NBER WORKING PAPER SERIES

SPARSE NETWORK ASYMPTOTICS FOR LOGISTIC REGRESSION

Bryan S. Graham

Working Paper 27962
http://www.nber.org/papers/w27962

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2020

Financial support from NSF Grant SES #1851647 is gratefully acknowledged. Some of the results contained in this paper were presented, albeit in more basic and preliminary forms, at an invited session of the 2018 Latin American Meetings of the Econometric Society, and at a plenary lecture of the 2019 meetings of the International Association of of Applied Econometrics. I am thankful to Michael Jansson for several very helpful conversations and to Konrad Menzel for feedback on the initial draft. All the usual disclaimers apply. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Sparse Network Asymptotics for Logistic Regression
Bryan S. Graham
NBER Working Paper No. 27962
October 2020
JEL No. C01,C31,C33,C55

ABSTRACT

Consider a bipartite network where N consumers choose to buy or not to buy M different products. This paper considers the properties of the logistic regression of the $N \times M$ array of "i-buys-j" purchase decisions, $[Y_{ij}]_{1 \leq i \leq N, \leq j \leq M}$, onto known functions of consumer and product attributes under asymptotic sequences where (i) both N and M grow large and (ii) the average number of products purchased per consumer is finite in the limit. This latter assumption implies that the network of purchases is sparse: only a (very) small fraction of all possible purchases are actually made (concordant with many real-world settings). Under sparse network asymptotics, the first and last terms in an extended Hoeffding-type variance decomposition of the score of the logit composite log-likelihood are of equal order. In contrast, under dense network asymptotics, the last term is asymptotically negligible. Asymptotic normality of the logistic regression coefficients is shown using a martingale central limit theorem (CLT) for triangular arrays. Unlike in the dense case, the normality result derived here also holds under degeneracy of the network graphon. Relatedly, when there "happens to be" no dyadic dependence in the dataset in hand, it specializes to recently derived results on the behavior of logistic regression with rare events and iid data. Sparse network asymptotics may lead to better inference in practice since they suggest variance estimators which (i) incorporate additional sources of sampling variation and (ii) are valid under varying degrees of dyadic dependence.

Bryan S. Graham
University of California - Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
bgraham@econ.berkeley.edu

Let $i = 1, \ldots, N$ index a random sample of consumers and $j = 1, \ldots, M$ a random sample of products. For each consumer-product pair ij we observe $Y_{ij} = 1$ if consumer i purchases product j and $Y_{ij} = 0$ otherwise. Let W_i be a vector of observed consumer attributes, X_j a vector of product attributes and $n = M + N$ the total number of sampled consumers and products. The conditional probability that consumer i buys product j is given by

$$\Pr\left(Y_{ij} = 1 | W_i, X_j\right) = e\left(\alpha_{0,n} + Z_{ij}'\beta_0\right), \tag{1}$$

where $Z_{ij} \overset{def}{\equiv} z\left(W_i, X_j\right)$ is a vector of known functions of W_i and X_j, $\alpha_{0,n}$ an "intercept" parameter (which may vary with n), β_0 a vector of fixed "slope" parameters, and $e\left(\cdot\right)$ a known increasing function mapping the real line into the unit interval. Below I will emphasize the logit case with $e\left(v\right) = \exp\left(v\right) / \left[1 + \exp\left(v\right)\right]$; this case is convenient and dominates empirical work, but nothing which follows hinges essentially upon it.

I am interested in settings where both the number of consumers, N, and the number of products, M, are very large. To motivate this focus, consider a large book retailer. Such a retailer may service many customers and also stock many books. Let x be the attribute vector associated with a newly released book, $\hat{\theta}_n = \left(\hat{\alpha}_n, \hat{\beta}'\right)'$ estimates of the parameters in (1), constructed from some training sample, and $\hat{e}_i\left(x\right) = e\left(\hat{\alpha}_n + z\left(W_i, x\right)'\hat{\beta}\right)$ the predicted probability that agent i purchases a book of type $X_j = x$. With this knowledge the retailer might use

$$\hat{\gamma}_n\left(x\right) = \sum_{i=1}^{N} \hat{e}_i\left(x\right) \tag{2}$$

to predict total unit sales for the new book. This prediction could be useful for making wholesale purchase decisions. Other objects of interest include various average partial effects (e.g., Chamberlain, 1984; Wooldridge, 2005).

In this paper I present a method of estimating the coefficient vector $\theta_{0,n} = \left(\alpha_{0,n}, \beta_0'\right)'$ as well as one for conducting inference on it. I also explore estimation and inference for aggregate effects, like (2), as well as for average effects. The econometric framework outlined below is designed to accommodate two peculiarities of the setting described above that appear to be important in practice and also consequential

for inference.

First, consider predicting whether randomly sampled consumer i purchases book j, say *The Clue in the Crossword Cipher*, the forty-fourth novel in the celebrated Nancy Drew mystery series. Knowledge of the frequency with which other consumers $k = 1, \ldots, i-1, i+1, \ldots, N$ purchase book j will generally alter the econometrician's prediction of whether i also purchases book j. That is, for any $k \neq i$,

$$\Pr\left(Y_{ij} = 1 \middle| Y_{kj} = 1\right) > \Pr\left(Y_{ij} = 1\right)$$

or that $Y_{i_1 j_1}$and $Y_{i_2 j_2}$ will covary whenever the two transactions correspond to a common book (such that $j_1 = j_2$).

Similarly, if the econometrician knew that consumer i was a frequent book buyer, she might conclude that this consumer is also more likely to purchase some other book (relative to the average consumer). That is $Y_{i_1 j_1}$and $Y_{i_2 j_2}$ will also covary whenever the transactions correspond to a common buyer (such that $i_1 = i_2$).

Importantly, dependence across $Y_{i_1 j_1}$ and $Y_{i_2 j_2}$ when $\{i_1, j_j\}$ and $\{i_2, j_2\}$ share a common buyer or book index may hold even conditional on observed consumer, W_i, and product attributes, X_j. Some consumers may have latent attributes (i.e., not contained in W_i) which induce them to buy many books and some books may be especially popular (for reasons not captured adequately by X_j). It might be, for example, that

$$\Pr\left(Y_{ij} = 1 \middle| Y_{kj} = 1, W_i, W_k, X_j\right) > \Pr\left(Y_{ij} = 1 \middle| W_i, X_j\right).$$

The structured form of dependence across the elements of $[Y_{ij}]_{1 \leq i \leq N, 1 \leq j \leq M}$ described above is a feature of separately exchangeable random arrays (Aldous, 1981; Hoover, 1979). The inferential implications of such dependence, in the context of subgraph counts, were first considered by Holland and Leinhardt (1976) almost fifty years ago. Bickel et al. (2011) make an especially important recent contribution in this area. In the context of regression models, the inferential implications of dyadic dependence have been considered by, among others, Fafchamps and Gubert (2007), Cameron and Miller (2014), Aronow et al. (2017), Graham (2020a), Davezies et al. (2020) and Menzel (2017) (see Graham (2020b, Section 4) for a review and references). Dyadic dependence will generate distinct issues here.

The second peculiarity explored here is suggested by the observation that, even

when presented with the opportunity to purchase many books, the typical customer will only purchase a few. Similarly, a retailer only sells a few copies of most titles in a given year. Put differently personal libraries are generally small and the market share of most books is, for all practical purposes, infinitesimally small. These observations have implications for what types of asymptotic approximations are likely to be useful in practice. In this paper I consider sequences where both N and M grow at the same rate such that, recalling that $n = M + N$,

$$M/n \to \phi \in (0, 1)$$

as $n \to \infty$.

Let $\rho_{0,n} \overset{def}{\equiv} \mathbb{E}\left[e\left(\alpha_{0,n} + Z_{ij}'\beta_0\right)\right]$ be probability that a randomly sampled consumer purchases a randomly sampled book. The average number of books purchased by the average consumer is then

$$\lambda_{0,n}^c \overset{def}{\equiv} M\rho_{0,n}. \tag{3}$$

In network parlance $\lambda_{0,n}^c$ corresponds to average consumer degree. If $\rho_{0,n}$ is bounded away from zero, then $\lambda_{0,n}^c \to \infty$ as $N, M \to \infty$. This implies that the number of actual book purchases and the number of possible book purchases should be of equal order. This not true in practice.[1] To develop a distribution theory which is concordant with the empirical regularity that consumers only purchase a small number of books (and, similarly, that retailers only sell a small number of copies of any given title) I let $\alpha_{0,n} \to -\infty$ at a rate which ensures that $\lambda_{0,n}^c$ converges to a non-zero and bounded constant $0 < \lambda_0^c < \infty$ as $N, M \to \infty$. In language of networks I consider bi-partite graphs which are *sparse*.

The asymptotic analysis in this paper is, to my knowledge, novel, but it does connect with two important areas of prior research by others. The first is the literature on subgraph counts and dyadic regression cited above. However, with the partial exception of Bickel et al.'s (2011) analysis of acyclic subgraph counts, this work has been, starting with Holland and Leinhardt (1976), limited to to dense networks.[2] The second connection is to the literature on "rare events" analysis (e.g., King and

[1] There are tens of millions of print titles available on, for example, Amazon, even consumers who buys hundreds of books in a year are completing only very small fraction of all possible purchases.

[2] Graham (2017) and Jochmans (2018) also considered regression in the context of graphs which are sparse in the limit. Both these papers utilize conditional likelihood type ideas; this has the effect of "conditioning away" some of the dependence which is central to the analysis below.

Zeng, 2001); an area of special concern in political science and epidemology, but also increasingly relevant in economics (especially in the era of "Big Data"). An interesting feature of Theorem 1 below is that it contains Wang's (2020) recent result for logistic regression with rare events and iid data as a special case.

The formal analysis of this paper is confined to bipartite networks, but adapting it to directed and/or undirected networks would be straightforward. Several consumer demand settings might be appropriately modeled with the methods described in this paper. Examples include (i) the listening behavior of streaming music service customers and (ii) the purchase behavior of big box store customers. A limitation vis-a-vis these applications, is that the basic set-up explored here is not useful for understanding complementary and substitution patterns across products (cf., Lewbel and Nesheim, 2019). Other possible applications include modeling plant locations in an industry where firms typically operate multiple plants (here $i = 1, \ldots, N$ would index firms and $j = 1, \ldots, M$ locations; see KPMG (2016)). Other many-to-many matching problems that have drawn economists' interest include (i) bank-firm lending relationships (e.g., Marotta et al., 2015), (ii) the matching of venture capital with start-ups (e.g., Bengtsson and Hsu, 2015), and (iii) supply chain settings with strong bi-partite structure (e.g., automakers and parts supplies as in Fox (2018)). When $i = 1, \ldots, N$ and $j = 1, \ldots, M$ index the same units with $M = N$, applications include the modeling of "rare events" in international relations data, such as interstate wars (e.g., King and Zeng, 2001). More generally, the methods developed in this paper, with minimal adaptation, can be used for link prediction in any sparse network setting: bi-partite, directed or undirected.[3] There are numerous applications of link prediction in the other social sciences, the bench sciences as well as in industry.

In what follows random variables are denoted by capital Roman letters, specific realizations by lower case Roman letters and their support by blackboard bold Roman letters. That is Y, y and $\mathbb{Y}$ respectively denote a generic random draw of, a specific value of, and the support of, Y. A "0" subscript on a parameter denotes its population value and may be omitted when doing so causes no confusion. In what follows I use graph, network and purchase graph to refer to $\mathbf{Y} \overset{def}{\equiv} [Y_{ij}]_{1 \leq i \leq N, 1 \leq j \leq M}$. All graph theory terms and notation used below are standard (e.g., Chartrand and Zhang, 2012).

[3]The methods outlined here are not appropriate for use in one-to-one matching settings.

1 Population and sampling assumptions

Let $i \in \mathbb{N}$ index *consumers* in an infinite population of interest. Associated with each consumer is the vector of observed attributes $W_i \in \mathbb{W} = \{w_1, \ldots, w_J\}$. Let $j \in \mathbb{M}$ index *products* in a second infinite population of interest. The model is a two population one (cf., Graham et al., 2018). Associated with each product is the vector of characteristics $X_i \in \mathbb{X} = \{x_1, \ldots, x_K\}$. The finite support assumption on $\mathbb{W}$ and $\mathbb{X}$ is not essential, but simplifies the discussion of exchangeability below.

Let $\sigma_w : \mathbb{N} \to \mathbb{N}$ be a permutation of a finite number of consumer indices which satisfies the restriction

$$\left[W_{\sigma_w(i)}\right]_{i \in \mathbb{N}} = [W_i]_{i \in \mathbb{N}} . \tag{4}$$

Restriction (4) implies that σ_w only permutes indices across observationally identical consumers (i.e., with the same values of W). Let $\sigma_x : \mathbb{M} \to \mathbb{M}$ be an analogously constrained permutation of a finite number of product indices. Adapting the terminology of Crane and Towsner (2018), I assume that the purchase graph is *W-X-exchangeable*

$$\left[Y_{\sigma_w(i)\sigma_x(j)}\right]_{i \in \mathbb{N}, j \in \mathbb{M}} \overset{D}{=} [Y_{ij}]_{i \in \mathbb{N}, j \in \mathbb{M}} . \tag{5}$$

Here $\overset{D}{=}$ denotes equality of distribution. One way to think about (5) is as a requirement that any probability law for $[Y_{ij}]_{i \in \mathbb{N}, j \in \mathbb{M}}$ should attach equal probability to all purchase graphs which are isomorphic as vertex-colored graphs. Here I associate W_i and X_j with the color of the corresponding consumer and product vertices in the overall purchase graph. Virtually all single-population micro-econometric models assume that agents are exchangeable, restriction (5) extends this idea to the two-population setting considered here. Our probability law for the model should not change if we re-label observationally identical units.

Graphon

It is well-known that exchangeability implies restrictions on the structure of dependence across observations in the cross-section setting (e.g., de Finetti, 1931). Aldous (1981), Hoover (1979) and Crane and Towsner (2018) showed that exchangeable random *arrays* also exhibit a special dependence structure. Let μ, $\{(W_i, A_i)\}_{i \geq 1}$, $\{(X_j, B_j)\}_{j \geq 1}$ and $\{V_{ij}\}_{i \geq 1, j \geq 1}$ be sequences of i.i.d. random variables, additionally

independent of one another, and consider the purchase graph $\left[Y_{ij}^*\right]_{i\in\mathbb{N},j\in\mathbb{M}}$, generated according to

$$Y_{ij}^* = h\left(\mu, W_i, X_j, A_i, B_j, V_{ij}\right) \tag{6}$$

with $h : [0,1] \times \mathbb{W} \times \mathbb{X} \times [0,1]^2 \to \{0,1\}$ a measurable function, henceforth referred to as a *graphon* (we can normalize μ, A_i, B_j and V_{ij} to have support on the unit interval, uniformly distributed, without loss of generality).

The results of Crane and Towsner (2018), which extend the earlier work of Aldous (1981) and Hoover (1979), show that, for any W-X-*exchangeable* random array $[Y_{ij}]_{i\in\mathbb{N},j\in\mathbb{M}}$, there exists another array $\left[Y_{ij}^*\right]_{i\in\mathbb{N},j\in\mathbb{M}}$, generated according to (6), such that the two arrays have the same distribution. An implication of this result is that we may use (6) as a nonparametric data generating process for $[Y_{ij}]_{i\in\mathbb{N},j\in\mathbb{M}}$.

Inspection of (6) indicates that exchangeability implies a particular pattern of dependence across the elements of $[Y_{ij}]_{i\in\mathbb{N},j\in\mathbb{M}}$. In particular $Y_{i_1j_1}$ and $Y_{i_2j_2}$ may covary whenever $i_1 = i_2$ or $j_1 = j_2$; this covariance may be present even conditional on consumer and product attributes. This is, of course, precisely the dependence structure discussed in the introduction.

Sampling process

Let $\mathbf{Y} = [Y_{ij}]_{1\leq i\leq N, 1\leq j\leq M}$ be the observed $N \times M$ matrix of consumer purchase decisions. Let $\mathbf{W}$ and $\mathbf{X}$ be the associated matrices of consumer and product regressors. I assume that $\mathbf{Y}$ is the adjacency matrix associated with the subgraph induced by a random sample of consumers and products from a W-X-*exchangeable* infinite population graph. Let $G_{\infty,\infty}$ denote this population network. Associated with this network is some graphon (6). Let $\mathcal{V}_c$ and $\mathcal{V}_p$ denote the set of consumers and products randomly sampled by the econometrician from $G_{\infty,\infty}$. We have $\mathbf{Y}$ equal to the adjacency matrix of the network:

$$G_{N,M} = G_{\infty,\infty}\left[\mathcal{V}_c, \mathcal{V}_p\right]. \tag{7}$$

An implication of (5), (6) and (7) is that we may proceed 'as if' the adjacency matrix in hand was generated according to

$$Y_{ij} = h\left(\mu, W_i, X_j, A_i, B_j, V_{ij}\right)$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, M$. The marginal probability of the event, random consumer i, purchases random product j, is thus

$$\rho_0 = \mathbb{E}\left[h\left(\mu, W_i, X_j, A_i, B_j, V_{ij}\right)\right]. \tag{8}$$

Let $\{G_{N,M}\}$ be a sequence of networks indexed by, respectively, the cardinality of the sampled consumer and product index sets, $N = |\mathcal{V}_c|$ and $M = |\mathcal{V}_p|$. The average number of products purchased per consumer, or *average consumer degree*,

$$\lambda_0^c = M\rho_0 \tag{9}$$

will diverge as $M \to \infty$ when $\rho_0 > 0$. Likewise the average number of times a given product is purchased, or *average product degree*,

$$\lambda_0^p = N\rho_0 \tag{10}$$

will also diverge as $N \to \infty$. A consequence of this divergence is that the number of possible purchases, and the number of actual purchases, will be of equal order. In practice, however, only a small fraction of all possible purchases are made. To capture this feature of the real world in our asymptotic approximations requires a slightly more elaborate thought experiment; which I outline next.

Instead of considering a sequence of graphs sampled from a *fixed* population, I consider a sequence of graphs sampled from a corresponding *sequence* of populations. The sequence of networks $\{G_{N,M}\}$ is one where both N and M grow at the same rate such that, recalling that $n = M + N$,

$$M/n \to \phi \in (0, 1)$$

as $n \to \infty$. For each N, M the graphon describing the infinite population sampled from is

$$Y_{ij} = h_{N,M}\left(\mu, W_i, X_j, A_i, B_j, V_{ij}\right). \tag{11}$$

This sequence of graphons/populations $\{h_{N,M}\}$ has the property that network *density*

$$\rho_{0,N,M} = \mathbb{E}_{N,M}\left[h_{N,M}\left(\mu, W_i, X_j, A_i, B_j, V_{ij}\right)\right]$$

may approach zero as $n \to \infty$. Under this setup the order of $\lambda^c_{0,N,M} = M\rho_{0,N,M}$ and $\lambda^p_{0,N,M} = N\rho_{0,N,M}$ will depend upon the speed with which $\rho_{0,N,M}$ approaches zero as $n \to \infty$. Here I use the notation $\mathbb{E}_{N,M}[\cdot]$ to emphasize that the probability law used to compute expectations may vary with the sample size.

As in other exercises in alternative asymptotics, indexing the population data generating process by the sample size is not meant to capture a literal feature of how the data are generated, rather it is done so that the limiting properties of the model share important features – in this case sparseness – with the actual finite network in hand. In other settings such an approach has led to more useful asymptotic approximations, a premise I maintain here (e.g., Staiger and Stock, 1997).

2 Composite likelihood estimator

The estimation target is the regression function of Y_{ij} given X_i and W_j. This is a predictive function and may, or may not, have structural economic meaning as well (see Graham (2020b, Sections 4-5)). I assume that this regression function takes the parametric form

$$e\left(\alpha_{0,n} + Z'_{ij}\beta_0\right) = \frac{\exp\left(\alpha_{0,n} + Z'_{ij}\beta_0\right)}{1 + \exp\left(\alpha_{0,n} + Z'_{ij}\beta_0\right)} \tag{12}$$

where $Z_{ij} \overset{def}{\equiv} z\left(W_i, X_j\right)$ is a finite vector of known functions of W_i and X_j. It would be interesting to extend what follows to semiparametric regression models, but this is not done here.

Assumption 1 formalizes the population and sampling set-up of the previous section.

Assumption 1. *(Sampling) The sampled network is the one induced by a random sample of N consumers and M products drawn from the nodes of the infinite W-X-*exchangeable bipartite random array $[Y_{ij}]_{i\in\mathbb{N},j\in\mathbb{M}}$ *with graphon* (11)*; N and M grow such that, for $n = M + N$,*

$$M/n \to \phi \in (0,1)$$

as $n \to \infty$.

To allow the probability of making a purchase decline with n, let $\alpha_{0,n} = \ln\left(\alpha_0/n\right)$.

This gives, after some manipulation,

$$e\left(\alpha_{0,n} + Z_{ij}'\beta_0\right) = \frac{\frac{\alpha_0}{n}\exp\left(Z_{ij}'\beta_0\right)}{1 + \frac{\alpha_0}{n}\exp\left(Z_{ij}'\beta_0\right)} \tag{13}$$

and hence an expression for average consumer degree (3) of, recalling that $M/n \approx \phi$,

$$\lambda_{0,n}^c = \alpha_0\phi\mathbb{E}\left[\exp\left(Z_{ij}'\beta_0\right)\right] + O\left(n^{-1}\right)$$

which converges to a bounded constant as $n \to \infty$ as long as $\mathbb{E}\left[\exp\left(Z_{ij}'\beta_0\right)\right] < \infty$. By allowing $\alpha_{0,n} \to -\infty$ as $n \to \infty$ we ensure that, in the limit, the bipartite graph $\mathbf{Y} = [Y_{ij}]$ is sparse. Similar devices are used by Owen (2007) and Wang (2020) to model "rare events" in cross-sectional binary outcome data.

Assumption 2. *(Logit Regression Function) The mean regression function (CEF)* $\mathbb{E}_{N,M}\left[Y_{ij}|\, W_i, X_j\right]$ *belongs to the parametric family* (12) *with* $\alpha_n = \ln\left(\alpha/n\right)$, $\theta = \left(\alpha, \beta'\right)' \in \mathbb{A} \times \mathbb{B} = \Theta$, $\mathbb{A}$ *and* $\mathbb{B}$ *compact, and* $Z_{ij} \in \mathbb{Z}$ *with* $\mathbb{Z}$ *a compact subset of* $\mathbb{R}^{\dim(\mathbb{Z}_{ij})}$. *The true parameter* $\theta_0 = \left(\alpha_0, \beta_0'\right)'$ *lies in the interior of the parameter space.*

The compact support assumption on Z_{ij} is not essential, but simplifies the proofs. Here I focus on estimation of $\theta_n = \left(\alpha_n, \beta'\right)'$ with $\alpha_n = \ln\left(\alpha/N\right)$. Observe that $\theta = \left(\alpha, \beta'\right)'$ does not vary with n, while θ_n does. Define $\theta_{0,n} = \left(\alpha_{0,n}, \beta_0'\right)'$ with $\alpha_{0,n} = \ln\left(\alpha_0/n\right)$. Let $\bar{\alpha} = \sup\mathbb{A}$, the parameter space for θ_n, $\Theta_n = \left(-\infty, \ln\left(\bar{\alpha}\right)\right] \times \mathbb{B}$, is the one induced by $\Theta = \mathbb{A} \times \mathbb{B}$ and the mapping from θ to θ_n.

To estimate $\theta_{0,n}$we maximize the composite log-likelihood function

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta_n} L_n\left(\theta\right)$$

with $L_n\left(\theta\right) \overset{def}{\equiv} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} l_{ij}\left(\theta\right)$ and

$$l_{ij}\left(\theta\right) = \left(2Y_{ij} - 1\right)R_{ij}'\theta - \ln\left[1 + \exp\left[\left(2Y_{ij} - 1\right)R_{ij}'\theta\right]\right]$$

the logit kernel function and $R_{ij} = \left(1, Z_{ij}'\right)'$. Observe that $\hat{\theta}_n$ is simply the coefficient vector associated with a logistic regression of Y_{ij} onto a constant and Z_{ij} using all NM dyads in the network. Although, by virtue of Assumption 2 above, $L_n\left(\theta\right)$ correctly represents the marginal (conditional) probability of Y_{ij} for each element of $\mathbf{Y}$, it

does not accurately reflect the dependence structure across these elements; hence the term "composite likelihood". See Lindsey (1988) an introduction to estimation by composite likelihood and Graham (2020b) for discussion in the contexts of network model estimation.

3 Sparse network asymptotics

If $\alpha_{0,n}$ equals a fixed constant, then $\rho_{0,n}$ - network density – will also be fixed such that the network will be dense in the limit. The limit distribution of $\hat{\theta}_n$ under such "dense network asymptotics" was derived by Graham (2020b). More general results for dyadic M-estimators under dense network asymptotics, including results on the bootstrap, can be found in Menzel (2017) and Davezies et al. (2020). None of these results apply here. To derive a result that does apply, begin with the mean value expansion

$$\sqrt{n}\left(\hat{\theta}_n - \theta_{0,n}\right) = \left[nH_n\left(\bar{\theta}_n\right)\right]^{+} \times n^{3/2} S_n\left(\theta_{0,n}\right).$$

where

$$S_n\left(\theta\right) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} s_{ij}\left(\theta\right), \tag{14}$$

with $s_{ij}\left(\theta\right) = \frac{\partial l_{ij}(\theta)}{\partial \theta} = \left(Y_{ij} - e_{ij}\left(\theta\right)\right) R_{ij}$ and $e_{ij}\left(\theta\right) = e\left(\alpha + Z_{ij}'\beta\right)$, corresponds to the score vector of the composite likelihood and

$$H_n\left(\theta\right) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \frac{\partial^2 l_{ij}\left(\theta\right)}{\partial\theta\partial\theta'} \tag{15}$$

the associated Hessian matrix. Here $\bar{\theta}_n$ is a mean value between $\theta_{0,n}$ and $\hat{\theta}_n$ which may vary from row to row.

Lemma 1, stated and proved in Appendix A, shows that, after re-scaling by n, that $nH_n\left(\theta\right)$ converges uniformly to

$$\Gamma\left(\theta\right) = -\alpha\mathbb{E}\left[\exp\left(Z_{12}'\beta\right)\begin{pmatrix} 1 & Z_{12}' \\ Z_{12} & Z_{12}Z_{12}' \end{pmatrix}\right]. \tag{16}$$

An intuition for why $H_n\left(\theta\right)$ needs to be rescaled to ensure convergence is that, under

sparse network asymptotics, information accrues at a slower rate: the effective sample size is not $NM = (n^2)$, but rather $O(n)$. I return to this point briefly at the end of the paper.

Assumption 3. *(IDENTIFICATION) The matrix* $\Gamma_0 \overset{def}{\equiv} \Gamma(\theta_0)$ *is of full rank.*

Assumption 3 is a standard identification condition (see, for example, Amemiya (1985, p. 270)). This assumption, in conjunction with Lemma 1, gives the linear approximation

$$\sqrt{n}\left(\hat{\theta}_n - \theta_n\right) = \Gamma_0^{-1} \times n^{3/2} S_n(\theta_{0,n}) + o_p(1).$$

To derive the limit distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$ I show that the distribution $n^{3/2} S_n(\theta_{0,n})$ is well-approximated by a Gaussian random variable. The main tool used is a martingale CLT for triangular arrays. That the variance stabilizing rate for $S_n(\theta_{0,n})$ is $n^{3/2}$, like the need to rescale the Hessian, is non-standard. The need to "blow up" $S_n(\theta_{0,n})$ at a faster than $\sqrt{n}$ rate is a consequence of the fact that the summands in $S_n(\theta_{0,n})$ are $O(n^{-1})$ since $\alpha_{0,n} \to -\infty$ as $n \to \infty$.

A detailed proof of Theorem 1, stated below, is provided in Appendix B. Here I outline the main arguments. Begin with the following three part decomposition of the score vector

$$S_n(\theta) = U_{1n}(\theta) + U_{2n}(\theta) + V_n(\theta) \tag{17}$$

where

$$U_{1n}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \bar{s}_{1i}^{c}(\theta) + \frac{1}{M}\sum_{j=1}^{M} \bar{s}_{1j}^{p}(\theta) \tag{18}$$

$$U_{2n}(\theta) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \left\{\bar{s}_{ij}(\theta) - \bar{s}_{1i}^{c}(\theta) - \bar{s}_{1j}^{p}(\theta)\right\} \tag{19}$$

$$V_n(\theta) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \left\{s_{ij}(\theta) - \bar{s}_{ij}(\theta)\right\} \tag{20}$$

with $\bar{s}_{ij}(\theta) = \bar{s}(W_i, X_j, A_i, B_j; \theta)$ with $\bar{s}(w, x, a, b; \theta) =$

$\mathbb{E}\left[s_{ij}(\theta)\middle| W_i = w, X_j = x, A_i = a, B_j = b\right]$ and

$$\begin{aligned}\bar{s}^c_{1i}(\theta) =& \bar{s}^c_1(W_i, A_i; \theta)\\ \bar{s}^p_{1j}(\theta) =& \bar{s}^p_1(X_j, B_j; \theta)\end{aligned}$$

with $\bar{s}^c_1(w, a; \theta) = \mathbb{E}\left[\bar{s}(w, X_j, a, B_j; \theta)\right]$ and $\bar{s}^p_1(x, b; \theta) = \mathbb{E}\left[\bar{s}(W_i, x, A_i, b; \theta)\right]$.

Decomposition (17) also features in Graham (2020a) and Menzel (2017).[4] It can be derived by first projecting $S_n(\theta)$ on to $\mathbf{A} = [A_i]_{1\leq i\leq N}$, $\mathbf{W} = [W_i]_{1\leq i\leq N}$, $\mathbf{B} = [B_j]_{1\leq j\leq M}$, and $\mathbf{X} = [X_i]_{1\leq j\leq N}$ as follows:

$$\begin{aligned}S_n(\theta) =& \mathbb{E}\left[S_n(\theta)\middle|\mathbf{W}, \mathbf{X}, \mathbf{A}, \mathbf{B}\right] + \left\{S_n(\theta) - \mathbb{E}\left[S_n(\theta)\middle|\mathbf{W}, \mathbf{X}, \mathbf{A}, \mathbf{B}\right]\right\}\\ =& \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\bar{s}_{ij}(\theta) + \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\{s_{ij}(\theta) - \bar{s}_{ij}(\theta)\right\}. \qquad (21)\end{aligned}$$

Next observe that $\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\bar{s}_{ij}(\theta)$ is a two sample U-Statistic, albeit one defined partially in terms of the latent variables A_i and B_j. Equation (18) corresponds to the Hajek Projection of this U-statistic onto (separately) $\{(W_i', A_i)\}_{i=1}^{N}$ and $\left\{\left(X_j', B_j\right)\right\}_{j=1}^{M}$. Equation (19) is the usual Hajek Projection error term.

Define $\phi_n = M/n$, $\bar{s}^c_{1ni} \overset{def}{\equiv} \bar{s}^c_{1i}(\theta_{0,n})$, $\bar{s}^p_{1nj} \overset{def}{\equiv} \bar{s}^p_{1j}(\theta_{0,n})$ and also $\bar{s}_{nij} \overset{def}{\equiv} \bar{s}_{ij}(\theta_{0,n})$. Similarly let $S_n = S_n(\theta_{0,n})$ and so on. Applying the variance operator to S_n yields:

$$\begin{aligned}\mathbb{V}(S_n) =& \mathbb{V}(U_{1n}) + \mathbb{V}(U_{2n}) + \mathbb{V}(V_n) \qquad (22)\\ =& \frac{\Sigma^c_{1n}}{N} + \frac{\Sigma^p_{1n}}{M} + \frac{1}{NM}\left[\Sigma_{2n} - \Sigma^c_{1n} - \Sigma^p_{1n}\right] + \frac{\Sigma_{3n}}{NM}\end{aligned}$$

where

$$\begin{aligned}&\Sigma^c_{1n} = \mathbb{E}\left[\bar{s}^c_{1ni}\left(\bar{s}^c_{1ni}\right)'\right] \quad \Sigma^p_{1n} = \mathbb{E}\left[\bar{s}^p_{1nj}\left(\bar{s}^p_{1nj}\right)'\right] \qquad (23)\\ &\Sigma_{2n} = \mathbb{E}\left[\bar{s}_{nij}\bar{s}_{nij}'\right] = \mathbb{V}\left(\mathbb{E}\left[s_{nij}\middle| W_i, X_j, A_i, B_j\right]\right)\\ &\Sigma_{3n} = \mathbb{E}\left[\left\{s_{nij} - \bar{s}_{nij}\right\}\left\{s_{nij} - \bar{s}_{nij}\right\}'\right] = \mathbb{E}\left[\mathbb{V}\left(s_{nij}\middle| W_i, X_j, A_i, B_j\right)\right].\end{aligned}$$

In the *dense* case Σ^c_{1n}, Σ^p_{1n}, Σ_{2n} and Σ_{3n} are all constant in n; hence the asymptotic properties of S_n coincide with those of U_{1n}. Since U_{1n} is a sum of independent random

[4] It is also implicit in the elegant proof in Bickel et al. (2011).

variables a standard argument gives

$$n^{1/2}S_n \xrightarrow{D} \mathcal{N}\left(0, \frac{\Sigma_1^c}{1-\phi} + \frac{\Sigma_1^p}{\phi}\right) \tag{24}$$

as long as Σ_1^c and/or Σ_1^p are non-zero.

Under the sparse network asymptotics considered here the order of Σ_{1n}^c, Σ_{1n}^p, Σ_{2n} and Σ_{3n} varies with n. This affects the order of the four variance terms in (22) and, consequently, which components of S_n contribute to its asymptotic properties. In Appendix A I show the order of the four terms in (22) are, respectively,

$$\begin{aligned}
\mathbb{V}(S_n) =& O\left(\frac{\rho_n^2}{N}\right) + O\left(\frac{\rho_n^2}{M}\right) + O\left(\frac{\rho_n^2}{MN}\right) + O\left(\frac{\rho_n}{MN}\right) \\
=& O\left(\left[\frac{\lambda_{0,n}^c}{\phi_n}\right]^2 \frac{1}{(1-\phi_n)}\frac{1}{n^3}\right) + O\left(\left[\frac{\lambda_{0,n}^c}{\phi_n}\right]^3 \frac{1}{n^3}\right) \\
&+ O\left(\left[\frac{\lambda_{0,n}^c}{\phi_n}\right]^2 \frac{1}{\phi_n(1-\phi_n)}\frac{1}{n^4}\right) + O\left(\frac{\lambda_{0,n}^c}{\phi_n^2(1-\phi_n)}\frac{1}{n^3}\right).
\end{aligned}$$

Since Σ_1^c and Σ_1^p are both $O(\rho_n^2) = O(n^{-2})$ we can multiply them by n^2 to stabilize them. Define $\tilde{\Sigma}_1^c$ to be the limit of $n^2\Sigma_{1n}$ and $\tilde{\Sigma}_1^p$ to be the limit of $n^2\Sigma_{1n}^p$. Similarly we can define $\tilde{\Sigma}_3$ to be the limit of $n\Sigma_{3n}$, all as $n \to \infty$. Normalizing (22) by $n^{3/2}$ therefore gives

$$\mathbb{V}\left(n^{3/2}S_n\right) = \frac{\tilde{\Sigma}_1^c}{1-\phi} + \frac{\tilde{\Sigma}_1^p}{\phi} + \frac{\tilde{\Sigma}_3}{\phi(1-\phi)} + O\left(n^{-1}\right) \tag{25}$$

where I also use the fact that $\Sigma_{2n} = O(n^{-2})$.

Under sparse network asymptotics both U_{1n} and V_n matter. In Appendix B I show that $U_{1n} + V_n$ is a martingale difference sequence (MDS) to which a martingale CLT can be applied; Theorem 1 then follows.

Theorem 1. *Under Assumptions 1, 2 and 3*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_n\right) \xrightarrow{D} \mathcal{N}\left(0, \Gamma_0^{-1}\left[\frac{\tilde{\Sigma}_1^c}{1-\phi} + \frac{\tilde{\Sigma}_1^p}{\phi} + \frac{\tilde{\Sigma}_3}{\phi(1-\phi)}\right]\Gamma_0^{-1}\right)$$

as $n \to \infty$.

Theorem 1 indicates that under sparse network asymptotics there are additional

sources of sampling variation in $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$ relative to those which appear in the dense case. Not incorporating these into inference procedures will lead to tests with incorrect size and/or confidence intervals with incorrect coverage. A further advantage of considering sparse network asymptotics is that Theorem 1 remains valid even under degeneracy of the graphon, $h_{N,M}(\mu, W_i, X_j, A_i, B_j, V_{ij})$. For example, if the graphon is constant in A_i and B_j such that $Y_{ij}$and Y_{ik} do not covary conditional on covariates (and likewise for Y_{ji} and Y_{ki}), then $\tilde{\Sigma}_1^c = \tilde{\Sigma}_1^p = 0$, but Theorem 1 nevertheless remains valid. In contrast, under dense network asymptotics, degeneracy – as elegantly shown by Menzel (2017) – generates additional complications. In that case the variance of U_{1n} is identically equal to zero, while that of U_{2n} and V_n are of equal order. In some cases, the behavior of U_{2n} may even induce a non-Gaussian limit distribution (see van der Vaart (2000)). In the sparse network cases, U_{2n} is always negligible relative to V_n. Furthermore V_n is well approximated – after suitable scaling – by a Gaussian distribution.

4 Extensions and discussion

In this section I connect Theorem 1 to prior work on rare events logistic analysis, sketch some results about the estimation of aggregate and average effects, and, finally, close with a few ideas about possible areas of additional research.

Rare events with iid data

King and Zeng (2001) discuss, with a focus on finite sample bias, the behavior of logistic regression under "rare events" with iid data. Evidently binary choice analyses where the marginal frequency of positive events is quite small are common in empirical work.[5] The properties of logistic regression under sequences where the number of "events" becomes small (i.e., "rare") relative to the sample size as it grows were recently characterized by Wang (2020) (see also Owen (2007)). The main result in Wang (2020) coincides with a special case Theorem 1 above. To see this observe that if the graphon is constant in A_i and B_j, then $\bar{s}_{nij}$ will be identically equal to zero for all $1 \leq i \leq N$ and $1 \leq j \leq M$. In this scenario there is no "dyadic dependence" (after conditioning on W_i and X_j) and $\tilde{\Sigma}_1^c = \tilde{\Sigma}_1^p = 0$. Inspection of the calculations

[5]The King and Zeng (2001) has close to four thousand citations on Google Scholar.

in Appendix A also reveals that, in this case, we further have an information-matrix type equality of $\tilde{\Sigma}_3 = \Gamma_0$. Under these conditions Theorem 1 specializes to

$$\sqrt{n}\left(\hat{\theta} - \theta_n\right) \xrightarrow{D} \mathcal{N}\left(0, \Gamma_0^{-1}\right),$$

as $n \to \infty$. This is precisely, up to some small differences in notation, the result given in Theorem 1 of Wang (2020).[6]

In his analysis Wang (2020) emphasizes that information accumulates more slowly under "rare event asymptotics". In the present setting this is reflected in the need to rescale the Hessian matrix by n to achieve convergence (see Lemma 1 in Appendix A). In the network setting dyadic dependence additionally slows down the rate of convergence (cf., Graham et al., 2019). If a researcher is working with a sparse network and concerned about dyadic dependence, then she should base inference on Theorem 1. If the graphon is degenerate or, more strongly, the elements of $[Y_{ij}]_{1\leq i\leq N, 1\leq j\leq M}$ are, in fact, iid, then her inferences will remain valid (since Theorem 1 specializes to the "rare events" result of Wang (2020) in that case).

Aggregate effects

Define $e_{ni}(x) = e\left(\alpha_{0,n} + z(W_i, x)'\beta_0\right)$ and $\hat{e}_{ni}(x) = e\left(\hat{\alpha}_n + z(W_i, x)'\hat{\beta}\right)$ and consider an estimate of total unit sales for a product with attribute vector $X_j = x$ of

$$\hat{\gamma}_n(x) = \sum_{i=1}^{N} \hat{e}_{ni}(x).$$

Under dense network asymptotics this statistic would diverge as $n \to \infty$. Under sparse network asymptotics the sum $\sum_{i=1}^{N} e_{ni}(x)$ behaves like an average because its summands are $O(N^{-1})$. Consequently, $\hat{\gamma}_n(x)$ has a well-defined probability limit of

$$\gamma_0(x) = \lim_{n\to\infty} \sum_{i=1}^{N} e_{ni}(x) = (1-\phi)\,\alpha_0 \mathbb{E}\left[\exp\left(z(W_i, x)'\beta_0\right)\right]. \tag{26}$$

This probability limit reflects the boundedness of average product degree in sparse networks (i.e., in expectation, total sales of a product are finite, even asymptotically).

[6] Wang (2020) scales by the square root of the number of events or "ones" in the dataset. This is, of course, of the same order as n as defined here. This difference leads to a minor difference in our two expressions for Γ_0. Making these adjustments the results coincide.

This result also holds within a sub-population of products with characteristics $X_j = x$. Parameter (26) corresponds to a conditional version of $\lambda_{0,n}^p$, the average product degree parameter defined in Section 1 above.

To derive the rate of convergence and limit distribution of $\hat{\gamma}_n(x)$ I proceed in the usual way. A mean value expansion and Theorem (1) together yield

$$\begin{aligned}\sqrt{n}\left(\hat{\gamma}_n(x) - \gamma_0(x)\right) \approx & \sqrt{n}\sum_{i=1}^{N}\left\{e_{ni}(x) - \gamma_0(x)\right\} \\ & + \sum_{i=1}^{N}\bar{e}_{ni}(x)\left[1 - \bar{e}_{ni}(x)\right]\begin{pmatrix} 1 & z(W_i,x)' \end{pmatrix}\Gamma_0^{-1} \\ & \times n^{3/2}S_n(\theta_{0,n}).\end{aligned} \tag{27}$$

By the conditional mean zero property of the score function, the two terms in (27) are asymptotically uncorrelated. The variance of the first term in (27) is

$$\mathbb{V}\left(\sqrt{n}\sum_{i=1}^{N}\left\{e_{ni}(x) - \gamma_0(x)\right\}\right) = O\left(n^2(1-\phi_n)\rho_n^2\right) = O(1),$$

whereas the Jacobian in the second term has the approximation

$$\begin{aligned}&\sum_{i=1}^{N}\bar{e}_{ni}(x)\left[1 - \bar{e}_{ni}(x)\right]\begin{pmatrix} 1 & z(W_i,x)' \end{pmatrix} = \\ &\qquad \frac{\alpha_0(1-\phi_n)}{N}\sum_{i=1}^{N}\exp\left(z(W_i,x)'\beta_0\right)\begin{pmatrix} 1 & z(W_i,x)' \end{pmatrix} + O_p\left(n^{-1}\right).\end{aligned}$$

Defining $\Phi_0(x) \overset{def}{\equiv} \alpha_0(1-\phi)\begin{pmatrix} \mathbb{E}\left[\exp\left(z(W_i,x)'\beta_0\right)\right] & \mathbb{E}\left[\exp\left(z(W_i,x)'\beta_0\right)z(W_i,x)\right]' \end{pmatrix}$, $\Lambda_0(x) = \lim_{n\to\infty} n^2(1-\phi)\,\mathbb{V}\left(e_{ni}(x) - \gamma_0(x)\right)$, and $\Omega_0 = \Gamma_0^{-1}\left[\frac{\tilde{\Sigma}_1^c}{1-\phi} + \frac{\tilde{\Sigma}_1^p}{\phi} + \frac{\tilde{\Sigma}_3}{\phi(1-\phi)}\right]\Gamma_0^{-1}$ suggest that

$$\sqrt{n}\left(\hat{\gamma}_n(x) - \gamma_0(x)\right) \to N\left(0, \Lambda_0(x) + \Phi_0(x)\Omega_0\Phi_0(x)'\right)$$

as $n \to \infty$. Aggregate effects are estimable with the same degree of precision as the logit coefficients themselves.

Average partial effects

Next consider estimating the *average* marginal effect of unit increases in the elements of Z_{ij} on making a purchase:

$$\hat{\gamma}_n = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \hat{e}_{nij} \left[1 - \hat{e}_{nij}\right] Z_{ij}. \tag{28}$$

Recall that $e_{nij} = e\left(\alpha_{0,n} + Z_{ij}\beta_0\right)$ and $\hat{e}_{nij} = e\left(\hat{\alpha}_n + Z_{ij}'\hat{\beta}\right)$. Interest in average partial effects of this type is widespread in modern micro-econometric empirical research (cf., Blundell and Powell, 2003; Wooldridge, 2005). Since (28) *is* an average of summands, each of which is $O\left(n^{-1}\right)$, we might expect some variance reduction relative to the aggregate case just discussed. In a certain sense, this conjecture appears to be correct.

Define $\gamma_{0,n} = \mathbb{E}_{N,M}\left[e_{nij}\left(1 - e_{nij}\right) Z_{ij}\right] = O\left(\rho_n\right)$; a mean-value expansion and some re-scaling yields

$$\begin{aligned} \rho_n n^{3/2} \left(\frac{\hat{\gamma}_n - \gamma_{0,n}}{\rho_n} \right) = {} & n^{3/2} \left\{ \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \left[e_{nij}\left(1 - e_{nij}\right) Z_{ij} - \gamma_{0,n} \right] \right\} \\ & + n \left\{ \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \bar{e}_{nij}\left[1 - \bar{e}_{nij}\right]\left[1 - 2\bar{e}_{nij}\right] \left[\begin{array}{cc} Z_{ij} & Z_{ij}Z_{ij}' \end{array} \right] \right\} \\ & \times n^{1/2} \left(\hat{\theta}_n - \theta_{0,n} \right). \end{aligned} \tag{29}$$

As above, the conditional mean zero property of the score function ensures that the two terms in (29) are asymptotically uncorrelated. We rescale the estimate and parameter using ρ_n since $\gamma_{0,n} \to 0$ as $n \to \infty$ (cf., Bickel et al., 2011). The need to rescale the Jacobian in (29) stems from the observation that

$$\begin{aligned} n \left\{ \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \bar{e}_{nij}\left[1 - \bar{e}_{nij}\right]\left[1 - 2\bar{e}_{nij}\right] \left[\begin{array}{cc} Z_{ij} & Z_{ij}Z_{ij}' \end{array} \right] \right\} = & \\ \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_0 \exp\left(Z_{ij}'\beta_0\right) \left[\begin{array}{cc} Z_{ij} & Z_{ij}Z_{ij}' \end{array} \right] + O_p\left(n\rho_n^2\right). & \end{aligned}$$

Next observe that first term in (29) is a two-sample U-Statistics. A Hoeffding variance decomposition gives

$$\mathbb{V}\left(\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} e_{nij}\left(1-e_{nij}\right)Z_{ij}\right) = \frac{\Lambda_{1n}^{c}}{N} + \frac{\Lambda_{1n}^{p}}{M} + \frac{1}{NM}\left[\Lambda_{2n} - \Lambda_{1n}^{c} - \Lambda_{1n}^{p}\right]$$

with

$$\begin{aligned}
\Lambda_{1n}^{c} &= \mathbb{C}\left(e_{nij}\left(1-e_{nij}\right)Z_{ij}, e_{nik}\left(1-e_{nik}\right)Z_{ik}'\right) \\
\Lambda_{1n}^{p} &= \mathbb{C}\left(e_{nji}\left(1-e_{nji}\right)Z_{ji}, e_{nki}\left(1-e_{nki}\right)Z_{ki}'\right) \\
\Lambda_{2n} &= \mathbb{V}\left(e_{nij}\left(1-e_{nij}\right)Z_{ij}\right).
\end{aligned}$$

Inspection indicates that $\Lambda_{1n}^{c} = O\left(\rho_n^2\right)$, $\Lambda_{1n}^{p} = O\left(\rho_n^2\right)$ and $\Lambda_{2n}^{c} = O\left(\rho_n^2\right)$ and hence that

$$n^3\mathbb{V}\left(\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} e_{nij}\left(1-e_{nij}\right)Z_{ij}\right) = \frac{\tilde{\Lambda}_1^{c}}{1-\phi} + \frac{\tilde{\Lambda}_1^{p}}{\phi} + O\left(n^{-1}\right).$$

Putting these calculations together suggests that

$$\rho_n n^{3/2}\left(\frac{\hat{\gamma}_n - \gamma_{0,n}}{\rho_n}\right) \xrightarrow{D} \mathcal{N}\left(\frac{\tilde{\Lambda}_1^{c}}{1-\phi} + \frac{\tilde{\Lambda}_1^{p}}{\phi} + \Phi_0\Omega_0\Phi_0'\right)$$

with $\Phi_0 \overset{def}{\equiv} \alpha_0\mathbb{E}\left[\exp\left(Z_{ij}'\beta_0\right)\left[\begin{array}{cc} Z_{ij} & Z_{ij}Z_{ij}' \end{array}\right]\right]$.

If we set $T = NM = O\left(n^2\right)$, then we have that $T^{1/4}\left(\hat{\theta}_n - \theta_{0,n}\right)$ has a Gaussian limit distribution. The rate of convergence of $\hat{\theta}_n$ toward $\theta_{0,n}$ is slow. For average partial effects we need to rescale in order ensure a meaningful probability limit. Let $\gamma_{0,n}^{*} = \gamma_{0,n}/\rho_n$ and similarly for $\hat{\gamma}_n$; the result above implies that $T^{1/4}\left(\hat{\gamma}_n^{*} - \gamma_{0,n}^{*}\right)$ is also Gaussian. In this sense the rates-of-convergence for the logit coefficients and their APEs coincide. However, if we think in terms of the resulting implied approximation to the finite sample distribution of the two parameter estimates, we have $\mathbb{V}\left(\hat{\theta}_n\right) = O\left(T^{-1/2}\right) = O\left(n^{-1}\right)$, but $\mathbb{V}\left(\hat{\gamma}_n\right) = O\left(T^{-3/2}\right) = O\left(n^{-3}\right)$. In this sense inference on APEs appears to be more precise.

Areas for additional research

For empirical researchers the main implication of this paper is to use an estimate for the variance of S_N that includes all components – even ones that are negligible under certain asymptotic sequences – when constructing standard errors. This is not a new idea. In the context of U-statistics it goes back to Hoeffding (1948). It is implicit in Holland and Leinhardt (1976) in their work on subgraph counts; see also the recent work on dyadic regression by Fafchamps and Gubert (2007), Cameron and Miller (2014) and Aronow et al. (2017), as well as that on density weighted average derivatives by Cattaneo et al. (2014). However, the small amount of extant formal limit theory for dyadic regression (cited earlier) suggests different approaches to variance estimation. This paper has outlined an asymptotic framework that provides formal justification for one of the leading "practical" approaches to inference in the presence of dyadic dependence. Graham (2020a) discusses variance estimation for dyadic regression in detail, advocating a variant of the estimate proposed by Fafchamps and Gubert (2007), Cameron and Miller (2014) and Aronow et al. (2017). Theorem 1 provides a formal justification for this recommendation.

Many outstanding questions remain. Can the above framework be generalized to a generic dyadic M-estimation setting? What is the "general" notion of "sparseness" needed for this? Extensions to semiparametric regression models are also of interest. In recent work, Menzel (2017) and Davezies et al. (2020) propose bootstrap procedures for dyadic regression. Are these procedures also valid under sparse network asymptotics and, if not, how might they be adapted to be so? The aggregate and average effect examples sketched above suggest that the systematic exploration of policy analysis questions – considered under dense network asymptotics by Graham (2020b) – would be interesting. Finally, although it seems likely that – in the absence of imposing more structure – that the composite maximum likelihood estimator is efficient, this is currently only a conjecture.

Appendix

The appendix includes proofs of the theorems stated in the main text as well as statements and proofs of supplemental lemmata – called limonata here. All notation is as established in the main text unless stated otherwise. Equation number continues in sequence with that established in the main text.

A Preliminary lemmata and proofs

Let $R_{ij} = \left(1, Z_{ij}'\right)'$ and note that, for $e(v) = \exp(v)/[1+\exp(v)]$, we have that $e'(v) = e(v)[1-e(v)]$ and $e''(v) = e(v)[1-e(v)][1-2e(v)]$. With this notation we can write the first three derivatives of the kernel function of the composite log-likelihood with respect θ_n as

$$s_{ij}(\theta_n) = (Y_{ij} - e_{ij}(\theta_n)) R_{ij} \tag{30}$$

$$\frac{\partial s_{ij}(\theta_n)}{\partial \theta'} = -e_{ij}(\theta_n)[1-e_{ij}(\theta_n)] R_{ij} R_{ij}' \tag{31}$$

$$\frac{\partial}{\partial \theta'}\left\{\frac{\partial s_{ij}(\theta_n)}{\partial \theta_p}\right\} = -e_{ij}(\theta_n)[1-e_{ij}(\theta_n)][1-2e_{ij}(\theta_n)] R_{ij} R_{ij}' R_{p,ij} \tag{32}$$

with (32) holding for for $p = 1, \ldots, \dim(\theta_n)$.

Let $\mathbf{t} = (\theta_n - \theta_{0,n})$ and recall that $\alpha_n = \ln(\alpha/n)$ and $\alpha_{0,n} = \ln(\alpha_0/n)$. This implies that $\mathbf{t} = \left(\ln(\alpha/\alpha_0), (\beta - \beta_0)'\right)'$ does not vary with n and hence that $\mathbf{t} \in \mathbb{T}$ with $\mathbb{T}$ compact by Assumption 2. Associated with any $\mathbf{t} \in \mathbb{T}$ is a $\theta_n \in \Theta_n$; furthermore associated with this θ_n is a $\theta \in \Theta$. With these preliminaries we can show that $nH_n(\theta_n)$ converges uniformly to $\Gamma(\theta)$, as defined in equation (16) of the main text.

Lemma 1. *(Uniform Hessian Convergence) Under Assumptions 1, 2 and 3*

$$\sup_{\theta \in \Theta} \|nH_n(\theta_n) - \Gamma(\theta)\| \xrightarrow{p} 0.$$

Proof. Recall that $\theta_n = \theta_{0,n} + \mathbf{t}$ and hence that $H_n(\theta_{0,n} + \mathbf{t}) = H_n(\theta_n)$. The mean value theorem, as well as compatibility of the Frobenius matrix norm with the Euclidean vector norm, gives for any $\mathbf{t}$ and $\bar{\mathbf{t}}$ both in $\mathbb{T}$,

$$\|H_n(\theta_{0,n} + \mathbf{t}) - H_n(\theta_{0,n} + \bar{\mathbf{t}})\|_{2,1} \leq \sum_{p=1}^{\dim(\theta_n)} \left\| \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\partial}{\partial \theta'} \left\{ \frac{\partial s_{ij}(\theta_{0,n} + \mathbf{t})}{\partial \theta_p} \right\} \right\|_F \|\mathbf{t} - \bar{\mathbf{t}}\|_2 .$$

Since $\mathbb{E}[e_{ij}(\theta_n)[1-e_{ij}(\theta_n)][1-2e_{ij}(\theta_n)]] = O(\rho_n) = O(n^{-1})$ we have that, inspect-

ing (31) above, for any $\mathbf{t} \in \mathbb{T}$,

$$\left\| \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\partial}{\partial \theta'} \left\{ \frac{\partial s_{ij}\left(\theta_{0,n}+\mathbf{t}\right)}{\partial \theta_p} \right\} \right\|_F = O_p\left(n^{-1}\right).$$

This gives $\left\| nH_N\left(\theta_{0,n}+\mathbf{t}\right) - nH_n\left(\theta_{0,n}+\bar{\mathbf{t}}\right) \right\|_{2,1} \leq O_p\left(1\right) \cdot \left\| \mathbf{t}-\bar{\mathbf{t}} \right\|_2$. Next, again recalling that $\theta_{0,n}+\mathbf{t} = \theta_n$, we have that

$$\begin{aligned} H_n\left(\theta_{0,n}+\mathbf{t}\right) &= -\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} e_{ij}\left(\theta_n\right)\left[1-e_{ij}\left(\theta_n\right)\right] R_{ij} R_{ij}' \\ &= -\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\alpha}{n} \exp\left(R_{ij}'\beta\right) R_{ij} R_{ij}' + O_p\left(\frac{1}{n^2}\right), \end{aligned}$$

which gives, using a law of large numbers for U-Statistics, $nH_n\left(\theta_n\right) \xrightarrow{p} \Gamma\left(\theta\right)$ for all $\mathbf{t} \in \mathbb{T}$. The claim then follows from an application of Lemma 2.9 of Newey and McFadden (1994, p. 2138). □

B Proof of Theorem 1

Asymptotic variance of the score

To prove (22), the decomposition of the variance of the score given in the main text, and hence that

$$\mathbb{V}\left(n^{3/2} S_n\right) = \frac{\tilde{\Sigma}_1^c}{1-\phi} + \frac{\tilde{\Sigma}_1^p}{\phi} + \frac{\tilde{\Sigma}_3}{\phi\left(1-\phi\right)} + O\left(n^{-1}\right)$$

Let $e_{ij} = e\left(\alpha_{0,n} + Z_{ij}'\beta_0\right)$; using the definitions given in (23) of the main text we have that

$$\begin{aligned} \Sigma_{1n}^c =& \mathbb{E}\left[\left(Y_{12}-e_{12}\right)\left(Y_{13}-e_{13}\right) R_{12} R_{13}'\right] \\ =& O\left(\rho_n^2\right) \end{aligned} \tag{33}$$

and also that

$$\begin{aligned}\Sigma_{1n}^{p} =&\mathbb{E}\left[\left(Y_{21}-e_{21}\right)\left(Y_{31}-e_{31}\right)R_{21}R_{31}'\right].\\ =&O\left(\rho_{n}^{2}\right).\end{aligned} \tag{34}$$

Turning to Σ_{2n} and Σ_{3n} we get that

$$\begin{aligned}\Sigma_{2n} =&\mathbb{E}\left[\mathbb{E}\left[\left(Y_{12}-e_{12}\right)R_{21}\middle|W_{1},X_{2},A_{1},B_{2}\right]\right.\\ &\left.\times\mathbb{E}\left[\left(Y_{12}-e_{12}\right)R_{21}\middle|W_{1},X_{2},A_{1},B_{2}\right]'\right]\\ =&O\left(\rho_{n}^{2}\right)\end{aligned} \tag{35}$$

and that

$$\begin{aligned}\Sigma_{3n} =&\mathbb{E}\left[\left\{s_{nij}-\bar{s}_{nij}\right\}\left\{s_{nij}-\bar{s}_{nij}\right\}'\right]\\ =&O\left(\rho_{n}\right)\end{aligned} \tag{36}$$

by virtue of the equality $Y_{ij}^{2}=Y_{ij}$(which holds because Y_{ij} is binary-valued).

From (13) we have that $\rho_{n}=O\left(n^{-1}\right)$, hence (33) implies that $n^{2}\Sigma_{1n}^{c}=O\left(1\right)$, (34) that $n^{2}\Sigma_{1n}^{p}=O\left(1\right)$, and (36) that $n^{2}\Sigma_{1n}^{p}=O\left(1\right)$. This gives

$$\begin{aligned}\mathbb{V}\left(S_{n}\right) =&O\left(\frac{\rho_{n}^{2}}{N}\right)+O\left(\frac{\rho_{n}^{2}}{M}\right)+O\left(\frac{\rho_{n}^{2}}{MN}\right)+O\left(\frac{\rho_{n}}{MN}\right)\\ =&O\left(\left[\frac{\lambda_{0,n}^{c}}{M}\right]^{2}\frac{1}{N}\right)+O\left(\left[\frac{\lambda_{0,n}^{c}}{M}\right]^{3}\right)+O\left(\left[\frac{\lambda_{0,n}^{c}}{M}\right]^{2}\frac{1}{MN}\right)+O\left(\frac{\lambda_{0,n}^{c}}{M}\frac{1}{MN}\right)\\ =&O\left(\left[\frac{\lambda_{0,n}^{c}}{\phi_{n}}\right]^{2}\frac{1}{\left(1-\phi_{n}\right)}\frac{1}{n^{3}}\right)+O\left(\left[\frac{\lambda_{0,n}^{c}}{\phi_{n}}\right]^{3}\frac{1}{n^{3}}\right)\\ &+O\left(\left[\frac{\lambda_{0,n}^{c}}{\phi_{n}}\right]^{2}\frac{1}{\phi_{n}\left(1-\phi_{n}\right)}\frac{1}{n^{4}}\right)+O\left(\frac{\lambda_{0,n}^{c}}{\phi_{n}^{2}\left(1-\phi_{n}\right)}\frac{1}{n^{3}}\right)\\ =&O\left(n^{3}\right)+O\left(n^{3}\right)+O\left(n^{4}\right)+O\left(n^{3}\right),\end{aligned}$$

as needed.

Triangular array setup

Consider the following triangular array $\{Z_{nt}\}$:

$$\begin{aligned}
Z_{n1} &= \frac{1}{N}\bar{s}^{c}_{1n1} \\
&\vdots \\
Z_{nN} &= \frac{1}{N}\bar{s}^{c}_{1nN} \\
Z_{nN+1} &= \frac{1}{M}\bar{s}^{p}_{1n1} \\
&\vdots \\
Z_{nN+M} &= \frac{1}{M}\bar{s}^{p}_{1nM} \\
Z_{nN+M+1} &= \frac{1}{NM}\left(s_{n11} - \bar{s}_{n11}\right) \\
&\vdots \\
Z_{nN+M+NM} &= \frac{1}{NM}\left(s_{nNM} - \bar{s}_{nNM}\right),
\end{aligned}$$

with $T = T(n) = N + M + NM$. For any vector X_i, let $X_1^t = (X_1, \ldots, X_t)'$. Iterated expectations, as well as the conditional independence relationships implied by dyadic dependence (Assumptions 1 and 2), yield

$$\mathbb{E}\left[Z_{ni} \middle| Z_{n1}^{i-1}\right] = 0,$$

establishing that $\{Z_{Ni}\}$ is a martingale difference sequence (MDS). The variance of this MDS is

$$\begin{aligned}
\bar{\Delta}_n &\overset{def}{\equiv} \mathbb{V}\left(\sum_{t=1}^{T} Z_{ni}\right) \\
&= \frac{\Sigma^{c}_{1n}}{N} + \frac{\Sigma^{p}_{1n}}{M} + \frac{\Sigma_{3n}}{NM}.
\end{aligned}$$

To show asymptotic normality of $N^{3/2}S_n(\theta_{0,n})$ I first show, recalling decomposition (17) in the main test, that, for a vector of constants c,

$$\left(c'\bar{\Delta}_n c\right)^{-1/2} c'S_n = \left(c'\bar{\Delta}_n c\right)^{-1/2} c'\left[U_{1n} + V_n\right] + o_p(1) \tag{37}$$

and subsequently that

$$\left(c'\bar{\Delta}_n c\right)^{-1/2} c'\left[U_{1n}+V_n\right] \xrightarrow{p} \mathcal{N}\left(0,1\right). \tag{38}$$

To show (37) observe that

$$\begin{aligned} c'\bar{\Delta}_n c &= O\left(\frac{\rho_n^2}{N}+\frac{\rho_n^2}{M}+\frac{\rho_n}{NM}\right) \\ &= O\left(\frac{\rho_n^2}{n}\left(\frac{1}{1-\phi_n}+\frac{1}{\phi_n}+\frac{1}{\left(1-\phi_n\right)\lambda_n^c}\right)\right) \\ &= O\left(\frac{\rho_n^2}{n}\right) \end{aligned}$$

and hence that $\left(c'\bar{\Delta}_N c\right)^{-1} = O\left(n\rho_n^{-2}\right)$ as long as $\lambda_n^c \geq C > 0$ and $\phi \in (0,1)$ (see Assumptions 1 and 2). Additionally using (35) yields

$$\begin{aligned} \left(c'\bar{\Delta}_n c\right)^{-1/2} c'U_{2n} &= O_p\left(n^{1/2}\rho_n^{-1}\right) O_p\left(\rho_n^2\right) \\ &= O_p\left(n^{1/2}\rho_n\right) \\ &= o_p\left(1\right), \end{aligned}$$

as long as $\rho_n = O\left(n^{-\alpha}\right)$ for $\alpha > \frac{1}{2}$, as is maintained here. This proves assertion (37).

Central limit theorem

To show (38) I verify the conditions of Corollary 5.26 of Theorem 5.24 in White (2001); specifically the Lyapunov condition, for $r > 2$

$$\sum_{t=1}^{T(n)} \mathbb{E}\left[\left(\frac{c'Z_{nt}}{\left(c'\bar{\Delta}_{nt}c\right)}\right)^r\right] = o\left(1\right) \tag{39}$$

and the stability condition

$$\sum_{t=1}^{T(n)} \frac{\left(c'Z_{Nt}\right)^2}{c'\bar{\Delta}_N c} \xrightarrow{p} 1. \tag{40}$$

I will show (39) for $r = 3$. Observe that

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{N}c'\bar{s}^c_{1ni}\right)^3\right] =& O\left(\frac{\rho_n^3}{N^3}\right) \\
\mathbb{E}\left[\left(\frac{1}{M}c'\bar{s}^p_{1ni}\right)^3\right] =& O\left(\frac{\rho_n^3}{M^3}\right) \\
\mathbb{E}\left[\left(\frac{1}{NM}c'\left(s_{n11}-\bar{s}_{n11}\right)\right)^3\right] =& O\left(\frac{\rho_n}{N^3M^3}\right)
\end{aligned}$$

These calculations, as well as independence of summands 1 to N, $N+1$ to $N+M$ and $N+M+1$ to $N+M+NM$, imply that

$$\begin{aligned}
\sum_{t=1}^{T(n)}\mathbb{E}\left[\left(\frac{c'Z_{Nt}}{\left(c'\bar{\Delta}_N c\right)}\right)^3\right] =& O_p\left(n^{3/2}\rho_N^{-3}\right)\left\{O\left(\frac{\rho_n^3}{N^2}\right)+O\left(\frac{\rho_n^3}{M^2}\right)+O\left(\frac{\rho_n}{N^2M^2}\right)\right\} \\
=& O_p\left(\frac{1}{\left(1-\phi_n\right)^2 n^{1/2}}\right)+O_p\left(\frac{1}{\phi_n^2 n^{1/2}}\right)+O_p\left(\frac{1}{\left(1-\phi_n\right)^2\lambda_n^c n^{1/2}}\right) \\
& O_p\left(n^{-1/2}\right) \\
& o_p\left(1\right)
\end{aligned}$$

as required.

To verify the stability condition (40) I re-write it as

$$\sum_{t=1}^{T(n)}\frac{1}{n\left(c'\bar{\Delta}_n c\right)}n\left\{\left(c'Z_{nt}\right)^2-\mathbb{E}\left[\left(c'Z_{nt}\right)^2\right]\right\}\xrightarrow{p}0 \tag{41}$$

Since $n\left(c'\bar{\Delta}_N c\right)^{-1} = O\left(n\cdot n\rho_N^{-2}\right) = O\left(1\right)$ the stability condition (40) will hold if the numerator in (41) – $\sum_{t=1}^{T(n)} n\left\{\left(c'Z_{nt}\right)^2-\mathbb{E}\left[\left(c'Z_{nt}\right)^2\right]\right\}$ – converges in probability to zero. Expanding the square we get that

$$\mathbb{E}\left[\left(n\left\{\left(c'Z_{nt}\right)^2-\mathbb{E}\left[\left(c'Z_{nt}\right)^2\right]\right\}\right)^2\right] = n^2\left\{\mathbb{E}\left[\left(c'Z_{nt}\right)^4\right]-\left(\mathbb{E}\left[\left(c'Z_{nt}\right)^2\right]\right)^2\right\}.$$

We then have

$$\mathbb{E}\left[\left(c'Z_{nt}\right)^2\right] = \begin{cases} \frac{1}{N^2}c'\Sigma_{1n}^c c = O\left(\left[\frac{\lambda_n^c}{\phi_n(1-\phi_n)}\right]^2 \frac{1}{n^4}\right), & t = 1,\ldots,N \\ \frac{1}{M^2}c'\Sigma_{1n}^p c = O\left(\left[\frac{\lambda_n^c}{\phi_n^2}\right]^2 \frac{1}{n^4}\right), & t = N+1,\ldots,N+M \\ \frac{1}{N^2M^2}c'\Sigma_{3N}c = O\left(\frac{\lambda_n^c}{\phi_n^3(1-\phi_n)^2}\frac{1}{n^5}\right), & t = N+M+1,\ldots,N+M+NM \end{cases}$$

and

$$\mathbb{E}\left[\left(c'Z_{nt}\right)^4\right] = \begin{cases} \frac{\mathbb{E}\left[\left(c'\bar{s}_{1n1}^c\right)^4\right]}{N^4} = O\left(\frac{1}{(1-\phi_n)^4}\frac{\rho_n^4}{n^4}\right), & t = 1,\ldots,N \\ \frac{\mathbb{E}\left[\left(c'\bar{s}_{1n1}^p\right)^4\right]}{M^4} = O\left(\frac{1}{\phi_n^4}\frac{\rho_n^4}{n^4}\right), & t = N+1,\ldots,N+M \\ \frac{\mathbb{E}\left[\left(c'(s_{n11}-\bar{s}_{n11})\right)^4\right]}{N^4M^4} = O\left(\frac{1}{\phi_n^4(1-\phi_n)^4}\frac{\rho_n}{n^8}\right), & t = N+M+1,\ldots,N+M+NM \end{cases}.$$

Since $T(n) = N + M + NM = O(n^2)$, the summands of $\frac{1}{T(n)}\sum_{t=1}^{T(n)} T(n)\, n\left\{\left(c'Z_{nt}\right)^2 - \mathbb{E}\left[\left(c'Z_{nt}\right)^2\right]\right\}$ all have variances which are $O(n^{-2})$ or smaller:

$$T(n)^2 n^2 \left\{\mathbb{E}\left[\left(c'Z_{nt}\right)^4\right] - \left(\mathbb{E}\left[\left(c'Z_{nt}\right)^2\right]\right)^2\right\} =$$

$$\begin{cases} T(n)^2 n^2 \left[O(n^{-8}) + O(n^{-8})\right] = O(n^{-2}), & t = 1,\ldots,N \\ T(n)^2 n^2 \left[O(n^{-8}) + O(n^{-8})\right] = O(n^{-2}), & t = N+1,\ldots,N+M \\ T(n)^2 n^2 \left[O(n^{-9}) + O(n^{-10})\right] = O(n^{-3}), & t = N+M+1,\ldots,N+M+NM \end{cases}$$

Since the summands of the numerator in (41) are all mean zero with variances shrinking to zero as $n \to \infty$ condition (41) holds as required.

References

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581 – 598.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Aronow, P. M., Samii, C., and Assenova, V. A. (2017). Cluster-robust variance estimation for dyadic data. *Political Analysis*, 23(4):564 – 577.

Bengtsson, O. and Hsu, D. H. (2015). Ethnic matching in the u.s. venture capital market. *Journal of Business Venturing*, 30(2):338 – 354.

Bickel, P. J., Chen, A., and Levina, E. (2011). The method of moments and degree distributions for network models. *Annals of Statistics*, 39(5):2280 – 2301.

Blundell, R. and Powell, J. L. (2003). *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, volume 2, chapter Endogeneity in nonparametric and semiparametric regression models, pages 312 – 357. Cambridge University Press.

Cameron, A. C. and Miller, D. L. (2014). Robust inference for dyadic data. Technical report, University of California - Davis.

Cattaneo, M., Crump, R., and Jansson, M. (2014). Small bandwidth asymptotics for density-weighted average derivatives. *Econometric Theory*, 30(1):176 – 200.

Chamberlain, G. (1984). *Handbook of Econometrics*, volume 2, chapter Panel Data, pages 1247 – 1318. North-Holland, Amsterdam.

Chartrand, G. and Zhang, P. (2012). *A First Course in Graph Theory*. Dover Publications.

Crane, H. and Towsner, H. (2018). Relatively exchangeable structures. *Journal of Symbolic Logic*, 83(2):416 – 442.

Davezies, L., d'Haultfoeuille, X., and Guyonvarch, Y. (2020). Empirical process results for exchangeable arrayes. *Annals of Statistics*.

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale*, 4:251 – 299.

Fafchamps, M. and Gubert, F. (2007). The formation of risk sharing networks. *Journal of Development Economics*, 83(2):326 – 350.

Fox, J. T. (2018). Estimating matching games with transfers. *Quantitative Economics*, 9(1):1 – 38.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica*, 85(4):1033 – 1063.

Graham, B. S. (2020a). *The Econometrics of Social and Economic Networks*, chapter Dyadic regression, pages 25 – 41. Elsevier, Amsterdam.

Graham, B. S. (2020b). *Handbook of Econometrics*, volume 7, chapter Network data. North-Holland, Amsterdam.

Graham, B. S., Imbens, G. W., and Ridder, G. (2018). Identification and efficiency bounds for the average match function under conditionally exogenous matching. *Journal of Business and Economic Statistics*.

Graham, B. S., Niu, F., and Powell, J. L. (2019). Kernel density estimation for undirected dyadic data. Technical report, University of California - Berkeley.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293 – 325.

Holland, P. W. and Leinhardt, S. (1976). Local structure in social networks. *Sociological Methodology*, 7:1 – 45.

Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. Technical report, Institute for Advanced Study, Princeton, NJ.

Jochmans, K. (2018). Semiparametric analysis of network formation. *Journal of Business and Economic Statistics*, 36(4):705 – 713.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2):137 – 163.

KPMG (2016). Competitive alternatives: Kpmg's guide to international business locations costs. Technical report, KMPG International Cooperative.

Lewbel, A. and Nesheim, L. (2019). Sparse demand systems: corners and complements. CeMMAP Working Paper CWP45/19, Centre for Microdata Methods and Practice.

Lindsey, B. G. (1988). Composite likelihood. *Contemporary Mathematics*, 80:221 – 239.

Marotta, L., Miccichè, S., Fujiwara, Y., Iyetomi, H., Aoyama, H., Gallegati, M., and Mantegna, R. N. (2015). Bank-firm credit network in japan: an analysis of a bipartite network. *Plos One*, 10(5):e0123079.

Menzel, K. (2017). Bootstrap with clustering in two or more dimensions. Technical Report 1703.03043v2, arXiv.

Newey, W. K. and McFadden, D. (1994). *Handbook of Econometrics*, volume 4, chapter Large sample estimation and hypothesis testing, pages 2111 – 2245. North-Holland, Amsterdam.

Owen, A. B. (2007). Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8:761 – 773.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557 – 586.

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

Wang, H. (2020). Pmlr. In *Proceedings of the 37 th International Conference on Machine Learning*, number 119.

White, H. (2001). *Asymptotic Theory for Econometricians*. Academic Press, San Diego.

Wooldridge, J. M. (2005). *Identification and inference for econometric models*, chapter Unobserved heterogeneity and the estimation of average partial effects, pages 27 – 55. Number 3. Cambridge University Press, Cambridge.