SIAM J. MATH. DATA SCI.
Vol. 2, No. 4, pp. 1034–1065

Moments of Uniform Random Multigraphs with Fixed Degree Sequences*

Philip S. Chodrow†

Abstract. We study the expected adjacency matrix of a uniformly random multigraph with fixed degree sequence $\mathbf{d} \in \mathbb{Z}_+^n$. This matrix arises in a variety of analyses of networked data sets, including modularity-maximization and mean-field theories of spreading processes. Its structure is well understood for large, sparse, simple graphs: the expected number of edges between nodes i and j is roughly $\frac{d_i d_j}{\sum_\ell d_\ell}$. Many network data sets are neither large, sparse, nor simple, and in these cases the standard approximation no longer applies. We derive a novel estimator using a dynamical approach: the estimator emerges from the stationarity conditions of a class of Markov Chain Monte Carlo algorithms for graph sampling. We derive error bounds for this estimator and provide an efficient scheme with which to compute it. We test the estimator on synthetic and empirical degree sequences, finding that it enjoys relative error against ground truth a full order of magnitude smaller than the standard approximation. We then compare modularity maximization techniques using both the standard and novel estimators, finding that the qualitative structure of the optimization landscape depends significantly on the estimator choice. Our results emphasize the importance of using carefully specified random graph models in data scientific applications.

Key words. random graphs, social networks, Markov Chain Monte Carlo, community structure, estimation

AMS subject classifications. 05C80, 05C82, 91D30, 62-07, 65C05

DOI. 10.1137/19M1288772

1. Introduction.

The language of graphs offers a standard formalism for representing systems of interrelated objects or agents. Simple graphs model agents connected by a single, usually static, relation, such as acquaintanceship, proximity, or similarity. In many data sets, however, agents are linked by multiple, discrete interactions. Two agents in a contact network may be in spatial proximity multiple times in the study period. Two agents in a communication network may exchange many emails over the course of a week. In an academic collaboration network, the same two authors may be jointly involved in tens or even hundreds of papers. In such cases, it is natural to draw a distinct edge between agents for each interaction event. Doing so results in a multigraph, in which any two nodes may be linked by an arbitrary, nonnegative, integer-valued number of edges.

A fundamental tool in network data science is null model comparison, which allows the analyst to evaluate whether a feature observed in a given network is surprising when compared to benchmark expectations. We therefore often compare observed networks against random graph null models—probability distributions over graphs. An especially common class of null

*Received by the editors September 23, 2019; accepted for publication (in revised form) August 4, 2020; published electronically October 20, 2020.
https://doi.org/10.1137/19M1288772
Funding: The author acknowledges support from the National Science Foundation under Graduate Research Fellowship grant 1122374.

†Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139 USA, and Department of Mathematics, University of California Los Angeles, Los Angeles, CA 90095 USA (phil@math.ucla.edu).

models is obtained by fixing the degree sequence $\mathbf{d}$ of the observed network, which encodes the number of interactions for each node. The degree sequence is known to constrain many of a network's macroscopic properties [41]. The least informative (or entropy-maximizing) distribution so obtained is the uniform distribution on the space of graphs with the specified degree sequence. The same construction goes through for multigraphs. When studying interaction networks, the corresponding random graph is the uniform distribution $\eta_{\mathbf{d}}$ on the set $\mathcal{G}_{\mathbf{d}}$ of multigraphs with degree sequence $\mathbf{d}$.

In many applications, a set of complete samples from $\eta_{\mathbf{d}}$ is not required—only some selected moments. An especially important set of moments is summarized by the expected adjacency matrix. We therefore consider the following question: if $\mathbf{W}$ is the (random) adjacency matrix of multigraph $G \sim \eta_{\mathbf{d}}$, what is the value of the expected adjacency matrix $\mathbf{\Omega} \triangleq \mathbb{E}[\mathbf{W}]$? The entry ω_{ij} of $\mathbf{\Omega}$ gives the expected number of edges between nodes i and j. These moments have several important applications in network science. Among these is community detection via modularity-maximization [38], which in many formulations includes a term for the expected number of edges between nodes under a suitably specified null model. Despite its simplicity and relevance for applications, this problem has received relatively little mathematical attention.

Before surveying existing approaches to the estimation of $\mathbf{\Omega}$, we fix some notation. Let $\mathcal{G}_{\mathbf{d}}$ refer to the set of multigraphs without self-loops with degree sequence $\mathbf{d} \in \mathbb{Z}_+^n$. From a modeling perspective, the exclusion of self-loops reflects an assumption that agents do not meaningfully interact with themselves. An element $G \in \mathcal{G}_{\mathbf{d}}$ has a fixed number n of nodes and $m = \frac{1}{2}\sum_i d_i$ of edges. We use bold uppercase symbols to denote matrices, bold lowercase symbols to denote vectors, and standard symbols to denote scalars. We do not notationally distinguish deterministic and random objects, instead relying on their associated definitions. We use Greek letters to denote expectations of random objects. An estimator of a quantity, either deterministic or stochastic, is distinguished by a hat. For example, $\mathbf{\Omega} = \mathbb{E}[\mathbf{W}]$ is the expectation of $\mathbf{W}$. An estimator of $\mathbf{\Omega}$, either deterministic or stochastic, may be written $\hat{\mathbf{\Omega}}$.

One approach to estimating $\mathbf{\Omega}$ is Monte Carlo sampling. We sample s independent and identically distributed (i.i.d.) samples $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(s)} \sim \eta_{\mathbf{d}}$ and construct the estimator

$$\hat{\mathbf{\Omega}}^{\mathrm{mc}}(\mathbf{d}) \triangleq \frac{1}{s}\sum_{\ell=1}^{s} \mathbf{W}^{(\ell)} . \tag{1.1}$$

The estimator $\hat{\mathbf{\Omega}}^{\mathrm{mc}}$ is a random function of $\mathbf{d}$, parameterized by the sample size s. The strong law of large numbers (SLLN) ensures that $\hat{\mathbf{\Omega}}^{\mathrm{mc}} \to \mathbf{\Omega}$ almost surely as the number of samples s grows large. Stronger results are possible: since each entry $\hat{\omega}_{ij}^{\mathrm{mc}}$ is bounded, the variance of $\hat{\omega}_{ij}^{\mathrm{mc}}$ is finite, and we can apply the central limit theorem to provide quantitative bounds on the convergence rate. This attractive picture is marred by a severe computational inconvenience: the size and complex combinatorial structure of $\mathcal{G}_d$ make exact sampling intractable. Markov Chain Monte Carlo (MCMC) methods [25] are therefore required. MCMC methods introduce a new complication: for any finite number of iterations, the samples produced will always be statistically dependent and may therefore over-represent some regions of $\mathcal{G}_{\mathbf{d}}$ and under-represent others. This dependence breaks the guarantees provided by the SLLN or central limit theorem. Control over the *mixing time* of the sampler is in principle sufficient to

ameliorate this issue; however, there are few known mixing time bounds on MCMC samplers of distributions on $\mathcal{G}_{\mathbf{d}}$. Available upper bounds on the mixing times [28, 29, 22] are too large for guarantees in many practical computations, and there are heuristic reasons to believe that there are limits on our ability to improve these bounds.

An alternative estimator $\hat{\mathbf{\Omega}}^0$, extremely common in the network science literature, is defined entrywise by a simple formula:

$$\hat{\omega}^0_{ij}(\mathbf{d}) = f_{ij}(\mathbf{d}) \triangleq \begin{cases} \frac{d_i d_j}{2m}, & i \neq j, \\ 0, & i = j. \end{cases} \tag{1.2}$$

The function f_{ij} plays an important role throughout this article. Unlike $\hat{\mathbf{\Omega}}^{\text{mc}}$, $\hat{\mathbf{\Omega}}^0$ is a deterministic function of $\mathbf{d}$ that is essentially free to compute. The functional form of $f_{ij}(\mathbf{d})$ can be derived in multiple ways. For example, it is the expected edge density between distinct nodes i and j in the model of Chung and Lu [17, 18] which preserves $\mathbf{d}$ in expectation rather than deterministically. We will therefore refer to (1.2) as the "CL estimate" after Chung and Lu, though we emphasize that these authors did not use this expression as an estimator for any of the models we consider here and indeed restricted their attention to graphs without parallel edges. The estimator $\hat{\mathbf{\Omega}}^0$ was also derived heuristically by Newman and Girvan when they introduced modularity maximization as a method for community detection in networks [40, 38]. In their derivation, we approximate the number of edges between i and j as follows. Node i has d_i edges. Each of these edges must connect to one of the $n-1$ other nodes. A "random edge" is attached to node j with probability roughly $\frac{d_j}{2m-d_i}$. Assuming that $d_i \ll 2m$ yields $\hat{\omega}^0_{ij}$ as an approximation. Importantly, this heuristic argument does not formalize any probability measure over a set of graphs. Thus, although $\hat{\mathbf{\Omega}}^0$ is sometimes described as the expectation of a "random graph with fixed degree sequence," this is not exactly true for any common models except that of Chung and Lu, in which degrees are fixed only in expectation. In particular, $\hat{\mathbf{\Omega}}^0$ possesses no guarantees related to its performance as an estimator for the uniform model $\eta_{\mathbf{d}}$, the most literal mathematical operationalization of the phrase "random graph with fixed degree sequence." As we will see, this performance can indeed be quite poor on data sets with high edge densities.

In this article, we construct an estimator of $\mathbf{\Omega}$ for dense multigraphs that is both scalable and accurate. By treating an MCMC sampler for $\eta_{\mathbf{d}}$ as a stochastic dynamical system whose state space is $\mathcal{G}_{\mathbf{d}}$, we derive stationarity conditions describing the desired moments. As we will show, there exists a vector $\boldsymbol{\beta} \in \mathbb{R}^n_+$ such that χ_{ij}, the probability that $w_{ij} \geq 1$, is given by

$$\chi_{ij} \triangleq \eta_{\mathbf{d}}(w_{ij} \geq 1) \approx \frac{\beta_i \beta_j}{\sum_i \beta_i} = f_{ij}(\boldsymbol{\beta})$$

for all $i \neq j$. The function f_{ij} in this approximation is the same as that which appears in the definition of the CL estimator in (1.2). Furthermore, the entries of $\mathbf{\Omega}$ are given approximately by

$$\omega_{ij} \approx \frac{\chi_{ij}}{1 - \chi_{ij}} .$$

Taken together, these two formulae provide a method for computing an estimate of $\boldsymbol{\Omega}$ given knowledge of the vector $\boldsymbol{\beta}$. We construct an estimator $\hat{\boldsymbol{\beta}}$ of this vector by solving the system of n nonlinear equations

$$\sum_j \frac{f_{ij}(\boldsymbol{\beta})}{1 - f_{ij}(\boldsymbol{\beta})} = d_i \,, \quad i = 1, \ldots, n \,.$$

We show that the solution to this equation, provided it exists, is unique within the realm of interpretable sequences $\boldsymbol{\beta}$ subject to mild regularity conditions, and that this solution can be found efficiently by a simple, iterative algorithm. From $\hat{\boldsymbol{\beta}}$ we construct an estimator $\hat{\boldsymbol{\Omega}}^1$ of $\boldsymbol{\Omega}$. As we show, this estimator is both easier to compute than $\hat{\boldsymbol{\Omega}}^{\mathrm{mc}}$ and much more accurate than $\hat{\boldsymbol{\Omega}}^0$. Furthermore, we can view the CL estimator $\hat{\boldsymbol{\Omega}}^0$ as an approximation of $\hat{\boldsymbol{\Omega}}^1$, obtained from the latter via a sequence of two linear approximations.

1.1. Outline. In section 2, we review two important null multigraph models—the configuration model and the uniform model—as well as a unified MCMC algorithm for sampling from each. The analysis of this algorithm forms the heart of our derivation of the estimate $\hat{\boldsymbol{\Omega}}^1$ in section 3. This estimator depends on the unknown vector $\boldsymbol{\beta}$, which must be learned from $\mathbf{d}$. We offer a simple scheme for doing so in section 4, including a qualified uniqueness guarantee on the resulting estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$; a description of its structure; and a numerical scheme for computing it efficiently. In section 5 we turn to experiments. We first study the behavior of our methods on two synthetic data sets, including a bootstrap-style test of the conjecture underlying our error bounds. We then check the accuracy of $\hat{\boldsymbol{\Omega}}^1$ on a subset of a high school contact network. Whereas $\hat{\boldsymbol{\Omega}}^0$ is significantly biased on this data set, $\hat{\boldsymbol{\Omega}}^1$ is nearly unbiased and decreases the mean relative error of the estimate by an order of magnitude. In our final experiment, we study the behavior of modularity maximization when the standard null expectation $\hat{\boldsymbol{\Omega}}^0$ is replaced by $\hat{\boldsymbol{\Omega}}^1$. We find that the behavior of a multiway spectral algorithm [50] depends strongly on both the choice of null expectation and the data set under study. We close in section 6 with a discussion and suggestions for future work.

2. Random graphs with fixed degree sequences.

Our interest will focus on the uniform model $\eta_{\mathbf{d}}$, but it will be useful to draw comparisons to the somewhat more commonly used configuration model [11].

Definition 2.1 (configurations). *For a fixed node set N and degree sequence $\mathbf{d} \in \mathbb{Z}_+^n$, let*

$$\Sigma_{\mathbf{d}} = \biguplus_{i=1}^{n} \{i_1, \ldots, i_{d_i}\} \,,$$

where $\uplus$ denotes multiset union. Thus, $\Sigma_{\mathbf{d}}$ contains d_i labeled copies of each node i. The copies $i_1, \ldots, i_{d_i}$ are called stubs *of node i. A* configuration *$C = (N, E)$ consists of the node set N and an edge set E which partitions $\Sigma_{\mathbf{d}}$ into unordered pairs. An edge in E of the form $\{i_k, i_\ell\}$ is called a* self-loop. *The process of forming C from $\Sigma_{\mathbf{d}}$ is often called* stub-matching.

Let $\mathcal{C}_{\mathbf{d}} \subset \Sigma_{\mathbf{d}}$ be the set of all configurations with degree sequence $\mathbf{d}$ that do not include any self-loops. There is a natural surjection $g : \mathcal{C}_{\mathbf{d}} \to \mathcal{G}_{\mathbf{d}}$. The image of $C \in \mathcal{C}_{\mathbf{d}}$ under g is

obtained by replacing all stubs with their corresponding nodes and consolidating the result as a multiset. The uniform distribution on $\mathcal{C}_{\mathbf{d}}$ induces a distribution on $\mathcal{G}_{\mathbf{d}}$ via g. Denote by $g^{-1} : \mathcal{G}_{\mathbf{d}} \to 2^{\mathcal{C}_{\mathbf{d}}}$ the function that assigns to each element of $\mathcal{G}_{\mathbf{d}}$ its preimage in $\mathcal{C}_{\mathbf{d}}$ under g.

Definition 2.2 (configuration model). *Let $\lambda_{\mathbf{d}}$ be the uniform distribution on $\mathcal{C}_{\mathbf{d}}$. The* configuration model *on $\mathcal{G}_{\mathbf{d}}$ is the distribution $\mu_{\mathbf{d}} = \lambda_{\mathbf{d}} \circ g^{-1}$.*

The distinction between $\eta_{\mathbf{d}}$ and $\mu_{\mathbf{d}}$—and its implications for data analysis—were recently highlighted by Fosdick et al. [25]. We have diverged from the terminology of the authors: our "uniform model" is their "configuration model on non-loopy, vertex-labeled multigraphs," and our "configuration model" is their "configuration model on non-loopy, stub-labeled multigraphs."

The distinction between uniform and configuration models lies in how they weight graphs with parallel edges. Let C_1 and C_2 be two configurations. Suppose that C_1 contains the matchings $(i_1, j_1), (i_2, j_2)$ and C_2 contains the matchings $(i_1, j_2), (i_2, j_1)$, and that they otherwise agree on all other stubs. Let $G = g(C_1) = g(C_2)$. Under the uniform model, G is considered to be a single state, weighted equally with all other states. Under the configuration model, on the other hand, the probability mass placed on G is proportional to $\left|g^{-1}(G)\right|$, reflecting both C_1 and C_2 as distinct states. In particular, the configuration model $\mu_{\mathbf{d}}$ will tend to place higher probabilistic weight on elements of $\mathcal{G}_{\mathbf{d}}$ with large numbers of parallel edges than will the uniform model $\eta_{\mathbf{d}}$.

In the absence of parallel edges, the uniform and configuration models are closely related. Let A be the event that G is simple, without self-loops or parallel edges. Then, it is direct to show (e.g., [11]) that, for all G, $\eta_{\mathbf{d}}(G|A) = \mu_{\mathbf{d}}(G|A)$. The reason is that, when G is simple, the sizes of the preimages $g^{-1}(G)$ depend only on the degree sequence $\mathbf{d}$. Since $\mathbf{d}$ is fixed in $\mathcal{G}_{\mathbf{d}}$, these preimages all have the same size. Thus, when a *simple* random graph is required, the uniform model $\eta_{\mathbf{d}}$ and configuration model $\mu_{\mathbf{d}}$ are in principle interchangeable in the sense that we can sample from $\eta_{\mathbf{d}}(\cdot|A)$ by repeatedly sampling from $\mu_{\mathbf{d}}$ until a simple graph is produced. Furthermore, when the degree sequence $\mathbf{d}$ grows slowly relative to n, $\mu_{\mathbf{d}}(A)$ is bounded away from zero by a function that depends on moments of $\mathbf{d}$ when n grows large [11, 36, 3]. This in turn provides an upper bound on the expected number of samples from $\mu_{\mathbf{d}}$ required to produce a single sample from $\eta_{\mathbf{d}}(\cdot|A)$. The computational importance of this relationship is that stub-matching for sampling from $\mu_{\mathbf{d}}$ is well understood and often fast.

For dense graphs, $\mu_{\mathbf{d}}(A)$ may be extremely small, and the number of samples required to produce a simple graph may be prohibitive. While it is possible to make post hoc edits to the graph to remove self-loops and multiple edges [37, 44], such methods can generate substantial and uncontrolled bias in finite graphs. Second, and more importantly for our context, there is no equivalence between the unconditional distributions $\eta_{\mathbf{d}}$ and $\mu_{\mathbf{d}}$ on spaces of multigraphs. Stub-matching cannot therefore be used to sample from $\eta_{\mathbf{d}}$ when modeling considerations allow the presence of multiple edges.

Before proceeding, we note that configuration-type models are not the only operationalizations of entropy-maximizing random graphs with flexibly specified degree sequences. An important alternative is the *$\boldsymbol{\beta}$-model*, introduced (in directed form) by [30] and given its present name by [15]. By definition, the $\boldsymbol{\beta}$-model for a given degree sequence $\mathbf{d}$ is the exponential family, defined over graphs, whose sufficient statistic is $\mathbf{d}$. The vector $\boldsymbol{\beta}$ appears as a sequence

of parameters controlling connection rates between edges.[1] It follows from the elementary theory of exponential families that the $\boldsymbol{\beta}$-model is the entropy-maximizing random graph whose degree sequence agrees with $\mathbf{d}$ in expectation [19]. These models possess interesting geometrical structure, which can often be leveraged to give efficient algorithms with useful guarantees [15, 43]. Interestingly for our purposes, the authors of [5] show that, under modest conditions, the densities of sufficiently large subgraphs in the $\boldsymbol{\beta}$-model and a uniformly random simple graph with fixed degree sequence agree in the limit of large graph size (Theorem 1.6). We discuss this connection further in section 6.

2.1. Markov Chain Monte Carlo. Edge-swap Markov chains provide the standard approach to sampling from random graphs and other data structures with deterministically fixed degree sequences. There exists a large constellation of related algorithms, including the sampling of marginal-constrained binary matrices [47, 4]; degree-regular [48, 35, 33] and degree-heterogeneous [14, 46, 9, 20] simple graphs; and graphs with degree-correlation constraints [1]. Most of these algorithms operate by repeatedly swapping edges in such a way as to preserve the required graph structure.

A variant formulated by Fosdick et al. [25] can sample from either the uniform model $\eta_{\mathbf{d}}$ or the configuration model $\mu_{\mathbf{d}}$ on $\mathcal{G}_{\mathbf{d}}$. We define an edge swap to be a random function of two edges that share no nodes.[2] An edge swap interchanges a node on the first edge with a node on the second:

$$\text{EdgeSwap}((i,j),(k,\ell)) = \begin{cases} (i,k),(j,\ell) & \text{with probability } 1/2, \\ (i,\ell),(j,k) & \text{with probability } 1/2. \end{cases}$$

An edge swap does not change the total number of edges incident to nodes i, j, k, and ℓ and therefore preserves $\mathbf{d}$. Starting from a graph $G_0 \in \mathcal{G}_{\mathbf{d}}$, repeated edge swaps can therefore be used to obtain a random sequence of elements of $\mathcal{G}_{\mathbf{d}}$. Since each element of this sequence depends stochastically only on its predecessor, this sequence is a Markov chain. We perform MCMC as follows. At each time step, we select two random edges (i,j) and (k,ℓ), uniformly selected from the set of pairs of edges with four distinct node indices. We then perform a pairwise edge swap of these edges with *acceptance probability*

$$a((i,j),(k,\ell)) \triangleq \begin{cases} 1 & \text{configuration model } \mu_{\mathbf{d}}, \\ (w_{ij}w_{k\ell})^{-1} & \text{uniform model } \eta_{\mathbf{d}}. \end{cases} \tag{2.1}$$

In the case that the edge swap is not accepted, we record the current state again and resample. Algorithm 1 formalizes this procedure. Let P_t denote the set of pairs of edges at time t that share no indices. This set is only empty if G_t can be written as a star graph, which we assume not to be the case.

For sufficiently large sample intervals δt, the output of Algorithm 1 will be approximately i.i.d. according to the target distribution ρ, as guaranteed by the following result.

[1] The "$\boldsymbol{\beta}$" in the $\boldsymbol{\beta}$-model is not the same as the "$\boldsymbol{\beta}$" defined in section 3 and discussed throughout this article, although both play the role of parameter vectors governing degrees and edge densities.

[2] Swaps involving edges that intersect are used when sampling from spaces that include self-loops [25].

Algorithm 1: MCMC sampling for $\eta_{\mathbf{d}}$ and $\mu_{\mathbf{d}}$.

Input: degree sequence $\mathbf{d}$, initial graph $G_0 \in \mathcal{G}_{\mathbf{d}}$, target distribution $\rho \in \{\eta_{\mathbf{d}}, \mu_{\mathbf{d}}\}$, sample interval $\delta t \in \mathbb{Z}_+$, sample size $s \in \mathbb{Z}_+$.

Initialization: $t \leftarrow 0$, $G \leftarrow G_0$
for $t = 1, 2, \ldots, s(\delta t)$ **do**
  sample (i, j) and (k, ℓ) uniformly at random from P_t
  if Uniform$([0, 1]) \leq a((i, j), (k, \ell))$ **then**
    $G_t \leftarrow \text{EdgeSwap}((i, j), (k, \ell))$
  else
    $G_t \leftarrow G_{t-1}$

Output: $\{G_t \text{ such that } t | \delta t\}$

Theorem 2.3 (Fosdick et al. [25]). *The Markov chain $\{G_t\}$ defined by Algorithm 1 is ergodic and reversible with respect to the input distribution ρ. As consequence, samples $\{G_t\}$ generated by Algorithm 1 are asymptotically i.i.d. according to ρ as $\delta t \to \infty$.*

These results provide a principled solution to the problem of asymptotically exact sampling from $\eta_{\mathbf{d}}$ and can therefore be used to construct an estimator $\hat{\mathbf{\Omega}}^{\text{mc}}$ of $\mathbf{\Omega}$, given by (1.1), with arbitrary levels of accuracy. It suffices to let the sample size s and sample interval δt grow large. There are two performance-related issues when using Algorithm 1 in practice, both of which are connected to the number of edges m. First is the question of how large δt should be to ensure that the samples are sufficiently close to independence. Heuristically, δt should scale with the mixing time of the chain, but very few bounds on mixing times for chains of this type appear to be available. In several recent papers, Greenhill [29, 28] and collaborators [22, 26] have derived the only bounds known to this author for edge-swap Markov chains. In the space of simple graphs, under certain regularity conditions on the degree sequence, they provide a mixing time bound with scaling $O({d_*}^{14} m^{10} \log m)$, where $d_* = \max_i d_i$. The scaling of this upper bound is very poor, especially with regard to m, and is therefore not reassuring for practical applications. The introduction of multigraphs also raises difficulties. This author is not aware of any mixing time bounds for distributions defined over multigraphs. However, many multigraphs of practical interest are relatively dense (m large), implying that the associated bounds may be extremely loose. Additionally, these multigraphs often have large numbers of edges w_{ij} between nodes i and j, which will in turn result in low acceptance rates in Algorithm 1. Indeed, supposing that a typical entry W_{ij} scales approximately linearly with m, a typical acceptance probability would scale roughly as m^{-2}. A standard coupon-collector argument shows that it takes roughly $O(m \log m)$ accepted transitions to ensure that each edge has been swapped at least once, which would appear to be a reasonable requirement for a well-mixed chain. We therefore conjecture that the overall mixing time of Algorithm 1 for the uniform model on dense multigraphs is no smaller than $O(m^3 \log m)$, though a more precise statement and proof would be welcome. While much better than the best-known proven results, such a scaling could likely be prohibitive for graphs of even modest size. These considerations suggest that forming the MCMC estimate $\hat{\mathbf{\Omega}}^{\text{mc}}$ may not be a computationally

practical way to estimate $\mathbf{\Omega}$ when m is large. Despite these limitations, Algorithm 1 lies at the heart of our main results in the next section.

3. A dynamical approach to model moments. We introduce some additional notation to facilitate calculations. The transpose of vector $\mathbf{u}$ is denoted by $\mathbf{u}^T$, and the inner product of $\mathbf{u}$ and $\mathbf{v}$ is denoted by $\mathbf{u}^T\mathbf{v}$. We denote the ith row or column of matrix $\mathbf{W}$ by $\mathbf{w}_i$; all matrices we encounter will be symmetric, and so no ambiguity will arise. Let $\mathbf{e}$ be the vector of ones; the dimension of $\mathbf{e}$ will be clear in context. Similarly, let $\mathbf{e}_i$ be the ith standard basis vector. All sums over node indices i, j, k, ℓ have implicit limits from 1 to n. Finally, $a \wedge b$ and $a \vee b$ denote the pairwise minimum and maximum of scalars a and b, respectively.

Algorithm 1 describes a stochastic dynamical update on the space $\mathcal{G}_{\mathbf{d}}$ of multigraphs, which we identify with the space of symmetric matrices with nonnegative integer entries and zero diagonals. Let $\mathbf{\Delta}(t) = \mathbf{W}(t+1) - \mathbf{W}(t)$ be the (random) increment in $\mathbf{W}$ in timestep $t+1$. We implicitly regard $\mathbf{W}$ and $\mathbf{\Delta}$ as functions of t, suppressing the argument for notational concision when there is no possibility of confusion. We can separate $\mathbf{\Delta} = \mathbf{\Delta}^+ - \mathbf{\Delta}^-$, where $\Delta^+_{ij} = (\Delta_{ij} \vee 0)$ and $\Delta^-_{ij} = (-\Delta_{ij} \vee 0)$. The first term Δ^+_{ij} describes the (random) number of edges flowing into the pair (i, j) and the second term Δ^-_{ij} the random number of edges flowing out. Conservation of edges implies that $\sum_{ij} \Delta^+_{ij} = \sum_{ij} \Delta^-_{ij}$. Since a pair of nodes can only gain or lose one edge at a time under the dynamics, the entries Δ^+_{ij} and Δ^-_{ij} are Bernoulli random variables. If a proposed swap is rejected in a given timestep, then all entries of both matrices are zero. If a proposed swap is accepted, then exactly two entries of each matrix are nonzero. Importantly, this implies that the entries of these matrices are not independent. Let $\boldsymbol{\delta}^+ = \mathbb{E}[\mathbf{\Delta}^+]$ and $\boldsymbol{\delta}^- = \mathbb{E}[\mathbf{\Delta}^-]$.

Two things must hold at stationarity of Algorithm 1. First, all moments of $\mathbf{W}$ must be constant in time. Second, since the stationary distribution of Algorithm 1 is the target distribution ρ by construction, these moments of $\mathbf{W}$ are the desired moments of ρ. We can therefore approximately compute moments of ρ by approximately solving conveniently chosen stationarity conditions. A useful set is given by

$$\mathbb{E}[w_{ij}(t+1)^p - w_{ij}(t)^p] = \mathbb{E}\left[(w_{ij}(t) + \Delta_{ij}(t))^p - w_{ij}(t)^p\right] = 0 \tag{3.1}$$

for positive integers p. These equations express directly the time-invariance of the moments $\mathbb{E}[w_{ij}(t)^p]$ at stationarity.

3.1. Illustration: The configuration model. We will derive a version of the CL estimator $\hat{\mathbf{\Omega}}^0$ for the configuration model by studying (3.1) when $p = 1$.

Theorem 3.1. *Under the configuration model $\mu_{\mathbf{d}}$, for all $i \neq j$,*

$$\omega_{ij} = \frac{d_i d_j - \mathbb{E}[\mathbf{w}_i^T \mathbf{w}_j] - \mathbb{E}[w_{ij}^2]}{2m - d_i - d_j}. \tag{3.2}$$

Proof. We first derive expressions for $\mathbf{\Delta}^-$ and $\mathbf{\Delta}^+$ by stepping through the stages of Algorithm 1. Beginning with the former, we have that $\Delta^-_{ij} = 1$ only if edge (i, j) is sampled in the first stage of the iteration. The probability that edges (i, j) and (k, ℓ) are sampled,

assuming that all four indices are distinct, is $z(\mathbf{W})^{-1}W_{ij}W_{k\ell}$, where

$$z(\mathbf{W}) = \sum_{\substack{i,j \\ k,\ell \notin \{i,j\}}} W_{ij}W_{k\ell}$$

gives the total number of ways to pick two edges with four distinct indices. Under the configuration model, $a((i,j),(k,\ell)) = 1$. Summing across k and ℓ and taking expectations, we obtain

$$\begin{aligned}\delta_{ij}^{-} &= \frac{1}{z(\mathbf{W})}\mathbb{E}\left[\sum_{k,\ell\notin\{i,j\}} w_{ij}w_{k\ell}\right] \\ &= \frac{1}{z(\mathbf{W})}\mathbb{E}\left[w_{ij}\left(\sum_{k,\ell} w_{k\ell} - \sum_{k}(w_{ki}+w_{kj}) - \sum_{\ell}(w_{i\ell}+w_{j\ell}) + 3w_{ij}\right)\right].\end{aligned}$$

Recalling constraints such as $\sum_{k,\ell} W_{k\ell} = 2m$ and $\sum_k W_{ki} = d_i$, this expression simplifies to

$$\delta_{ij}^{-} = \frac{1}{z(\mathbf{W})}\left(2\omega_{ij}(m - d_i - d_j) + 3\mathbb{E}[w_{ij}^2]\right).$$

We can derive a similar expression for δ_{ij}^{+}. Fix two additional indices k and ℓ, such that all four indices i,j,k,ℓ are distinct. A new edge (i,j) can be generated from selecting for swap either of the pairs $\{(i,k),(\ell,j)\}$ or $\{(i,\ell),(k,j)\}$. These events occur with probabilities $z(\mathbf{W})^{-1}w_{ik}w_{\ell j}$ and $z(\mathbf{W})^{-1}w_{i\ell}w_{kj}$, respectively. Having selected edges $\{(i,k),(\ell,j)\}$, edges $\{(i,j),(k,\ell)\}$ are formed by the swap with probability $\frac{1}{2}$; otherwise $\{(i,\ell),(k,j)\}$ are formed. Summing across k and ℓ and computing expectations, we have

$$\begin{aligned}\delta_{ij}^{+} &= \frac{1}{2z(\mathbf{W})}\mathbb{E}\left[\sum_{\substack{k,\ell\notin\{i,j\}\\ k\neq\ell}} w_{ik}w_{\ell j} + \sum_{\substack{k,\ell\notin\{i,j\}\\ k\neq\ell}} w_{i\ell}w_{kj}\right] \\ &= \frac{1}{z(\mathbf{W})}\mathbb{E}\left[\sum_{\substack{k,\ell\notin\{i,j\}\\ k\neq\ell}} w_{ik}w_{\ell j}\right] \\ &= \frac{1}{z(\mathbf{W})}\mathbb{E}\left[\left(\sum_k w_{ik}\right)\left(\sum_\ell w_{\ell j}\right) - w_{ij}\sum_k(w_{ik}+w_{jk}) - \sum_k w_{ik}w_{kj} + w_{ij}^2\right] \\ &= \frac{1}{z(\mathbf{W})}\left(d_i d_j \ - \omega_{ij}(d_i+d_j) - \mathbb{E}[\mathbf{w}_i^T\mathbf{w}_j] + \mathbb{E}[w_{ij}^2]\right).\end{aligned}$$

Choosing $p = 1$ in (3.1), we must have $\delta_{ij}^{+} = \delta_{ij}^{-}$ at stationarity. Inserting our derived expressions and solving for ω_{ij} yields the result. ∎

Theorem 3.1 does not give an explicit operational solution for ω_{ij}, since the right-hand side contains higher moments of $\mathbf{W}$. Progress can be made in the "large, sparse regime," in which we assume that n is large and the entries of $\mathbf{W}$ and $\mathbf{d}$ small relative to m. Recalling that $\hat{\omega}^0_{ij} = \frac{d_i d_j}{2m}$, we can rewrite (3.2) as

$$\omega_{ij} = \left(1 - \frac{d_i + d_j}{2m}\right)^{-1} \left(1 - \frac{\mathbb{E}[\mathbf{w}_i^T \mathbf{w}_j] - \mathbb{E}[w_{ij}^2]}{d_i d_j}\right) \hat{\omega}^0_{ij}.$$

In the large, sparse heuristic, each entry of $\mathbf{d}$ is small in comparison to m, and the first error factor is near unity. Similarly, the expression $\mathbb{E}[\mathbf{w}_i^T \mathbf{w}_j] - \mathbb{E}[w_{ij}^2]$ implicitly contains up to $n-1$ nonzero products of entries of $\mathbf{W}$. On the other hand, the denominator contains $(n-1)^2$ such terms, and we therefore expect the second error factor to also lie near unity. We therefore expect that $\omega_{ij} \to \hat{\omega}^0_{ij}$ "in the large, sparse regime." This statement can be made precise by specifying the asymptotic behavior of $\mathbf{W}$ with respect to n, which is beyond our present scope. Through analysis of Algorithm 1, we have derived both $\hat{\mathbf{\Omega}}^0$ and explicit error terms that are often elided in the network science literature.

3.2. Moments of the uniform model. The analysis of the uniform model is somewhat more subtle. As seen in the configuration model above, many of the sums that appear in the calculations of $\boldsymbol{\delta}^+$ and $\boldsymbol{\delta}^-$ reduced to fixed constants due to the degree constraints. Unfortunately, there is no analogous simplification in the uniform model $\eta_{\mathbf{d}}$. Because of this, we require some additional technology in order to make progress.

Define the binary matrix $\mathbf{X} \in \{0,1\}^{n \times n}$ entrywise by $x_{ij} = \mathbb{1}(w_{ij} \geq 1)$. For convenience, we adopt the convention $0/0 = 0$ under which the identity $w_{ij}/w_{ij} = x_{ij}$ holds even when $w_{ij} = 0$. We can interpret $\mathbf{X}$ as the adjacency matrix of the simple graph obtained by collapsing all sets of parallel edges in a multigraph into single edges. Let $\mathbf{b} = \mathbf{X}\mathbf{e}$, the vector of row sums of $\mathbf{X}$. The vector $\mathbf{b}$ is interpretable as the collapsed degree sequence, whose ith entry gives the number of distinct neighbors of node i. Let $y = \frac{1}{2}\mathbf{e}^T\mathbf{b}$ give the total number of collapsed edges. The expectations of $\mathbf{X}$, $\mathbf{b}$, and y play important roles in our analysis. We denote them

$$\boldsymbol{\chi} = \mathbb{E}[\mathbf{X}]\,, \quad \boldsymbol{\beta} = \mathbb{E}[\mathbf{b}]\,, \text{ and } \quad \psi = \mathbb{E}[y]\,.$$

The objects $\boldsymbol{\chi}$, $\boldsymbol{\beta}$, and ψ are all implicitly deterministic functions of $\mathbf{d}$. Throughout this section, we let $I = (i_1, \ldots, i_p)$ be a set of p not-necessarily-distinct indices and let $K = ((k_1, \ell_1), \ldots, (k_q, \ell_q))$ be a set of q not-necessarily-distinct dyadic indices. If $\mathbf{v}$ is a vector, we let $\mathbf{v}_I$ denote the vector with entries $(v_{i_1}, \ldots, v_{i_p})$. Similarly, if $\mathbf{A}$ is a matrix, we let $\mathbf{A}_K$ denote the vector with entries $(a_{k_1,\ell_1}, \ldots, a_{k_q,\ell_q})$.

We first require control over the behavior of $\boldsymbol{\beta}$ with respect to $\mathbf{d}$.

Definition 3.2 (regularity constant for $\boldsymbol{\beta}$). *Let $u(\mathbf{d})$ be the smallest real number such that, for any degree sequence $\mathbf{d}'$ such that $\mathbf{d}' \geq \mathbf{d}$ entrywise, and for all distinct indices i and j,*

$$\left\|\boldsymbol{\beta}(\mathbf{d}' + \mathbf{e}_i + \mathbf{e}_j) - \boldsymbol{\beta}(\mathbf{d}')\right\|_\infty \leq u(\mathbf{d}).$$

Intuitively, $u(\mathbf{d})$ provides a bound on the sensitivity of $\boldsymbol{\beta}$ to increments in entries of the degree sequence. Indeed, $u(\mathbf{d})/\sqrt{2}$ is by definition the Lipschitz constant (with respect to the

ℓ^2 and ℓ^∞ norms) for the restriction of $\boldsymbol{\beta}$ to the set $\{\mathbf{d}' : \mathbf{d}' \geq \mathbf{d}\}$. Since $0 \leq \beta_\ell(\mathbf{d}) \leq n-1$, we have trivially that $u(\mathbf{d}) \leq n-1$. On the other hand, we can also produce degree sequences $\mathbf{d}$ such that $u(\mathbf{d}) \geq 1$. For example, if $d_i = d_j = 0$, then $\beta_i(\mathbf{d}) = \beta_j(\mathbf{d}) = 0$. On the other hand, $\beta_i(\mathbf{d} + \mathbf{e}_i + \mathbf{e}_j) = \beta_j(\mathbf{d} + \mathbf{e}_i + \mathbf{e}_j) = 1$.

We also define v as the smallest real number such that, for all i and j,

$$|\psi(\mathbf{d} + \mathbf{e}_i + \mathbf{e}_j) - \psi(\mathbf{d})| \leq v(\mathbf{d}) \ .$$

Similarly to the above, we have in general that $v(\mathbf{d}) \leq n(n-1)$, since $n(n-1)$ is the largest possible number of nonzero entries of $\mathbf{X}$. We can also produce sequences $\mathbf{d}$ such that $v(\mathbf{d}) \geq 1$; the same example as given above for $\boldsymbol{\beta}$ works here as well.

Conjecture 3.3. *For all* $\mathbf{d}$*, we have* $u(\mathbf{d}) \leq 1$ *and* $v(\mathbf{d}) \leq 1$.

The intuition behind this conjecture is as follows. If $G \in \mathcal{G}_{\mathbf{d}}$, the sequence $\mathbf{d} + \mathbf{e}_i + \mathbf{e}_j$ can be instantiated by a graph $G' \in \mathcal{G}_{\mathbf{d}+\mathbf{e}_i+\mathbf{e}_j}$ in which a single edge has been added between nodes i and j. Trivially, this operation does not decrease the degrees of any nodes, does not increase the degrees of any nodes by more than one, and does not increase the total number of edges by more than one. Conjecture 3.3 states that the same is true of the expected collapsed degrees $\boldsymbol{\beta}$ and expected number of collapsed edges ψ. The bounds we present below do not formally depend on the truth of Conjecture 3.3, but they are not guaranteed to be meaningful unless u and v are indeed small in comparison to n. Unfortunately, the complex combinatorial structure of $\mathcal{G}_{\mathbf{d}}$ renders a proof of our conjecture obscure, and we leave such a proof to future work. In subsection 5.1, we will show anecdotal numerical experiments consistent with this conjecture. For now, we define

$$u_*(\mathbf{d}) = \max_{\mathbf{d}} \max\{u(\mathbf{d}), 1\} \quad \text{and} \quad v_*(\mathbf{d}) = \max_{\mathbf{d}} \max\{v(\mathbf{d}), 1\}.$$

Conjecture 3.3 then states that $u_* = v_* = 1$ for all $\mathbf{d}$.

Theorem 3.4. *Suppose that* $\eta_{\mathbf{d}}(\mathbf{X}_K = \mathbf{e}) > 0$*. Then, for any* $i \in [n]$*,*

$$|\mathbb{E}[b_i | \mathbf{X}_K = \mathbf{e}] - \beta_i(\mathbf{d})| \leq 2q u_*(\mathbf{d}). \tag{3.3}$$

Additionally,

$$|\mathbb{E}[y | \mathbf{X}_K = \mathbf{e}] - \psi(\mathbf{d})| \leq 2q v_*(\mathbf{d}). \tag{3.4}$$

Proof. We will prove (3.3); the proof of (3.4) is parallel. Fix $k, \ell \in [n]$. The distribution $\eta_{\mathbf{d}+\mathbf{e}_k+\mathbf{e}_\ell}$ is supported on graphs with n nodes and $m+1$ edges. Let us condition $\eta_{\mathbf{d}+\mathbf{e}_k+\mathbf{e}_\ell}$ on the event $w_{k\ell} \geq 1$. Then, there exists at least one edge (k, ℓ). Since $\eta_{\mathbf{d}+\mathbf{e}_k+\mathbf{e}_\ell}$ is itself uniform, the conditioned distribution, in which the remaining m edges, is also uniform. Indeed, since we have already assigned an edge incident to nodes k and ℓ, the conditional distribution is uniform over configurations of the remaining m edges in which the degrees sum to $\mathbf{d}$. This is exactly $\eta_{\mathbf{d}}$, and we therefore obtain the identity

$$\eta_{\mathbf{d}}(G) = \eta_{\mathbf{d}+\mathbf{e}_k+\mathbf{e}_\ell}(G \uplus \{(k, \ell)\} | w_{k\ell} \geq 1), \tag{3.5}$$

where $\uplus$ denotes multiset union. This identity relates the operations of conditioning and degree sequence modification. Now, let $\mathbf{v}(K) = \sum_{s=1}^{q}(\mathbf{e}_{k_s} + \mathbf{e}_{\ell_s})$. By iterating (3.5), we obtain

$$\eta_{\mathbf{d}}(G) = \eta_{\mathbf{d}+\mathbf{v}(K)}\left(G \uplus \biguplus_{s=1}^{q}\{k_s, \ell_s\} \Big| \mathbf{W}_K \geq \mathbf{e}\right).$$

The conditioning event can be equivalently written $\mathbf{X}_K = \mathbf{e}$.

For the remainder of this proof, let the symbol $\mathbb{E}_{\mathbf{z}}$ denote expectations with respect to $\eta_{\mathbf{z}}$. We use (3.5) to estimate $\mathbb{E}_{\mathbf{d}+\mathbf{v}(K)}[b_i|\mathbf{X}_K = \mathbf{e}]$ via a two-step experiment. We first sample $G \sim \eta_{\mathbf{d}}$ and compute b_i. We then add the q edges $\{(k_s, \ell_s)\}$ sequentially. Doing so does not decrease b_i and can increase b_i by no more than $q \leq qu_*$. Taking expectations, we obtain the bound

$$\beta_i(\mathbf{d}) \leq \mathbb{E}_{\mathbf{d}+\mathbf{v}(K)}[b_i|\mathbf{X}_K = \mathbf{e}] \leq \beta_i(\mathbf{d}) + qu_*.$$

Applying Conjecture 3.3 inductively, we also have

$$\beta_i(\mathbf{d}) - qu_* \leq \beta_i(\mathbf{d} + \mathbf{v}(K)) \leq \beta_i(\mathbf{d}) + qu_*.$$

We infer that

$$\left|\mathbb{E}_{\mathbf{d}+\mathbf{v}(K)}[b_i|\mathbf{X}_K = \mathbf{e}] - \beta_i(\mathbf{d} + \mathbf{v}(K))\right| \leq 2qu_*.$$

Since $\mathbf{d}$ and K were arbitrary, we can absorb $\mathbf{v}(K)$ into $\mathbf{d}$, obtaining the required statement

$$|\mathbb{E}_{\mathbf{d}}[b_i|\mathbf{X}_K = \mathbf{e}] - \beta_i(\mathbf{d})| \leq 2qu_*.$$

The only subtlety in this case is that the expectation must exist. For this it is sufficient that $\eta_{\mathbf{d}}(\mathbf{X}_K = \mathbf{e}) > 0$, as assumed by hypothesis. ∎

Theorem 3.4 is our primary tool for proving second-moment bounds on the entries of $\mathbf{b}$. From this point forward, we will assume that $\mathbf{d}$ is fixed. The symbols η and $\mathbb{E}$ will refer to the uniform model with degree sequence $\mathbf{d}$ and expectations with respect to that model, respectively.

Lemma 3.5. *Let $i \neq j \neq k$. The following bounds hold:*

$$|\mathbb{E}[b_i x_{jk}] - \beta_i \chi_{jk}| \leq 2u_* \chi_{jk}, \tag{3.6}$$

$$|\mathbb{E}[y x_{jk}] - \psi \chi_{jk}| \leq 2v_* \chi_{jk}, \tag{3.7}$$

$$|\mathbb{E}[b_i b_j] - \beta_i \beta_j| \leq 2u_*(\beta_i \wedge \beta_j), \tag{3.8}$$

$$\mathbb{E}[b_i^2 b_j^2] \leq (\beta_i + 6u_*)^2(\beta_j + 6u_*)^2, \tag{3.9}$$

$$\operatorname{var}(y) \leq 2v_* \psi. \tag{3.10}$$

Proof. To prove (3.6), write

$$\mathbb{E}[b_i x_{jk}] = \eta(x_{jk} = 1)\mathbb{E}[b_i|x_{jk} = 1] = \chi_{jk}\mathbb{E}[b_i|x_{jk} = 1]$$

and apply Theorem 3.4. To prove (3.7), we similarly write

$$\mathbb{E}[yx_{jk}] = \eta(x_{jk} = 1)\mathbb{E}[y|x_{jk} = 1] = \chi_{jk}\mathbb{E}[y|x_{jk} = 1]$$

and apply Theorem 3.4. To prove (3.8), write

$$\mathbb{E}[b_i b_j] = \sum_\ell \mathbb{E}[b_i x_{j\ell}].$$

Now, applying (3.6), we obtain

$$|\mathbb{E}[b_i b_j] - \beta_i\beta_j| \leq 2u_* \sum_\ell \chi_{j\ell} = 2u_*\beta_j.$$

Since we could have expanded b_j instead of b_i, we can choose the smaller of these, and the result follows. The proof of (3.9) is similar. We expand the sums and apply Theorem 3.4. The first step is

$$\begin{aligned}\mathbb{E}[b_i^2 b_j^2] &= \sum_{k,\ell,h} \mathbb{E}[x_{ik}x_{i\ell}x_{jh}]\mathbb{E}[b_j|x_{ik}x_{i\ell}x_{jh} = 1] \\ &\leq \sum_{k,\ell,h} \mathbb{E}[x_{ik}x_{i\ell}x_{jh}](\beta_j + 6u_*) \\ &= (\beta_j + 6u_*)\mathbb{E}[b_i^2 b_j].\end{aligned}$$

Repeating this procedure three more times proves the result. Finally, to prove (3.10), write

$$\operatorname{var}(y) = \mathbb{E}[y^2] - \psi^2 = \frac{1}{2}\sum_{ij}\chi_{ij}\mathbb{E}[y|x_{ij} = 1] - \psi^2 \leq \frac{1}{2}\sum_{ij}\chi_{ij}(\psi + 2v_*) - \psi^2 = 2v_*\psi.$$

We have used (3.7) in the inequality. ∎

Lemma 3.6. *We have*

$$\left|\delta_{ij}^- - \frac{2\psi\chi_{ij}}{z(\mathbf{W})}\right| \leq \frac{\epsilon_{ij}^-}{z(\mathbf{W})} \quad \text{and} \quad \left|\delta_{ij}^+ - \frac{\beta_i\beta_j}{z(\mathbf{W})}\right| \leq \frac{\epsilon_{ij}^+}{z(\mathbf{W})},$$

where

$$\begin{aligned}\epsilon_{ij}^- &\triangleq \chi_{ij}(\beta_i + \beta_j + 3 + 4v_* + 2u_*), \\ \epsilon_{ij}^+ &\triangleq \chi_{ij}(\beta_i + \beta_j + 4u_* - 1) + (2u_* + 1)(\beta_i \wedge \beta_j).\end{aligned}$$

Proof. We require expressions for $\mathbf{\Delta}^-$ and $\mathbf{\Delta}^+$. These are as in the calculation for the configuration model in subsection 3.1, except that there now appears an acceptance probability $a((i,j),(k,\ell)) = \frac{1}{w_{ij},w_{k\ell}}$ that modifies the swap probabilities. The acceptance probability has the effect of replacing instances of w_{ij} with x_{ij}. Performing the algebra and simplifying, we find that

$$\delta_{ij}^- = \frac{1}{z(\mathbf{W})}\mathbb{E}\left[2x_{ij}(y - b_i - b_j) + 3x_{ij}\right], \tag{3.11}$$

$$\delta_{ij}^+ = \frac{1}{z(\mathbf{W})}\mathbb{E}\left[b_i b_j - x_{ij}(b_i + b_j) - \mathbf{x}_i^T\mathbf{x}_j + x_{ij}\right]. \tag{3.12}$$

These are indeed the same expressions as in the configuration model in subsection 3.1, with $\mathbf{W}$ replaced by $\mathbf{X}$. We have used the identity $x_{ij}^2 = x_{ij}$. Computing the expectation of the first line yields

$$\delta_{ij}^- = \frac{1}{z(\mathbf{W})}\left(2\mathbb{E}[yx_{ij}] - 2\mathbb{E}[b_i x_{ij}] - 2\mathbb{E}[b_j x_{ij}] + 3\chi_{ij}\right).$$

Applying (3.6) and (3.7), we obtain the bound

$$\left|\delta_{ij}^- - \frac{1}{z(\mathbf{W})}\left(\chi_{ij}(2(\psi - \beta_i - \beta_j) + 3)\right)\right| \le \frac{4}{z(\mathbf{W})}\chi_{ij}(v_* + 2u_*).$$

We similarly compute

$$\delta_{ij}^+ = \frac{1}{z(\mathbf{W})}\left(\mathbb{E}[b_i b_j] - \mathbb{E}[b_i x_{ij}] - \mathbb{E}[b_j x_{ij}] - \mathbb{E}[\mathbf{x}_i^T \mathbf{x}_j] + \chi_{ij}\right).$$

Since $\mathbf{X}$ is binary, $0 \le \mathbf{x}_i^T \mathbf{x}_j \le b_i \wedge b_j$, and therefore $0 \le \mathbb{E}[\mathbf{x}_i^T \mathbf{x}_j] \le \beta_i \wedge \beta_j$. Applying this observation in concert with (3.8) and (3.6), we find

$$\left|\delta_{ij}^+ - \frac{1}{z(\mathbf{W})}\left(\beta_i\beta_j - \chi_{ij}\beta_i - \chi_{ij}\beta_j + \chi_{ij}\right)\right| \le \frac{1}{z(\mathbf{W})}\left((2u_* + 1)(\beta_i \wedge \beta_j) + 4u_*\chi_{ij}\right).$$

Moving the unwanted terms to the right-hand side in both bounds proves the lemma. ∎

Theorem 3.7 (expectations of $\mathbf{X}$). *We have*

$$\left|\chi_{ij} - \frac{\beta_i\beta_j}{2\psi}\right| \le \epsilon_{ij}(\boldsymbol{\beta}) \triangleq \frac{\epsilon_{ij}^+(\boldsymbol{\beta}) + \epsilon_{ij}^-(\boldsymbol{\beta})}{2\psi}.$$

Furthermore,

$$\epsilon_{ij}(\boldsymbol{\beta}) = \frac{2\chi_{ij}(\beta_i + \beta_j + 3u_* + 2v_* + 2) + (2u_* + 1)(\beta_i \wedge \beta_j)}{2\psi}.$$

Proof. Setting $p = 1$ in (3.1) again yields $\delta_{ij}^+ - \delta_{ij}^- = 0$. Applying Lemma 3.6 and the triangle inequality, we obtain

$$\left|\frac{\beta_i\beta_j}{z(\mathbf{W})} - \frac{2\psi\chi_{ij}}{z(\mathbf{W})}\right| \le \frac{\epsilon_{ij}^- + \epsilon_{ij}^+}{z(\mathbf{W})}.$$

Multiplying through by $\frac{z(\mathbf{W})}{2\psi}$ proves the first claim. The expression for ϵ_{ij} is obtained by inserting the expressions for ϵ_{ij}^- and ϵ_{ij}^+ from Lemma 3.6 and simplifying. ∎

Theorem 3.7 provides an asymptotic error bound of the form

$$\chi_{ij} = \frac{\beta_i\beta_j}{2\psi}\left(1 + O\left(\frac{\chi_{ij}v_*}{\beta_i\beta_j} + \chi_{ij}\frac{\beta_i + \beta_j}{\beta_i\beta_j} + \frac{u_*}{\beta_i \vee \beta_j}\right)\right) \tag{3.13}$$

as β_i and β_j grow large. This bound is admittedly relatively loose, even assuming that u_* and v_* are indeed small. In light of the numerical results presented below, we conjecture that much better bounds may be possible. This appears to be a promising direction for future work.

We can recognize the leading term in (3.13):

$$f_{ij}(\boldsymbol{\beta}) = \frac{\beta_i \beta_j}{2\psi},$$

the same functional form f_{ij} as in the CL estimator defined in (1.2). Speaking somewhat figuratively, we can interpret Theorem 3.7 as indicating that $\mathbf{X}$, the matrix of the projected simple graph, approximately agrees in expectation with the CL model (on off-diagonal entries) with parameter vector $\boldsymbol{\beta}$. However, it would be incorrect to state that $\mathbf{X}$ is distributed according to any model that deterministically preserves a collapsed degree sequence. First, $\boldsymbol{\beta}$ does not in general possess integer entries. Second, the collapsed degrees b_i are still stochastic, preserved only approximately in expectation.

3.3. First moments of W. In the case of the configuration model, approximately solving the $p = 1$ stationarity condition yielded an approximation for $\boldsymbol{\Omega}$ in terms of the known vector $\mathbf{d}$. However, in the uniform model we derived an approximation only for $\boldsymbol{\chi}$ in terms of the unknown vector $\boldsymbol{\beta}$. Computing another equilibrium condition will allow us to both estimate $\boldsymbol{\Omega}$ from $\boldsymbol{\chi}$ and estimate $\boldsymbol{\beta}$ from $\mathbf{d}$. Take $p = 2$ in (3.1), obtaining

$$2\mathbb{E}[w_{ij}\Delta_{ij}] + \mathbb{E}[\Delta_{ij}^2] = 0. \tag{3.14}$$

Study of this condition yields the following result.

Theorem 3.8. *Assume that $f_{ij}(\boldsymbol{\beta}) < 1$. Then,*

$$\left| \omega_{ij} - \frac{f_{ij}(\boldsymbol{\beta})}{1 - f_{ij}(\boldsymbol{\beta})} \right| \leq \frac{1}{1 - f_{ij}(\boldsymbol{\beta})} \left(\frac{2\epsilon'_{ij}(\boldsymbol{\beta}) + \epsilon_{ij}(\boldsymbol{\beta})}{2\psi} + \frac{\epsilon_{ij}(\boldsymbol{\beta})}{2} \right),$$

where $\epsilon_{ij}(\boldsymbol{\beta})$ is as in Theorem 3.7 and

$$\epsilon'_{ij}(\boldsymbol{\beta}) \triangleq \frac{2u_*}{\beta_i \vee \beta_j} + \frac{\sigma_{ij}}{\omega_{ij}} \frac{\sqrt{(\beta_i + 6u_*)^2(\beta_j + 6u_*)^2 - (\beta_i\beta_j - 2u_*(\beta_i \wedge \beta_j))^2}}{\beta_i\beta_j} + \sigma_{ij}\sqrt{2v_*\psi_i} + \omega_{ij}(\beta_i + \beta_j) + \sigma_{ij}\sqrt{2u_*}(\sqrt{\beta_i} + \sqrt{\beta_j})].$$

The proof of Theorem 3.8 proceeds similarly to that of Theorem 3.7, albeit with more involved algebra. It is provided in the supplementary information. While it is notationally convenient to leave the final (inside the square root) term unexpanded, the term $\beta_i^2\beta_j^2$ cancels. The entire expression is therefore of polynomial order $-\frac{1}{2}$ in the entries of $\boldsymbol{\beta}$ and again goes to zero as these entries grow large.

Informally, Theorem 3.8 states that

$$\omega_{ij} \approx \frac{f_{ij}(\boldsymbol{\beta})}{1 - f_{ij}(\boldsymbol{\beta})}. \tag{3.15}$$

Recall that $f_{ij}(\boldsymbol{\beta}) \approx \chi_{ij}$ by Theorem 3.7, and that $\chi_{ij} = \eta(w_{ij} \geq 1)$ by definition. Then, (3.8) states that ω_{ij} is approximately equal to the odds that there is at least one edge present between nodes i and j. As we will see, this approximation gives us a method to compute the vector $\boldsymbol{\beta}$ in terms of the vector $\mathbf{d}$, thereby obtaining an approximation for the moments of $\mathbf{W}$. As in Theorem 3.7, the derived bounds are relatively loose, and substantially better ones may perhaps be obtained from further analysis.

3.4. Second moments. Before proceeding, we briefly comment on the $p = 3$ stationarity condition. From this case on, it becomes quite tedious to control the error terms associated with factoring expectations. Omitting them, we obtain the approximation

$$\mathbb{E}[w_{ij}^2] \approx \omega_{ij}\left(\omega_{ij} + \frac{1}{1-\chi_{ij}}\right).$$

It follows that

$$\sigma_{ij}^2 = \operatorname{var}(w_{ij}) \approx \frac{\chi_{ij}}{(1-\chi_{ij})^2} \approx \omega_{ij}(\omega_{ij}+1). \tag{3.16}$$

Under this approximation, $\sigma_{ij}^2 > \omega_{ij}$ whenever $\chi_{ij} > 0$. It is common to model the entries of the adjacency matrix as Poisson random variables, for which the mean and variance are equal. The formula (3.16) suggests that this approach will be approximately correct for the uniform model when $\omega_{ij} \ll 1$ but will systematically underestimate the variance for larger values.

4. Estimation of $\boldsymbol{\beta}$. We now possess approximate formulae for the low-order moments of $\mathbf{W}$ in terms of the vector $\boldsymbol{\beta}$. In practice, we do not observe $\boldsymbol{\beta}$ and must therefore estimate it from $\mathbf{d}$. To do so, we impose the degree constraint $\sum_j \omega_{ij} = d_i$ and insert the approximation given by Theorem 3.8. Eliding the error terms, we obtain

$$d_i \approx \sum_j \frac{f_{ij}(\boldsymbol{\beta})}{1-f_{ij}(\boldsymbol{\beta})}.$$

We therefore define the function $\mathbf{h} : \mathbb{R}_+^n \to \mathbb{R}^n$ componentwise as

$$h_i(\boldsymbol{\beta}) \triangleq \sum_j \frac{f_{ij}(\boldsymbol{\beta})}{1-f_{ij}(\boldsymbol{\beta})} \tag{4.1}$$

and aim to solve the equation

$$\mathbf{h}(\boldsymbol{\beta}) = \mathbf{d} \tag{4.2}$$

for $\boldsymbol{\beta}$. We define the estimator $\hat{\boldsymbol{\beta}}$ as the solution of (4.2). We then use the estimators $\hat{\chi}_{ij} \triangleq f_{ij}(\hat{\boldsymbol{\beta}})$ and $\hat{\omega}_{ij}^1 \triangleq \frac{f_{ij}(\hat{\boldsymbol{\beta}})}{1-f_{ij}(\hat{\boldsymbol{\beta}})}$ supplied by Theorems 3.7 and 3.8 as estimates of the moments of $\mathbf{W}$. In general, $\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$, since we have discarded the error terms derived in the previous section. We should therefore expect that $\hat{\boldsymbol{\beta}}$ is a biased estimator of $\boldsymbol{\beta}$, and that $\hat{\boldsymbol{\Omega}}^1$ is a biased estimator of $\boldsymbol{\Omega}$. Experiments, however, will show that these biases are substantially smaller than those of $\hat{\boldsymbol{\Omega}}^0$.

To get some intuition on the behavior of (4.2), it is useful to consider two contrasting cases. First, consider the degree sequence $\mathbf{d} = d\mathbf{e}$. In this case, $\mathcal{G}_{\mathbf{d}}$ is the set of regular graphs in which all nodes have the same degree d. We can find a solution of (4.2) analytically. Due to the symmetry of $\mathbf{d}$, we restrict our search to solutions $\boldsymbol{\beta} = \beta\mathbf{e}$, where β is a scalar. Then, (4.2) reads

$$\frac{(n-1)\beta^2}{n\beta - \beta^2} = d.$$

Solving for β yields the estimator $\hat{\beta}$:

$$\hat{\beta} = \frac{d}{1 + n^{-1}(d-1)}.$$

We see that, in a sparse limit in which we let $n \to \infty$ while $d = o(n)$, $\hat{\beta} \to d$. This reflects the asymptotic equivalence of uniform and configuration models under large, sparse limits.

Our second example illustrates a case in which no interpretable solution to (4.2) exists. Consider the star graph, which possesses $k \geq 2$ leaves (labeled 1 through k) and a central node (labeled $k+1$). A single edge connects each leaf to node $k+1$. Node $k+1$ has degree k, while each leaf has degree 1. Every edge is incident to node $k+1$. Because of this, there are no pairs of edges with four distinct indices and therefore no valid edge swaps under Algorithm 1. The corresponding null space $\mathcal{G}_{\mathbf{d}}$ therefore contains only one element. We can thus read off the correct expected collapsed degree sequence: $\beta_{k+1} = k$ and $\beta_j = 1$ for $1 \leq j \leq k$. However, this sequence does not solve (4.2). Indeed, letting β_L denote the unknown shared collapsed degree for each leaf and β_C the collapsed degree of node $k+1$, we can write (4.2) as

$$\begin{aligned} k &= k\frac{\beta_C\beta_L}{2\psi - \beta_C\beta_L}, \\ 1 &= \frac{\beta_C\beta_L}{2\psi - \beta_C\beta_L} + (k-1)\frac{\beta_L^2}{2\psi - \beta_L^2}. \end{aligned}$$

The first line requires that $\frac{\beta_C\beta_L}{2\psi - \beta_C\beta_L} = 1$. In conjunction with the second line, this implies that $\beta_L = 0$, which in turn contradicts the first line unless $\beta_C = 0$ as well. We conclude that no solution to (4.2) exists which respects the symmetries of the star graph. On the other hand, simply adding a second copy of the star graph is sufficient to introduce a solution. For example, in the union of two 5-stars, the algorithm we develop below to solve (4.2) finds that $\beta_C \approx 3.40$ and $\beta_L \approx 0.93$, with mean-square error (MSE) below machine precision. In light of these examples, the conditions such that $\hat{\boldsymbol{\beta}}$ exists constitute an interesting direction for future research.

4.1. Properties of $\hat{\boldsymbol{\beta}}$. While existence remains an open question, it is possible to provide a qualified uniqueness guarantee for (4.2). We will also prove several simple results about the "shape" of the entries of $\hat{\boldsymbol{\beta}}$ as functions of the entries of $\mathbf{d}$. Throughout this section, we assume that $\hat{\boldsymbol{\beta}}$ is sorted, so that $\hat{\beta}_1 \leq \hat{\beta}_2 \leq \cdots \leq \hat{\beta}_n$.

Definition 4.1. *A vector $\boldsymbol{\beta}$ is* physical *if $\mathbf{e} \leq \boldsymbol{\beta} \leq (n-1)\mathbf{e}$ entrywise. A vector $\boldsymbol{\beta}$ is* well behaved with parameter $\delta > 0$ *if, in addition, $\beta_n^2 \leq \mathbf{e}^T\boldsymbol{\beta} - \delta$.*

The bounds imposed by the physicality condition are in fact obeyed by the true expected collapsed degree vector $\mathbb{E}_\eta[\mathbf{b}]$, provided that $\mathbf{d} \geq \mathbf{e}$ entrywise. Well-behavedness with parameter $\delta > 0$ is sufficient, but not necessary, to ensure that $\hat{\omega}_{ij} = f_{ij}(\hat{\boldsymbol{\beta}})(1 - f_{ij}(\hat{\boldsymbol{\beta}}))^{-1} > 0$ for all i and j. Let $\mathcal{B}_\delta$ denote the set of all physical, well-behaved vectors of (implied) fixed size n with a fixed parameter $\delta > 0$. Throughout, we will assume that δ is "sufficiently small"; this will not pose problems due to the inclusion $\mathcal{B}_{\delta'} \subset \mathcal{B}_\delta$ whenever $\delta' < \delta$. By construction, the function $\mathbf{h}$ defined by (4.1) is continuous, and indeed smooth, on $\mathcal{B}_\delta$.

We will show that (4.2) possesses at most one solution on $\mathcal{B}_\delta$. Let

$$\mathcal{L}(\boldsymbol{\beta}) = \|\mathbf{h}(\boldsymbol{\beta}) - \mathbf{d}\|_2^2 \tag{4.3}$$

be the square error associated with approximating $\mathbf{d}$ by $\mathbf{h}(\boldsymbol{\beta})$. Then, the problem

$$\min_{\boldsymbol{\beta} \in \mathcal{B}_\delta} \mathcal{L}(\boldsymbol{\beta}) \tag{4.4}$$

achieves its minimum value of 0 at the solutions of (4.2) in $\mathcal{B}_\delta$, provided there are any. We will show that (4.4) possesses at most one such minimum.

Our proof relies on an elementary form of the Mountain Pass Theorem [2], given as Lemma 6.1 in [8]. A closely related statement is given as Theorem 5.2 in [32].

Definition 4.2 (Palais–Smale condition [8]). *Let $q : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Let $\{\mathbf{a}_n\}$ be a sequence of points in $\mathbb{R}^n$ such that $q(\mathbf{a}_n)$ is bounded and $\|\nabla q(\mathbf{a}_n)\| \to 0$. The function q* satisfies the Palais–Smale condition *if any such $\{\mathbf{a}_n\}$ possesses a convergent subsequence.*

Theorem 4.3 (Mountain Pass Lemma in $\mathbb{R}^n$ [8, 2]). *Suppose that function $q : \mathbb{R}^n \to \mathbb{R}$ satisfies the Palais–Smale condition. Suppose further that the following hold:*

1. *$q(\mathbf{a}_0) = 0$.*
2. *There exist an $r > 0$ and $\alpha > 0$ such that $q(\mathbf{a}) \geq \alpha$ for all $\mathbf{a}$ with $\|\mathbf{a} - \mathbf{a_0}\| = r$.*
3. *There exists $\mathbf{a}'$ such that $\|\mathbf{a}' - \mathbf{a}_0\| > r$ and $q(\mathbf{a}') \leq 0$.*

Then, q possesses a critical point $\tilde{\mathbf{a}}$ with $q(\tilde{\mathbf{a}}) \geq \alpha$.

Our strategy is as follows. We will first show that all critical points of $\mathcal{L}$ are solutions of (4.2). We will then show that all such critical points are, furthermore, isolated local minima of (4.4). The existence of two such isolated local minima would trigger Theorem 4.3, implying the existence of an additional critical point with $\mathcal{L}(\boldsymbol{\beta}) > 0$. Since this is a contradiction, we will conclude that only one such minimum exists.

Our first step is to lower-bound the eigenvalues of the Jacobian $\mathbf{J}$ matrix of $\mathbf{h}$ at an arbitrary point $\boldsymbol{\beta}$. This Jacobian may be written

$$\mathbf{J} = (\mathbf{S} + \mathbf{D}) \left(\mathbf{B}^{-1} - \frac{1}{4\psi} \mathbf{E} \right). \tag{4.5}$$

In this expression, $\mathbf{S}$ is the matrix with entries

$$s_{ij} = \begin{cases} \frac{f_{ij}(\boldsymbol{\beta})}{(1 - f_{ij}(\boldsymbol{\beta}))^2}, & i \neq j, \\ 0, & i = j. \end{cases}$$

We have also defined $\mathbf{D} = \operatorname{diag} \mathbf{Se}$ and $\mathbf{B} = \operatorname{diag}(\boldsymbol{\beta})$. We note as a point of curiosity that $s_{ij} \approx \operatorname{var}(w_{ij})$ by (3.16), although our results here do not depend on this relationship. A derivation of (4.5) is supplied in the supplementary information. Let $\lambda_i(\mathbf{M})$ denote the ith eigenvalue of the matrix $\mathbf{M}$, sorted in ascending order. Thus, $\lambda_1(\mathbf{M})$ is the smallest eigenvalue of $\mathbf{M}$, and $\lambda_n(\mathbf{M})$ the largest.

Lemma 4.4. *Assume $n \geq 5$. Then,*

$$\lambda_1(\mathbf{J}) \geq \frac{1}{n(n-1)}\left(1 - \frac{2}{\sqrt{5}}\right) > 0. \tag{4.6}$$

In particular, $\mathbf{J}$ is positive-definite, and its eigenvalues are bounded away from zero on $\mathcal{B}_\delta$.

The proof proceeds by applying two eigenvalue lower bounds to the form of $\mathbf{J}$ given by (4.5) [7, 31], and is given in the supplementary information.

Lemma 4.5. *If $n \geq 5$ and $\boldsymbol{\beta}$ is a critical point of $\mathcal{L}$, then*
(a) *$\boldsymbol{\beta}$ solves (4.2), and*
(b) *the Hessian $\mathbf{H}$ of $\mathcal{L}$ at $\boldsymbol{\beta}$ is positive-definite.*

Proof. To prove (a), we compute the gradient of $\mathcal{L}$:

$$\nabla \mathcal{L}(\boldsymbol{\beta}) = 2(\mathbf{h}(\boldsymbol{\beta}) - \mathbf{d})^T \mathbf{J}(\boldsymbol{\beta}). \tag{4.7}$$

By Lemma 4.4, $\mathbf{J}(\boldsymbol{\beta})$ is positive-definite and therefore full-rank on $\mathcal{B}_\delta$. It follows that $\nabla \mathcal{L}(\boldsymbol{\beta}) = 0$ iff $\mathbf{h}(\boldsymbol{\beta}) = \mathbf{d}$, or, equivalently, iff $\mathcal{L}(\boldsymbol{\beta}) = 0$.

To prove (b), we calculate the entries of the Hessian. These are

$$\mathbf{H}(\boldsymbol{\beta})_{ij} = 2\sum_{\ell=1}^{n}\left[(h_\ell(\boldsymbol{\beta}) - d_\ell)\frac{\partial^2 h_\ell}{\partial \beta_i \partial \beta_j} + \frac{\partial h_\ell}{\partial \beta_i}\frac{\partial h_\ell}{\partial \beta_j}\right].$$

The first term vanishes at critical points. Recognizing the second as an outer product of the rows of $\mathbf{J}$, we can write the Hessian at critical points as

$$\mathbf{H}(\boldsymbol{\beta}) = 2\sum_{\ell=1}^{n} \mathbf{J}_\ell(\boldsymbol{\beta})\mathbf{J}_\ell(\boldsymbol{\beta})^T.$$

Since $\mathbf{J}$ is full-rank by Lemma 4.4, the sum is full-rank and therefore positive-definite. This completes the proof. ∎

We immediately obtain the following corollary.

Corollary 1. *If $n \geq 5$, then each critical point of $\mathcal{L}$ is an isolated local minimum, and there are finitely many of them.*

For the second clause, we rely on the fact that $\mathcal{B}_\delta$ is closed and bounded.

Lemma 4.6. *If $n \geq 5$, the restriction of $\mathbf{h}$ to $\mathcal{B}_\delta$ satisfies the Palais–Smale condition.*

Proof. Taking norms in (4.7) and lower-bounding the right-hand side, we obtain

$$\|\nabla \mathcal{L}(\boldsymbol{\beta})\|_2 \geq \lambda_1\left(\mathbf{J}(\boldsymbol{\beta})\right) \|h(\boldsymbol{\beta}) - \mathbf{d}\|_2 .$$

Since $\lambda_1(\mathbf{J}(\boldsymbol{\beta}))$ is bounded away from zero on $\mathcal{B}_\delta$ by Lemma 4.4, the only sequences $\{\boldsymbol{\beta}_t\}$ in $\mathcal{B}_\delta$ that satisfy $\|\nabla\mathcal{L}(\boldsymbol{\beta}_t)\|_2 \to 0$ must also satisfy $\mathbf{h}(\boldsymbol{\beta}_t) \to \mathbf{d}$. By Lemma 4.5, there are finitely many solutions to (4.2), and therefore any such sequence has a finite number of limit points. The sequence $\{\boldsymbol{\beta}_t\}$ then possesses a subsequence that converges to each of these limit points, which completes the proof. ∎

Theorem 4.7. *If $n \geq 5$, there exists at most one $\hat{\boldsymbol{\beta}}$ in the set $\mathcal{B}_\delta$ such that $\mathbf{h}(\hat{\boldsymbol{\beta}}) = \mathbf{d}$.*

Proof. Suppose that there were two solutions $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ in $\mathcal{B}_\delta$. Since $\mathcal{L}$ satisfies the Palais–Smale condition (Lemma 4.6), we check that conditions 1–3 of Theorem 4.3 are satisfied. Condition 1 requires that $\mathcal{L}(\boldsymbol{\beta}_0) = 0$, which is true by hypothesis. Condition 2 requires that there exist $r > 0$ and $\alpha > 0$ such that $h(\boldsymbol{\beta}) \geq \alpha$ for all $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = r$. This follows from Taylor-expanding $\mathcal{L}$ around $\boldsymbol{\beta}_0$ and using the positive-definiteness of $\mathbf{H}$. Applying Lemma 4.5 yields the existence of such an r and α and further implies that r may be taken to be arbitrarily small. In particular, r may be taken to be smaller than $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1\|$, which in turn supplies condition 3. Applying Theorem 4.3, we conclude that there exists a critical point $\tilde{\boldsymbol{\beta}}$ of $\mathcal{L}$ such that $\mathcal{L}(\tilde{\boldsymbol{\beta}}) \geq \alpha > 0$. But this contradicts Lemma 4.5. We conclude that at most one solution to (4.2) exists in $\mathcal{B}_\delta$, as was to be shown. ∎

Numerical experiments suggest that the solution to (4.2), if it exists, may be unique in the positive orthant $\mathbb{R}^n_+$. If true, this would be a stronger result than that provided by Theorem 4.7, which requires physicality and well-behavedness. Extending Theorem 4.7 to cover the full positive orthant would be an interesting direction of future work.

The following theorem specifies several properties of $\hat{\boldsymbol{\beta}}$, provided that it exists.

Theorem 4.8. *Let $n \geq 5$. Suppose that $\hat{\boldsymbol{\beta}} \in \mathcal{B}_\delta$ solves (4.2). Then the following hold:*

(a) *The map $d_i \mapsto \hat{\beta}_i$ is nondecreasing.*

(b) *Furthermore, $\hat{\beta}_i - \hat{\beta}_j \leq d_i - d_j$.*

(c) *Finally, $\hat{\boldsymbol{\beta}} \leq \mathbf{d}$ entrywise.*

A proof of this result is given in the supplementary information. A plot of $\hat{\beta}_i$ as a function of d_i, such as provided in panel (c) of Figure 2, is a helpful way to visualize these results. Theorem 4.8 asserts of such a plot that the curve is nondecreasing, that its slope is no greater than unity, and that it lies below the line of equality.

4.2. Algorithms. Having proven some properties of the solutions of (4.2), it remains to develop an algorithm to find these solutions. While it is possible to use standard gradient-based methods, this task is complicated by the ill-conditioned Jacobian of $\mathbf{h}$. Ill-conditioning arises from dramatic heterogeneity in the entries of $\mathbf{S}$. For example, in the experiments shown in Figure 2 in the next section, the observed and estimated values of σ_{ij} span four orders of magnitude, implying that entries of $s_{ij} \approx \sigma_{ij}^2$ span roughly eight. Because of this, methods based on the full Jacobian, such as standard implementations of gradient descent or Newton's method, may require impractically small step-sizes in order to avoid pathological behavior.

We instead adopt a coordinatewise approach. Suppose we have a current estimate $\hat{\boldsymbol{\beta}}^{(t-1)}$. We obtain an estimate of $\mathbf{d}$ given by $\hat{\mathbf{d}}^{(t-1)} = \mathbf{h}(\hat{\boldsymbol{\beta}}^{(t-1)})$. To update the ith coordinate of $\hat{\boldsymbol{\beta}}^{(t-1)}$, we hold all other $n-1$ coordinates fixed and define $\hat{\beta}_i^{(t)}$ to be the value of b that solves

the equation

$$h_i\left(\hat{\beta}_1^{(t-1)}, \dots, \hat{\beta}_{i-1}^{(t-1)}, b, \hat{\beta}_{i+1}^{(t-1)}, \dots, \hat{\beta}_n^{(t-1)}\right) = d_i. \tag{4.8}$$

We repeat this process for each of the n coordinates, obtaining a fully updated new estimate $\hat{\boldsymbol{\beta}}^{(t)}$. We iterate this sweep over the coordinates until a desired error function drops below a user-specified tolerance. A standard choice for the error function is the MSE $n^{-1}\mathcal{L}(\boldsymbol{\beta})$, with $\mathcal{L}$ as in (4.3). Algorithm 2 formalizes the solution method. The call Solve_b solves a single-variable equation for b. In the accompanying code (see "Software"), we implement Solve_b with options to use either the `root_scalar()` function supplied by `Python`'s `scipy` package or a bespoke Newton-type method.

Algorithm 2: Computation of $\hat{\beta}$.

Input: degree sequence $\mathbf{d} \in \mathbb{Z}_+^n$, initial guess $\hat{\boldsymbol{\beta}}^{(0)} \in \mathbb{R}_+^n$, tolerance ϵ

Initialization: $t \leftarrow 0$, $\gamma \leftarrow \infty$

while $\gamma^{(t)} > \epsilon$ **do**

  for $i = 1, \dots, n$ **do**

    $\hat{\beta}_i^{(t)} \leftarrow \text{Solve}_b\{h_i(\hat{\beta}_1^{(t-1)}, \dots, \hat{\beta}_{i-1}^{(t-1)}, b, \hat{\beta}_{i+1}^{(t-1)} \dots, \hat{\beta}_n^{(t-1)}) = d_i\}$

  $\gamma^{(t)} \leftarrow n^{-1}\mathcal{L}(\hat{\boldsymbol{\beta}}^{(t)})$

  $t \leftarrow t + 1$

Output: $\hat{\boldsymbol{\beta}}^{(t)}$

In order to ensure that this algorithm is well defined, we will show that the update given by (4.8) possesses a unique solution under mild conditions.

Lemma 4.9. *Assume that* $\mathbf{d} > \mathbf{0}$ *and* $\boldsymbol{\beta}^{(t-1)} > \mathbf{0}$ *entrywise. Then, for each* i, (4.8) *possesses a unique solution in* b *on the open interval* $\left(0, \frac{2\psi^{(t-1)}}{\max_{\ell \neq i} \beta_\ell^{(t-1)}}\right)$.

Remark 4.10. The hypotheses of Lemma 4.9 can be ensured by removing degree-zero nodes from $\mathbf{d}$ and initializing $\boldsymbol{\beta}^{(0)} > \mathbf{0}$.

Proof. To prove existence, we first observe that h_i is a continuous function of β_i. We have $h_i(\beta_1, \dots, 0, \dots \beta_n) = 0$ and

$$\lim_{\beta \to \frac{2\psi^{(t-1)}}{\max_{\ell \neq i} \beta_\ell}} h_i(\beta_1, \dots, \beta, \dots, \beta_n) = \infty.$$

The Intermediate Value Theorem then provides existence.

To show uniqueness, it suffices to check the derivative (cf. (4.5))

$$\frac{\partial h_i(\boldsymbol{\beta})}{\partial \beta_i} = \left(\frac{1}{\beta_i} - \frac{1}{2\psi}\right) \sum_{\ell \neq i} \frac{f_{i\ell}(\boldsymbol{\beta})}{(1 - f_{i\ell}(\boldsymbol{\beta}))^2}.$$

When $\boldsymbol{\beta} > 0$, this expression is strictly positive. The function h_i is therefore strictly increasing on I, proving uniqueness. ∎

While we have existence, uniqueness, and convergence guarantees for each coordinate update, we possess no such guarantees for Algorithm 2 as a whole. Additionally, it may be the case that some elements of the sequence $\{\hat{\boldsymbol{\beta}}^{(t)}\}$ produce estimates $\hat{\omega}^{(t)}$ of the adjacency matrix in which some entries are negative. However, we have never observed Algorithm 2 to fail to converge to a solution in which all entries of $\hat{\omega}^{(t)}$ are positive. Additionally, when a solution to (4.2) exists in $\mathcal{B}_\delta$ for some δ, we have never observed Algorithm 2 to fail to find this solution. In practice, an analyst can assess the success of the algorithm by checking that (a) the MSE is near zero and that (b) the corresponding estimate $\hat{\boldsymbol{\Omega}}$ has nonnegative entries. Both such checks are implemented in the accompanying software.

5. Experiments.

In this section, we describe a sequence of experiments exploring the behavior of Algorithm 2; the accuracy of the estimator $\hat{\boldsymbol{\beta}}$; the disparity between $\hat{\boldsymbol{\Omega}}^0$ and $\hat{\boldsymbol{\Omega}}^1$ on empirical networks; and implications for downstream tasks such as modularity maximization.

5.1. Synthetic data.

To study the convergence behavior of Algorithm 2, we test it on two synthetic degree sequences. The "uniform" sequence consists of 200 independent copies of $2(u+1)$, where u is a discrete uniform random variable on the interval $[0, 50]$. We contrast this with a "Zipf" sequence $\mathbf{d}_2$, generated by sampling 200 copies of $2z$, where z is distributed as a Zipf random variable with parameter $\alpha = 2$. These degree sequences are shown in Figure 1(a). By design, the uniform sequence is relatively homogeneous in its degrees, while the Zipf sequence possesses a small number of extremely high-degree nodes.

We then estimated $\hat{\boldsymbol{\beta}}$ for each of these degree sequences using Algorithm 2, initialized with $\boldsymbol{\beta}^{(0)} = \mathbf{e}$. The estimates for the uniform sequence $\mathbf{d}_1$ converge rapidly, as shown in panel (b), and after two rounds the iterates cannot be distinguished by eye from the final estimate. The final estimate $\hat{\boldsymbol{\beta}}$ is both physical and well behaved. By Theorem 4.7, it is the only such solution to (4.2). In contrast, the estimates for the Zipf-distributed sequence $\mathbf{d}_2$, shown in panel (c), require many rounds to converge. Figure 1(d) compares the differing convergence rates. The vertical axis gives the MSE $\frac{1}{n}\left\|\mathbf{h}(\boldsymbol{\beta}) - \mathbf{d}\right\|_2$. While the MSE for the uniform degree sequence converges to within machine precision after 14 rounds, the Zipf iterates require over 200 iterations to reach an MSE below 10^{-6}. The resulting estimate $\hat{\boldsymbol{\beta}}$ is physical but not well behaved, and Theorem 4.7 is therefore insufficient to provide a uniqueness guarantee.

Algorithm 2 also allows us to perform some bootstrap-style tests of Conjecture 3.3. Recall that this conjecture asserts that the scalar $u(\mathbf{d})$, which bounds the effect of perturbations of $\mathbf{d}$ on β, is no larger than 1. The size of $u(\mathbf{d})$ in turn influences the tightness of the error bounds derived in section 3. Figure 1(e)–(f) shows the results of a simple experiment in which we use our estimator $\hat{\boldsymbol{\beta}}$ as a surrogate for $\boldsymbol{\beta}$. For each degree sequence, we repeatedly sample i and j from $\binom{[n]}{2}$. We then compute $\hat{\boldsymbol{\beta}}' = \hat{\boldsymbol{\beta}}(\mathbf{d} + \mathbf{e}_i + \mathbf{e}_j)$ and compare it to $\hat{\boldsymbol{\beta}}$. Filled dots show the maximum absolute change, $\left\|\hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}}\right\|_\infty$, while empty dots give the mean absolute change $\frac{1}{n}\left\|\hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}}\right\|_1$. Under Conjecture 3.3, we would expect that $\left\|\hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}}\right\|_\infty \leq 1$, which is indeed the case for both degree sequences. These results may be viewed as heuristic supports of Conjecture 3.3. Additionally, $\frac{1}{n}\left\|\hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}}\right\|_1 \ll 1$. This observation suggests the possibility of substantially tightening the error bounds given in section 3 by controlling the ℓ^1-norm rather than the ℓ^∞-norm, an interesting problem which we leave to future work.

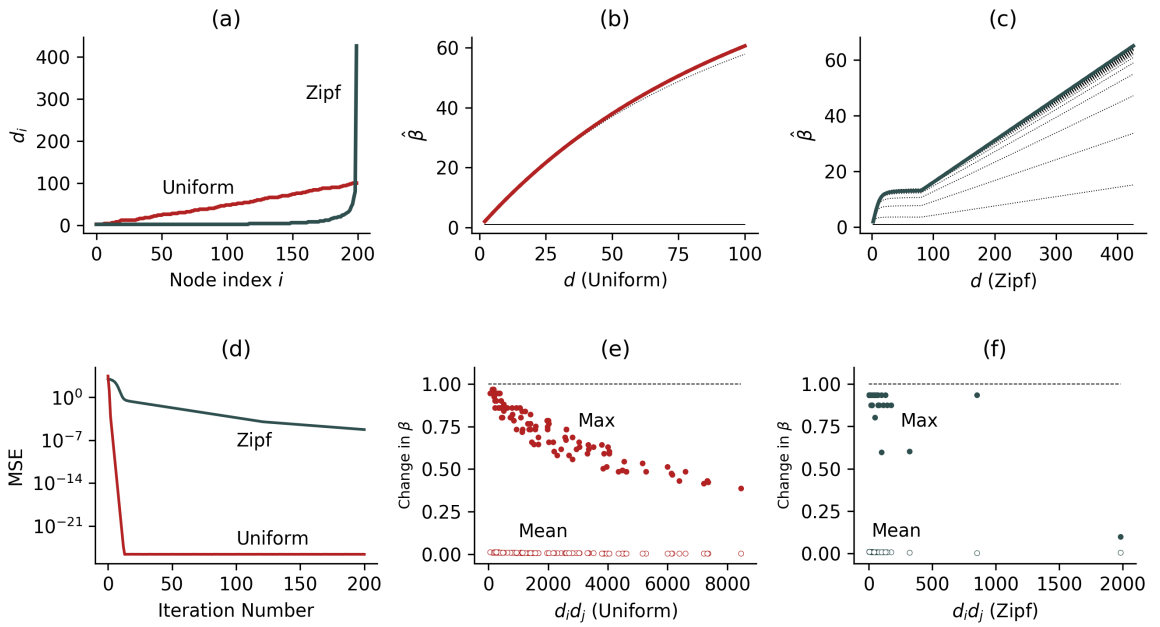

Figure 1. (a) *The uniform and Zipf degree sequences described in the text.* (b) *Iterates of Algorithm 2 for the uniform degree sequence. The final iterate is highlighted.* (c) *Iterates of Algorithm 2 for the Zipf degree sequence. The final iterate is highlighted.* (d) *MSE in Algorithm 2 for the uniform (red) and Zipf (gray) degree distributions as a function of the iteration number.* (e) *Bootstrap estimates of* $\|\boldsymbol{\beta}(\mathbf{d}+\mathbf{e}_i+\mathbf{e}_j)-\boldsymbol{\beta}\|_\infty$ *(filled points) and* $n^{-1}\|\boldsymbol{\beta}(\mathbf{d}+\mathbf{e}_i+\mathbf{e}_j)-\boldsymbol{\beta}\|_1$ *(empty circles) for the uniform degree sequence. Each point corresponds to a uniformly random choice of distinct indices* i *and* j*.* (f) *As in* (e)*, for the Zipf degree sequence.*

5.2. Evaluation on an empirical contact network. Our evaluation data set is a contact network among students in a French secondary school, called `contact-high-school` [34, 6]. During data collection, each student wore a proximity sensor. An interaction between two students was logged by their respective sensors when the students were face-to-face and within approximately 1.5m of each other. Edges are time-stamped, although we do not use any temporal information in the present experiments. The original data set contains $n = 327$ nodes and $m = 189,928$ distinct interactions.

We first test the accuracy of the estimator $\hat{\boldsymbol{\Omega}}^1$, using $\hat{\boldsymbol{\Omega}}^{\text{mc}}$ as a reliable estimate of the true mean $\boldsymbol{\Omega}$. Because of the scaling issues associated with estimating $\hat{\boldsymbol{\Omega}}^{\text{mc}}$ on $m \approx 2 \times 10^5$ edges, we constructed a data subset based on a temporal threshold τ, chosen to incorporate approximately the last 5% of the original interaction volume. The resulting subnetwork has 268 nodes and 10,026 edges. To estimate the ground-truth moments $\eta_{\mathbf{d}}$, we estimated $\hat{\boldsymbol{\Omega}}^{\text{mc}}$ on the subnetwork from 10^7 samples at intervals of 10^3 steps. This computation required approximately one week on a single thread of a modern server.

In Figure 2(a)–(b), we show the distributions of degrees and entries of $\mathbf{w}$ for this subnetwork. Figure 2(a) depicts the heterogeneous degree distribution, with standard deviation larger than the average degree. While most nodes have small degrees, there are twelve whose degree exceeds n. Figure 2(b) shows the clumping of edges between pairs of nodes. On

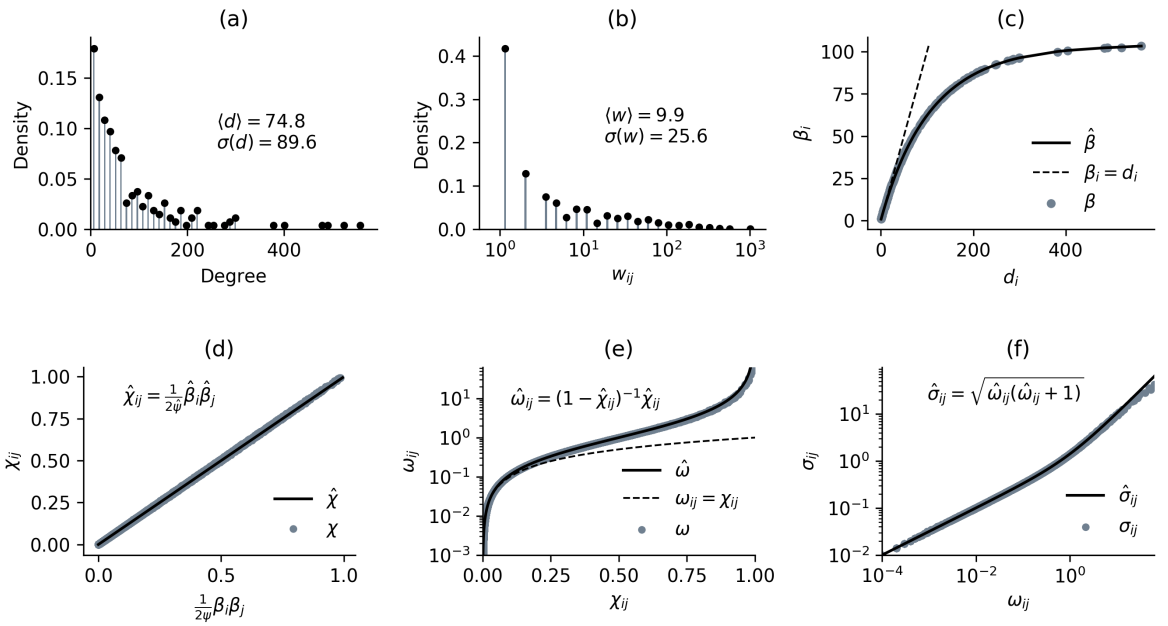

Figure 2. (a) *Degree distribution of the* `contact-high-school` *subnetwork. The mean degree* $\langle d \rangle$ *and standard deviation of the degree* $\sigma(d)$ *are shown.* (b) *Distribution of the entries of* $\mathbf{w}$*. Note the logarithmic horizontal axis.* (c) *Collapsed degree sequence* $\hat{\boldsymbol{\beta}}$ *learned from* $\mathbf{d}$ *via Algorithm 2. Dashes give the line of equality.* (d) *Approximation of* $\boldsymbol{\chi}$ *via* (3.13). (e) *Approximation of* $\boldsymbol{\Omega}$ *via* (3.15)*. Note the logarithmic vertical axis. Dashes give the line of equality.* (f) *Approximation of* $\sigma_{ij} = \sigma(W_{ij})$ *via* (3.16)*. Note the log-log axis. In* (c)-(f)*, simulated moments (gray dots) are obtained via the Monte Carlo estimator using Algorithm 1; see main text for details.*

average, two students who interact at all interact nearly ten times, but there is substantial deviation around this average. Almost half of all pairs interact just once. In contrast, a small number of pairs interact over 100 times, and one over 1,000.

In Figure 2(c)–(f), we show the construction of estimators for the moments of $\boldsymbol{\Omega}$ under the uniform random graph model with the observed degree sequence $\mathbf{d}$. In Figure 2(c), the solid line shows the estimate $\hat{\boldsymbol{\beta}}$ output by Algorithm 2, plotted against the degree sequence. Points give the MCMC estimate for $\boldsymbol{\beta}$. The agreement is almost exact. The estimate $\hat{\boldsymbol{\beta}}$ is physical and well behaved. Theorem 4.7 implies that it is the only physical, well-behaved solution to (4.2).

In Figure 2(d), we estimate $\hat{\chi}_{ij} \approx f_{ij}(\hat{\boldsymbol{\beta}}) = \dfrac{\hat{\beta}_i \hat{\beta}_j}{2\hat{\psi}}$, again finding the agreement to be near exact. In (e), we estimate $\hat{\omega}_{ij} \approx (1 - \hat{\chi}_{ij})^{-1} \hat{\chi}_{ij}$. The agreement with data is again excellent, although there is a small amount of visible overestimation of ω_{ij} when χ_{ij} is large. Finally, (f) uses (3.16) to compute an estimator $\hat{\sigma}_{ij} = \sqrt{\hat{\omega}_{ij}(\hat{\omega}_{ij} + 1)}$ of σ_{ij} the standard deviation of W_{ij}. The agreement is strong through roughly $\omega_{ij} \approx 10$ and begins to overestimate σ_{ij} for larger values.

Figure 2(c) and Figure 2(e) also highlight the relationship of $\hat{\boldsymbol{\Omega}}^1$ and $\hat{\boldsymbol{\Omega}}^0$. The dashed lines in these figures represent two linear approximations that can be made to yield the latter from

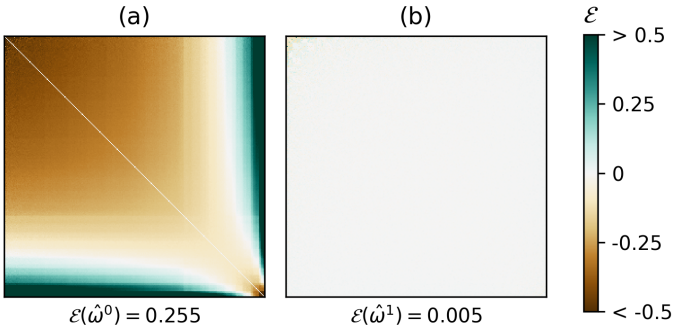

Figure 3. *Shading gives the relative error for approximating* $\mathbf{\Omega}$ *under the uniform model for the* `contact-high-school` *subnetwork.* (a) *Using the CL estimator* $\hat{\mathbf{\Omega}}^0$. (b) *Using the present estimator* $\hat{\mathbf{\Omega}}^1$. *Node degrees in each matrix increase left to right and top to bottom. The "ground truth" is provided by* $\hat{\mathbf{\Omega}}^{\text{mc}}$, *computed as in Figure 2.*

the former. First, we approximate $\boldsymbol{\beta} = \mathbf{d}$ (dashed line, Figure 2(c)). This approximation holds well when d_i is small, since then the number of parallel edges incident to node i should be small, i.e., $W_{ij} \approx X_{ij}$. Then, we approximate $\mathbf{\Omega} = \boldsymbol{\chi}$ (dashed line, Figure 2(e)). This approximation should hold for small entries of $\mathbf{\Omega}$, since in this case χ_{ij} is small and $(1 - \chi_{ij})^{-1} \approx 1$. As the plots indicate, these approximations are indeed accurate when d_i and ω_{ij} are small. These conditions correspond roughly to the "large, sparse" heuristics used frequently in the literature. We can therefore view $\hat{\mathbf{\Omega}}^0$ as a first-order approximation to $\hat{\mathbf{\Omega}}^1$ near the large, sparse regime. Conversely, we can view $\hat{\mathbf{\Omega}}^1$ as a nonlinear correction to $\hat{\mathbf{\Omega}}^0$ as we depart from that regime.

Figure 3 compares the overall performance of the estimators $\hat{\mathbf{\Omega}}^0$ and $\hat{\mathbf{\Omega}}^1$. We compute the entrywise relative error $\mathcal{E}_{ij}(\hat{\mathbf{\Omega}}) = (\hat{\omega}_{ij}^{\text{mc}})^{-1}(\hat{\omega}_{ij} - \hat{\omega}_{ij}^{\text{mc}})$ when approximating $\hat{\mathbf{\Omega}}^{\text{mc}} \approx \mathbf{\Omega}$ with both methods. Cells are shaded according to the magnitude and sign of the error. The CL estimator in (a) displays systematic bias, underestimating the density of edges between nodes of similar degrees and overestimating the density of edges between nodes with highly disparate degrees. The mean absolute relative error of the CL estimate is $\mathcal{E}(\hat{\mathbf{\Omega}}^0) = \binom{n}{2}^{-1} \sum_{ij} \left| \mathcal{E}_{ij}(\hat{\mathbf{\Omega}}^0) \right| \approx 0.255$, indicating that a typical entry of $\hat{\mathbf{\Omega}}^0$ is off by over 25%. In contrast, $\hat{\mathbf{\Omega}}^1$ evaluated in (b) has almost no visible bias. Some large residuals are visible for entries ω_{ij} in which both d_i and d_j are small (top left corner), although it is difficult to evaluate to what extent these residuals reflect error in $\hat{\mathbf{\Omega}}^1$ or in the challenge to $\hat{\mathbf{\Omega}}^{\text{mc}}$ to estimate these edge densities in practical runtime. The mean absolute relative error $\mathcal{E}(\hat{\mathbf{\Omega}}^1)$ is roughly 0.5%, an improvement over $\hat{\mathbf{\Omega}}^0$ of a full order and a half of magnitude.

5.3. Application: Modularity maximization in dense contact networks. Let $\ell : N \to \mathcal{L}$ be a function that assigns to node i a label $\ell_i \in \mathcal{L}$. The *modularity* of the partition ℓ with respect to matrix $\mathbf{w}$ and null model ρ is given by

$$Q(\ell; \rho) = \frac{1}{2m} \sum_{ij} \left[w_{ij} - \mathbb{E}_\rho[W_{ij}] \right] \delta(\ell_i, \ell_j) \,. \tag{5.1}$$

The normalization by $2m$ ensures that $-1 \leq Q(\ell;\rho) \leq 1$. Intuitively, $Q(\ell;\rho)$ is high when nodes that are more densely connected than expected by chance (under the specified null) are grouped together. Maximizing this quantity with respect to ℓ may therefore be reasonably expected to identify modular ("community") structure in the network [38, 40]. Exact modularity maximization is NP-hard [12] and subject to theoretical limitations in networks with modules of heterogeneous sizes [24]. Despite this, it remains one of the most popular methods for practical community detection at scale [10].

In most implementations, ρ is not explicitly specified; rather, the expectation $\mathbb{E}_\rho[W_{ij}]$ is "hard-coded" as equal to $\hat{\omega}^0_{ij}$. From a statistical perspective, this reflects an implicit choice of ρ as the CL model [18], which preserves expected degrees and indeed possesses the given first moment.[3] Modifications are possible; the best known is perhaps the resolution adjustment that replaces $\hat{\mathbf{\Omega}}^0$ with $\gamma\hat{\mathbf{\Omega}}^0$ for some $\gamma > 0$ [42]. Other adjustments may incorporate spatial structure [23] or adjust for the inclusion of self-loops in the null space [13]. When we wish to perform modularity maximization against a null that deterministically preserves degree sequences, however, $\hat{\mathbf{\Omega}}^0$ is at best an approximation. We expect this approximation to perform adequately for the configuration model (cf. Theorem 3.1) and very poorly for the uniform model (previous subsection). The Monte Carlo estimate $\hat{\mathbf{\Omega}}^{\mathrm{mc}}$ can be used for very small data sets but rapidly becomes computationally infeasible for larger ones. In these cases, we can use the present estimator $\hat{\mathbf{\Omega}}^1$ instead.

Recent work has highlighted the importance of studying the *modularity landscape*, especially the set of local maxima of Q, rather than restricting attention to a single partition. One reason for this is the phenomenon of *degeneracy*; in many practical contexts, a given network will possess many distinct local maxima with modularity comparable to the global maximum [27]. A second reason is model-specification. As shown in [39], maximization of Q is related to maximum-likelihood inference in a planted-partition stochastic block model. When the planted-partition model is unrealistic as a generative model for the data, modularity-maximization amounts to inference in a mis-specified model. Degeneracy is a common symptom of this problem, but observing it requires locally optimizing Q multiple times. For these reasons among others, ensemble-based methods that implicitly average over local optima, such as those of [49], may be preferable. With these considerations in mind, our aim in this section is not to show that the use of $\hat{\mathbf{\Omega}}^1$ is strictly superior for this task when compared to $\hat{\mathbf{\Omega}}^0$ for either one-shot or ensemble-based modularity maximization. Rather, we will argue that the corresponding modularity landscapes are significantly different on data sets of practical interest, and that it is therefore methodologically unsafe to interchange these estimators without carefully scrutinizing the results.

Our first data set for this experiment is the full `contact-high-school` network, consisting of $n = 327$ nodes and $m = 189,928$ edges as described in the previous subsection. The computation of $\hat{\mathbf{\Omega}}^{\mathrm{mc}}$ is indeed infeasible for a graph this dense, and we therefore use $\hat{\mathbf{\Omega}}^1$ as an estimate. This setting highlights the utility of $\hat{\mathbf{\Omega}}^1$, since otherwise we would have no practical way to compute the uniform expectation.

[3] We note that alternative justifications of the use of $\hat{\mathbf{\Omega}}^0$ exist, including connections to the stability of Markov chains [21] and to stochastic block models [39].

To approximately maximize (5.1), we employ the multiway spectral partitioning (MSP) algorithm of [50], which generalizes the spectral bipartitioning algorithm of [38]. While greedy methods often enjoy superior performance [10], spectral methods have the advantage of depending strongly on the structure of the observed graph and the null model employed and are relatively insensitive to choices made during the runtime of the algorithm. Spectral methods are therefore ideal for highlighting differences in the modularity landscapes induced by alternative null models. The algorithm requires the analyst to specify a null model and a desired number of communities k. The core of the approach is to use a low-rank approximation of the *modularity matrix*, $\mathbf{M} = \mathbf{w} - \mathbb{E}_\rho[\mathbf{W}]$. This approximation induces a map from the vertices of G to a low-dimensional vector space. Vectors in this space are clustered according to their relative angles using a procedure reminiscent of k-means to produce the community assignment. Because the clustering algorithm involves a stochastic starting condition, it is useful to run the algorithm multiple times and choose the highest modularity partition from among the repetitions. We refer the reader to [50] for details, and to the code accompanying this paper for an implementation of MSP for arbitrary modularity matrices (see "Software").

We ran this algorithm using both the CL modularity matrix $\mathbf{M}^0 = \mathbf{w} - \hat{\mathbf{\Omega}}^0$ and the approximate uniform modularity matrix $\mathbf{M}^1 = \mathbf{w} - \hat{\mathbf{\Omega}}^1$. We refer to these two algorithmic variants as MSP^0 and MSP^1, respectively. Since $\hat{\mathbf{\Omega}}^0$ and $\hat{\mathbf{\Omega}}^1$ produce very different null matrices, the modularity matrices $\mathbf{M}^0$ and $\mathbf{M}^1$ are themselves very different; the mean absolute relative error of using the latter to estimate the former is approximately 32%. We would therefore expect MSP^0 and MSP^1 to behave very differently in this task. We allowed the number of communities k to vary between 2 and 10. For each value of k, we ran MSP^0 and MSP^1 in 100 batches of 50 repetitions. From each batch of 50, the highest-modularity partition was chosen, resulting in 100 partitions per value of k. Figure 4(a) shows that MSP^1 tends to find higher modularity partitions than MSP^0 on this data set. The difference is especially large when k is small, but a substantial difference between the means is noticeable even for larger values. While partitions under $\mathbf{M}^0$ exist that are comparable to those under $\mathbf{M}^1$, it appears to be more difficult for MSP^0 to find them. Panels (b) and (c) shed some light on the differing behavior of the two algorithms. Partitions under MSP^0 tend to display a larger, less cohesive community ((b), top left) alongside smaller, more tightly interconnected ones. Partitions under MSP^1 (c) tend to display communities that are slightly more uniform in size.

It is reasonable to object that modularity values under MSP^1 and MSP^0 should not be compared, since these objectives are defined with respect to differing null matrices. In this specific case, the objection is not borne out numerically, however; "cross-evaluating" the partitions on the opposite matrices changes the modularities only minimally. Evaluating the MSP^0 partition in Figure 4(b) on the modularity matrix $\mathbf{M}^1$ gives $Q = 0.699$, while evaluating the MSP^1 partition on $\mathbf{M}^0$ yields $Q = 0.731$. On this data set, MSP^1 searches the energy landscape of MSP^0 more efficiently than does MSP^0 itself.

It should be noted that this behavior is data set dependent. The opposite case occurs in the `contact-primary-school` network [45, 6], which used similar sensors to construct an interaction network among students in a French primary school. On this data, MSP^0 and MSP^1 perform similarly for $k \leq 6$ communities (Figure 5), with the former consistently

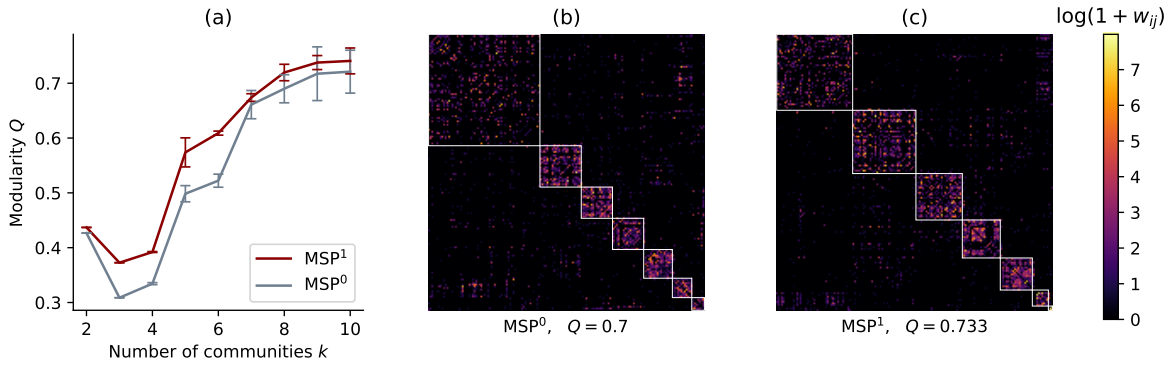

Figure 4. (a) *Performance of MSP on the full* `contact-high-school` *network using the CL modularity matrix* $\mathbf{M}^0$ *and the approximate uniform modularity matrix* $\mathbf{M}^1$*, over* 100 *batches of* 50 *repetitions each. Solid lines give the average modularity, and error bars give two standard deviations from the mean.* (b) *Example partition using* $\mathbf{M}^0$*.* (c) *Example partition using* $\mathbf{M}^1$*. To generate* (b) *and* (c)*, the best partition of* 500 *runs was chosen for each algorithm variant. Each run was initialized with* $k = 8$*; in the best partition, however, only* 7 *labels are actually used. Colors are shown on a log scale.*

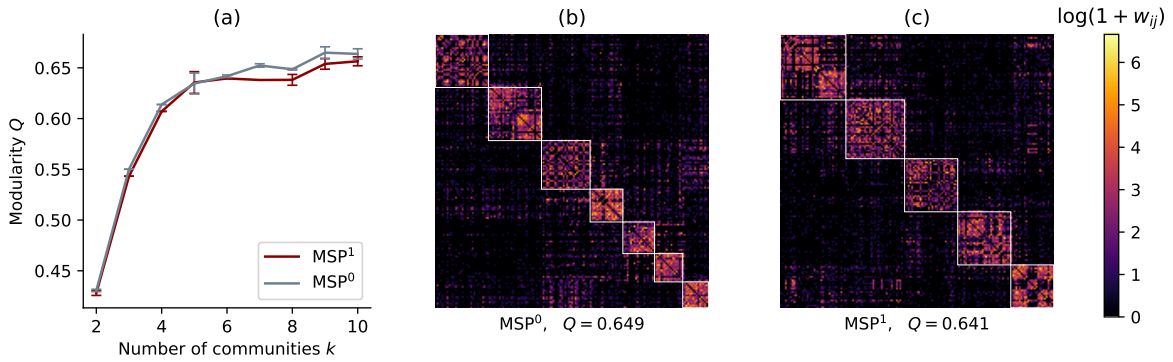

Figure 5. *This figure is in all methodological details identical to Figure 4, using the study data set* `contact-primary-school` [45, 6]*. In* (b)–(c)*, both algorithms were initialized with* $k = 8$*.*

outperforming the latter for $k \geq 7$. The illustrative partitions in panels (b) and (c) suggest that MSP^1 tends to prefer partitions with fewer communities. Whereas MSP^0 chooses a partition with 7 communities, MSP^1 chooses one with just 5 (both having been initialized at $k = 8$). These illustrations emphasize that MSP^1 and MSP^0 explore different modularity landscapes; that the relative advantages of each algorithm depend on the data; and that the landscape for MSP^1 can be tractably computed under the methodology we have introduced here.

6. Discussion.

Much existing network theory is explicitly designed for large, sparse data. However, many networks of interest are sufficiently dense to diverge significantly from the predictions of large, sparse theory. We have highlighted this phenomenon in the context of dense multigraphs, with a focus on estimating the expected adjacency matrix $\mathbf{\Omega}$ of a random multigraph with specified degree sequence. We have shown that, rather than falling back

to computationally expensive MCMC, we can construct an accurate estimator $\hat{\mathbf{\Omega}}^1$ using an indirect, dynamical approach. Use of this estimator can in turn have a significant impact on the results of downstream data analyses.

We have left open several interesting questions. From a theoretical perspective, the most important outstanding problem is the improvement of the error bounds derived in section 3, including a proof of Conjecture 3.3. As previously noted, the error bounds on $\hat{\boldsymbol{\chi}}$ and $\hat{\mathbf{\Omega}}^1$ derived in section 3 appear quite loose when compared against the empirical results in Figures 1 and 2, and we speculate that significant improvement is possible. One interesting angle of attack runs through the $\boldsymbol{\beta}$-model [30, 15]. Recall that the $\boldsymbol{\beta}$-model is the maximum-entropy random graph with specified *expected* degree sequence (and that $\boldsymbol{\beta}$ in the $\boldsymbol{\beta}$-model is not directly related to $\boldsymbol{\beta}$ as primarily discussed throughout this article). As shown in [5], there is a connection between the densities of large subgraphs in the $\boldsymbol{\beta}$-model and the configuration model on simple graphs as n grows large. The authors show that, subject to mild restrictions on the node degrees, the densities of subgraphs with $\Theta(n^2)$ edges are nearly equal as n grows large. One might hope to use this connection to import some results from the relatively well-developed theory of estimation in $\boldsymbol{\beta}$-models to the uniform model, resulting in better error control. At least two steps would be required to carry out this program. First, it would be necessary to extend the connecting result of [5] to the case of multigraphs. Second, it would be necessary to somehow relate guarantees of *large* subgraph densities to guarantees on *small* subgraph densities, such as a single edge or the neighborhood of a single node. The development of tighter error bounds, through these or other methods, would be of substantial utility for practitioners using the uniform model.

There are also several intriguing questions concerning the application of our work in data scientific practice. For example, a more systematic study of the impact of the choice between $\hat{\mathbf{\Omega}}^1$ and $\hat{\mathbf{\Omega}}^0$ on downstream analyses would be extremely welcome. We have seen that the choice of null expectation can substantially change the performance of MSP, and that the direction of this effect depends on the data set. A better understanding of the properties of the data or algorithm that make certain estimators highlight better solutions would be most welcome. Another promising direction concerns the derivation of estimators for small subgraph densities. While we focused our attention on the derivation of an estimator for $\mathbf{\Omega}$, it may also be possible to derive expressions for higher moments using the same methodology. Examples of interest may include wedge densities $\mathbb{E}[w_{ij}w_{jk}]$ and triangle densities $\mathbb{E}[w_{ij}w_{jk}w_{ik}]$. Parsing the stationarity conditions for these more complicated moments may be correspondingly more difficult. An alternative would be to construct mean-field estimates by computing the relevant statistics on $\hat{\mathbf{\Omega}}^1$ itself. An evaluation of the accuracy of this approach would potentially augment or replace computationally intensive MCMC sampling to estimate subgraph counts in certain settings.

Finally, it may also be of interest to develop a similar theory for uniform distributions over related spaces of graphs. For example, one might study a uniform model including self-loops. The associated analysis would be nontrivial due to required modifications in the MCMC sampling procedure (see [25]), but one might reasonably hope to obtain parallel results. Directed graphs offer another important direction of generalization. It would be natural to define a uniform distribution over spaces of directed multigraphs with fixed in-degree and

out-degree sequences. In this case, one might expect analysis to produce expressions for the moments of this distribution in terms of two collapsed degree sequences, corresponding to in- and out-degrees. These and other generalizations offer promising avenues of future work.

Software and data. We used the implementation of MCMC in [16] to conduct simulation experiments. All additional code used in this study may be freely accessed at

https://github.com/PhilChodrow/multigraph_moments.

The data used in this study was collected by [34] and neatly packaged by the authors of [6] at

https://www.cs.cornell.edu/~arb/data/index.html.

REFERENCES

[1] G. Amanatidis, B. Green, and M. Mihail, *Graphic Realizations of Joint-Degree Matrices*, preprint, https://arxiv.org/abs/1509.07076, 2015.

[2] A. Ambrosetti and P. H. Rabinowitz, *Dual variational methods in critical point theory and applications*, J. Funct. Anal., 14 (1973), pp. 349–381.

[3] O. Angel, R. van der Hofstad, and C. Holmgren, *Limit Laws for Self-Loops and Multiple Edges in the Configuration Model*, preprint, https://arxiv.org/abs/1603.07172, 2016.

[4] Y. Artzy-Randrup and L. Stone, *Generating uniformly distributed random networks*, Phys. Rev. E, 72 (2005), 056708.

[5] A. Barvinok and J. A. Hartigan, *The number of graphs and a random graph with a given degree sequence*, Random Structures Algorithms, 42 (2013), pp. 301–348.

[6] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, *Simplicial closure and higher-order link prediction*, Proc. Natl. Acad. Sci. USA, 115 (2018), pp. 11221–11230.

[7] R. Bhatia, *Matrix Analysis*, Grad. Texts in Math. 169, Springer, New York, 1997.

[8] J. Bisgard, *Mountain passes and saddle points*, SIAM Rev., 57 (2015), pp. 275–292, https://doi.org/10.1137/140963510.

[9] J. Blitzstein and P. Diaconis, *A sequential importance sampling algorithm for generating random graphs with prescribed degrees*, Internet Math., 6 (2011), pp. 489–522.

[10] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large networks*, J. Stat. Mech. Theory Experiment, 10 (2008), pp. 1–12.

[11] B. Bollobás, *A probabilistic proof of an asymptotic formula for the number of labelled regular graphs*, European J. Combin., 1 (1980), pp. 311–316.

[12] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, *On finding graph clusterings with maximum modularity*, in International Workshop on Graph-Theoretic Concepts in Computer Science, Springer, New York, 2007, pp. 121–132.

[13] S. Cafieri, P. Hansen, and L. Liberti, *Loops and multiple edges in modularity maximization of networks*, Phys. Rev. E, 81 (2010), 046102.

[14] C. J. Carstens, *Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast curveball algorithm*, Phys. Rev. E, 91 (2015), 042812.

[15] S. Chatterjee, P. Diaconis, and A. Sly, *Random graphs with a given degree sequence*, Ann. Appl. Probab., 21 (2011), pp. 1400–1435.

[16] P. S. Chodrow, *Configuration models of random hypergraphs*, J. Complex Networks, 8 (2020), pp. 1–20.

[17] F. Chung and L. Lu, *Connected components in random graphs with given expected degree sequences*, Ann. Combin., 6 (2002), pp. 125–145.

[18] F. Chung and L. Lu, *The average distances in random graphs with given expected degrees*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 15879–15882.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 2012.
[20] C. I. Del Genio, H. Kim, Z. Toroczkai, and K. E. Bassler, *Efficient and exact sampling of simple graphs with given arbitrary degree sequence*, PloS ONE, 5 (2010), e10012.
[21] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, *Stability of graph communities across time scales*, Proc. Natl. Acad. Sci. USA, 107 (2010), pp. 12755–12760.
[22] P. L. Erdős, C. Greenhill, T. R. Mezei, I. Miklós, D. Soltész, and L. Soukup, *The Mixing Time of the Swap (Switch) Markov Chains: A Unified Approach*, preprint, https://arxiv.org/abs/1903.06600, 2019.
[23] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, *Uncovering space-independent communities in spatial networks*, Proc. Natl. Acad. Sci. USA, 108 (2011), pp. 7663–7668.
[24] S. Fortunato and M. Barthélemy, *Resolution limit in community detection*, Proc. Natl. Acad. Sci. USA, 104 (2006), pp. 36–41.
[25] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander, *Configuring random graph models with fixed degree sequences*, SIAM Rev., 60 (2018), pp. 315–355, https://doi.org/10.1137/16M1087175.
[26] P. Gao and C. Greenhill, *Mixing Time of the Switch Markov Chain and Stable Degree Sequences*, preprint, https://arxiv.org/abs/2003.08497, 2020.
[27] B. H. Good, Y.-A. De Montjoye, and A. Clauset, *Performance of modularity maximization in practical contexts*, Phys. Rev. E, 81 (2010), 046106.
[28] C. Greenhill, *A polynomial bound on the mixing time of a Markov chain for sampling regular directed graphs*, Electron. J. Combin., 18 (2011), 234.
[29] C. Greenhill, *The switch Markov chain for sampling irregular graphs*, in Proceedings of the 2015 Annual ACM–SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 2015, pp. 1564–1572, https://doi.org/10.1137/1.9781611973730.103.
[30] P. W. Holland and S. Leinhardt, *An exponential family of probability distributions for directed graphs*, J. Amer. Statist. Assoc., 76 (1981), pp. 33–65.
[31] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 2012.
[32] Y. Jabri, *The Mountain Pass Theorem: Variants, Generalizations and Some Applications*, Encyclopedia Math. Appl. 95, Cambridge University Press, Cambridge, UK, 2003.
[33] M. Jerrum and A. Sinclair, *Fast uniform generation of regular graphs*, Theoret. Comput. Sci., 73 (1990), pp. 91–100.
[34] R. Mastrandrea, J. Fournet, and A. Barrat, *Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys*, PLoS ONE, 10 (2015), e0136497.
[35] B. D. McKay and N. C. Wormald, *Uniform generation of random regular graphs of moderate degree*, J. Algorithms, 11 (1990), pp. 52–67.
[36] M. Molloy and B. Reed, *A critical point for random graphs with a given degree sequence*, Random Structures Algorithms, 6 (1995), pp. 161–180.
[37] M. Molloy and B. Reed, *The size of the giant component of a random graph with a given degree sequence*, Combin. Probab. Comput., 7 (1998), pp. 295–305.
[38] M. E. J. Newman, *Modularity and community structure in networks*, Proc. Natl. Acad. Sci. USA, 103 (2006), pp. 8577–8582.
[39] M. E. J. Newman, *Equivalence between modularity optimization and maximum likelihood methods for community detection*, Phys. Rev. E, 94 (2016), 052315.
[40] M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Phys. Rev. E, 69 (2004), 026113.
[41] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Random graphs with arbitrary degree distributions and their applications*, Phys. Rev. E, 64 (2001), 026118.
[42] J. Reichardt and S. Bornholdt, *Statistical mechanics of community detection*, Phys. Rev. E, 74 (2006), 016110.
[43] A. Rinaldo, S. Petrović, and S. E. Fienberg, *Maximum likelihood estimation in the β-model*, Ann. Statist., 41 (2013), pp. 1085–1110.
[44] J. Sjöstrand, *Making Multigraphs Simple by a Sequence of Double Edge Swaps*, preprint, https://arxiv.org/abs/1904.06999, 2019.

[45] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems, *High-resolution measurements of face-to-face contact patterns in a primary school*, PLoS ONE, 6 (2011), e23176.

[46] G. Strona, D. Nappo, F. Boccacci, S. Fattorini, and J. San-Miguel-Ayanz, *A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals*, Nature Commun., 5 (2014), 4114.

[47] N. D. Verhelst, *An efficient MCMC algorithm to sample binary matrices with fixed marginals*, Psychometrika, 73 (2008), pp. 705–728.

[48] F. Viger and M. Latapy, *Efficient and simple generation of random simple connected graphs with prescribed degree sequence*, in Proceedings of the International Computing and Combinatorics Conference, Springer, New York, 2005, pp. 440–449.

[49] P. Zhang and C. Moore, *Scalable detection of statistically significant communities and hierarchies, using message passing for modularity*, Proc. Natl. Acad. Sci. USA, 111 (2014), pp. 18144–18149.

[50] X. Zhang and M. E. J. Newman, *Multiway spectral community detection in networks*, Phys. Rev. E, 92 (2015), 052808.