Semiparametric Analysis of Network Formation

Koen Jochmans*
Department of Economics, Sciences Po, Paris

(First version: September 18, 2015)

January 4, 2017

Abstract

We consider a statistical model for directed network formation that features both node-specific parameters that capture degree heterogeneity and common parameters that reflect homophily among nodes. The goal is to perform statistical inference on the homophily parameters while treating the node-specific parameters as fixed effects. Jointly estimating all parameters leads to incidental-parameter bias and incorrect inference. 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. The estimator is easy to compute and can be applied to both dense and sparse networks, and is shown to have desirable asymptotic properties under sequences of growing networks. We illustrate the improvements of this estimator over maximum likelihood and bias-corrected estimation in a series of numerical experiments. The technique is applied to explain the import and export patterns in a dense network of countries and to estimate a more sparse advice network among attorneys in a corporate law firm.

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

*Address for correspondence: Sciences Po, Département d'économie, 28 rue des Saints Pères, 75007 Paris, France. E-mail: koen.jochmans@sciencespo.fr.

I am grateful to the Joint Editor, an Associate Editor, and two referees for constructive comments. I would also like to thank Bryan Graham for comments and suggestions, Thierry Magnac for discussion on semiparametric identification of binary-choice models, and Martin Weidner for discussion on the estimation of models from network data. Financial support from the European Research Council through Starting Grant nº 715787 is gratefully acknowledged.

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. This issue has received quite some attention, not only in economics, but also in sociology and statistics; Snijders (2011) and de Paula (2016) provide extensive overviews and many references.

Estimating models of network formation can be subject to a variety of complications. These range from incompleteness and lack of point identification (see Sheng 2012 and de Paula et al. 2015) over computational intractability (as in exponential random graph models; see Robins et al. 2007 and Robins et al. 2009) to the presence of a large number of parameters to estimate relative to the sample size (as in Holland and Leinhardt 1978, 1981 or Graham 2015). These difficulties explain why there is relatively little statistical theory available (Goldenberg et al. 2010).

In this paper we study a model that is sufficiently tractable to allow estimation and hypothesis testing, yet is able to replicate several key features of networks typically observed in economic data. These features are degree heterogeneity—that is, the observation that the number of links can differ substantially across nodes—and homophily—the feature that nodes are more likely to establish a link between them if they are more similar to one another.

The model under study is a directed Erdős and Rényi (1959, 1960) random-graph model where the probability of link formation is heterogenous. Moreover, the probability that a link is established between two nodes is a function of parameters that are specific to each of the nodes as well as of a set of observable characteristics that are specific to the pair of nodes. The node-specific parameters capture degree heterogeneity while the presence of dyad characteristics can be used to study homophily patterns in the data. The model is an extension of the classic model by Holland and Leinhardt (1981) and is also studied by Dzemski (2014) and, more recently, Yan et al. (2016). Graham (2015) considers a similar

model as do we but for undirected random graphs.

The main parameter of interest in our model is the homophily parameter. Our aim is to perform statistical inference on this parameter from observing a single network, treating the node-specific parameters as fixed effects. Graham (2015) has illustrated the importance of controlling for degree heterogeneity in this way when assessing homophily. Inference is non-standard because the number of parameters grows with the sample size. This results in an incidental-parameter problem (Neyman and Scott, 1948) that is similar in nature as in the estimation of two-way models for panel data; see Fernández-Val and Weidner (2016) for a characterization of the resulting bias and methods to correct for it. Dzemski (2014) adapts their techniques to perform inference in our network-formation model. A limitation of this bias-correction approach is that it has not been developed for sparse networks, where the number of links is small relative to the sample size. Such networks are nonetheless prevalent in economic settings. Moreover, very few results are available on the accuracy of statistical inference on fixed-effect models estimated from sparse networks; Jochmans and Weidner (2016) study the linear model.

As an alternative to the work of Dzemski (2014) 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 that does not depend on the node-specific parameters. This approach can be understood to be a generalization of the celebrated conditional-likelihood argument of Rasch (1960, 1961) for panel data binary-choice models with fixed effects; see also Chamberlain (1980). However, standard theory for conditional-likelihood estimators (Andersen, 1970) does not apply to our case. We use results from Jochmans (2016) to establish consistency and to derive the limit distribution of our estimator. Importantly, because the objective function is free of node-specific parameters, these parameters need not be estimated. This implies that our estimator can be used on sparse networks. A similar approach is taken in contemporaneous work by Graham (2015) for an undirected version of our model.

An example where our model can be useful is in the analysis of trade networks. In that context, a lot of effort has been made in understanding the drivers behind the import and export patterns of countries. Typical drivers looked at include geographical distance as well as a set of indicators of closeness, such as whether or not the countries have a free trade agreement or whether they share a border. Head and Mayer (2014) survey the literature. At least since the work of Anderson and van Wincoop (2003) models for trade flows feature country-specific parameters. Moreover, the estimated equation in Helpman et al. (2008) is essentially an application of the network-formation model under study here. However, the statistical methods used there do not properly account for the presence of the node-specific fixed effects. We estimate such a model as an empirical illustration of our techniques and find smaller effects (in magnitude) of dyad characteristics on the log-odds of countries engaging in trade.

As a second empirical application we infer the determinants of an advice network in a corporate law partnership. Here, the dyad covariates measure differences in position in the firm, location of employment, gender, tenure, and age of the attorneys. While the trade data from the first application yields a network that can be considered dense, the advice network is rather sparse, making this application on small informal networks a useful complement to illustrate the scope of our approach. Similar analysis could be performed to study risk sharing (Fafchamps and Gubert, 2007) or microfinance (Banerjee et al., 2013), for example.

An important feature of our setup is that link decisions are conditionally independent. This can be a reasonable assumption if the dominant drivers behind link creation are node and dyad characteristics. As such, the model postulated here is not well-suited for situations where link decisions are influenced by link decisions made by other nodes or for data where one observes a high degree of transitivity in links. Models for interdependent network formation typically fail to be point identifying. Achieving (point) identification while allowing for transitivity will typically require observing the network at multiple time

periods; see, e.g., Graham (2013, 2016).

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

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

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. Throughout we treat $\{\alpha_i, \gamma_i\}_n$ as fixed, that is, we condition on them. Equations (2.1)–(2.2) state that nodes form links by maximizing the joint surplus of a link. As such, the model under study is a cooperative model of network formation. Furthermore, the decision rule is compatible with the direct-transfer network-formation game studied in Bloch and Jackson (2007).

Suppose that the ϵ_{ij} are independent and identically distributed and follow the standard logistic distribution $F(\epsilon) = (1 + \exp(-\epsilon))^{-1}$. The logistic distribution has a long history in the analysis of network formation and arises naturally in several classic models (Zermelo 1929, Bradley and Terry 1952). The probability of observing a link from i to j given the characteristics of the nodes is

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

Thus, the data generating process of interest yields an Erdős and Rényi (1959, 1960) type random graph where the probability of link formation between i and j is heterogeneous

across both i and j. Our model is an extension of the classic model of Holland and Leinhardt (1981) for network formation. The extension lies in the presence of characteristics at the dyad level on top of the node-specific parameters. 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.

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

$$\log\left(\frac{\Pr(y_{ij}=1|x_{ij})}{\Pr(y_{ij}=0|x_{ij})}\right) = x_{ij}'\theta_0 + \alpha_i + \gamma_j,$$

this allows evaluating the importance of dyad characteristics on the probability that the nodes form a link between them. Knowledge of θ_0 is valuable in learning about homophily, that is, to what extent nodes with similar characteristics are more likely to establish links between them (McPherson et al., 2001). Homophily is a common phenomenon and is well recognized to be important in economic models (see, e.g., Currarini et al. 2009 and Golub and Jackson 2012). Empirical analysis of homophily has precedent in economics (see, e.g., De Weerdt 2004, Fafchamps and Gubert 2007, Attanasio et al. 2012). However, most specifications do not take into account the presence of degree heterogeneity, that is, the fact that the number of links a node is involved in can vary substantially across nodes. Such heterogeneity is a very frequent phenomenon, and Graham (2015) shows that ignoring it will typically lead to erroneous inference. In our model—as in those of Holland and Leinhardt (1981), Rinaldo et al. (2013), and Graham (2015)—degree heterogeneity is captured by the node-specific parameters.

Our model differs from that in Graham (2015) in that we look at directed networks. The appropriate choice of model specification depends on the application at hand. With directed data it is natural to allow for heterogeneity in the number of links send as well as in the number of links received. In (2.2), this is done by including two different fixed effects;

α_i captures heterogeneity in outgoing links while γ_i reflects heterogeneity in incoming links. For example, in our first empirical application we study trade flows between exporters and importers, which calls for a directed model. In that context, the importer and exporter fixed effects are typically referred to as multilateral resistance terms, following the seminal work of Anderson and van Wincoop (2003). In our second application we estimate an advice network, which is clearly directed in nature. Other applications where a directed model is suitable are the study of how information flows through a network (Jackson and López-Pintado 2013), and also the analysis of risk sharing (Fafchamps and Gubert 2007, Jackson et al. 2012) and financial contagion (Allen and Gale 2000, Acemoglu et al. 2015).

Dzemski (2014) studies the same model as we do here. Adapting techniques introduced by Fernández-Val and Weidner (2016) he constructs an estimator for θ_0 that is applicable to more general specifications of the distribution of the idiosyncratic disturbance than the logistic distribution. On the other hand, his estimation approach is designed for dense networks. The estimation strategy developed below is targeted to the logistic specification but can explicitely handle sparse networks, where the probability of link formation shrinks to zero with the sample size.

Note that, by independence of the idiosyncratic errors in (2.2), conditional on node and dyad characteristics, links between nodes are formed independently. This means that any observed dependence across link decisions must come from the presence of the node and dyad characteristics. Moreover, the model does not reflect clustering phenomena like transitivity, where two nodes are more likely to be linked if there is more overlap between the sets of nodes they are already linked to. Transitivity is the subject of a recent literature; see, for example, Jackson and Rogers (2007) for theoretical work on social networks and Chaney (2014) and Morales et al. (2015) for work in the context of international trade. However, it also creates identification challenges when only a single network is observed (Goldsmith-Pinkham and Imbens 2013, Graham 2013, 2016). A specification test for our model is given by Dzemski (2014) (extending an approach by Holland and Leinhardt 1978),

and our estimator can serve as a useful plug-in estimator to his test statistic. This test statistic, however, requires estimates of the node-specific fixed effects, and its asymptotic properties are only known for dense networks.

3 Conditional likelihood

Treating $\{\alpha_i, \gamma_i\}_n$ as parameters and jointly estimating them with the common parameter θ_0 leads to an incidental-parameter problem (Neyman and Scott, 1948). For dense networks, where the probability of link formation is bounded away from zero and one, Dzemski (2014) characterizes the asymptotic bias in the maximum-likelihood estimator of θ_0 and develops bias-reduction methods by building on the work of Fernández-Val and Weidner (2016) on two-way models for panel data. For the sparse case, where the probability of link formation is allowed to shrink to zero with n, the behavior of the maximum-likelihood estimator is more complicated and no results are available. The problem here is that the node-specific parameters may not be consistently estimable or may be estimable only at a very slow rate. The statistical properties of the maximum-likelihood estimator in such cases are not obvious and are currently an open question.

On the other hand, Charbonneau (2014) shows the existence of a sufficient statistic for the pair (α_i, γ_j) in our setting by building on the work of Cox (1958), Rasch (1960, 1961), and Hirji et al. (1987). This allows to bypass estimation of the fixed effects to infer θ_0. Our aim here is to develop the implied estimator and to derive its statistical properties. To motivate the estimator we first present the sufficiency result developed by Charbonneau (2014). We turn to estimation and inference from observed network data in the next section.

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)**.**

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

Proof. See the Appendix or Charbonneau (2014). □

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 (quasi) 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.

The conclusion of Lemma 1 depends crucially on the fact the node-specific heterogeneity parameters α_i, γ_j enter the surplus u_{ij} in an additive manner. A more general specification of our model would have $u_{ij} = x_{ij}'\theta_0 + d(\alpha_i, \gamma_j) - \epsilon_{ij}$, where $d : \mathcal{R}^2 \to \mathcal{R}$ is a known function. For such a specification, Lemma 1 fails to difference-out the node-specific parameters, in general. The estimation of models with node-specific parameters entering the surplus in a non-additive manner is the subject of an active literature. Following Fernández-Val and Weidner (2016), recent work by Chen et al. (2014) looks at the case where we have $d(\alpha, \gamma) = \alpha\gamma$ in the context of panel data models with two-way fixed effects. Having access to repeated measurements can offer a solution here. Indeed, suppose we observe the network for multiple time periods; at time t the surplus is $u_{ij,t} = x_{ij,t}'\theta_0 + d(\alpha_i, \gamma_j) - \epsilon_{ij,t}$. If the errors are independent across time, we can collapse the data across (i, j) to get $u_{\sigma\{i,j\},t} = x_{\sigma\{i,j\},t}'\theta_0 + \delta_{\sigma\{i,j\}} - \epsilon_{\sigma\{i,j\},t}$, where $\sigma : \mathbb{N}_n \times \mathbb{N}_{n-1} \to \mathbb{N}_{n(n-1)}$ ranges across all dyads and $\delta_{\sigma\{i,j\}} = d(\alpha_i, \gamma_j)$. This is a one-way panel data model (albeit with cross-sectional dependence) to which the conditioning argument of Rasch (1960, 1961) can be applied. On the other hand, if we wish to allow for time effects, an alternative specification would have $u_{ij,t} = x_{ij,t}'\theta_0 + d(\alpha_i, \gamma_j) + \eta_t - \epsilon_{ij,t}$. Again collapsing the data across all dyads (i, j) gives a model to which Lemma 1 can be applied.

Panel data can also allow to model certain forms of reciprocity and transitivity in network formation; see Graham (2016).

The sufficiency result in Lemma 1 also uses the logistic specification and independence

of the errors ϵ_{ij} across both i and j. Relaxing this assumption is possible to a certain extent, and is the topic of ongoing work.

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

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

such quadruples. It will prove useful to introduce a function σ that maps these quadruples to the index set $\mathbb{N}_{m_n} = \{1, 2, \ldots, m_n\}$. Thus, each distinct quadruple of nodes $\{i_1, i_2; j_1, j_2\}$ corresponds to a unique $\sigma\{i_1, i_2; j_1, j_2\} \in \mathbb{N}_{m_n}$. 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 $\mathbb{N}_{m_n}$.

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

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

This objective function is a standard logit log-likelihood applied to the

$$m_n^* = \sum_{\sigma \in \mathbb{N}_{m_n}} 1\{z_\sigma \in \{-1, 1\}\}$$

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 $\{z_\sigma, r_\sigma\}_{m_n}$, which is easy to do.

Note that $L_n(\theta)$ is a quasi likelihood. It can be shown that the number of incoming and outgoing links of all nodes forms a sufficient statistic for the node-specific parameters $\{\alpha_i, \gamma_i\}_n$, in the sense that the conditional-likelihood function does not depend on the fixed effects. This extends results on the fixed-effect logit model in Chamberlain (1980). However, the resulting likelihood function is computationally intractable. This is why we work with the quasi likelihood for quadruples.

The conditional-logit estimator is consistent under weak conditions.

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

This assumption is a natural sampling scheme for network data. It permits dependence of the covariates across dyads that have nodes in common. Note that we do not require that nodes are sampled from the same distribution.

The second assumption is conventional for establishing consistency in nonlinear models; see, for example, Newey and McFadden (1994).

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

The third assumption requires the existence of second moments.

Assumption 3 (Moments)**.** *For all $(i, j) \in \mathbb{N}_n \times \mathbb{N}_n$, $E(\|x_{ij}\|^2) < C$, where C is a finite constant.*

The fourth assumption ensures that θ_0 is identified. To state it, note that m_n^* is a random variable; we write

$$p_n = \frac{E(m_n^*)}{m_n} = \frac{\sum_{\sigma \in \mathbb{N}_{m_n}} \Pr(z_\sigma \in \{-1, 1\})}{m_n}$$

for the expected fraction of quadruples in the data that contribute to the log-likelihood. We denote the logistic density function by f.

Assumption 4 (Identification)**.** *$np_n \to \infty$ as $n \to \infty$ and*

$$\operatorname{rank}\left\{\lim_{n\to\infty}(m_n p_n)^{-1}\sum_{\sigma\in\mathbb{N}_{mn}} E(r_\sigma r_\sigma' f(r_\sigma'\theta_0)\,1\{z_\sigma \in \{-1,1\}\})\right\} = \dim\theta.$$

The first part of Assumption 4 allows p_n, the (expected) fraction of informative quadruples, to shrink to zero as the sample size grows. Note that $\Pr(z_\sigma \in \{-1,1\})$ depends on the fixed effects involved in the quadruple σ. If these parameters become unbounded as n grows, adding nodes to the network may not provide additional information on θ_0. Assumption 4 allows for such sequences and, as such, our approach can be applied to sparse networks. The requirement that p_n does not shrink faster than n^{-1} is needed to ensure uniform convergence of $L_n(\theta)$. Note that, as $m_n = O(n^4)$, this rate condition implies that $E(m_n^*) = m_n p_n \to \infty$, so that the accumulation of informative quadruples does not cease as the sample grows. The second part of Assumption 4 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.

To gain some insight into the rate condition in Assumption 4, consider sequences of fixed effects where α_i and γ_i tend to $-\infty$ and suppose that the covariates have bounded support. By the exponential tails of the logistic distribution we have that, as n increases,

$$p_n \sim \left(\frac{\sum_{i=1}^n e^{\alpha_i}}{n}\right)^2 \left(\frac{\sum_{i=1}^n e^{\gamma_i}}{n}\right)^2,$$

and we can translate the condition that $np_n \to \infty$ into a restriction on the growth rate of the fixed effects. One example that satisfies our condition would be $\alpha_i = -a_n \log i$ and $\gamma_i = -b_n \log i$ for non-negative sequences a_n and b_n that are bounded from above by $1/4$. To link our rate condition to the link probability $q_n = \sum_{i=1}^n \sum_{j\neq i} \Pr(y_{ij}=1)/n(n-1)$, note that, in the left tail,

$$q_n \sim \frac{\sum_{i=1}^n e^{\alpha_i}}{n}\,\frac{\sum_{i=1}^n e^{\gamma_i}}{n}.$$

Hence, $q_n \sim \sqrt{p_n}$ in this case.

A similar exercise can be done for sequences of fixed effects growing to ∞ 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. From a statistical perspective such networks are equally challenging as sparse networks as, again, variation in the number of links diminishes and information on θ_0 may not accrue or only do so very slowly.

Theorem 1 formally states our consistency result.

Theorem 1 (Consistency)**.** *Let Assumptions 1–4 hold. Then* $\theta_n \xrightarrow{p} \theta_0$ *as* $n \to \infty$.

Proof. See the Appendix. □

To perform hypothesis testing on the homophily parameter we move on to deriving distribution theory for θ_n. We note that, 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. First, a sandwich-form variance estimator will be required; recall that $L_n(\theta)$ is a quasi log-likelihood, so the information equality will not hold. Second, the score vector involves a sum over quadruples of nodes, with the same nodes showing up in multiple quadruples. This induces dependence across the summands in $S_n(\theta)$ that cannot be ignored. An estimation problem with the same structure arises in Jochmans (2016), and we follow a similar strategy as taken there in deriving the distribution theory to follow.

To do so we strengthen the moment requirement in Assumption 3 as follows.

Assumption 5 (Moments, cont'd)**.** *For all* $(i,j) \in \mathbb{N}_n \times \mathbb{N}_n$, $E(\|x_{ij}\|^6) < C$, *where* C *is a finite constant.*

Introduce

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

Note that $s(\sigma\{i_1,i_2;j_1,j_2\};\theta)$ is permutation invariant in both senders (i_1,i_2) and receivers (j_1,j_2). We may then write the score vector as

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

The key to characterizing the limit distribution of the conditional-logit estimator is the result that $\Upsilon_n(\theta_0)^{-1/2} S_n(\theta_0) \xrightarrow{d} N(0, I)$, where

$$\Upsilon_n(\theta) = \sum_{i=1}^{n} \sum_{j \neq i} v_{ij}(\theta)\, v_{ij}(\theta)', \qquad v_{ij}(\theta) = \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} s(\sigma\{i, i'; j, j'\}; \theta),$$

and I denotes the $\dim \theta \times \dim \theta$ identity matrix. The Hessian matrix, in turn, is given by

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

and, on defining

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

we arrive at the following result.

Theorem 2 (Asymptotic distribution)**.** *Let Assumptions 1–5 hold. Then* $\|\theta_n - \theta_0\| = O_p(1/\sqrt{n(n-1)p_n})$ *and*

$$\Omega_n^{-1/2} \left(\theta_n - \theta_0\right) \xrightarrow{d} N(0, I)$$

as $n \to \infty$*.*

Proof. See the Appendix. □

In the dense case, where p_n is bounded away from zero, $\|\theta_n - \theta_0\| = O_p(n^{-1})$ holds, and so the conditional-logit estimator converges at the parametric rate. In the sparse case p_n shrinks with n and we have a slower rate depending on how slow the number of informative quadruples grows. In general, $\|\theta_n - \theta_0\| = O_p(n/\sqrt{E(m_n^*)})$. Because Assumption 4 requires that $E(m_n^*)$ grows at least at the rate n^3, the convergence rate of the estimator can be arbitrarily close to, but will be faster than, $n^{-1/2}$. In the statement of the theorem the estimator is self-normalized so, in practice, inference proceeds in the same way in the dense and sparse case.

The result of Theorem 2 is qualitatively similar to that of Graham (2015, Theorem 1) for his estimator for undirected networks.

5 Simulations

We evaluated the small-sample performance of the conditional-logit estimator through a series of numerical experiments. Here, we present results for designs similar to those in Dzemski (2014) and Yan et al. (2016). Other designs yielded the same conclusions. We generate the single regressor as

$$x_{ij} = -|u_i - u_j|,$$

where $u_i = v_i - \frac{1}{2}$ for $v_i \sim \text{Beta}(2,2)$, and set $\theta_0 = 1$. Note that the covariate is generated in such a way that it is dependent across both senders and receivers of links. The fixed effects are set as a deterministic function of the sample size, as

$$\alpha_i = -\frac{n-i}{n-1} C_n, \qquad \gamma_i = \alpha_i,$$

for a constant C_n that depends on n. We will consider sample sizes $n \in \{25, 50\}$ and constants $C_n \in \{0, \log(\log(n)), \log(n)^{1/2}, \log(n)\}$ in our experiments. These configurations are chosen to illustrate estimation of dense and sparse networks.

In Table 1 we provide summary statistics of the (average) degree distributions for the different constants C_n and sample sizes n. Note that the in-degree and out-degree have the same distribution by symmetry of the data generating process, and so the table applies to both. The table provides the mean, the quartiles, and the minimum and maximum of the degree (normalized by n) distributions (as computed by simulation). We also show the average value of np_n—which needs to grow with n to satisfy Assumption 4—and of nq_n—(the inverse of) the order of the standard deviation of the conditional-likelihood estimator as $n \to \infty$.

When $C_n = 0$ node heterogeneity is absent and the model gives rise to dense networks. The probability of link formation, q_n, is bounded away from zero and one and our estimator converges at the parametric rate. For $C_n > 0$ we have that q_n and p_n shrink to zero as n grows. The larger C_n, the more the degree distribution concentrates mass closer to zero, and the less variability in link decisions will be observed in the data. As such, larger values

of C_n yield more sparse networks. The severity of the sparsity for the different choices of C_n is clear from Table 1. For $C_n = \log(\log(n))$ the average network has roughly half as many links as in the dense configuration, and this ratio goes down to about one third for $C_n = \log(n)^{1/2}$. When $C_n = \log(n)$ it drops to about 10%. While the former two designs are sparse in that q_n decreases as n grows, np_n grows (albeit slower than n) and so these designs behave as required by Assumption 4. In contrast, Table 1 shows that, for $C_n = \log(n)$, np_n stays roughly constant as n doubles. As such, this design serves to capture the knife-edge case where our regularity condition on the degree of sparsity breaks down.

Table 1: Degree distributions for simulated data

C_n	mean	1st quartile	median	3$^{\text{th}}$ quartile	minimum	maximum	np_n	nq_n
				$n = 25$				
0	0.439	0.372	0.438	0.506	0.247	0.637	3.020	10.980
$\log(\log(n))$	0.208	0.137	0.200	0.272	0.045	0.418	1.230	5.203
$\log(n)^{1/2}$	0.138	0.071	0.126	0.195	0.007	0.343	0.585	3.457
$\log(n)$	0.062	0.005	0.042	0.097	0.000	0.233	0.118	1.545
				$n = 50$				
0	0.438	0.389	0.438	0.486	0.279	0.600	6.030	21.897
$\log(\log(n))$	0.181	0.122	0.174	0.234	0.041	0.373	1.985	9.073
$\log(n)^{1/2}$	0.122	0.064	0.110	0.170	0.008	0.310	0.970	6.086
$\log(n)$	0.043	0.002	0.025	0.067	0.000	0.191	0.120	2.159

We give simulation results for the conditional-likelihood estimator (CMLE), as well as for the maximum-likelihood estimator (MLE) and its bias-corrected version (BC), obtained using the formula in Dzemski (2014) and Yan et al. (2016). For each of these estimators we compute the mean, median, standard deviation (std), and interquartile range (iqr) over $1,000$ Monte Carlo replications for all designs. We also give the ratio of the average estimated standard error to the Monte Carlo standard deviation (se/std) for each of these

estimators. For CMLE the standard error is computed as discussed in the previous section. For MLE and BC we use the inverse of the Fisher information as estimated by maximum likelihood.

Table 2 contains the simulation results for all designs and sample sizes. The MLE clearly suffers from upward bias in all designs. Bias correction is effective in recentering the point estimator when fixed effects are small—that is, in dense networks—but its performance deteriorates as C_n increases and the fixed effects become harder to estimate. For example, for $C_n = \log(n)$, BC is effectively more biased than MLE, both for $n = 25$ and for $n = 50$. CMLE performs similarly as does BC, in terms of both location and spread, in the dense case. However, it does not suffer from a dramatic increase in bias in the other cases. Furthermore, the variance estimator of CMLE captures well the small-sample variability in the point estimator. Consequently, the asymptotic argument of Theorem 2 yields reliable inference.

6 Empirical applications

6.1 A trade network

As a first 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 4 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

Table 2: Simulation results

	MLE	CMLE	BC	MLE	CMLE	BC	MLE	CMLE	BC	MLE	CMLE	BC
	$n = 25$						$n = 50$					
	$C_n = 0$			$C_n = \log(\log(n))$			$C_n = 0$			$C_n = \log(\log(n))$		
mean	1.110	1.022	1.021	1.074	0.986	0.970	1.057	1.016	1.015	1.039	0.995	0.991
median	1.097	1.023	1.009	1.067	0.979	0.961	1.054	1.012	1.012	1.043	0.995	0.995
std	0.652	0.604	0.599	0.812	0.756	0.733	0.289	0.277	0.277	0.384	0.368	0.365
iqr	0.835	0.786	0.768	1.024	0.947	0.930	0.398	0.385	0.382	0.513	0.490	0.486
se/std	0.938	1.058	1.020	0.958	1.087	1.061	0.976	1.039	1.017	0.979	1.048	1.028
	$C_n = \log(n)^{1/2}$			$C_n = \log(n)$			$C_n = \log(n)^{1/2}$			$C_n = \log(n)$		
mean	1.088	0.978	0.948	1.134	0.968	0.817	1.035	0.987	0.976	1.0639	0.9919	0.9134
median	1.120	1.016	0.979	1.118	0.972	0.799	1.050	0.997	0.991	1.0473	0.9831	0.8765
std	1.042	0.956	0.911	1.896	1.702	1.323	0.483	0.462	0.455	0.8675	0.8136	0.7451
iqr	1.331	1.179	1.190	2.161	1.886	1.566	0.659	0.627	0.620	1.0962	1.0116	0.9534
se/std	0.912	1.062	1.044	0.835	1.018	1.197	0.940	1.016	0.997	0.9199	1.0368	1.0709

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).

About 50% of all potential bilateral-trade routes are open. All countries in the data trade at least with one country. Table 3 provides summary statistics of the out-degree and in-degree distributions, as well as of their difference and absolute difference. The table reveals some heterogeneity in the number of export and import partners.

Table 3: Degree distributions for trade data

	mean	1st quartile	median	3$^{\text{th}}$ quartile	minimum	maximum
out degree	0.524	0.348	0.485	0.641	0.156	1
in degree	0.524	0.289	0.441	0.785	0.089	1
difference	0	-0.063	0.000	0.067	-0.259	0.344
abs. difference	0.085	0.030	0.067	0.133	0.000	0.259

Table 4: Descriptive statistics for trade data

	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

Table 5: Trade estimates

	MLE	CMLE
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)

the point estimates), are collected in Table 5. 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.

A, perhaps, 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 dyads) have established preferential trade agreements; see Table 4. In the raw data, the dyads that allow to discriminate between the impact of *common border* and *preferential trade agreement* 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.

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. These findings are in line with the Monte Carlo results reported on above. 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%. Thus the difference is quite sizeable. This confirms the importance of appropriately controlling for the presence of country fixed effects in trade applications.

6.2 An advice network

As a second empirical illustration we estimate an advice network among 71 attorneys employed in a Northeastern U.S. law firm, with offices in Boston, Hartford, and Providence. The data are a survey taken from Lazega (2001). For dyad (i, j), the outcome variable, *advice*, is a binary indicator that takes the value one if i has indicated that he or she has consulted j for professional advice. The degree distributions are summarized in Table 6. It may be noted that the average number of links is much smaller relative to the sample size than it was in the previous application. As such, the advice network serves as an illustration of a more sparse network.

The data set contains information on the status of the attorneys in the firm (whether they are a partner or associate) and which of the three offices (either Boston, Hartford, or Providence) they work in, as well as their gender, tenure in the firm, and their age. From this we construct the regressors *same status*, *same gender*, *same office* , *difference in tenure*, and *difference in age* at the dyad level. The definition of each of these regressors is obvious. Note that, for the last two of these variables, we take the absolute value of the difference in tenure and age, respectively. As such, the dyad characteristics are symmetric in (i, j). Table 7 contains descriptive statistics for each of the variables.

Table 6: Degree distributions for advice data

	mean	1st quartile	median	3$^{\text{th}}$ quartile	minimum	maximum
out degree	0.179	0.075	0.157	0.282	0.000	0.529
in degree	0.179	0.100	0.171	0.243	0.000	0.429
difference	0.000	-0.086	0.000	0.082	-0.343	0.300
abs. difference	0.114	0.029	0.086	0.168	0.000	0.343

Table 7: Descriptive statistics for advice data

	mean	standard deviation
advice	0.1759	0.3838
same status	0.4930	0.5000
same gender	0.6161	0.4864
same office	0.5252	0.4994
difference in tenure	10.4773	8.6519
difference in age	11.6821	8.5912

Table 8: Advice estimates

	MLE	CMLE
same status	0.9577	0.9409
	(0.1259)	(0.1349)
same gender	0.2438	0.1801
	(0.1254)	(0.1303)
same office	2.2098	1.9570
	(0.1251)	(0.1380)
difference in tenure	−0.0401	−0.0330
	(0.0103)	(0.0120)
difference in age	−0.0165	−0.0150
	(0.0085)	(0.0092)

We again estimated the parameters of this model by maximum likelihood and by the conditional-likelihood approach developed here. Table 8 contains the point estimates and standard errors. The estimated signs are all in line with what would be expected and confirm the presence of homophily among the attorneys. Moreover, in order of estimated importance, an attorney is more likely to be consulted if he works in the same regional office, if he has the same status, and if he is of the same gender. The latter estimate is not significantly different from zero at the 5% level, however. He is less likely to be asked for advice the larger are the differences in tenure and age, with tenure being the more dominant of the two, and the importance of age is not significantly different from zero at the 5% level.

Appendix

Proof of Lemma 1. Equations (2.1)–(2.2) together with the functional form of the 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. By virtue of Assumption 4, θ_0 is the unique global maximizer of the limit quantity $\lim_{n\to\infty}(m_np_n)^{-1}E(L_n(\theta))$ on Θ. Because this function is concave, $\theta_n \xrightarrow{p} \theta_0$ will follow from pointwise convergence in probability of $(m_n^*)^{-1}L_n(\theta)$ (the normalized objective function) to $(m_np_n)^{-1}E(L_n(\theta))$ (Newey and McFadden, 1994, Theorem 2.7). We proceed by showing that this is the case.

Write

$$L_n(\theta) = \sum_{\sigma\in\mathbb{N}_{mn}} \ell_\sigma(\theta),$$

where $\ell_\sigma(\theta)$ denotes the log-likelihood contribution of quadruple σ. Then

$$\frac{L_n(\theta)}{m_n^*} - \frac{E(L_n(\theta))}{E(m_n^*)} = \frac{\sum_{\sigma\in\mathbb{N}_{mn}}\ell_\sigma(\theta) - E(\ell_\sigma(\theta))}{E(m_n^*)} + \frac{\sum_{\sigma\in\mathbb{N}_{mn}}\ell_\sigma(\theta)}{E(m_n^*)}\left(\frac{E(m_n^*)}{m_n^*}-1\right) \quad \text{(A.1)}$$

and it suffices to show that each of the right-hand side terms in this expression converges to zero in probability.

For the first right-hand side term in (A.1), note that $|\ell_\sigma(\theta)| \leq \log 2 + 2\|r_\sigma\| \, \|\theta\|$. Because $E(\|r_\sigma\|^2)$ is finite and Θ is compact, it follows that the variance of $\ell_\sigma(\theta)$ exists and is uniformly bounded in σ. Therefore, by Chebychev's inequality, it holds that, for any $\epsilon > 0$,

$$\Pr\left(\left|\frac{\sum_{\sigma\in\mathbb{N}_{mn}} \ell_\sigma(\theta) - E(\ell_\sigma(\theta))}{E(m_n^*)}\right| > \epsilon\right) \leq \frac{1}{\epsilon^2} \frac{E(|\sum_{\sigma\in\mathbb{N}_{mn}} \ell_\sigma(\theta) - E(\ell_\sigma(\theta))|^2)}{E(m_n^*)^2},$$

for each $\theta \in \Theta$. Now, $E(|\sum_{\sigma\in\mathbb{N}_{mn}} \ell_\sigma(\theta) - E(\ell_\sigma(\theta))|^2)$ equals

$$E\left(\left(\sum_{\sigma\in\mathbb{N}_{mn}} \ell_\sigma(\theta) - E(\ell_\sigma(\theta))\right)\left(\sum_{\sigma'\in\mathbb{N}_{mn}} \ell_{\sigma'}(\theta) - E(\ell_{\sigma'}(\theta))\right)\right)$$

and a pair of quadruples $\sigma = \sigma\{i_1, i_2, j_1, j_2\}$ and $\sigma' = \sigma\{i_1', i_2', j_1', j_2'\}$ can 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. By using the Cauchy-Schwarz inequality and Jensen's inequality their contribution to the total variance is found to be bounded by a multiple of

$$n^3 \sum_{\sigma\in\mathbb{N}_{mn}} E((\ell_\sigma(\theta) - E(\ell_\sigma(\theta)))^2) = O(n^3 m_n p_n),$$

where the last equality follows because $E((\ell_\sigma(\theta) - E(\ell_\sigma(\theta)))^2) = O(\Pr(z_\sigma \in \{-1, 1\}))$ for each $\theta \in \Theta$ and all $\sigma \in \mathbb{N}_{mn}$. As $E(m_n^*) = m_n p_n$ and $m_n = O(n^4)$ we find that

$$\frac{E(|\sum_{\sigma\in\mathbb{N}_{mn}} \ell_\sigma(\theta) - E(\ell_\sigma(\theta))|^2)}{E(m_n^*)^2} = O\left(\frac{1}{np_n}\right),$$

which converges to zero by Assumption 4. Therefore,

$$\lim_{n\to\infty} \Pr\left(\left|\frac{\sum_{\sigma\in\mathbb{N}_{mn}} \ell_\sigma(\theta) - E(\ell_\sigma(\theta))}{E(m_n^*)}\right| > \epsilon\right) = 0$$

for any $\epsilon > 0$ and all $\theta \in \Theta$.

For the second right-hand side term in (A.1), recall that

$$m_n^* = \sum_{\sigma \in \mathbb{N}_{m_n}} 1\{z_\sigma \in \{-1, 1\}\}.$$

The summands are bounded uniformly in σ and do not depend on θ. Following the same argument as in the previous paragraph it is readily verified that $(m_n^*/m_n - p_n) \xrightarrow{p} 0$, and so $m_n^*/E(m_n^*) \xrightarrow{p} 1$.

By (A.1) we have $\lim_{n\to\infty} \Pr(|(m_n^*)^{-1}L_n(\theta) - (m_n p_n)^{-1}E(L_n(\theta))| > \epsilon) = 0$ for any $\epsilon > 0$ and all $\theta \in \Theta$, so that $\theta_n \xrightarrow{p} \theta_0$ as $n \to \infty$. The proof is complete. □

Proof of Theorem 2. The proof of the theorem proceeds in four main steps. First we show that the score vector, evaluated at the true parameter value and properly normalized, is asymptotically equivalent to its Hájek projection (conditional on the covariates). Second we establish the limit distribution of this projection and show that the matrix $\Upsilon_n(\theta_n)$ is a consistent estimator of its variance. Third we prove that the Hessian of the conditional likelihood, normalized by m_n^*, converges to a well-behaved limit uniformly on Θ. Finally, we collect these results and combine them with a mean-value expansion of the first-order condition around the true value in the usual manner to arrive at the limit distribution given in Theorem 2.

(i) Projection of the score vector. The Hájek projection of $S_n(\theta_0)$, conditional on the covariate sequence $\{x_{ij}\}_{n,n}$, is

$$V_n = \sum_{i=1}^{n} \sum_{i' \neq i} \sum_{j \neq i,i'} \sum_{j' \neq i,i',j} E(s(\sigma\{i, i'; j, j'\}) |\, y_{ij}, \{x_{ij}\}_{n,n}) = \sum_{i=1}^{n} \sum_{j \neq i} v_{ij}$$

where we have introduced the random variables

$$v_{ij} = \frac{w_{ij}}{\Pr(y_{ij} = 1|x_{ij})\ \Pr(y_{ij} = 0|x_{ij})} \left(y_{ij} - \Pr(y_{ij} = 1|x_{ij})\right)$$

with

$$w_{ij} = \sum_{i' \neq i,j} \sum_{j' \neq i,j,i'} r(\sigma\{i, i'; j, j'\})\, q(\sigma\{i, i'; j, j'\}), \quad q(\sigma) = \frac{\Pr(z_\sigma = 1|x_\sigma)\ \Pr(z_\sigma = -1|x_\sigma)}{\Pr(z_\sigma \in \{-1, 1\}|x_\sigma)}.$$

Here, we abuse notation slightly by denoting by x_σ the collection of covariates for the nodes in the quadruple σ. This result follows from a small calculation and uses the fact that

$$\begin{aligned} \Pr(z_\sigma = \quad 1|x_\sigma) = \quad & F(r_\sigma'\theta_0) \qquad \Pr(z_\sigma \in \{-1,1\}|\, x_\sigma), \\ \Pr(z_\sigma = -1|x_\sigma) = \big(1 - & F(r_\sigma'\theta_0)\big) \Pr(z_\sigma \in \{-1,1\}|\, x_\sigma), \end{aligned} \tag{A.2}$$

which follows from Lemma 1. By iterating expectations, we find $E(v_{ij}) = 0$, and so $E(V_n) = 0$. Also, because link decisions are conditionally independent, $E(v_{ij}v_{i'j'}') = 0$ unless $i = i'$ and $j = j'$. Therefore,

$$\Upsilon = E(V_n V_n') = \sum_{i=1}^{n}\sum_{j\neq i} E(v_{ij}v_{ij}') = \sum_{i=1}^{n}\sum_{j\neq i} E\left(\frac{w_{ij}\, w_{ij}'}{\Pr(y_{ij}=1|x_{ij})\,\Pr(y_{ij}=0|x_{ij})}\right).$$

Below we show that $\Upsilon^{-1/2}V_n \xrightarrow{d} N(0,I)$. Here we show that $\Upsilon^{-1/2}V_n$ and $\Upsilon^{-1/2}S_n(\theta_0)$ are asymptotically equivalent.

To establish asymptotic equivalence we show that

$$\lim_{n\to\infty} \Upsilon^{-1/2} E\big((V_n - S_n(\theta_0))\,(V_n - S_n(\theta_0))'\big)\,\Upsilon^{-1/2} = 0. \tag{A.3}$$

The main step in doing so is calculating the variance of the score vector, $E(S_n(\theta_0)S_n(\theta_0)')$. Because $E[s(\sigma;\theta_0)|x_\sigma] = 0$ for all $\sigma \in \mathbb{N}_{m_n}$ 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 the leading term of $E(S_n(\theta_0)S_n(\theta_0)')$ 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. Note that, 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. The leading term of $E(S_n(\theta_0)S_n(\theta_0)')$ then is

$$A = \sum_{i=1}^{n}\sum_{j\neq i}\left(\sum_{i'\neq i,j}\sum_{j'\neq i,i',j}\sum_{i''\neq i,j}\sum_{j''\neq i,i'',j} E\left(s(\sigma\{i,i';j,j'\};\theta_0)\, s(\sigma\{i,i'';j,j''\};\theta_0)'\right)\right),$$

where we have exploiting symmetry of $s(\sigma;\theta)$ in senders and receivers once again to expand the sums. Fix $\sigma = \sigma\{i, i'; j, j'\}$ and $\sigma' = \sigma\{i, i''; j, j''\}$. Then

$$\begin{aligned} s(\sigma;\theta_0)\, s(\sigma';\theta_0)' = {} & r_\sigma r'_{\sigma'}\, 1\{z_\sigma = \ \ 1, z_{\sigma'} = \ \ 1\}\, (1 - F(r'_\sigma\theta_0))\, (1 - F(r'_{\sigma'}\theta_0)) \\ & + r_\sigma r'_{\sigma'}\, 1\{z_\sigma = -1, z_{\sigma'} = -1\} \quad F(r'_\sigma\theta_0) \quad\ F(r'_{\sigma'}\theta_0) \\ & - r_\sigma r'_{\sigma'}\, 1\{z_\sigma = \ \ 1, z_{\sigma'} = -1\}\, (1 - F(r'_\sigma\theta_0)) \quad F(r'_{\sigma'}\theta_0) \\ & - r_\sigma r'_{\sigma'}\, 1\{z_\sigma = -1, z_{\sigma'} = \ \ 1\} \quad F(r'_\sigma\theta_0) \quad (1 - F(r'_{\sigma'}\theta_0)). \end{aligned} \tag{A.4}$$

Take expectations given covariates. The last two terms on the right-hand side of (A.4) drop out, while the expectations of the first and second right-hand side term are equal to

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

and

$$r_\sigma\, r'_{\sigma'}\, \frac{F(r'_\sigma\theta_0)\, (1 - F(r'_\sigma\theta_0))\, F(r'_{\sigma'}\theta_0)\, (1 - F(r'_{\sigma'}\theta_0))}{\Pr(y_{ij} = 0|x_{ij})} \Pr(z_\sigma \in \{-1, 1\}|\, x_\sigma)\ \Pr(z_{\sigma'} \in \{-1, 1\}|\, x_{\sigma'}),$$

respectively. By (A.2), and recalling 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

$$E(s(\sigma;\theta_0)\, s(\sigma';\theta_0)'\,|\, x_\sigma, x_{\sigma'}) = r_\sigma r'_{\sigma'}\, \frac{q(\sigma)\, q(\sigma')}{\Pr(y_{ij} = 1|x_{ij})\Pr(y_{ij} = 0|x_{ij})}.$$

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

$$A = \sum_{i=1}^{n} \sum_{j \neq i} E\left(\frac{w_{ij} w'_{ij}}{\Pr(y_{ij} = 1|x_{ij})\Pr(y_{ij} = 0|x_{ij})} \right) = \Upsilon.$$

Thus, $\Upsilon^{-1/2} E(S_n(\theta_0)\, S_n(\theta_0)')\, \Upsilon^{-1/2} = I + o(1)$. Making use of the above calculations, it is readily deduced that we equally have that $\Upsilon^{-1/2} E(V_n\, S_n(\theta_0)')\, \Upsilon^{-1/2} = I + o(1)$, that is, that the asymptotic covariance between V_n and $S_n(\theta_0)$ equals their variance. Put together, these results imply (A.3).

(ii) Limit distribution of the projection. Recall that the v_{ij} are zero mean and independent conditional on $\{x_{ij}\}_{n,n}$. Let

$$\Upsilon_X = \sum_{i=1}^{n} \sum_{j \neq i} E(v_{ij} v_{ij}' | \{x_{ij}\}_{n,n}) = \sum_{i=1}^{n} \sum_{j \neq i} \frac{w_{ij} w_{ij}'}{\Pr(y_{ij} = 1 | x_{ij}) \Pr(y_{ij} = 0 | x_{ij})}.$$

By a conditional version of Lyapunov's central limit theorem (see, e.g., Prakasa Rao 2009),

$$\Upsilon_X^{-1/2} V_n \xrightarrow{d} N(0, I) \tag{A.5}$$

conditional on the covariates. Now, using Assumption 5, it is easy to see that $\|\Upsilon_X - \Upsilon\| \xrightarrow{p} 0$ as $n \to \infty$. Hence, the limit result is independent of the covariate values, and (A.5) continues to hold unconditionally, with Υ replacing Υ_X.

The matrix $\Upsilon_n(\theta_n)$ as defined in the main text is a plug-in estimator of Υ based on the matrix A given above. Using the same arguments as those used to establish convergence of the normalized Hessian in the next section it is straightforward to show that this estimator is consistent. Therefore,

$$\Upsilon_n(\theta_n)^{-1/2} V_n \xrightarrow{d} N(0, I)$$

as $n \to \infty$ by an application of Slutsky's theorem.

(iii) Convergence of the Hessian. Recall that the Hessian is

$$H_n(\theta) = \sum_{\sigma \in \mathbb{N}_{m_n}} r_\sigma r_\sigma' \, f(r_\sigma' \theta) \, 1\{z_\sigma \in \{-1, 1\}\}.$$

We need to show that

$$\sup_{\theta \in \Theta} \left\| \frac{H_n(\theta)}{m_n^*} - \frac{E(H_n(\theta))}{m_n p_n} \right\| \xrightarrow{p} 0$$

as $n \to \infty$. The matrix $\lim_{n \to \infty} (m_n p_n)^{-1} E(H_n(\theta_0))$ 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

$$\frac{\sup_{\theta \in \Theta} \|H_n(\theta) - E(H_n(\theta))\|}{m_n p_n} \xrightarrow{p} 0$$

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

$$\frac{\|H_n(\theta_1) - H_n(\theta_2)\|}{m_n p_n} \leq \left((m_n p_n)^{-1} \sum_{\sigma \in \mathbb{N}_{mn}} \|r_\sigma\|^3 \, 1\{z_\sigma \in \{-1, 1\}\} \right) \sup_{\epsilon \in \mathcal{R}} \left| \frac{\partial f(\epsilon)}{\partial \epsilon} \right| \|\theta_1 - \theta_2\|$$

for any $\theta_1, \theta_2 \in \Theta$. Next, using the same arguments as those used to establish Theorem 1 we find that

$$(m_n p_n)^{-1} \sum_{\sigma \in \mathbb{N}_{mn}} \|r_\sigma\|^3 \, 1\{z_\sigma \in \{-1, 1\}\} = O_p(1),$$

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

$$\frac{\|H_n(\theta_1) - H_n(\theta_2)\|}{m_n p_n} = O_p(1) \, \|\theta_1 - \theta_2\|$$

for any $\theta_1, \theta_2 \in \Theta$. Thus, the Hessian matrix is stochastically equicontinuous. This implies that uniform convergence follows from pointwise convergence on Θ. Assumption 5 implies that $E(\|r_\sigma\|^4 | z_\sigma \in \{-1, 1\})$ is uniformly bounded in σ while f is bounded uniformly on $\mathcal{R}$. Therefore, the same arguments as those used to establish Theorem 1 yield the convergence result

$$\frac{\|H_n(\theta) - E(H_n(\theta))\|}{m_n p_n} \xrightarrow{p} 0$$

for all $\theta \in \Theta$. Uniform convergence has been shown.

(iv) Limit distribution of the estimator. An expansion of the first-order condition to the log-likelihood optimization problem around θ_0 together with the results obtained above yields

$$\Omega_n^{-1/2}(\theta_n - \theta_0) = -\Omega_n^{-1/2} H_n(\theta_*)^{-1} S_n(\theta_0) \xrightarrow{d} N(0, I)$$

as $n \to \infty$ by an application of Slutsky's theorem. Here, $\theta_* \in \Theta$ is a value that lies between θ_n and θ_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 $\Upsilon = O(n(n-1)p_n)$. This rate

result follows from the same argument as the convergence rate of $(m_n p_n)^{-1} L_n(\theta)$ to its expectation in the proof of Theorem 1 given above and can readily be deduced from the expression for A given above. The proof of Theorem 2 is thus complete. □

References

Acemoglu, D., A. Ozdaglar, and A. Tahbaz-Salehi (2015). Systemic risk and stability in financial networks. *American Economic Review 105*, 564–608.

Allen, F. and D. Gale (2000). Financial contagion. *Journal of Political Economy 108*, 1–33.

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

Anderson, J. E. and E. van Wincoop (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review 93*, 170–192.

Attanasio, O., A. Barr, J. C. Cardenas, G. Genicot, and C. Meghir (2012). Risk pooling, risk preferences, and social networks. *American Economic Journal: Applied Economics 4*, 134–167.

Banerjee, A., A. Chandrasekhar, E. Duflo, and M. O. Jackson (2013). The diffusion of microfinance. *Science 341*(6144).

Bloch, F. and M. O. Jackson (2007). The formation of networks with transfers among players. *Journal of Economic Theory 113*, 83–110.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika 39*, 324–325.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies 47*, 225–238.

Chaney, T. (2014). The network structure of international trade. *American Economic Review 104*, 3600–3634.

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

Chen, M., I. Fernández-Val, and M. Weidner (2014). Nonlinear panel models with interactive effects. Mimeo.

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

Currarini, S., M. Jackson, and P. Pin (2009). An economic model of friendship" Homophily, minorities and segregation. *Econometrica 77*, 1003–1045.

De Weerdt, J. (2004). Risk-sharing and endogenous network formation. In S. Dercon (Ed.), *Insurance Against Poverty*, pp. 197–216. Oxford University Press.

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

Erdős, P. and A. Rényi (1959). On random graphs. *Publicationes Mathematicae 6*, 290–297.

Erdős, P. and A. Rényi (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5*, 17–61.

Fafchamps, M. and F. Gubert (2007). Risk sharing and network formation. *American Economic Review 97*, 75–79.

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

Goldenberg, A., A. X. Zheng, S. E. Fienberg, and E. M. Airoldi (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning 2*, 129–233.

Goldsmith-Pinkham, P. and G. W. Imbens (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics 31*, 253—264.

Golub, B. and M. O. Jackson (2012). How homophily affects the speed of learning and best-response dynamics. *Quarterly Journal of Economics 127*, 1287–1338.

Graham, B. S. (2013). Comment. *Journal of Business & Economic Statistics 31*, 266–270.

Graham, B. S. (2015). An econometric model of link formation with degree heterogeneity. CeMMAP Working Paper 43/15.

Graham, B. S. (2016). Homophily and transitivity in dynamic network formation. CeMMAP Working Paper 16/16.

Head, K. and T. Mayer (2014). Gravity equations: Workhorse, toolkit, and cookbook. In G. Gopinath, E. Helpman, and K. Rogoff (Eds.), *Handbook of International Economics 4*, Chapter 3, pp. 131–195. Elsevier.

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.

Holland, P. W. and S. Leinhardt (1978). An omnibus test for social structure using triads. *Sociological Methodology 7*, 227–256.

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. and B. Rogers (2007). Meetings strangers and friends of friends: How random are social networks? *American Economic Review 97*, 890–915.

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

Jackson, M. O. and D. López-Pintado (2013). Diffusion and contagion in networks with heterogeneous agents and homophily. *Network Science 1*, 49–67.

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.

Jochmans, K. (2016). Two-way models for gravity. Forthcoming in *Review of Economics and Statistics*.

Jochmans, K. and M. Weidner (2016). Fixed-effect regressions on network data. CeMMAP Working Paper 32/16.

Lazega, E. (2001). *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*. Oxford University Press.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27*, 415–444.

Morales, E., G. Sheu, and A. Zahler (2015). Extended gravity. Mimeo.

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.

de Paula, A. (2016). Econometrics of network models. CeMMAP Working Paper 06/16.

de Paula, A., S. Richards-Shubik, and E. Tamer (2015). Identification of preferences in network formation games. CeMMAP Working Paper 29/15.

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.

Robins, G., P. Pattison, Y. Kalish, and D. Lusher (2007). An introduction to exponential random graph (p^*) models for social networks. *Social Networks 29*, 173–191.

Robins, G., T. Snijders, P. Wang, M. Handcock, and P. Pattison (2009). Recent developments in exponential random graph (p^*) models for social networks. *Social Networks 29*, 192–215.

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

Sheng, S. (2012). Identification and estimation of network formation games. Mimeo.

Snijders, T. A. B. (2011). Statistical models for social networks. *Annual Review of Sociology 37*, 129–151.

Yan, T., B. Jiang, S. E. Fienberg, and C. Leng (2016). Statistical inference in a directed network model with covariates. Mimeo.

Zermelo, E. (1929). Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift 29*, 436–460.