SciencesPo
DEPARTMENT OF ECONOMICS

Discussion paper 2016-02

SEMIPARAMETRIC ANALYSIS OF NETWORK FORMATION

Koen Jochmans

Sciences Po Economics Discussion Papers

Semiparametric Analysis of Network Formation

Koen Jochmans*
Department of Economics, Sciences Po, Paris

(First version: September 18, 2015)

February 9, 2016

Abstract

We consider a statistical model for network formation that features both node-specific heterogeneity parameters and common parameters that reflect homophily among nodes. The goal is to perform statistical inference on the homophily parameters while allowing the distribution of the node heterogeneity to be unrestricted, that is, by treating the node-specific parameters as fixed effects. Jointly estimating all the parameters leads to asymptotic bias that renders conventional confidence intervals incorrectly centered. As an alternative, we develop an approach based on a sufficient statistic that separates inference on the homophily parameters from estimation of the fixed effects. This estimator is easy to compute and is shown to have desirable asymptotic properties. In numerical experiments we find that the asymptotic results provide a good approximation to the small-sample behavior of the estimator. As an empirical illustration, the technique is applied to explain the import and export patterns in a cross-section of countries.

Keywords: conditional inference, degree heterogeneity, directed random graph, fixed effects, homophily, U-statistic.

*Sciences Po, Département d'économie, 28 rue des Saints Pères, 75007 Paris, France
koen.jochmans@sciencespo.fr

1 Introduction

It is well recognized that network connections are important determinants of economic and social outcomes (Jackson, 2008). Therefore, it is important to understand what drives network formation. Models of link formation tend to be characterized by a large number of parameters. For example, the classic model of Holland and Leinhardt (1981), a statistical model for directed random graphs between n nodes, features $O(n)$ parameters. Thus, under asymptotics where $n \to \infty$, the number of parameters grows with the sample size. Such a problem is reminiscent of the classic incidental-parameter problem of Neyman and Scott (1948) and presents a serious challenge for statistical inference. The first theoretical results for maximum-likelihood estimation of Holland and Leinhardt type models have only been obtained recently (Fernández-Val and Weidner 2015, Yan et al. 2015). Related work by Chatterjee et al. (2011), Rinaldo et al. (2013), Yan and Xu (2013), and Graham (2015) provides similar results for the undirected version of the Holland and Leinhardt model, known as the β-model.[1]

In this paper we study a version of the Holland and Leinhardt (1981) model that incorporates a set of observable dyad characteristics along with sender- and receiver-specific effects. The motivation for this model is as in Graham (2015) and Dzemski (2014). The aim is to conduct statistical inference on the effect of dyad characteristics on the probability of link formation. This allows to investigate the importance of homophily in network formation while controlling for (degree and other) unobserved heterogeneity among senders and receivers of links. We treat the node-specific effects as fixed. While joint estimation of these effects and the common homophily parameters leads to consistent estimators as

[1] Alternative approaches aim to reduce the dimensionality of the problem and are based on stochastic blockmodels (Holland et al. 1983), finite-mixture models (Hoff et al. 2002; Vu et al. 2013), and models with random effects (van Duijn et al. 2004).

the network grows large (Yan et al. 2015), the estimator of the homophily parameters is asymptotically biased (Fernández-Val and Weidner 2015). Moreover, the maximum-likelihood estimator suffers from an incidental-parameter problem akin to the one described in Li et al. (2003) and Sartori (2003) in the context of longitudinal data analysis with stratum-specific nuisance parameters.

We build on the conditioning argument of Hirji et al. (1987) and Charbonneau (2014) to set up a statistical objective function for the homophily parameters. The objective function can be understood to be a generalization of Rasch (1960, 1961) and does not depend on the sender- and receiver-specific parameters. However, the objective function takes the form of a U-statistic in both the senders and receivers of links and so standard theory for conditional-likelihood estimators (Andersen, 1970) does not apply. Nonetheless, we show that the estimator converges to the true parameter value at the rate n^{-1} and that its asymptotic distribution is normal, with a variance that can be consistently estimated. In numerical experiments we find that the asymptotic theory provides a good approximation to the finite-sample behavior of the estimator.

As an empirical illustration we apply the estimator to investigate the importance of geographical distance and other measures of proximity—such as the participation in a preferential trade agreement, and sharing a common border and a common language—as determinants of import and export patterns observed in a cross section of 136 countries. Understanding the drivers behind these patterns of trade has recently received substantial attention in the international-trade literature (see, e.g., Helpman et al. 2008). However, the statistical methods used thus far do not properly account for the presence of the implied incidental-parameter bias. We find smaller effects (in magnitude) of dyad characteristics on the log-odds.

2 Network formation

In this section we put forth our probabilistic model of network formation. Introduce a set of n nodes, $\mathbb{N}_n = \{1, 2, \ldots, n\}$, and consider the decision of two distinct nodes i and j in $\mathbb{N}_n$ to form an edge from i to j . Let u_{ij} denote the joint surplus of the dyad (i, j) from creating an edge from i to j. Then the decision takes on the simple threshold-crossing form

$$y_{ij} = \begin{cases} 1 & \text{if } u_{ij} \geq 0 \\ 0 & \text{if } u_{ij} < 0 \end{cases}. \tag{2.1}$$

Suppose the surplus decomposes as

$$u_{ij} = x'_{ij}\theta_0 + \alpha_i + \gamma_j - \epsilon_{ij}, \tag{2.2}$$

where x_{ij} is a vector of observable attributes of the dyad and θ_0 is a parameter vector of conformable dimension, α_i and γ_j are unobserved characteristics specific to the nodes, and ϵ_{ij} is an unobserved idiosyncratic component. If the ϵ_{ij} are independent and identically distributed with distribution F, the probability of observing a link from i to j given the characteristics of the nodes becomes

$$p_{ij} = \Pr(y_{ij} = 1 | x_{ij}, \alpha_i, \gamma_j) = F(x'_{ij}\theta_0 + \alpha_i + \gamma_j).$$

Following most of the literature since Holland and Leinhardt (1981) (see, e.g., Chatterjee et al. 2011, Rinaldo et al. 2013, and Yan and Xu 2013) we will work with the logistic specification

$$F(\epsilon) = \frac{1}{1 + \exp(-\epsilon)}.$$

Our interest lies in estimation of and inference about the parameter θ_0. As the log-odds ratio is

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = F^{-1}(p_{ij}) = x'_{ij}\theta_0 + \alpha_i + \gamma_j,$$

this allows evaluating the importance of dyad characteristics on the probability that the nodes form a link between them.

The covariates in our network-formation model capture homophily between nodes. In a typical application, they will be measures of distance, similarity, or divergence between sender i and receiver j. In our trade application, they include a measure of geographical distance as well as several indicators of closeness, such as whether or not countries i and j share a common language and have established a preferential trade agreement. In the work of Jackson et al. (2012) on favor exchange among Indian villagers, the covariates measure such things as whether or not the sender and receiver are members of the same caste and have a common religion, and if they have a similar age, education level, and employment background.

The model postulated in (2.1)–(2.2) also permits rich patterns of degree heterogeneity. Moreover, unobserved node-specific factors affect edge creation in the same way as in the model of Holland and Leinhardt (1981). While link formation decisions are independent conditional on dyad and node characteristics, the distribution of $\{y_{ij}\}_n$ given $\{x_{ij}\}_n$ will exhibit large dependence due to the presence of the $\{\alpha_i, \gamma_j\}_n$. Thus, the model allows for heterogeneity in link formation among observationally-equivalent nodes, and can equally rationalize the large dependence between links across different nodes typically observed in network data.

Treating $\{\alpha_i, \gamma_j\}_n$ as random effects by specifying their distribution conditional on $\{x_{ij}\}_n$ gives a model as in van Duijn et al. (2004) and Hoff (2005). Here we aim to perform statistical inference on the determinants of network formation without making such functional-form restrictions. Thus, throughout the analysis, we treat $\{\alpha_i, \gamma_j\}_n$ as fixed effects, that is, we condition on them. Henceforth, for notational convenience, we no longer make this conditioning explicit.

3 Identification

Treating $\{\alpha_i, \gamma_j\}_n$ as parameters and jointly estimating them with the common parameter θ_0 leads to incidental-parameter bias (Neyman and Scott, 1948) in the maximum-likelihood estimator. Fernández-Val and Weidner (2015) characterize the asymptotic bias in the maximum-likelihood estimator of θ_0 and consider bias-reduction methods. On the other hand, Charbonneau (2014) shows the existence of a sufficient statistic for the pair (α_i, γ_j) by building on the work of Cox (1958), Rasch (1960, 1961), and Hirji et al. (1987). Our aim here is to develop the implied estimator and to derive its statistical properties. We first give an alternative and more direct derivation to the sufficiency result of Charbonneau (2014).

Fix a quadruple of distinct nodes $\{i_1, i_2; j_1, j_2\}$ from $\mathbb{N}_n$ and define the random variable

$$z = \frac{(y_{i_1 j_1} - y_{i_1 j_2}) - (y_{i_2 j_1} - y_{i_2 j_2})}{2},$$

and collect $x = (x_{i_1 j_1}, x_{i_1 j_2}, x_{i_2 j_1}, x_{i_2 j_2})$. Note that z can take on values from the set $\{-1, -1/2, 0, 1/2, 1\}$. Conditional on x and the event $z \in \{-1, 1\}$, z follows a Bernoulli distribution with

$$\Pr(z = 1|x, z \in \{-1, 1\}) = \frac{\Pr(z = 1|x)}{\Pr(z = 1|x) + \Pr(z = -1|x)} = \frac{1}{1 + \frac{\Pr(z=-1|x)}{\Pr(z=\ \ 1|x)}}.$$

Equations (2.1)–(2.2) together with the functional form of the logistic distribution imply that

$$\frac{\Pr(z = -1|x)}{\Pr(z = \ \ 1|x)} = \exp(-r'\theta_0),$$

where we introduce $r = (x_{i_1 j_1} - x_{i_1 j_2}) - (x_{i_2 j_1} - x_{i_2 j_2})$. This yields the following simple lemma.

Lemma 1 (Sufficiency)**.**

$$\Pr(z = 1|x, z \in \{-1, 1\}) = \frac{1}{1 + \exp(-r'\theta_0)} = F(r'\theta_0).$$

Proof. See the Appendix. $\square$

Lemma 1 states that, conditional on x and $z \in \{-1, 1\}$, the distribution of z is logistic and does not depend on the parameters $\alpha_{i_1}, \alpha_{i_2}$ and $\gamma_{j_1}, \gamma_{j_2}$. The conditional log-likelihood of the quadruple is

$$1\{z = 1\} \log F(r'\theta_0) + 1\{z = -1\} \log(1 - F(r'\theta_0)) \tag{3.3}$$

and can form the basis for the construction of a conditional maximum-likelihood estimator for θ_0.

The conditioning event $z \in \{-1, 1\}$ corresponds to only 2 of the 2^4 possible realizations of the quadruple of link decisions. These are

$$\begin{pmatrix} y_{i_1 j_1} & y_{i_1 j_2} \\ y_{i_2 j_1} & y_{i_2 j_2} \end{pmatrix} \in \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\},$$

and so cover quadruples in which the senders i_1, i_2 form only one out of two possible links to j_1, j_2 and make opposite decisions about the creation of these edges. This is an intuitive generalization of the conditional-likelihood approach in the model of Rasch (1960, 1961), where children answer two tests and only children who get one test right and the other wrong contribute to the conditional likelihood. This subpopulation of observations in the Rasch model is also frequently referred to as movers. In the current setting, the movers are taken in pairs, and only those pairs consisting of movers in opposite directions are retained for construction of the conditional likelihood. As such, the conditioning is akin to a difference-in-differences strategy.

4 Estimation and inference

The argument from the previous section suggests estimating θ_0 by maximizing the empirical counterpart to (3.3) obtained on considering all distinct quadruples $\{i_1, i_2; j_1, j_2\}$ from $\mathbb{N}_n$. There are

$$\rho = \binom{n}{2}\binom{n-2}{2} = \frac{n(n-1)(n-2)(n-3)}{4} = \frac{1}{4}\frac{n!}{(n-4)!}$$

such quadruples. It will prove useful to introduce a function σ that maps these quadruples to the index set $\mathbb{Q}_\rho = \{1, 2, \ldots, \rho\}$. Thus, each distinct quadruple $\{i_1, i_2; j_1, j_2\}$ corresponds to a unique $\sigma\{i_1, i_2; j_1, j_2\} \in \mathbb{Q}_\rho$. We may then extend our notation by defining the random variables

$$\begin{aligned} z(\sigma\{i_1, i_2; j_1, j_2\}) &= \frac{(y_{i_1 j_1} - y_{i_1 j_2}) - (y_{i_2 j_1} - y_{i_2 j_2})}{2}, \\ r(\sigma\{i_1, i_2; j_1, j_2\}) &= (x_{i_1 j_1} - x_{i_1 j_2}) - (x_{i_2 j_1} - x_{i_2 j_2}). \end{aligned}$$

When the dependence of these random variables on four nodes can be left implicit we will use the simpler shorthand notation z_σ, r_σ, where σ ranges over the set $\mathbb{Q}_\rho$.

With this notation at hand, our estimator may be written as

$$\theta_n = \arg\max_{\theta \in \Theta} L_n(\theta),$$

where Θ is the parameter space searched over, and

$$L_n(\theta) = \rho^{-1} \sum_{\sigma \in \mathbb{Q}_\rho} 1\{z_\sigma = 1\} \log F(r_\sigma' \theta) + 1\{z_\sigma = -1\} \log(1 - F(r_\sigma' \theta)).$$

This objective function is a standard logit log-likelihood restricted to quadruples of data for which $z_\sigma \in \{-1, 1\}$. Hence, the estimator can be computed using standard statistical software. The researcher is only required to construct the variables $\{z_\sigma, r_\sigma\}_\rho$, which is easy to do.

The conditional-logit estimator is consistent under weak conditions. We denote the logistic density function by f.

Assumption 1 (Sampling)**.** *The n nodes in $\mathbb{N}_n$ are sampled independently.*

Assumption 2 (Parameter space)**.** *θ_0 is interior to Θ, a compact subset of $\mathcal{R}^{\dim\theta}$.*

Assumption 3 (Identification)**.** *For all σ, $E[\|r_\sigma\|^2] < C$, where C is a finite constant, and*

$$\operatorname{rank}\left\{\lim_{\rho\to\infty}\rho^{-1}\sum_{\sigma\in\mathbb{Q}_\rho} E\big[r_\sigma r_\sigma' \, f(r_\sigma'\theta_0)\, 1\{z_\sigma \in \{-1,1\}\}\big]\right\} = \dim\theta.$$

Assumptions 1 is a natural sampling scheme for network data. It permits dependence of the covariates across dyads that have nodes in common. Assumption 2 is conventional. Assumption 3 is a standard identification condition. Together with concavity of $L_n(\theta)$, the rank requirement implies that θ_0 is the global maximizer of the large-sample conditional likelihood. Assumption 3 implies that

$$\lim_{\rho\to\infty}\rho^{-1}\sum_{\sigma\in\mathbb{Q}_\rho}\Pr(z_\sigma \in \{-1,1\}) > 0.$$

This means that the accummulation of informative quadruples does not cease as the sample grows.

Theorem 1 formally states our consistency result.

Theorem 1 (Consistency)**.** *Let Assumptions 1–3 hold. Then $\theta_n \xrightarrow{p} \theta_0$ as $n\to\infty$.*

Proof. See the Appendix. □

Although $L_n(\theta)$ has the form of the log-likelihood for a standard cross-sectional logit model, the conventional standard-error formula is not valid for θ_n. Indeed, the score (when evaluated at the true parameter value) is not a simple sample average of independent and identically distributed random variables. Moreover, it is an average over quadruples of nodes, with the same nodes showing up in multiple quadruples. This implies that some more work is to be done to perform statistical inference. We use the following moment condition in deriving distribution theory.

Assumption 4 (Moments)**.** *For all $\sigma \in \mathbb{Q}_\rho$,*

$$E[\|r_\sigma\|^6] \leq C,$$

where C is a finite constant.

The key to deriving the asymptotic distribution of θ_n is to note that the score function has the form of a U-statistic in both the senders and receivers of edges. Moreover, we have

$$S_n(\theta) = \frac{\partial L_n(\theta)}{\partial \theta} = \rho^{-1} \sum_{i_1} \sum_{i_1 < i_2} \sum_{j_1 \neq i_1, i_2} \sum_{\substack{j_1 < j_2 \\ j_2 \neq i_1, i_2}} s(\sigma\{i_1, i_2; j_1, j_2\}; \theta),$$

where we introduce the kernel function

$$s(\sigma; \theta) = r_\sigma \left\{ 1\{z_\sigma = 1\} \left(1 - F(r_\sigma' \theta)\right) - 1\{z_\sigma = -1\} F(r_\sigma' \theta) \right\}.$$

Note that the kernel $s(\sigma\{i_1, i_2; j_1, j_2\}; \theta)$ is permutation invariant in both senders (i_1, i_2) and receivers (j_1, j_2).

Standard results on U-statistics (as in, e.g., van der Vaart 2000, Chapter 12) are not directly applicable to the current setup because the data are not identically distributed and they are not independent across dyads. Nonetheless, under our conditions, the limit distribution of the normalized score vector evaluated at the true parameter value co-incides with that of its Hájek projection (van der Vaart, 2000, Section 11.3) conditional on the covariates. This projection is

$$U_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \upsilon_{ij},$$

where

$$\upsilon_{ij} = w_{ij} \left(\frac{y_{ij}\,(1-p_{ij}) - (1-y_{ij})\,p_{ij}}{p_{ij}\,(1-p_{ij})} \right),$$

and we define

$$w_{ij} = \frac{4}{(n-2)(n-3)} \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} r(\sigma\{i,i';j,j'\})\, q(\sigma\{i,i';j,j'\})$$

for

$$q(\sigma\{i,i';j,j'\}) = \frac{p_{ij}\,(1-p_{ij})\,p_{ij'}\,(1-p_{ij'})\,p_{i'j}\,(1-p_{i'j})p_{i'j'}\,(1-p_{i'j'})}{p_{ij}\,(1-p_{ij'})\,(1-p_{i'j})\,p_{i'j'} + (1-p_{ij})\,p_{ij'}\,p_{i'j}\,(1-p_{i'j'})}.$$

Because $p_{ij} = E[y_{ij}|x_{ij}]$, we have that $E[\upsilon_{ij}] = 0$ and $E[\upsilon_{ij}\,\upsilon'_{i'j'}] = 0$ unless $i = i'$ and $j = j'$. Furthermore,

$$\sqrt{n(n-1)}\,U_n \xrightarrow{d} \mathcal{N}(0, \varUpsilon), \tag{4.4}$$

where

$$\varUpsilon = \lim_{n\to\infty} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} E[\upsilon_{ij}\upsilon'_{ij}] = \lim_{n\to\infty} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} E\left[\frac{w_{ij}\,w'_{ij}}{p_{ij}(1-p_{ij})}\right],$$

which exists by Assumption 4.

The Hessian matrix, in turn, is

$$H_n(\theta) = \frac{\partial^2 L_n(\theta)}{\partial\theta\partial\theta'} = -\rho^{-1} \sum_{\sigma \in \mathbb{Q}_\rho} r_\sigma r'_\sigma\, f(r'_\sigma\theta)\; 1\{z_\sigma \in \{-1,1\}\},$$

and Assumption 4 is sufficient to ensure that $H_n(\theta_n)$ converges in probability to the matrix

$$\begin{aligned} H(\theta_0) &= \lim_{\rho\to\infty} -\rho^{-1} \sum_{\sigma \in \mathbb{Q}_\rho} E[r_\sigma r'_\sigma\, f(r'_\sigma\theta_0)\; 1\{z_\sigma \in \{-1,1\}\}] \\ &= \lim_{\rho\to\infty} -\rho^{-1} \sum_{\sigma \in \mathbb{Q}_\rho} E[r_\sigma r'_\sigma\, q(\sigma)], \end{aligned} \tag{4.5}$$

which is non-singular by Assumption 3. It is apparent from inspection of Υ and $H(\theta_0)$ that the information equality does not hold.

Combining Equations (4.4) and (4.5) with a mean-value expansion of $S_n(\theta_n)$ around θ_0 and letting

$$\Omega = H(\theta_0)^{-1}\Upsilon H(\theta_0)^{-1}$$

yields the following theorem.

Theorem 2 (Asymptotic distribution)**.** *Let Assumptions 1–4 hold. Then*

$$\sqrt{n(n-1)}\,(\theta_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega),$$

as $n \to \infty$.

Proof. See the Appendix. □

To perform statistical inference, a consistent estimator of Ω is needed. Our assumptions imply consistency of the estimator

$$\Omega_n = H_n(\theta_n)^{-1}\Upsilon_n H_n(\theta_n)^{-1},$$

where $H_n(\theta_n)$ is readily obtained when optimization of $L_n(\theta)$ is performed using standard Newton-Raphson type algorithms, and

$$\Upsilon_n = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\hat{\upsilon}_{ij}\hat{\upsilon}_{ij}'$$

with

$$\hat{\upsilon}_{ij} = \frac{4}{(n-2)(n-3)}\sum_{i'\neq i,j}\;\sum_{j'\neq i,j,i'} s(\sigma\{i,i';j,j'\};\theta_n).$$

In contrast to $H_n(\theta_n)$, the latter covariance matrix does not follow directly as output to any numerical optimization routine and needs to be obtained in an additional post-estimation step.

5 Simulations

In this section we report results from a Monte Carlo experiment. We focus on inferring a single homophily parameter, θ_0, so that

$$u_{ij} = x_{ij}\theta_0 + \alpha_i + \gamma_j - \epsilon_{ij}.$$

We will generate the unobserved-heterogeneity parameters (α_i, γ_j) independently from the dyad characteristic x_{ij}. Experiments with correlated heterogeneity yielded similar results. Under independence, the total variance of the surplus factors as

$$\operatorname{var} u_{ij} = \operatorname{var} x_{ij}\theta_0 + \operatorname{var}(\alpha_i + \gamma_j) + \operatorname{var} \epsilon_{ij},$$

and we can vary the relative importance of homophily and unobserved degree heterogeneity in forming a match by varying the relative contribution of $x_{ij}\theta_0$ and $\alpha_i + \gamma_j$ to the variance of u_{ij}.

We fix $\theta_0 = 1$ throughout and consider symmetric data generating processes in which

$$x_{ij} = \delta\, v_i\, v_j = x_{ji},$$

and

$$\alpha_i \sim \mathcal{N}(0, \beta^2), \qquad \gamma_i \sim \mathcal{N}(0, \beta^2),$$

where δ and β are positive scale parameters and v_i for $i \in \mathbb{N}_n$ is a random variable on which matching between nodes is based. We generated this variable as $v_i \sim \mathcal{N}(0, 1)$. The model implies positive assortative matching in the sense that the propensity to form a link from i to j is larger when both v_i and v_j are larger and of the same sign. The scale parameters δ and β allow to vary the contributions of, respectively, homophily and latent heterogeneity to the surplus. We consider three different choices for these parameters, yielding designs where homophily is of relatively equal, more, and less importance than the degree heterogeneity parameters (α_i, γ_j). Table 1 gives an overview of the three designs (#). It states the relative contribution of each of the three components of the match surplus to its total variance, together with the associated parameter values. The parameter values are stated up to the factor of proportionality $\pi^2/3$, which is the variance of the logistic distribution.

Table 1: Design parameters for simulations

#	$\mathrm{var}\, x_{ij}\theta_0$	$\mathrm{var}\,(\alpha_i + \gamma_j)$	$\mathrm{var}\, \epsilon_{ij}$	$\delta^2 \propto \pi^2/3$	$\beta^2 \propto \pi^2/3$
1	1/4	1/4	2/4	1/2	1/4
2	2/6	1/6	3/6	2/3	1/6
3	1/6	2/6	3/6	1/3	1/3

These designs are difficult in the sense that, with relative contributions of 1/4, 1/3, and 1/6 respectively, homophily contributes only little to the total variance of the link surplus. Each of the designs was ran with $n = 25$ and $n = 50$, yielding $25 \times 24 = 600$ and $50 \times 49 = 2450$ link decisions, respectively.

We estimated θ_0 by our conditional logit estimator (logit) and by (full-information) maximum likelihood (mle), that is, estimating the nuisance parameters $\{\alpha_i, \gamma_j\}_n$ jointly with the parameter θ_0. In Table 2 we report the mean, median, standard deviation (std), and interquartile range (iqr) of the two estimators, computed over $1,000$ Monte Carlo replications with all random variables redrawn in each iteration. The table also contains the ratio of the (average) estimated standard error to the standard deviation of the Monte Carlo estimates (se/std) and the coverage rate of the associated 95% confidence intervals (coverage).

Table 2: Simulation results

#	mean		median		std		iqr		se/std		coverage	
						$n = 25$						
	mle	logit	mle	logit	mle	logit	mle	logit	mle	logit	mle	logit
1	1.124	1.022	1.121	1.018	.157	.159	.205	.209	.910	1.069	.860	.959
2	1.136	1.021	1.125	1.012	.147	.138	.200	.174	.901	1.128	.839	.972
3	1.125	1.023	1.112	1.009	.173	.179	.228	.239	.920	1.055	.895	.966
						$n = 50$						
	mle	logit	mle	logit	mle	logit	mle	logit	mle	logit	mle	logit
1	1.055	1.003	1.052	.999	.066	.071	.087	.097	.995	1.015	.881	.950
2	1.058	1.001	1.058	1.001	.064	.065	.087	.065	.959	1.028	.857	.952
3	1.057	1.006	1.055	1.006	.076	.082	.106	.079	.949	1.034	.886	.968

The results confirm that the maximum-likelihood estimator is biased and that the bias is $O(n^{-1})$. Consequently, the bias is not negligible relative to the standard error and confidence bounds based on the asymptotic distribution are not centered around the true parameter value. This is apparent from inspection of the empirical coverage rates, which show substantial undercoverage. For the designs with $n = 25$ the undercoverage problem is exacerbated by an underestimation of the actual standard deviation of the point estimates by the standard error. In contrast, the conditional estimator has bias that is small compared to its standard error for all designs considered. The associated confidence intervals are somewhat too wide when $n = 25$, and so inference is conservative, because, here, the asymptotic-variance formula tends to slightly overestimate the small-sample variability in the point estimates. When n increases the variance approximation rapidly improves and the empirical coverage rate converges to the theoretical coverage rate of .95. Note also that the standard deviation of the conditional estimator is very similar to that of maximum likelihood. This suggests that, at least in the designs considered here, little to no information is lost by conditioning.

Finally, Figure 1 contains plots of the histograms of the $1,000$ Monte Carlo replications of the Studentized estimator

$$\sqrt{n(n-1)}\, \mathit{\Omega}_n^{-1/2}(\theta_n - \theta_0)$$

for each of the designs in Table 1 (top to bottom) and the two sample sizes considered ($n \in \{25, 50\}$; left and right, respectively). In each plot, the histogram is accompanied by the standard-normal density as a point of reference for the asymptotic approximation in Theorem 2. The plots reveal that the asymptotic approximation is fairly accurate even for the small sample sizes considered here.

6 Application

As an empirical application we investigate the determinants of trade from country-level trade data. The network-formation model we estimate follows closely Helpman et al. (2008), who provide a theoretical foundation for it. Our data set consists of a cross section of 136 countries. For each country pair (i, j) the outcome variable, *trade decision*, is a dummy variable that registers whether or not trade occured from i to j. The data also contain various dyad characteristics that we use as explanatory variables. All these variables are measures of closeness between the two countries. Table 3 contains descriptive statistics. *log distance* is the (log of the) geographical distance between the capitals of countries i and j. *common border* and *common language* are dummy variables that take on the value one if i and j share, respectively, a physical boundary or a common language. *colonial ties* takes

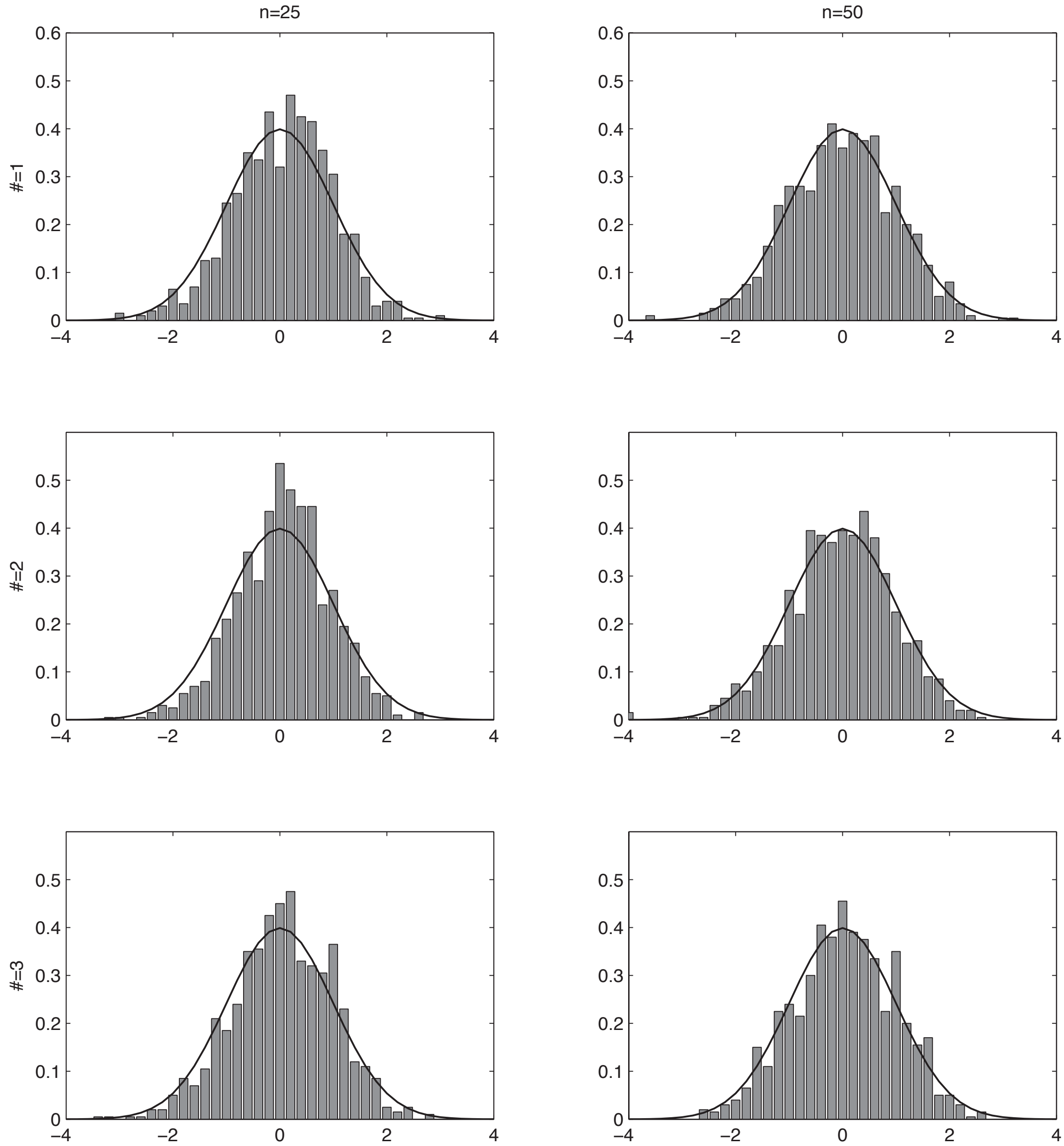

Figure 1: Distributions of Studentized estimator

on the value one if, at some point, i colonized j (or vice versa) and zero otherwise. Finally, *preferential trade agreement* is a binary variable that indicates whether i and j take part in a joint preferential trade agreement. Original data sources and additional details on the data are available in Santos Silva and Tenreyro (2006).

Table 3: Descriptive statistics

	mean	standard deviation
trade decision	0.5236	0.4995
log distance	8.7855	0.7418
common border	0.0196	0.1387
common language	0.2097	0.4071
colonial ties	0.1705	0.3761
preferential trade agreement	0.0155	0.1234

We estimated the parameters of this model by maximum likelihood and by conditional logit. The point estimates, along with their standard errors (stated in parentheses below the point estimates), are collected in Table 4. The signs of all parameter estimates agree with those of Helpman et al. (2008). Geographical distance decreases the propensity to trade while homophily tends to increase the likelihood of trade. Indeed, speaking a common language and having a colonial history positively affect the probability of trading. Trade agreements have a large positive impact on trade decisions, which more than offsets the negative influence of distance. A, perhaps, somewhat surprising finding is the negative point estimate on *common border*. It should be noted that, when not controlling for preferential trade agreements, the sign of this coefficient changes. Also, of the $136 \times 135 = 18,360$ country dyads in the data, relatively few (360 dyads) share a border and even less (285

Table 4: Trade estimates

	mle	logit
log distance	-1.3490	-1.0920
	(0.0504)	(0.0573)
common border	-1.2070	-0.8220
	(0.2089)	(0.2668)
common language	0.5851	0.4672
	(0.0906)	(0.1031)
colonial ties	0.5206	0.5925
	(0.0962)	(0.1047)
preferential trade agreement	2.0444	1.3038
	(0.3056)	(0.2913)

dyads) have established preferential trade agreements; see Table 3. In the raw data, the dyads that allow to discriminate between the impact of *common border* and *preferential trade agreement* we have the following pattern. Of the country pairs that do not have a common border but have established a preferential trade agreement, 85% are engaged in trade. On the other hand, of the country pairs that do have a common border but have not established a preferential trade agreement, only 58% trade. Again, the positive effect of a preferential trade agreement outweighs the negative border effect. On comparing the maximum-likelihood estimates with those obtained by conditional logit we see that the latter tend to be smaller (in absolute value), with similar standard errors. The one exception is *colonial ties*, where the difference is nonetheless very small and statistically insignificant at conventional significance levels. The ratio of the other conditional estimates to their maximum-likelihood counterparts ranges from 63% to 81%.

Appendix

Proof of Lemma 1. Equations (2.1)–(2.2) together with the functional form of the standard logistic distribution imply that

$$\begin{aligned}\Pr(z=1|x) &= \frac{1}{1+\exp(-\alpha_{i_1}-\gamma_{j_1}-x'_{i_1j_1}\theta_0)}\frac{\exp(-\alpha_{i_1}-\gamma_{j_2}-x'_{i_1j_2}\theta_0)}{1+\exp(-\alpha_{i_1}-\gamma_{j_2}-x'_{i_1j_2}\theta_0)}\\ &\times\frac{\exp(-\alpha_{i_2}-\gamma_{j_1}-x'_{i_2j_1}\theta_0)}{1+\exp(-\alpha_{i_2}-\gamma_{j_1}-x'_{i_2j_1}\theta_0)}\frac{1}{1+\exp(-\alpha_{i_2}-\gamma_{j_2}-x'_{i_2j_2}\theta_0)}\end{aligned}$$

and, similarly, that

$$\begin{aligned}\Pr(z=-1|x) &= \frac{\exp(-\alpha_{i_1}-\gamma_{j_1}-x'_{i_1j_1}\theta_0)}{1+\exp(-\alpha_{i_1}-\gamma_{j_1}-x'_{i_1j_1}\theta_0)}\frac{1}{1+\exp(-\alpha_{i_1}-\gamma_{j_2}-x'_{i_1j_2}\theta_0)}\\ &\times\frac{1}{1+\exp(-\alpha_{i_2}-\gamma_{j_1}-x'_{i_2j_1}\theta_0)}\frac{\exp(-\alpha_{i_2}-\gamma_{j_2}-x'_{i_2j_2}\theta_0)}{1+\exp(-\alpha_{i_2}-\gamma_{j_2}-x'_{i_2j_2}\theta_0)}.\end{aligned}$$

Therefore,

$$\frac{\Pr(z=-1|x)}{\Pr(z=\ \ 1|x)} = \frac{\exp(-\alpha_{i_1}-\gamma_{j_1}-x'_{i_1j_1}\theta_0)\,\exp(-\alpha_{i_2}-\gamma_{j_2}-x'_{i_2j_2}\theta_0)}{\exp(-\alpha_{i_1}-\gamma_{j_2}-x'_{i_1j_2}\theta_0)\,\exp(-\alpha_{i_2}-\gamma_{j_1}-x'_{i_2j_1}\theta_0)} = \exp(-r'\theta_0),$$

from which Lemma 1 follows. □

Proof of Theorem 1. Let

$$L(\theta) = \rho^{-1}\sum_{\sigma\in\mathbb{Q}_\rho} E\left[1\{z_\sigma=1\}\,F(r'_\sigma\theta) + 1\{z_\sigma=-1\}\,\log(1-F(r'_\sigma\theta))\right].$$

Note that, by Assumption 3, θ_0 is the unique global maximizer of $\lim_{n\to\infty} L(\theta)$ on Θ. Because $L_n(\theta)$ is concave, $\theta_n \xrightarrow{p} \theta_0$ will follow from pointwise convergence in probability of $L_n(\theta)$ to $L(\theta)$ (Newey and McFadden, 1994, Theorem 2.7).

Write,

$$L_n(\theta) - L(\theta) = \rho^{-1}\sum_{\sigma\in\mathbb{Q}_\rho} \ell_\sigma(\theta) - E[\ell_\sigma(\theta)].$$

Because $|\ell_\sigma(\theta)| \leq \log 2 + 2\|r_\sigma\| \, \|\theta\|$, and $E[\|r_\sigma\|^2]$ is finite and Θ is compact, the variance of $\ell_\sigma(\theta)$ exists and is uniformly bounded in σ. Therefore, by Chebychev's inequality, for any $\epsilon > 0$,

$$\Pr(|L_n(\theta) - L(\theta)| > \epsilon) \leq \frac{E(|L_n(\theta) - L(\theta)|)^2}{\epsilon^2},$$

for each $\theta \in \Theta$. Now,

$$E(|L_n(\theta) - L(\theta)|)^2 = E\left(\left(\rho^{-1} \sum_{\sigma \in \mathbb{Q}_\rho} \ell_\sigma(\theta) - E[\ell_\sigma(\theta)]\right)\left(\rho^{-1} \sum_{\sigma' \in \mathbb{Q}_\rho} \ell_{\sigma'}(\theta) - E[\ell_{\sigma'}(\theta)]\right)\right).$$

A pair of quadruples $\sigma = \sigma\{i_1, i_2, j_1, j_2\}$ and $\sigma' = \sigma\{i'_1, i'_2, j'_1, j'_2\}$ will deliver a non-zero contribution to this covariance as long as σ and σ' have at least one node in common. Quadruples involving only distinct nodes are independent by Assumption 1. There are $O(n^7)$ terms with at least one node in common. The number of terms with two or more nodes in common is $O(n^6)$. Because, $\rho = O(n^4)$ we have

$$E(|L_n(\theta) - L(\theta)|)^2 = \frac{O(n^7)}{\rho^2} = \frac{O(n^7)}{O(n^8)} = O(n^{-1}),$$

and so $\lim_{n\to\infty} \Pr(|L_n(\theta) - L(\theta)| > \epsilon) = 0$ for any $\epsilon > 0$ and all $\theta \in \Theta$. Therefore, $\theta_n \xrightarrow{p} \theta_0$ as $n \to \infty$ and the proof is complete. □

Proof of Theorem 2. A mean-value expansion around θ_0 gives

$$\begin{aligned}\sqrt{n(n-1)}\,(\theta_n - \theta_0) &= H(\theta_0)^{-1}\sqrt{n(n-1)}\,S_n(\theta_0) + o_p(1)\\ &= H(\theta_0)^{-1}\sqrt{n(n-1)}\,U_n + o_p(1),\end{aligned}$$

where the first equality follows from the uniform convergence of $H_n(\theta)$ to $H(\theta)$ and Theorem 1, and the second equality follows from the asymptotic equivalence of $\sqrt{n(n-1)}\,S_n(\theta_0)$ and $\sqrt{n(n-1)}\,U_n$. The validity of each of these two transitions is shown below (under *(i)* and *(ii)*, respectively). As will also be discussed below (under *(iii)*), our assumptions further imply that

$$\sqrt{n(n-1)}\,U_n \xrightarrow{d} \mathcal{N}(0, \varUpsilon),$$

as $n \to \infty$. Therefore, Slutsky's theorem yields

$$\sqrt{n(n-1)}\,(\theta_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, H(\theta_0)^{-1}\varUpsilon H(\theta_0)^{-1}),$$

as $n \to \infty$, which is the result of Theorem 2. We now turn to demonstrating Points *(i)*–*(iii)* in turn.

(i) Convergence of the Hessian. We need to show that $H_n(\theta_*) \xrightarrow{p} H(\theta_0)$ as $n \to \infty$, for any θ_* that lies in between θ_n and θ_0. Because $\theta_n \xrightarrow{p} \theta_0$ as $n \to \infty$ and $H_n(\theta)$ is a continuous function of θ, it suffices to show that

$$\sup_{\theta \in \varTheta} \|H_n(\theta) - H(\theta)\| = o_p(1).$$

To show this we verify the conditions of Lemma 2.9 of Newey and McFadden (1994). Because

$$H_n(\theta) = -\rho^{-1} \sum_{\sigma \in \mathbb{Q}_\rho} r_\sigma r_\sigma'\, f(r_\sigma'\theta)\, 1\{z_\sigma \in \{-1, 1\}\},$$

and f has bounded derivative f',

$$\begin{aligned}
\|H_n(\theta_1) - H_n(\theta_2)\| &\leq \left(\rho^{-1} \sum_{\sigma \in \mathbb{Q}_\rho} \|r_\sigma\|^3\, 1\{z_\sigma \in \{-1, 1\}\}\right) \sup_\epsilon f'(\epsilon)\; \|\theta_1 - \theta_2\| \\
&= O_p(1)\, \|\theta_1 - \theta_2\|,
\end{aligned}$$

for any $\theta_1, \theta_2 \in \varTheta$. Here, the second transition follows from the moment requirements in Assumption 4, as can be shown using the same steps as those used to prove Theorem 1. Therefore, the matrix $H_n(\theta)$ is stochastically equicontinuous, and uniform convergence follows from pointwise convergence on Θ. Now, as $E\left[\|r_\sigma\|^4 | z_\sigma \in \{-1,1\}\right]$ is uniformly bounded in σ by Assumption 4, and because the density f is bounded on $\mathcal{R}$, the same argument as above equally gives $\|H_n(\theta) - H(\theta)\| = o_p(1)$ as $n \to \infty$ for all $\theta \in \Theta$. Thus, uniform convergence has been shown.

(ii) Projection of the score. We will first show that the $S_n(\theta_0)$ is asymptotically equivalent to its Hájek projection, U_n, *conditional* on the covariate sequence $\{x_{ij}\}_n$. Introduce $\overline{E}$ as a notational shorthand for the expectation given $\{x_{ij}\}_n$ (and the fixed effects). The expression for the projection of $S_n(\theta_0)$ given in the main text follows from a small calculation of the expectation in

$$U_n = \frac{4}{n(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \sum_{i' \neq i} \sum_{j \neq i,i'} \sum_{j' \neq i,i',j} \overline{E}[s(\sigma\{i,i';j,j'\})|\, y_{ij}]$$

and uses the fact that

$$\begin{aligned} \Pr(z_\sigma = 1|x_\sigma) &= \phantom{\big(1-{}} F(r_\sigma'\theta_0) \Pr(z_\sigma \in \{-1,1\}|\, x_\sigma), \\ \Pr(z_\sigma = -1|x_\sigma) &= \big(1 - F(r_\sigma'\theta_0)\big) \Pr(z_\sigma \in \{-1,1\}|\, x_\sigma), \end{aligned} \tag{A.1}$$

where we abuse notation slightly by denoting by x_σ the collection of covariates for the nodes in the quadruple σ. To show that the scaled score vector $\sqrt{n(n-1)}\, S_n(\theta_0)$ is asymptotically equivalent to its projection $\sqrt{n(n-1)}\, U_n$, conditional on covariates, we need to verify that

$$n^2\, \overline{E}[(U_n - S_n(\theta_0))\, (U_n - S_n(\theta_0))'] = o(1), \tag{A.2}$$

as $n \to \infty$.

The main task in establishing (A.2) is the calculation of the asymptotic variance of the normalized score. Moreover, we need to show that

$$n(n-1)\, \overline{E}[S_n(\theta_0)\, S_n(\theta_0)'] = \Upsilon + o(1),$$

as $n \to \infty$. Because $E[s(\sigma;\theta_0)|x_\sigma] = 0$ for all $\sigma \in \mathbb{Q}_\rho$ and link decisions are conditionally independent,

$$E[s(\sigma;\theta_0)\, s(\sigma';\theta_0)'|x_\sigma, x_{\sigma'}] = 0$$

unless σ and σ' have at least one dyad in common. There are $O(n^6)$ terms with only one dyad in common. The number of terms with more than one dyad in common is $o(n^6)$. Therefore, $\operatorname{var} S_n(\theta_0) = \rho^{-1}\, O(n^6) = O(n^{-2})$ and its leading term is comprised of correlations between $s(\sigma;\theta_0)$, and $s(\sigma';\theta_0)$ for which the quadruples σ, σ' have exactly one dyad in common. By symmetry of $s(\sigma,\theta)$ in the sender and receiver nodes, we can fix this to be the first sender-receiver dyad and multiply the expression for $s(\sigma;\theta_0)$ through by 4. We may then write the dominant part of $n(n-1)\, \overline{E}[S_n(\theta_0)\, S_n(\theta_0)']$ as

$$\frac{4^2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \left(\sum_{i' \neq i,j} \sum_{j' \neq i,i',j} \sum_{i'' \neq i,j} \sum_{j'' \neq i,i'',j} \frac{\overline{E}\left[s(\sigma\{i,i';j,j'\};\theta_0)\, s(\sigma\{i,i'';j,j''\};\theta_0)'\right]}{(n-2)^2(n-3)^2} \right).$$

Fix $\sigma = \sigma\{i,i';j,j'\}$ and $\sigma' = \sigma\{i,i'';j,j''\}$. Then

$$\begin{aligned} s(\sigma;\theta_0)\, s(\sigma';\theta_0)' = {} & r_\sigma r'_{\sigma'}\, 1\{z_\sigma = 1, z_{\sigma'} = 1\}\, (1 - F(r'_\sigma\theta_0))\, (1 - F(r'_{\sigma'}\theta_0)) \\ & + r_\sigma r'_{\sigma'}\, 1\{z_\sigma = -1, z_{\sigma'} = -1\} \quad F(r'_\sigma\theta_0) \qquad F(r'_{\sigma'}\theta_0) \\ & - r_\sigma r'_{\sigma'}\, 1\{z_\sigma = 1, z_{\sigma'} = -1\}\, (1 - F(r'_\sigma\theta_0)) \quad F(r'_{\sigma'}\theta_0) \\ & - r_\sigma r'_{\sigma'}\, 1\{z_\sigma = -1, z_{\sigma'} = 1\} \quad F(r'_\sigma\theta_0) \quad (1 - F(r'_{\sigma'}\theta_0)). \end{aligned} \tag{A.3}$$

Take expectations (given covariates). The last two terms on the right-hand side of (A.3) drop out, while the expectations of the first and second right-hand side term are equal to

$$r_\sigma\, r'_{\sigma'}\, \frac{F(r'_\sigma\theta_0)\, (1 - F(r'_\sigma\theta_0))\, F(r'_{\sigma'}\theta_0)\, (1 - F(r'_{\sigma'}\theta_0))}{p_{ij}} \Pr(z_\sigma \in \{-1,1\})\, \Pr(z_{\sigma'} \in \{-1,1\})$$

and

$$r_\sigma\, r'_{\sigma'}\, \frac{F(r'_\sigma\theta_0)\, (1 - F(r'_\sigma\theta_0))\, F(r'_{\sigma'}\theta_0)\, (1 - F(r'_{\sigma'}\theta_0))}{1 - p_{ij}} \Pr(z_\sigma \in \{-1,1\})\, \Pr(z_{\sigma'} \in \{-1,1\}),$$

respectively. By (A.1), and observing that

$$q(\sigma) = \frac{\Pr(z_\sigma = 1 |\, x_\sigma)\, \Pr(z_\sigma = -1 |\, x_\sigma)}{\Pr(z_\sigma \in \{-1, 1\} |\, x_\sigma)},$$

we therefore have

$$E[s(\sigma; \theta_0)\, s(\sigma'; \theta_0)' \,|\, x_\sigma, x_{\sigma'}] = r_\sigma r'_{\sigma'} \, \frac{q(\sigma)\, q(\sigma')}{p_{ij}(1 - p_{ij})}.$$

Averaging across all quadruples and using the definition of w_{ij} given in the main text we find

$$n(n-1)\, \overline{E}[S_n(\theta_0)\, S(\theta_0)'] = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{w_{ij} w'_{ij}}{p_{ij}(1 - p_{ij})} + o(1) = \Upsilon_X + o(1), \text{ (say)}. \quad \text{(A.4)}$$

Assumptions 1 and 4 imply that $\|\Upsilon_X - \Upsilon\| = O_p(n^{-1/2})$; the proof of this result follows the same pattern as that of the pointwise-convergence statement in the proof of Theorem 1. Therefore,

$$\lim_{n \to \infty} n(n-1)\, \overline{E}[S(\theta_0)\, S(\theta_0)'] = \lim_{n \to \infty} n(n-1)\, \overline{E}[U_n\, U'_n] = \Upsilon,$$

as claimed.

Making use of the above calculations, it is readily deduced that we equally have that

$$\lim_{n \to \infty} n(n-1)\, \overline{E}[U_n\, S_n(\theta_0)'] = \Upsilon,$$

that is, that the asymptotic covariance between U_n and $S_n(\theta_0)$ equals their variance. Put together, these results imply (A.2).

(iii) Asymptotic normality. To conclude the proof of Theorem 2 it remains only to show that

$$\sqrt{n(n-1)}\, \Upsilon^{-1/2}\, U_n \xrightarrow{d} \mathcal{N}(0, I) \quad \text{(A.5)}$$

as $n \to \infty$, where I denotes the identity matrix of conformable dimension. Recall that

$$U_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \upsilon_{ij}$$

and that the υ_{ij} are independent conditional on the covariates $\{x_{ij}\}_n$ (and the fixed effects). A conditional central limit theorem (e.g., Prakasa Rao 2009, Theorem 8) then implies that

$$\sqrt{n(n-1)}\, \varUpsilon_X^{-1/2}\, U_n \xrightarrow{d} \mathcal{N}(0, I), \tag{A.6}$$

conditional on the covariates. Now, $\overline{E}[U_n] = E[U_n] = 0$ and, as was established above, $\|\varUpsilon_X - \varUpsilon\| = O_p(n^{-1/2})$. Therefore, the limit distribution is independent of the covariate values, and (A.6) continues to hold unconditionally, with $\varUpsilon$ replacing $\varUpsilon_X$. This is (A.5). This concludes the proof of Theorem 2. $\square$

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B 32*, 283–301.

Charbonneau, K. B. (2014). Multiple fixed effects in binary response panel data models. Working Paper 2014–17, Bank of Canada.

Chatterjee, S., P. Diaconis, and A. Sly (2011). Random graphs with a given degree sequence. *Annals of Applied Probability 21*, 1400–1435.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society, Series B 20*, 215–242.

Dzemski, A. (2014). An empirical model of dyadic link formation in a network with unobserved heterogeneity. Mimeo.

Fernández-Val, I. and M. Weidner (2015). Individual and time effects in nonlinear panel data models with large N, T. CeMMAP Working Paper 17/15.

Graham, B. S. (2015). An econometric model of link formation with degree heterogeneity. Mimeo.

Helpman, E., M. Melitz, and Y. Rubinstein (2008). Estimating trade flows: Trading partners and trading volumes. *Quarterly Journal of Economics 123*, 441–487.

Hirji, K. F., C. R. Mehta, and N. R. Patel (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association 82*, 1110–1117.

Hoff, P. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association 100*, 286–295.

Hoff, P., A. Raftery, and M. Handcock (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association 97*, 1090–1098.

Holland, P., K. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: Some first steps. *Social Networks 5*, 109–137.

Holland, P. W. and S. Leinhardt (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association 76*, 33–65.

Jackson, M. O. (2008). *Social and Economic Networks*. Princeton University Press.

Jackson, M. O., T. Rodriguez-Barraquer, and X. Tan (2012). Social capital and social quilts: Network patterns of favor exchange. *American Economic Review 102*, 1857–1897.

Li, H., B. Lindsay, and R. Waterman (2003). Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society, Series B 65*, 191–208.

Newey, W. K. and D. L. McFadden (1994). Large sample estimation and hypothesis testing. In R. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, Chapter 36, pp. 2111–2245. Elsevier.

Neyman, J. and E. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1–32.

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

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Unpublished report, The Danish Institute of Educational Research, Copenhagen.

Rasch, G. (1961). On the general laws and the meaning of measurement in psychology. In *The Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 4, pp. 321–333. University of California Press, Berkeley and Los Angeles.

Rinaldo, A., S. Petrovic, and S. Fienberg (2013). Maximum likelihood estimation in the β-model. *Annals of Statistics 41*, 1085–1110.

Santos Silva, J. M. C. and S. Tenreyro (2006). The log of gravity. *Review of Economics and Statistics 88*, 641–658.

Sartori, N. (2003). Modified profile likelihood in models with stratum nuisance parameters. *Biometrika 90*, 533–549.

van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge University Press.

van Duijn, M. A. J., T. A. B. Snijders, and B. J. H. Zijlstra (2004). p_2: A random effects model with covariates for directed graphs. *Statistica Neerlandica 58*, 234–254.

Vu, D. Q., D. R. Hunter, and M. Schweinberger (2013). Model-based clustering of large networks. *Annals of Applied Statistics 7*, 1010–1039.

Yan, T., C. Leng, and J. Zhu (2015). Asymptotics in directed exponential random graph models with an increasing bi-degree sequence. Forthcoming in *Annals of Statistics*.

Yan, T. and J. Xu (2013). A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices. *Biometrika 100*, 519–524.