setminus \{(i,i) : i = 1, \ldots, N\}$[2].

The individual fixed effects for units $i$ and $j$ are taken into account through vectors of unspecified dimensions $\boldsymbol{\nu}_i$ and $\boldsymbol{\omega}_j$ that contain unobserved random variables or effects that might be arbitrarily related to the covariates $\boldsymbol{x}_{ij}$. Therefore, they can be seen as nuisance parameters. The conditional distribution of $y_{ij}$ given $(\boldsymbol{x}_{ij}, \boldsymbol{\nu}_i, \boldsymbol{\omega}_j)$ is given by:

$$F_{y_{ij}}\left(y \mid x_{ij}, v_i, w_j\right) = \Lambda\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_0(y) + \alpha\left(\boldsymbol{\nu}_i, y\right) + \gamma\left(\boldsymbol{\omega}_j, y\right)\right), \quad y \in \mathcal{Y}, \quad (i,j) \in \mathcal{D}, \tag{1}$$

where $\Lambda(\cdot)$ is a known link function assumed to be the logistic distribution throughout this paper. $\boldsymbol{\beta}_0(y)$ is an unknown parameter vector of interest that varies with the levels of $y$; and $\alpha\left(\boldsymbol{\nu}_i, y\right)$ and $\gamma\left(\boldsymbol{\omega}_j, y\right)$ are unspecified measurable functions that can be seen as the unobserved individual fixed effects at a given level of $y$. This model is naturally semiparametric, not only because the parameters are allowed to vary with the output levels but also because it does not restrict how the individual unobserved effects correlate with the covariates. As shown in Appendix A,

[1] Thoughout the paper, I use units, nodes, or individuals interchangeably.

[2] We consider that all the nodes are senders and receivers, but the method in this paper also allows for cases where the nodes that are senders differs from the nodes that are receivers, i.e., $i = 1, \ldots, I$ and $j = 1, \ldots J$, with $I \neq J$; and also for self-links to be formed.

modeling the conditional distribution by this link function is equivalent to reparametrizing the problem in terms of log-odds ratios as initially proposed by Foresi and Peracchi (1995).

A key feature of models of dyadic interaction is the introduction of the two-way fixed effects. Given the double indices nature of the model, it is reasonable to assume that it exhibits a two-way error component structure captured by both units' nuisance terms. This structure incorporates essential aspects of networks since it accounts for part of the dependence across dyads.[3] For instance, the outcome determined from the pairwise interaction between units $i$ and $j$ can be correlated with the outcome resulting from the interaction between $i$ and $k$ due to the fixed effect for unit $i$ and possible correlations in the covariates that share the index $i$ in common. Finally, by allowing the fixed effects for senders and receivers to be different, together with $y_{ij}$ and $\boldsymbol{x}_{ij}$ not necessarily being equal to $y_{ji}$ and $\boldsymbol{x}_{ji}$, this model allows for directed networks. However, notice that the model outlined in this Section and the estimator proposed in the following Section can be easily modified to accommodate undirected and bipartite networks.

Finally, the conditional distribution $F_{y_{ij}}(y \mid \boldsymbol{x}_{ij}, \boldsymbol{\nu}_i, \boldsymbol{\omega}_j)$ can be written as:

$$\begin{aligned} F_{y_{ij}}\left(y \mid \boldsymbol{x}_{ij}, \boldsymbol{\nu}_i, \boldsymbol{\omega}_j\right) &= \mathbb{E}[1\{y_{ij} \leq y\} \mid \boldsymbol{x}_{ij}, \boldsymbol{\nu}_i, \boldsymbol{\omega}_j] \\ &= \Pr[\tilde{y}_{ij} = 1 \mid \boldsymbol{x}_{ij}, \boldsymbol{\nu}_i, \boldsymbol{\omega}_j] \\ &= \Lambda\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_0(y) + \alpha\left(\boldsymbol{\nu}_i, y\right) + \gamma\left(\boldsymbol{\omega}_j, y\right)\right), \end{aligned} \tag{2}$$

where the first equality follows from the definition of a conditional distribution function, and the third equality follows from the specified model in Equation (1). The second equality follows from the probability of an outcome $y_{ij}$ being smaller or equal to a threshold $y$ can be expressed as the probability of a binary variable $\tilde{y}_{ij}$, which is one if the outcome is below $y$ and zero otherwise, being equal to one. Therefore, by constructing a collection of binary variables $\tilde{y}_{ij} = 1\{y_{ij} \leq y\}$, for all pairs $(i, j) \in \mathcal{D}$ and all points in the region of interest $\mathcal{Y}$, $y \in \mathcal{Y}$, the parameters of the DR model can be estimated by a sequence (or continuum) of binary (logistic) regressions with

[3] In this model, part of the dependence also stems from the possibility that the covariates for a given node are correlated.

two-way fixed effects. By varying the threshold $y$ value, it is possible to obtain the estimated slopes for the entire region of interest $\mathcal{Y}$. In practice, when $\mathcal{Y}$ is not finite, it is replaced by a finite subset $\overline{\mathcal{Y}}$.

**Remark 1.** As in Chernozhukov et al. (2020), this approximation works provided that the Hausdorff distance between $\overline{\mathcal{Y}}$ and $\mathcal{Y}$ goes to zero at a rate faster than $1/\sqrt{n}$. In practice, if $\mathcal{Y}$ is an interval $[y, \bar{y}]$, $\overline{\mathcal{Y}}$ can be a fine mesh of $\sqrt{n}\log\log n$ equidistant points covering $\mathcal{Y}$, i.e., $\overline{\mathcal{Y}} = \{y, y+d, y+2d, \ldots, \bar{y}\}$ for $d = (\bar{y}-\underline{y})/(\sqrt{n}\log\log n)$. Alternatively, if $\mathcal{Y}$ is the support of $y_{ij}$, $\overline{\mathcal{Y}}$ can be a grid of $\sqrt{n}\log\log \bar{n}$ sample quantiles with equidistant indexes.

Despite the equivalence between the conditional distribution function and the conditional quantile function in terms of characterizing the conditional distribution of an outcome variable (namely, one function is the inverse of the other), there are important reasons behind choosing the DR over a quantile regression (QR)[4]. In particular, the linear-in-parameters QR may provide a poor approximation to the conditional distribution when the outcome variable does not have a smooth conditional density. In contrast, the DR does not require such smoothness since the approximation to the conditional distribution is made pointwise at each defined threshold in the support of the outcome. Thus, the DR method is well-suited for cases where the variable of interest is censored, or discrete, or has, in general, point masses in its distribution[5]. Finally, to my knowledge there are no available methods for QR models that can allow for two-way fixed effects.

For a given threshold, a discrete choice model, inherently non-linear, with two-way fixed effects, should be estimated to employ the DR method. It is well-known in the literature that the traditional maximum likelihood approach suffers from the incidental parameter problem in these cases. In this particular setting, where both dimensions of the pseudo-panel tend to

[4]Notice that the estimated coefficients of a DR and a QR are only relatable if the set of covariates is rich enough (Chernozhukov et al. 2013).

[5]Even though there is literature on the estimation of QR for censored dependent variables, such as in Galvao Jr (2011) and Chernozhukov et al. (2015), there are no available methods (to my knowledge) for network models. Moreover, the DR approach is more general because it allows for arbitrary mass points in the distribution of the outcome.

infinity at the same rate, the incidental parameter problem manifests as an asymptotic bias in the limiting distribution of the slope parameter $\boldsymbol{\beta}_0(y)$, leading to incorrect inference. In Appendix B, I demonstrate how the asymptotic biases arise in this framework, based on results from Fernández-Val and Weidner (2016).

## 3. Estimation method

As outlined in the previous Section, the main challenge in the estimation of the sequence of binary regressions given by Equation (2) is that, even for a single binary regression, the incidental parameter problem (Neyman and Scott 1948) stems from the presence of the two-way fixed effects. To circumvent this problem, I propose to estimate the parameters of the model $\boldsymbol{\beta}(y)$ for each threshold point (for a given level $y$), independently, with the conditional maximum-likelihood method suggested by Charbonneau (2017) (for directed networks) and concurrently by Graham (2017) (for undirected networks). The core of this approach is to extend the conditional maximum likelihood method for logistic standard panel data models with one fixed effect in Rasch (1960) and Chamberlain (2010)[6] to models with two-way fixed effects, accommodating dyadic structures. The method relies on the existence of a set of conditions for quadruples of nodes in the observed network that differences out the fixed effects from the likelihood when the link function follows a logistic distribution. Estimating binary panel data models by a conditional logit estimation is a well-known semiparametric technique since it avoids specifying the distribution of the fixed effects conditional on covariates. While it seems restrictive due to the distributional assumption, Chamberlain (2010) showed that estimating such models at a parametric rate is only possible when the error terms are logistic. Moreover, it is only possible to difference out the fixed effects under a logistic specification.

The approach of Charbonneau (2017) was initially proposed for network formation models. However, it is applicable to the model that I consider for a single cutoff point since it resembles that of a bilateral network formation model. This follows because it is a discrete choice model

[6] Also refer to Arellano and Honoré (2001) for a survey.

with fixed effects for each node and dyad characteristics. For the sake of simplicity, and without loss of generality, I denote $\boldsymbol{\beta}(y) = \boldsymbol{\beta}_y$, $\alpha_{i,y} = \alpha(\boldsymbol{\nu}_i, y)$ and $\gamma_{j,y} = \gamma(\boldsymbol{\omega}_j, y)$.

From the conditional distribution $F_{y_{ij}}$ and the constructed binary variables $\tilde{y}_{ij}$:

$$\tilde{y}_{ij} = 1\{\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{y,0} + \alpha_{i,y} + \gamma_{j,y} + \varepsilon_{ij} \geq 0\}, \quad (i,j) \in \mathcal{D}$$

where $\alpha_{i,y}$ and $\gamma_{j,y}$ are fixed effects that depend on the level of the threshold $y$, and we assume that $\varepsilon_{ij}$ follows a logistic distribution. Therefore:

$$\begin{aligned} \mathbb{E}[1\{y_{ij} \leq y\} \mid \boldsymbol{x}_{ij}, \alpha_{i,y}, \gamma_{j,y}] &= \Pr[\tilde{y}_{ij} = 1 \mid \boldsymbol{x}_{ij}, \alpha_{i,y}, \gamma_{j,y}] \\ &= \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{y,0} + \alpha_{i,y} + \gamma_{j,y})}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{y,0} + \alpha_{i,y} + \gamma_{j,y})} \end{aligned} \tag{3}$$

**Proposition 1.** *Under the model specification given by Equation (3), the sums across each dimension of the pseudo panel, $\sum_{j=1}^{N} \tilde{y}_{ij}$ and $\sum_{i=1}^{N} \tilde{y}_{ij}$, are sufficient statistics for $\alpha_{i,y}$ and $\gamma_{j,y}$.*

*Proof.* shown in Appendix C.

While this statement is previously proved for the standard panel case with one fixed effect, Graham (2017) only implicitly provides this result for undirected networks. Even though one could propose a conditional maximum likelihood estimator based on the sufficient statistics, the maximization problem might be intractable. Fortunately, Charbonneau (2017) provides a more tractable solution by showing that it is possible to difference out the two-way fixed effects by further conditioning the above probability on the set of events $\{\tilde{y}_{ij} + \tilde{y}_{ik} = 1, \tilde{y}_{lj} + \tilde{y}_{lk} = 1, \tilde{y}_{ij} + \tilde{y}_{lk} = 1\}$ for different indices of senders and receivers $\{i, l; j, k\}$, such that:

$$\begin{aligned} &\Pr[\tilde{y}_{ij} = 1 \mid \boldsymbol{x}_{ij}, \alpha_{i,y}, \gamma_{j,y}, \tilde{y}_{ij} + \tilde{y}_{ik} = 1, \tilde{y}_{lj} + \tilde{y}_{lk} = 1, \tilde{y}_{ij} + \tilde{y}_{lk} = 1] \\ &= \frac{\exp(((\boldsymbol{x}_{ij} - \boldsymbol{x}_{ik}) - (\boldsymbol{x}_{lj} - \boldsymbol{x}_{lk}))'\boldsymbol{\beta}_{y,0})}{1 + \exp(((\boldsymbol{x}_{ij} - \boldsymbol{x}_{ik}) - (\boldsymbol{x}_{lj} - \boldsymbol{x}_{lk}))'\boldsymbol{\beta}_{y,0})}, \end{aligned} \tag{4}$$

which also no longer depends on the fixed effects. This result is obtained by applying the same trick as usually used for the logit estimation in a static standard panel model with a single fixed


Figure 1: Main configurations of informative subgraphs. Solid arrows indicate links that should be present, while dashed arrows indicate links that should not be present.


Subgraph 1: Informative

Subgraph 2: Not informative

Figure 2: Examples of an informative and a non-informative configuration of quadruples to the likelihood.

effect twice.

The last expression is then applied to all quadruples of observations that satisfy the conditions above. Hence, the function to be maximized is given by:

$$\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\sum_{l,k\in Z_{ij}} \log\left(\frac{\exp(((\boldsymbol{x}_{ij}-\boldsymbol{x}_{ik})-(\boldsymbol{x}_{lj}-\boldsymbol{x}_{lk}))'\boldsymbol{\beta}_y)}{1+\exp(((\boldsymbol{x}_{ij}-\boldsymbol{x}_{ik})-(\boldsymbol{x}_{lj}-\boldsymbol{x}_{lk}))'\boldsymbol{\beta}_y)}\right), \tag{5}$$

where $Z_{ij}$ is the set of all potential nodes $k$ and $l$ that satistifies the conditions $\{\tilde{y}_{ij}+\tilde{y}_{ik}=1, \tilde{y}_{lj}+\tilde{y}_{lk}=1, \tilde{y}_{ij}+\tilde{y}_{lk}=1\}$ for the pair $ij$. In the next Section, I show that a simple pairwise differences transformation of the outcomes $\tilde{y}_{ij}$ and the covariates $\boldsymbol{x}_{ij}$ followed by a logit estimation leads to the implementation of this estimator.

As in the case of the static logit considered in Rasch (1960) and Chamberlain (2010), the units that are informative for the likelihood are considered movers. In other words, a given quadruple $\{i,l;j,k\}$, will only be informative for the likelihood if the outcomes, for instance, for node $i$ have variation. To illustrate this argument, consider Figure 1, where every possible wiring rendering an informative quadruple delivers a set of outcomes for each unit (considering a unit both as a

sender and receiver) that has variation. More specifically, Figure 2 shows a configuration that is informative for the likelihood (Subgraph 1), and a not informative configuration. It is clear that, for instance, in Subgraph 1, the node $i$ connects with node $j$, while it does not connect with node $k$. On the contrary, in Subgraph 2 there is no variation in the outcomes for node $i$. For further intuition for the identification of the common parameters, I refer to Appendix D.

## 4. Asymptotic properties

Throughout this Section, I treat the sequence of individual effects $\{\alpha_i, \gamma_j\}_N$ as fixed since I always condition on them. Moreover, I consider asymptotic approximations where both dimensions of the pseudo-panel tend to infinity at the same rate.

The asymptotic properties for a single threshold value of the conditional distribution follow from results provided in Jochmans (2018) for the estimator of Charbonneau (2017). Define the following random variables by fixing a quadruple of distinct nodes $\{i, l; j, k\}$ from $\mathcal{N}$:

$$z(\sigma\{i,l;j,k\}) = \frac{(\tilde{y}_{ij} - \tilde{y}_{ik}) - (\tilde{y}_{lj} - \tilde{y}_{lk})}{2}$$

$$\boldsymbol{r}(\sigma\{i,l;j,k\}) = (\boldsymbol{x}_{ij} - \boldsymbol{x}_{ik}) - (\boldsymbol{x}_{lj} - \boldsymbol{x}_{lk}),$$

where the function $\sigma(\cdot)$ maps a quadruple to the index set $\mathcal{M}_n = \{1, 2, \ldots, M_n\}$, $M_n$ denoting the number of distinct quadruples from $\mathcal{N}$, i.e, $M_n = \binom{N}{2}\binom{N-2}{2} = \frac{N(N-1)(N-2)(N-3)}{4}$.[7] Each distinct quadruple of nodes $\{i, l; j, k\}$ corresponds to a unique $\sigma\{i, l; j, k\} \in \mathcal{M}_n$. In the remainder of this Section, I will use the shortcut notation $z_\sigma$ and $\boldsymbol{r}_\sigma$.

Notice that the transformed dependent variable can take values from the set $\{-1, -1/2, 0, 1/2, 1\}$, and that the event that $z \in \{-1, 1\}$ corresponds to the condition $\{\tilde{y}_{ij} + \tilde{y}_{ik} = 1, \tilde{y}_{lj} + \tilde{y}_{lk} = 1, \tilde{y}_{ij} + \tilde{y}_{lk} = 1\}$. Therefore, by collecting $\boldsymbol{x} = (\boldsymbol{x}_{ij}, \boldsymbol{x}_{ik}, \boldsymbol{x}_{lj}, \boldsymbol{x}_{lk})$, the results in the previous Section leads to the following Lemma:

[7] Notice that the number of quadruples reflect the fact that the senders are permutation invariant, and the receivers as well.

**Lemma 1.** *(Sufficiency)*

$$Pr[z_\sigma = 1 \mid \boldsymbol{x}, z_\sigma \in \{-1, 1\}] = \frac{\exp(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_{y,0})}{1 + \exp(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_{y,0})}$$

As before, conditional on $\boldsymbol{x}$ and on $z_\sigma \in \{-1, 1\}$, the distribution is logistic and does not depend on fixed effects. The conditional log-likelihood of a quadruple is:

$$1\{z_\sigma = 1\}\log\Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_{y,0}) + 1\{z_\sigma = -1\}\log(1 - \Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_{y,0})),$$

which form the basis of the construction of the quasi-conditional maximum likelihood estimator for $\boldsymbol{\beta}_y$. Hence, the model is estimated by maximizing the empirical counterpart of this conditional log-likelihood for all distinct quadruples in $\mathcal{M}$. The estimator can be written as:

$$\hat{\boldsymbol{\beta}}_y = \arg\max_{\boldsymbol{\beta}_y \in \Theta} L_n(\boldsymbol{\beta}_y),$$

where $\Theta$ is the parameter space searched over, and

$$L_n(\boldsymbol{\beta}_y) = \sum_{\sigma \in \mathcal{M}_n} 1\{z_\sigma = 1\}\log\Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y) + 1\{z_\sigma = -1\}\log(1 - \Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y)).$$

It is clear at this point that the objective function is the same as the standard logit log-likelihood function applied to all quadruples that satisfy $z_\sigma \in \{-1, 1\}$. I denote the number of quadruples satisfying it by $M_n^* = \sum_{\sigma \in \mathcal{M}_n} 1\{z_\sigma \in \{-1, 1\}\}$.

The following set of (weak) standard assumptions are needed to establish consistency of the estimator:

**Assumption 4.1.** *(Sampling) The N nodes in $\mathcal{N}$ are sampled independently.*

**Assumption 4.2.** *(Parameter Space) $\boldsymbol{\beta}_{y,0}$ is interior to $\Theta$, a compact subset of $\mathcal{R}^{dim(\boldsymbol{\beta}_y)}$.*

**Assumption 4.3.** *(Moments) For all $(i, j) \in \mathcal{D}$, $\mathbb{E}(||\boldsymbol{x}_{ij}||^2) < C_1$, where $C_1$ is a finite constant.*

Define the expected fraction of quadruples that contribute to the log-likelihood as:

$$p_n = \frac{\mathbb{E}(M_n^*)}{M_n} = \frac{\sum_{\sigma \in \mathcal{M}_n} 1\{z_\sigma \in \{-1, 1\}\}}{M_n}.$$

**Assumption 4.4.** *(Identification)* $Np_n \to \infty$ *as* $N \to \infty$ *and the matrix*

$$\lim_{N \to \infty} (M_n p_n)^{-1} \sum_{\sigma \in \mathcal{M}_n} \mathbb{E}(\boldsymbol{r}_\sigma \boldsymbol{r}_\sigma' f(\boldsymbol{r}_\sigma' \boldsymbol{\beta}_{y,0}) 1\{z_\sigma \in \{-1, 1\}\})$$

*where* $f$ *is the logistic density function has maximal rank.*

Assumption 4.1 allows for dependence of the covariates across dyads that have nodes in common, a key feature in network models. That is, the network dependence of the data arises because not only do the same fixed effects appear across different pairs but also, the covariates of a dyad might be correlated with those of a different dyad with one node in common. Note that this accommodates for settings such as in Graham (2017), where it is assumed that the covariates are of the form $x_{ij} = g(x_i, x_j)$, where $g(\cdot)$ is a measurable function, but it is more general than that. Assumption 4.2 is standard for establishing consistency in non-linear models. Assumption 4.4 allows for the expected fraction of informative quadruples to shrink as $N$ grows, allowing for sparse networks in the context of network formation models. In the DR context, it means that the probabilities of the events $\{y_{ij} \leq y\}$ need not be bounded away from zero and one.

However, $p_n$ should not shrink faster than $N^{-1}$, implying that the accumulation of informative quadruples should not cease as the sample grows. Notice that $p_n$, which can be expressed as $\Pr(z_\sigma \in \{-1, 1\})$, depends on the set of fixed effects of the quadruple $\sigma$. Thus, if the parameters become unbounded as $N$ grows, adding more nodes to the network may not provide additional information for the likelihood. Assumption 4.4 allows for such sequences, such that the method is robust to sparse networks. More specifically, as shown by Jochmans (2018), considering sequences of fixed effects where $\alpha_{i,y}$ and $\gamma_{i,y}$ tend to $-\infty$, and supposing that covariates have bounded

support, by the exponential tails of the logistic distribution we have that, as $N$ increases,

$$p_n \sim \left( \frac{\sum_{i=1}^{N} e^{\alpha_{i,y}}}{N} \right)^2 \left( \frac{\sum_{i=1}^{n} e^{\gamma_{i,y}}}{N} \right)^2.$$

Thus, we can translate the condition that $Np_n \to \infty$ into a restriction on the growth rate of the fixed effects. It is also possible to link the rate condition to the probability of links being formed, $q_n = \sum_{i=1}^{N} \sum_{j \neq i} \Pr\{\tilde{y}_{ij} = 1\}/N(N-1)$, that is, in the DR setting, the probability $q_n = \sum_{i=1}^{N} \sum_{j \neq i} \Pr\{y_{ij} \leq y\}/N(N-1)$, since, in the left tail,

$$q_n \sim \frac{\sum_{i=1}^{N} e^{\alpha_{i,y}}}{n} \frac{\sum_{i=1}^{N} e^{\gamma_{i,y}}}{n}.$$

Hence, $q_n \sim \sqrt{p_n}$. Importantly, in my setting (especially for the application and Monte Carlo exercises), and for the sake of completeness of the argument for sparse networks in general, a similar exercise can be done for sequences of fixed effects growing to $\infty$ with $N$. In this case, $q_n \to 1$ and $p_n \sim (1- \ q_n)^2 \to 0$ at the same rates as above. That is, if the linking probability approaches one, there is less accumulation of quadruples that are informative to the likelihood as well. Based on the previous Section of this paper, the intuition behind it is that in both scenarios, there will be less variation in the set of outcomes of units, rendering less informative quadruples.

Assumption 4.2 and the second part of Assumption 4.4 are standard regularity conditions to establish consistency in non-linear models (Newey and McFadden 1994). From an application of Chebyshev's inequality, the following Theorem holds:

**Theorem 1.** *(Consistency) Let Assumptions 4.1-4.4 hold. Then* $\hat{\boldsymbol{\beta}}_y \xrightarrow{p} \boldsymbol{\beta}_{y,0}$ *as* $N \to \infty$.

*Proof.* Follows from Jochmans (2018), a more detailed and slightly modified proof is available in Appendix E.

Even though the empirical counterpart of the conditional log-likelihood has the form of a standard static logit model for the cross-sectional case, the conventional standard errors are not valid for the estimated $\hat{\boldsymbol{\beta}}_y$. Because the score vector involves sums over quadruples of nodes,

such that each node appears in different summands, it leads to dependences over such summands that need to be considered. This results in an estimator based on a quasi-likelihood, where the information matrix equality does not hold. To derive the asymptotic distribution of the estimator, we first need to strengthen the moment requirements:

**Assumption 4.5.** *(Moments) For all* $(i,j) \in \mathcal{D}$, $\mathbb{E}(||\boldsymbol{x}_{ij}||^6) < C_2$, *where* $C_2$ *is a finite constant.*

Then, each summand of the score vector is introduced as:

$$\boldsymbol{s}(\sigma, \boldsymbol{\beta}_y) = \boldsymbol{r}_\sigma \{1\{z_\sigma = 1\}(1 - \Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y)) + 1\{z_\sigma = -1\}\Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y)\}.$$

Hence, given the permutation invariance of senders and receivers, the score vector is:

$$\boldsymbol{S}_n(\boldsymbol{\beta}_y) = \sum_{i}^{N} \sum_{j \neq i} \sum_{\substack{l > i \\ l \neq j}} \sum_{\substack{k > j \\ k \neq i,l}} \boldsymbol{s}(\sigma\{i,l;j,k\}, \boldsymbol{\beta}_y)$$

The main result to characterize the distribution of the estimator is that $\boldsymbol{\Upsilon}_n(\boldsymbol{\beta}_{y,0})^{-1/2} \boldsymbol{S}_n(\boldsymbol{\beta}_{y,0}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{I})$, where:

$$\boldsymbol{\Upsilon}_n(\boldsymbol{\beta}_y) = \sum_{i} \sum_{j \neq i} \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \sum_{i'' \neq i,j,i',j'} \sum_{j'' \neq i,j,i',j',i''} 16 \times \mathbb{E}\left[\boldsymbol{s}(\sigma\{i,j,i',j'\}, \boldsymbol{\beta}_y) \boldsymbol{s}(\sigma\{i,j,i'',j''\}, \boldsymbol{\beta}_y)'\right] \tag{6}$$

This result, combined with the Hessian that is given by:

$$\boldsymbol{H}_n(\boldsymbol{\beta}_y) = -\sum_{\sigma \in \mathcal{M}_n} \boldsymbol{r}_\sigma \boldsymbol{r}'_\sigma f(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y 1\{z_\sigma \in \{-1,1\}\}).$$

And, finally, defining:

$$\hat{\boldsymbol{\Omega}} = \boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_y)^{-1} \boldsymbol{\Upsilon}_n(\hat{\boldsymbol{\beta}}_y) \boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_y)^{-1}$$

We have:

**Theorem 2.** *(Asymptotic distribution) Let Assumptions 1-5 hold. Then* $||\hat{\boldsymbol{\beta}}_y - \boldsymbol{\beta}_{y,0}|| = O_p(1/\sqrt{N(N-1)p_n})$

*and*

$$\hat{\boldsymbol{\Omega}}^{-1/2}(\hat{\boldsymbol{\beta}}_y - \boldsymbol{\beta}_{y,0}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I})$$

*as* $N \to \infty$.

*Proof.* Follows from Jochmans (2018), a more detailed and slightly modified proof is available in Appendix E.

The proof for Theorem 2 relies on steps that are akin to the ones taken when establishing the limit distribution of a U-statistics, as in Graham (2017). We first propose a projection for the score vector, that resembles a Hájek projection, but it is not formally one, since the kernel of this projection is not symmetric, and we do not only condition on observable and unobservable attributes of a specific dyad $ij$. However, similar steps ara taken when compared to the U-statistics literature: we show that the score evaluated at the true parameter value is asymptotically equivalent to the projection (conditional on covariates), by defining the asymptotic variance of the projection and the score; and, by arguments of conditional independence, it is possible to derive the limiting distribution. The main argument is that, following traditional dyadic models, the probability of $\tilde{y}_{ij} = 1$ for a given dyad $i, j$ is conditionally independent of the probability for the remaining dyads, conditioning on the node (fixed effects) and dyad (covariates) characteristics[8]. Therefore, this model belongs to the class of conditionally independent dyadic (CID) models (Fafchamps and Gubert (2007), Graham (2020) for a review). Importantly, Theorem 2 shows that pointwise (for each threshold $y$), the estimator converges at a parametric rate to the true parameter value. Furthermore, it provides an estimate for the asymptotic variance that delivers valid inference.

This paper is not the first in the literature to propose an estimation method for DR in a network framework. Considering the same setting as in this paper, Chernozhukov et al. (2020) propose to estimate the parameters of the model $\boldsymbol{\theta}(y) := (\boldsymbol{\beta}(y), \alpha_1(y), \dots \alpha_I(y), \gamma_1(y), \dots \gamma_J(y))$

[8]One drawback is that transitivity across the probabilities is not taken into account by this model. It rules out interdependent link preferences, where individuals' preferences over a link may vary with the presence or absence of links elsewhere in the network. However, it is shown by Dzemski (2019) that such a dyadic structure can recover the transitivity observed in some datasets.

also separately for each threshold. The key difference to my approach is that they employ a maximum likelihood method with analytical bias corrections initially proposed by Fernández-Val and Weidner (2016) to the standard panel data setting with two-way fixed effects (additive or interactive) and both dimensions tending to infinity (large N and T) and later applied to the context of a network by Yan et al. (2019) and Dzemski (2019).

Even though their approach encompasses a broader class of models, it does not completely eliminate the asymptotic bias. In comparison, the pairwise difference eliminates it entirely by differencing out the nuisance parameters. Besides, as mentioned before, in the context of a network formation model, the asymptotic bias corrections require that the underlying network is dense, meaning that, in the DR context, the conditional probabilities of the events $\{y_{ij} \leq y\}$ are bounded away from zero and one. Therefore, in the extreme quantiles of the distribution, such an approach might lead to a remaining asymptotic bias. As Assumption 4.4 indicates, the method proposed in this paper allows for the expected fraction of quadruples to shrink to zero as N grows, which implies that the probability of links forming can approach zero or one. Hence, in the DR setting, it is also suitable for more extreme quantiles as opposed to the analytical bias correction method.

## 5. Monte Carlo simulations

### *5.1. Monte Carlo simulations for a single threshold*

In this section, I propose a set of Monte Carlo simulation studies for a single threshold $y$, which boils down to a network formation model. The aim is to compare the performance of the analytical bias correction methods to that of my approach under different levels of sparsity of the network in finite samples. I follow a standard set of data generating processes (DGPs) for directed networks, similar to those in Jochmans (2018).

The outcome variable is generated as follows:

$$y_{ij} = \mathbb{1}(x_{ij}\beta_0 + \alpha_i + \gamma_j - \epsilon_{ij} \geq 0)$$

where $x_{ij}$ is a single regressor, the true parameter value $\beta_0$ is set to one, $\alpha_i$ and $\gamma_j$ are sequences of fixed effects, and $\epsilon_{ij}$ are $N(N-1)$ draws from the standard logistic distribution, with $N$ the number of nodes in the network.

The single regressor is generated as:

$$x_{ij} = -\mid u_i - u_j \mid,$$

where $u_i = \nu_i - \frac{1}{2}$ for $\nu_i \sim Beta(2,2)$. The fixed effects are a deterministic function of the sample size:

$$\alpha_i = -\frac{N-i}{N-1} C_n, \quad \gamma_i = \alpha_i$$

where the constant $C_n$ usually depends on $N$. Specifically, the larger the value of $C_n$, the sparser the generated network, where sparsity here is defined as the fraction of possible links that are observed. I follow most of the literature and consider the following set of values for it: $C_n \in \{0, log(log(N)), log(N)^{1/2}, log(N)\}$. Importantly, the source of the dependence across dyads comes from both the covariates structure and the inclusion of the fixed effects. Moreover, I consider a set of number of nodes: $N \in \{25, 50, 70, 100\}$.

For each specification of the Design above, I run $S = 1000$ Monte Carlo repetitions and compute the estimated coefficients, the estimated standard errors, and the size of the corresponding $t$-statistic for the maximum likelihood estimator (MLE, logistic regression without corrections), the analytical bias correction method (BC), and the conditional maximum likelihood estimator (PD, standing for pairwise differences).

**Remark 2.** The computation of the bias-corrected estimates differs slightly from that of Chernozhukov et al. (2020). I propose to estimate the coefficient $\beta_1$ and the analytical bias corrections taking into account only the subsample consisting of nodes $i$ and $j$ for which there is variation in the outcome variables (thus preventing the perfect prediction problem). While the MLE estimates of $\beta_1$ are, as expected, invariant to whether one considers the entire sample or only the subsample, the estimates of the biases are not. This follows because, although the estimated

fixed effects of the nodes that present no variation are not taken into account in the analytical expression of the bias, the estimates of all fixed effects demonstrate a substantial variation when (i) considering the entire sample or only a subsample or (ii) due to the choice of normalization of the fixed effects. This sensitivity is aggravated when the normalized unit corresponds to a node without variation in the outcome.

Table 1: Network statistics for simulated data according to Design 1. Based on 1000 Monte Carlo replications. Percentage quadruples refer to the average percentage of informative quadruples for the likelihood across the simulations, and percentage links refers to the average percentage of links in the network across the simulations.

| N | $C_N$ | percentage quadruples | $p_N$ | percentage Links | $q_N$ | average In-degree | average Out-degree |
|---|---|---|---|---|---|---|---|
| 25 | 0 | 0.1206 | 1.0000 | 0.4376 | 1.0000 | 10.5024 | 10.5024 |
| 50 | 0 | 0.1205 | 1.0000 | 0.4372 | 1.0000 | 21.4217 | 21.4217 |
| 70 | 0 | 0.1205 | 1.0000 | 0.4366 | 1.0000 | 30.1231 | 30.1231 |
| 100 | 0 | 0.1204 | 1.0000 | 0.4363 | 1.0000 | 43.1970 | 43.1970 |
| 25 | $log(log(N))$ | 0.0493 | 0.1232 | 0.2061 | 0.3509 | 4.9466 | 4.9466 |
| 50 | $log(log(N))$ | 0.0396 | 0.0898 | 0.1803 | 0.2996 | 8.8349 | 8.8349 |
| 70 | $log(log(N))$ | 0.0361 | 0.0788 | 0.1705 | 0.2808 | 11.7674 | 11.7674 |
| 100 | $log(log(N))$ | 0.0329 | 0.0696 | 0.1616 | 0.2638 | 15.9966 | 15.9966 |
| 25 | $log(N)^{1/2}$ | 0.0238 | 0.0486 | 0.1360 | 0.2206 | 3.2633 | 3.2633 |
| 50 | $log(N)^{1/2}$ | 0.0194 | 0.0369 | 0.1210 | 0.1922 | 5.9282 | 5.9282 |
| 70 | $log(N)^{1/2}$ | 0.0175 | 0.0328 | 0.1142 | 0.1810 | 7.8821 | 7.8821 |
| 100 | $log(N)^{1/2}$ | 0.0160 | 0.0291 | 0.1085 | 0.1706 | 10.7426 | 10.7426 |
| 25 | $log(N)$ | 0.0047 | 0.0089 | 0.0596 | 0.0946 | 1.4313 | 1.4313 |
| 50 | $log(N)$ | 0.0025 | 0.0043 | 0.0425 | 0.0654 | 2.0800 | 2.0800 |
| 70 | $log(N)$ | 0.0018 | 0.0031 | 0.0364 | 0.0557 | 2.5105 | 2.5105 |
| 100 | $log(N)$ | 0.0014 | 0.0023 | 0.0311 | 0.0475 | 3.0793 | 3.0793 |
| 25 | $2log(N)$ | 0.0003 | 0.0008 | 0.0171 | 0.0288 | 0.4098 | 0.4098 |
| 50 | $2log(N)$ | 0.0002 | 0.0003 | 0.0114 | 0.0184 | 0.5590 | 0.5590 |
| 70 | $2log(N)$ | 0.0001 | 0.0002 | 0.0095 | 0.0152 | 0.6536 | 0.6536 |
| 100 | $2log(N)$ | 0.0001 | 0.0002 | 0.0081 | 0.0127 | 0.8027 | 0.8027 |

Table 2: Simulation results for Design 1. MLE refers to the Logit maximum-likelihood estimator, BC refers to the analytical bias-correction estimates, and PD refers to the conditional maximum-likelihood method proposed in this paper.

| | | Mean Bias | | | Median Bias | | | Size t-test | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N | $C_N$ | MLE | BC | PD | MLE | BC | PD | MLE | BC | PD | MLE | BC | PD |
| 25 | 0 | 0.1232 | 0.0273 | -0.0235 | 0.1218 | 0.0259 | -0.0051 | 0.0400 | 0.0240 | 0.0280 | 0.6053 | 0.5421 | 0.5996 |
| 50 | 0 | 0.0482 | 0.0050 | 0.0014 | 0.0562 | 0.0126 | 0.0046 | 0.0750 | 0.0480 | 0.0340 | 0.3024 | 0.2862 | 0.2750 |
| 70 | 0 | 0.0300 | -0.0001 | 0.0047 | 0.0243 | -0.0056 | 0.0069 | 0.0540 | 0.0470 | 0.0540 | 0.2088 | 0.2005 | 0.2004 |
| 100 | 0 | 0.0240 | 0.0032 | 0.0027 | 0.0210 | 0.0003 | 0.0012 | 0.0600 | 0.0490 | 0.0560 | 0.1391 | 0.1342 | 0.1381 |
| 25 | $log(log(N))$ | 0.1258 | 0.0187 | -0.0068 | 0.1425 | 0.0347 | -0.0298 | 0.0530 | 0.0300 | 0.0290 | 0.7842 | 0.6995 | 0.7673 |
| 50 | $log(log(N))$ | 0.0482 | 0.0002 | 0.0147 | 0.0508 | 0.0027 | 0.0139 | 0.0660 | 0.0550 | 0.0380 | 0.3994 | 0.3780 | 0.3670 |
| 70 | $log(log(N))$ | 0.0228 | -0.0103 | -0.0112 | 0.0211 | -0.0114 | -0.0068 | 0.0600 | 0.0420 | 0.0380 | 0.2778 | 0.2678 | 0.2611 |
| 100 | $log(log(N))$ | 0.0251 | 0.0019 | -0.0030 | 0.0198 | -0.0035 | 0.0003 | 0.0590 | 0.0470 | 0.0430 | 0.1932 | 0.1872 | 0.1930 |
| 25 | $log(N)^{1/2}$ | 0.1257 | 0.0088 | 0.0339 | 0.1249 | 0.0087 | 0.0152 | 0.0560 | 0.0290 | 0.0230 | 1.0054 | 0.8933 | 0.9238 |
| 50 | $log(N)^{1/2}$ | 0.0557 | 0.0021 | -0.0060 | 0.0644 | 0.0118 | -0.0107 | 0.0600 | 0.0450 | 0.0480 | 0.4780 | 0.4503 | 0.4595 |
| 70 | $log(N)^{1/2}$ | 0.0359 | -0.0016 | 0.0020 | 0.0380 | 0.0003 | 0.0085 | 0.0430 | 0.0350 | 0.0500 | 0.3286 | 0.3145 | 0.3244 |
| 100 | $log(N)^{1/2}$ | 0.0362 | 0.0099 | -0.0044 | 0.0407 | 0.0148 | -0.0066 | 0.0550 | 0.0470 | 0.0550 | 0.2328 | 0.2242 | 0.2344 |
| 25 | $log(N)$ | 0.2036 | 0.0668 | 0.0030 | 0.1769 | 0.0517 | 0.0739 | 0.0780 | 0.0440 | 0.0180 | 1.8694 | 1.6546 | 1.7137 |
| 50 | $log(N)$ | 0.0896 | 0.0261 | 0.0376 | 0.0694 | 0.0081 | 0.0526 | 0.0710 | 0.0520 | 0.0280 | 0.8845 | 0.8284 | 0.7629 |
| 70 | $log(N)$ | 0.0937 | 0.0466 | -0.0192 | 0.0809 | 0.0336 | -0.0103 | 0.0610 | 0.0500 | 0.0340 | 0.6434 | 0.6102 | 0.5820 |
| 100 | $log(N)$ | 0.0610 | 0.0276 | 0.0101 | 0.0606 | 0.0286 | 0.0105 | 0.0540 | 0.0490 | 0.0460 | 0.4380 | 0.4205 | 0.4351 |
| 25 | $2log(N)$ | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 50 | $2log(N)$ | 0.2178 | 0.1468 | -0.0176 | 0.1904 | 0.1322 | 0.0753 | 0.0620 | 0.0440 | 0.0220 | 2.2601 | 2.1286 | 2.2657 |
| 70 | $2log(N)$ | 0.1444 | 0.0973 | 0.0050 | 0.1578 | 0.1051 | -0.0050 | 0.0640 | 0.0490 | 0.0250 | 1.4514 | 1.3892 | 1.3792 |
| 100 | $2log(N)$ | 0.0485 | 0.0177 | -0.0056 | 0.0886 | 0.0574 | 0.0082 | 0.0570 | 0.0490 | 0.0290 | 0.9786 | 0.9490 | 0.9352 |

Table 1 shows the network statistics obtained from the simulated data. The results show that, as expected, as the value of the parameter $C_N$ increases, the percentage of informative quadruples and the percentage of links and averages in and out-degree decrease. Overall, this indicates that the larger the $C_N$, the sparser is the underlying network. Finally, the values of the calculated $p_N$ and $q_N$ are a good approximation of the average percentage of informative quadruples and links in the network when the network is sparser.

Table 2 shows the results obtained with the logit maximum-likelihood estimator (MLE), the bias correction method (BC), and the method proposed in this paper (PD). Both in terms of the mean and the median biases, when $C_N = log(N)$ or $C_N = 2log(N)$, the MLE performs poorly for any sample size (when the network is denser, the MLE mostly performs poorly when N=25 or 50). While the BC estimator reduces part of the biases, it is clear that the PD is more effective at reducing the bias, especially in small sample sizes. Note, however, that when $N = 25$, there are no available estimates since the number of informative quadruples for the likelihood (for the PD estimator) and the number of units displaying variation in the outcomes (for the BC estimator) is minimal.

In terms of the size of the t-test, surprisingly, the MLE shows the correct size throughout the specifications. However, these results are misleading due to the sizeable bias seen in denser specifications. Concerning the PD estimator, in general, the estimated variance is somewhat overestimated, a finding also seen in Jochmans (2018).

Moreover, in the following, I deviate from the standard specifications for $C_N$ such that it is possible to have a better comparison between the two methods. Namely, I vary the constant $C_N$ for different sample sizes indicated by the number of nodes $N$ such that the number of informative quadruples for the Charbonneau (2017) estimator remains constant. Tables 3 shows the networks statistics such as the percentage of informative quadruples to the likelihood, the average in and out-degree, and the number of informative quadruples, that as expected, remains constant across the configurations. Table 4 show the mean bias, median bias, size of the t-test and the RMSE obtained with the simulations.

The simulation exercise shows that when increasing both $N$ and $C_N$, while the number of quadruples contributing to the likelihood function of the Pairwise Differencing (PD) estimator remains reasonably constant, the sparsity in the network increases (as it is reflected by the average percentage links). Note that, as expected from before, the mean bias of the PD estimates are generally smaller in magnitude than that of the Bias Corrected (BC) estimator for the sparser settings. This difference is even bigger in magnitude when considering the Median Bias.

Table 3: Network statistics for alternative simulated data for Design 1. Based on 1000 Monte Carlo replications. Percentage quadruples refer to the average percentage of informative quadruples for the likelihood across the simulations, and percentage links refers to the average percentage of links in the network across the simulations.

| N | $C_N$ | percentage quadruples | $p_N$ | percentage links | $q_N$ | average In-degree | average. Out-degree | # informative quadruples |
|---|---|---|---|---|---|---|---|---|
| 25 | 3.65 | 0.00301 | 0.00590 | 0.04843 | 0.07684 | 1.16220 | 1.16220 | 228.78 |
| 50 | 7.75 | 0.00016 | 0.00035 | 0.01155 | 0.01868 | 0.56614 | 0.56614 | 228.19 |
| 70 | 10.85 | 0.00004 | 0.00009 | 0.00593 | 0.00964 | 0.40920 | 0.40920 | 228.47 |
| 100 | 15.65 | 0.00001 | 0.00002 | 0.00288 | 0.00468 | 0.28499 | 0.28499 | 225.74 |

Table 4: Alternative simulation results for Design 1. MLE refers to the Logit maximum-likelihood estimator, BC refers to the analytical bias-correction estimates, and PD refers to the conditional maximum-likelihood method proposed in this paper.

| | | Mean Bias | | | Median Bias | | | Size t-test | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N | $C_N$ | MLE | BC | PD | MLE | BC | PD | MLE | BC | PD | MLE | BC | PD |
| 25 | 3.65 | 0.1961 | 0.0568 | -0.0205 | 0.2643 | 0.1347 | 0.0220 | 0.0680 | 0.0350 | 0.0150 | 2.4219 | 2.1588 | 1.9873 |
| 50 | 7.75 | 0.1966 | 0.1268 | -0.0953 | 0.1812 | 0.1211 | -0.0015 | 0.0660 | 0.0460 | 0.0260 | 2.2303 | 2.1006 | 2.2124 |
| 70 | 10.85 | 0.1490 | 0.1010 | -0.0875 | 0.2146 | 0.1737 | -0.0600 | 0.0760 | 0.0640 | 0.0210 | 2.3345 | 2.2377 | 1.8724 |
| 100 | 15.65 | 0.2652 | 0.2282 | -0.0264 | 0.2910 | 0.2571 | 0.1159 | 0.0590 | 0.0520 | 0.0230 | 2.2051 | 2.1407 | 2.1044 |

*5.2. Monte Carlo simulations for the entire distribution*

To analyze the finite sample properties of both DR estimators, I follow the same Monte Carlo simulations setting as Chernozhukov et al. (2020), which is calibrated to the empirical application in the following Section for gravity models of international trade. More specifically, I set the outcome to be generated by a censored logistic process

$$y_{ij}^s = \max\left\{x_{ij}'\hat{\beta} + \widehat{\alpha}_i + \widehat{\gamma}_j + \widehat{\sigma}\Lambda^{-1}\left(u_{ij}^s\right)/\sigma_L, 0\right\}, \quad (i,j) \in \mathcal{D}$$

where $\mathcal{D} = \{(i,j) : 1 \leq i,j \leq 157, i \neq j\}, x_{ij}$ is the value of the covariates for the observational unit $(i,j)$ in the trade data set, $y_{ij}^s$ is the level of exports from country $i$ to $j$, $\sigma_L = \pi/\sqrt{3}$, the standard deviation of the logistic distribution, and $\left(\widehat{\beta}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_I, \widehat{\gamma}_1, \ldots, \widehat{\gamma}_J, \widehat{\sigma}\right)$ are Tobit fixed effect estimates of the parameters in the trade data set with lower censoring point at zero. Moreover, I set the errors to be independently drawn from a uniform distribution $\mathcal{U}(0,1)$. For simplicity, in this simulation exercise, I consider only one covariate, the log of the distance between countries.

Importantly, it can be shown that the conditional distribution of the dependent variable $y_{ij}^s$ is equivalent to a DR model as defined before, where:

$$\beta(y) = \sigma_L\left(e_1 y - \widehat{\beta}\right)/\widehat{\sigma}, \quad \alpha_i(y) = -\sigma_L\widehat{\alpha}_i/\widehat{\sigma}, \quad \text{and} \quad \gamma_j(y) = -\sigma_L\widehat{\gamma}_j/\widehat{\sigma}$$

with $e_1$ the unit vector of dimension $d_x$ with a one in the first component. The results are based on 250 simulations for now due to computational limitations (*it is to be expanded in next versions of this paper*).

Figure 3 shows the absolute bias, absolute median bias, and RMSE obtained with a naive fixed effects logit estimation (Uncorrected Logit UL), the Bias Correction (BC) method of Chernozhukov et al. (2020) and with the proposed Pairwise Differences (PD) estimator in this paper. Both BC and PD reduce the bias significantly in finite samples compared to the UL estimates. However, as expected, in the extreme quantiles, the BC method does not seem to fully correct

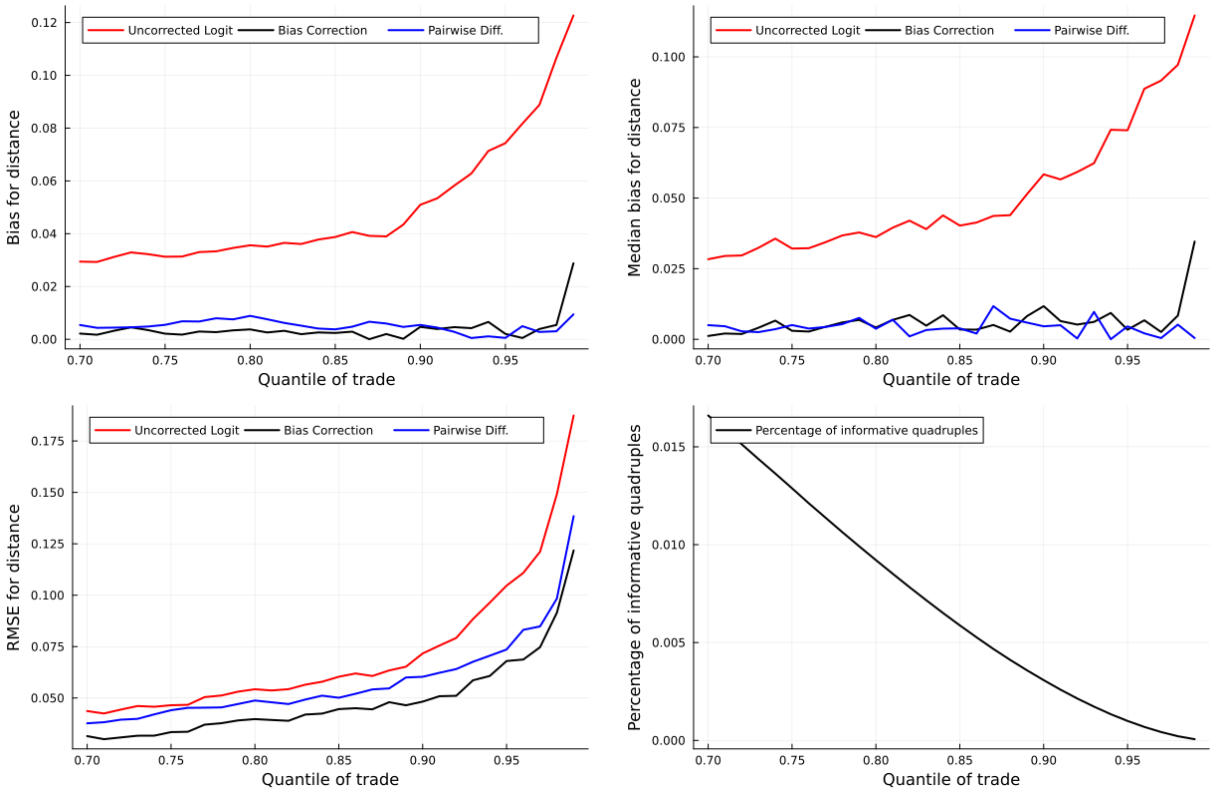


Figure 3: Simulation results for the DR coefficients of log distance. Based on 250 Monte Carlo simulations.

for the bias (both mean and median), while the PD biases do not increase relative to the other quantiles. Moreover, as expected, the RMSE for the UL is the biggest due to the high magnitude of the biases. Also, since the PD estimator is not efficient, its RMSE is larger than that of the BC.

Figure 3 also displays the percentage of informative quadruples to the likelihood of the PD estimator. Naturally, as the quantiles increase, the percentage and number of quadruples decrease significantly. However, at the 99% quantile, there are still about 10000 informative quadruples, which renders the PD estimation robust at the tail of the outcome distribution.

## 6. Application to gravity models of international trade

One key feature of models for bilateral international trade (and many other applications) is that the outcome of interest, which in this case is the volume of trade, is bounded below at zero and contains many zeros. Another important aspect in such models is the inclusion of

importer and exporter country effects, which, in the international trade literature, are known as the multilateral resistance terms (Anderson and Van Wincoop 2003), i.e., the barriers to trade that each country faces with all its trading partners that are not observable.

Despite the abundance of datasets and models with this feature, estimating models reflecting such a structure in the outcome of interest remains challenging, especially when there are very few nonzero observations, such that the network is sparse at the outcome level.

**Remark 3.** Notably, this notion of sparsity differs from the one mentioned in the previous Sections of this paper. Here, I refer to the degree of sparsity as the frequency of zero observed outcomes (trade flows in this application) relative to the amount of strictly positive outcomes. Earlier in this paper, the degree of sparsity referred to the frequency of zeros (or ones) in a binary variable obtained after constructing the variables $\tilde{y}_{ij}$ for each threshold in the support of the outcome variable. I denote the former form of sparsity as *first-degree sparsity*, and the latter as *second-degree sparsity*.

To illustrate the challenges that arise with the first-degree sparsity, consider the following two-way fixed effects model with a possible selection bias:

$$y_{1,ij} = y_{2,ij}\left(\boldsymbol{x}'_{1,ij}\boldsymbol{\beta}_{1,0} + \alpha_i + \gamma_j + u_{ij}\right) \tag{7}$$

$$y_{2,ij} = \mathbb{1}\left(y^*_{2,ij} > 0\right) \tag{8}$$

$$y^*_{2,ij} = \boldsymbol{x}'_{2,ij}\boldsymbol{\beta}^*_{2,0} + \xi^*_i + \zeta^*_j + \eta^*_{ij}, \tag{9}$$

$$(i = 1, \dots N; j = 1, \dots N, i \neq j)$$

where $\alpha_i, \gamma_j, \xi^*_i$ and $\zeta^*_j$ are individual fixed effects. Moreover, pairs $(i, j)$ first decide whether to form a link, in which occasion $y_{2,ij} = 1$ and then a nonzero outcome $y_{1,ij}$ is observed, and zero otherwise. Therefore, this model generates outcomes $y_{1,ij}$ with potentially many zeros. The unobservable individual-specific effects might arbitrarily depend on the observable explanatory variables in both equations. Thus, they are considered nuisance parameters to be estimated in a semi-parametric model. The errors in the equations ($u_{ij}$ and $\eta_{ij}$) might be correlated, in which

case sample selectivity should be addressed. In the gravity model case, the equations above are obtained after a log-linearization of the original model, which has a multiplicative form.

Currently, there are two strands in the literature of gravity equations on how to take into account the zeros: (i) modeling through a sample selection model (Helpman et al. 2008); or (ii) considering the model in its multiplicative form through a Poisson pseudo maximum likelihood (PPML) estimator (Silva and Tenreyro 2006). However, due to the introduction of the two-way fixed effects and the non-linearities in both models, both estimations suffer from the incidental parameter problem. While bias-correction methods have been proposed for both the PPML (Fernández-Val and Weidner 2016)[9] and the first stage of the sample selection model (Dzemski (2019) and Yan et al. (2019)), the estimates are consistent and valid inference is available only under dense networks, as mentioned before. The first case refers to the concept of first-degree sparsity, and the second case refers to second-degree sparsity. Notice that the method of Charbonneau (2017) can be employed for the first stage of the sample selection, providing asymptotically unbiased estimates even when the network is sparse. However, this method does not provide estimates of the fixed effects. Hence, the estimation of the second stage is infeasible.

Therefore, the DR approach proposed in this paper fills this gap in the literature, displaying advantages compared to the previously available methods. Namely, (i) it allows for zero outcomes and other censoring points by relying on approximating the conditional distribution of the outcomes pointwise at given thresholds and avoiding strong assumptions on how the zeros are generated, (ii) it allows for the presence of many zeros being suitable for sparse networks (in both notions of sparsity), and (iii) it can accommodate conditional heteroskedasticity, which is also a topic of discussion in the estimation of gravity models for international trade. As a by-product, this approach also delivers conclusions concerning the possible heterogeneity of the effects of observable trade barriers (covariates) across the distribution of trade flows.

I consider the estimation of gravity equations for bilateral trade between countries, using the same data as Helpman et al. (2008), Jochmans (2018) and Chernozhukov et al. (2020). It

[9]This estimator accommodates both binary choice models as well as other non-linear models, encompassing a wide class of models.

| | Mean | Std. Dev. |
|---|---|---|
| Trade | 0.45 | 0.50 |
| Trade volume | 84.54 | 1,082,219 |
| Log distance | 4.18 | 0.78 |
| Legal | 0.37 | 0.48 |
| Language | 0.29 | 0.45 |
| Religion | 0.17 | 0.25 |
| Border | 0.02 | 0.13 |
| Currency | 0.01 | 0.09 |
| FTA | 0.01 | 0.08 |
| Colony | 0.01 | 0.10 |
| Country pairs | | |

Table 5: Descriptive statistics. Source: Helpman et al. (2008).

contains information on bilateral trade flows and covariates for 157 countries in 1986 (Congo is excluded due to the lack of variation in the dependent variable since it does not export to any other country in 1986). $i$ and $j$ index each country as an exporter and an importer. There are no trade flows for 55% of the country pairs in the dataset, which, together with the fact that the volume of trade variable exhibits a much larger standard deviation than its mean, as indicated in Table 5, indicates the presence of a very heavy upper tail in the distribution. This feature, combined with the arguments highlighted previously, makes DR methods especially well-suited for this application on robustness grounds.

The outcome $y_{ij}$ is the volume of trade in thousands of constant 2000 U.S. dollars from country $i$ to country $j$. The bilateral trade flows data are obtained from Feenstra's "World Trade Flows, 1970-1992", and it was transformed to constant 2000 U.S. dollars using the U.S. CPI by Helpman et al. (2008). The covariates $\boldsymbol{x}_{ij}$ include the following determinants of bilateral trade flows:

1. Distance: the logarithm of the distance in kilometers between the capitals of exporter $i$ and importer $j$.

2. Colonial ties: a binary variable that takes the value one if country $i$ colonized country $j$

(or vice-versa), and zero otherwise.

3. Currency union: a binary variable that takes value one if countries $i$ and $j$ use the same currency or if the currencies were interchangeable at a 1:1 exchange rate for an extended period of time, and zero otherwise.

4. Regional free trade area (FTA): a binary variable that takes value one if the countries $i$ and $j$ belong to a common regional trade agreement and zero otherwise.

5. Border: a binary variable that equals one if the countries $i$ and $j$ are neighbors and zero otherwise.

6. Legal system: a binary variable that takes value one if the countries $i$ and $j$ share the same legal origin, and zero otherwise.

7. Language: a binary variable that takes value one if the countries $i$ and $j$ share the same language, and zero otherwise.

8. Religion: index of common religion constructed as (% Protestants in country $i$ $\times$ % Protestants in country $j$) + (% Catholics in country $i$ $\times$ % Catholics in country $j$) + (% Muslims in country $i$ $\times$ % Muslims in country $j$).

The variables distance, colonial ties, border, legal system, language, and religion were constructed by Helpman et al. (2008) using the CIA's World Factbook; and the variables currency union and FTA were constructed using data from Rose (2000) and Glick and Rose (2002).

Figure 4 shows estimates and 95% pointwise confidence intervals for the DR coefficients of log distance plotted against the quantile indexes of the trade volume. I plot three different curves: one obtained by using bias-corrected (BC) fixed effects estimates for each considered quantile (the approach of Chernozhukov et al. (2020)); one obtained by employing the method proposed in this paper, using the pairwise differencing (PD) of outcomes and regressors; and one obtained when using the uncorrected FE logit estimation at each quantile. The $x$-axis starts at .54, the maximum quantile index corresponding to zero trade volume. The region of interest

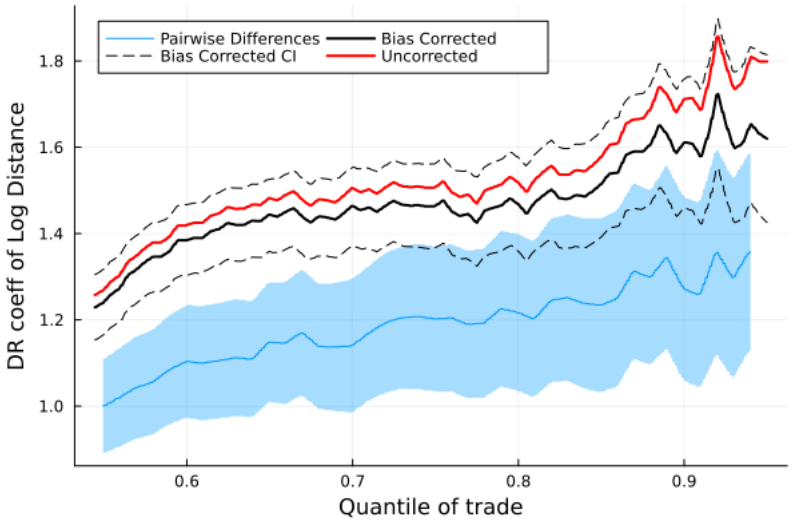


Figure 4: Estimates and 95% pointwise confidence intervals for the DR-coefficients of log distance.

$\mathcal{Y}$ corresponds to the interval between zero and the 0.95-quantile of the trade volume. Note that the sign of the effect in terms of trade volume, $y_{ij}$ , is the opposite of the sign of the DR coefficient. Figure 5 shows the analogous estimates for the DR coefficients of the legal system.

As in Chernozhukov et al. (2020), notice that the difference between the uncorrected and the bias-corrected estimates is of the same order as the magnitude of the width of the confidence intervals for log distance. Moreover, the largest estimated bias when comparing the two and also when comparing with the pairwise difference estimator lies on the upper quantiles of trade, where the constructed binary variables have less variation. However, our estimates suggest an even higher bias in magnitude across the entire distribution. Interestingly, when compared to the BC estimates, the difference between the two estimates (PD and BC) seems to be constant across the distribution. Notably, the PD estimates suggest that the effect of distance across the distribution is significant but of a smaller magnitude. A similar conclusion is drawn in general for the coefficients for the legal system. An exception is at the upper quantiles, where the PD estimates suggest a smaller bias relative to the BC estimates. Finally, as expected, the confidence intervals around the proposed estimator are wider than that of the BC estimator since this estimator is not as efficient as standard MLE.

Even though it would be desirable to compute counterfactual effects for the estimated distri-

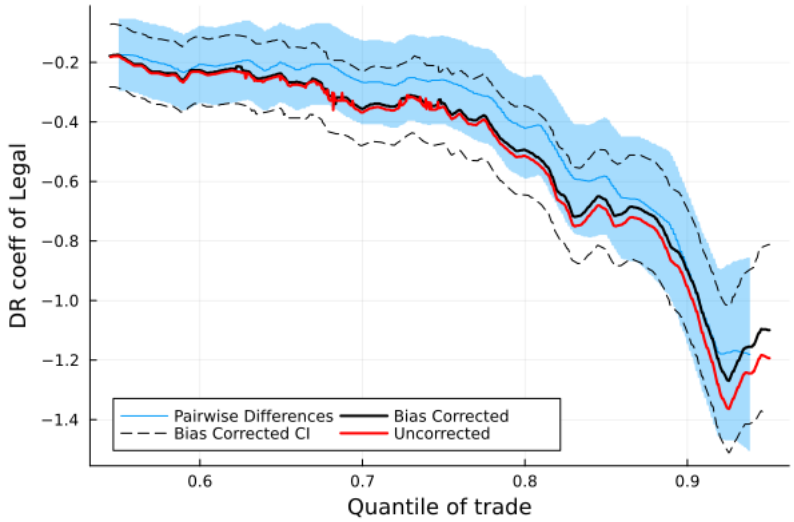


Figure 5: Estimates and 95% pointwise confidence intervals for the DR-coefficients of legal.

bution function (or quantile functions), a drawback of this approach is that such an estimation is infeasible. This occurs since there are no available estimates for the fixed effects or average marginal effects in general (for which the estimated fixed effects are an input), as opposed to the bias correction method of Chernozhukov et al. (2020). However, there is a further interpretation for the estimated coefficients in terms of the derivatives of the conditional quantiles under certain conditions.

### *6.1. Relation to the conditional quantile function*

The conditional distribution of $y_{ij}$ given the covariates and the unobserved effects can be represented by either the conditional distribution function or the conditional quantile function. While these equivalent representations correspond to two alternative approaches for estimation, there are relevant links between DR and quantile regression (QR) estimates, as shown by Koenker et al. (2013). In particular, following results from Chernozhukov et al. (2020) for our framework, it is possible to show that the common parameters of the model are related to the derivatives of the conditional quantiles under certain conditions.

When $y_{ij}$ is continuous, the model given by Equation 1 has the representation as an implicit

nonseparable model by the probability integral transform:

$$\Lambda\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}\left(y_{ij}\right)+\alpha\left(\boldsymbol{\nu}_i,y_{ij}\right)+\gamma\left(\boldsymbol{\omega}_j,y_{ij}\right)\right)=u_{ij},\quad u_{ij}\mid\boldsymbol{x}_{ij},\boldsymbol{\nu}_i,\boldsymbol{\omega}_j\sim U(0,1). \tag{10}$$

where, as commonly seen in DR or QR approaches, $u_{ij}$ represents the unobserved ranking of the observation $y_{ij}$ in the conditional distribution. Let $Q\left(u|\boldsymbol{x}_{ij},\boldsymbol{\nu}_i,\boldsymbol{\omega}_j\right)$ be the $u$-quantile of $y_{ij}$ conditional on $(\boldsymbol{x}_{ij},\boldsymbol{\nu}_i,\boldsymbol{\omega}_j)$. This quantile function can be defined as:

$$Q\left(u\mid\boldsymbol{x}_{ij},\boldsymbol{\nu}_i,\boldsymbol{\omega}_j\right)=\inf\left\{y\in\mathcal{Y}:F_{y_{ij}}\left(y\mid\boldsymbol{x}_{ij},\boldsymbol{\nu}_i,\boldsymbol{\omega}_j\right)\geq u\right\}\wedge\sup\{y\in\mathcal{Y}\}. \tag{11}$$

It can be shown that if (i) $F_{y_{ij}}\left(y\mid\boldsymbol{x}_{ij},\boldsymbol{\nu}_i,\boldsymbol{\omega}_j\right)$ is strictly increasing in the support of $y_{ij}$; (ii) $\partial\Lambda(z)/\partial z>0$ for all $y$ in the support of $y_{ij}$; and (iii) $Q\left(u\mid\boldsymbol{x}_{ij},\boldsymbol{\nu}_i,\boldsymbol{\omega}_j\right)$ is differentiable, then the DR coefficients are proportional to (minus) derivatives of the conditional quantile function, and ratios of the DR coefficients correspond to ratios of derivatives:

$$\left.\frac{\beta_\ell(y)}{\beta_k(y)}\right|_{y=Q(u|\boldsymbol{x}_{ij},\boldsymbol{\nu}_i,\boldsymbol{\omega}_j)}=\frac{\partial_{x_{ij}^\ell}Q\left(u\mid\boldsymbol{x}_{ij},\boldsymbol{\nu}_i,\boldsymbol{\omega}_j\right)}{\partial_{x_{ij}^k}Q\left(u\mid\boldsymbol{x}_{ij},\boldsymbol{\nu}_i,\boldsymbol{\omega}_j\right)},\quad \ell,k=1,\ldots,d_x \tag{12}$$

Therefore, based on the figures above for the estimated coefficients, it is possible to infer that the marginal effects of distance on the quantile functions are larger in magnitude than that of the legal variable.

Finally, there is an ongoing debate in the international trade literature regarding the homogeneity of trade elasticities, which ultimately affects welfare gains from trade (Arkolakis et al. (2012), Melitz and Redding (2015), Chen and Novy (2022)). The method presented in this paper provides a straightforward way (provided the confidence bands for the estimates) to test the heterogeneity of trade elasticities across different quantiles of trade distribution, proving to be of empirical and theoretical importance in the literature.

## 7. Conclusion

A novel method for estimating distribution regressions in a network setting is introduced. To accommodate the network structure, it envolves a semiparametric approach, treating two-way unit-specific effects as fixed parameters, and I address the incidental parameter problem using a conditional maximum-likelihood approach initially proposed for network formation models (Charbonneau (2017), Jochmans (2018)). The proposed method provides consistent estimates and robust inference pointwise, particularly for extremum quantiles of the distribution. This approach fills a gap in the econometrics of network model literature by accommodating zero outcomes and sparse networks without relying on strong assumptions regarding how the zero outcomes are generated. Moreover, the empirical application demonstrates its practical relevance, allowing, for instance, to test whether the elasticities of gravity models of international trade are heterogeneous across thresholds. A current drawback of the proposed method is that estimates of counterfactual distributions are infeasible. This is because the estimates of fixed effects are unavailable, and the average (marginal) effects of network formation models remain set-identified when the network is sparse - which is the case of the underlying network in the extreme quantiles of the distribution of outcomes. Therefore, future research would involve obtaining estimates for bounds on the partially identified average effects of network models.

## References


Alfaro-Urena, A., J. Castro-Vincenzi, S. Fanelli, and E. Morales (2023): "Firm export dynamics in interdependent markets," Tech. rep., National Bureau of Economic Research.

Anderson, J. E. and E. Van Wincoop (2003): "Gravity with gravitas: A solution to the border puzzle," *American economic review*, 93, 170–192.

Arellano, M. and J. Hahn (2007): "Understanding bias in nonlinear panel models: Some recent developments," *Econometric Society Monographs*, 43, 381.

Arellano, M. and B. Honoré (2001): "Panel data models: some recent developments," in *Handbook of econometrics*, Elsevier, vol. 5, 3229–3296.

Arkolakis, C., A. Costinot, and A. Rodríguez-Clare (2012): "New trade models, same old gains?" *American Economic Review*, 102, 94–130.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "The diffusion of microfinance," *Science*, 341, 1236498.

Bonhomme, S., T. Lamadon, and E. Manresa (2019): "A distributional framework for matched employer employee data," *Econometrica*, 87, 699–739.

Chamberlain, G. (2010): "Binary response models for panel data: Identification and information," *Econometrica*, 78, 159–168.

Charbonneau, K. B. (2017): "Multiple fixed effects in binary response panel data models," *The Econometrics Journal*, 20, S1–S13.

Chen, N. and D. Novy (2022): "Gravity and heterogeneous trade cost elasticities," *The Economic Journal*, 132, 1349–1377.

Chernozhukov, V., I. Fernández-Val, and A. E. Kowalski (2015): "Quantile regression with censoring and endogeneity," *Journal of Econometrics*, 186, 201–221.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "Inference on counterfactual distributions," *Econometrica*, 81, 2205–2268.

Chernozhukov, V., I. Fernandez-Val, and M. Weidner (2020): "Network and panel quantile effects via distribution regression," *Journal of Econometrics*.

Dzemski, A. (2019): "An empirical model of dyadic link formation in a network with unobserved heterogeneity," *Review of Economics and Statistics*, 101, 763–776.

Fafchamps, M. and F. Gubert (2007): "The formation of risk sharing networks," *Journal of development Economics*, 83, 326–350.

Fernández-Val, I. and M. Weidner (2016): "Individual and time effects in nonlinear panel models with large N, T," *Journal of Econometrics*, 192, 291–312.

Foresi, S. and F. Peracchi (1995): "The conditional distribution of excess returns: An empirical analysis," *Journal of the American Statistical Association*, 90, 451–466.

Galvao Jr, A. F. (2011): "Quantile regression for dynamic panel data with fixed effects," *Journal of Econometrics*, 164, 142–157.

Glick, R. and A. K. Rose (2002): "Does a currency union affect trade? The time-series evidence," *European economic review*, 46, 1125–1151.

Graham, B. S. (2017): "An econometric model of network formation with degree heterogeneity," *Econometrica*, 85, 1033–1063.

——— (2020): "Dyadic regression," in *The Econometric Analysis of Network Data*, Elsevier, 23–40.

Helpman, E., M. Melitz, and Y. Rubinstein (2008): "Estimating trade flows: Trading partners and trading volumes," *The quarterly journal of economics*, 123, 441–487.

Jochmans, K. (2018): "Semiparametric analysis of network formation," *Journal of Business & Economic Statistics*, 36, 705–713.

Koenker, R., S. Leorato, and F. Peracchi (2013): "Distributional vs. quantile regression," .

Melitz, M. J. and S. J. Redding (2015): "New trade models, new welfare implications," *American Economic Review*, 105, 1105–1146.

Newey, W. K. and D. McFadden (1994): "Chapter 36 large sample estimation and hypothesis testing. volume 4 of Handbook of Econometrics," *Elsevier*, 12, 2111–2245.

Neyman, J. and E. L. Scott (1948): "Consistent estimates based on partially consistent observations," *Econometrica: Journal of the Econometric Society*, 1–32.

Peracchi, F. (2002): "On estimating conditional quantiles and distribution functions," *Computational statistics & data analysis*, 38, 433–447.

Rao, B. P. (2009): "Conditional independence, conditional mixing and conditional association," *Annals of the Institute of Statistical Mathematics*, 61, 441–460.

Rasch, G. (1960): *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.*, Nielsen & Lydiche.

Rose, A. K. (2000): "One money, one market: the effect of common currencies on trade," *Economic policy*, 15, 08–45.

Serfling, R. J. (2009): *Approximation theorems of mathematical statistics*, vol. 162, John Wiley & Sons.

Silva, J. S. and S. Tenreyro (2006): "The log of gravity," *The Review of Economics and statistics*, 88, 641–658.

Van der Vaart, A. W. (2000): *Asymptotic statistics*, vol. 3, Cambridge university press.

Yan, T., B. Jiang, S. E. Fienberg, and C. Leng (2019): "Statistical inference in a directed network model with covariates," *Journal of the American Statistical Association*, 114, 857–868.

## Appendix A. An introduction to Distribution Regression (DR)

In this section I present an introduction to the DR approach, following the initial proposal by Foresi and Peracchi (1995), and further discussed in Peracchi (2002) and Koenker et al. (2013). Consider the problem of estimating the conditional distribution of a random variable $Y$ given a vector of $X$ covariates in a standard cross-sectional case. Note that the interest is not in merely a few quantiles but in the entire conditional distribution, $F(y|x)$.

It is proposed to select $J$ distinct values $-\infty < y_1 < \cdots < y_J < \infty$ in the range of interest of $Y$ (which is related to the quantiles of the distribution of Y), and estimate $J$ functions

$F_1(x), \dots F_J(x)$, with $F_j(x) = F(y_j|x), j = 1, \dots J$. It is argued that by suitably choosing $J$ and their position, one can get a reasonably accurate description of $F(y|x)$.

If the conditional distribution of $Y$ is continuous with support on the entire real line, then at any point $x$ in the support of $X$, the sequence of conditional distribution functions must satisfy:

$$0 < F_j(x) < 1, \quad j = 1, \dots, J, \tag{A.1}$$

$$0 < F_1(x) < \cdots < F_J(x) < 1. \tag{A.2}$$

To impose the condition given by Equation A.1, it is suggested to not model $F_j(x)$ directly, but rather to estimate the log-odds $\eta_j(x) = \ln[F_j(x)/(1 - F_j(x))]$. Then, given this estimate of the $\eta_j(x)$, one can estimate the conditional distribution at the threshold $j$ by:

$$\hat{F}_j(x) = \frac{\exp \hat{\eta}_j(x)}{1 + \exp \hat{\eta}_j(x)}. \tag{A.3}$$

Let $\mathcal{H}$ be the class of functions of $x$ that are possible candidates for the log-odds ratio. Since the random variable $1\{Y \leq y_j\}$ has a Bernoulli distribution with parameter $F_j(x)$, by the definition of the cumulative conditional distribution, we can define the best Kullback-Leibler approximation $\eta_j^*(x)$ to $\eta_j(x)$ in the class of functions $\mathcal{H}$ as the minimizer of $\mathcal{K}(\eta, \eta_j) = l(\eta_j) - l(\eta)$, with:

$$\begin{aligned} l(\eta) &= \mathbb{E}[1\{Y \leq y_j\}\eta(X) - \ln(1 + \exp \eta(X))] \\ &= \mathbb{E}[F_j(X)\eta(X) - \ln(1 + \exp \eta(X))]. \end{aligned} \tag{A.4}$$

The first expectation is taken with respect to the joint distribution of $(X, Y)$, and the second with respect to the marginal distribution of $X$. Therefore, the function $\eta_j^*$ maximizes $l(\eta)$ over the class $\mathcal{H}$. If $\eta_j \in \mathcal{H}$, then $\eta_j^* = \eta_j$. Importantly, if $X$ is a scalar random variable, and $\mathcal{H}$ is the class of functions linear in $x$, then the best Kullback-Leibler approximation to $\eta_j(x)$ is of a

linear form $\eta_j^*(x) = \gamma_j + x\delta_j$, where $(\gamma_j, \delta_j)$ are such that the approximation error

$$F_j(X) - \frac{\exp n_j^*(X)}{1 + \exp n_j^*(X)}$$

has mean zero and is uncorrelated with $X$. Therefore, $\eta_j^*$ can be estimated by maximizing the sample log-likelihood:

$$L(\eta) = n^{-1} \sum_{i=1}^{n} [1\{Y_i \leq y_j\}\eta(X_i) - \ln(1 + \exp \eta(X_i))].$$

over the linear functions in the class $\mathcal{H}$. Clearly, this is obtained by fitting $J$ separate logistic regressions, one for each binary random variable $1\{Y_i \leq y_j\}, \quad j = 1, \ldots J$. Alternative specifications for the class of functions $\mathcal{H}$, for instance, non-linear specifications, entails alternative estimation methods. One caveat of this approach is that while it satisfies the condition given by Equation A.1, by modeling the log-odds ratio, it does not guarantee the monotonicity condition given by Equation A.2.

## Appendix B. The incidental parameter problem

As highlighted by Arellano and Hahn (2007) in a standard panel data regression with one way fixed effects and dimensions $i = 1, \ldots N$ and $t = 1, \ldots T$, if $T$ is fixed and $N \to \infty$, there will be an estimation error in the estimates of the fixed effects, as only a finite number $T$ of observations are available to estimate each fixed effect. As we allow for the fixed effects to be correlated with the exogenous regressors (and its distribution is left unspecified), this estimation error contaminates the estimates of the other parameters as well, as they are not informationally orthogonal. For large enough $T$, this bias should be small. However, even under $T \to \infty$ and $N \to \infty$, the fixed effects estimator will be asymptotically biased, leading to incorrect inference over the parameters and the average partial effects.

The same argument holds for the present framework of a dyadic regression with two-way fixed effects. In our panel data model, we have two dimensions: $i = 1, ..N, j = 1, \ldots N$. However,

both the dimensions grow at rate $N$. I will consider asymptotic results such that $N \to \infty$.

Note as well that for each new country in the dataset, the number of observations is increased by $2(N-1)$. Moreover, for each fixed effect in Equation 2 there are $(N-1)$ observations available for their estimation. I will now use results shown by Fernández-Val and Weidner (2016) to demonstrate how the incidental parameter problem arises in this framework, delivering consistent but asymptotic biased estimators, keeping in mind that as $N \to \infty$ both dimensions $i$ and $j$ go to infinity and also the number of observations $N(N-1)$ go to infinity.

Given the dataset of $N(N-1)$ observations $\left\{ \left( \tilde{y}_{ij}, x'_{ij} \right)' : 1 \leq i \leq N, 1 \leq j \leq N, i \neq j \right\}$ with $\tilde{y}_{ij} = \mathbb{1}\left( \tilde{y}^*_{ij} > 0 \right)$, we have that $\tilde{y}_{ij}$ is generated by the process:

$$\tilde{y}_{ij} \mid x_{ij}, \alpha_y, \gamma_y, \beta_y \sim f_Y \left( \cdot \mid x_{ij}, \alpha_y, \gamma_y, \beta_y \right)$$

where: $\alpha_y = (\alpha_{1,y}, \ldots \alpha_{N,y})$, $\gamma_y = (\gamma_{1,y}, \ldots \gamma_{N,y})$, $f_Y$ is a known probability function and $\alpha_{i,y}, \gamma_{j,y}$ are the unobserved fixed effects. I assume for simplicity a single regressorn and a single $\beta_y$. Note here that this approach is semi-parametric in the sense that is does not specify the distribution of the fixed effects or their relationship with the explanatory variables.

We can further model the conditional distribution of $y_{2,ij}$ using a single-index specification with fixed effects, since it is a binary response model:

$$\begin{aligned} f_Y \left( \tilde{y}_{ij} \mid x_{ij}, \alpha_y, \gamma_y, \beta_y \right) = & F \left( x'_{ij} \beta_y + \alpha_{i,y} + \gamma_{j,y} \right)^{\tilde{y}_{ij}} \\ & \times \left[ 1 - F \left( x'_{ij} \beta_y + \alpha_{i,y} + \gamma_{j,y} \right) \right]^{1 - \tilde{y}_{ij}}, \end{aligned}$$

where, clearly $\tilde{y}_{ij} \in \{0, 1\}$ and $F$ is a cumulative distribution function, defined to be a standard logistic.

I can then collect all the fixed effects to be estimated in the vector $\omega_{NN,y} = (\alpha_{1,y}, \ldots \alpha_{N,y}, \gamma_{1,y}, \ldots \gamma_{N,y})'$, which can be seen as a nuisance parameter vector. Then, the true values of the parameters, denoted by $\beta_{y,0}$ and $\omega_{NN,y,0}$ are the solution to the population conditional maximum likelihood

maximization:

$$\max_{(\beta_y,\omega_{NN,y})\in\mathbb{R}^{\dim\beta_y+\dim\omega_{NN,y}}} \mathbb{E}_\omega\left[\mathcal{L}\left(\beta_y,\omega_{NN,y}\right)\right]$$

with

$$\begin{aligned}&\mathcal{L}\left(\beta_y,\omega_{NN,y}\right)\\&=\left(N(N-1)\right)^{-1}\left\{\sum_{i=1}^{N}\sum_{j\neq i}\log f_Y\left(\tilde{y}_{ij}\mid x_{ij},\alpha_y,\gamma_y,\beta_y\right)-b\left(\iota_{NN}'\omega_{NN,y}\right)^2/2\right\}\end{aligned}$$

where $\mathbb{E}_\omega$ denotes the expectation with respect to the distribution of the data conditional on the unobserved effects and strictly exogenous variables, $b>0$ is an arbitrary constant, $\iota_{NN}=(1_N',-1_N')'$ and $1_N$ denotes a vector of ones of dimension $N$.

The second term of $\mathcal{L}$ relates to a penalty that imposes a normalization to identify the fixed effects in models with two-way fixed effects that enter in the log-likelihood function as $\alpha_{i,y}+\gamma_{j,y}$. To be more specific, in this case, adding a constant to all $\alpha_{i,y}$ and subtracting the same constant from all $\gamma_{j,y}$ would not change $\alpha_{i,y}+\gamma_{j,y}$. Thus, without this normalization, the parameters $\alpha_{i,y}$ and $\gamma_{j,y}$ are not identifiable.

To estimate the parameters, we solve the sample analogue of the following equation:

$$\max_{(\beta_y,\omega_{NN,y})\in\mathbb{R}^{\dim\beta_y+\dim\omega_{NN,y}}} \mathcal{L}\left(\beta_y,\omega_{NN,y}\right)$$

In order to analyze the statistical properties of $\beta_y$, we first concentrate out the nuisance parameters $\omega_{NN,y}$, such that for given $\beta_y$, the optimal $\hat{\omega}_{NN,y}\left(\beta_y\right)$ is:

$$\hat{\omega}_{NN,y}\left(\beta_y\right)=\operatorname{argmax}_{\omega_{NN,y}\in\mathbb{R}^{\dim\omega_{NN,y}}}\mathcal{L}\left(\beta_y,\omega_{NN,y}\right)$$

Thus, the fixed effects estimator of $\beta_y$ and $\omega_{NN,y}$ are, by plugging in the previous expression for $\hat{\omega}_{NN,y}\left(\beta_y\right)$ :

$$\hat{\beta}_y=\operatorname{argmax}_{\beta_y\in\mathbb{R}^{\dim\beta_y}}\mathcal{L}\left(\beta_y,\hat{\omega}_{NN,y}\left(\beta_y\right)\right) \tag{B.1}$$

$$\hat{\omega}_{NN,y}\left(\beta_y\right)=\hat{\omega}_{NN,y}\left(\hat{\beta}_y\right) \tag{B.2}$$

The source of the problem is that the dimension of the nuisance parameters $\omega_{NN,y}$ increases with the sample size under asymptotic approximations where $N \to \infty$. To further describe the incidental parameter problem, denote:

$$\bar{\beta}_y = \text{argmax}_{\beta_y \in \mathbb{R}^{\dim \beta_y}} \mathbb{E}_\omega \left[ \mathcal{L} \left( \beta_y, \hat{\omega}_{NN,y} \left( \beta_y \right) \right) \right]$$

Using an asymptotic expansion for smooth likelihoods under appropriate regularity conditions, provided by Fernández-Val and Weidner (2016), we have that:

$$\bar{\beta}_y = \beta_{y,0} + \frac{\bar{B}_\infty}{(N-1)} + \frac{\bar{D}_\infty}{(N-1)} + \text{o}_P\left((N-1)^{-1}\right) \tag{B.3}$$

For some constants $\bar{B}_\infty$ and $\bar{D}_\infty$. The derivation for this expression can be found in the Appendix of Fernández-Val and Weidner (2016). As explained by the authors, the expansion is obtained by first taking a firstorder Taylor expansion of the Equation B.1 around the true value $\beta_{y,0}$, as it is usually done to obtain the asymptotic properties of such estimator. Then, one should additionally take a second-order Taylor expansion of the obtained term $\frac{\partial \mathcal{L}(\beta_{y,0}, \hat{\omega}_{NN})}{\partial \beta_y}$ around the true values of the nuisance terms. Intuitively, this second step demonstrates how the estimates of the fixed effects affect the estimates of the structural parameter $\beta_y$. To obtain the exact form of the expressions $\bar{B}_\infty$ and $\bar{D}_\infty$ a quite involved derivation is needed. However, this is not the focus of our study, since we show later that there are other possibilities to correct for the asymptotic bias generated by these terms other than deriving the biases themselves.

Moreover, by the properties of the maximum likelihood estimator we have that, under regularity conditions:

$$\sqrt{N(N-1)} \left( \hat{\beta}_y - \bar{\beta}_y \right) \xrightarrow{d} N\left(0, \bar{V}_{B\infty}\right)$$

For some $\bar{V}_{B\infty}$. By substituting the expression for $\beta_{y,0}$ obtained in Equation B.3, we obtain that,

by Slutsky's theorem:

$$
\begin{aligned}
&\sqrt{N(N-1)}\left(\hat{\beta}_y - \beta_{y,0}\right) \\
&= \sqrt{N(N-1)}\left(\hat{\beta}_y - \bar{\beta}_y\right) \\
&+ \sqrt{N(N-1)}\left(\frac{\bar{B}_\infty}{(N-1)} + \frac{\bar{D}_\infty}{(N-1)} + \mathrm{o}_P\left((N-1)^{-1}\right)\right) \\
&\xrightarrow{d} N\left(\bar{B}_\infty + \bar{D}_\infty, \bar{V}_{B\infty}\right)
\end{aligned}
$$

We can see from Equation B.3 that, as $N \to \infty, \hat{\beta}_y \xrightarrow{p} \beta_{y,0}$ ($\beta_{y,0}$ being the true value of the parameter), thus, the estimates of $\beta_{y,0}$ are consistent. However, from the equation above we see that the estimates converge to a distribution that is not centered at zero, which leads to incorrect asymptotic confidence intervals. This demonstrates the incidental parameters problem, that boils down to an asymptotic bias in the estimates of $\beta_{y,0}$. This asymptotic bias arises as the order of the bias is higher than the inverse of the sample size because of the small rate of convergence of the fixed effects.

## Appendix C. Sufficient statistics

In this section, I provide a proof that $\sum_{j=1}^{N} \tilde{y}_{ij}$ and $\sum_{i=1}^{N} \tilde{y}_{ij}$ are indeed the sufficient statistics for $\boldsymbol{\alpha}_{i,y}$ and $\boldsymbol{\gamma}_{j,y}$. In the following, for the sake of simplification of notation, I omit the subscript $y$ that denotes the threshold of the outcome variable. Denoting by $\tilde{\boldsymbol{Y}}$ the vector of all observations $(\tilde{y}_{11}, \dots \tilde{y}_{IJ})$; $\boldsymbol{r}$ the vector of sums of rows $(r_1, \dots, r_I)$ where $r_i = \sum_{j=1}^{J} \tilde{y}_{ij}$; $\boldsymbol{c}$ the vector of sums of columsn $(c_1, \dots, c_J)$ where $c_j = \sum_{i=1}^{I} \tilde{y}_{ij}$; and $\boldsymbol{x}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ the vectors of covariates and fixed effects, we have that:

$$
Pr[\tilde{\boldsymbol{Y}} \mid \boldsymbol{r}, \boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}; \boldsymbol{\beta}_y] = \frac{Pr[\tilde{\boldsymbol{Y}} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}; \boldsymbol{\beta}_y]}{Pr[\boldsymbol{r}, \boldsymbol{c} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}; \boldsymbol{\beta}_y]},
$$

with $Pr[\boldsymbol{r}, \boldsymbol{c} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}; \boldsymbol{\beta}_y] = \sum_{\bar{\tilde{\boldsymbol{Y}}} \in Q} Pr[\tilde{\boldsymbol{Y}} = \bar{\tilde{\boldsymbol{Y}}} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}; \boldsymbol{\beta}_y]$, where $Q$ is the set of all possible combinarions of $\tilde{y}_{ij}$ in $\tilde{\boldsymbol{Y}}$ such that the sum of the rows is given by $\boldsymbol{r}$ and the sum of columns by $\boldsymbol{c}$.

Following the proposed model for the constructed binary variables $\tilde{y}_{ij}$:

$$\tilde{y}_{ij} = 1\left\{\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y + \alpha_i + \gamma_j + \varepsilon_{ij} \geq 0\right\} \quad i = 1, \ldots, I, j = 1, \ldots, J$$

we have that:

$$Pr[\tilde{y}_{ij} \mid \boldsymbol{x}_{ij}, \alpha_i, \gamma_j] = \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y + \alpha_i + \gamma_j)^{\tilde{y}_{ij}}}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y + \alpha_i + \gamma_j)}.$$

Therefore, the joint probability of all the outcomes, conditional on $\sum_{j=1}^{N} y_{ij}$ and $\sum_{i=1}^{N} y_{ij}$ is:

$$\begin{aligned} Pr[\tilde{\boldsymbol{Y}} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}; \boldsymbol{\beta}_y] &= \prod_{i \neq j} Pr[\tilde{y}_{ij} \mid \boldsymbol{x}_{ij}, \alpha_i, \gamma_j; \boldsymbol{\beta}_y] \\ &= \frac{\exp(\sum_{i \neq j} \tilde{y}_{ij}\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y + \sum_{i \neq j} \tilde{y}_{ij}(\alpha_i + \gamma_j))}{\prod_{i \neq j}[1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y + \alpha_i + \gamma_j)]} \\ &= \frac{\exp(\sum_{i \neq j} \tilde{y}_{ij}\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y) \exp(\sum_{i=1}^{I} \alpha_i \sum_{j=1}^{J} \tilde{y}_{ij}) \exp(\sum_{j=1}^{J} \gamma_j \sum_{i=1}^{I} \tilde{y}_{ij})}{\prod_{i \neq j}[1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y + \alpha_i + \gamma_j)]} \end{aligned}$$

And analogously for $Pr[\bar{\tilde{\boldsymbol{Y}}} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}; \boldsymbol{\beta}_y]$. Then, we can write:

$$\frac{Pr[\tilde{\boldsymbol{Y}} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}; \boldsymbol{\beta}_y]}{\sum_{\bar{\tilde{Y}} \in Q} Pr[\bar{\tilde{\boldsymbol{Y}}} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}; \boldsymbol{\beta}_y]} = \frac{\exp(\sum_{i \neq j} \tilde{y}_{ij}\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y) \exp(\sum_{i=1}^{I} \alpha_i \sum_{j=1}^{J} \tilde{y}_{ij}) \exp(\sum_{j=1}^{J} \gamma_j \sum_{i=1}^{I} \tilde{y}_{ij})}{\sum_{\bar{\tilde{Y}} \in Q} \exp(\sum_{i \neq j} \bar{\tilde{y}}_{ij}\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y) \exp(\sum_{i=1}^{I} \alpha_i \sum_{j=1}^{J} \bar{\tilde{y}}_{ij}) \exp(\sum_{j=1}^{J} \gamma_j \sum_{i=1}^{I} \bar{\tilde{y}}_{ij})}$$

Finally, independently of which set in Q we consider, we have that, by the construction of the set, $\sum_{i=1}^{I} \bar{\tilde{y}}_{ij} = \sum_{i=1}^{I} \tilde{y}_{ij}$ and $\sum_{j=1}^{J} \bar{\tilde{y}}_{ij} = \sum_{j=1}^{J} \tilde{y}_{ij}$, such that:

$$\frac{Pr[\tilde{\boldsymbol{Y}} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}; \boldsymbol{\beta}_y]}{\sum_{\bar{\tilde{Y}} \in Q} Pr[\bar{\tilde{\boldsymbol{Y}}} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}; \boldsymbol{\beta}_y]} = \frac{\exp(\sum_{i \neq j} \tilde{y}_{ij}\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y)}{\sum_{\bar{\tilde{Y}} \in Q} \exp(\sum_{i \neq j} \bar{\tilde{y}}_{ij}\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y)}$$

which does not depend on the fixed effects, rendering the result.

## Appendix D. Identification

A loose intuition for the identification of the common parameters is analogous to that of Graham (2017) for the undirected case. The heterogeneity parameters (fixed effects) account for the in-degree and out-degree distributions of the network (the quantity of one's for a given

node when the node is a sender or a receiver). Therefore, the precise location of the ones (or links) is driven by the variation provided by the covariates and the common parameters ($\boldsymbol{x}'_{ij}\boldsymbol{\beta}_y$). Thus, conditioning on the set $\{\tilde{y}_{ij} + \tilde{y}_{ik} = 1, \tilde{y}_{lj} + \tilde{y}_{lk} = 1, \tilde{y}_{ij} + \tilde{y}_{lk} = 1\}$ provides the ground for an estimator that is based on the relative probability of different types of subgraphs configurations with identical degree sequences, giving the necessary variation to identify the common parameters.

For instance, assuming that the dashed wirings are not present in the above Subgraphs 1 and 2 in Figure 1 provide the same contribution of the unobserved heterogeneity to the likelihood, such that the conditional frequency to which each is observed depends only on the variation given by the covariates associated with each. In other words, in conditioning on the degree sequences of tetrads (since they are the same in both subgraphs), the only variation is the location of the links. This intuition aligns with the fact that the sums across each dimension are sufficient statistics for the fixed effects. At the same time, the conditioning events guarantee that for a node there is variation in the outcomes such that the common parameters can be identified. This feature cannot be seen in Figure 2, where Subgraph 2 is not informative to the likelihood, and the outcomes for all nodes do not present variation (i.e., a node is always sending a link or never sending a link in each Subgraph).

## Appendix E. Proofs

### *Appendix E.1. Proof of Theorem 1*

The derivation of this Theorem follows very closely that of Jochmans (2018), including only some minor modifications and more details. Using Lemma 2.2. in Newey and McFadden (1994), we can show that the limit of objective function

$$\lim_{N\to\infty} (M_n p_n)^{-1} \mathbb{E}(L_n(\boldsymbol{\beta}_y))$$

has a unique maximizer $\boldsymbol{\beta}_{y,0}$ on $\Theta$. We start by showing that the partial-likelihood contributions are bounded.

The partial-likelihood contributions are given by:

$$l_\sigma(\boldsymbol{\beta}_y) = \mathbb{1}\{z_\sigma = 1\} \log(\Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y)) + \mathbb{1}\{z_\sigma = -1\} \log(1 - \Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y)).$$

And thus:

$$|l_\sigma(\boldsymbol{\beta}_y)| \leq |\log(\Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y))| + |\log(-\Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y))|.$$

From a mean-value expansion around $\boldsymbol{\beta}_y = \mathbf{0}$:

$$\begin{aligned} |\log(\Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y))| =& |\log(\Lambda(0)) + \lambda(\boldsymbol{r}'_\sigma \tilde{\boldsymbol{\beta}}_y) \boldsymbol{r}'_\sigma \boldsymbol{\beta}_y| \\ \leq& |\log(\Lambda(0))| + C(1 + |\boldsymbol{r}'_\sigma \tilde{\boldsymbol{\beta}}_y|) |\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y| \\ \leq& |\log(\Lambda(0))| + C(1 + \|\boldsymbol{r}_\sigma\| \|\boldsymbol{\beta}_y\|) \|\boldsymbol{r}_\sigma\| \|\boldsymbol{\beta}_y\|, \end{aligned}$$

and, since $1 - \Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y) = \Lambda(-\boldsymbol{r}'_\sigma \boldsymbol{\beta}_y)$ and $\tilde{y}$ is bounded:

$$|l_\sigma(\boldsymbol{\beta}_y)| \leq 2 \left[ |\log(\Lambda(0))| + C(1 + \|\boldsymbol{r}_\sigma\| \|\boldsymbol{\beta}_y\|) \|\boldsymbol{r}_\sigma\| \|\boldsymbol{\beta}_y\| \right]. \tag{E.1}$$

The existence of second moments of $\boldsymbol{r}_\sigma$ given by Assumption 4.4., implies that $\mathbb{E}[|l_\sigma(\boldsymbol{\beta}_y|]$ is finite. Together $\boldsymbol{\beta}_y$, Newey and McFadden (1994, Lemma 2.2) applies.

Because the limit of the objective function is concave, $\hat{\boldsymbol{\beta}}_y \xrightarrow{p} \boldsymbol{\beta}_{y,0}$ will follow from a pointwise convergence in probability of $(M_n^*)^{-1} L_n(\boldsymbol{\beta}_y)$, which is the normalized objective function, to $(M_n p_n)^{-1} \mathbb{E}(L_n(\boldsymbol{\beta}_y))$, following Newey and McFadden (1994, Theorem 2.7).

By writing:

$$L_n(\boldsymbol{\beta}_y) = \sum_{\sigma \in \mathcal{M}_n} l_\sigma(\boldsymbol{\beta}_y),$$

where $l_\sigma(\boldsymbol{\beta}_y)$ is the log-likelihood contribution of quadruple $\sigma$. Then, noticing that:

$$\frac{L_n(\boldsymbol{\beta}_y)}{M_n^*} - \frac{\mathbb{E}(L_n(\boldsymbol{\beta}_y))}{\mathbb{E}(M_n^*)} = \frac{\sum_{\sigma \in \mathcal{M}_n} l_\sigma(\boldsymbol{\beta}_y) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))}{\mathbb{E}(M_n^*)} + \frac{\sum_{\sigma \in \mathcal{M}_n} l_\sigma(\boldsymbol{\beta}_y)}{\mathbb{E}(M_n^*)} \left( \frac{\mathbb{E}(M_n^*)}{M_n^*} - 1 \right),$$

it is sufficient to show that each of the terms in the RHS of this expression converges to zero in

probability.

For the first term in the RHS, using Equation (E.1), because $\mathbb{E}(\|\boldsymbol{r}_\sigma\|^2)$ is finite by Assumption 4.4, and $\Theta$ is compact, it follows that the variance of $l_\sigma(\boldsymbol{\beta}_y)$ exists and is uniformly bounded in $\sigma$. Then, by Chebyshev's inequality, it holds that for any $\epsilon > 0$:

$$\Pr\left(\left|\frac{\sum_{\sigma\in\mathcal{M}_n} l_\sigma(\boldsymbol{\beta}_y) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))}{\mathbb{E}(M_n^*)}\right| > \epsilon\right) \leq \frac{1}{\epsilon^2}\frac{\mathbb{E}\left(\left|\sum_{\sigma\in\mathcal{M}_n} l_\sigma(\boldsymbol{\beta}_y) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))\right|^2\right)}{\mathbb{E}(M_n^*)^2},$$

for each $\boldsymbol{\beta}_y \in \Theta$. Also,

$$\mathbb{E}\left(\left|\sum_{\sigma\in\mathcal{M}_n} l_\sigma(\boldsymbol{\beta}_y) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))\right|^2\right) = \mathbb{E}\left(\left(\sum_{\sigma\in\mathcal{M}_n} l_\sigma(\boldsymbol{\beta}_y) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))\right)\left(\sum_{\sigma'\in\mathcal{M}_n} l_{\sigma'}(\boldsymbol{\beta}_y) - \mathbb{E}(l_{\sigma'}(\boldsymbol{\beta}_y))\right)\right).$$

Notice that a pair of quadruples $\sigma = \sigma\{i,l;j,k\}$ and $\sigma' = \sigma\{i',l';j',k'\}$ only contributes to the covariance if they share at least one node in common, i.e., a pair of quadruples with only distinct nodes are independent by Assumption 4.1.

More specifically, the product above contains $O(N^8)$ terms. Consider $\sigma = \sigma\{i,l;j,k\}$ and $\sigma\{i',l';j',k'\}$ having indices in common, and therefore, having dependence and contributing to the covariance. If they share the index $i = i'$, there are $\binom{N-1}{2}$ choices for $l$ and $l'$ to be distinct, and $\binom{N-2}{4}$ choices of $j, j', k, k'$ to be distinct. Since there are $N$ different combinations for $i = i'$, the number of quadruples that share at least one node in common is of the order $O(N^7)$ (given by the fact that the quadruples that share two or more nodes in common are of smaller order than the ones sharing one node in common). This also implies that for a quadruple $\sigma = \sigma\{i,l;j,k\}$ there are $O(N^3)$ other quadruples sharing at least one index. By writing $\mathbb{1}\{\sigma\cap\sigma' \neq \emptyset\}$ to indicate

the quadruples that share common indices:

$$
\begin{aligned}
&\mathbb{E}\left(\left(\sum_{\sigma\in\mathcal{M}_n} l_\sigma(\boldsymbol{\beta}_y) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))\right)\left(\sum_{\sigma'\in\mathcal{M}_n} l_{\sigma'}(\boldsymbol{\beta}_y) - \mathbb{E}(l_{\sigma'}(\boldsymbol{\beta}_y))\right)\right) \\
&= \sum_{\sigma\in\mathcal{M}_n}\sum_{\sigma'\in\mathcal{M}_n} \mathbb{E}\left(\left(l_\sigma(\boldsymbol{\beta}_y) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))\right)\left(l_{\sigma'}(\boldsymbol{\beta}_y) - \mathbb{E}(l_{\sigma'}(\boldsymbol{\beta}_y))\right)\right) \\
&= \sum_{\sigma\in\mathcal{M}_n}\sum_{\sigma'\in\mathcal{M}_n} \mathbb{1}\{\sigma\cap\sigma'\neq\emptyset\}\mathbb{E}\left(\left(l_\sigma(\boldsymbol{\beta}_y) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))\right)\left(l_{\sigma'}(\boldsymbol{\beta}_y) - \mathbb{E}(l_{\sigma'}(\boldsymbol{\beta}_y))\right)\right) \\
&\leq \sum_{\sigma\in\mathcal{M}_n}\sum_{\sigma'\in\mathcal{M}_n} \mathbb{1}\{\sigma\cap\sigma'\neq\emptyset\}\sqrt{\mathrm{Var}(l_\sigma(\boldsymbol{\beta}_y)) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))}\sqrt{\mathrm{Var}(l_{\sigma'}(\boldsymbol{\beta}_y)) - \mathbb{E}(l_{\sigma'}(\boldsymbol{\beta}_y))} \\
&= O(N^3 M_n p_n), \qquad \text{(E.2)}
\end{aligned}
$$

where the inequality follows from the fact that the variance is bounded, and the last equality follows from the fact that, in expectation, there are $M_n p_n$ informative quadruples, and $O(N^3)$ quadruples in $\sigma'$ sharing at least one index with $\sigma$.

Moreover, as $\mathbb{E}(M_n^*) = M_n p_n$ and $M_n = O(N^4)$, we find that:

$$
\frac{\mathbb{E}\left(\left|\sum_{\sigma\in\mathcal{M}_n} l_\sigma(\boldsymbol{\beta}_y) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))\right|^2\right)}{\mathbb{E}(M_n^*)^2} = O\left(\frac{1}{Np_n}\right),
$$

which converges to zero by Assumption 4.4, as $p_n \in (0,1)$ and $N \to \infty$. Therefore:

$$
\lim_{N\to\infty} \Pr\left(\left|\frac{\sum_{\sigma\in\mathcal{M}} l_\sigma(\boldsymbol{\beta}_y) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))}{\mathbb{E}(M_n^*)}\right| > \epsilon\right) \leq \frac{1}{\epsilon^2}\frac{\mathbb{E}\left(\left|\sum_{\sigma\in\mathcal{M}} l_\sigma(\boldsymbol{\beta}_y) - \mathbb{E}(l_\sigma(\boldsymbol{\beta}_y))\right|^2\right)}{\mathbb{E}(M_n^*)^2} = 0
$$

for any $\epsilon > 0$ and all $\boldsymbol{\beta}_y \in \Theta$.

For the second term in the RHS, the summands are bounded uniformly in $\sigma$ and do not depend on $\boldsymbol{\beta}_y$. Following the same arguments as before, it is easy to verify that $\left(\frac{M_n^*}{M_n} - p_n\right) \xrightarrow{p} 0$, and therefore, $\frac{M_n^*}{\mathbb{E}(M_n^*)} \xrightarrow{p} 1$.

*Appendix E.2. Proof of Theorem 2*

We first rewrite the score vector as:

$$\boldsymbol{S}_n(\boldsymbol{\beta}_y) = \sum_{i}^{N} \sum_{j \neq i} \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \underbrace{\frac{\partial}{\partial \boldsymbol{\beta}_y} l_\sigma(\boldsymbol{\beta}_y)}_{\boldsymbol{s}_\sigma}, \tag{E.3}$$

where, as before,

$$s(\sigma; \boldsymbol{\beta}_y) = \boldsymbol{r}_\sigma \{1\{z_\sigma = 1\}(1 - \Lambda(\boldsymbol{r}_\sigma' \boldsymbol{\beta}_y)) - 1\{z_\sigma = -1\}\Lambda(\boldsymbol{r}_\sigma' \boldsymbol{\beta}_y)\} \tag{E.4}$$

Notice that twice as many elements than are needed are being summed over, due to the permutation invariance of senders and receivers, but I apply the similar summation over the elements of the Hessian in the following. Thus the normalization will be correct.

**Step 1.**

The first step of the proof consists on defining a projection of the score vector. Defining the information set $\mathcal{F}_n = \{\{\boldsymbol{x}_{ij}\}_{NN}, \{\alpha_{i,y}\}_N, \{\gamma_{j,y}\}_N\}$, the projection of the score is given by:

$$\begin{aligned} \boldsymbol{U}_n(\boldsymbol{\beta}_{y,0}) &= \sum_{i}^{N} \sum_{j \neq i} \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \mathbb{E}(\boldsymbol{s}(\sigma\{i, i'; j, j'\}; \boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij}, \mathcal{F}_n) \\ &= \sum_{i} \sum_{j \neq i} \boldsymbol{v}_{ij}(\boldsymbol{\beta}_{y,0}) \end{aligned} \tag{E.5}$$

where $\boldsymbol{v}_{ij}(\boldsymbol{\beta}_{y,0}) = \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \mathbb{E}(\boldsymbol{s}(\sigma\{i, i'; j, j'\}, \boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij}, \mathcal{F}_n)$. Note that while this projection resembles a Hájek projection, as in (Serfling 2009), it is not formally one, since (i) the kernel is not symmetric in the arguments, and (ii) the conditioning terms are more involved - since they are not only related to the dyad $i, j$.

Before moving to the next steps of the proof, I present some intermediate results that will become relevant.

*Intermediate result 1.*

Note that $\mathbb{E}(\boldsymbol{s}(\sigma; \boldsymbol{\beta}_{y,0})) = 0$, since, from sufficiency, we have that, by defining $\mathcal{F}_\sigma$ to be the

collection of covariates and fixed effects for the nodes in the quadruple $\sigma$:

$$Pr(z_\sigma = 1 \mid \mathcal{F}_\sigma) = \Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_{y,0}) Pr(z_\sigma\{-1,1\} \mid \mathcal{F}_\sigma)$$

$$Pr(z_\sigma = -1 \mid \mathcal{F}_\sigma) = (1 - \Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_{y,0})) Pr(z_\sigma\{-1,1\} \mid \mathcal{F}_\sigma)$$

Then,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{s}(\sigma, \boldsymbol{\beta}_{y,0})] &= \mathbb{E}\left[\mathbb{E}[\mathbb{1}\{z_\sigma = 1\}(1 - \Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_{y,0}))\boldsymbol{r}'_\sigma - \mathbb{1}\{z_\sigma = -1\}\Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_{y,0})\boldsymbol{r}'_\sigma \mid \mathcal{F}_\sigma]\right] \\
&= \mathbb{E}\left[\mathbb{E}[\mathbb{1}\{z_\sigma = 1\} \mid \mathcal{F}_\sigma](1 - \Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_{y,0}))\boldsymbol{r}'_\sigma - \mathbb{E}[\mathbb{1}\{z_\sigma = -1\} \mid \mathcal{F}_\sigma]\Lambda(\boldsymbol{r}'_\sigma \boldsymbol{\beta}_{y,0})\boldsymbol{r}'_\sigma\right] \\
&= 0. \qquad \text{(E.6)}
\end{aligned}$$

*Intermediate result 2.*

From the first intermediate result, it also follows that:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{v}_{ij}(\boldsymbol{\beta}_{y,0})] &= \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \mathbb{E}\left[\mathbb{E}[\boldsymbol{s}(\sigma\{i,i';j,j'\}, \boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij}, \mathcal{F}_n]\right] \\
&= \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \mathbb{E}[\boldsymbol{s}(\sigma\{i,i';j,j'\}, \boldsymbol{\beta}_{y,0})] = 0 \qquad \text{(E.7)}
\end{aligned}$$

And, importantly, from this result, it follows that $\mathbb{E}[\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})] = 0$.

*Intermediate result 3.*

First, note that, similarly to Graham (2017) link decisions are conditionally independent, that is, considering nodes $i$, $j$, and $k$, conditional on these agents observed and unobserved characteristics, the events that *$i$ and $j$ are connected*, *$i$ and $k$ are connected*, and *$j$ and $k$ are connected* are independent of each other. Then, we have that:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{v}_{ij}(\boldsymbol{\beta}_{y,0})\boldsymbol{v}_{i'j'}(\boldsymbol{\beta}_{y,0})] &= \mathbb{E}\left[\mathbb{E}[\boldsymbol{v}_{ij}(\boldsymbol{\beta}_{y,0})\boldsymbol{v}_{i'j'}(\boldsymbol{\beta}_{y,0}) \mid \mathcal{F}_n]\right] \\
&= \mathbb{E}\left[\mathbb{E}[\boldsymbol{v}_{ij}(\boldsymbol{\beta}_{y,0}) \mid \mathcal{F}_n]\mathbb{E}[\boldsymbol{v}_{i'j'}(\boldsymbol{\beta}_{y,0}) \mid \mathcal{F}_n]\right] = 0 \qquad \text{(E.8)}
\end{aligned}$$

unless $i = i'$ and $j = j'$.

*Intermediate result 4.*

Also from conditional independence of links, $\mathbb{E}[\boldsymbol{s}_\sigma(\boldsymbol{\beta}_{y,0})\boldsymbol{s}_{\sigma'}(\boldsymbol{\beta}_{y,0})'] = \mathbb{E}\left[\mathbb{E}[\boldsymbol{s}_\sigma(\boldsymbol{\beta}_{y,0})\boldsymbol{s}_{\sigma'}(\boldsymbol{\beta}_{y,0})' \mid \mathcal{F}_n]\right] = 0$ unless $\sigma$ and $\sigma'$ share one dyad in common.

**Step 2.**

The second step of the proof consists of showing the asymptotic equivalence between the normalized score vector and the normalized projection. That is, to show that $\boldsymbol{\Upsilon}^{-1/2}\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})$ and $\boldsymbol{\Upsilon}^{-1/2}\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})$ are asymptotically equivalent, where $\boldsymbol{\Upsilon}$ is the covariance matrix of the projection.

First, note that the covariance matrix of the projection is given by, given the Intermediate results 1 and 2:

$$
\begin{aligned}
\boldsymbol{\Upsilon} &= \mathbb{E}[\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})'] \\
&= \sum_i \sum_{j \neq i} \mathbb{E}[\boldsymbol{v}_{ij}(\boldsymbol{\beta}_{y,0})\boldsymbol{v}_{i'j'}(\boldsymbol{\beta}_{y,0})] \\
&= \sum_i \sum_{j \neq i} \mathbb{E}\left[ \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \mathbb{E}[\boldsymbol{s}(\sigma\{i,i';j,j'\},\boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij}, \mathcal{F}_n] \sum_{i'' \neq i,j} \sum_{j'' \neq i,j,i''} \mathbb{E}[\boldsymbol{s}(\sigma\{i,i'';j,j''\},\boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij}, \mathcal{F}_n]' \right] \\
&= \sum_i \sum_{j \neq i} \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \sum_{i'' \neq i,j} \sum_{j'' \neq i,j,i''} \mathbb{E}\left[ \mathbb{E}[\boldsymbol{s}(\sigma\{i,i';j,j'\},\boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij}, \mathcal{F}_n] \mathbb{E}[\boldsymbol{s}(\sigma\{i,i'';j,j''\},\boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij}, \mathcal{F}_n]' \right]
\end{aligned}
\tag{E.9}
$$

This matrix is positive definite, and therefore $\boldsymbol{\Upsilon}^{1/2}$ exists and is positive definite and invertible. Furthermore, from Intermediate result 4, we know that the scores are conditionally uncorrelated, unless they share one dyad in common. Following the reasoning in Jochmans (2018), in the expression above, there are $O(N^6)$ terms with only one dyad in common, and the number of terms with more than one dyad in common in $o(N^6)$. Therefore, the leading term of $\mathbb{E}[\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})']$ is comprised of correlations between $\mathbb{E}[\boldsymbol{s}(\sigma\{i,i';j,j'\},\boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij}, \mathcal{F}_n]$ and $\mathbb{E}[\boldsymbol{s}(\sigma\{i,i'';j,j''\},\boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij}, \mathcal{F}_n]$ for which quadruples $\{i,i';j,j'\}$, $\{i,i'';j,j''\}$ share exactly one dyad in common.

Moreover, given that, conditional on $\mathcal{F}_n$ and $\tilde{y}_{ij}$, $\boldsymbol{s}(\sigma\{i,i';j,j'\},\boldsymbol{\beta}_{y,0})$ and $\boldsymbol{s}(\sigma\{i,i'';j,j''\},\boldsymbol{\beta}_{y,0})$

are independent if $i' \neq i''$ and $j' \neq j''$, we can characterize the leading term as:

$$\begin{aligned}
\boldsymbol{\Upsilon}_l &= \sum_i \sum_{j \neq i} \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \sum_{i'' \neq i,j,i',j'} \sum_{j'' \neq i,j,i',j',i''} \mathbb{E}\left[\mathbb{E}[\boldsymbol{s}(\sigma\{i,i';j,j'\},\boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij},\mathcal{F}_n]\mathbb{E}[\boldsymbol{s}(\sigma\{i,i'';j,j''\},\boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij},\mathcal{F}_n]'\right] \\
&= \sum_i \sum_{j \neq i} \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \sum_{i'' \neq i,j,i',j'} \sum_{j'' \neq i,j,i',j',i''} \mathbb{E}\left[\mathbb{E}[\boldsymbol{s}(\sigma\{i,i';j,j'\},\boldsymbol{\beta}_{y,0})\boldsymbol{s}(\sigma\{i,i'';j,j''\},\boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij},\mathcal{F}_n]'\right] \\
&= \sum_i \sum_{j \neq i} \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \sum_{i'' \neq i,j,i',j'} \sum_{j'' \neq i,j,i',j',i''} \mathbb{E}\left[\boldsymbol{s}(\sigma\{i,i';j,j'\},\boldsymbol{\beta}_{y,0})\boldsymbol{s}(\sigma\{i,i'';j,j''\},\boldsymbol{\beta}_{y,0})'\right]
\end{aligned} \tag{E.10}$$

where the second equality follows from conditional independence. This matrix is positive definite, and therefore $\boldsymbol{\Upsilon}_l^{1/2}$ exists and is positive definite and invertible.

To show asymptotic equivalence, we want to establish that, following Van der Vaart (2000):

$$\lim_{N \to \infty} \Upsilon^{-1/2} \mathbb{E}\left[(\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0}) - \boldsymbol{S}_n(\boldsymbol{\beta}_{y,0}))(\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0}) - \boldsymbol{S}_n(\boldsymbol{\beta}_{y,0}))'\right] \Upsilon^{-1/2} = 0.$$

Or,

$$\lim_{N \to \infty} \Upsilon^{-1/2} \mathbb{E}\left[\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})' - \boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})' - \boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})' + \boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})'\right] \Upsilon^{-1/2} = 0.$$

The next step is to show that the (leading term of the) covariance matrix of the score $\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})$ is the same as the (leading term of the) covariance matrix for the projection, that is given above. The covariance of the score is given by $\mathbb{E}[\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})']$. However, by similar arguments as before, we can focus on the leading term, which comprises of quadruples sharing one dyad in common. By symmetry of the scores in the sender and receiver nodes, we can fix this to be the first sender-receiver $ij$ dyad and multiply the expression for the scores by 4. The leading term of $\mathbb{E}[\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})']$ is then given by:

$$\Upsilon_{s,l} = \sum_i \sum_{j \neq i} \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \sum_{i'' \neq i,j,i',j'} \sum_{j'' \neq i,j,i',j',i''} 16 \times \mathbb{E}\left[\boldsymbol{s}(\sigma\{i,i';j,j'\},\boldsymbol{\beta}_{y,0})\boldsymbol{s}(\sigma\{i,i'';j,j''\},\boldsymbol{\beta}_{y,0})'\right] \tag{E.11}$$

Thus, we immediately have that $\Upsilon_{s,l} = 16 \times \Upsilon_l$. It also follows not only that $\Upsilon^{-1/2}\mathbb{E}[\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})']\Upsilon^{-1/2} = I + o(1)$, but also from the equivalence of the leading terms, it also follows that $\Upsilon^{-1/2}\mathbb{E}[\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})']\Upsilon^{-1/2} = I + o(1)$. We can also see that, given:

$$\mathbb{E}[\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})'] = \sum_i \sum_{j \neq i} \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \sum_{i'' \neq i,j} \sum_{j'' \neq i,j,i''} \mathbb{E}[\boldsymbol{s}(\sigma\{i,i';j,j'\},\boldsymbol{\beta}_{y,0})]\mathbb{E}[\boldsymbol{s}(\sigma\{i,i'';j,j''\},\boldsymbol{\beta}_{y,0}) \mid \tilde{y}_{ij}, \mathcal{F}_n]'$$

and applying the same reasoning as before, $\Upsilon^{-1/2}\mathbb{E}[\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})']\Upsilon^{-1/2} = I + o(1)$, and analogously, we have that $\Upsilon^{-1/2}\mathbb{E}[\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0})\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0})']\Upsilon^{-1/2} = I + o(1)$, which proves the asymptotic equivalence above.

*The next steps follow exactly Jochmans (2018), but they are illustrated below.*

**Step 3.**

Recall that $\boldsymbol{v}_{ij}$ are zero mean and independent conditional on the sequence of covariates (this relaxes the intermediate result 3, since the fixed effects are differenced out). Let

$$\boldsymbol{\Upsilon}_X = \sum_{i=1}^{N} \sum_{j \neq i} E\left(\boldsymbol{v}_{ij}\boldsymbol{v}_{ij}' \mid \{\boldsymbol{x}_{ij}\}_{N,N}\right).$$

By a conditional version of the Lyapunov's CLT (Rao 2009):

$$\boldsymbol{\Upsilon}_X^{-1/2}\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0}) \xrightarrow{d} N(0, I)$$

conditional on the covariates. By Assumption 4.5, it is easy to see that $||\boldsymbol{\Upsilon}_X - \boldsymbol{\Upsilon}|| \xrightarrow{p} 0$. Defining a matrix $\boldsymbol{\Upsilon}_n(\hat{\boldsymbol{\beta}}_y)$ to be the plug-in estimator of $\Upsilon$ based on the matrix $\boldsymbol{\Upsilon}_{s,l}$ above, such that:

$$\boldsymbol{\Upsilon}_n(\hat{\boldsymbol{\beta}}_y) = \sum_i \sum_{j \neq i} \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} \sum_{i'' \neq i,j,i',j'} \sum_{j'' \neq i,j,i',j',i''} 16 \times \left[\boldsymbol{s}(\sigma\{i,j,i',j'\},\hat{\boldsymbol{\beta}}_y)\boldsymbol{s}(\sigma\{i,j,i'',j''\},\hat{\boldsymbol{\beta}}_y)'\right] \tag{E.12}$$

Using the same arguments as the ones in the next section to establish convergence of the nor-

malized Hessian, we can show that this estimator is consistent, and therefore:

$$\Upsilon_n(\hat{\beta}_y)^{-1/2}\boldsymbol{U}_n(\boldsymbol{\beta}_{y,0}) \xrightarrow{d} N(0,I)$$

as $N \to \infty$, by applying the Slutsky's theorem. Note that his also implies that, due to the asymptotic equivalence result:

$$\Upsilon_n(\hat{\beta}_y)^{-1/2}\boldsymbol{S}_n(\boldsymbol{\beta}_{y,0}) \xrightarrow{d} N(0,I)$$

**Step 4.**

We proceed by showing the convergence of the Hessian. Recall that the Hessian is

$$\boldsymbol{H}_n(\boldsymbol{\beta}_y) = \sum_{\sigma\in\mathcal{M}_n} \boldsymbol{r}_\sigma \boldsymbol{r}_\sigma' f\left(\boldsymbol{r}_\sigma'\theta\right) 1\left\{\boldsymbol{z}_\sigma \in \{-1,1\}\right\}.$$

We need to show that

$$\sup_{\beta_y\in\Theta} \left\| \frac{\boldsymbol{H}_n(\boldsymbol{\beta}_y)}{M_n^*} - \frac{\mathbb{E}\left(\boldsymbol{H}_n(\boldsymbol{\beta}_y)\right)}{M_n p_n} \right\| \xrightarrow{p} 0$$

as $N \to \infty$. The matrix $\lim_{N\to\infty} (M_n p_n)^{-1} E\left(\boldsymbol{H}_n(\boldsymbol{\beta}_{y,0})\right)$ is the matrix given in Assumption 4. Because we have shown in the proof of Theorem 1 that $(M_n^*/M_n - p_n) \xrightarrow{p} 0$ as $N \to \infty$ it suffices to show

$$\frac{\sup_{\beta_y\in\Theta} \left\|\boldsymbol{H}_n(\boldsymbol{\beta}_y) - E\left(\boldsymbol{H}_n(\boldsymbol{\beta}_y)\right)\right\|}{M_n p_n} \xrightarrow{p} 0$$

as $N \to \infty$. To show this we verify the conditions of Lemma 2.9 of Newey and McFadden (1994). First, a Taylor expansion gives

$$\frac{\left\|\boldsymbol{H}_n(\boldsymbol{\beta}_{y,1}) - \boldsymbol{H}_n(\boldsymbol{\beta}_{y,2})\right\|}{M_n p_n} \leq \left( (M_n p_n)^{-1} \sum_{\sigma\in\mathcal{M}_n} \|\boldsymbol{r}_\sigma\|^3 1\left\{z_\sigma \in \{-1,1\}\right\} \right) \sup_{\epsilon\in\mathcal{R}} \left| \frac{\partial f(\epsilon)}{\partial \epsilon} \right| \left\|\boldsymbol{\beta}_{y,1} - \boldsymbol{\beta}_{y,2}\right\|$$

for any $\beta_{y,1}, \beta_{y,2} \in \Theta$. Next, using the same arguments as those used to establish Theorem 1 we find that

$$(M_n p_n)^{-1} \sum_{\sigma\in\mathcal{M}_n} \|\boldsymbol{r}_\sigma\|^3 1\left\{z_\sigma \in \{-1,1\}\right\} = O_p(1)$$

where we use the moment condition in Assumption 5. Because the derivative of $f$ is bounded uniformly on $\mathcal{R}$ we obtain

$$\frac{\|\boldsymbol{H}_n(\boldsymbol{\beta}_{y,1}) - \boldsymbol{H}_n(\boldsymbol{\beta}_{y,2})\|}{M_n p_n} = O_p(1)\,\|\boldsymbol{\beta}_{y,1} - \boldsymbol{\beta}_{y,2}\|$$

for any $\boldsymbol{\beta}_{y,1}, \boldsymbol{\beta}_{y,2} \in \Theta$. Thus, the Hessian matrix is stochastically equicontinuous. This implies that uniform convergence follows from pointwise convergence on $\Theta$. Assumption 5 implies that $E\left(\|\boldsymbol{r}_\sigma\|^4 \mid z_\sigma \in \{-1, 1\}\right)$ is uniformly bounded in $\sigma$ while $f$ is bounded uniformly on $\mathcal{R}$. Therefore, the same arguments as those used to establish Theorem 1 yield the convergence result

$$\frac{\|\boldsymbol{H}_n(\boldsymbol{\beta}_y) - E\left(\boldsymbol{H}_n(\boldsymbol{\beta}_y)\right)\|}{M_n p_n} \xrightarrow{p} 0$$

for all $\boldsymbol{\beta}_y \in \Theta$. Uniform convergence has been shown.

**Step 5.**

Limit distribution of the estimator. An expansion of the first-order condition to the log-likelihood optimization problem around $\boldsymbol{\beta}_{y,0}$ together with the results obtained above yields

$$\boldsymbol{\Omega}_n^{-1/2}\left(\hat{\boldsymbol{\beta}}_y - \boldsymbol{\beta}_{y,0}\right) = -\boldsymbol{\Omega}_n^{-1/2}\boldsymbol{H}_n\left(\boldsymbol{\beta}_y^*\right)^{-1}\boldsymbol{S}_n\left(\boldsymbol{\beta}_{y,0}\right) \xrightarrow{d} N(0, I)$$

as $N \to \infty$ by an application of Slutsky's theorem. Here, $\boldsymbol{\beta}_y^* \in \Theta$ is a value that lies between $\hat{\boldsymbol{\beta}}_y$ and $\boldsymbol{\beta}_{y,0}$. This conclusion is the limit result stated in Theorem 2. The statement on the convergence rate in the theorem is implied by the fact that $\boldsymbol{\Upsilon} = O\left(N(N-1)p_n\right)$. This rate result follows from the same argument as the convergence rate of $(M_n p_n)^{-1} L_n(\beta_y)$ to its expectation in the proof of Theorem 1 given above and can readily be deduced from the expression for $\boldsymbol{\Upsilon}_{s,l}$ given above. The proof of Theorem 2 is thus complete.