1-2014

Co-clustering separately exchangeable network data

David Choi
Carnegie Mellon University, davidch@andrew.cmu.edu

Patrick J. Wolfe
University College London

Published In
Annals of Statistics, 42, 1, 29-63.

The Annals of Statistics
2014, Vol. 42, No. 1, 29–63
DOI: 10.1214/13-AOS1173

CO-CLUSTERING SEPARATELY EXCHANGEABLE NETWORK DATA[1]

BY DAVID CHOI AND PATRICK J. WOLFE

Carnegie Mellon University and University College London

This article establishes the performance of stochastic blockmodels in addressing the co-clustering problem of partitioning a binary array into subsets, assuming only that the data are generated by a nonparametric process satisfying the condition of separate exchangeability. We provide oracle inequalities with rate of convergence $\mathcal{O}_P(n^{-1/4})$ corresponding to profile likelihood maximization and mean-square error minimization, and show that the blockmodel can be interpreted in this setting as an optimal piecewise-constant approximation to the generative nonparametric model. We also show for large sample sizes that the detection of co-clusters in such data indicates with high probability the existence of co-clusters of equal size and asymptotically equivalent connectivity in the underlying generative process.

1. Introduction. Blockmodels are popular tools for network modeling that see wide and rapidly growing use in analyzing social, economic and biological systems; see Zhao, Levina and Zhu (2011) and Fienberg (2012) for recent overviews. A blockmodel dictates that the probability of connection between any two network nodes is determined only by their respective block memberships, parameterized by a latent categorical variable at each node.

Fitting a blockmodel to a binary network adjacency matrix yields a clustering of network nodes, based on their shared proclivities for forming connections. More generally, fitting a blockmodel to any binary array involves partitioning it into blocks. In this way, blockmodels represent a piecewise-constant approximation to a latent function that generates network connection probabilities. This in turn can be viewed as a histogram-like approximation to a nonparametric generative process for binary arrays; fitting such models is termed co-clustering [Flynn and Perry (2012), Rohe and Yu (2012)].

This article analyzes the performance of stochastic blockmodels for co-clustering under model misspecification, assuming only an underlying generative

Received December 2012; revised September 2013.

[1]Supported in part by the US Army Research Office under PECASE Award W911NF-09-1-0555 and MURI Award W911NF-11-1-0036; by the UK EPSRC under Mathematical Sciences Established Career Fellowship EP/K005413/1 and Institutional Sponsorship Award EP/K503459/1; by the UK Royal Society under a Wolfson Research Merit Award; and by Marie Curie FP7 Integration Grant PCIG12-GA-2012-334622 within the 7th European Union Framework Program.

MSC2010 subject classifications. Primary 62G05; secondary 05C80, 60B20.

Key words and phrases. Bipartite graph, network clustering, oracle inequality, profile likelihood, statistical network analysis, stochastic blockmodel and co-blockmodel.

process that satisfies the condition of separate exchangeability [Diaconis and Janson (2008)]. This significantly generalizes known results for the blockmodel and its co-clustering variant, which have been established only recently under the requirement of correct model specification [Bickel and Chen (2009), Bickel, Chen and Levina (2011), Chatterjee (2012), Choi, Wolfe and Airoldi (2012), Fishkind et al. (2013), Flynn and Perry (2012), Rohe, Chatterjee and Yu (2011), Rohe and Yu (2012), Zhao, Levina and Zhu (2012)].

We show that blockmodels for co-clustering satisfy consistency properties and remain interpretable whenever separate exchangeability holds. Exchangeability is a natural condition satisfied by many network models: it characterizes permutation invariance, implying that the ordering of nodes carries no information [Bickel and Chen (2009), Hoff (2009)]. A blockmodel is an exchangeable model in which the connection probabilities are piecewise constant. Blockmodels also provide a simplified parametric approximation in the more general nonparametric setting [Bickel, Chen and Levina (2011)].

In addition to providing oracle inequalities for blockmodel M-estimators corresponding to profile likelihood and least squares optimizations, we show that it is possible to identify clusterings in data—what practitioners term *network communities*—even when the actual generative process is far from a blockmodel. The main statistical application of our results is to enable co-clustering under model misspecification. Much effort has been devoted to the task of community detection [Fortunato and Barthélemy (2007), Newman (2006), Zhao, Levina and Zhu (2011), Fienberg (2012)], but the drawing of inferential conclusions in this setting has been limited by the need to assume a correctly specified model.

Our results imply that community detection can be understood as finding a best piecewise-constant or simple function approximation to a flexible nonparametric process. In settings where the underlying generative process is not well understood and the specification of models is thus premature, such an approach is a natural first step for exploratory data analysis. This has been likened to the use of histograms to characterize exchangeable data in nonnetwork settings [Bickel and Chen (2009)].

The article is organized as follows. In Section 2, we introduce our nonparametric setting and model. In Section 3 we present oracle inequalities for co-clustering based on blockmodel fitting. In Section 4 we give our main technical result, and discuss a concrete statistical application: quantifying how the collection of co-clusterings of the data approaches that of a generative nonparametric process. We prove our main result in Section 5, by combining a construction used to establish a theory of graph limits [Borgs et al. (2006, 2008, 2012)] with statistical learning theory results on U-statistics [Clémençon, Lugosi and Vayatis (2008)]. In Section 6 we illustrate our results via a simulation study, and in Section 7 we relate them to other recent work. Appendices A–C contain additional proofs and technical lemmas.

2. Model elicitation. Recall that fitting a blockmodel to a binary array involves partitioning it into blocks. Denote by $G = (V_1, V_2, E)$ a bipartite graph with edge set E and vertex sets (V_1, V_2), where assignments of vertices to V_1 or V_2 are known. For example, V_1 and V_2 might represent people and locations, with edge (i, j) denoting that person i frequents location j. See Flynn and Perry (2012) and Rohe and Yu (2012) for additional examples.

2.1. *Exchangeable graph models.* For a bipartite graph G represented as a binary array A, the appropriate notion of exchangeability is as follows.

DEFINITION 2.1 (Separate exchangeability [Diaconis and Janson (2008)]). An array $\{A_{ij}\}_{i,j=1}^{\infty}$ of binary random variables is separately exchangeable if

$$P(A_{ij} = X_{ij}, 1 \le i, j \le n) = P(A_{ij} = X_{\Pi_1(i)\Pi_2(j)}, 1 \le i, j \le n)$$

for all $n = 1, 2, \dots,$ all permutations Π_1, Π_2 of for all $n = 1, 2, \dots,$ all permutations Π_1, Π_2 of $1, \dots, n$, and all $X \in \{0, 1\}^{n\times n}$.

If we identify a finite set of rows and columns of A with the adjacency matrix of an observed bipartite graph G, then it is clear that the notion of separate exchangeability encompasses a broad class of network models. Indeed, given a single observation of an unlabeled graph, it is natural to consider the class of all models that are invariant to permutation of its adjacency matrix; see Bickel and Chen (2009) and Hoff (2009) for discussion.

The assumption of separate exchangeability is the only one we will require for our results to hold. A representation of models in this class will be given by the Aldous–Hoover theorem for separately exchangeable binary arrays.

DEFINITION 2.2 (Exchangeable array model). Fix a measurable mapping $\omega : [0, 1]^3 \to [0, 1]$. Then the following model generates an exchangeable random bipartite graph $G = (V_1, V_2, E)$ through its adjacency matrix A:

(1) generate $\alpha \sim \text{Uniform}(0, 1)$;

(2) fix $m = |V_1|$ and $n = |V_2|$, and generate each element of $\xi = (\xi_1, \dots, \xi_m)$ and $\zeta = (\zeta_1, \dots, \zeta_n) \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1)$;

(3) for $i = 1, \dots, m$, and $j = 1, \dots, n$, generate $A_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\omega^{\alpha}(\xi_i, \zeta_j))$, where $\omega(x, y) \equiv \omega^{\alpha}(x, y)$ denotes the function $(x, y) \mapsto \omega(\alpha, x, y)$. If $A_{ij} = 1$, then connect vertices $i \in V_1$ and $j \in V_2$.

The Aldous–Hoover theorem states that this representation is sufficient to describe any separately exchangeable network distribution.

THEOREM 2.1 [Diaconis and Janson (2008)]. *Let $\{A_{ij}\}_{i,j=1}^{\infty}$ be a separately exchangeable binary array. Then there exists some $\omega : [0, 1]^3 \to [0, 1]$, unique up to measure-preserving transformation, which generates $\{A_{ij}\}_{i,j=1}^{\infty}$.*

The interpretation of the exchangeable graph model of Definition 2.2 is that each vertex has a latent parameter in [0, 1] (ξ_i for vertex i in V_1, and ζ_j for vertex j in V_2) which determines its affinity for connecting to other vertices, while α is a network-wide connectivity parameter (nonidentifiable from a single network observation). Because ξ and ζ are latent, $\omega(x, y)$ itself is identifiable only up to measure-preserving transformation, and is hence indistinguishable from any mapping $(x, y) \mapsto \omega(\alpha, \pi_1(x), \pi_2(y))$ for which π_1, π_2 are in the set $\mathcal{P}$ of measure-preserving bijective maps of [0, 1] to itself.

2.2. *The stochastic co-blockmodel.* Many popular network models can be recognized as instances of Definition 2.2. For example, Airoldi et al. (2008), Hoff, Raftery and Handcock (2002) and Kim and Leskovec (2012) all present models in which the resulting $\omega(\alpha, x, y)$ is constant in α, while Miller, Griffiths and Jordan (2009) require the full parameterization $\omega(\alpha, x, y)$. The stochastic co-blockmodel specifies $\omega(\alpha, x, y)$ constant in α and also piecewise-constant in x and y, and thus can be viewed as a simple function approximation to $\omega(x, y)$ in Definition 2.2.

DEFINITION 2.3 (Stochastic co-blockmodel [Rohe and Yu (2012)]). Fix integers $K_1, K_2 > 0$, a matrix $\theta \in [0, 1]^{K_1 \times K_2}$ and discrete probability measures μ and ν on $\{1, \dots, K_1\}$ and $\{1, \dots, K_2\}$. Then the *stochastic co-blockmodel* generates an exchangeable bipartite graph $G = (V_1, V_2, E)$ through the matrix A as follows:

(1) Fix $m = |V_1|$ and $n = |V_2|$, and generate $S = (S(1), \dots, S(m)) \overset{\text{i.i.d.}}{\sim} \mu$ and $T = (T(1), \dots, T(n)) \overset{\text{i.i.d.}}{\sim} \nu$.

(2) For $i = 1, \dots, m$, and $j = 1, \dots, n$, generate $A_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta_{S(i)T(j)})$. If $A_{ij} = 1$, then connect vertices $i \in V_1$ and $j \in V_2$.

Additionally, given co-blockmodel parameters $\phi \equiv (\mu, \nu, \theta)$, define

$$\omega_\phi(x, y) = \theta_{F_\mu^{-1}(x) F_\nu^{-1}(y)}, \qquad x, y \in [0, 1]$$

as the mapping corresponding to Definition 2.2, with $F_\mu^{-1}(x) = \inf_z\{F_\mu(z) \geq x\}$ the inverse distribution function corresponding to a given distribution μ.

Without loss of generality we assume $K_1 = K_2 = K$ in what follows, noting that our results do not depend in any crucial way on this assumption. Thus, a stochastic blockmodel's vertices in V_1 belong to one of K latent classes, as do those in V_2. Vectors $S \in \{1, \dots, K\}^m$ and $T \in \{1, \dots, K\}^n$ of categorical variables specify these class memberships. The matrix $\theta \in [0, 1]^{K \times K}$ indexes the corresponding connection affinities between classes in V_1 and V_2. Because S and T are latent, the stochastic co-blockmodel is identifiable only up to a permutation of its class labels.

3. Oracle inequalities for co-clustering. If we assume that the separately exchangeable data model of Definition 2.2 is in force, then a natural first step is to approximate $\omega(x,y)$ by way of some piecewise-constant $\omega_\phi(x,y)$, according to the stochastic co-blockmodel of Definition 2.3. This approximation task is equivalent to fixing K and estimating $\phi=(\mu,\nu,\theta)$ by co-clustering the entries of an observed adjacency matrix $A\in\{0,1\}^{m\times n}$.

3.1. *Sets of co-clustering parameters.* To accomplish this task, we consider M-estimators that involve an optimization over the latent categorical variable vectors $S\in\{1,\dots,K\}^m$ and $T\in\{1,\dots,K\}^n$. The resulting blockmodel estimates will reside in a set Φ containing triples $(\mu,\nu,\theta)\in\Omega_m\times\Omega_n\times[0,1]^{K\times K}$, where we define Ω_m to be the set of all probability distributions over $\{1,\dots,K\}$ whose elements are integer multiples of $1/m$,

$$\Omega_m=\left\{p\in\left\{0,\frac{1}{m},\frac{2}{m},\dots,1\right\}^K:\sum_{a=1}^K p_a=1\right\},$$

and likewise for Ω_n. Note that Ω_m and Ω_n are subsets of the standard $K-1$-simplex, chosen to contain all measures μ and ν that can be obtained by empirically co-clustering the elements of an $m\times n$-dimensional binary array. Thus, by construction, any estimator $\hat\phi(A)=(\hat\mu,\hat\nu,\hat\theta)$ based on an empirical co-clustering of an observed binary array $A\in\{0,1\}^{m\times n}$ has codomain Φ.

Given a specific μ and ν, let $\mathcal{Q}_\mu^m$ denote the set of all node-to-class assignment functions that partition the set $\{1,\dots,m\}$ into K classes in a manner that respects the proportions dictated by $\mu=(\mu_1,\dots,\mu_K)\in\Omega_m$,

$$\mathcal{Q}_\mu^m=\{v\in\{1,\dots,K\}^m:|v^{-1}(a)|=m\mu_a,a=1,\dots,K\},$$

and likewise for $\mathcal{Q}_\nu^n$.

3.2. *Oracle inequalities.* We now establish that, for L^2 risk and Kullback–Leibler divergence, there exist M-estimators that enable us to determine, with rate of convergence $n^{-1/4}$, optimal piecewise-constant approximations of the generative $\omega(x,y)$, up to quantization due to the discreteness of Φ.

THEOREM 3.1 (Oracle inequalities for co-clustering). *Let $A\in\{0,1\}^{m\times n}$ be a separately exchangeable array generated by some ω in accordance with Definition 2.2, and consider fitting a K-class stochastic co-blockmodel parameterized by $\phi\equiv(\mu,\nu,\theta)$ to A. Then as $n\to\infty$, with K and m/n fixed*:

(1) *For the least squares co-blockmodel M-estimator*

$$\hat\phi=\operatorname*{argmin}_{\phi\in\Phi}\left\{\min_{S\in\mathcal{Q}_\mu^m,T\in\mathcal{Q}_\nu^n}\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n|\theta_{S(i)T(j)}-A_{ij}|^2\right\}\tag{3.1}$$

relative to the L^2 risk

$$R_\omega(\phi) = \inf_{\pi_1,\pi_2\in\mathcal{P}} \int_{[0,1]^2} \big|\omega\big(\pi_1(x),\pi_2(y)\big) - \omega_\phi(x,y)\big|^2\,dx\,dy,$$

we have that

$$R_\omega(\hat{\phi}) - \inf_{\phi\in\Phi} R_\omega(\phi) = \mathcal{O}_P\big(n^{-1/4}\big);$$

(2) *Given any $\phi = (\mu,\nu,\theta)$, let $B(\phi) = \max_{1\le a,b,\le K} |\log(\theta_{ab}/(1-\theta_{ab}))|$. Consider the profile likelihood co-blockmodel M-estimator*

$$\hat{\phi} = \operatorname*{argmax}_{\phi\in\Phi}\Bigg\{\max_{S\in\mathcal{Q}_\mu^m, T\in\mathcal{Q}_\nu^n} \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n \big\{A_{ij}\log(\theta_{S(i)T(j)}) + (1-A_{ij})\log(1-\theta_{S(i)T(j)})\big\}\Bigg\} \tag{3.2}$$

relative to

$$L_\omega(\phi) = \sup_{\pi_1,\pi_2\in\mathcal{P}} \int_{[0,1]^2} \big\{\omega\big(\pi_1(x),\pi_2(y)\big)\log\omega_\phi(x,y) + \big[1-\omega\big(\pi_1(x),\pi_2(y)\big)\big]\log\big(1-\omega_\phi(x,y)\big)\big\}\,dx\,dy.$$

If $\phi^* = \operatorname{argmax}_{\phi\in\Phi} L_\omega(\phi)$ exists, and $B(\phi^*)$ and $B(\hat{\phi})$ are finite, then

$$\frac{\max_{\phi\in\Phi} L_\omega(\phi) - L_\omega(\hat{\phi})}{B(\phi^*) + B(\hat{\phi})} = \mathcal{O}_P\big(n^{-1/4}\big). \tag{3.3}$$

Theorem 3.1 can be viewed as analyzing maximum likelihood techniques in the context of model misspecification [White (1982)], and is proved in Appendix A. It establishes that minimization of the squared error between a fitted co-blockmodel and an observed binary array according to (3.1) serves as a proxy for approximation of ω by ω_ϕ in mean square, and that fitting a stochastic co-blockmodel via profile likelihood according to (3.2) is equivalent to minimizing the average Kullback–Leibler divergence of the approximation $\omega_\phi(x,y)$ from the generative $\omega(x,y)$.

The existence of a limiting object $\omega(x,y)$ implies that we are in the dense graph regime, with expected network degree values increasing linearly as a function of m or n. Given a correctly specified generative blockmodel, profile likelihood estimators are known to be consistent even in the sparse graph setting of polynomial or poly-logarithmic expected degree growth [Bickel and Chen (2009)]. In our setting, however, the generative model is no longer necessarily a blockmodel; in this context, both Borgs et al. (2008) and Chatterjee (2012) leave open the question of consistently estimating sparse network parameters, while Bickel, Chen and Levina (2011) give an identifiability result extending to the sparse case. The simulation study reported in Section 6 below suggests that the behavior of blockmodel estimators is qualitatively similar across at least some families of dense and sparse models.

3.3. *Additional remarks on Theorem* 3.1. In essence, Theorem 3.1 implies that the binary array A yields information on its underlying generative $\omega(x, y)$ at a rate of at least $n^{-1/4}$. While the necessary optimizations in (3.1) and (3.2) are not currently known to admit efficient exact algorithms, they strongly resemble existing objective functions for community detection for which many authors have reported good heuristics [Fortunato and Barthélemy (2007), Newman (2006), Zhao, Levina and Zhu (2011)]. Furthermore, polynomial-time spectral algorithms are known in certain settings to find correct labelings under the assumption of a generative blockmodel [Fishkind et al. (2013), Rohe, Chatterjee and Yu (2011)], suggesting that efficient algorithms may exist when distinct clusterings or community divisions are present in the data. In this vein, Chatterjee (2012) has recently proposed a universal thresholding procedure based on the singular value decomposition.

REMARK 3.1. We may replace the objective function of (3.2) with the full profile likelihood function $\max_{S\in\mathcal{Q}_\mu^m, T\in\mathcal{Q}_\nu^n}\{\sum_{i=1}^m \log\mu_{S(i)} + \sum_{j=1}^n \log\nu_{T(j)} + \sum_{i=1}^m\sum_{j=1}^n\{A_{ij}\log\theta_{S(i)T(j)} + (1-A_{ij})\log(1-\theta_{S(i)T(j)})\}\}$. The same rate of convergence can then be established with respect to the corresponding term for $L_\omega(\phi)$, adapting the proofs in Appendices A and B.

REMARK 3.2. Assume $\phi^* = \operatorname{argmax}_{\phi\in\Phi} L_\omega(\phi)$ exists. Terms $B(\phi^*)$ and $B(\hat{\phi})$ in (3.3) show that elements of θ^* and $\hat{\theta}$ must not approach 0 or 1 too quickly as $n\to\infty$; otherwise $L_\omega(\hat{\phi})$ can be much smaller than $L_\omega(\phi^*)$.

This is a natural consequence of the fact that the Kullback–Leibler divergence of ω_ϕ from ω is finite if and only if ω is absolutely continuous with respect to ω_ϕ. To see the implication, consider ξ, ζ, and A generated according to Definition 2.2 with $\omega(x, y) = 1\{x \le 1/2\}1\{y \le 1/2\}$. Let $\mu_1 = m^{-1}\sum_{i=1}^m 1\{\xi_i \le 1/2\}$ and $\nu_1 = n^{-1}\sum_{j=1}^n 1\{\zeta_i \le 1/2\}$. Then the maximum-likelihood two-class blockmodel fit to A will yield $\omega_{\hat{\phi}}(x, y) = 1\{x \le \mu_1\}1\{y \le \nu_1\}$, and so $L_\omega(\hat{\phi})$ diverges to $-\infty$ unless $\mu_1 = \nu_1 = 1/2$.

4. Convergence of co-cluster estimates. We now give our main technical result and show its statistical application in enabling us to interpret the convergence of co-cluster estimates. The estimators of Theorem 3.1 require optimizations over the set of all possible co-clusterings of the data; that is, over vectors S and T that map the observed vertices to $1,\dots,K$. Analogously, one may also envision an uncountable set of co-clusterings of the generative model, which map the unit interval [0, 1] to $1,\dots,K$. We define these two sets of co-clusterings more formally and then give a result showing in what sense they become close with increasing m and n, so that optimizing over co-clusters of the data is asymptotically equivalent to optimizing over co-clusters of the generative model. This result yields the rate of convergence $\mathcal{O}_P(n^{-1/4})$ appearing in Theorem 3.1, and also has a geometric interpretation that sheds light on the estimators defined by (3.1) and (3.2).

4.1. *Relating co-clusterings of A to those of ω.* Given a bipartite graph $G = (V_1, V_2, E)$ with adjacency matrix $A \in \{0, 1\}^{m\times n}$, recall that the latent class vectors $S \in \{1, \dots, K\}^m$ and $T \in \{1, \dots, K\}^n$ respectively partition V_1 and V_2 into K subsets each. To relate an empirical co-clustering of A to a piecewise-constant approximation of some ω, we first define the matrix $A/ST \in [0, 1]^{K\times K}$ to index the proportion of edges spanning each of the K^2 subset pairs defined by S and T,

$$(A/ST)_{ab} = \frac{1}{mn} \sum_{i\in S^{-1}(a)} \sum_{j\in T^{-1}(b)} A_{ij}, \qquad a, b = 1, \dots, K.$$

Second, we define mappings $\sigma, \tau : [0, 1] \to \{1, \dots, K\}$, which will play a role analogous to S and T. Given some $\omega : [0, 1]^2 \to [0, 1]$, this allows us to define a matrix $\omega/\sigma\tau \in [0, 1]^{K\times K}$ which encodes the mass of ω assigned to each of the K^2 subset pairs defined by σ and τ as follows:

$$(\omega/\sigma\tau)_{ab} = \int_{\sigma^{-1}(a)\times\tau^{-1}(b)} \omega(x, y)\, dx\, dy, \qquad a, b = 1, \dots, K.$$

We will use the $K \times K$ matrices A/ST and $\omega/\sigma\tau$ to index all possible co-clusterings that can be induced by partitioning an observed binary array $A \in \{0, 1\}^{m\times n}$ into K^2 blocks. To link these sets of co-clusters, recall from Section 3 the sets $\mathcal{Q}_\mu^m$ and $\mathcal{Q}_\nu^n$ of all node-to-class assignment functions that partition $\{1, \dots, m\}$ and $\{1, \dots, n\}$ into K classes in manners that respect the proportions dictated by $\mu = (\mu_1, \dots, \mu_K) \in \Omega_m$ and $\nu = (\nu_1, \dots, \nu_K) \in \Omega_n$. Analogously, we define $\mathcal{Q}_\mu$ (resp., $\mathcal{Q}_\nu$) to be the set of partitions of $[0, 1]$ into K subsets whose cardinalities are of proportions $\mu_1, \dots, \mu_K$:

$$\mathcal{Q}_\mu = \{\sigma : [0, 1] \to \{1, \dots, K\} \text{ such that } |\sigma^{-1}(a)| = \mu_a, a = 1, \dots, K\}.$$

We are now equipped to introduce sets $\mathcal{F}_{\mu\nu}^A$ and $\mathcal{F}_{\mu\nu}^\omega$, which describe all possible co-clusterings that can be induced from A and ω with respect to $(\mu, \nu) \in \Omega_m \times \Omega_n$, and to define the related notion of a support function.

DEFINITION 4.1 (Sets $\mathcal{F}_{\mu\nu}^A$ and $\mathcal{F}_{\mu\nu}^\omega$ of admissible co-clusterings). For fixed discrete probability distributions μ and ν over $1, \dots, K$, we define the sets $\mathcal{F}_{\mu\nu}^A, \mathcal{F}_{\mu\nu}^\omega \subset \mathbb{R}^{K\times K}$ of all co-clustering matrices A/ST and $\omega/\sigma\tau$, induced respectively by $(S, T) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n$ and $(\sigma, \tau) \in \mathcal{Q}_\mu \times \mathcal{Q}_\nu$, as follows:

$$\mathcal{F}_{\mu\nu}^A = \{A/ST \in [0, 1]^{K\times K} : S \in \mathcal{Q}_\mu^m, T \in \mathcal{Q}_\nu^n\},$$

$$\mathcal{F}_{\mu\nu}^\omega = \{\omega/\sigma\tau \in [0, 1]^{K\times K} : \sigma \in \mathcal{Q}_\mu, \tau \in \mathcal{Q}_\nu\}.$$

DEFINITION 4.2 (Support functions of $\mathcal{F}^A_{\mu\nu}$ and $\mathcal{F}^\omega_{\mu\nu}$). Let $\mathcal{F} \subset \mathbb{R}^{K\times K}$ be nonempty and with $\langle F, F'\rangle = \operatorname{tr}(F^T F')$. Its support function $h_{\mathcal{F}} : \mathbb{R}^{K\times K} \to \mathbb{R} \cup \{+\infty\}$ is defined as $h_{\mathcal{F}}(\Gamma) = \sup_{F\in\mathcal{F}}\langle\Gamma, F\rangle$ for any $\Gamma \in \mathbb{R}^{K\times K}$, whence

$$h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) = \max_{(S,T)\in\mathcal{Q}^m_\mu\times\mathcal{Q}^n_\nu}\langle\Gamma, A/ST\rangle, \tag{4.1a}$$

$$h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma) = \sup_{(\sigma,\tau)\in\mathcal{Q}_\mu\times\mathcal{Q}_\nu}\langle\Gamma, \omega/\sigma\tau\rangle. \tag{4.1b}$$

We will show below that $\sup_{\Gamma\in[-1,1]^{K\times K}} |h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma)|$ converges in probability to zero at a rate of at least $n^{-1/4}$, and this result in turn gives rise to Theorem 3.1. To see why, observe that for any $(\mu, \nu, \theta) \in \Phi$, the least squares objective function of (3.1) can be expressed using $h_{\mathcal{F}^A_{\mu\nu}}(\theta)$ as follows:

$$\begin{aligned}
&\min_{(S,T)\in\mathcal{Q}^m_\mu\times\mathcal{Q}^n_\nu}\left\{\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n\left(\theta^2_{S(i)T(j)} - 2\theta_{S(i)T(j)}A_{ij} + A^2_{ij}\right)\right\}\\
&\quad= \sum_{a=1}^K\sum_{b=1}^K \mu_a\nu_b\theta^2_{ab} - 2\max_{(S,T)\in\mathcal{Q}^m_\mu\times\mathcal{Q}^n_\nu}\langle\theta, A/ST\rangle + \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n A_{ij}\\
&\quad= \sum_{a=1}^K\sum_{b=1}^K \mu_a\nu_b\theta^2_{ab} - 2h_{\mathcal{F}^A_{\mu\nu}}(\theta) + \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n A_{ij}.
\end{aligned}$$

As we prove in Appendix A, this line of argument establishes the following.

LEMMA 4.1. *For any $(\mu, \nu, \theta) \in \Phi$, the difference between the least squares objective function of* (3.1) *and the L^2 risk R_ω is equal to*

$$2\big(h_{\mathcal{F}^\omega_{\mu\nu}}(\theta) - h_{\mathcal{F}^A_{\mu\nu}}(\theta)\big) + \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n A^2_{ij} - \int_{[0,1]^2}\omega(x, y)^2\,dx\,dy,$$

and the difference between the profile likelihood function of (3.2) *and L_ω is $B(\theta)(h_{\mathcal{F}^A_{\mu\nu}}(\Gamma_\theta) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma_\theta))$ whenever $0 < \theta_{ab} < 1$ for all $a, b = 1, \ldots, K$, with $\Gamma_\theta \in [-1, 1]^{K\times K}$ given element-wise by $(\Gamma_\theta)_{ab} = \log(\theta_{ab}/(1 - \theta_{ab}))/B(\theta)$.*

4.2. *A general result on consistency of co-clustering*. From Lemma 4.1 we see that closeness of $h_{\mathcal{F}^A_{\mu\nu}}$ to $h_{\mathcal{F}^\omega_{\mu\nu}}$ implies closeness (up to constant terms) of the least squares objective function of (3.1) to the L^2 risk $R_\omega(\phi)$, and of the profile likelihood of (3.2) to the average Kullback–Leibler divergence of $\omega_\phi(x, y)$ from the generative $\omega(x, y)$. Equipped with this motivation, we now state our main technical result, which serves to establish the rate of convergence $\mathcal{O}_P(n^{-1/4})$ in Theorem 3.1. Its proof follows in Section 5 below.

THEOREM 4.1. *Let $A \in \{0,1\}^{m \times n}$ be a separately exchangeable array generated by some ω in accordance with Definition 2.2. Then for each K and each ratio m/n, there exists a universal constant C such that as $n \to \infty$,*

$$\mathbb{P}\left(\max_{(\mu,\nu) \in \Omega_m \times \Omega_n} \left\{ \sup_{\Gamma \in [-1,1]^{K \times K}} \left| h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma) \right| \right\} \geq \frac{C}{n^{1/4}} \right) = o(1).$$

The support functions $h_{\mathcal{F}^A_{\mu\nu}}$ and $h_{\mathcal{F}^\omega_{\mu\nu}}$ also have a geometric interpretation: for any fixed $\Gamma \in \mathbb{R}^{K \times K}$, they define the supporting hyperplanes of the sets $\mathcal{F}^A_{\mu\nu}$ and $\mathcal{F}^\omega_{\mu\nu}$ in the direction specified by Γ. Each supporting hyperplane is induced by a point in $\mathcal{F}^A_{\mu\nu}$, or in the closure of $\mathcal{F}^\omega_{\mu\nu}$ respectively; these points are extremal in that they cannot be written as a convex combination of any other points in their respective sets. Evidently, it is only the extreme points which determine convergence properties for the risk functionals considered here. Equivalently, for any fixed parameter triple $\phi \in \Phi$, the values of these functionals depend only on the maximizing choices of (S, T) or (σ, τ).

Formally, Theorem 4.1 has the following geometric interpretation:

COROLLARY 4.1. *The result of Theorem 4.1 is equivalent to the following: The Hausdorff distance between the convex hulls of $\mathcal{F}^A_{\mu\nu}$ and $\mathcal{F}^\omega_{\mu\nu}$ is $\mathcal{O}_P(n^{-1/4})$.*

PROOF. Consider $\mathcal{F}, \mathcal{F}' \subset \mathbb{R}^{K \times K}$, and denote by $\|F\| = \sqrt{\operatorname{tr}(F^T F)}$ the Frobenius norm (i.e., the Hilbert–Schmidt metric on $\mathbb{R}^{K \times K}$ induced by $\langle \cdot, \cdot \rangle$). The Hausdorff distance between $\mathcal{F}$ and $\mathcal{F}'$, based on the metric $\|\cdot\|$, is then

$$d_{\text{Haus}}(\mathcal{F}, \mathcal{F}') = \max\left\{ \sup_{F \in \mathcal{F}} \left\{ \inf_{F' \in \mathcal{F}'} \|F - F'\| \right\}, \sup_{F' \in \mathcal{F}'} \left\{ \inf_{F \in \mathcal{F}} \|F - F'\| \right\} \right\}.$$

This measures the maximal shortest distance between any two elements of $\mathcal{F}$ and $\mathcal{F}'$. If these subsets of $\mathbb{R}^{K \times K}$ are furthermore nonempty and bounded, then the Hausdorff distance between their convex hulls $\operatorname{conv}(\mathcal{F})$ and $\operatorname{conv}(\mathcal{F}')$ can be expressed in terms of their support functions $h_{\mathcal{F}}, h_{\mathcal{F}'}$,

$$d_{\text{Haus}}\big(\operatorname{conv}(\mathcal{F}), \operatorname{conv}(\mathcal{F}')\big) = \sup_{\Gamma \in \mathbb{R}^{K \times K} : \|\Gamma\| = 1} \left| h_{\mathcal{F}}(\Gamma) - h_{\mathcal{F}'}(\Gamma) \right|;$$

see, for example, Schneider (1993), as applied to the convex hulls of the closures of $\mathcal{F}$ and of $\mathcal{F}'$. In this way, $d_{\text{Haus}}(\cdot, \cdot)$ is a natural measure of distance between two convex bodies. Recalling the equivalence of norms on $\mathbb{R}^{K^2}$, we see that

$$\begin{aligned} \sup_{\|\Gamma\|=1} \left| h_{\mathcal{F}}(\Gamma) - h_{\mathcal{F}'}(\Gamma) \right| &\leq \sup_{\Gamma \in [-1,1]^{K \times K}} \left| h_{\mathcal{F}}(\Gamma) - h_{\mathcal{F}'}(\Gamma) \right| \\ &\leq K \sup_{\|\Gamma\|=1} \left| h_{\mathcal{F}}(\Gamma) - h_{\mathcal{F}'}(\Gamma) \right|. \end{aligned}$$

Since Theorem 4.1 holds for $\sup_{\Gamma \in [-1,1]^{K \times K}} |h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma)|$, the leftmost inequality implies that it also holds for $\sup_{\|\Gamma\|=1} |h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma)|$. Now suppose instead that Theorem 4.1 holds for $\sup_{\|\Gamma\|=1} |h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma)|$; by the rightmost inequality, it then also holds for $K^{-1} \sup_{\Gamma\in[-1,1]^{K\times K}} |h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma)|$. Thus the result of Theorem 4.1 is equivalent to the statement that

$$\max_{(\mu,\nu)\in\Omega_m\times\Omega_n} d_{\mathrm{Haus}}\big(\mathrm{conv}(\mathcal{F}^A_{\mu\nu}), \mathrm{conv}(\mathcal{F}^\omega_{\mu\nu})\big) = \mathcal{O}_P\big(n^{-1/4}\big). \qquad \square$$

This geometric interpretation is helpful in relating our work to a series of papers by Borgs et al. (2006, 2008, 2012), which explore dense graph limits in depth and statistical applications thereof. Very broadly speaking, Borgs et al. (2008), Theorem 2.9 and Borgs et al. (2012), Theorem 4.6, analyze sets termed quotients, which resemble $\bigcup_{\mu,\nu} \mathcal{F}^A_{\mu\nu}$ and $\bigcup_{\mu,\nu} \mathcal{F}^\omega_{\mu\nu}$. The authors show convergence of these sets in the Hausdorff metric at rate $\mathcal{O}(\log^{-1/2} n)$, based on a distance termed the cut metric, and detail implications that can also be related to those of Bickel, Chen and Levina (2011).

In fixing μ and ν through our M-estimators, we are studying what Borgs et al. term the microcanonical quotients. Because our results require only convergence of the closed convex hulls of $\mathcal{F}^A_{\mu\nu}$ and $\mathcal{F}^\omega_{\mu\nu}$, we are able to obtain an exponentially faster bound on the rate of convergence.

4.3. *Interpreting convergence of blockmodel estimates.* Recall that the M-estimators of Theorem 3.1 each involve an optimization over the set $\mathcal{F}^A_{\mu\nu}$ by way of its support function, which in turn represents its convex hull. Suppose that $\hat\phi \equiv (\hat\mu, \hat\nu, \hat\theta)$ optimizes either objective function in Theorem 3.1. Then the following corollary of Theorem 3.1 shows that $\hat\phi$ is interpretable, in that there will exist a partition $\hat\sigma, \hat\tau$ of ω yielding co-clusters of equal size and asymptotically equivalent connectivity.

COROLLARY 4.2. *Let $\hat\phi = (\hat\mu, \hat\nu, \hat\theta)$ minimize the least squares criterion of* (3.1). *Then there exists some pair $(\hat\sigma, \hat\tau) \in \mathcal{Q}_{\hat\mu} \times \mathcal{Q}_{\hat\nu}$ such that*

$$\sum_{a=1}^K \sum_{b=1}^K \hat\mu_a \hat\nu_b \left| \frac{(\omega/\hat\sigma\hat\tau)_{ab}}{\hat\mu_a\hat\nu_b} - \hat\theta_{ab} \right|^2 = \mathcal{O}_P\big(n^{-1/4}\big).$$

Similarly, if $\hat\phi = (\hat\mu, \hat\nu, \hat\theta)$ maximizes the profile likelihood criterion of (3.2) *and $\phi^* = \operatorname{argmax}_{\phi\in\Phi} L_\omega(\phi)$ exists, then there is some $(\hat\sigma, \hat\tau) \in \mathcal{Q}_{\hat\mu} \times \mathcal{Q}_{\hat\nu}$ with*

$$\frac{1}{B(\phi^*) + B(\hat\phi)} \sum_{a=1}^K \sum_{b=1}^K \hat\mu_a \hat\nu_b D\left(\frac{(\omega/\hat\sigma\hat\tau)_{ab}}{\hat\mu_a\hat\nu_b} \Big\| \hat\theta_{ab} \right) = \mathcal{O}_P\big(n^{-1/4}\big),$$

where $D(p\|p') = p\log(p/p') + (1-p)\log[(1-p)/(1-p')] \ge 0$ is the Kullback–Leibler divergence of a Bernoulli(p') *distribution from a* Bernoulli(p) *one.*

PROOF. We show the latter result; parallel arguments yield the former. Since $\omega_{\hat{\phi}}(x,y) = \hat{\theta}_{F_{\hat{\mu}}^{-1}(x)F_{\hat{\nu}}^{-1}(y)}$ for the co-blockmodel, by letting σ and τ satisfy $\sigma(x) = F_{\hat{\mu}}^{-1}(\pi_1(x))$ and $\tau(y) = F_{\hat{\nu}}^{-1}(\pi_2(y))$ we may express $L_\omega(\hat{\phi})$ as

$$\sup_{(\sigma,\tau)\in\mathcal{Q}_{\hat{\mu}}\times\mathcal{Q}_{\hat{\nu}}} \sum_{a=1}^{K}\sum_{b=1}^{K} \int_{\sigma^{-1}(a)\times\tau^{-1}(b)} \{\omega(x,y)\log\hat{\theta}_{ab} + [1-\omega(x,y)]\log(1-\hat{\theta}_{ab})\}\,dx\,dy.$$

Thus, for any $\varepsilon > 0$, there exists some choice of $(\hat{\sigma},\hat{\tau}) \in \mathcal{Q}_{\hat{\mu}} \times \mathcal{Q}_{\hat{\nu}}$ such that

$$L_\omega(\hat{\phi}) - \varepsilon \le \sum_{a=1}^{K}\sum_{b=1}^{K}\{(\omega/\hat{\sigma}\hat{\tau})_{ab}\log\hat{\theta}_{ab} + [\hat{\mu}_a\hat{\nu}_b - (\omega/\hat{\sigma}\hat{\tau})_{ab}]\log(1-\hat{\theta}_{ab})\}.$$

If we now take $\hat{\theta}_{ab}^{(\omega)} = (\omega/\hat{\sigma}\hat{\tau})_{ab}/(\hat{\mu}_a\hat{\nu}_b)$ for $a,b = 1,\dots,K$, we see by a similar argument that since $L_\omega(\phi^*) = \max_{\phi\in\Phi} L_\omega(\phi)$, we have in turn that

$$\begin{aligned} L_\omega(\phi^*) &\ge L_\omega((\hat{\mu},\hat{\nu},\hat{\theta}^{(\omega)})) \\ &\ge \sum_{a=1}^{K}\sum_{b=1}^{K}\{(\omega/\hat{\sigma}\hat{\tau})_{ab}\log\hat{\theta}_{ab}^{(\omega)} + [\hat{\mu}_a\hat{\nu}_b - (\omega/\hat{\sigma}\hat{\tau})_{ab}]\log(1-\hat{\theta}_{ab}^{(\omega)})\}. \end{aligned}$$

Expanding $D(\hat{\theta}_{ab}^{(\omega)}\|\hat{\theta}_{ab})$ in accordance with its definition, we then see that

$$0 \le \sum_{a=1}^{K}\sum_{b=1}^{K}\hat{\mu}_a\hat{\nu}_b D(\hat{\theta}_{ab}^{(\omega)}\|\hat{\theta}_{ab}) \le L_\omega(\phi^*) - L_\omega(\hat{\phi}) + \varepsilon.$$

Choosing $\varepsilon = o(n^{-1/4})$ and applying Theorem 3.1 completes the proof. □

Corollary 4.2 ensures that co-blockmodel fits remain interpretable, even in the setting of model misspecification. It establishes that the identification of co-clusters in an observed exchangeable binary array A indicates with high probability the existence of co-clusters of equal size and asymptotically equivalent connectivity in the underlying generative process ω.

5. Proof of Theorem 4.1. Our proof strategy is inspired by Borgs et al. (2008) and adapts certain of its tools, but also requires new techniques in order to attain polynomial rates of convergence. Most significantly, we do not use the Szemerédi regularity lemma, which typically features strongly in the graph-theoretic literature, and provides a means of partitioning any large dense graph into a small number of regular clusters. Results in this direction are possible, but instead we use a Rademacher complexity bound for U-statistics adapted from Clémençon, Lugosi and Vayatis (2008), allowing us to achieve the improved rates of convergence described above.

5.1. *Establishing pointwise convergence.* The main step in proving Theorem 4.1 is to establish pointwise convergence of $h_{\mathcal{F}_{\mu\nu}^{A}}(\Gamma)$ to $h_{\mathcal{F}_{\mu\nu}^{\omega}}(\Gamma)$ for any fixed Γ. We do this through Proposition 5.1 below, after which we may apply it to a union bound over a covering of all $\Gamma \in [-1, 1]^{K\times K}$ to deduce the result of Theorem 4.1. Appendix B provides a formal statement and proof of this argument, along with proofs of all supporting lemmas.

PROPOSITION 5.1 [Pointwise convergence of $h_{\mathcal{F}_{\mu\nu}^{A}}(\Gamma)$ to $h_{\mathcal{F}_{\mu\nu}^{\omega}}(\Gamma)$]. *Assume the setting of Theorem 4.1, fixing $m = \rho n$. Then there exist constants C_K, n_K such that, given any $\Gamma \in [-1, 1]^{K\times K}$, μ, ν, ω, and $A \in \{0, 1\}^{m\times n}$ generated from ω, it holds for all $n \geq n_K$ that*

$$\mathbb{P}\left(\left|h_{\mathcal{F}_{\mu\nu}^{A}}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^{\omega}}(\Gamma)\right| \geq \frac{C_K}{n^{1/4}}\right) \leq 2e^{-\sqrt{n}[2\rho/(\rho+1)]}[1 + o(1)].$$

PROOF. To obtain the claimed result, we must establish lower and upper bounds on the support function $h_{\mathcal{F}_{\mu\nu}^{A}}(\Gamma)$ that show its convergence to $h_{\mathcal{F}_{\mu\nu}^{\omega}}(\Gamma)$ at rate $\mathcal{O}_P(n^{-1/4})$. Recalling the definitions of $h_{\mathcal{F}_{\mu\nu}^{A}}(\Gamma)$ and $h_{\mathcal{F}_{\mu\nu}^{\omega}}(\Gamma)$ in (4.1), we first require a statement of Lipschitz conditions on $\langle\Gamma, A/ST\rangle$ and $\langle\Gamma, \omega/\sigma\tau\rangle$. Its proof follows by direct inspection.

LEMMA 5.1. *Define for measurable mappings σ, σ' over* $[0, 1]$ *the metric*

$$d_{\mathrm{Ham}}(\sigma, \sigma') = \int_{[0,1]} 1\{\sigma(x) \neq \sigma'(x)\}\, dx,$$

and analogously the standard Hamming distance for sequences, with respect to normalized counting measure. Then for any $\Gamma \in [-1, 1]^{K\times K}$ and $A, A' \in [0, 1]^{m\times n}$, with $(S, T, \omega, \sigma, \tau)$ as defined in Section 4.1, we have that:

(1) $|\langle\Gamma, A/ST\rangle - \langle\Gamma, A/S'T'\rangle| \leq 2[d_{\mathrm{Ham}}(S, S')/m + d_{\mathrm{Ham}}(T, T')/n]$;
(2) $|\langle\Gamma, \omega/\sigma\tau\rangle - \langle\Gamma, \omega/\sigma'\tau'\rangle| \leq 2[d_{\mathrm{Ham}}(\sigma, \sigma') + d_{\mathrm{Ham}}(\tau, \tau')]$;
(3) $|\langle\Gamma, A/ST\rangle - \langle\Gamma, A'/ST\rangle| \leq 1/(mn)$ *if A, A' differ by a single entry.*

In conjunction with McDiarmid's inequality, these Lipschitz conditions yield the following lower bound on $h_{\mathcal{F}_{\mu\nu}^{A}}(\Gamma)$, proved in Appendix B.1.

LEMMA 5.2 [Lower bound on $h_{\mathcal{F}_{\mu\nu}^{A}}(\Gamma)$]. *Assume the setting of Theorem 4.1. Then there exist constants C'_K, n'_K such that, given any $\Gamma \in [-1, 1]^{K\times K}$, μ, ν, ω, and $A \in \{0, 1\}^{\rho n\times n}$ generated from ω, for all $n \geq n'_K$,*

$$\mathbb{P}\left(h_{\mathcal{F}_{\mu\nu}^{\omega}}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^{A}}(\Gamma) \geq \frac{C'_K}{n^{1/4}}\right) \leq 2e^{-\sqrt{n}[2\rho/(\rho+1)]}[1 + o(1)].$$

The upper bound comes by way of Rademacher complexity arguments. The remainder of this section and Appendix B is devoted to its proof.

LEMMA 5.3 [Upper bound on $h_{\mathcal{F}^A_{\mu\nu}}(\Gamma)$]. *Assume the setting of Theorem 4.1. Then there exist constants C''_K, n''_K such that, given any $\Gamma \in [-1,1]^{K\times K}$, μ, ν, ω and $A \in \{0,1\}^{\rho n\times n}$ generated from ω, for all $n \geq n''_K$,*

$$\mathbb{P}\biggl(h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma) \geq \frac{C''_K}{n^{1/4}}\biggr) \leq 2e^{-\sqrt{n}[2\rho/(\rho+1)]}[1+o(1)].$$

Proposition 5.1 now follows simply by combining Lemmas 5.2 and 5.3. □

5.2. *Establishing an upper bound on $h_{\mathcal{F}^A_{\mu\nu}}(\Gamma)$.* Lemma 5.3 represents the main technical hurdle in obtaining the polynomial rate of convergence given in Theorems 3.1 and 4.1. To illustrate the main ideas as clearly as possible, we will introduce our Rademacher complexity arguments below for the case $K = 2$, deferring the necessary generalizations to Appendix B.

We first define $W \in [0,1]^{m\times n}$ with reference to Definition 2.2 as

$$W_{ij} = \omega(\xi_i, \zeta_j), \qquad i \in 1, \dots, m, j \in 1, \dots, n;$$

and then define, in direct analogy to $h_{\mathcal{F}^A_{\mu\nu}}(\Gamma)$,

$$h_{\mathcal{F}^W_{\mu\nu}}(\Gamma) = \max_{(S,T)\in\mathcal{Q}^m_\mu\times\mathcal{Q}^n_\nu} \langle \Gamma, W/ST\rangle = \max_{(S,T)\in\mathcal{Q}^m_\mu\times\mathcal{Q}^n_\nu} \Biggl\{\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n W_{ij}\Gamma_{S(i)T(j)}\Biggr\}.$$

The matrix W serves as an empirical realization of the mapping ω, with its support function $h_{\mathcal{F}^W_{\mu\nu}}(\Gamma)$ defined with respect to co-blockmodel partitions $(S,T) \in \mathcal{Q}^m_\mu \times \mathcal{Q}^n_\nu$. As proved in Appendix B.2, Lemma 5.4 enables us to bound $|h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - \mathbb{E}h_{\mathcal{F}^W_{\mu\nu}}(\Gamma)|$ using the Lipschitz conditions in Lemma 5.1.

LEMMA 5.4. *Fix some measurable $\omega : [0,1]^2 \to [0,1]$, with $W \in [0,1]^{m\times n}$ generated by ω and $A \in \{0,1\}^{m\times n}$ generated by W, and some $\Gamma \in [-1,1]^{K\times K}$. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\bigl(\bigl|h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - \mathbb{E}h_{\mathcal{F}^W_{\mu\nu}}(\Gamma)\bigr| \geq 2\varepsilon\bigr) \leq 2e^{-2mn\varepsilon^2/(m+n)} + 2K^{m+n}e^{-2mn\varepsilon^2}. \tag{5.1}$$

Having bounded $|h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - \mathbb{E}h_{\mathcal{F}^W_{\mu\nu}}(\Gamma)|$, we must upper-bound $\mathbb{E}h_{\mathcal{F}^W_{\mu\nu}}(\Gamma)$ in terms of $h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma)$. We do this in a series of steps, first bounding $\mathbb{E}h_{\mathcal{F}^W_{\mu\nu}}(\Gamma)$ using a result adapted from Alon et al. (2003) and proved in Appendix B.3.

LEMMA 5.5. *Let $\mathcal{I}$ and $\mathcal{J}$ be sets of deterministic size, whose elements are sampled without replacement from $1,\dots,m$ and $1,\dots,n$. Let W be generated as* *in Lemma 5.4, and fix* $\Gamma \in [-1,1]^{K\times K}$. *Given* $W, \mathcal{I}, \mathcal{J}$ *and* $(Q,R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n$, *let* $\hat{S}^R \equiv \hat{S}^{R,\mathcal{J},W}$ *and* $\hat{T}^Q \equiv \hat{T}^{Q,\mathcal{I},W}$ *denote partitions satisfying*

$$\hat{S}^R = \operatorname*{argmax}_{S\in\mathcal{Q}_\mu^m}\left\{\sum_{i=1}^m \sum_{j\in\mathcal{J}} W_{ij}\Gamma_{S(i)R(j)}\right\}, \tag{5.2}$$

$$\hat{T}^Q = \operatorname*{argmax}_{T\in\mathcal{Q}_\nu^n}\left\{\sum_{i\in\mathcal{I}} \sum_{j=1}^n W_{ij}\Gamma_{Q(i)T(j)}\right\}. \tag{5.3}$$

Then

$$\begin{aligned}\mathbb{E}h_{\mathcal{F}_{\mu\nu}^W}(\Gamma) \le \mathbb{E}\Big(&\max_{(Q,R)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n}\langle\Gamma, W/\hat{S}^R\hat{T}^Q\rangle\Big)\\ &+ K\sqrt{2\pi}\big(|\mathcal{I}|^{-1/2}+|\mathcal{J}|^{-1/2}\big).\end{aligned} \tag{5.4}$$

To bound the right-hand side of (5.4) relative to $h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$, we will introduce an additional construction comprising several steps. Specifically, for fixed (Q,R) and Γ, we will define function classes $\mathcal{Q}_U$ and $\mathcal{Q}_V$, and a random functional $G_{\sigma\tau}$ which approximates $\langle\Gamma, W/\hat{S}^R\hat{T}^Q\rangle$ for some $(\hat{\sigma},\hat{\tau}) \in \mathcal{Q}_U\times\mathcal{Q}_V$. By a Rademacher complexity argument, $G_{\hat{\sigma}\hat{\tau}}$ will concentrate for all (Q,R) near its expectation, which itself will be bounded by $h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$.

For the case $K=2$, define U by

$$U(x) = \sum_{j\in\mathcal{J}} \omega(x,\zeta_j)(\Gamma_{1R(j)} - \Gamma_{2R(j)}).$$

It follows that

$$\hat{S}^R = \operatorname*{argmax}_{S\in\mathcal{Q}_\mu^m}\sum_{i=1}^m U(\xi_i)1\{S(i)=1\},$$

and so $\hat{S}^R$ will assign to class 1 the $\mu_1 m$ largest elements of $U(\xi_1),\dots,U(\xi_m)$. If U is invertible, this set can be written $\{\xi_i : U(\xi_i) < t\}$ for some t. To treat non-invertible U, define $\mathcal{Q}_U$ to be the class of functions $\{1_u : u\in[0,1]\}$, with 1_u a one-sided interval on the range of U with lexicographic "tie-breaking":

$$1_u(x) = \begin{cases} 2, & \text{if either } U(x) < U(u), \text{ or } U(x)=U(u) \text{ and } x<u;\\ 1, & \text{if either } U(x) > U(u), \text{ or } U(x)=U(u) \text{ and } x\ge u.\end{cases}$$

Then there exists $\hat{\sigma}\in\mathcal{Q}_U$ such that $\hat{S}^R$ can be chosen to satisfy

$$\hat{S}^R(i) = \hat{\sigma}(\xi_i), \qquad i=1,\dots,m.$$

Let V denote a function defined analogously to U as follows:

$$V(y) = \sum_{i\in\mathcal{I}} \omega(\xi_i,y)(\Gamma_{Q(i)1} - \Gamma_{Q(i)2}),$$

and likewise define $\mathcal{Q}_V$ so that there exists $\hat{\tau} \in \mathcal{Q}_V$ such that $\hat{T}^Q$ can be chosen to satisfy

$$\hat{T}^Q(j) = \hat{\tau}(\zeta_j), \qquad j = 1, \dots, n.$$

We are now ready to define $G_{\sigma\tau}$. Given any $\sigma \in \mathcal{Q}_U$ and $\tau \in \mathcal{Q}_V$, let

$$G_{\sigma\tau}(\xi, \zeta) = \frac{1}{mn} \sum_{i \in \overline{\mathcal{I}}} \sum_{j \in \overline{\mathcal{J}}} \omega(\xi_i, \zeta_j) \Gamma_{\sigma(\xi_i)\tau(\zeta_j)},$$

where $\overline{\mathcal{I}}$ is the complement of $\mathcal{I}$ in $\{1, \dots, m\}$, and $\overline{\mathcal{J}}$ the complement of $\mathcal{J}$ in $\{1, \dots, n\}$. Comparing $G_{\sigma\tau}$ to Lemma 5.5, we see that $G_{\hat{\sigma}\hat{\tau}}$ well approximates $\langle \Gamma, W/\hat{S}^R \hat{T}^Q \rangle$ whenever $|\mathcal{I}|$ and $|\mathcal{J}|$ are small; and indeed, we will later set $|\mathcal{I}| = |\mathcal{J}| = n^{1/2}$ in order to obtain an upper bound for $h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma)$.

By construction, the random classes $\mathcal{Q}_U$ and $\mathcal{Q}_V$ are independent of the random variables $\{\xi_i\}_{i \in \overline{\mathcal{I}}}$ and $\{\zeta_j\}_{i \in \overline{\mathcal{J}}}$ appearing in the summand of $G_{\sigma\tau}$. As a result, we may bound the deviation δ_{UV} of $G_{\sigma\tau}$ from its expectation,

$$\delta_{UV} = \sup_{(\sigma,\tau) \in \mathcal{Q}_U \times \mathcal{Q}_V} \left| G_{\sigma\tau}(\xi, \zeta) - \mathbb{E}\big(G_{\sigma\tau}(\xi, \zeta) | U, V\big) \right|,$$

using Rademacher complexity results for U-statistics due to Hoeffding (1963) and Clémençon, Lugosi and Vayatis (2008), Lemma A.1, applied to the class of one-sided interval functions.

LEMMA 5.6. *Assume the setting of Lemma 5.5, and set $\ell = \min(m - |\mathcal{I}|, n - |\mathcal{J}|)$. Then the deviation δ_{UV} of $G_{\sigma\tau}$ from its expectation satisfies*

$$\mathbb{E}\Big(\max_{(Q,R) \in \mathcal{Q}^m_\mu \times \mathcal{Q}^n_\nu} \delta_{UV} \Big) \le 4 \sqrt{\frac{(|\mathcal{I}| + |\mathcal{J}|) \log K + 2\binom{K}{2} \log(\ell + 1) + \log 2}{2\ell}}.$$

Lemma 5.6 is proved in Appendix B.5 to hold for arbitrary K, under the appropriate generalization of $\mathcal{Q}_U$, $\mathcal{Q}_V$, and quantities that depend on them.

Similarly, we may bound δ_U, defined for $K = 2$ as the maximum discrepancy between the expected and empirical class frequency in $\mathcal{Q}_U$,

$$\delta_U = \sup_{\sigma \in \mathcal{Q}_U} \left\{ \max_{1 \le a \le K} \left| |\sigma^{-1}(a)| - \frac{1}{m} \sum_{i=1}^m 1\{\sigma(\xi_i) = a\} \right| \right\},$$

with δ_V defined mutatis mutandis. We then have the following result, proved for arbitrary K (with appropriate redefinitions of δ_U, δ_V) in Appendix B.6.

LEMMA 5.7. *Assume the setting of Lemma 5.5. Then*

$$\mathbb{E}\Big(\max_{R \in \mathcal{Q}^n_\nu} \delta_U \Big) \le 4 \sqrt{\frac{(|\mathcal{J}| + 1) \log K + \binom{K}{2} \log(m + 1) + \log 2}{2m}},$$

$$\mathbb{E}\Big(\max_{Q \in \mathcal{Q}^m_\mu} \delta_V \Big) \le 4 \sqrt{\frac{(|\mathcal{I}| + 1) \log K + \binom{K}{2} \log(n + 1) + \log 2}{2n}}.$$

We state and prove a final auxiliary lemma prior to the proof of Lemma 5.3.

LEMMA 5.8. *Assume the setting of Lemma 5.5. Then*

$$\begin{aligned}
&\mathbb{E}\Big(\max_{(Q,R)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n}\langle\Gamma, W/\hat{S}^R\hat{T}^Q\rangle\Big) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)\\
&\qquad\le 2\{m^{-1}|\mathcal{I}| + n^{-1}|\mathcal{J}|\} + \mathbb{E}\Big(\max_{(Q,R)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n}\delta_{UV}\Big)\\
&\qquad\quad + 2K\mathbb{E}\Big(\max_{(Q,R)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n}\delta_U + \delta_V\Big).
\end{aligned}$$

PROOF. Let $\mathring{\sigma}$ and $\mathring{\tau}$ denote the mappings in $\mathcal{Q}_\mu$ and $\mathcal{Q}_\nu$ that are respectively closest in the metric d_{Ham} to $\hat{\sigma}$ and $\hat{\tau}$. Observe that we may then expand and upper-bound the left-hand side of the lemma statement by

$$\begin{aligned}
&\underbrace{\mathbb{E}\Big(\max_{Q,R}\langle\Gamma, W/\hat{S}^R\hat{T}^Q\rangle - G_{\hat\sigma\hat\tau}(\xi,\zeta)\Big)}_{\text{(i)}} + \underbrace{\mathbb{E}\Big(\max_{Q,R} G_{\hat\sigma\hat\tau}(\xi,\zeta) - \langle\Gamma,\omega/\hat\sigma\hat\tau\rangle\Big)}_{\text{(ii)}}\\
&\quad + \underbrace{\mathbb{E}\Big(\max_{Q,R}\langle\Gamma,\omega/\hat\sigma\hat\tau\rangle - \langle\Gamma,\omega/\mathring{\sigma}\mathring{\tau}\rangle\Big)}_{\text{(iii)}} + \underbrace{\mathbb{E}\Big(\max_{Q,R}\langle\Gamma,\omega/\mathring{\sigma}\mathring{\tau}\rangle\Big) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)}_{\text{(iv)}},
\end{aligned}$$

after which we may upper-bound terms (i)–(iv) in turn as follows.

First, since $|\omega(x,y)\Gamma_{\hat\sigma(x)\hat\tau(y)}| \le 1$ for all (x,y), it follows from their respective definitions that $\langle\Gamma, W/\hat{S}^R\hat{T}^Q\rangle - G_{\hat\sigma\hat\tau}(\xi,\zeta)$ is deterministically bounded above by $|\mathcal{I}|/m + |\mathcal{J}|/n$. Hence, term (i) is bounded by the same quantity.

Second, observe that by definition, $G_{\hat\sigma\hat\tau}(\xi,\zeta) - \mathbb{E}(G_{\hat\sigma\hat\tau}(\xi,\zeta)|U,V) \le \delta_{UV}$. Since for fixed σ,τ we have $\mathbb{E}(G_{\sigma\tau}(\xi,\zeta)|U,V) = [|\overline{\mathcal{I}}||\overline{\mathcal{J}}|/(mn)]\langle\Gamma,\omega/\sigma\tau\rangle$, with $|\langle\Gamma,\omega/\sigma\tau\rangle| \le 1$, it holds deterministically that $\mathbb{E}(G_{\hat\sigma\hat\tau}(\xi,\zeta)|U,V) - \langle\Gamma,\omega/\hat\sigma\hat\tau\rangle \le |\mathcal{I}|/m + |\mathcal{J}|/n$. Thus term (ii) is bounded above by the quantity $\mathbb{E}(\max_{(Q,R)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n}\delta_{UV}) + |\mathcal{I}|/m + |\mathcal{J}|/n$.

Third, by the second Lipschitz condition of Lemma 5.1, we have that $\langle\Gamma,\omega/\hat\sigma\hat\tau\rangle - \langle\Gamma,\omega/\mathring{\sigma}\mathring{\tau}\rangle \le 2[d_{\mathrm{Ham}}(\hat\sigma,\mathring{\sigma}) + d_{\mathrm{Ham}}(\hat\tau,\mathring{\tau})]$. Observe that

$$d_{\mathrm{Ham}}(\hat\sigma,\mathring{\sigma}) \le \sum_{a=1}^K \big||\hat\sigma^{-1}(a)| - \mu_a\big| \le \sum_{a=1}^K \left||\hat\sigma^{-1}(a)| - \frac{1}{m}\sum_{i=1}^m 1\{\hat\sigma(\xi_i) = a\}\right| \le K\delta_U,$$

where the second inequality holds as $\hat{S}^R \in \mathcal{Q}_\mu^m$. By the same argument for $d_{\mathrm{Ham}}(\hat\tau,\mathring{\tau})$, we see term (iii) is bounded by $2K\mathbb{E}(\max_{(Q,R)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n}\delta_U + \delta_V)$.

To conclude, note term (iv) is deterministically upper-bounded by 0. □

We may now establish the claimed upper bound on $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$.

PROOF OF LEMMA 5.3. Combining the results of Lemmas 5.4–5.8 yields directly that, with probability at least $1 - 2e^{-2mn\varepsilon^2/(m+n)} - 2K^{m+n}e^{-2mn\varepsilon^2}$,

$$\begin{aligned} h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma) \le{}& 2\varepsilon + K\sqrt{2\pi}\{|\mathcal{I}|^{-1/2} + |\mathcal{J}|^{-1/2}\} + 2\{m^{-1}|\mathcal{I}| + n^{-1}|\mathcal{J}|\} \\ &+ f\left(|\mathcal{I}| + |\mathcal{J}|, \ell, 2\binom{K}{2}\right) \\ &+ 2K\left\{f\left(|\mathcal{I}| + 1, n, \binom{K}{2}\right) + f\left(|\mathcal{J}| + 1, m, \binom{K}{2}\right)\right\}, \end{aligned}$$

where $f(p,q,r) = 4\{[p\log K + r\log(q+1) + \log 2]/(2q)\}^{1/2}$, and $\ell = \min(m - |\mathcal{I}|, n - |\mathcal{J}|)$ as in Lemma 5.6. Letting $\varepsilon = n^{-1/4}$, $|\mathcal{I}| = |\mathcal{J}| = n^{1/2}$, and fixing $m = \rho n$ as assumed in the hypothesis of Lemma 5.3, it follows that for $n \ge 2$,

$$\begin{aligned} h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma) \le{}& \frac{2 + 2K(2\pi)^{1/2} + (4\sqrt{2} + 8K)(2\log K)^{1/2}}{n^{1/4}} \\ &+ \frac{4 + 12(K^2\log(\rho n + 1) + 2)^{1/2}}{n^{1/2}} \end{aligned}$$

with probability at least $1 - 2e^{-\sqrt{n}[2\rho/(\rho+1)]} - 2K^{(\rho+1)n}e^{-2\rho n^{3/2}}$. Thus we have established the claimed upper bound on $h_{\mathcal{F}^A_{\mu\nu}}(\Gamma)$ in terms of $h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma)$. $\square$

6. Simulation study. We now present a brief simulation study which investigates empirical rates of convergence as model misspecification increases. We control the degree of misspecification through a sigmoidal functional form $f_\beta(x) : [0,1] \to [-1/2, 1/2]$, parameterized by $\beta \ge 1$,

$$\begin{aligned} f_\beta(x) &= Z_\beta^{-1}\left(\frac{x^\beta}{x^\beta + (1-x)^\beta} - \frac{1}{2}\right), \qquad 0 \le x \le 1; \\ Z_\beta &= 4\int_0^{1/2}\left|\frac{x^\beta}{x^\beta + (1-x)^\beta} - \frac{1}{2}\right| dx. \end{aligned}$$

Each $f_\beta(x)$ describes a strictly monotone increasing sigmoidal curve on $[0, 1]$, proportional to $x - 1/2$ for $\beta = 1$ and to $1\{x > 1/2\} - 1/2$ in the limit as $\beta \to \infty$. Normalization by Z_β maintains constant area under $|f_\beta|$.

To explore sparse graph regimes, we introduce an additional n-dependent parameter $\rho_n \in (0, 1)$, and take the outer product $f_\beta(x)f_\beta(y)$ to obtain a separable generative function $\rho_n\omega_\beta(x,y) = \rho_n(f_\beta(x)f_\beta(y) + 1/2)$. As $\beta \to \infty$, this tends to a stochastic co-blockmodel, with two classes of equal size.

Figure 1 shows a number of simulation results based on this model. Specifically, for $\beta \in \{1, 3, 5\}$ and $\rho_n \in \{0.5, n^{-2/3}, n^{-1}\log^2 n\}$, one thousand separable $n \times n$ binary arrays were generated from the corresponding $\rho_n\omega_\beta(x,y)$, for network sizes ranging from 100–500 for dense graphs (left column), and 100–2200

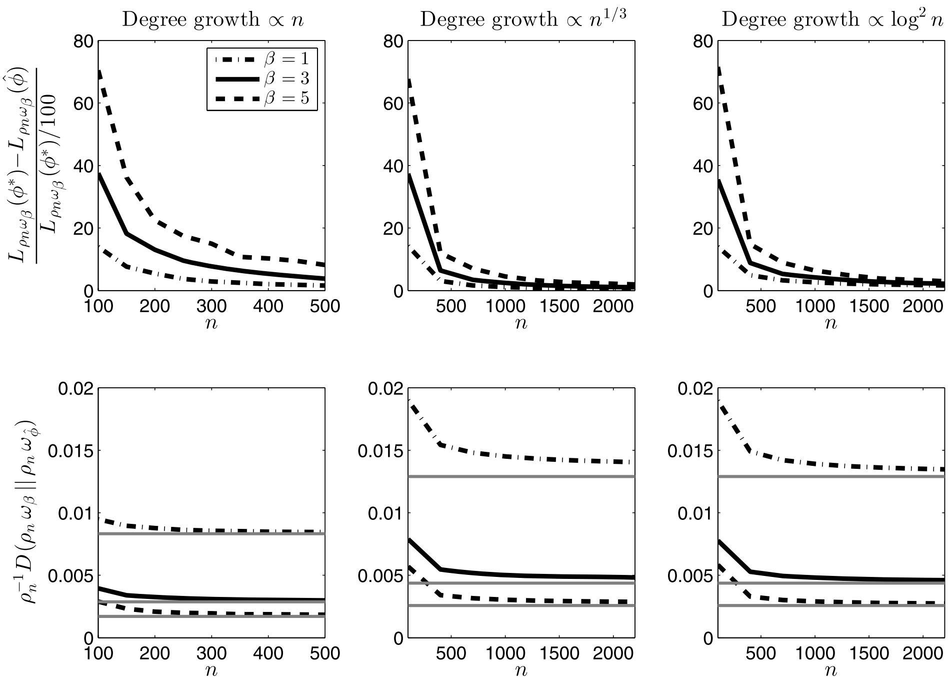

FIG. 1. *Median performance of approximate profile likelihood maximization according to* (3.2), *for* $\rho_n \in \{1/2, n^{-2/3}, n^{-1}\log^2 n\}$ (*left column*, *middle*, *right*). *Top row*: *percent relative excess risk*, *decaying toward zero. Bottom row*: *Kullback–Leibler divergence normalized by* ρ_n, *decaying toward its asymptotic optimum in* n (*grey horizontal lines*).

for sparse graphs (right columns). We see immediately that the simulation results of Figure 1 are qualitatively similar for all three regimes, suggesting that at least in some cases, co-blockmodel estimators will converge despite model misspecification in sparse as well as dense graph regimes.

Each of the $n \times n$ arrays described above was fitted by a two-class co-blockmodel, whose parameters $\hat{\phi} = (\hat{\mu}, \hat{\nu}, \hat{\theta})$ were obtained by heuristically optimizing the profile likelihood criterion of (3.2) using an algorithmic approach based on simulated annealing [Choi, Wolfe and Airoldi (2012)]. Parameter values were initialized to coincide with the optimal blockmodel approximation based on $\rho_n\omega_\beta/\sigma^*\tau^*$, where $\sigma^*, \tau^* : [0, 1] \to \{1, 2\}$ each map the interval $[0, 1/2)$ to class 1 and the interval $[1/2, 1]$ to class 2.

Lemma C.1 establishes that $\phi^* = \operatorname{argmax}_{\phi\in\Phi} L_{\rho_n\omega_\beta}(\phi)$ exists in this setting, and that $L_{\rho_n\omega_\beta}(\phi)$ may be straightforwardly computed for any triple $\phi = (\mu, \nu, \theta)$ of two-class co-blockmodel parameters. Corollary C.1 then yields a finite set containing $\phi^* = \operatorname{argmax}_{\phi\in\Phi} L_{\rho_n\omega_\beta}(\phi)$, from which we found that ϕ^* corresponded to the blockmodel induced by σ^* and τ^*. Thus we were able to evaluate the relative excess risk $[L_{\rho_n\omega_\beta}(\phi^*) - L_{\rho_n\omega_\beta}(\hat{\phi})]/L_{\rho_n\omega_\beta}(\phi^*)$, shown as a percentage in the top row of Figure 1, and seen to decay toward 0.

The bottom row of Figure 1 shows the normalized Kullback–Leibler divergences $\rho_n^{-1} D(\rho_n\omega_\beta \| \rho_n\omega_{\hat{\phi}})$ decaying toward the grey horizontal lines representing the limiting values of $D(\rho_n\omega_\beta \| \rho_n\omega_{\phi^*})$ as $\rho_n \to 0$. These are order-one quantities, obtained through a Taylor expansion of $D(\rho_n\omega_\beta \| \rho_n\omega_{\phi^*})$. Smaller divergences are achieved when β is large, reflecting the fact that as β increases, $\rho_n\omega_\beta(x, y)$ becomes closer to a co-blockmodel.

Overall, we see that the simulation results shown in Figure 1 are consistent with the behavior predicted by Theorem 3.1 for profile likelihood maximization; qualitatively similar results were also obtained for the least squares setting of Theorem 3.1 and hence are omitted for brevity.

7. Discussion. In this article we have addressed the case of network co-clustering, in which the inference task is to group two sets of network nodes into classes based on their observed relations. Our results significantly generalize known consistency results for the blockmodel and its co-blockmodel variant: they do not require the data to be generated (even approximately) by a co-blockmodel, and they achieve improved rates of convergence relative to results from the graph limits literature, through the use a Rademacher complexity bound for U-statistics adapted from Clémençon, Lugosi and Vayatis (2008). The assumption of a nonparametric generative model is both more general and more realistic, and to our knowledge Theorems 3.1 and 4.1 are the first for this regime to establish polynomial rates of convergence.

In the work of Clémençon, Lugosi and Vayatis (2008), these Rademacher complexity results are used to derive convergence rates for learning pairwise rankings. This setting is related to ours, but differs in some important ways. Those authors seek a rule $r : \mathcal{X} \times \mathcal{X} \to \{-1, +1\}$ such that, given $X, X' \in \mathcal{X}$, r indicates which has the higher rank. In this setting, X and X' can be thought of as covariates describing the two objects for which a relative ranking is desired, and $\mathcal{X}$ represents the space of allowable covariate values. In our network setting, the nonparametric model $\omega : [0, 1]^2 \to [0, 1]$ is analogous to a ranking rule, with $\mathcal{X}$ taken to be $[0, 1]$. However, X and X' are never observed in the data, and effectively must be imputed up to measure-preserving transformation.

The recent work of Flynn and Perry (2012) analyzes the consistency of co-clustering with model misspecification, but in a rather different setting, with the data matrix A assumed to be real valued, along with a real-valued generalization of the co-blockmodel. This generalization utilizes discrete latent class variables S and T; conditioned on $S(i)$ and $T(j)$, the distribution of A_{ij} is assumed to have mean $\theta_{S(i)T(j)}$, but may otherwise be arbitrary up to technical conditions, and may be misspecified in the estimator. Under these assumptions, it is shown that the latent classes can be estimated consistently if their number is known. In the case where A is binary, the conditions of Flynn and Perry (2012) are equivalent to assuming a generative co-blockmodel with a known number of classes.

Finally, the very recent work of Chatterjee (2012) derives a simple and elegant spectral method to consistently estimate the matrix W defined in the proof of Lemma 5.3 in Section 5.2, that is, the mapping $\omega(x, y)$, evaluated at the values of the latent variables $\xi_1, \dots, \xi_m$, and $\zeta_1, \dots, \zeta_n$. This implies consistency of estimation of ω in the L^2 sense, and while rates of convergence are not given for general ω, they can be established for particular instances, such as under the assumption of a generative blockmodel whose number of classes K is growing with n. Our setting is distinct, in that we desire only the best blockmodel approximation to ω, and so are able to establish L^2 rates of convergence that are independent of ω.

APPENDIX A: PROOF OF THEOREM 3.1 AND LEMMA 4.1

To prove Theorem 3.1, we first denote the objective functions of (3.1) and (3.2) by $R_A(\phi)$ and $L_A(\phi)$, respectively. Lemma 4.1, proved below, relates $R_A(\phi) - R_\omega(\phi)$ and $L_A(\phi) - L_\omega(\phi)$ to the support functions $h_{\mathcal{F}^A_{\mu\nu}}(\cdot)$ and $h_{\mathcal{F}^\omega_{\mu\nu}}(\cdot)$, after which the result follows directly from Theorem 4.1.

To see this, let $\hat{\phi} \equiv (\hat{\mu}, \hat{\nu}, \hat{\theta}) = \operatorname{argmin}_{\phi \in \Phi} R_A(\phi)$. For any $\phi \in \Phi$, we have

$$\begin{aligned}
&R_\omega(\hat{\phi}) - R_\omega(\phi) \\
&\quad = R_\omega(\hat{\phi}) - R_A(\hat{\phi}) + R_A(\hat{\phi}) - R_A(\phi) + R_A(\phi) - R_\omega(\phi) \\
&\quad \le R_\omega(\hat{\phi}) - R_A(\hat{\phi}) + R_A(\phi) - R_\omega(\phi) \\
&\quad \le 2\big|h_{\mathcal{F}^A_{\mu\nu}}(\hat{\theta}) - h_{\mathcal{F}^\omega_{\mu\nu}}(\hat{\theta})\big| + 2\big|h_{\mathcal{F}^\omega_{\mu\nu}}(\theta) - h_{\mathcal{F}^A_{\mu\nu}}(\theta)\big|,
\end{aligned}$$

where the first inequality holds because $R_A(\hat{\phi}) - R_A(\phi) \le 0$, and the second holds by the triangle inequality and Lemma 4.1. Applying Theorem 4.1 and choosing ϕ to satisfy $R_\omega(\phi) \le \inf_{\phi' \in \Phi} R_\omega(\phi') + n^{-1/4}$ then yields the result.

Now, assume $\phi^* = \operatorname{argmax}_{\phi \in \Phi} L_\omega(\phi)$ exists, and set $\hat{\phi} = \operatorname{argmax}_{\phi \in \Phi} L_A(\phi)$. Whenever $0 < \hat{\theta}_{ab}, \theta^*{}_{ab} < 1$ for all $a, b = 1, \dots, K$, the second result $[L_\omega(\phi^*) - L_\omega(\hat{\phi})]/[B(\theta^*) + B(\hat{\theta})] = \mathcal{O}_P(n^{-1/4})$ of Theorem 3.1 follows similarly from

$$\begin{aligned}
0 &\le L_\omega(\phi^*) - L_\omega(\hat{\phi}) \\
&= L_\omega(\phi^*) - L_A(\phi^*) + L_A(\phi^*) - L_A(\hat{\phi}) + L_A(\hat{\phi}) - L_\omega(\hat{\phi}) \\
&\le L_\omega(\phi^*) - L_A(\phi^*) + L_A(\hat{\phi}) - L_\omega(\hat{\phi}) \\
&\le B(\theta^*)\big|h_{\mathcal{F}^\omega_{\mu^*\nu^*}}(\Gamma_{\theta^*}) - h_{\mathcal{F}^A_{\mu^*\nu^*}}(\Gamma_{\theta^*})\big| + B(\hat{\theta})\big|h_{\mathcal{F}^A_{\hat{\mu}\hat{\nu}}}(\Gamma_{\hat{\theta}}) - h_{\mathcal{F}^\omega_{\hat{\mu}\hat{\nu}}}(\Gamma_{\hat{\theta}})\big|.
\end{aligned}$$

PROOF OF LEMMA 4.1. We show the results of the lemma directly,

$$\begin{aligned}
R_A(\phi) &= \min_{(S,T)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n} \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n |\theta_{S(i)T(j)} - A_{ij}|^2 \\
&= \min_{F\in\mathcal{F}_{\mu\nu}^A}\left\{\sum_{a=1}^K\sum_{b=1}^K -2F_{ab}\theta_{ab} + \mu_a\nu_b\theta_{ab}^2\right\} + \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n A_{ij}^2 \\
&= \left\{-2h_{\mathcal{F}_{\mu\nu}^A}(\theta) + \sum_{a=1}^K\sum_{b=1}^K \mu_a\nu_b\theta_{ab}^2\right\} + \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n A_{ij}^2,
\end{aligned}$$

where the second line follows from the definition of $\mathcal{F}_{\mu\nu}^A$, and the last line from that of $h_{\mathcal{F}_{\mu\nu}^A}$. Letting (σ,τ) satisfy $\sigma(x) = F^{-1}_{\mu(\pi_1(x))}$ and $\tau(y) = F^{-1}_{\nu(\pi_2(y))}$,

$$\begin{aligned}
R_\omega(\phi) &= \inf_{\pi_1,\pi_2\in\mathcal{P}}\int_{[0,1]^2}\left|\omega\big(\pi_1(x),\pi_2(y)\big) - \omega_\phi(x,y)\right|^2 dx\,dy \\
&= \inf_{(\sigma,\tau)\in\mathcal{Q}_\mu\times\mathcal{Q}_\nu}\sum_{a=1}^K\sum_{b=1}^K\int_{\sigma^{-1}(a)\times\tau^{-1}(b)}\left|\omega(x,y)-\theta_{ab}\right|^2 dx\,dy \\
&= \inf_{F\in\mathcal{F}_{\mu\nu}^\omega}\left\{\sum_{a=1}^K\sum_{b=1}^K -2F_{ab}\theta_{ab} + \mu_a\nu_b\theta_{ab}^2\right\} + \int_{[0,1]^2}\omega(x,y)^2\,dx\,dy \\
&= \left\{-2h_{\mathcal{F}_{\mu\nu}^\omega}(\theta) + \sum_{a=1}^K\sum_{b=1}^K\mu_a\nu_b\theta_{ab}^2\right\} + \int_{[0,1]^2}\omega(x,y)^2\,dx\,dy.
\end{aligned}$$

Following similar steps, we show the second result as follows:

$$\begin{aligned}
L_A(\phi) &= \max_{(S,T)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n}\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n\{A_{ij}\log(\theta_{S(i)T(j)}) \\
&\qquad\qquad + (1-A_{ij})\log(1-\theta_{S(i)T(j)})\} \\
&= \max_{F\in\mathcal{F}_{\mu\nu}^A}\sum_{a=1}^K\sum_{b=1}^K\left\{F_{ab}\log\left(\frac{\theta_{ab}}{1-\theta_{ab}}\right) + \mu_a\nu_b\log(1-\theta_{ab})\right\} \\
&= B(\theta)h_{\mathcal{F}_{\mu\nu}^A}(\Gamma_\theta) + \sum_{a=1}^K\sum_{b=1}^K\mu_a\nu_b\log(1-\theta_{ab}),
\end{aligned}$$

since $\max_{F\in\mathcal{F}_{\mu\nu}^A}\sum_{a,b}F_{ab}B(\theta)(\Gamma_\theta)_{ab} = B(\theta)h_{\mathcal{F}_{\mu\nu}^A}(\Gamma_\theta)$, and similarly

$$\begin{aligned}
L_\omega(\phi) = \sup_{\pi_1,\pi_2\in\mathcal{P}}\int_{[0,1]^2}&\{\omega\big(\pi_1(x),\pi_2(y)\big)\log\omega_\phi(x,y) \\
&+ \big[1-\omega\big(\pi_1(x),\pi_2(y)\big)\big]\log\big(1-\omega_\phi(x,y)\big)\}\,dx\,dy
\end{aligned}$$

$$= \sup_{(\sigma,\tau)\in\mathcal{Q}_\mu\times\mathcal{Q}_\nu} \sum_{a=1}^{K}\sum_{b=1}^{K}\int_{\sigma^{-1}(a)\times\tau^{-1}(b)} \{\omega(x,y)\log\theta_{ab} + (1-\omega(x,y))\log(1-\theta_{ab})\}\,dx\,dy$$

$$= \sup_{F\in\mathcal{F}^{\omega}_{\mu\nu}} \sum_{a=1}^{K}\sum_{b=1}^{K}\left\{F_{ab}\log\left(\frac{\theta_{ab}}{1-\theta_{ab}}\right) + \mu_a\nu_b\log(1-\theta_{ab})\right\}$$

$$= B(\theta)h_{\mathcal{F}^{\omega}_{\mu\nu}}(\Gamma_\theta) + \sum_{a=1}^{K}\sum_{b=1}^{K}\mu_a\nu_b\log(1-\theta_{ab}). \qquad \square$$

APPENDIX B: AUXILIARY PROOFS FOR THEOREM 4.1

Below we provide proofs of all supporting lemmas for Theorem 4.1, and state and prove the covering argument used to establish the theorem:

(1) First, in Sections B.1–B.3 below, we prove auxiliary Lemmas 5.2, 5.4 and 5.5 as stated in Section 5.

(2) Then, in Section B.4, we generalize the definitions of $\mathcal{Q}_U$ and $\mathcal{Q}_V$, given in Section 5.2 for $K = 2$, to arbitrary K; this induces generalizations of the quantities δ_U, δ_V and δ_{UV} in the natural way.

(3) Then, in Sections B.5 and B.6, we prove Lemmas 5.6 and 5.7, which depend on $(\mathcal{Q}_U, \mathcal{Q}_V, \delta_U, \delta_V, \delta_{UV})$ as defined for arbitrary K.

(4) Finally, in Section B.7, we extend the pointwise convergence result of Proposition 5.1 by way of a covering argument for all $\Gamma \in [-1,1]^{K\times K}$.

B.1. Proof of Lemma 5.2. For fixed Γ, let $(\sigma^*,\tau^*) \in \mathcal{Q}_\mu \times \mathcal{Q}_\nu$ satisfy

$$\langle\Gamma, \omega/\sigma^*\tau^*\rangle > h_{\mathcal{F}^{\omega}_{\mu\nu}}(\Gamma) - \frac{1}{n^{1/4}}, \tag{B.1}$$

so that $\omega/\sigma^*\tau^*$ is within $n^{-1/4}$ of the supporting hyperplane. Define

$$S^*(i) = \sigma^*(\xi_i), \qquad T^*(j) = \tau^*(\zeta_j); \qquad i = 1,\dots,m,\ j = 1,\dots,n.$$

By the arguments of Lemma 5.4 as proved in Section B.2 below, applying McDiarmid's inequality with the Lipschitz conditions of Lemma 5.1 yields

$$\mathbb{P}(|\langle\Gamma, A/S^*T^*\rangle - \langle\Gamma, \omega/\sigma^*\tau^*\rangle| \geq 2\varepsilon) \leq 2e^{-2mn\varepsilon^2/(m+n)} + 2e^{-2mn\varepsilon^2}. \tag{B.2}$$

While (S^*, T^*) many not be in $\mathcal{Q}^m_\mu \times \mathcal{Q}^n_\nu$, a Chernoff bound implies that

$$\mathbb{P}\left(\left|\frac{S^{*-1}(a)}{m} - \mu_a\right| \geq \varepsilon\right) \leq 2e^{-2m\varepsilon^2}, \qquad a = 1,\dots,K.$$

The analogous bound also holds for $|T^{*-1}(b)/n - \nu_b|$. Applying these results in conjunction with a union bound yields

$$\mathbb{P}\Big(\max_{1\le a,b\le K}\Big\{\Big|\frac{S^{*-1}(a)}{m} - \mu_a\Big| + \Big|\frac{T^{*-1}(b)}{n} - \nu_b\Big|\Big\} \ge 2\varepsilon\Big) \le K\big(2e^{-2m\varepsilon^2} + 2e^{-2n\varepsilon^2}\big).$$

Therefore, with probability at least $1 - K(2e^{-2m\varepsilon^2} + 2e^{-2n\varepsilon^2})$, there exists a pair $(\mathring{S}, \mathring{T}) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n$ such that

$$\frac{1}{m} d_{\mathrm{Ham}}(S^*, \mathring{S}) + \frac{1}{n} d_{\mathrm{Ham}}(T^*, \mathring{T}) \le 2K\varepsilon,$$

which by the first condition of Lemma 5.1 implies that

$$\big|\langle \Gamma, A/\mathring{S}\mathring{T}\rangle - \langle \Gamma, A/S^*T^*\rangle\big| \le 4K\varepsilon. \tag{B.3}$$

Recalling that $h_{\mathcal{F}^A_{\mu\nu}} = \max_{(S,T)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n} \langle \Gamma, A/ST\rangle$, we have that

$$h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) \ge \langle \Gamma, A/\mathring{S}\mathring{T}\rangle,$$

following which (B.3), (B.2) and (B.1) in turn imply that with probability at least $1 - 2e^{-2mn\varepsilon^2/(m+n)} - 2e^{-2mn\varepsilon^2} - K(2e^{-2m\varepsilon^2} + 2e^{-2n\varepsilon^2})$, we have

$$\begin{aligned} h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) &\ge \langle \Gamma, A/S^*T^*\rangle - 4K\varepsilon \\ &\ge \langle \Gamma, \omega/\sigma^*\tau^*\rangle - (4K+2)\varepsilon \\ &\ge h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma) - n^{-1/4} - (4K+2)\varepsilon. \end{aligned}$$

Now letting $m = \rho n$ as in the statement of the lemma, and setting $\varepsilon = n^{-1/4}$, we see that with probability at least $1 - 2e^{-\sqrt{n}[2\rho/(\rho+1)]}[1 + o(1)]$,

$$h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) \ge h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma) - \frac{4K+3}{n^{1/4}},$$

providing the necessary lower bound on $h_{\mathcal{F}^A_{\mu\nu}}(\Gamma)$ in terms of $h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma)$.

B.2. Proof of Lemma 5.4. Recalling the definitions of $h_{\mathcal{F}^A_{\mu\nu}}$ and $h_{\mathcal{F}^W_{\mu\nu}}$,

$$\begin{aligned} &\mathbb{P}\big(\big|h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^W_{\mu\nu}}(\Gamma)\big| \ge \varepsilon\big) \\ &\quad = \mathbb{P}\Big(\Big|\max_{(S,T)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n} \langle \Gamma, A/ST\rangle - \max_{(S,T)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n} \langle \Gamma, W/ST\rangle\Big| \ge \varepsilon\Big) \\ &\quad \le \mathbb{P}\Big(\max_{(S,T)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n} \big|\langle \Gamma, A/ST\rangle - \langle \Gamma, W/ST\rangle\big| \ge \varepsilon\Big) \end{aligned}$$

$$\le \sum_{(S,T)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n} \mathbb{P}\big(\big|\langle \Gamma, A/ST\rangle - \langle \Gamma, W/ST\rangle\big| \ge \varepsilon\big) \tag{B.4}$$

$$= \sum_{(S,T)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n} \mathbb{P}\big(\big|\langle \Gamma, A/ST\rangle - \mathbb{E}(\langle \Gamma, A/ST\rangle)\big| \ge \varepsilon\big), \tag{B.5}$$

where (B.4) follows by a union bound, and (B.5) by considering $\langle \Gamma, A/ST \rangle$ as a function of the mn independent random variables $\{A_{ij}\}$, which shows that $\mathbb{E}(\langle \Gamma, A/ST \rangle) = \langle \Gamma, W/ST \rangle$ for each (S, T), as $W_{ij} = \omega(\xi_i, \zeta_j) = \mathbb{E}(A_{ij})$.

Next, recall the final Lipschitz condition of Lemma 5.1, which states that $|\langle \Gamma, A/ST \rangle - \langle \Gamma, A'/ST \rangle| \le 1/(mn)$ if A and A' differ by a single entry. Thus we may apply McDiarmid's inequality to bound each term in (B.5), and since $|\mathcal{Q}_\mu^m| \le K^m$ and $|\mathcal{Q}_\nu^n| \le K^n$, we obtain after summing that

$$\mathbb{P}\big(|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \ge \varepsilon\big) \le K^{m+n} \cdot 2e^{-2mn\varepsilon^2}.$$

Now consider $h_{\mathcal{F}_{\mu\nu}^W}(\Gamma) = \max_{(S,T)\in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \langle \Gamma, W/ST \rangle$ as a function of the $m + n$ independent random variables $\xi_1, \dots, \xi_m$ and $\zeta_1, \dots, \zeta_n$. Changing a single component of ξ or ζ affects only a single row or column of W, respectively, and thus alters $\langle \Gamma, W/ST \rangle$ and hence $h_{\mathcal{F}_{\mu\nu}^W}$ by at most $1/m$ or $1/n$. It therefore follows directly from McDiarmid's inequality that

$$\mathbb{P}\big(|h_{\mathcal{F}_{\mu\nu}^W}(\Gamma) - \mathbb{E}h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \ge \varepsilon\big) \le 2e^{-2mn\varepsilon^2/(m+n)}.$$

Combining these inequalities via a union bound yields the statement of the lemma, since by the triangle inequality we must have $|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \ge \varepsilon$ or $|h_{\mathcal{F}_{\mu\nu}^W}(\Gamma) - \mathbb{E}h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \ge \varepsilon$ in order that $|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - \mathbb{E}h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \ge 2\varepsilon$.

B.3. Proof of Lemma 5.5. Recall from the statement of the lemma that $\mathcal{I}$ and $\mathcal{J}$ denote sets of deterministic size whose elements are sampled without replacement from $1, \dots, m$ and $1, \dots, n$, respectively. We adopt the notation that $\mathbb{E}_\mathcal{I}$ denotes an expectation taken over $\mathcal{I}$, with all other random variables held constant, and define $\mathbb{E}_\mathcal{J}$ and $\mathbb{E}_{\mathcal{I}\mathcal{J}}$ in the same manner.

To prove the lemma, it suffices to show that for all W, T, S,

$$\mathbb{E}_\mathcal{J}(\langle \Gamma, W/\hat{S}^T T \rangle) \ge \langle \Gamma, W/S^T T \rangle - K\sqrt{2\pi/|\mathcal{J}|}, \tag{B.6}$$

$$\mathbb{E}_\mathcal{I}(\langle \Gamma, W/S\hat{T}^S \rangle) \ge \langle \Gamma, W/ST^S \rangle - K\sqrt{2\pi/|\mathcal{I}|}, \tag{B.7}$$

where $\hat{S}^T$ and $\hat{T}^S$ are respectively defined in (5.2) and (5.3), and

$$S^T = \operatorname*{argmax}_{S \in \mathcal{Q}_\mu^m} \langle \Gamma, W/ST \rangle, \qquad T^S = \operatorname*{argmax}_{T \in \mathcal{Q}_\nu^n} \langle \Gamma, W/ST \rangle.$$

This is because (B.6) and (B.7) imply that for all $(U, V) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n$,

$$\begin{aligned}
\langle \Gamma, W/UV \rangle &\le \langle \Gamma, W/UT^U \rangle \\
&\le \mathbb{E}_\mathcal{I}(\langle \Gamma, W/U\hat{T}^U \rangle) + K\sqrt{2\pi/|\mathcal{I}|} \\
&\le \mathbb{E}_\mathcal{I}(\langle \Gamma, W/S^{\hat{T}^U}\hat{T}^U \rangle) + K\sqrt{2\pi/|\mathcal{I}|}
\end{aligned}$$

$$
\begin{aligned}
&\le \mathbb{E}_{\mathcal{I}}\mathbb{E}_{\mathcal{J}}\big(\langle \Gamma, W/\hat{S}^{\hat{T}^U}\hat{T}^U\rangle\big) + K\sqrt{2\pi/|\mathcal{I}|} + K\sqrt{2\pi/|\mathcal{J}|} \\
&\le \mathbb{E}_{\mathcal{I}\mathcal{J}}\Big(\max_{(Q,R)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n}\langle \Gamma, W/\hat{S}^R\hat{T}^Q\rangle\Big) \\
&\qquad + K\sqrt{2\pi}\big(|\mathcal{I}|^{-1/2} + |\mathcal{J}|^{-1/2}\big).
\end{aligned}
$$

Recalling the definition of $h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)$, and noting that the right-hand side above is deterministic for fixed W, with no dependence on U or V, we may write

$$
\begin{aligned}
h_{\mathcal{F}_{\mu\nu}^W}(\Gamma) &= \max_{(U,V)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n}\langle \Gamma, W/UV\rangle \\
&\le \mathbb{E}_{\mathcal{I}\mathcal{J}}\Big(\max_{(Q,R)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n}\langle \Gamma, W/\hat{S}^R\hat{T}^Q\rangle\Big) \\
&\qquad + K\sqrt{2\pi}\big(|\mathcal{I}|^{-1/2} + |\mathcal{J}|^{-1/2}\big).
\end{aligned}
$$

Taking expectations on both sides over W gives the statement of the lemma.

We now establish (B.6), noting that (B.7) will follow by parallel arguments. For fixed W and T, define for any $a = 1, \dots, K$ the difference

$$
\Delta_i^a = \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} W_{ij}\Gamma_{aT(j)} - \frac{1}{n}\sum_{j=1}^n W_{ij}\Gamma_{aT(j)}.
$$

It follows that $\mathbb{E}_{\mathcal{J}}(\Delta_i^a) = 0$, and by a Chernoff bound,

$$
\mathbb{P}\big(|\Delta_i^a| \ge t\big) \le 2e^{-2t^2|\mathcal{J}|}.
$$

As $|\Delta_i^a|$ is nonnegative, the identity $\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X \ge t)\,dt$ for X taking only nonnegative values can be used to bound its expectation according to

$$
\mathbb{E}_{\mathcal{J}}\big(|\Delta_i^a|\big) \le \sqrt{\pi/(2|\mathcal{J}|)},
$$

which implies

$$
\mathbb{E}_{\mathcal{J}}\Big(\max_{1\le a\le K}|\Delta_i^a|\Big) \le K\sqrt{\pi/(2|\mathcal{J}|)}. \tag{B.8}
$$

For fixed W and $\mathcal{J}$, define the function

$$
f_W(S,T) = \frac{1}{m|\mathcal{J}|}\sum_{i=1}^m\sum_{j\in\mathcal{J}} W_{ij}\Gamma_{S(i)T(j)},
$$

and for fixed W and T, let

$$
\Delta = \max_{S\in\mathcal{Q}_\mu^m}\big|f_W(S,T) - \langle \Gamma, W/ST\rangle\big|. \tag{B.9}
$$

From the definition of Δ it follows that

$$\begin{aligned}\Delta &= \max_{S \in \mathcal{Q}_\mu^m} \left\{ \frac{1}{m} \left| \sum_{i=1}^m \left(\frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} W_{ij} \Gamma_{S(i)T(j)} - \frac{1}{n} \sum_{j=1}^n W_{ij} \Gamma_{S(i)T(j)} \right) \right| \right\} \\ &\le \frac{1}{m} \left| \sum_{i=1}^m \max_{1 \le a \le K} \left\{ \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} W_{ij} \Gamma_{aT(j)} - \frac{1}{n} \sum_{j=1}^n W_{ij} \Gamma_{aT(j)} \right\} \right| \\ &= \frac{1}{m} \left| \sum_{i=1}^m \max_{1 \le a \le K} \{\Delta_i^a\} \right| \le \frac{1}{m} \sum_{i=1}^m \max_{1 \le a \le K} |\Delta_i^a|.\end{aligned}$$

Taking expectations of both sides over $\mathcal{J}$ and substituting (B.8) yields

$$\mathbb{E}_{\mathcal{J}}(\Delta) \le \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\mathcal{J}}\Big(\max_{1 \le a \le K} |\Delta_i^a| \Big) \le K \sqrt{\pi / (2|\mathcal{J}|)}. \tag{B.10}$$

Finally, to show (B.6), observe that since $\hat{S}^T$ from (5.2) maximizes $f_W(\cdot, T)$, and S^T as defined above maximizes $\langle \Gamma, W / \cdot\, T \rangle$, we have from (B.9) that

$$\begin{aligned}0 &\le \langle \Gamma, W/S^T T \rangle - \langle \Gamma, W/\hat{S}^T T \rangle \\ &\le \langle \Gamma, W/S^T T \rangle - f_W(S^T, T) + f_W(\hat{S}^T, T) - \langle \Gamma, W/\hat{S}^T T \rangle \le 2\Delta,\end{aligned}$$

and so $\langle \Gamma, W/\hat{S}^T T \rangle \ge \langle \Gamma, W/S^T T \rangle - 2\Delta$. Taking expectations of both sides of this expression over $\mathcal{J}$, and then substituting (B.10), yields the inequality

$$\mathbb{E}_{\mathcal{J}}(\langle \Gamma, W/\hat{S}^T T \rangle) \ge \langle \Gamma, W/S^T T \rangle - 2K \sqrt{\pi / (2|\mathcal{J}|)},$$

which is the statement of (B.6). That of (B.7) follows by parallel arguments.

B.4. Definition of $\mathcal{Q}_U$ and $\mathcal{Q}_V$ for arbitrary K. In order to redefine $\mathcal{Q}_U$ and $\mathcal{Q}_V$ to accommodate arbitrary K, we first redefine the mappings U and V. Given $\zeta_{\mathcal{J}} = \{\zeta_j : j \in \mathcal{J}\}$ and an assignment function $R : \{1, \ldots, n\} \to \{1, \ldots, K\}$, define the mapping $U : [0, 1] \to \mathbb{R}^K$ by

$$U_a(x) = \sum_{j \in \mathcal{J}} \omega(x, \zeta_j) \Gamma_{aR(j)}, \qquad x \in [0, 1], a = 1, \ldots, K.$$

Analogously, given $\xi_{\mathcal{I}}$ and Q, define $V : [0, 1] \to \mathbb{R}^K$ by

$$V_a(y) = \sum_{i \in \mathcal{I}} \omega(\xi_i, y) \Gamma_{Q(i)a}, \qquad y \in [0, 1], a = 1, \ldots, K.$$

Given $a, b \in \{1, \ldots, K\}$ and the mapping U, define the relation $\succeq^{U,a,b}$ by

$$\begin{aligned}&x_1 \succeq^{U,a,b} x_2 \\ &\quad \equiv \begin{cases} U_a(x_1) - U_b(x_2) > U_a(x_2) - U_b(x_1), & \text{or} \\ U_a(x_1) - U_b(x_2) = U_a(x_2) - U_b(x_1), & \text{if } (a - b)(x_1 - x_2) \ge 0. \end{cases}\end{aligned}$$

Informally, $x_1 \succeq^{U,a,b} x_2$ implies that, given the choice of assigning either x_1 or x_2 to group a, with the other relegated to group b, x_1 is at least as attractive as x_2. The latter tie-breaker condition results in a symmetric definition: if $x_1 \succeq^{U,a,b} x_2$, then $x_2 \succeq^{U,b,a} x_1$. We define $\succ^{U,a,b}$ analogously to $\succeq^{U,a,b}$, except that the inequality $(a-b)(x_1-x_2) > 0$ is strict.

Let $\mathcal{S}$ denote the set of symmetric matrices in $[0,1]^{K\times K}$. Given $t \in \mathcal{S}$ and the mapping U, we define the function $\sigma_t : [0,1] \to \{1,\dots,K\}$ as the mapping which satisfies the following:

$$\sigma_t^{-1}(a) = \{x : x \succeq^{U,a,b} t_{ab}\ \forall b > a, x \succ^{U,a,b} t_{ab}\ \forall b < a\}, \qquad a = 1,\dots,K,$$

with the convention that σ_t is undefined whenever the above rule does not map all of $[0,1]$ to $\{1,\dots,K\}$.

We define the function class $\mathcal{Q}_U$ as follows:

$$\mathcal{Q}_U = \{\sigma_t : t \in \mathcal{S} \text{ and } \sigma_t \text{ is defined}\}.$$

Given $t \in \mathcal{S}$ and the mapping V as defined above, we define $\succ^{V,a,b}$, τ_t and $\mathcal{Q}_V$ analogously. We then have the following.

LEMMA B.1. *Given U induced by $\zeta_{\mathcal{J}}$ and R, and given W induced by ξ and ζ, define $\hat{S}^R$ by (5.2). Then there exists $\hat{\sigma} \in \mathcal{Q}_U$ such that*

$$\hat{S}^R(i) = \hat{\sigma}(\xi_i), \qquad i = 1,\dots,m.$$

Likewise, given V induced by $\xi_{\mathcal{I}}$ and Q, and given W induced by ξ and ζ, define $\hat{T}^Q$ by (5.3). Then there exists $\hat{\tau} \in \mathcal{Q}_V$ such that

$$\hat{T}^Q(j) = \hat{\tau}(\zeta_j), \qquad j = 1,\dots,n.$$

PROOF. Let $\hat{S}^R$ be chosen lexicographically from the set of all maximizers of (5.2), where S lexicographically precedes S' if and only if $S(i_1),\dots,S(i_m)$ lexicographically precedes $S'(i_1),\dots,S(i_m)$, where $i_1,\dots,i_m$ are in order of increasing $\xi_{i_1},\dots,\xi_{i_m}$.

Since $\hat{S}^R$ maximizes (5.2), it holds for all $i, j = 1,\dots,m$ that

$$U_{\hat{S}^R(i)}(\xi_i) + U_{\hat{S}^R(j)}(\xi_j) \geq U_{\hat{S}^R(i)}(\xi_j) + U_{\hat{S}^R(j)}(\xi_i);$$

otherwise switching labels for i and j would increase the value of the objective function. As $\hat{S}^R$ is chosen lexicographically, for any i, j such that

$$U_{\hat{S}^R(i)}(\xi_i) + U_{\hat{S}^R(j)}(\xi_j) = U_{\hat{S}^R(i)}(\xi_j) + U_{\hat{S}^R(j)}(\xi_i),$$

it holds that $(\hat{S}^R(i) - \hat{S}^R(j))(\xi_i - \xi_j) \geq 0$, with equality if and only if $\xi_i = \xi_j$. Otherwise, switching labels would improve the lexicographic ordering.

Since $\xi_i \neq \xi_j$ for $i \neq j$ except on a set of measure zero, it follows that

$$(\hat{S}^R)^{-1}(a) \succ^{U,a,b} (\hat{S}^R)^{-1}(b), \qquad a, b = 1,\dots,K, a \neq b,$$

where we have let $(\hat{S}^R)^{-1}(a)$ denote $\{\xi_i : \hat{S}^R(\xi_i) = a\}$. As a result, for each a and b we may choose $t_{ab} = t_{ba} \in [0,1]$ such that $(\hat{S}^R)^{-1}(a) \succ^{U,a,b} t_{ab}$ and $(\hat{S}^R)^{-1}(b) \succ^{U,b,a} t_{ba}$, implying that $\hat{S}^R(i) = \hat{\sigma}(\xi_i)$ for some $\hat{\sigma} \in \mathcal{Q}_U$. As parallel arguments hold for $\hat{T}^Q$, the statement of the lemma follows. $\square$

B.5. Proof of Lemma 5.6. Recall the definition of δ_{UV} from Section 5.2, which we can now interpret for arbitrary K according to the definitions of $\mathcal{Q}_U$ and $\mathcal{Q}_V$ in Section B.4 above. We use a symmetrization argument of Hoeffding [Clémençon, Lugosi and Vayatis (2008), Hoeffding (1963)] to bound $\mathbb{E}(\max_{(Q,R)\in\mathcal{Q}_\mu^m\times\mathcal{Q}_\nu^n} \delta_{UV})$. Let $\mathcal{M}_\mathcal{I}$ denote the set of permutations of $1,\dots,m$ which map $1,\dots,m-|\mathcal{I}|$ to $i \notin \mathcal{I}$, and let $\mathcal{M}_\mathcal{J}$ be defined analogously for permutations on $1,\dots,n$. Let $\mathcal{M} = \mathcal{M}_\mathcal{I} \times \mathcal{M}_\mathcal{J}$ and let $Z = |\mathcal{M}|$. Let ξ', ζ' be identically distributed as ξ and ζ, and independent of U and V. Let $\xi_\mathcal{I}$ and $\zeta_\mathcal{J}$ be defined as in Section B.4. To abbreviate the notation, let $g_{\sigma\tau}(x,y) = \omega(x,y)\Gamma_{\sigma(x)\tau(y)}$, and let $\mathcal{Q} = \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n \times \mathcal{Q}_U \times \mathcal{Q}_V$. It holds for $(Q,R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n$ that

$$\mathbb{E}\Big(\max_{Q,R} \delta_{UV}\Big) = \mathbb{E}\Big(\sup_{(Q,R,\sigma,\tau)\in\mathcal{Q}} \big|G_{\sigma,\tau}(\xi,\zeta) - \mathbb{E}\big(G_{\sigma\tau}(\xi',\zeta')\big|U,V\big)\big| \big|\xi_\mathcal{I},\zeta_\mathcal{J}\Big),$$

which by convexity can be upper-bounded by

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{(Q,R,\sigma,\tau)\in\mathcal{Q}} \big|G_{\sigma,\tau}(\xi,\zeta) - G_{\sigma\tau}(\xi',\zeta')\big| \big|\xi_\mathcal{I},\zeta_\mathcal{J}\Big) \\
&\quad = \mathbb{E}\left(\sup_{(Q,R,\sigma,\tau)\in\mathcal{Q}} \left|\frac{1}{mn}\sum_{i\notin\mathcal{I}}\sum_{j\notin\mathcal{J}} g_{\sigma\tau}(\xi_i,\zeta_j) - g_{\sigma\tau}(\xi_i',\zeta_j')\right| \middle| \xi_\mathcal{I},\zeta_\mathcal{J}\right) \\
&\quad = \mathbb{E}\left(\sup_{(Q,R,\sigma,\tau)\in\mathcal{Q}} \left|\frac{|\overline{\mathcal{I}}||\overline{\mathcal{J}}|}{Zmn} \sum_{\pi,\eta\in\mathcal{M}} \frac{1}{\ell}\sum_{i=1}^{\ell} g_{\sigma\tau}(\xi_{\pi(i)},\zeta_{\eta(j)}) \right.\right. \\
&\qquad\qquad\qquad\qquad \left.\left. - g_{\sigma\tau}(\xi_{\pi(i)}',\zeta_{\eta(j)}')\right| \middle| \xi_\mathcal{I},\zeta_\mathcal{J}\right),
\end{aligned}$$

since the permutations π and η weight each (i,j) term equally for $i \notin \mathcal{I}$ and $j \notin \mathcal{J}$; by convexity again, and then linearity of expectation, we have

$$\begin{aligned}
&\leq \mathbb{E}\left(\frac{|\overline{\mathcal{I}}||\overline{\mathcal{J}}|}{Zmn} \sum_{\pi,\eta\in\mathcal{M}} \sup_{(Q,R,\sigma,\tau)\in\mathcal{Q}} \left|\frac{1}{\ell}\sum_{i=1}^{\ell} g_{\sigma\tau}(\xi_{\pi(i)},\zeta_{\eta(j)}) - g_{\sigma\tau}(\xi_{\pi(i)}',\zeta_{\eta(j)}')\right| \middle| \xi_\mathcal{I},\zeta_\mathcal{J}\right) \\
&= \frac{|\overline{\mathcal{I}}||\overline{\mathcal{J}}|}{mn} \mathbb{E}\left(\sup_{(Q,R,\sigma,\tau)\in\mathcal{Q}} \left|\frac{1}{\ell}\sum_{i=1}^{\ell} g_{\sigma\tau}(\xi_i,\zeta_i) - g_{\sigma\tau}(\xi_i',\zeta_i')\right| \middle| \xi_\mathcal{I},\zeta_\mathcal{J}\right).
\end{aligned}$$

We may now introduce independent and identically distributed Rademacher variables $r_1,\dots,r_\ell$, and use standard Rademacher symmetrization arguments [see, e.g., Bousquet, Boucheron and Lugosi (2004)] to show that the final quantity above is equal to

$$\begin{aligned}
&\frac{|\overline{\mathcal{I}}||\overline{\mathcal{J}}|}{mn}\mathbb{E}\left(\sup_{(Q,R,\sigma,\tau)\in\mathcal{Q}}\left|\frac{1}{\ell}\sum_{i=1}^{\ell} r_i\big(g_{\sigma\tau}(\xi_i,\zeta_i)-g_{\sigma\tau}(\xi_i',\zeta_i')\big)\right|\Big|\xi_{\mathcal{I}},\zeta_{\mathcal{J}}\right)\\
&\quad\le\frac{|\overline{\mathcal{I}}||\overline{\mathcal{J}}|}{mn}\mathbb{E}\left(\sup_{(Q,R,\sigma,\tau)\in\mathcal{Q}}\left|\frac{1}{\ell}\sum_{i=1}^{\ell} r_i g_{\sigma\tau}(\xi_i,\zeta_i)\right|+\left|\frac{1}{\ell}\sum_{i=1}^{\ell} r_i g_{\sigma\tau}(\xi_i',\zeta_i')\right|\Big|\xi_{\mathcal{I}},\zeta_{\mathcal{J}}\right)\\
&\quad\le 2\frac{|\overline{\mathcal{I}}||\overline{\mathcal{J}}|}{mn}\mathbb{E}\left(\sup_{(Q,R,\sigma,\tau)\in\mathcal{Q}}\left|\frac{1}{\ell}\sum_{i=1}^{\ell} r_i g_{\sigma\tau}(\xi_i,\zeta_i)\right|\Big|\xi_{\mathcal{I}},\zeta_{\mathcal{J}}\right).
\end{aligned}$$

To bound this expectation, note that for fixed $\mathcal{I}$, $\mathcal{J}$, Q, R (inducing a fixed U and V), and fixed $(\sigma,\tau)\in\mathcal{Q}_U\times\mathcal{Q}_V$, a Hoeffding inequality gives

$$\mathbb{P}\left(\left|\frac{1}{\ell}\sum_{i=1}^{\ell} r_i g_{\sigma\tau}(\xi_i,\zeta_j)\right|\ge\varepsilon\Big|\xi_{\mathcal{I}},\zeta_{\mathcal{J}}\right)\le 2e^{-2\ell\varepsilon^2}. \tag{B.11}$$

We may now apply (B.11) in conjunction with a union bound over all $(Q,R,\sigma,\tau)\in\mathcal{Q}$ as follows. For fixed Q,R,a,b, the set $\{i:\xi_i\succeq^{U,a,b} t_{ab}\}$ can be chosen at most $\ell+1$ ways by varying t_{ab}. As a result, the set $\xi_1,\dots,\xi_\ell$ can be partitioned at most $(\ell+1)^{\binom{K}{2}}$ ways by varying $\sigma\in\mathcal{Q}_U$. Analogously, the set $\zeta_1,\dots,\zeta_\ell$ can be partitioned the same number of ways by varying $\tau\in\mathcal{Q}_V$. For fixed $\mathcal{I}$, $\mathcal{J}$, the functions U and V can be chosen $K^{|\mathcal{I}|+|\mathcal{J}|}$ different ways by varying Q and R. Hence, a union bound gives

$$\mathbb{P}\left(\sup_{(Q,R,\sigma,\tau)\in\mathcal{Q}}\left|\frac{1}{\ell}\sum_{i=1}^{\ell} r_i g_{\sigma\tau}(\xi_i\zeta_i)\right|\ge\varepsilon\Big|\xi_{\mathcal{I}},\zeta_{\mathcal{J}}\right)\le K^{|\mathcal{I}|+|\mathcal{J}|}(\ell+1)^{2\binom{K}{2}}\cdot 2e^{-2\ell\varepsilon^2}.$$

Since this expression is of the form $\mathbb{P}(X\ge t)\le f(t)$ for X nonnegative, we may apply the inequality $\mathbb{E}(X)\le\int_0^\infty\min\{1,f(t)\}\,dt$ to yield

$$\begin{aligned}
&2\frac{|\overline{\mathcal{I}}||\overline{\mathcal{J}}|}{mn}\mathbb{E}\left(\sup_{(Q,R,\sigma,\tau)\in\mathcal{Q}}\left|\frac{1}{\ell}\sum_{i=1}^{\ell} r_i g_{\sigma\tau}(\xi_i,\zeta_i)\right|\Big|\xi_{\mathcal{I}},\zeta_{\mathcal{J}}\right)\\
&\qquad\le 4\sqrt{\frac{(|\mathcal{I}|+|\mathcal{J}|)\log K+2\binom{K}{2}\log(\ell+1)+\log 2}{2\ell}}.
\end{aligned}$$

Since the bound holds for any $\xi_{\mathcal{I}}$, $\zeta_{\mathcal{J}}$, the same bound holds when the conditioning is removed and $\xi_{\mathcal{I}}$, $\zeta_{\mathcal{J}}$ are chosen randomly, thus proving the lemma.

B.6. Proof of Lemma 5.7. To abbreviate notation, let $\mathcal{Q}=\mathcal{Q}_\nu^n\times\mathcal{Q}_U$. Let $r_1,\dots,r_m$ be Rademacher variables as in the proof of Lemma 5.6. By a standard

Rademacher symmetrization,

$$\mathbb{E}\left(\sup_{(R,\sigma)\in\mathcal{Q}}\left\{\max_{1\le a\le K}\left|\sigma^{-1}(a)-\frac{1}{m}\sum_{i=1}^{m}1\{\sigma(\xi_i)=a\}\right|\right\}\Big|\zeta_{\mathcal{J}}\right)$$
$$\le 2\mathbb{E}\left(\sup_{(R,\sigma)\in\mathcal{Q}}\left\{\max_{1\le a\le K}\left|\frac{1}{m}\sum_{i=1}^{m}r_i 1\{\sigma(\xi_i)=a\}\right|\right\}\Big|\zeta_{\mathcal{J}}\right).$$

As in the proof of Lemma 5.6, a Hoeffding inequality and union bound yield

$$\mathbb{P}\left(\sup_{(R,\sigma)\in\mathcal{Q}}\left\{\max_{1\le a\le K}\left|\sigma^{-1}(a)-\frac{1}{m}\sum_{i=1}^{m}1\{\sigma(\xi_i)=a\}\right|\right\}\ge\varepsilon\Big|\zeta_{\mathcal{J}}\right)$$
$$\le K^{|\mathcal{J}|}(m+1)^{\binom{K}{2}}K\cdot 2e^{-2m\varepsilon^2},$$

and applying $\mathbb{E}(|X|)\le\int_0^\infty\min\{1,f(t)\}\,dt$ for $\mathbb{P}(|X|\ge t)\le f(t)$ then gives

$$2\mathbb{E}\left(\sup_{(R,\sigma)\in\mathcal{Q}}\left\{\max_{1\le a\le K}\left|\sigma^{-1}(a)-\frac{1}{m}\sum_{i=1}^{m}1\{\sigma(\xi_i)=a\}\right|\right\}\Big|\zeta_{\mathcal{J}}\right)$$
$$\le 4\sqrt{\frac{(|\mathcal{J}|+1)\log K+\binom{K}{2}\log(m+1)+\log 2}{2m}}.$$

As in the proof of Lemma 5.6, removing the conditioning on $\zeta_{\mathcal{J}}$ does not alter the bound. Parallel arguments apply to $\tau\in\mathcal{Q}_V$, and the lemma follows.

B.7. Covering argument to establish Theorem 4.1. The establishment of Theorem 4.1 from Proposition 5.1 proceeds as follows. For $\mathcal{F}\subset[0,1]^{K\times K}$, recall that $h_{\mathcal{F}}(\Gamma)=\sup_{F\in\mathcal{F}}\langle\Gamma,F\rangle=\sup_{F\in\mathcal{F}}\operatorname{tr}(\Gamma^T F)$. By the Cauchy–Schwarz inequality, $h_{\mathcal{F}}$ is Lipschitz continuous,

$$|h_{\mathcal{F}}(\Gamma)-h_{\mathcal{F}}(\Gamma')|\le\sup_{F\in\mathcal{F}}|\langle\Gamma-\Gamma',F\rangle|\le K\|\Gamma-\Gamma'\|.$$

Let $\mathcal{B}_\varepsilon$ denote an ε-cover in $\|\cdot\|$ for $[-1,1]^{K\times K}$, with $\Gamma^{\mathcal{B}}$ the closest point in $\mathcal{B}_\varepsilon$ to a given Γ. The triangle inequality, Lipschitz condition and $\mathcal{B}_\varepsilon$ imply

$$\begin{aligned}
&\sup_{\Gamma\in[-1,1]^{K\times K}}|h_{\mathcal{F}^A_{\mu\nu}}(\Gamma)-h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma)|\\
&\quad\le\sup_{\Gamma\in[-1,1]^{K\times K}}\{|h_{\mathcal{F}^A_{\mu\nu}}(\Gamma)-h_{\mathcal{F}^A_{\mu\nu}}(\Gamma^{\mathcal{B}})|\\
&\qquad\qquad+|h_{\mathcal{F}^A_{\mu\nu}}(\Gamma^{\mathcal{B}})-h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma^{\mathcal{B}})|+|h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma^{\mathcal{B}})-h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma)|\}\\
&\quad\le\sup_{\Gamma\in[-1,1]^{K\times K}}\{|h_{\mathcal{F}^A_{\mu\nu}}(\Gamma^{\mathcal{B}})-h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma^{\mathcal{B}})|+2K\|\Gamma-\Gamma^{\mathcal{B}}\|\}
\end{aligned}$$

$$\leq \sup_{\Gamma \in [-1,1]^{K\times K}} \left| h_{\mathcal{F}^A_{\mu\nu}}(\Gamma^{\mathcal{B}}) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma^{\mathcal{B}}) \right| + 2K\varepsilon$$

$$= \max_{\Gamma \in \mathcal{B}_\varepsilon} \left| h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma) \right| + 2K\varepsilon.$$

Now let C_K and n_K be defined as in Proposition 5.1, and set $\varepsilon = C_K/n^{1/4}$. It follows by the above relation, a union bound, and Proposition 5.1 that

$$\mathbb{P}\Big(\max_{(\mu,\nu)\in\Omega_{\rho n}\times\Omega_n} \Big\{ \sup_{\Gamma\in[-1,1]^{K\times K}} \left| h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma) \right| \Big\} \geq 3\varepsilon \Big)$$

$$\leq \mathbb{P}\Big(\max_{(\mu,\nu)\in\Omega_{\rho n}\times\Omega_n} \Big\{ \max_{\Gamma\in\mathcal{B}_{\varepsilon/K}} \left| h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma) \right| \Big\} \geq \varepsilon \Big)$$

$$\leq \sum_{(\mu,\nu)\in\Omega_{\rho n}\times\Omega_n} \sum_{\Gamma\in\mathcal{B}_{\varepsilon/K}} \mathbb{P}\big(\left| h_{\mathcal{F}^A_{\mu\nu}}(\Gamma) - h_{\mathcal{F}^\omega_{\mu\nu}}(\Gamma) \right| \geq \varepsilon \big)$$

$$\leq |\Omega_{\rho n}||\Omega_n||\mathcal{B}_{\varepsilon/K}| 2e^{-\sqrt{n}[2\rho/(\rho+1)]}\big[1 + o(1)\big]$$

for all $n \geq n_K$. The result of Theorem 4.1 then follows, since we have that $|\Omega_n| = \binom{n+K-1}{K-1}$, and $\mathcal{B}_{\varepsilon/K}$ can be chosen such that $|\mathcal{B}_{\varepsilon/K}| \leq (1 + K^2/\varepsilon)^{K^2}$.

APPENDIX C: STATEMENT AND PROOF OF LEMMA C.1

To evaluate the excess risk quantities reported in Section 6, we require both that $\phi^* = \operatorname{argmax}_{\phi\in\Phi} L_{\rho_n\omega_\beta}(\phi)$ exist, and that $L_{\rho_n\omega_\beta}(\phi)$ be computable. The following lemma establishes this, using the fact that each $\rho_n\omega_\beta(x, y)$ is a separable function plus a constant. Given a triple $\phi = (\mu, \nu, \theta)$ of two-class blockmodel parameters, it shows that $L_{\rho_n\omega_\beta}(\phi)$, which nominally involves an optimization over all measure-preserving maps of [0, 1], can be reduced to a maximization over four cases, and thus evaluated tractably.

LEMMA C.1. *Given $\mu \in [0, 1]^2$, let $\sigma^{(1)}, \sigma^{(2)} \in \mathcal{Q}_\mu$ denote the mappings*

$$\sigma^{(1)}(x) = \begin{cases} 1, & \text{if } 0 \leq x < \mu_1, \\ 2, & \text{if } \mu_1 \leq x \leq 1; \end{cases} \qquad \sigma^{(2)}(x) = \begin{cases} 1, & \text{if } 1-\mu_2 \leq x \leq 1, \\ 2, & \text{if } 0 \leq x < 1-\mu_2; \end{cases}$$

and let $\tau^{(1)}, \tau^{(2)} \in \mathcal{Q}_\nu$ be defined analogously, given $\nu \in [0, 1]^2$. Given $\phi = (\mu, \nu, \theta)$, terms $L_{\rho_n\omega_\beta}(\phi)$ and $R_{\rho_n\omega_\beta}(\phi)$ from Theorem 3.1 equal

$$L_{\rho_n\omega_\beta}(\phi) = \max_{(i,j)\in\{1,2\}^2} \sum_{a=1}^{2}\sum_{b=1}^{2} \big(\rho_n\omega/\sigma^{(i)}\tau^{(j)}\big)_{ab} \log\Big(\frac{\theta_{ab}}{1-\theta_{ab}}\Big) + \mu_a\nu_b \log(1-\theta_{ab}),$$

$$R_{\rho_n\omega_\beta}(\phi) = \min_{(i,j)\in\{1,2\}^2} \Big\{ \sum_{a=1}^{2}\sum_{b=1}^{2} -2\big(\rho_n\omega/\sigma^{(i)}\tau^{(j)}\big)_{ab}\theta_{ab} + \mu_a\nu_b\theta_{ab}^2 \Big\} + \int_{[0,1]^2} \omega(x, y)^2\, dx\, dy.$$

PROOF. Below we establish the claimed expression for $L_{\rho_n\omega_\beta}(\phi)$; analogous arguments yield the result for $R_{\rho_n\omega_\beta}(\phi)$. First, define $L_\omega(\phi;\sigma,\tau)$ as

$$L_\omega(\phi;\sigma,\tau)=\sum_{a=1}^{2}\sum_{b=1}^{2}(\omega/\sigma\tau)_{ab}\log\left(\frac{\theta_{ab}}{1-\theta_{ab}}\right)+\mu_a\nu_b\log(1-\theta_{ab}).$$

Next, let $\sigma^*|\tau=\operatorname{argmax}_{\sigma\in\mathcal{Q}_\mu}L_\omega(\phi;\sigma,\tau)$, with the convention that $\operatorname{argmax}_{\sigma\in\mathcal{Q}_\mu}(\cdot)$ is undefined if no maximizer exists. We then see that

$$\begin{aligned}\sigma^*|\tau&=\operatorname*{argmax}_{\sigma\in\mathcal{Q}_\mu}\sum_{a=1}^{2}\sum_{b=1}^{2}(\omega/\sigma\tau)_{ab}\log\left(\frac{\theta_{ab}}{1-\theta_{ab}}\right)\\&=\operatorname*{argmax}_{\sigma\in\mathcal{Q}_\mu}\sum_{a=1}^{2}\sum_{b=1}^{2}\int_{\sigma^{-1}(a)\times\tau^{-1}(b)}\omega(x,y)\,dx\,dy\log\left(\frac{\theta_{ab}}{1-\theta_{ab}}\right)\\&=\operatorname*{argmax}_{\sigma\in\mathcal{Q}_\mu}\sum_{a=1}^{2}\int_{\sigma^{-1}(a)}g_a(x)\,dx\\&\qquad\qquad\text{with } g_a(x)=\sum_{b=1}^{2}\int_{\tau^{-1}(b)}\omega(x,y)\log\left(\frac{\theta_{ab}}{1-\theta_{ab}}\right)dy\\&=\operatorname*{argmax}_{\sigma\in\mathcal{Q}_\mu}\int_{[0,1]}g_2(x)\,dx+\int_{[0,1]}\bigl(g_1(x)-g_2(x)\bigr)1\{\sigma(x)=1\}\,dx.\end{aligned}$$

It can be seen that $\sigma^*|\tau$ is always defined and assigns the μ_1-quantile of $g_1(x)-g_2(x)$ to class 1. Since $\rho_n\omega_\beta(x,y)=\rho_n(f_\beta(x)f_\beta(y)+1/2)$, $g_1(x)-g_2(x)$ is affine in $f_\beta(x)$, and can be written as $mf_\beta(x)+c$ for some scalars m and c. As f_β is monotone, the μ_1-quantile will either be $[0,\mu_1]$ or $[1-\mu_1,1]$—depending on the sign of m—meaning that $\sigma^*|\tau$ equals either $\sigma^{(1)}$ or $\sigma^{(2)}$ for any τ. Analogously, $\tau^*|\sigma$ equals either $\tau^{(1)}$ or $\tau^{(2)}$ for any σ. Hence,

$$\begin{aligned}L_{\rho_n\omega_\beta}(\phi)&=\sup_{\sigma\in\mathcal{Q}_\mu}\sup_{\tau\in\mathcal{Q}_\nu}L_{\rho_n\omega_\beta}(\phi;\sigma,\tau)\\&=\sup_{\sigma\in\mathcal{Q}_\mu}L_{\rho_n\omega_\beta}\bigl(\phi;\sigma,(\tau^*|\sigma)\bigr)\le\sup_{\sigma\in\mathcal{Q}_\mu}L_{\rho_n\omega_\beta}\bigl(\phi;\bigl(\sigma^*|(\tau^*|\sigma)\bigr),(\tau^*|\sigma)\bigr)\\&=\max_{(i,j)\in\{1,2\}^2}L_{\rho_n\omega_\beta}\bigl(\phi;\sigma^{(i)},\tau^{(j)}\bigr).\end{aligned}$$

Thus $L_{\rho_n\omega_\beta}(\phi)$, which nominally involves a supremum over every pair $(\sigma,\tau)\in\mathcal{Q}_\mu\times\mathcal{Q}_\nu$, is reduced to a maximization over $\sigma^{(1)},\sigma^{(2)}$ and $\tau^{(1)},\tau^{(2)}$. □

COROLLARY C.1. *The quantity* $\sup_{\phi\in\Phi}L_{\rho_n\omega_\beta}(\phi)$ *is achieved by* $\phi^{(ij)}=(\mu,\nu,\omega/\sigma^{(i)}\tau^{(j)})$ *for some* $(\mu,\nu)\in\Omega_m\times\Omega_n$ *and* $(i,j)\in\{1,2\}^2$.

PROOF. For any $\phi = (\mu, \nu, \theta) \in \Phi$, it holds that

$$
\begin{aligned}
L_{\rho_n\omega_\beta}(\phi) &= \max_{(i,j)\in\{1,2\}^2} L_{\rho_n\omega_\beta}(\phi; \sigma^{(i)}, \tau^{(j)}) \\
&\le \max_{(i,j)\in\{1,2\}^2} L_{\rho_n\omega_\beta}(\phi^{(ij)}; \sigma^{(i)}, \tau^{(j)}) \\
&= \max_{(i,j)\in\{1,2\}^2} L_{\rho_n\omega_\beta}(\phi^{(ij)}),
\end{aligned}
$$

where the first line holds by Lemma C.1, the second because $p\log x + (1-p)\log(1-x)$ is maximized over $0 \le x \le 1$ by $x = p$, and the third by the definition of $L_{\rho_n\omega_\beta}(\cdot;\cdot,\cdot)$. $\square$

Acknowledgment. The first author wishes to thank Peter Bickel for helpful advice and feedback.

REFERENCES

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

ALON, N., FERNANDEZ DE LA VEGA, W., KANNAN, R. and KARPINSKI, M. (2003). Random sampling and approximation of MAX-CSPs. *J. Comput. System Sci.* **67** 212–243. MR2022830

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 2280–2301. MR2906868

BORGS, C., CHAYES, J., LOVÁSZ, L., SÓS, V. T., SZEGEDY, B. and VESZTERGOMBI, K. (2006). Graph limits and parameter testing. In *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing* 261–270. ACM, New York. MR2277152

BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T. and VESZTERGOMBI, K. (2008). Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. *Adv. Math.* **219** 1801–1851. MR2455626

BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T. and VESZTERGOMBI, K. (2012). Convergent sequences of dense graphs. II. Multiway cuts and statistical physics. *Ann. of Math.* (2) **176** 151–219. MR2925382

BOUSQUET, O., BOUCHERON, S. and LUGOSI, G. (2004). Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning* (O. Bousquet, U. von Luxburg and G. Rätsch, eds.) 169–207. Springer, Berlin.

CHATTERJEE, S. (2012). Matrix estimation by universal singular value thresholding. Preprint. Available at arXiv:1212.1247.

CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253

CLÉMENÇON, S., LUGOSI, G. and VAYATIS, N. (2008). Ranking and empirical minimization of U-statistics. *Ann. Statist.* **36** 844–874. MR2396817

DIACONIS, P. and JANSON, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (7) **28** 33–61. MR2463439

FIENBERG, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *J. Comput. Graph. Statist.* **21** 825–839. MR3005799

FISHKIND, D. E., SUSSMAN, D. L., TANG, M., VOGELSTEIN, J. T. and PRIEBE, C. E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* **34** 23–39. MR3032990

FLYNN, C. J. and PERRY, P. O. (2012). Consistent biclustering. Preprint. Available at arXiv:1206.6927.

FORTUNATO, S. and BARTHÉLEMY, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA* **104** 36–41.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

HOFF, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computat. Math. Org. Theory* **15** 261–272.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262

KIM, M. and LESKOVEC, J. (2012). Multiplicative attribute graph model of real-world networks. *Internet Math.* **8** 113–160. MR2900491

MILLER, K. T., GRIFFITHS, T. L. and JORDAN, M. I. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems* 22 (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta, eds.) 1276–1284. MIT Press, Cambridge, MA.

NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8582.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

ROHE, K. and YU, B. (2012). Co-clustering for directed graphs: The stochastic co-blockmodel and a spectral algorithm. Preprint. Available at arXiv:1204.2296.

SCHNEIDER, R. (1993). *Convex Bodies: The Brunn–Minkowski Theory. Encyclopedia of Mathematics and Its Applications* **44**. Cambridge Univ. Press, Cambridge. MR1216521

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

ZHAO, Y., LEVINA, E. and ZHU, J. (2011). Community extraction for social networks. *Proc. Natl. Acad. Sci. USA* **108** 7321–7326.

ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083

HEINZ COLLEGE OF PUBLIC POLICY AND MANAGEMENT
CARNEGIE MELLON UNIVERSITY
5000 FORBES AVE, HAMBURG HALL
PITTSBURGH, PENNSYLVANIA 15213-3890
USA
E-MAIL: davidch@andrew.cmu.edu

DEPARTMENT OF STATISTICAL SCIENCE
UNIVERSITY COLLEGE LONDON
GOWER STREET
LONDON WC1E 6BT
UNITED KINGDOM
E-MAIL: p.wolfe@ucl.ac.uk