Sparse network asymptotics for logistic regression under possible misspecification

Bryan S. Graham[*]

First Draft: September 2020, This Draft: September 2022

Abstract

[*] Department of Economics, University of California - Berkeley, 530 Evans Hall #3380, Berkeley, CA 94720-3888, e-mail: bgraham@econ.berkeley.edu, web: http://bryangraham.github.io/econometrics/. Financial support from NSF Grant SES #1851647 is gratefully acknowledged. Portions of the research reported here were also undertaken at CEMFI with support from the Spanish State Research Agency under the María de Maeztu Unit of Excellence Programme (Project No: CEX2020-001104-M). I am extraordinarily grateful to CEMFI for this support. 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. This revision has further benefited from the comments, questions and suggestions of Harold Chiang, three especially insightful anonymous reviewers, and the Guest Co-Editor. All the usual disclaimers apply.

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 logit fit of the $N \times M$ array of "i-buys-j" purchase decisions, $\mathbf{Y} = [Y_{ij}]_{1 \leq i \leq N, 1 \leq j \leq M}$, onto a vector of known functions of consumer and product attributes under asymptotic sequences where (i) both N and M grow large, (ii) the average number of products purchased per consumer is finite in the limit, (iii) there exists dependence across elements in the same row or same column of $\mathbf{Y}$ (i.e., dyadic dependence) and (iv) the true conditional probability of making a purchase may, or may not, take the assumed logit form. Condition (ii) implies that the limiting network of purchases is *sparse*: only a vanishing fraction of all possible purchases are actually made. Under sparse network asymptotics, I show that the parameter indexing the logit approximation solves a particular Kullback–Leibler Information Criterion (KLIC) minimization problem (defined with respect to a certain Poisson population). This finding provides a simple characterization of the logit pseudo-true parameter under general misspecification. With respect to sampling theory, sparseness implies that the first and last terms in an extended Hoeffding-type variance decomposition of the score of the logit pseudo 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. Simulation results suggest that sparse network asymptotics better approximate the finite network distribution of the logit estimator.

JEL Codes: C31, C33, C35

Keywords: Networks, Exchangeable Random Arrays, Dyadic Clustering, Sparse Networks, Logistic Regression, Rare Events, Bipartite Network, Alternative Asymptotics, Sparse Network Asymptotics

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 $W_i \in \mathbb{W}$ be a vector of observed consumer attributes, $X_j \in \mathbb{X}$ a vector of product attributes and $n \overset{def}{\equiv} M + N$ the total number of sampled consumers and products. The conditional probability that consumer i buys product j is given by

$$\Pr\left(Y_{ij} = 1 \middle| W_i, X_j\right) = g_n\left(W_i, X_j\right) \tag{1}$$

with $g_n : \mathbb{W} \times \mathbb{X} \to \{0, 1\}$ an unknown regression function. In this paper I consider the statistical properties of (a sequence of) parametric logit approximations of $g_n(w, x)$ when (i) both N and M grow large at the same rate (i.e., $M/n \to \phi \in (0, 1)$ as $n \to \infty$), (ii) the limiting purchase graph $\mathbf{Y} \overset{def}{\equiv} [Y_{ij}]_{1 \leq i \leq N, 1 \leq j \leq M}$ is *sparse,* and (iii) there exists *dyadic dependence* (i.e., $Y_{i_1 j_1}$ and $Y_{i_2 j_2}$ may covary whenever $\{i_1, j_j\}$ and $\{i_2, j_2\}$ share a common consumer or product index). Dyadic dependence arises in the presence of unobserved consumer- and/or product-specific heterogeneity.

The novelty relative to prior work on dyadic regression by Graham (2020a,b), Menzel (2021), Davezies et al. (2021) and others involves (i) the introduction of "sparse network asymptotics" and (ii) an analysis which accommodates misspecification of the regression function. The sparse network thought experiment introduced in this paper leads to novel asymptotic approximations which appropriately account for effects of dyadic dependence when present, while simultaneously being robust to its absence (and other forms of degeneracy).[1] Accommodating misspecification allows researchers to conduct inference on well-defined pseudo-true parameters in settings where their model for (1) is only an approximation (as is invariably the case in practice).

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).

Sparseness

Let $\rho_n = \mathbb{E}_n[Y_{ij}]$ be the probability of the event that (randomly sampled) consumer i buys (randomly sampled) product j. The notation $\mathbb{E}_n[\cdot]$ is used to emphasize that the probability law used to evaluate the expectation may vary with n (below I use the notation $\mathbb{E}_0[\cdot]$ to indicate an average with respect to the limiting probability law as

[1] An important precedent for the asymptotic thought experiment considered below is the work Bickel et al. (2011). They study the properties of acyclic subgraph frequencies under sparseness.

$n \to \infty$). Sparseness of the limit graph implies that the average consumer purchases only a finite number of products in the limit:

$$\lambda_n^c \overset{def}{\equiv} M\rho_n \to \lambda_0^c \text{ with } 0 < \lambda_0^c < \infty \text{ as } n \to \infty. \tag{2}$$

Condition (2) is concordant with the fact that, for example, although consumers choose from tens of millions of books, it is rare for individual libraries to exceed a few hundred volumes (i.e., average consumer degree λ_n^c is small). Similarly, the lifetime sales of most books rarely exceed several hundred copies, such that

$$\lambda_n^p \overset{def}{\equiv} N\rho_n \to \lambda_0^p \text{ with } 0 < \lambda_0^p < \infty \text{ as } n \to \infty \tag{3}$$

(i.e., average product degree λ_n^p is also small).

Conditions (2) and (3) restrict the sequence of regression functions (1) such that

$$\mathbb{E}_n \left[g_n \left(W_i, X_j\right)\right] \overset{def}{\equiv} \rho_n = O\left(n^{-1}\right). \tag{4}$$

Equation (4) implies that the number of purchases actually made is negligible relative to the set of all possible purchases that could have been made; the purchase graph $\mathbf{Y}$ is *sparse*. If, instead, the marginal purchase probability ρ_n was fixed at, or converged to, a constant between zero and one, then the number of actual book purchases and the number of possible book purchases would be of equal order (the so-called *dense* case). Sparseness is a property of a *sequence* of graphs, each with an increasing number of vertices. It is used here in the context of a particular asymptotic approximation argument, motivated by the fact that in many real world graphs the number of edges present is small relative to the number that could be present (e.g., Newman, 2010).

Dyadic dependence

Dyadic dependence refers to a particular pattern of dependence across the rows and columns of $\mathbf{Y}$. 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 product 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), Tabord-Meehan (2018), Graham (2020a), Davezies et al. (2021) and Menzel (2021) (see Graham (2020b, Section 4) for a review and references). This work generally considers the dense case. Dyadic dependence, in the context of the sparse network asymptotics explored below, generates new issues.

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 (see 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 formally maintained below, but invoking it here simplifies the discussion of exchangeability.

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

$$\left[W_{\sigma_w(i)}\right]_{i \in \mathbb{N}} = [W_i]_{i \in \mathbb{N}}. \tag{5}$$

Restriction (5) implies that σ_w only permutes indices across observationally identical consumers (i.e., those homogenous in 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*

$$\left[Y_{\sigma_w(i)\sigma_x(j)}\right]_{i\in\mathbb{N},j\in\mathbb{M}} \overset{D}{=} \left[Y_{ij}\right]_{i\in\mathbb{N},j\in\mathbb{M}}. \tag{6}$$

Here $\overset{D}{=}$ denotes equality of distribution. One way to think about (6) 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 W_i and X_j are associated 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 (6) 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

$$Y^*_{ij} = h\left(\mu, W_i, X_j, A_i, B_j, V_{ij}\right) \tag{7}$$

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 (7), such that the two arrays have the same distribution. An implication of this result is that we may use (7) as a nonparametric data generating process for $[Y_{ij}]_{i\in\mathbb{N},j\in\mathbb{M}}$.

Inspection of (7) 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 observed consumer and product attributes. This is, of course, precisely the dependence structure discussed earlier.

The aggregate shock, μ, in (7) is analogous to the latent mixing variable appearing in de Finetti's (1931) original theorem. The distribution of μ is never identified and the inference results described below may be (informally) thought of as being conditional on its realization; see Menzel (2021) for additional relevant discussion. Formally, the analysis which follows works with a restriction of (7) which excludes μ:

$$Y_{ij}^{*} = h\left(W_i, X_j, A_i, B_j, V_{ij}\right). \tag{8}$$

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* with graphon (8). Let $G_{\infty,\infty}$ denote this population network. 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:

$$G_{N,M} = G_{\infty,\infty}\left[\mathcal{V}_c, \mathcal{V}_p\right]. \tag{9}$$

The marginal probability of the event, random consumer i, purchases random product j, is thus

$$\rho_0 = \mathbb{E}\left[h\left(W_i, X_j, A_i, B_j, V_{ij}\right)\right]. \tag{10}$$

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*,

$$\lambda_n^c = M\rho_0 \tag{11}$$

will diverge as $M \to \infty$ when $0 < \rho_0 < 1$. Likewise the average number of times a given product is purchased, or *average product degree*,

$$\lambda_n^p = N\rho_0 \tag{12}$$

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, as discussed earlier, only a small fraction of all possible purchases are made in many real world settings. To capture this qualitatively in my 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 $n = M + N$,

$$M/n \to \phi \in (0, 1) \tag{13}$$

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

$$Y_{ij} = h_n \left(W_i, X_j, A_i, B_j, V_{ij} \right). \tag{14}$$

This sequence of graphons/populations $\{h_n\}$ has the property that network *density*

$$\rho_n = \mathbb{E}_n \left[h_n \left(W_i, X_j, A_i, B_j, V_{ij} \right) \right]$$

may approach zero as $n \to \infty$. (It would be technically more appropriate to index the sequence $\{h_n\}$ by both N and M, as opposed to just n, however doing so adds no real insight and clutters the notation.) Under this setup the order of $\lambda_n^c = M\rho_n$ and $\lambda_n^p = N\rho_n$ will depend upon the speed with which ρ_n approaches zero as $n \to \infty$.

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 qualitative 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).

The following two assumptions provide the foundation for the sparse network limit theory presented below.

Assumption 1. *(Sampling) (i) $i = 1, \ldots, N$ and $j = 1, \ldots, M$ index independent random samples of consumers ($\mathbb{N}$) and products ($\mathbb{M}$) respectively; (ii) $W_i \in \mathbb{W}$, with $\mathbb{W}$ a compact subset of $\mathbb{R}^{\dim(W_i)}$ and $f_W(w)$ bounded and bounded away from zero on $\mathbb{W}$; similarly $X_j \in \mathbb{X}$, with $\mathbb{X}$ a compact subset of $\mathbb{R}^{\dim(X_j)}$ and $f_X(x)$ bounded and bounded away from zero on $\mathbb{X}$; (iii) $[Y_{ij}]_{1 \leq i \leq N, 1 \leq j \leq M}$ is generated according to* (14)*; (iv) the sequence of samples is such that $\frac{M}{M+N} \to \phi \in (0,1)$ as $N, M \to \infty$.*

The sequence of graphons $\{h_n\}$ is left nonparametric, but restricted such that in the limit the graph is sparse (i.e., conditions (2) and (3) above hold). To ensure this property I impose the stronger condition, observing that $\mathbb{E}_n \left[h_n \left(W_i, X_j, A_i, B_j, V_{ij} \right) \middle| W_i, X_j \right] = g_n \left(W_i, X_j \right)$:

Assumption 2. *(Conditional Sparseness) : The graphon sequence $\{h_n\}$ is such that (i)*

$$n g_n (w, x) = \lambda_0 (w, x) + o\left(n^{-1}\right)$$

with $0 < \lambda_0 (w, x) < \infty$ for all $(w, x) \in \mathbb{W} \times \mathbb{X}$ and (ii) $n g_n (w, x) \leq k (w, x)$ for all n and $(w, x) \in \mathbb{W} \times \mathbb{X}$ with $\mathbb{E}\left[k \left(W_i, X_j\right)\right] < \infty$.

Assumption 2 implies that the conditional probability that a type $W_i = w$ customer buys a type $X_j = x$ product is $O\left(n^{-1}\right)$ for all $(w, x) \in \mathbb{W} \times \mathbb{X}$. This restriction has two important implications for the analysis which follows.

First, it ensures, as desired, that the limiting graph is *sparse.* Let $\lambda_0 = \lambda_0^c + \lambda_0^p$ equal the sum of the limiting average consumer and product degrees. Note that $n\rho_n \to \lambda_0$ and further that $\lambda_0 = \mathbb{E}\left[\lambda_0 \left(W_i, X_j\right)\right]$. In what follows I will call $\lambda_0 \left(w, x\right)$ the (limiting) *conditional degree function.*

Second, it implies that consumer and product attributes do not affect the *order* of the probability that an edge forms. It rules out, for example, the existence of observable subpopulations of products, say those with $X_j = x$, that are purchased by a non-trivial fraction of consumers of, say, type $W_i = w$. This can be restrictive: if i indexes moviegoers and j films, then it rules out film types $X_j = x$ which "everyone" sees. In contrast, if i indexes econometricians and j research articles, it seems reasonable to assume that there are no observable econometrician-article combinations, $W_i = w$, $X_j = x$, where the event i cites j occurs with high probability.[2] Indeed, sparseness of the type imposed by Assumption 2 appears to be a useful description of many real world graphs (Newman, 2010).

2 Pseudo composite likelihood estimator

The estimation target is the coefficient vector indexing (an approximation of) the regression function $g_n\left(w, x\right)$. Other than the sparseness restrictions imposed by Assumption 2, the form of $g_n\left(w, x\right)$ is left unspecified. Let $Z_{ij} = z\left(W_i, X_j\right)$ be a vector of known basis functions in the underlying consumer and product attributes W_i and X_j (excluding the constant) and consider the *sequence* – indexed by n – of parametric logit models:

$$e_n\left(W_i, X_j; \theta\right) = \frac{\exp\left(\alpha + Z_{ij}'\beta - \ln n\right)}{1 + \exp\left(\alpha + Z_{ij}'\beta - \ln n\right)}, \tag{15}$$

where $\theta = \left(\alpha, \beta'\right)'$.

Sequence (15) has the feature that

$$ne_n\left(W_i, X_j; \theta\right) \to \exp\left(\alpha + Z_{ij}'\beta\right)$$

as $n \to \infty$ and hence shares the sparseness features of the population graphon $g_n\left(w, x\right)$. Its implied (limiting) average consumer and product degrees are

$$\lambda^c\left(\phi, \theta\right) = \phi\mathbb{E}\left[\exp\left(\alpha + Z_{ij}'\beta\right)\right], \ \lambda^p\left(\phi, \theta\right) = \left(1 - \phi\right)\mathbb{E}\left[\exp\left(\alpha + Z_{ij}'\beta\right)\right].$$

For large n, the logit model is shown to provide a well-defined approximation of the conditional degree function $\lambda_0\left(w, x\right)$. Furthermore, the pseudo-true parameter values indexing this approximation are consistently estimable with a Gaussian limit distribution.

[2] Feedback from the referees was especially helpful in formulating Assumption 2.

Note that in the event that $g_n(w,x)$ happens to take the logit form, Assumption 2 holds since, with $g_n(w,x) = e_n(W_i, X_j; \theta_0)$ and $\lambda_0(w,x) = \exp\left(\alpha_0 + Z'_{ij}\beta_0\right)$, we have

$$\begin{aligned} ng_n(w,x) - \lambda_0(w,x) &= \left[\frac{\exp\left(\alpha_0 + Z'_{ij}\beta_0\right)}{1 + \frac{1}{n}\exp\left(\alpha_0 + Z'_{ij}\beta_0\right)} - \exp\left(\alpha_0 + Z'_{ij}\beta_0\right)\right] \\ &= -\exp\left(\alpha_0 + Z'_{ij}\beta_0\right)\left[\frac{\frac{1}{n}\exp\left(\alpha_0 + Z'_{ij}\beta_0\right)}{1 + \frac{1}{n}\exp\left(\alpha_0 + Z'_{ij}\beta_0\right)}\right] \\ &= o\left(n^{-1}\right). \end{aligned}$$

(we can also set $k(w,x) = \lambda_0(w,x)$).

Estimation

To estimate θ I propose maximizing the pseudo composite log-likelihood function

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L_n(\theta) \tag{16}$$

with $L_n(\theta) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} l_{ij,n}(\theta)$ and $l_{ij,n}(\theta)$ the logit kernel function:

$$l_{ij,n}(\theta) = (2Y_{ij} - 1)\left(R'_{ij}\theta - \ln n\right) - \ln\left(1 + \exp\left((2Y_{ij} - 1)\left[R'_{ij}\theta - \ln n\right]\right)\right) \tag{17}$$

for $R_{ij} \overset{def}{\equiv} \left(1, Z'_{ij}\right)'$. The use of the word 'composite' emphasizes that the criterion function only models the data at the dyad level; no attempt is made to model the precise structure of dependence across dyads (see Lindsey, 1988; Cox and Reid, 2004). The use of the word 'pseudo' emphasizes the allowance for misspecification of the dyad-level regression function. Indeed the analysis in this paper is potentially compatible with a wide variety of actual network generating process; whether the estimated regression function approximation has any structural economic significance or is simply a predictor for Y_{ij} given W_i and X_j will vary from application to application.

Consistency

Let $\theta_0 = (\alpha_0, \beta'_0)'$ denote the pseudo-true value of θ; θ_0 indexes a unique "best approximation" of the conditional degree function $\lambda_0(w,x)$. To characterize this "best approximation" Lemma 1 below provides a uniform convergence result for the pseudo composite log-likelihood function. This result is used to both characterize the population approximation problem for which θ_0 is the unique solution and to demonstrate consistency of the maximum pseudo composite likelihood estimate $\hat{\theta}$ for θ_0.

In addition to Assumptions 1 and 2 above, I require a standard identification condition (e.g., Amemiya, 1985, p. 270).

Assumption 3. *(Identification)*
(i) $\theta_0 = (\alpha_0, \beta_0')' \in \mathbb{A} \times \mathbb{B} = \Theta$, $\mathbb{A}$ *and* $\mathbb{B}$ *compact;*
(ii) $Z_{ij} \in \mathbb{Z}$ *with* $\mathbb{Z}$ *a compact subset of* $\mathbb{R}^{\dim(Z_{ij})}$ *with* $f_Z(z)$ *bounded on* $z \in \mathbb{Z}$*;*
(iii) $\mathbb{E}\left[Z_{ij}Z_{ij}'\right]$ *is a finite non-singular matrix;*
(iv) $\mathbb{V}\left(\left|Y_{ij}\right| W_i, X_j, A_i, B_j\right) \geq \kappa > 0$.

Let $f_0\left(v\right| w, x)$ be the Poisson probability mass function (pmf) with rate parameter $\lambda_0(x, w)$ and $f\left(v\right| w, x; \theta)$ the one with rate parameter $\lambda(z;\theta) = \exp(\alpha + z'\beta)$. The corresponding distribution functions are F_0 and F_θ. Let $\delta_n \overset{def}{\equiv} \frac{\ln(n)}{NM}\sum_{i=1}^N \sum_{j=1}^M Y_{ij}$; in Appendix A I show:

Lemma 1. *(Limiting Objection Function) Let* $L_n^*(\theta) = L_n(\theta) + \delta_n$*. Under Assumptions 1, 2 and 3*

$$\sup_{\theta\in\Theta}\left|nL_n^*(\theta) - L_0(\theta)\right| \xrightarrow{p} 0$$

as $n \to \infty$ *with*

$$L_0(\theta) = -\mathbb{D}_{KL}\left(F_0\|\, F_\theta\right) + \mathbb{S}(F_0),$$

where $\mathbb{D}_{KL}\left(F_0\|\, F_\theta\right) \overset{def}{\equiv} \mathbb{E}_0\left[\ln\left\{\frac{f_0(V_{ij}|W_i, X_j)}{f(V_{ij}|W_i, X_j;\theta)}\right\}\right]$ *in the Kullback–Leibler divergence from* F_θ *to* F_0 and $\mathbb{S}(F_0) \overset{def}{\equiv} \mathbb{E}_0\left[\lambda_0(W_i, X_j)\ln\lambda_0(W_i, X_j)\right] - \mathbb{E}_0\left[\lambda_0(W_i, X_j)\right]$ does not vary with θ.

The addition of δ_n to $L_n(\theta)$ ensures the existence of a well-defined limit; since it does not change the value of $\hat{\theta}$, replacing $L_n(\theta)$ with $L_n^*(\theta)$ does not change inference. The $\mathbb{E}_0[\cdot]$ notation in the definition of $\mathbb{D}_{KL}\left(F_0\|\, F_\theta\right)$ indicates that V_{ij} is (conditionally) Poisson with rate parameter $\lambda_0(X_i, W_j)$; which may or may not coincide with $\lambda(Z_{ij};\theta) = \exp\left(\alpha + Z_{ij}'\beta\right)$.

Lemma 1 suggests the follow pseudo-true parameter as a target for estimation:

$$\theta_0 = \arg\min_{\theta\in\Theta}\mathbb{D}_{KL}\left(F_0\|\, F_\theta\right). \tag{18}$$

Equation (18) indicates that θ_0 indexes the best approximation, in the (Poisson) Kullback–Leibler divergence sense, of $\lambda_0(x, w)$ – averaged over the distribution of W_i and X_j – in the family of exponential parametric conditional degree functions $\{\exp(\alpha + Z_{12}'\beta) : \alpha \in \mathbb{A}, \beta \in \mathbb{B}\}$. If $e_n(w, x;\theta_0) = g_n(w, x)$ for all $(w, x) \in \mathbb{W} \times \mathbb{X}$, then θ_0 indexes the true probability law for the graph.

The purchase graph $[Y_{ij}]_{1\leq i\leq N, 1\leq j\leq M}$ coincides with the outcome of NM dependent and heterogenous Bernoulli trials, each with $O\left(n^{-1}\right)$ success probabilities. Given this structure it is (perhaps) ex post unsurprising that the limiting criterion function, and hence the form of the pseudo-true parameter θ_0, is related to the Poisson distribution. The Bernoulli distribution with small success probabilities is well-approximated by the

Poisson distribution (Mises, 1921; Hodges and Le Cam, 1960). The take away for the analysis at hand, is that $\lambda(z;\theta_0) = \exp(\alpha_0 + z'\beta_0)$ is as close as possible to $\lambda_0(x,w)$ over $(w,x) \in \mathbb{W} \times \mathbb{X}$ in a well-defined and interpretable way.

Theorem 1. *(Consistency) Under Assumptions 1, 2 and 3 (i) θ_0 is the unique maximizer of $L_0(\theta)$, as defined in Lemma 1, and (ii) the maximum pseudo composite likelihood estimate $\hat{\theta} \xrightarrow{p} \theta_0$.*

Proof. See Appendix A. □

Asymptotic normality

The limit distribution of $\hat{\theta}$ under dense network asymptotics was derived by Graham (2020b,a). More general results for dyadic M-estimators under dense network asymptotics, including results on the bootstrap and cross-fitting, can be found in Menzel (2021), Davezies et al. (2021) and Chiang et al. (2022a). None of these results apply here. To derive a result that does apply, begin with the mean value expansion

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \left[nH_n\left(\bar{\theta}\right)\right]^{+} \times n^{3/2} S_n(\theta_0),$$

where F^{+} denotes a generalized inverse if the matrix F and

$$S_n(\theta) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} s_{ij,n}(\theta), \tag{19}$$

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

$$H_n(\theta) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \frac{\partial^2 l_{ij,n}(\theta)}{\partial \theta \partial \theta'} \tag{20}$$

to the associated Hessian matrix. Here $\bar{\theta}$ is a mean value between θ_0 and $\hat{\theta}$ which may vary from row to row.

Lemma 2, stated and proved in Appendix A, shows that, after re-scaling by n, that $nH_n(\theta)$ converges uniformly to the negative of

$$\tilde{\Gamma}(\theta) = \mathbb{E}\left[\exp(\alpha + Z_{12}'\beta)\begin{pmatrix} 1 & Z_{12}' \\ Z_{12} & Z_{12}Z_{12}' \end{pmatrix}\right]. \tag{21}$$

An intuition for why $H_n(\theta)$ 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 $NM = O(n^2)$, but rather $O(n)$, an order of magnitude lower. Note that, under part (iii) of Assumption 3, the matrix $\tilde{\Gamma}_0 \overset{def}{\equiv} \tilde{\Gamma}(\theta_0)$ is of full rank. This fact, in conjunction with Lemma 2 (stated in Appendix A), gives the linear approximation

$$\sqrt{n}\left(\hat{\theta}_n-\theta_n\right)=-\tilde{\Gamma}_0^{-1}\times n^{3/2}S_n\left(\theta_0\right)+o_p\left(1\right).$$

To derive the limit distribution of $\sqrt{n}\left(\hat{\theta}_n-\theta_n\right)$ I show that the distribution $n^{3/2}S_n\left(\theta_0\right)$ 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 $S_n\left(\theta_0\right)$ is $n^{3/2}$, like the need to rescale the Hessian, is non-standard. The need to "blow up" $S_n\left(\theta_0\right)$ at a faster than $\sqrt{n}$ rate is a consequence of the fact that the summands in $S_n\left(\theta_0\right)$ are $O_p\left(n^{-1}\right)$. A second complication is that, for any fixed n, $S_n\left(\theta_0\right)$ is not mean zero. This bias reflects the discrepancy between the finite network pseudo composite log-likelihood criterion and the limiting population problem described by Lemma 1 above.

A detailed proof of Theorem 2, stated below, is provided in Appendix B. Here I outline the main arguments, which begin with the following four part decomposition of the score vector

$$S_n\left(\theta\right)=U_{1n}\left(\theta\right)+U_{2n}\left(\theta\right)+V_n\left(\theta\right)+b_n\left(\theta\right) \tag{22}$$

where

$$\begin{aligned}
U_{1n}\left(\theta\right)=&\frac{1}{N}\sum_{i=1}^{N}\left[\bar{s}_{1i,n}^{c}\left(\theta\right)-b_n\left(\theta\right)\right]+\frac{1}{M}\sum_{j=1}^{M}\left[\bar{s}_{1j,n}^{p}\left(\theta\right)-b_n\left(\theta\right)\right] && (23)\\
U_{2n}\left(\theta\right)=&\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\{\bar{s}_{ij,n}\left(\theta\right)-b_n\left(\theta\right)-\left[\bar{s}_{1i,n}^{c}\left(\theta\right)-b_n\left(\theta\right)\right]\right. && (24)\\
&\left.-\left[\bar{s}_{1j,n}^{p}\left(\theta\right)-b_n\left(\theta\right)\right]\right\}\\
V_n\left(\theta\right)=&\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\{s_{ij,n}\left(\theta\right)-\bar{s}_{ij,n}\left(\theta\right)\right\} && (25)\\
b_n\left(\theta\right)=&\ \mathbb{E}\left[S_n\left(\theta\right)\right] && (26)
\end{aligned}$$

with $\bar{s}_{ij,n}\left(\theta\right)=\bar{s}_n\left(W_i,X_j,A_i,B_j;\theta\right)$ with $\bar{s}_n\left(w,x,a,b;\theta\right)=\mathbb{E}\left[\left.s_{ij,n}\left(\theta\right)\right|W_i=w,X_j=x,A_i=a,B_j=b\right]$ and

$$\begin{aligned}
\bar{s}_{1i,n}^{c}\left(\theta\right)=&\bar{s}_{1,n}^{c}\left(W_i,A_i;\theta\right)\\
\bar{s}_{1j,n}^{p}\left(\theta\right)=&\bar{s}_{1,n}^{p}\left(X_j,B_j;\theta\right)
\end{aligned}$$

with $\bar{s}_{1,n}^{c}\left(w,a;\theta\right)=\mathbb{E}\left[\bar{s}_n\left(w,X_j,a,B_j;\theta\right)\right]$ and $\bar{s}_{1,n}^{p}\left(x,b;\theta\right)=\mathbb{E}\left[\bar{s}_n\left(W_i,x,A_i,b;\theta\right)\right]$.

A variant of decomposition (22) also features in Graham (2020a), Menzel (2021) and Graham et al. (2022). It can be derived by first projecting $S_n\left(\theta\right)$ on to $\mathbf{A}=\left[A_i\right]_{1\leq i\leq N}$,

$\mathbf{W} = [W_i]_{1\leq i\leq N}$, $\mathbf{B} = [B_j]_{1\leq j\leq M}$, and $\mathbf{X} = [X_i]_{1\leq j\leq N}$ as follows:

$$\begin{aligned} S_n(\theta) &= \mathbb{E}\left[\left. S_n(\theta)\right| \mathbf{W}, \mathbf{X}, \mathbf{A}, \mathbf{B}\right] + \left\{S_n(\theta) - \mathbb{E}\left[\left. S_n(\theta)\right| \mathbf{W}, \mathbf{X}, \mathbf{A}, \mathbf{B}\right]\right\} \\ &= \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\bar{s}_{ij,n}(\theta) + \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\{s_{ij,n}(\theta) - \bar{s}_{ij,n}(\theta)\right\}. \end{aligned} \tag{27}$$

Next observe that $\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\bar{s}_{ij,n}(\theta)$ is a two sample U-Statistic, albeit one defined partially in terms of the latent variables A_i and B_j. Equation (23) 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 (24) is the usual Hajek Projection error term.

The final term in (22), $b_n(\theta)$, arises because – for any fixed n – $b_n(\theta_0) = \mathbb{E}[S_n(\theta_0)]$ is not mean zero. Instead we have, after some manipulation, that

$$\begin{aligned} b_n(\theta_0) =& \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\mathbb{E}\left[\left(Y_{ij} - e_{ij,n}(\theta_0)\right)R_{ij}\right] \\ =& \frac{1}{n}\mathbb{E}\left[\left(\lambda_0(W_1, X_2) - \exp\left(R_{12}'\theta_0\right)\right)R_{12}\right] + \frac{1}{n}\mathbb{E}\left[\left(ng_n(W_1, X_2) - \lambda_0(W_1, X_2)\right)R_{12}\right] \\ &+ \frac{1}{n}\mathbb{E}\left[\left(\exp\left(R_{12}'\theta_0\right) - ne_{12,n}(\theta_0)\right)R_{12}\right] \\ =& \frac{1}{n}\mathbb{E}\left[\left(ng_n(W_1, X_2) - \lambda_0(W_1, X_2)\right)R_{12}\right] \\ &+ \frac{1}{n}\mathbb{E}\left[\left(\exp\left(R_{12}'\theta_0\right)\left[1 - \frac{1}{1 + \frac{1}{n}\exp\left(R_{12}'\theta_0\right)}\right]\right)R_{12}\right] \end{aligned} \tag{28}$$

which, by Assumption 2, is $o\left(n^{-2}\right)$.[3]

Define $\phi_n \overset{def}{\equiv} M/n$, $\bar{s}_{1i,n}^{c} \overset{def}{\equiv} \bar{s}_{1i,n}^{c}(\theta_0)$, $\bar{s}_{1j,n}^{p} \overset{def}{\equiv} \bar{s}_{1j,n}^{p}(\theta_0)$ and also $\bar{s}_{ij,n} \overset{def}{\equiv} \bar{s}_{ij,n}(\theta_0)$. Similarly let $S_n = S_n(\theta_0)$ and so on. Applying the variance operator to S_n yields:

$$\begin{aligned} \mathbb{V}(S_n) =& \mathbb{V}(U_{1n}) + \mathbb{V}(U_{2n}) + \mathbb{V}(V_n) \qquad (29) \\ =& \frac{\Sigma_{1n}^{c}}{N} + \frac{\Sigma_{1n}^{p}}{M} + \frac{1}{NM}\left[\Sigma_{2n} - \Sigma_{1n}^{c} - \Sigma_{1n}^{p}\right] + \frac{\Sigma_{3n}}{NM} \end{aligned}$$

where

$$\begin{aligned} \Sigma_{1n}^{c} &= \mathbb{V}\left(\bar{s}_{1i,n}^{c}\right) \quad \Sigma_{1n}^{p} = \mathbb{V}\left(\bar{s}_{1j,n}^{p}\right) \\ \Sigma_{2n} &= \mathbb{V}\left(\bar{s}_{ij,n}\right) = \mathbb{V}\left(\mathbb{E}\left[\left. s_{ij,n}\right| W_i, X_j, A_i, B_j\right]\right) \\ \Sigma_{3n} &= \mathbb{E}\left[\mathbb{V}\left(\left. s_{ij,n}\right| W_i, X_j, A_i, B_j\right)\right]. \end{aligned} \tag{30}$$

In the *dense* case Σ_{1n}^{c}, Σ_{1n}^{p}, Σ_{2n} and Σ_{3n} are all constant in n; hence the asymptotic

[3] While not developed in the theory which follows, equation (28) suggests that part of the bias in $S_n(\theta_0)$ is estimable (namely the second term to the right of the last equality in (28)). This, in turn, suggests that it might be fruitful to explore methods of bias reduction.

properties of S_n coincide with those of U_{1n} (the bias term is also zero in this case). Since U_{1n} is a sum of independent random variables a standard argument gives

$$n^{1/2}S_n \xrightarrow{D} \mathcal{N}\left(0, \frac{\Sigma_1^c}{1-\phi} + \frac{\Sigma_1^p}{\phi}\right) \tag{31}$$

as long as Σ_1^c and/or Σ_1^p are non-zero (see Graham (2020a) or Davezies et al. (2021)). In the degenerate – but still dense – case, as emphasized by Menzel (2021), the limiting behavior of $n^{1/2}S_n$ may be degenerate and, after appropriate rescaling, may also be non-Gaussian.

Under the sparse network asymptotics considered here, the orders of Σ_{1n}^c, Σ_{1n}^p, Σ_{2n} and Σ_{3n} vary with n. This affects the order of the four variance terms in (29) and, consequently, which components of S_n contribute to its asymptotic properties. In Appendix B I show the order of the four terms in (29) are, respectively,

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

Since Σ_1^c and Σ_1^p are both $O\left(\rho_n^2\right) = O\left(n^{-2}\right)$ 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 (29) by $n^{3/2}$ therefore gives

$$\mathbb{V}\left(n^{3/2}S_n\right) = \frac{\tilde{\Sigma}_1^c}{1-\phi} + \frac{\tilde{\Sigma}_1^p}{\phi} + \frac{\tilde{\Sigma}_3}{\phi\left(1-\phi\right)} + O\left(n^{-1}\right) \tag{32}$$

where I also use the fact that $\Sigma_{2n} = O\left(n^{-2}\right)$. We also have, from Assumption 2, that $\mathbb{E}\left[n^{3/2}S_n\right]^2 = \mathbb{E}\left[n^{3/2}b_n\right]^2 = o\left(n^{-1}\right)$.

Under sparse network asymptotics both U_{1n} and V_n matter. In Appendix B I further show that $U_{1n} + V_n$ is a martingale difference sequence (MDS) to which a martingale CLT can be applied; Theorem 2 then follows.

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

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \tilde{\Gamma}_0^{-1}\left[\frac{\tilde{\Sigma}_1^c}{1-\phi} + \frac{\tilde{\Sigma}_1^p}{\phi} + \frac{\tilde{\Sigma}_3}{\phi\left(1-\phi\right)}\right]\tilde{\Gamma}_0^{-1}\right)$$

as $n \to \infty$.

Proof. See Appendix B. □

Theorem 2 indicates that under sparse network asymptotics there are additional sources of sampling variation in $\sqrt{n}\left(\hat{\theta}-\theta_0\right)$ 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 2 remains valid even under degeneracy of the graphon, $h_n\left(W_i, X_j, A_i, B_j, V_{ij}\right)$. 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 2 nevertheless remains valid (condition (iv) of 3 ensures that $\tilde{\Sigma}_3$ will be positive definite). In contrast, under dense network asymptotics, degeneracy – as elegantly shown by Menzel (2021) – 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 case, U_{2n} is always negligible relative to V_n. Furthermore V_n is – after suitable scaling – approximately a Gaussian random variable.

Limit theory under correct specification

Theorem 2 holds for a general nonparametric regression function $g_n\left(w, x\right)$, with θ_0 a vector of pseudo-true parameters as defined by equation (18) above. If, in fact, $g_n\left(w, x\right) = e_n\left(w, x; \theta_0\right)$ for all $\left(w, x\right) \in \mathbb{W} \times \mathbb{X}$, then calculations in the Appendix B indicate the asymptotic variance simplifies to

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \tilde{\Gamma}_0^{-1}\left[\frac{\tilde{\Sigma}_1^c}{1-\phi}+\frac{\tilde{\Sigma}_1^p}{\phi}\right]\tilde{\Gamma}_0^{-1}+\frac{\tilde{\Gamma}_0^{-1}}{\phi\left(1-\phi\right)}\right),$$

which follows from an information matrix type equality result of $n\mathbb{V}\left(s_{ij,n}\right) \rightarrow \tilde{\Gamma}_0$ as $n \rightarrow \infty$.

Relationship with rare events analysis using 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.[4] 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 2 above.[5] To see this observe that if the graphon is constant in A_i

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

[5] In fact, Theorem 2 is a bit more general even in the special case of no dyadic dependence as it also accommodate misspecification of the the regression function.

and B_j, then $\bar{s}_{ij,n}$ 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 $\tilde{\Sigma}_1^c = \tilde{\Sigma}_1^p = 0$. Under these conditions, also maintaining correct specification, Theorem 1 specializes to

$$\sqrt{n}\left(\hat{\theta} - \theta_n\right) \xrightarrow{D} \mathcal{N}\left(0, \frac{\tilde{\Gamma}_0^{-1}}{\phi(1-\phi)}\right),$$

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 2 in Appendix A). In the network setting dyadic dependence additionally reduces the asymptotic precision with which θ_0 may be estimated (cf., Graham et al., 2022). If a researcher is working with a sparse network and concerned about dyadic dependence, then she should base inference on Theorem 2. 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 2 specializes to the "rare events" result of Wang (2020) in that case).

3 Simulation experiments

In this section I report the results of a small set of simulation experiments. An annotated Python Jupyter Notebook with replication code is available in the Supplemental Materials. The Monte Carlo experiments utilize the 'bilogit' estimation command included in the Python 'netrics' package; available on GitHub (https://github.com/bryangraham/netrics). The goal of these experiments is to assess the finite sample quality of the sparse network asymptotic approximations developed above in a stylized setting. The question of precisely how to best conduct inference when analyzing sparse networks (e.g., assessing the relative merits of different methods of variance estimation) is largely open and not directly addressed (see Chiang et al. (2022b)).

For the Monte Carlo experiments I set the graphon, $h_n(W_i, X_j, A_i, B_j, V_{ij})$, equal to

$$Y_{ij} = 1\left(\alpha + z(W_i, X_j)'\beta + \ln(A_i) + \ln(B_j) - \ln(n) \geq V_{ij}\right)$$

with V_{ij} a standard exponential random variable. Averaging over V_{ij} yields

$$\mathbb{E}_n\left[Y_{ij} | W_i, X_j, A_i, B_j\right] = \frac{1}{n}\exp\left(\alpha + z(W_i, X_j)'\beta\right) A_i B_j.$$

[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 variance expressions. Making these adjustments the results coincide.

I set $\{A_i\}_{i=1}^{N}$ and $\{B_j\}_{j=1}^{M}$ to be iid log-normal sequences of random variables with $\mu = -1/12$ and $\sigma = 1/\sqrt{6}$. This implies that both A_i and B_j are mean one and, furthermore, that the variance of $\ln(A_i) + \ln(B_j)$ is one third that of V_{ij}. This generates meaningful, but not overpowering, cross dyad dependence. Under these assumptions the regression function equals

$$g_n(w,x) = \frac{1}{n}\exp\left(\alpha + z(W_i, X_j)'\beta\right).$$

Finally I set $z(W_i, X_j) = \begin{pmatrix} W_i & X_j & W_iX_j \end{pmatrix}'$ with $\{W_i\}_{i=1}^{N}$ iid Bernoulli with a success probability $\pi_w = 1/\sqrt{3}$ and $\{X_j\}_{j=1}^{M}$ iid Bernoulli with a success probability $\pi_x = 1/\sqrt{3}$. This implies that one third of dyads are of the $W_i = X_j = 1$ type.

I simulate data for five sample sizes: $n = 64$, 144, 256, 576 and 1024 with $N = M$ in all cases. I set $\alpha = \ln(64 \times 0.04)$, $\beta_w = \beta_x = 0$ and $\beta_{wx} = \ln 4 \approx 1.3863$. This implies that $\rho_n = 0.08$, 0.036, 0.020, 0.009 and 0.005 across the five designs. Note that θ_0 is fixed across these designs, but the triangular array structure of the DGP induces a decline in density with n. For each design I perform $5,000$ Monte Carlo replications.

The design is a stylized version of how a researcher might analyze data from a simple consumer-product promotion experiment. Let A_i be consumer-specific heterogeneity, B_j product quality heterogeneity, $W_i = 1$ if consumer i was randomly invited to participate in a 'sale' and zero otherwise and $X_j = 1$ if product j was randomly determined to be 'sale eligible' and zero otherwise. The treatment effect of being invited to participate in the sale increases the purchases probability for sale eligible items by a factor of four ($\beta_{wx} = \ln 4$); there is no spillover effect onto non-eligible items ($\beta_w = 0$). Likewise there is no direct effect of an item being 'sale eligible' on the probability of making a purchase ($\beta_x = 0$). In what follows I focus on estimation of, and inference on, the interaction coefficient β_{wx}.

In the experiments, the logit approximation does not coincide with the population regression function for any fixed n, however the approximation error declines as $n \to \infty$. Therefore the pseudo composite maximum likelihood estimates of $\hat{\theta}$ are consistent for their population analogs. However, we would expect to observe noticeable bias in small samples. This is shown in the first two rows of Table 1: for smaller samples mean and median bias are modestly large relative to the standard deviation of $\hat{\beta}_{wx}$ across the $5,000$ Monte Carlo replications (row 3). As predicted, this bias declines with n.

The theoretical rate-of-convergence results outlined above suggest that the standard deviation of $\hat{\beta}_{wx}$ in design 2 should be two thirds of that in design 1. In practice we have that $\frac{0.4221}{0.7039} \approx 0.60 \approx \frac{\frac{1}{\sqrt{144}}}{\frac{1}{\sqrt{64}}} = \frac{2}{3}$, which is close. That in design 3 should be three quarters of that in design 2 (actual: $\frac{0.2968}{0.4221} \approx 0.70 \approx \frac{\frac{1}{\sqrt{256}}}{\frac{1}{\sqrt{144}}} = \frac{3}{4}$); design 4 two thirds of that in design 3 (actual: $\frac{0.1972}{0.2968} \approx 0.66 \approx \frac{\frac{1}{\sqrt{576}}}{\frac{1}{\sqrt{256}}} = \frac{2}{3}$); and design 5 three quarters of that in design 4 (actual:

Table 1: Monte Carlo Results, β_{wx}

	(1)	(2)	(3)	(4)	(5)
	$n=64$ $\rho_n=0.080$	$n=144$ $\rho_n=0.036$	$n=256$ $\rho_n=0.020$	$n=576$ $\rho_n=0.009$	$n=1,024$ $\rho_n=0.005$
Mean Bias	0.1209	0.0615	0.0396	0.0171	0.0119
Median Bias	0.1632	0.0635	0.0406	0.0149	0.0127
Std. Dev.	0.7039	0.4221	0.2968	0.1972	0.1516
Mean S.E. - Sparse	0.6779	0.4638	0.3445	0.2340	0.1783
Coverage (95% CI) - 'Sparse'	0.8754	0.9286	0.9442	0.9496	0.9434
Coverage (95% CI) - 'Dense'	0.3468	0.3620	0.3506	0.3208	0.2922

NOTES: Results based on 5,000 replications of the data generating process described in the text. The Monte Carlo standard deviation of the point estimates (row 3) is a robust measure (the difference between 95th and 5th percentiles of the estimated coefficient's Monte Carlo distribution divided by the corresponding quantile differences of a standard normal variate). The standard deviation of the simulation error on the coverage estimates is $\sqrt{\alpha(1-\alpha)/5000} \approx 0.003$ for $\alpha = 0.05$. See the text for additional information.

$\frac{0.1516}{0.1972} \approx 0.77 \approx \frac{\frac{1}{\sqrt{1024}}}{\frac{1}{\sqrt{576}}} = \frac{3}{4}$). Overall the Monte Carlo rate-of-convergence estimates track theoretical predictions well.

The final two rows of Table 1 report the actual coverage of two different nominal 95 percent Wald-based confidence intervals. The sparse intervals are Wald ones which use a variance estimate suggested by Cameron and Miller (2014). This estimate can also be thought of as a bias-corrected version of the usual jackknife variance estimate (see Efron and Stein (1981); Cattaneo et al. (2014); Graham (2020b)). A description of the variance estimate, which is a direct analog estimate of the asymptotic variance presented in Theorem 2, is given in Appendix C. The 'dense' intervals are based upon the analog estimate of the dense asymptotic variance given by Graham (2020a) (see also Appendix C).

In the designs with smaller samples, the sparse confidence intervals undercover slightly, but once n is large enough such that bias is negligible, their actual and nominal coverage coincide. As suggested by the theory, the actual coverage of the dense asymptotic intervals are well below nominal levels in all designs.

Table 2 summarizes the sampling behavior of the components of

$$n^{3/2}S_n(\theta_0) = n^{3/2}U_{1n}(\theta_0) + n^{3/2}U_{2n}(\theta_0) + n^{3/2}V_n(\theta_0) + n^{3/2}b_n(\theta_0).$$

For each Monte Carlo draw I construct each component of $n^{3/2}S_n(\theta_0)$ analytically (see the Python Jupyter Notebook in the Supplemental Materials). The variance of these components is then estimated by their sampling variance across the 5,000 Monte Carlo draws (i.e., by Monte Carlo integration). Table 2 reports the mean and standard deviation of each of the components $n^{3/2}S_n(\theta_0)$ in the the $n = 256$ and $n = 1,024$ designs; specifically the elements corresponding to the interaction coefficient β_{wx}.

Table 2 indicates that, for the designs considered here, $n^{3/2}U_{1n}(\theta_0)$ and $n^{3/2}V_n(\theta_0)$ are of equal order, while – as asserted by the theoretical analysis – $n^{3/2}U_{2n}(\theta_0)$ is of lower order. The closeness of the Monte Carlo standard deviations across the two samples also indicates that $n^{3/2}$ is the correct variance stabilizing rate. The Monte Carlo estimate of the bias in $n^{3/2}S_{1n}(\theta_0)$ also closely tracks its theoretical counterpart. Most importantly, the normal approximation to $n^{3/2}[U_{1n}(\theta_0) + V_n(\theta_0)]$, which underlies Theorem 2, appears to be quite accurate. Normalized by its standard deviation, the tail frequencies of $n^{3/2}[U_{1n}(\theta_0) + V_n(\theta_0)]$ are close to those of a standard normal random variable (especially for the larger sample size).

Table 2: Accuracy Sparse Network Asymptotics for $\hat{\beta}_{wx}$

	(1)	(2)	(3)	(4)	(5)	(6)
	$n^{3/2}S_n(\theta_0)$	$n^{3/2}U_{1n}(\theta_0)$	$n^{3/2}U_{2n}(\theta_0)$	$n^{3/2}V_n(\theta_0)$	$n^{3/2}[U_{1n}(\theta_0)+V_n(\theta_0)]$	$n^{3/2}b_n(\theta_0)$
Panel A: $n=256$						
Mean	2.164	0.0446	-0.0045	0.0227	0.0672	2.101
Std. Dev.	5.2165	3.8460	0.3196	3.6122	5.2090	-
$\Pr(T \geq 1.645)$	0.0578	0.0546	0.0422	0.0542	0.0576	-
$\Pr(T \leq 1.645)$	0.0400	0.0432	0.0502	0.0472	0.0412	-
$\Pr(T \geq 1.96)$	0.0324	0.0290	0.0282	0.0308	0.0304	-
$\Pr(T \leq 1.96)$	0.0154	0.0184	0.0360	0.0246	0.0166	-
Panel B: $n=1,024$						
Mean	1.116	0.0399	-0.0025	-0.0019	0.0380	1.081
Std. Dev.	5.3091	3.8169	0.1555	3.7162	5.3123	-
$\Pr(T \geq 1.645)$	0.0502	0.0526	0.0432	0.0508	0.0504	-
$\Pr(T \leq 1.645)$	0.0490	0.0490	0.0522	0.0476	0.0490	-
$\Pr(T \geq 1.96)$	0.0276	0.0244	0.0266	0.0236	0.0268	-
$\Pr(T \leq 1.96)$	0.0236	0.0234	0.0362	0.0234	0.0244	-

NOTES: Results based on 5,000 replications of the data generating process described in the text. The forms of $S_n(\theta_0)$, $U_{1n}(\theta_0)$, $U_{2n}(\theta_0)$, $V_n(\theta_0)$ and , $b_n(\theta_0)$ are based on pencil and paper calculations and the details of the simulated data generating process (see the Python Jupyter Notebook in the Supplemental Materials for details).

Appendix

The appendix includes proofs of the formal results stated in the main text as well as statements and proofs of supplemental results. All notation is as established in the main text unless stated otherwise. Equation numbering continues in sequence with that established in the main text.

A Identification and consistency

Proof of Lemma 1 (Representation result for θ_0)

To show Lemma 1 is is convenient to observe that $L_0(\theta) = \mathbb{E}\left[\lambda_0(X_i, W_j) R'_{ij}\theta\right] - \mathbb{E}\left[\exp\left(R'_{ij}\theta\right)\right]$. To see this equality note that

$$\begin{aligned}
L_0(\theta) =& \mathbb{E}\left[\lambda_0(X_i, W_j) R'_{ij}\theta\right] - \mathbb{E}\left[\exp\left(R'_{ij}\theta\right)\right] \\
=& \mathbb{E}_0\left[V_{ij}\ln\left(\frac{\exp\left(R'_{ij}\theta\right)}{\lambda_0(X_i, W_j)}\right)\right] + \mathbb{E}\left[\lambda_0(X_i, W_j)\right] \\
&- \mathbb{E}\left[\exp\left(R'_{ij}\theta\right)\right] + \mathbb{E}\left[V_{ij}\ln\left(\lambda_0(X_i, W_j)\right)\right] - \mathbb{E}\left[\lambda_0(X_i, W_j)\right] \\
=& \mathbb{E}_0\left[\ln\left\{\frac{f\left(V_{ij}\middle| W_i, X_j; \theta\right)}{f_0\left(V_{ij}\middle| W_i, X_j\right)}\right\}\right] + \mathbb{E}\left[\lambda_0(X_i, W_j)\ln\left(\lambda_0(X_i, W_j)\right)\right] - \mathbb{E}\left[\lambda_0(X_i, W_j)\right] \\
=& -\mathbb{D}_{KL}\left(F_0 \| F_\theta\right) + \mathbb{S}(F_0).
\end{aligned}$$

To show uniform convergence of $nL^*_n(\theta)$ to $L_0(\theta)$ write $L^*_n(\theta) = L_n(\theta) + \delta_n$ as the average

$$L^*_n(\theta) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} l^*_{ij,n}(\theta) \tag{33}$$

with kernel, recalling that $R_{ij} = \left(1, Z'_{ij}\right)'$,

$$l^*_{ij,n}(\theta) = Y_{ij}R'_{ij}\theta - \ln\left(1 + \frac{1}{n}\exp\left(R'_{ij}\theta\right)\right). \tag{34}$$

The form of (34) follows from the fact that, manipulating (17) in the main text

$$\begin{aligned}
l^*_{ij,n}(\theta) &= (2Y_{ij} - 1)\left(R'_{ij}\theta - \ln n\right) - \ln\left(1 + \exp\left((2Y_{ij} - 1)\left[R'_{ij}\theta - \ln n\right]\right)\right) + Y_{ij}\ln n \\
&= Y_{ij}\left(R'_{ij}\theta - \ln n\right) - \ln\left(1 + \exp\left(R'_{ij}\theta - \ln n\right)\right) + Y_{ij}\ln n \\
&= Y_{ij}R'_{ij}\theta - \ln\left(1 + \frac{1}{n}\exp\left(R'_{ij}\theta\right)\right).
\end{aligned}$$

First I show that

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\left[nl^{*}_{ij,n}\left(\theta\right)\right]=&L_0\left(\theta\right) && (35)\\ =&\mathbb{E}\left[\lambda_0\left(X_i,W_j\right)R'_{ij}\theta\right]-\mathbb{E}\left[\exp\left(R'_{ij}\theta\right)\right]\end{aligned}$$

pointwise in $\theta\in\Theta$. By part (ii) of Assumption 1, part (ii) of Assumption 2 and parts (i) and (ii) of Assumption 3 we have the dominating function

$$\left|ng_n\left(w,x\right)r'\theta f_W\left(w\right)f_X\left(x\right)\right|\leq k\left(w,x\right)\times\sup_{r\in(1,\mathbb{Z}),\theta\in\Theta}\left|r'\theta\right|\times f_W\left(w\right)f_X\left(x\right)<\infty.$$

Part (i) of Assumption 2 implies that $ng_n\left(w,x\right)r'\theta$ converges pointwise to $\lambda_0\left(x,w\right)r'\theta$. The Dominated Convergence Theorem then yields

$$\lim_{n\to\infty}\mathbb{E}\left[ng_n\left(W_i,X_j\right)R'_{ij}\theta\right]\to\mathbb{E}\left[\lambda_0\left(X_i,W_j\right)R'_{ij}\theta\right]. \qquad (36)$$

Next, the exponential function characterization $\exp x=\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$ and continuity of the $\ln\left(\cdot\right)$ function yield the limit

$$\lim_{n\to\infty}\ln\left(1+\frac{1}{n}\exp\left(r'\theta\right)\right)^n=\exp\left(r'\theta\right).$$

To verify the stronger equality

$$\lim_{n\to\infty}\mathbb{E}\left[\ln\left(1+\frac{1}{n}\exp\left(R'_{ij}\theta\right)\right)^n\right]=\mathbb{E}\left[\exp\left(R'_{ij}\theta\right)\right] \qquad (37)$$

it suffices to show that

$$\sup_{w\in\mathbb{W},x\in\mathbb{X}}\left|\ln\left(1+\frac{1}{n}\exp\left(r'\theta\right)\right)^n f_W\left(w\right)f_X\left(x\right)-\exp\left(r'\theta\right)f_W\left(w\right)f_X\left(x\right)\right|\to 0$$

as $n\to\infty$. Under part (ii) of Assumption 1 and parts (i) and (ii) of Assumption 3 this follows if

$$\sup_{x\in[\underline{x},\bar{x}]}\left|\ln\left(1+\frac{1}{n}\exp\left(x\right)\right)^n-\exp\left(x\right)\right|\to 0 \qquad (38)$$

with $[\underline{x},\bar{x}]$ the support of possible values the index $r'\theta$. Let $b_n\left(x\right)=\ln\left(1+\frac{1}{n}\exp\left(x\right)\right)^n-\exp\left(x\right)$; since $b'_n\left(x\right)=\exp\left(x\right)\left[\frac{1}{1+\frac{1}{n}\exp(x)}-1\right]<0$ on $x\in[\underline{x},\bar{x}]$ condition (38) holds since both $b_n\left(\underline{x}\right)$ and $b_n\left(\bar{x}\right)$ converge to zero. Condition (35) follows directly from (36) and (37).

Second, since (35) also gives $\lim_{n\to\infty}\mathbb{E}\left[nL^{*}_n\left(\theta\right)\right]=L_0\left(\theta\right)$, the mean square error decomposition

$$\mathbb{E}\left[\left(nL^{*}_n\left(\theta\right)-L_0\left(\theta\right)\right)^2\right]=\left(\mathbb{E}\left[nL^{*}_n\left(\theta\right)\right]-L_0\left(\theta\right)\right)^2+\mathbb{V}\left(nL^{*}_n\left(\theta\right)\right)$$

implies convergence of $nL_n^*(\theta)$ to $L_0(\theta)$ in mean square if $\mathbb{V}(nL_n^*(\theta)) \to 0$ as $n \to \infty$. This follows under Assumptions 2 and 3 since

$$\begin{aligned}\mathbb{V}(nL_n^*(\theta)) &= \frac{n^2}{N}O\left(\rho_n^2\right) + \frac{n^2}{M}O\left(\rho_n^2\right) + \frac{n^2}{NM}O(\rho_n) \\ &= O\left(n^{-1}\right) + O\left(n^{-1}\right) + O\left(n^{-1}\right).\end{aligned}$$

By concavity of $L_n^*(\theta)$ in θ, this convergence is uniform in $\theta \in \Theta$. Lemma (1) follows directly with some algebra.

Proof of Theorem 1: consistency of $\hat{\theta}$ for θ_0

The result follows by verifying conditions (i) to (iv) of Theorem 2.1 in Newey and McFadden (1994, p. 2121). Part (ii) of follows from Assumption 3, part (iii) follows by inspection, part (iv) was shown in Lemma 1. Part (i) requires demonstrating uniqueness of the solution

$$\theta_0 = \arg\max_{\theta\in\Theta} L_0(\theta). \tag{39}$$

For this to hold it suffices to verify global concavity of $L_0(\theta)$ in θ. Direct calculation yields first and second order conditions equal to

$$\begin{aligned}\mathbb{E}\left[\frac{\partial L_0(\theta)}{\partial\theta}\right] &= \mathbb{E}\left[\left(\lambda_0(X_i, W_j) - \exp\left(R_{ij}'\theta\right)\right)R_{ij}\right] \\ \mathbb{E}\left[\frac{\partial^2 L_0(\theta)}{\partial\theta\partial\theta'}\right] &= -\mathbb{E}\left[\exp\left(R_{ij}'\theta\right)R_{ij}R_{ij}'\right] \overset{def}{\equiv} \Gamma(\theta).\end{aligned} \tag{40}$$

Under Assumption 3 the matrix $\Gamma(\theta)$ is negative definite for all $\theta \in \Theta$; therefore $L_0(\theta)$ is globally concave in $\theta \in \Theta$ with unique maximum θ_0.

Hessian convergence

Note that for $e_n(v) = \exp(v - \ln n)/[1 + \exp(v - \ln n)]$, we have that $e_n'(v) = e_n(v)[1 - e_n(v)]$ and $e_n''(v) = e_n(v)[1 - e_n(v)][1 - 2e_n(v)]$. Further let $e_{ij,n}(\theta) = e_n\left(R_{ij}'\theta\right)$; with this notation we can write the first three derivatives of the kernel function of the composite log-likelihood with respect θ as

$$s_{ij,n}(n) = (Y_{ij} - e_{ij,n}(\theta))R_{ij} \tag{41}$$

$$\frac{\partial s_{ij,n}(\theta)}{\partial\theta'} = -e_{ij,n}(\theta)[1 - e_{ij,n}(\theta)]R_{ij}R_{ij}' \tag{42}$$

$$\frac{\partial}{\partial\theta'}\left\{\frac{\partial s_{ij,n}(\theta)}{\partial\theta_p}\right\} = -e_{ij,n}(\theta)[1 - e_{ij,n}(\theta)][1 - 2e_{ij,n}(\theta)]R_{ij}R_{ij}'R_{p,ij} \tag{43}$$

with (43) holding for for $p = 1, \ldots, \dim(\theta)$.

Let $\mathbf{t} = (\theta - \theta_0)$ and note that $\mathbf{t} \in \mathbb{T}$ with $\mathbb{T}$ compact by Assumption 3. Associated with any $\mathbf{t} \in \mathbb{T}$ is a $\theta \in \Theta$. With these preliminaries we can show that $nH_n(\theta)$ converges uniformly to $\tilde{\Gamma}(\theta)$, as defined in equation (21) of the main text.

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

$$\sup_{\theta \in \Theta} \left\| nH_n(n) - \tilde{\Gamma}(\theta) \right\| \xrightarrow{p} 0.$$

Proof. Let $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{M} A_{ij}^2}$ denote the $\ell_{2,1}$ matrix norm. Note that $\theta = \theta_0 + \mathbf{t}$ and hence that $H_n(\theta_0 + \mathbf{t}) = H_n(\theta)$. 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 $\mathbb{T}$,

$$\left\| H_n(\theta_0 + \mathbf{t}) - H_n(\theta_0 + \bar{\mathbf{t}}) \right\|_{2,1} \leq \sum_{p=1}^{\dim(\theta)} \left\| \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\partial}{\partial \theta'} \left\{ \frac{\partial s_{ij,n}(\theta_0 + \mathbf{t})}{\partial \theta_p} \right\} \right\|_F \left\| \mathbf{t} - \bar{\mathbf{t}} \right\|_2.$$

Since $\mathbb{E}\left[e_{ij,n}(\theta)\left[1 - e_{ij,n}(\theta)\right]\left[1 - 2e_{ij,n}(\theta)\right]\right] = O\left(n^{-1}\right)$ we have that, inspecting (43) above, for any $\mathbf{t} \in \mathbb{T}$,

$$\left\| \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\partial}{\partial \theta'} \left\{ \frac{\partial s_{ij,n}(\theta_0 + \mathbf{t})}{\partial \theta_p} \right\} \right\|_F = O_p\left(n^{-1}\right).$$

This gives $\left\| nH_n(\theta_0 + \mathbf{t}) - nH_n(\theta_0 + \bar{\mathbf{t}}) \right\|_{2,1} \leq O_p(1) \cdot \left\| \mathbf{t} - \bar{\mathbf{t}} \right\|_2$. Next, again recalling that $\theta_0 + \mathbf{t} = \theta$, we have that

$$\begin{aligned} H_n(\theta_0 + \mathbf{t}) &= -\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} e_{ij,n}(\theta)\left[1 - e_{ij,n}(\theta)\right] R_{ij} R_{ij}' \\ &= -\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{1}{n} \exp\left(R_{ij}'\theta\right) R_{ij} R_{ij}' + O_p\left(\frac{1}{n^2}\right), \end{aligned}$$

which gives, using a law of large numbers for U-Statistics, $nH_n(\theta) \xrightarrow{p} \Gamma(\theta)$ 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 2

To show Theorem 2 I first verify the rate-of-convergence analysis for S_n given in the main next. Next I show asymptotic normality of $U_{1n} + V_n$, after normalization. I then prove the main result.

Asymptotic variance of the score

To prove (29), the decomposition of the variance of the score given in the main text, and hence that

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

use the definitions given in (30) of the main text and observe that

$$\begin{aligned}\Sigma_{1n}^c =& \mathbb{E}\left[\left(Y_{12}-e_{12,n}\right)\left(Y_{13}-e_{13,n}\right)R_{12}R_{13}'\right] - b_n^2 \\ =& O\left(\rho_n^2\right) + o\left(n^{-4}\right)\end{aligned} \tag{44}$$

and also that

$$\begin{aligned}\Sigma_{1n}^p =& \mathbb{E}\left[\left(Y_{21}-e_{21,n}\right)\left(Y_{31}-e_{31,n}\right)R_{21}R_{31}'\right] - b_n^2. \\ =& O\left(\rho_n^2\right) + o\left(n^{-4}\right).\end{aligned} \tag{45}$$

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

$$\begin{aligned}\Sigma_{2n} =& \mathbb{E}\left[\mathbb{E}\left[\left(Y_{12}-e_{12,n}\right)R_{21}\right|W_1,X_2,A_1,B_2\right] \\ & \times\mathbb{E}\left[\left(Y_{12}-e_{12,n}\right)R_{21}\right|W_1,X_2,A_1,B_2\left]'\right] - b_n^2 \\ =& O\left(\rho_n^2\right) + o\left(n^{-4}\right)\end{aligned} \tag{46}$$

and further that

$$\begin{aligned}\Sigma_{3n} =& \mathbb{E}\left[\left\{s_{ij,n}-\bar{s}_{ij,n}\right\}\left\{s_{ij,n}-\bar{s}_{ij,n}\right\}'\right] \\ =& O\left(\rho_n\right)\end{aligned} \tag{47}$$

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

From Assumption 2 we have that $\rho_n = O\left(n^{-1}\right)$, hence (44) implies that $n^2\Sigma_{1n}^c = O\left(1\right)$, (45) that $n^2\Sigma_{1n}^p = O\left(1\right)$, and (47) that $n\Sigma_{3n} = O\left(1\right)$. This gives

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

and hence the form of the variance expression stated in the Theorem.

Variance simplification when $g_n(w, x)$ takes the logit form

Observe that $\mathbb{V}(s_{ij,n}) = \Sigma_{2n} + \Sigma_{3n}$. Therefore when $g_n(W_i, X_j) = e_n\left(\alpha_0 + Z_{ij}'\beta_0\right)$ we have that

$$\begin{aligned} n\mathbb{V}(s_{ij,n}) &= n\mathbb{E}\left[(Y_{ij} - e_{ij,n})^2 R_{ij}R_{ij}'\right] - nb_n^2 \\ &= n\mathbb{E}\left[e_{ij,n}(1 - e_{ij,n}) R_{ij}R_{ij}'\right] + o\left(n^{-3}\right) \\ &\to \tilde{\Gamma}_0, \end{aligned}$$

and hence the alternative limiting variance expression

$$\begin{aligned} \mathbb{V}\left(n^{3/2}S_n\right) &= \frac{n^2\Sigma_{1n}^c}{1-\phi_n} + \frac{n^2\Sigma_{1n}^p}{\phi_n} + \frac{n(\Sigma_{2n} + \Sigma_{3n})}{\phi_n(1-\phi_n)} + O\left(n^{-1}\right) \\ &\to \frac{\tilde{\Sigma}_1^c}{1-\phi} + \frac{\tilde{\Sigma}_1^p}{\phi} + \frac{\tilde{\Gamma}_0}{\phi(1-\phi)} \end{aligned}$$

as $n \to \infty$.

Triangular array setup

Observe that $U_{1n} + V_n = \sum_{t=1}^{T} Z_{nt}$,where the triangular array $\{Z_{nt}\}$ is defined as follows:

$$\begin{aligned} Z_{n1} &= \frac{1}{N}\left(\bar{s}_{11,n}^c - b_n\right) \\ &\vdots \\ Z_{nN} &= \frac{1}{N}\left(\bar{s}_{1N,n}^c - b_n\right) \\ Z_{nN+1} &= \frac{1}{M}\left(\bar{s}_{11,n}^p - b_n\right) \\ &\vdots \\ Z_{nN+M} &= \frac{1}{M}\left(\bar{s}_{1M,n}^p - b_n\right) \\ Z_{nN+M+1} &= \frac{1}{NM}\left(s_{11,n} - \bar{s}_{11,n}\right) \\ &\vdots \\ Z_{nN+M+NM} &= \frac{1}{NM}\left(s_{NM,n} - \bar{s}_{NM,n}\right), \end{aligned}$$

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

$$\mathbb{E}\left[\left. Z_{nt} \right| Z_{n1}^{t-1}\right] = 0,$$

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

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

To show asymptotic normality of $n^{3/2}S_n(\theta_0)$ I first show, recalling decomposition (22) in the main test, that, for a vector of constants c,

$$\left(c'\bar{\Delta}_n c\right)^{-1/2} c' S_n = \left(c'\bar{\Delta}_n c\right)^{-1/2} c' \left[U_{1n} + V_n\right] + o_p(1) \tag{48}$$

and subsequently that

$$\left(c'\bar{\Delta}_n c\right)^{-1/2} c' \left[U_{1n} + V_n\right] \xrightarrow{p} \mathcal{N}(0,1). \tag{49}$$

To show (48) observe that

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

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 (46) yields

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

as long as $\rho_n = O\left(n^{-\alpha}\right)$ for $\alpha > \frac{1}{2}$, as is maintained here. We also have that $\left(c'\bar{\Delta}_n c\right)^{-1/2} c' b_n = O_p\left(n^{1/2}\rho_n^{-1}\right) o\left(n^{-2}\right) = o_p(1)$. These two results imply assertion (48).

Central limit theorem

To show (49) I verify the conditions of Corollary 5.26 of Theorem 5.24 in White (2001); specifically the Lyapunov condition, for $r > 2$

$$\sum_{t=1}^{T(n)} \mathbb{E}\left[\left(\left|\frac{c'Z_{nt}}{\left(c'\bar{\Delta}_n c\right)^{1/2}}\right|\right)^r\right] = o\left(1\right) \tag{50}$$

and the stability condition

$$\sum_{t=1}^{T(n)} \frac{\left(c'Z_{Nt}\right)^2}{c'\bar{\Delta}_n c} \xrightarrow{p} 1. \tag{51}$$

I will show (50) for $r = 3$. Observe that

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{N}c'\left(\bar{s}^c_{1i,n} - b_n\right)\right)^3\right] =& O\left(\frac{\rho_n^3}{N^3}\right) \\
\mathbb{E}\left[\left(\frac{1}{M}c'\left(\bar{s}^p_{1j,n} - b_n\right)\right)^3\right] =& O\left(\frac{\rho_n^3}{M^3}\right) \\
\mathbb{E}\left[\left(\frac{1}{NM}c'\left(s_{11,n} - \bar{s}_{11,n}\right)\right)^3\right] =& O\left(\frac{\rho_n}{N^3M^3}\right)
\end{aligned}$$

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

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

as required.

To verify the stability condition (51) I re-write it as

$$\sum_{t=1}^{T(n)} \frac{1}{n\left(c'\bar{\Delta}_n c\right)} n\left\{\left(c'Z_{nt}\right)^2 - \mathbb{E}\left[\left(c'Z_{nt}\right)^2\right]\right\} \xrightarrow{p} 0 \tag{52}$$

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 (51) will hold if the numerator in (52) – $\sum_{t=1}^{T(n)} n\left\{\left(c'Z_{nt}\right)^2 - \mathbb{E}\left[\left(c'Z_{nt}\right)^2\right]\right\}$ – 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

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

and

$$\mathbb{E}\left[\left(c'Z_{nt}\right)^4\right]=\begin{cases}\frac{\mathbb{E}\left[\left(c'\bar{s}_{1n1}^{c}\right)^4\right]}{N^4}=O\left(\left[\frac{\lambda_n^c}{(1-\phi_n)\phi_n}\right]^4\frac{1}{n^8}\right), & t=1,\ldots,N\\ \frac{\mathbb{E}\left[\left(c'\bar{s}_{1n1}^{p}\right)^4\right]}{M^4}=O\left(\left[\frac{\lambda_n^c}{\phi_n^2}\right]^4\frac{1}{n^8}\right), & t=N+1,\ldots,N+M\\ \frac{\mathbb{E}\left[\left(c'\left(s_{n11}-\bar{s}_{n11}\right)\right)^4\right]}{N^4M^4}=O\left(\frac{\lambda_n^c}{\phi_n^5(1-\phi_n)^4}\frac{1}{n^9}\right), & t=N+M+1,\ldots,N+M+NM\end{cases}.$$

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:

$$T(n)^2n^2\left\{\mathbb{E}\left[\left(c'Z_{nt}\right)^4\right]-\left(\mathbb{E}\left[\left(c'Z_{nt}\right)^2\right]\right)^2\right\}=$$
$$\begin{cases}T(n)^2n^2\left[O(n^{-8})+O(n^{-8})\right]=O(n^{-2}), & t=1,\ldots,N\\ T(n)^2n^2\left[O(n^{-8})+O(n^{-8})\right]=O(n^{-2}), & t=N+1,\ldots,N+M\\ T(n)^2n^2\left[O(n^{-9})+O(n^{-10})\right]=O(n^{-3}), & t=N+M+1,\ldots,N+M+NM\end{cases}$$

Since the summands of the numerator in (52) are all mean zero with variances shrinking to zero as $n\to\infty$ condition (52) holds as required.

Next observe that

$$n^3\bar{\Delta}_n\to\frac{\tilde{\Sigma}_1^c}{1-\phi}+\frac{\tilde{\Sigma}_1^p}{\phi}+\frac{\tilde{\Sigma}_3}{\phi(1-\phi)}$$

as $n\to\infty$, such that, using (48) and the Cramér-Wold Theorem, $n^{3/2}S_n\overset{D}{\to}\mathcal{N}\left(0,\frac{\tilde{\Sigma}_1^c}{1-\phi}+\frac{\tilde{\Sigma}_1^p}{\phi}+\frac{\tilde{\Sigma}_3}{\phi(1-\phi)}\right)$. The result then follows from Lemma 2 and Slutsky's Theorem.

C Variance estimation

In this appendix I describe the variance estimators used in the Monte Carlo experiments reported in the main text. Graham (2020a) and Graham (2020b) both discuss variance estimation under dyadic dependence and provide references to the primary literature.

We have that $\Sigma_{1n}^{c} = \mathbb{C}_n\left(s_{ij,n}s_{ik,n}\right)$ for $j \neq i$. For each of the $i = 1, \ldots, N$ consumers there are $\binom{M}{2} = \frac{2}{M(M-1)}$ pairs of products j and k, yielding a sample covariance of

$$\hat{\Sigma}_{1n}^{c} = \frac{2}{NM\left(M-1\right)} \sum_{i=1}^{N} \sum_{j=1}^{M-1} \sum_{k=j+1}^{M} \hat{s}_{ij,n}\hat{s}_{ik,n}'. \tag{53}$$

A similar argument gives

$$\hat{\Sigma}_{1n}^{p} = \frac{2}{MN\left(N-1\right)} \sum_{j=1}^{M} \sum_{i=1}^{N-1} \sum_{k=i+1}^{N} \hat{s}_{ij,n}\hat{s}_{kj,n}'. \tag{54}$$

The 'dense', Wald-based, confidence intervals whose coverage properties are analyzed by Monte Carlo are based on the limit distribution for $n^{1/2}S_n$ given in equation (31) of the main text (with (53), (54) and ϕ_n replacing their populating/limiting values). Under dense asymptotics it is also the case that $\hat{\Gamma}_n \overset{def}{\equiv} H_n\left(\hat{\theta}\right)$ converges to, say, Γ_0, without rescaling by n. From these two observations a simple sandwich variance estimator can be constructed and inference based on the approximation (see, for example, Graham (2020a)):

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{approx}{\sim} \mathcal{N}\left(0, \hat{\Gamma}_n^{-1}\hat{\Omega}_n^{\mathrm{D}}\hat{\Gamma}_n^{-1}\right), \tag{55}$$

with $\hat{\Omega}_n^{\mathrm{D}} = \frac{\hat{\Sigma}_{1n}^{c}}{1-\phi_n} + \frac{\hat{\Sigma}_{1n}^{p}}{\phi_n}$.

Next define

$$\hat{\bar{s}}_{1i,n}^{c} = \frac{1}{M}\sum_{j=1}^{M} \hat{s}_{ij,n}$$

$$\hat{\bar{s}}_{1j,n}^{p} = \frac{1}{N}\sum_{i=1}^{N} \hat{s}_{ij,n}.$$

The 'jackknife' estimate of Σ_{1n}^{c} is

$$\breve{\Sigma}_{1n}^{c} = \frac{1}{N}\sum_{i=1}^{N} \hat{\bar{s}}_{1i,n}^{c}\hat{\bar{s}}_{1i,n}^{c\prime}. \tag{56}$$

See, for example, Efron and Stein (1981). Basic manipulation gives

$$
\begin{aligned}
\breve{\Sigma}_{1n}^{c} =& \frac{1}{N}\frac{1}{M^2}\sum_{i=1}^{N}\left[\sum_{j=1}^{M}\hat{s}_{ij,n}\right]\left[\sum_{j=1}^{M}\hat{s}_{ij,n}\right]' \\
=& \frac{1}{N}\frac{1}{M^2}\sum_{i=1}^{N}\left[\sum_{j=1}^{M}\hat{s}_{ij,n}\hat{s}_{ij,n}' + 2\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}\hat{s}_{ij,n}\hat{s}_{ik,n}'\right] \\
=& \frac{1}{M}\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{s}_{ij,n}\hat{s}_{ij,n}' + \frac{1}{N}\frac{2}{M^2}\sum_{i=1}^{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}\hat{s}_{ij,n}\hat{s}_{ik,n}' \\
=& \frac{1}{M}\widehat{\Sigma_{2n}+\Sigma_{3n}} + \frac{M-1}{M}\hat{\Sigma}_{1n}^{c}
\end{aligned}
$$

where I define $\widehat{\Sigma_{2n}+\Sigma_{3n}} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{s}_{ij,n}\hat{s}_{ij,n}'$.

These calculations give the equality

$$
\hat{\Sigma}_{1n}^{c} = \frac{M}{M-1}\left[\breve{\Sigma}_{1n}^{c} - \frac{1}{M}\widehat{\Sigma_{2n}+\Sigma_{3n}}\right].
$$

Analogous calculations yield

$$
\breve{\Sigma}_{1}^{p} = \frac{1}{M}\sum_{j=1}^{M}\hat{\bar{s}}_{1i,n}^{p}\hat{\bar{s}}_{1i,n}^{p\prime} = \frac{1}{N}\widehat{\Sigma_{2n}+\Sigma_{3n}} + \frac{N-1}{N}\hat{\Sigma}_{1n}^{p}
$$

and hence that

$$
\hat{\Sigma}_{1n}^{p} = \frac{N}{N-1}\left[\breve{\Sigma}_{1n}^{p} - \frac{1}{N}\widehat{\Sigma_{2n}+\Sigma_{3n}}\right].
$$

The jackknife estimate for $\mathbb{V}\left(n^{1/2}S_n\right)$ in the dense case is thus

$$
\begin{aligned}
\hat{\Omega}_{n}^{\mathrm{JK}} &= \frac{\breve{\Sigma}_{1n}^{c}}{1-\phi_n} + \frac{\breve{\Sigma}_{1n}^{p}}{\phi_n} \\
&= \frac{M-1}{M}\frac{\hat{\Sigma}_{1n}^{c}}{1-\phi_n} + \frac{N-1}{N}\frac{\hat{\Sigma}_{1n}^{p}}{\phi_n} + \frac{1}{M}\frac{\widehat{\Sigma_{2n}+\Sigma_{3n}}}{N/n} + \frac{1}{N}\frac{\widehat{\Sigma_{2n}+\Sigma_{3n}}}{M/n} \\
&= \frac{\hat{\Sigma}_{1n}^{c}}{1-\phi_n} + \frac{\hat{\Sigma}_{1n}^{p}}{\phi_n} + \frac{2n}{NM}\widehat{\Sigma_{2n}+\Sigma_{3n}} - \frac{1}{M}\frac{\hat{\Sigma}_{1n}^{c}}{N/n} - \frac{1}{N}\frac{\hat{\Sigma}_{1n}^{p}}{M/n} \\
&= \frac{\hat{\Sigma}_{1n}^{c}}{1-\phi_n} + \frac{\hat{\Sigma}_{1n}^{p}}{\phi_n} + \frac{1}{n}\frac{1}{\phi_n\left(1-\phi_n\right)}\left(2\left[\widehat{\Sigma_{2n}+\Sigma_{3n}}\right] - \hat{\Sigma}_{1n}^{c} - \hat{\Sigma}_{1n}^{p}\right).
\end{aligned}
$$

This suggests the bias corrected estimate of $\mathbb{V}\left(n^{1/2}S_n\right)$ equal to

$$
\begin{aligned}
\hat{\Omega}_{n}^{\mathrm{JK-BC}} &= \frac{\breve{\Sigma}_{1n}^{c}}{1-\phi_n} + \frac{\breve{\Sigma}_{1n}^{p}}{\phi_n} - \frac{1}{n}\frac{\widehat{\Sigma_{2n}+\Sigma_{3n}}}{\phi_n\left(1-\phi_n\right)} \\
&= \frac{\hat{\Sigma}_{1n}^{c}}{1-\phi_n} + \frac{\hat{\Sigma}_{1n}^{p}}{\phi_n} + \frac{1}{n}\frac{1}{\phi_n\left(1-\phi_n\right)}\left(\widehat{\Sigma_{2n}+\Sigma_{3n}} - \hat{\Sigma}_{1n}^{c} - \hat{\Sigma}_{1n}^{p}\right).
\end{aligned}
$$

See Cattaneo et al. (2014) for a related estimator in the context of density weighted average derivatives. [7]

To estimate $\mathbb{V}\left(n^{3/2}S_n\right)$, as required for sparse network inference, I use $n^2\hat{\Omega}^{\mathrm{JK-BC}}$ since

$$n^2\hat{\Omega}_n^{\mathrm{JK-BC}} = \frac{n^2\hat{\Sigma}_{1n}^c}{1-\phi_n} + \frac{n^2\hat{\Sigma}_{1n}^p}{\phi_n} + \frac{n}{\phi_n\left(1-\phi_n\right)}\left(\widehat{\Sigma_{2n}+\Sigma}_{3n} - \hat{\Sigma}_{1n}^c - \hat{\Sigma}_{1n}^p\right)$$

which, under suitable conditions, should be such that

$$n^2\hat{\Omega}_n^{\mathrm{JK-BC}} \to \frac{\tilde{\Sigma}_1^c}{1-\phi} + \frac{\tilde{\Sigma}_1^p}{\phi} + \frac{\tilde{\Sigma}_3}{\phi\left(1-\phi\right)} + O\left(\frac{1}{n}\right).$$

To estimate $\tilde{\Gamma}_0$ I use $-nH_n\left(\hat{\theta}\right)$. To ensure that $\hat{\Omega}_n^{\mathrm{JK-BC}}$ is positive definite I threshold negative eigenvalues as suggested by Cameron and Miller (2014).

The above estimators seem to be obvious places to start based on the prior work on dyadic clustering surveyed in Graham (2020a) and Graham (2020b). However, exploring the strengths and weakness of alternative methods of sparse network inference formally is a topic for future research.

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.

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.

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.

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

[7] Note that$n^2\hat{\Omega}^{\mathrm{JK}}$ appears to be a conservative estimate of $\mathbb{V}\left(n^{3/2}S_n\right)$ under sparsity (again see Cattaneo et al. (2014) for helpful discussion in a different context).

Chiang, H. D., Kato, K., Ma, Y., and Sasaki, Y. (2022a). Multiway cluster robust double/debiased machine learning. *Journal of Business and Economic Statistics*, 40(3):1046 – 1056.

Chiang, H. D., Matsushita, Y., and Otsu, T. (2022b). Multiway empirical likelohood. Technical report, London School of Economics.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729 – 737.

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. (2021). Empirical process results for exchangeable arrays. *Annals of Statistics*, 49(2):845 – 862.

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.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Annals of Statistics*, 9(3):586 – 596.

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

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, pages 111 – 218. North-Holland, Amsterdam, 1st edition.

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. (2022). Kernel density estimation for undirected dyadic data. *Journal of Econometrics (forthcoming).*

Hodges, J. L. and Le Cam, L. (1960). The poisson approximation to the poisson binomial distribution. *Annals of Mathematical Statistics*, 31(3):737 – 740.

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.

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

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

Menzel, K. (2021). Bootstrap with cluster-dependence in two or more dimensions. *Econometrica*, 89(5):2143 – 2188.

Mises, R. V. (1921). über die wahrscheinlichkeit seltener ereignisse. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 1(2):121 – 124.

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.

Newman, M. E. J. (2010). *Networks: An Introduction.* Oxford University Press, Oxford.

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.

Tabord-Meehan, M. (2018). Inference with dyadic data: Asymptotic behavior of the dyadic-robust t-statistic. *Journal of Business and Economic Statistics.*

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

Wang, H. (2020). *Proceedings of the 37th International Conference on Machine Learning*, volume 119, chapter Logistic regression for massive data with rare events, pages 9829 – 9836.

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