Network histograms and universality of blockmodel approximation

Sofia C. Olhede and Patrick J. Wolfe[1]

Departments of Statistical Science and Computer Science, University College London, London WC1E 6BT, United Kingdom

Edited* by Peter J. Bickel, University of California, Berkeley, CA, and approved September 2, 2014 (received for review January 8, 2014)

In this paper we introduce the network histogram, a statistical summary of network interactions to be used as a tool for exploratory data analysis. A network histogram is obtained by fitting a stochastic blockmodel to a single observation of a network dataset. Blocks of edges play the role of histogram bins and community sizes that of histogram bandwidths or bin sizes. Just as standard histograms allow for varying bandwidths, different blockmodel estimates can all be considered valid representations of an underlying probability model, subject to bandwidth constraints. Here we provide methods for automatic bandwidth selection, by which the network histogram approximates the generating mechanism that gives rise to exchangeable random graphs. This makes the blockmodel a universal network representation for unlabeled graphs. With this insight, we discuss the interpretation of network communities in light of the fact that many different community assignments can all give an equally valid representation of such a network. To demonstrate the fidelity-versus-interpretability tradeoff inherent in considering different numbers and sizes of communities, we analyze two publicly available networks—political weblogs and student friendships—and discuss how to interpret the network histogram when additional information related to node and edge labeling is present.

community detection | graphons | nonparametric statistics | graph limits | sparse networks

The purpose of this paper is to introduce the network histogram—a nonparametric statistical summary obtained by fitting a stochastic blockmodel to a single observation of a network dataset. A key point of our construction is that it is not necessary to assume the data to have been generated by a blockmodel. This is crucial, because networks provide a general means of describing relationships between objects. Given n objects under study, a total of $\binom{n}{2}$ pairwise relationships are possible. When only a small fraction of these relationships are present—as is often the case in modern high-dimensional data analysis across scientific fields—a network representation simplifies our understanding of this dependency structure.

One fundamental characterization of a network comes through the identification of community structure (1), corresponding to groups of nodes that exhibit similar connectivity patterns. The canonical statistical model in this setting is the stochastic blockmodel (2): It posits that the probability of an edge between any two network nodes depends only on the community groupings to which those nodes belong. Grouping nodes together in this way serves as a natural form of dimensionality reduction: As n grows large, we cannot retain an arbitrarily complex view of all possible pairwise relationships. Describing how the full set of n objects interrelate is then reduced to understanding the interactions of $k \ll n$ communities. Studying the properties of fitted blockmodels is thus important (3, 4).

Despite the popularity of the blockmodel, and its clear utility, scientists have observed that it often fails to describe all of the structure present in a network (5–8). Indeed, as a network becomes larger, it is no longer reasonable to assume that a majority of its structure can be explained by a blockmodel with a fixed number of blocks. Extensions to the blockmodel have focused on capturing additional variability, for example through mixed community membership (5) and degree correction (6, 9). However, the simplest and most natural method of extending the descriptiveness of the blockmodel is to add blocks, so that k grows with n. As more and more blocks are fitted, we expect an increasing degree of structure in the data to be explained. The natural questions to ask then are many: What happens as we fit more blocks to an arbitrary network dataset, if the true data-generating mechanism is not a blockmodel? At what rate should we increase the number of blocks used, depending on the variability of the network? We discuss these and other questions in this paper.

We will stipulate how the dimension k of the fitted blockmodel should be allowed to increase with the size n of the network. This increase will be dictated by a tradeoff between the sparsity of the network and its heterogeneity or smoothness. If one assumes that a k-community blockmodel is the actual data-generating mechanism, then theory has already been developed that allows k to grow with n (10–12), and methods have been suggested for choosing the number of blocks based on the data (13, 14). General theory for the case when the blockmodel is merely approximating the observed network structure is nascent; ref. 15 treated the case of dense bipartite graphs with a fixed number of blocks, and ref. 16 established the first such results for the setting of relevance here.

From Stochastic Networks to Histograms

A Simple Stochastic Network Model. We encode the relationships between n objects using $\binom{n}{2}$ binary random variables. Each of these variables indicates the presence or absence of an edge between two nodes and can be collected into an $n \times n$ adjacency matrix A, such that $A_{ij} = 1$ if nodes i and j are connected and $A_{ij} = 0$ otherwise, with $A_{ii} = 0$. This yields what is known as a simple random graph.

Models for unlabeled graphs are strongly related to the statistical notion of exchangeability, a fundamental concept describing random variables whose ordering is without information. To relate to exchangeable variables, we appeal to the Aldous–Hoover

Significance

Representing and understanding large networks remains a major challenge across the sciences, with a strong focus on communities: groups of network nodes whose connectivity properties are similar. Here we argue that, independently of the presence or absence of actual communities in the data, this notion leads to something stronger: a histogram representation, in which blocks of network edges that result from community groupings can be interpreted as two-dimensional histogram bins. We provide an automatic procedure to determine bin widths for any given network and illustrate our methodology using two publicly available network datasets.

Author contributions: S.C.O. and P.J.W. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: p.wolfe@ucl.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1400374111/-/DCSupplemental.

theorem (3) and model our network hierarchically using three components:

i) A fixed, symmetric function $f(x,y)$ termed a graphon (17), which behaves like a probability density function for $0<x,y<1$;
ii) For each n, a random sample ξ of n uniform random variables $\{\xi_1,\ldots,\xi_n\}$ which will serve to index the graphon $f(x,y)$; and
iii) For each n, a deterministic scaling constant $\rho_n>0$, specifying the expected fraction of edges $\binom{n}{2}^{-1}\mathbb{E}\sum_{i<j}A_{ij}$ in the network.

For each n, our simple stochastic network model is then

$$A_{ij}\big|\xi_i,\xi_j\sim\text{Bernoulli}\big(\rho_n f\big(\xi_i,\xi_j\big)\big),\quad 1\le i<j\le n, \qquad [1]$$

where for statistical identifiability of ρ_n we assume

$$\iint_{(0,1)^2} f(x,y)dx\,dy=1. \qquad [2]$$

In this way we model the network structure itself—rather than the particular ordering in which the network's nodes are arranged in A. As an example, Fig. 1 shows three different orderings of the adjacency matrix of a network of US political weblogs recorded in 2005 (18), each emphasizing a different aspect of the network.

We see from this generative mechanism that any (symmetric) rearrangement of the x and y axes of f will lead to the same probability distribution on unlabeled graphs, and in fact a graphon describes an entire equivalence class of functions. We assume that at least one member of this equivalence class is Hölder-continuous, which we refer to as f without loss of generality; that f is bounded away from 0 and $\rho_n f$ is bounded away from 1; and that the sequence ρ_n is monotone nonincreasing and decays more slowly than $n^{-1}\log^3 n$, so that the average network degree grows faster than $\log^3 n$.

To summarize the network we therefore wish to estimate the graphon $f(x,y)$, up to rearrangement of its axes. By inspection,

$$\mathbb{P}\big(A_{ij}=1\big)=\mathbb{E}A_{ij}=\rho_n\iint_{(0,1)^2}f(x,y)dx\,dy=\rho_n,$$

and so we may estimate ρ_n via the sample proportion estimator

$$\hat\rho_n=\binom{n}{2}^{-1}\sum_{i<j}A_{ij}. \qquad [3]$$

The Network Histogram. Given a single adjacency matrix A of size $n\times n$, we will estimate $f(x,y)$ (up to rearrangement of its axes) using a stochastic blockmodel with a single, prespecified community size h, to yield a network histogram. Choosing the bandwidth h is equivalent to choosing a specific number of communities k—corresponding to the number of bins in an ordinary histogram setting.

To define the network histogram, we first write the total number of network nodes n in terms of the integers h, k, and r as $n=hk+r$, where $k=\lfloor n/h\rfloor$ is the total number of communities; h is the corresponding bandwidth, ranging from 2 to n; and $r=n \bmod h$ is a remainder term between 0 and $h-1$. To collect together the nodes of our network that should lie in the same group, we introduce a community membership vector z of length n. All components of z will take values in $\{1,\ldots,k\}$ and will share the same values whenever nodes are assigned to the same community.

The main challenge in forming a network histogram lies in estimating the community assignment vector z from A. To this end, for each n, let the set $\mathcal{Z}_k\subseteq\{1,\ldots,k\}^n$ contain all community assignment vectors z that respect the given form of $n=hk+r$. Thus, $\mathcal{Z}_k$ consists of all vectors z with h components equal to each of the integers from 1 to $k-1$ (up to relabeling) and $h+r$ components equal to k (again, up to relabeling). In this way, $\mathcal{Z}_k$ indexes all possible histogram arrangements of network nodes into $k-1$ communities of equal size h, plus an additional community of size $h+r$.

Many ways of estimating z from a single observed adjacency matrix A have been explored in the literature. In essence, nodes that exhibit similar connectivity patterns are likely to be grouped together (an idea that can be exploited directly if multiple observations of the same network are available; see ref. 19). We can formalize this notion through the method of maximum likelihood, by estimating

$$\hat z=\underset{z\in\mathcal{Z}_k}{\text{argmax}}\sum_{i<j}\big\{A_{ij}\log\overline{A}_{z_iz_j}+\big(1-A_{ij}\big)\log\big(1-\overline{A}_{z_iz_j}\big)\big\}, \qquad [4]$$

where for all $1\le a,b\le k$ we define the histogram bin heights

$$\overline{A}_{ab}=\frac{\sum_{i<j}A_{ij}\mathbb{I}\big(\hat z_i=a\big)\mathbb{I}\big(\hat z_j=b\big)}{\sum_{i<j}\mathbb{I}\big(\hat z_i=a\big)\mathbb{I}\big(\hat z_j=b\big)}. \qquad [5]$$

Each bin height $\overline{A}_{ab}$ is the proportion of successes (edges present) in the histogram bin corresponding to a block of Bernoulli trials, with the grouping of nodes into communities determined by the objective function in Eq. **4**. Because A is symmetric, we have $\overline{A}_{ab}=\overline{A}_{ba}$.

Combining Eqs. **3** and **5**, we obtain our network histogram:

$$\hat f(x,y;h)=\hat\rho_n^{+}\overline{A}_{\min\left(\lceil nx/h\rceil,k\right)\min\left(\lceil ny/h\rceil,k\right)},\quad 0<x,y<1, \qquad [6]$$

with $\hat\rho_n^{+}$ the generalized inverse of $\hat\rho_n$.

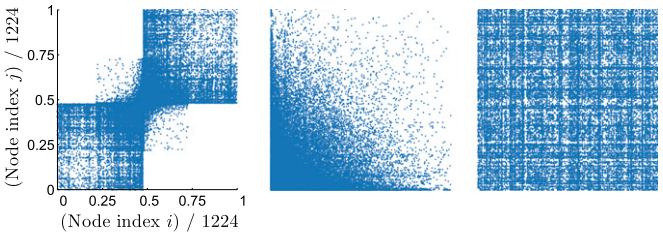

Fig. 1. Three adjacency matrix representations of the political weblog data of ref. 18, each showing all 1,224 weblogs with at least one link to another weblog in the dataset (links denoted by blue dots). (*Left*) The first 586 weblogs are categorized by ref. 18 as liberal and the remaining 638 as conservative; note the sparsity of cross-linkages. (*Center*) The same data, ordered by decreasing number of links. (*Right*) How a random labeling obscures structure.

Universality of Blockmodel Approximation

Blockmodel Approximations of Unlabeled Graphs. To understand the performance of blockmodel approximation we must compare $\hat f$ to f in a way that is invariant to all symmetric rearrangements of the axes of f. We will base our comparison on the graph-theoretic notion of cut distance, which in mathematical terminology defines a compact metric space on graphons (17). Just as our notion of unlabeled graphs treats any two adjacency matrices as the same if one can be obtained by symmetrically permuting the rows and columns of the other, we will compare two graphons via an invertible, symmetric rearrangement of the x and y axes that relates one graphon to the other. We call $\mathcal{M}$ the set of all such rearrangements—formally, it is the set of all measure-preserving bijections of the form $[0,1]\to[0,1]$.

In ref. 16 we formulated convergence rates at which the resulting error between $\hat f$ and f shrinks to zero as $n\to\infty$ under the assumptions above. Here we consider mean integrated square error (MISE), typically used in standard histogram theory (see, e.g., ref. 20) and take its greatest lower bound over all possible rearrangements $\sigma\in\mathcal{M}$:

$$\text{MISE}\big(\hat f\big)=\mathbb{E}\inf_{\sigma\in\mathcal{M}}\iint_{(0,1)^2}\Big|f(x,y)-\hat f(\sigma(x),\sigma(y);h)\Big|^2dx\,dy. \qquad [7]$$

This definition factors out the unknown ordering of the data A induced by $\{\xi_1,\ldots,\xi_n\}$ in the model of Eq. **1**, accounting for the fact that A may represent an unlabeled graph. The appearance of σ may at first seem counterintuitive, but its introduction is necessary once we use Eq. **1** to model A. In contrast to the optimization of Eq. **7** over all $\sigma\in\mathcal{M}$, which is purely conceptual, the vector $\hat z$ results from the algorithmic optimization of Eq. **4**

STATISTICS

given an observed adjacency matrix A and determines which entries of A are averaged to estimate $f(x,y)$.

Using a single bandwidth h to form $\hat{f}(x,y;h)$ in Eq. **7** represents a conceptual paradigm shift away from the standard use of the stochastic blockmodel. Instead of representing community structure, a blockmodel can be used as a universal mechanism to represent an arbitrary unlabeled network. In practice, of course, we may well have information that implies certain labelings or orderings of the network nodes. The assumption of exchangeability models our ignorance of this information as a baseline, just as we may choose to cluster a Euclidean dataset without taking into account any accompanying labels. Thus, we require our error metric to respect this ignorance, even if we later choose to interpret a fitted histogram in light of node labels (as one might with Euclidean data clusters, and as we shall do below).

The goal in the setting of exchangeable networks is therefore no longer to discover latent community structure, but rather simply to group together nodes whose patterns of interactions are similar. Thus, the interpretation of the fitted groups has altered. Instead of uncovering true underlying communities that might have given rise to the data, our blocks now approximate the generative process, up to a resolution chosen by the user—namely the bandwidth, h. This can be related to previous understanding of the error behavior when the data are generated by a blockmodel, both in the regimes of ρ_n corresponding to growing degrees (11) as well as even sparser ones (21).

The Oracle Network Labeling. We next show how the ideal or oracle labeling information, were it to be available, would yield the optimal bandwidth parameter h for any given network histogram. This oracle information arises from the latent random variables $\{\xi_1,\ldots,\xi_n\}$ present in the generative model of Eq. **1**. In this setting, instead of fitting blocks of varying sizes to the network, to be interpreted as community structure, we rely on the fact that the simplest type of blockmodel will suffice, with only a single tuning parameter h. The existence of a smooth limiting object—namely the graphon $f(x,y)$—implies that a single community size or bandwidth will provide an adequate summary of the entire network.

To choose h, we therefore use the notion of a network oracle. As in standard statistical settings (20), the oracle provides information that is not ordinarily available, thereby serving to bound the performance of any data-driven estimation procedure. The oracle estimator for each histogram bin height takes the same form as Eq. **5** but uses a unique (almost surely) labeling $\tilde{z}$ calculated from the latent random vector ξ. This labeling is given by $\tilde{z}_i=\min\{\lceil (i)^{-1}/h,k\rceil\}$, where $(i)^{-1}$ is the rank, from smallest to largest, of the ith element of ξ. Thus, $\tilde{z}$ orders elements of the unobserved vector ξ, sorts the indices of the data according to this ordering, and then groups these indices into sets of size h, with one additional set of size $h+r$.

With the oracle labeling $\tilde{z}$, we may define the graphon oracle estimator from the block averages $\overline{A}^*_{ab}$ according to

$$\overline{A}^*_{ab}=\frac{\sum_{i<j}A_{ij}\mathbb{I}(\tilde{z}_i=a)\mathbb{I}(\tilde{z}_j=b)}{\sum_{i<j}\mathbb{I}(\tilde{z}_i=a)\mathbb{I}(\tilde{z}_j=b)},$$

$$\hat{f}^*(x,y;h)=\rho_n^{-1}\overline{A}^*_{\min(\lceil nx/h\rceil,k)\min(\lceil ny/h\rceil,k)}. \quad [8]$$

Comparing Eq. **8** with its counterpart in Eq. **6**, we see that the oracle serves to replace the estimators of Eqs. **3** and **4** with their ideal quantities. Thus, the oracle estimator is based on a priori knowledge of the sparsity parameter ρ_n and the latent vector ξ. In this sense, it shows the best performance that can be achieved for a fixed bandwidth h, by providing knowledge of the scaling and ordering necessary for the estimator to become a linear function of the data.

Determining the Histogram Bandwidth

Oracle Mean-Square Error Bound. By making use of the network oracle we can determine what performance limits are possible and in turn derive a rule of thumb for selecting the bandwidth h. We assume here that f is differentiable, noting that this result extends to Hölder-continuous functions, as shown in *SI Appendix*.

Theorem 1 (Network Histogram Oracle Bandwidth Selection). *Assume that h grows more slowly than n, and that the graphon $f(x,y)$ is differentiable, with a gradient magnitude bounded by M. Then as n grows the oracle mean integrated square error satisfies the bound*

$$\mathrm{MISE}\left(\hat{f}^*\right)\le M^2\left\{2\left(\frac{h}{n}\right)^2+\frac{1}{n}+\frac{1}{M^2}\left(\frac{1}{h^2\rho_n}\right)\right\}\{1+o(1)\}.$$

The right-hand side of this expression is minimized at $h=h^*$:

$$h^*=\left(2M^2\rho_n\right)^{-1/4}\cdot\sqrt{n}, \quad [9]$$

whence $\mathrm{MISE}(\hat{f}^*)$ evaluated at h^* decays at the rate $1\Big/\sqrt{\binom{n}{2}\rho_n}$:

$$\mathrm{MISE}\left(\hat{f}^*\right)\Big|_{h=h^*}\le M^2\left[\frac{2}{M}\left\{\binom{n}{2}\rho_n\right\}^{-1/2}+\frac{1}{n}\right]\{1+o(1)\}. \quad [10]$$

Proof: We evaluate Eq. **7** with $\hat{f}$ set equal to $\hat{f}^*$ as defined in Eq. **8**, and with $\sigma(x)$ set equal to x to obtain an upper bound on the error criterion $\mathrm{MISE}\left(\hat{f}^*\right)$. This yields the bias–variance decomposition

$$\begin{aligned}\mathrm{MISE}\left(\hat{f}^*\right)&\le\mathbb{E}\iint_{(0,1)^2}\left|f(x,y)-\hat{f}^*(x,y;h)\right|^2dx\,dy\\&=\sum_{a,b=1}^{k}\iint_{\omega_{ab}}\left\{\left|f(x,y)-\rho_n^{-1}\mathbb{E}\,\overline{A}^*_{ab}\right|^2+\rho_n^{-2}\mathrm{Var}\overline{A}^*_{ab}\right\}dx\,dy,\end{aligned}$$

with ω_{ab} the domain of integration corresponding to the block $\overline{A}_{ab}$. Now let $\bar{f}_{ab}=|\omega_{ab}|^{-1}\iint_{\omega_{ab}}f(x,y)dx\,dy$ be the average value of f over ω_{ab}, and $\overline{f^2}_{ab}$ the average value of f^2. Using the assumed smoothness of f in a manner quantified by *Proposition 1* in *SI Appendix*, we substitute for $\mathrm{Var}\overline{A}^*_{ab}$ and $\mathbb{E}\,\overline{A}^*_{ab}$ to obtain

$$\begin{aligned}\mathrm{MISE}\left(\hat{f}^*\right)&\le\sum_{a,b=1}^{k}\iint_{\omega_{ab}}\left[\left|\{f(x,y)-\bar{f}_{ab}\}+\left\{\bar{f}_{ab}-\rho_n^{-1}\mathbb{E}\,\overline{A}^*_{ab}\right\}\right|^2\right.\\&\qquad\left.+\frac{\bar{f}_{ab}-\rho_n\overline{f^2}_{ab}}{\rho_nh^2_{ab}}+\frac{M\{1+o(1)\}}{\rho_nh^2_{ab}(2n)^{1/2}}+\frac{M^2}{2n}\right]dx\,dy\\&\le\sum_{a,b=1}^{k}\left[\iint_{\omega_{ab}}\left|f(x,y)-\bar{f}_{ab}\right|^2dx\,dy+\left\{\frac{M^2\{1+o(1)\}}{2n}\right.\right.\\&\qquad\left.\left.+\frac{\bar{f}_{ab}-\rho_n\overline{f^2}_{ab}}{\rho_nh^2_{ab}}+\frac{M\{1+o(1)\}}{(2n)^{1/2}}\frac{1}{\rho_nh^2_{ab}}+\frac{M^2}{2n}\right\}\frac{h^2_{ab,r}}{n^2}\right],\end{aligned}$$

with $h^2_{ab}=\sum_{i<j}\mathbb{I}(\tilde{z}_i=a)\mathbb{I}(\tilde{z}_j=b)$ and $h^2_{ab,r}=\{h+r\,\mathbb{I}(a=k)\}\{h+r\,\mathbb{I}(b=k)\}$. Applying *Lemma 1* in *SI Appendix* to each $\iint_{\omega_{ab}}\left|f(x,y)-\bar{f}_{ab}\right|^2dx\,dy$,

$$\sum_{a,b=1}^{k}\iint_{\omega_{ab}}\left|f(x,y)-\bar{f}_{ab}\right|^2dx\,dy\le M^2\cdot2\left(\frac{h}{n}\right)^2\left\{1+\mathcal{O}\left(\frac{h}{n}\right)\right\},$$

with the $\mathcal{O}(h/n)$ term due to the grouping of size $h+r$. Using Eq. **2**,

$$\sum_{a,b=1}^{k} \frac{\bar{f}_{ab}}{\rho_n h_{ab}^2} \frac{h_{ab,r}^2}{n^2} = \sum_{a,b=1}^{k} \frac{1}{\rho_n h_{ab}^2} \iint_{\omega_{ab}} f(x,y)dx\,dy = \frac{1}{\rho_n h^2}\{1+o(1)\}.$$

Combining these simplifications yields the stated expression.

This theorem informs the selection of a network histogram bandwidth h. It quantifies how the oracle integrated mean square error depends on the smoothness of the graphon f, relative to the size and sparsity of the observed adjacency matrix A. The theorem decomposes this error into three contributions: smoothing bias, which scales as $M^2(h/n)^2$; resolution bias, which scales as M^2/n; and variance contributions, which scale as the inverse of the effective degrees of freedom $h^2\rho_n$ of each bin. As shown in ref. 16, ensuring that $h^2\rho_n$ grows faster than $\log^3 n$ will enable consistent estimation of the graphon when z is estimated according to Eq. **4**; this accounts for the additional variance involved in estimating z in the nonoracle setting.

Theorem 1 subsequently enables us to choose a bandwidth h that respects the global properties of the network. If we were to know ρ_n and M, then the theorem provides directly for an oracle choice of bandwidth h^* according to Eq. **9**. From this expression we see that for the case of a dense network, with $\rho_n \propto 1$, the oracle choice of bandwidth h^* scales as $\sqrt{n}$. More generally, we observe that as the sparsity of the network increases h^* must also increase, whereas as the gradient magnitude of the graphon increases h^* must decrease. If f is not differentiable but is still Hölder-continuous, then the Hölder exponent will appear in the theorem expressions, leading to a smaller bandwidth for a given n and ρ_n.

Finally, *Theorem 1* provides for an upper bound on the oracle mean integrated square error when the network histogram bandwidth is set equal to h^*. This bound reveals the best possible estimation performance we might achieve for given values of n, ρ_n, and M.

Automatic Bandwidth Selection. *Theorem 1* is important for our theoretical understanding of the bandwidth selection problem, because it shows the tradeoffs between sparsity, smoothness, and sample size. It suggests that h should grow at a rate proportional to $\rho_n^{-1/4}\sqrt{n}$, with ρ_n estimated via Eq. **3**, and with a constant of proportionality depending on the squared magnitude M^2 of the graphon gradient.

To estimate M^2 from A, we will form a simple one-dimensional approximation of the graphon f using the vector d of sorted degrees. This yields a nonparametric estimator for what is referred to as the canonical version of $\int_0^1 f(x,y)dy$ (3). Whenever the smoothness of this canonical marginal is equivalent to that of f, then this procedure yields a suitable estimator $\widehat{M^2}$ according to the steps below. In some instances, however, the marginal may be smoother than f; for example, let $B(x)$ denote the distribution function of a Beta(a,b) random variable, and suppose $f(x,y) \propto B^{-1}(x)B^{-1}(y) + B^{-1}(1-x)B^{-1}(1-y)$. Then the marginal is constant if $a = b$, but the corresponding M^2 (and indeed the Hölder regularity of f) will depend on a and b.

To proceed, assume that the rows and column of A have been reordered such that $d_i = \sum_{j\neq i} A_{ij}$ is increasing with i. Enumerating the sampled elements $f(\xi_i, \xi_j)$ of the graphon in a $n \times n$ matrix F under this same reordering, we obtain in analogy to Eq. **6** a rank-one estimate of the sampled graphon as $\hat{F} \propto \hat{\rho}_n^+ dd^T$. Minimizing the Frobenius norm $\|\hat{F} - \hat{\rho}_n^+ A\|$ then leads to $\hat{F} = [\{(d^Td)^+\}^2 \hat{\rho}_n^+ d^TAd]dd^T$.

We then use $\hat{F}$ to estimate the bandwidth h as follows:

i) Compute the vector d of degrees of A; sort its entries.

ii) Estimate the slope of the ordered d over indices $\lfloor n/2 \rfloor \pm \lfloor c\sqrt{n} \rfloor$ for some choice of c; normally $c = 4$ is appropriate. Treating the ordered entries of d near $\lfloor n/2 \rfloor$ as a set of observations, fit a line with slope m and intercept b using the system of equations

$$d_{\lfloor n/2 \rfloor + j} = jm + b, \quad j = \lfloor -c\sqrt{n} \rfloor, -\lfloor c\sqrt{n}+1 \rfloor, \ldots, \lfloor c\sqrt{n} \rfloor.$$

By the method of least squares, this yields estimates $\hat{m}$ and $\hat{b}$.

iii) Define the vector-valued function of first differences

$$\Delta f(x,y) = \left(f(x,y) - f\left(x + \frac{1}{n+1}, y\right) \quad f(x,y) - f\left(x, y + \frac{1}{n+1}\right) \right)^T,$$

leading to the following gradient estimate:

$$\widehat{\Delta f} = \left[\left\{ (d^Td)^+ \right\}^2 \hat{\rho}_n^+ d^TAd \right] (\hat{m}\hat{b} \quad \hat{m}\hat{b})^T.$$

Via $\left\| \widehat{\Delta f} \right\|^2$, we estimate the average squared magnitude of Δf:

$$\widehat{M^2} = 2n^2 \left\{ (d^Td)^+ \right\}^4 (\hat{\rho}_n^+)^2 (d^TAd)^2 \hat{m}^2 \hat{b}^2 \{1+o(1)\}. \qquad [11]$$

iv) Substituting $\widehat{M^2}$ into Eq. **9**, we obtain the bandwidth estimate

$$\hat{h}^* = \left(2\widehat{M^2}\hat{\rho}_n\right)^{-\frac{1}{4}} \sqrt{n} = \left(2\left\{ (d^Td)^+ \right\}^2 d^TAd \cdot \hat{m}\hat{b}\right)^{-\frac{1}{2}} \hat{\rho}_n^{\frac{1}{4}}. \qquad [12]$$

Equipped with this rule of thumb for selecting the bandwidth h, we can now calculate the network histogram $\hat{f}(x,y;\hat{h}^*)$.

Data Analysis Using Network Histograms

Data analysis software to calculate the network histogram is available at github.com/p-wolfe/network-histogram-code.

Political Weblog Data. To demonstrate the utility of the network histogram, we first analyze a well-studied dataset of political weblogs described in ref. 18 and illustrated in Fig. 1. This dataset was collected to quantify the degree of interaction between liberal and conservative blogs around the time of the 2004 US presidential election and consists of a snapshot of nearly 1,500 weblogs from February 8, 2005. An edge is considered to be present between two weblogs whenever at least one of the weblogs' front page links to the other.

The relative sparsity of conservative–liberal weblog linkages in this dataset is clearly apparent from Fig. 1. Thus, it is often used to illustrate the notion of network community structure (see, e.g., ref. 7). At the same time, Fig. 1 also makes clear that the dataset exhibits additional heterogeneity not fully captured by a simple division of its weblogs into two communities, and indeed recent work also provides evidence of its additional block structure (21). Thus, the network histogram provides a natural tool to explore the data.

Fig. 2 shows a fitted histogram $\hat{f}(x,y)$ obtained from the $n = 1{,}224$ weblogs with at least one link to another weblog in the dataset. From Eq. **11** we obtained an estimate $\widehat{M^2}$ in the range 1.1–1.25 for c in the range 3–5, and so the estimated oracle error bound of Eq. **10** evaluates to $\sim 1.8 \times 10^{-2}$. The bandwidth $\hat{h}^*$ was then determined using Eq. **12** and was found to evaluate to 72–74 for c in the range 3–5. We rounded this to $h = 72$ to obtain the $k = 17$ equal-sized histogram bins that comprise Figs. 2 and 3. The marginal edge probability estimator $\hat{\rho} = \sum_{i<j} A_{ij} \Big/ \binom{n}{2}$ evaluates to 16,715/748,476 = 2.2332×10^{-2}, implying that each off-diagonal histogram bin has ~116 effective degrees of freedom.

Because exact maximization of the likelihood of Eq. **4** is known to be computationally infeasible, we obtained the fit shown in Fig. 2 by implementing a simple stochastic search algorithm that swaps pairs and triples of node group memberships selected at random until a local optimum is reached in the likelihood of Eq. **4**. The log-likelihood of the data under the fitted model, normalized by the estimated effective degrees of freedom $\binom{n}{2}\hat{\rho}$, is –2.8728. To explore as full a range as possible of local likelihood optima, we

started from several hundred random configurations, inspected the largest 5% of returned local maxima, and then repeatedly reoptimized after randomly swapping up to 100 group membership pairs in the best returned solution.

The histogram bin index, relative to the x and y axes of Fig. 2, allows comparison with the leftmost panel of Fig. 1. Bin indices are arranged first by majority grouping—liberal or conservative—and then by the strength of each fitted group's cross-party connections. Each node's political affiliation can be viewed as an observed binary covariate that partially explains the network structure. Below we will consider the more general setting of multiple categorical covariates.

As summarized in Figs. 2 and 3, the coarsest feature of this network is its polarization into sets of dense linkages within the two political blocs of liberal and conservative ideologies. We also observe from Fig. 2 that nearly 40% of the histogram bins are empty, in keeping with the sparsity pattern of the data observed in the leftmost panel of Fig. 1. The most densely connected groups of weblogs in both parties show considerable cross-party linkage structure. This is apparent both from the center region of Fig. 2, as well as the groupings of Fig. 3, in which the most influential weblogs identified by ref. 18 are seen to be placed in the center of the histogram. Such features are examples of network microstructure, corresponding to variation at scales smaller than the large fractions of a network that would be captured by a blockmodel with a fixed number of groups.

Student Friendship Data. Network datasets often have additional covariates measured at nodes or edges. To illustrate how to use such information to interpret network histograms, we analyze a student friendship network from the US National Longitudinal Study of Adolescent Health (Add Health) (22). As part of this study, students were asked to identify their sex, race, and school year (grades 7–12) and then to nominate up to five friends of each sex. We consider an undirected version of the resulting network, with a link present whenever either of a pair of students has nominated the other.

We chose to analyze School 44 from the Add Health study, a relative large and racially diverse example among the over 80 schools for which data were collected (23), and one that has been previously analyzed in ref. 24 using exponential random graph models. It comprises a main high school with grades 9–12 and a sister "feeder" school with grades 7 and 8. We removed 21 zero-degree nodes as well as five nodes corresponding to students for which any two of sex, grade, or race covariates were missing, yielding $n = 1,122$ nodes.

To fit the histogram shown in Fig. 4, we used the same bandwidth selection procedure and optimization algorithm as above. This yielded a bandwidth $\widehat{h^*}$ in the range 69–70 for c in the range 3–5, which we rounded down to $h = 66$ to obtain $k = 17$ equal-sized histogram bins. This is sparser than the political weblog network considered above, but at the same time $\widehat{M^2}$ evaluates to 3.2–3.5, indicating relatively less smoothness. The estimated oracle

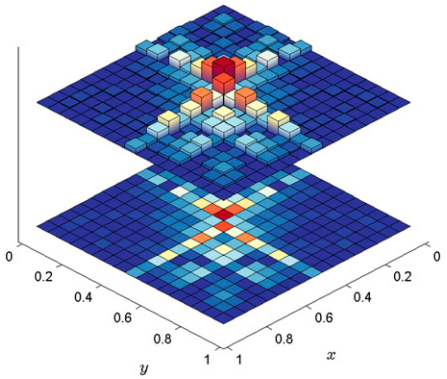

Fig. 2. Network histogram $\hat{f}(x,y)^{\frac{1}{2}}$ fitted to political weblog data. The square root stabilizes the variance of the bin heights and is solely for ease of visualization.

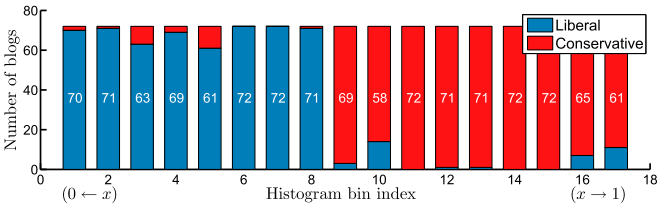

Fig. 3. Political affiliation of weblogs within each fitted group, ordered relative to Fig. 2. Affiliation counts are shown in white, out of 72 weblogs per group. Nineteen of the 20 most influential liberal weblogs identified by ref. 18 are assigned to group 8, and 17 of the top 20 conservative blogs are assigned to group 9.

error bound of Eq. **10** is then $\sim 5.6 \times 10^{-2}$, and our fit yielded a normalized data log-likelihood of −4.1714. For this example, the marginal edge probability estimator $\hat{\rho} = \sum_{i<j} A_{ij} \Big/ \binom{n}{2}$ evaluates to $5{,}048/628{,}881 = 8.0270 \times 10^{-3}$, implying that each off-diagonal histogram bin has ∼35 effective degrees of freedom.

To explore the fitted groups, we ordered them post hoc via the mean covariate value per bin for race (coded 0–5), grade (coded 6–12), and number of friends nominated (coded 0–10). The resulting histograms are shown in the top row of Fig. 4, and the bottom row shows the number of covariate categories comprising each bin. In the leftmost column of Fig. 4 we observe that the connectivity structure associated with race divides most of the white and black students into two separate groupings, with a decreased tendency to link across these categories. In the middle column we observe a similar effect for grade, as well as an even stronger effect between the two separate schools: Students in grades 7–8 have relatively few interactions with students in grades 9–12. There is evidence for more mixing within the latter school, with the exception of grade 12, whereas in the former school the division between grades 7 and 8 is strong. Finally, in the rightmost column of Fig. 4 we see a strong effect associated with the number of friends nominated, which serves as a rough proxy for the degree of each network node. Diagonal bins in this histogram are ordered almost exclusively from smallest to largest, and we see none of the assortativity associated with race or grade that was so apparent in the previous histogram orderings.

From this example we conclude that the network histogram can provide not only an effective summary of network interactions, but one which is also interpretable in the context of additional covariate information. This type of aggregate summary allows a fine-grained but concise view of adolescent student friendship networks, and suggests that aggregate statistics on race and grade within a particular school may not be sufficient to give a full picture of the reported social interactions among its students.

Discussion

We argue that the blockmodel is universal as a tool for representing interactions in an unlabeled network. As we use more blocks in our representation, we improve our approximation of the underlying data-generating mechanism, albeit at the cost of increasing complexity. The results in this paper give us insight into how to control the tradeoff between complexity and precision, leading to a flexible nonparametric summary of a network akin to an ordinary histogram.

There is a clear philosophical distinction between the network histogram and the stochastic blockmodel. The network histogram yields a nonparametric summary of link densities across a network. In contrast, the stochastic blockmodel was originally conceived as a generative statistical model, meaning that it is typically analyzed in settings where it is presumed to be correctly specified as the data-generating mechanism. We have instead shown how it can be useful in the case when the blockmodel serves simply to approximate the generating mechanism of the network—a much milder assumption.

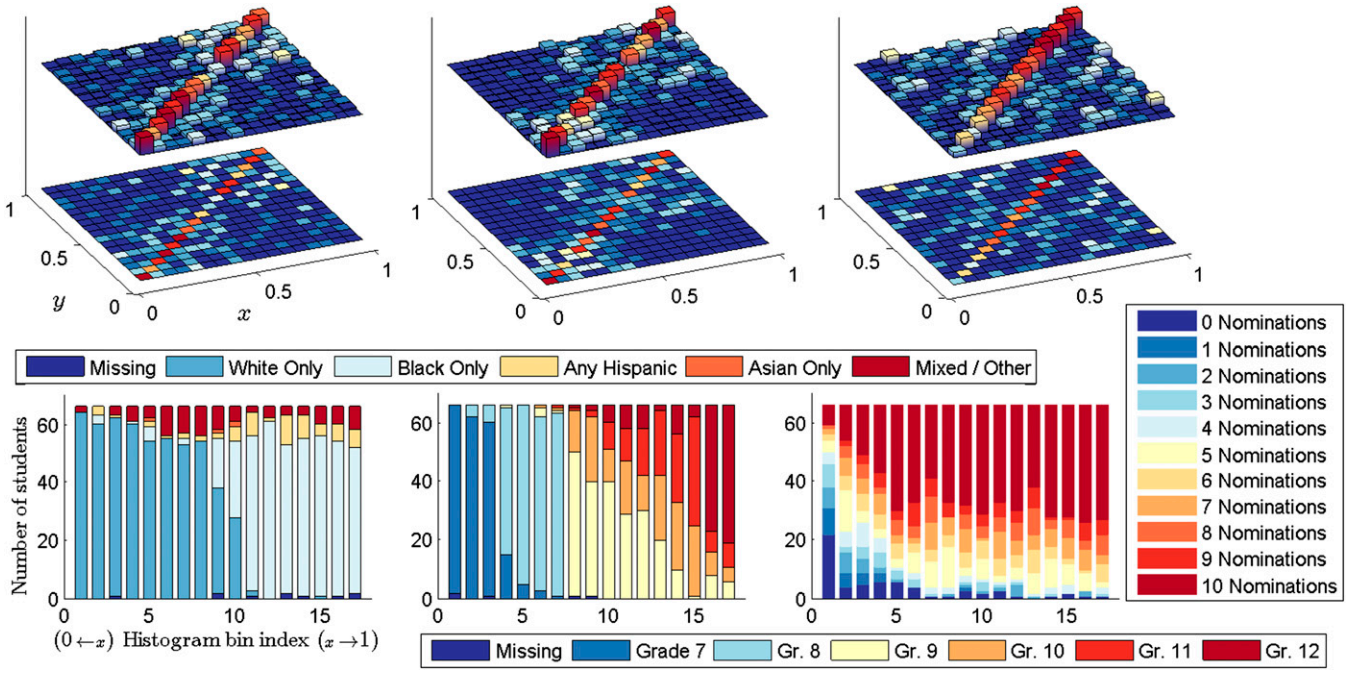

Fig. 4. Network histogram $\hat{f}(x,y)^{\frac{1}{2}}$ fitted to student friendship data (*Upper*), with bins ordered according to mean covariate value for race (*Lower Left*), school year (*Lower Center*), and number of friend nominations (*Lower Right*). The histogram structure visible with respect to each of these three covariates is discussed in the text.

To make the network histogram into a useful practical tool, we have derived a procedure for automatically selecting an analysis bandwidth under the assumption of a smooth (Hölder-continuous) graphon. If the graphon has finitely many discontinuities parallel to its x and y axes, for example if it corresponds to an actual blockmodel, then good estimation properties can still be achieved, in analogy to ordinary histogram estimates (25). In such scenarios the rates at which estimation errors decay are not yet established; indeed, exploring different graphon smoothness classes, and the networks they give rise to, remains an important avenue of future investigation.

As a final point, networks are rarely explored in the absence of other data. A network histogram is defined only up to permutation of its bins, and so to aid in its interpretation we may use other observed variables, labels, or covariates to inform our choice of bin ordering. As our second data analysis example has shown in the context of student friendship networks, multiple representations can be useful in different ways, and more than one such visual representation can yield insight into the generating mechanism of the network. In this way the universality of the blockmodel representation is a key piece in the puzzle of general network understanding. Our results suggest a fundamental rethinking of the interpretation of network communities, in light of the fact that many different community assignments can all give an equally valid representation of the network.

ACKNOWLEDGMENTS. This work was supported in part by the US Army Research Office under Presidential Early Career Award for Scientists and Engineers W911NF-09-1-0555 and Multidisciplinary University Research Initiative Award W911NF-11-1-0036; by the US Office of Naval Research under Award N00014-14-1-0819; by the UK Engineering and Physical Sciences Research Council under Mathematical Sciences Leadership Fellowship EP/I005250/1, Established Career Fellowship EP/K005413/1, and Developing Leaders Award EP/L001519/1; by the UK Royal Society under a Wolfson Research Merit Award; and by Marie Curie FP7 Integration Grant PCIG12-GA-2012-334622 within the 7th European Union Framework Program.

1. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99(12):7821–7826.
2. Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: First steps. *Soc Networks* 5(2):109–137.
3. Bickel PJ, Chen A (2009) A nonparametric view of network models and Newman-Girvan and other modularities. *Proc Natl Acad Sci USA* 106(50):21068–21073.
4. Zhao Y, Levina E, Zhu J (2011) Community extraction for social networks. *Proc Natl Acad Sci USA* 108(18):7321–7326.
5. Airoldi EM, Blei DM, Fienberg SE, Xing EP (2008) Mixed membership stochastic blockmodels. *J Mach Learn Res* 9:1981–2014.
6. Karrer B, Newman MEJ (2011) Stochastic blockmodels and community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 83(1 Pt 2):016107.
7. Newman MEJ (2011) Communities, modules and large-scale structure in networks. *Nat Phys* 8:25–31.
8. Gopalan PK, Blei DM (2013) Efficient discovery of overlapping communities in massive networks. *Proc Natl Acad Sci USA* 110(36):14534–14539.
9. Zhao Y, Levina E, Zhu J (2012) Consistency of community detection in networks under degree-corrected stochastic block models. *Ann Stat* 40(4):2266–2292.
10. Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. *Ann Stat* 39(4):1878–1915.
11. Choi DS, Wolfe PJ, Airoldi EM (2012) Stochastic blockmodels with a growing number of classes. *Biometrika* 99(2):273–284.
12. Chatterjee S (2012) Matrix estimation by universal singular value thresholding. arXiv:1212.1247.
13. Fishkind DE, Sussman DL, Tang M, Vogelstein JT, Priebe CE (2013) Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J Matrix Anal Appl* 34(1):23–39.
14. Bickel PJ, Sarkar P (2013) Hypothesis testing for automated community detection in networks. arXiv:1311.2694.
15. Choi DS, Wolfe PJ (2014) Co-clustering separately exchangeable network data. *Ann Stat* 42(1):29–63.
16. Wolfe PJ, Olhede SC (2013) Nonparametric graphon estimation. arXiv:1309.5936.
17. Lovász L (2012) *Large Networks and Graph Limits* (Am. Mathematical Soc, Providence, RI).
18. Adamic L, Glance N (2005) The political blogosphere and the 2004 US election: Divided they blog. *Proceedings of the 3rd International Workshop on Link Discovery* (ACM, New York), pp 36–43.
19. Airoldi EM, Costa TB, Chan SH (2013) Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY), Vol 26, pp 692–700.
20. Tsybakov AB (2009) *Introduction to Nonparametric Estimation* (Springer, Berlin).
21. Krzakala F, et al. (2013) Spectral redemption in clustering sparse networks. *Proc Natl Acad Sci USA* 110(52):20935–20940.
22. Resnick MD, et al. (1997) Protecting Adolescents from Harm: Findings from the National Longitudinal Study on Adolescent Health. *JAMA* 278(10):823–832.
23. Moody J (2001) Race, school integration, and friendship segregation in America. *Am J Sociol* 107(3):679–716.
24. Hunter DR, Goodreau SM, Handcock MS (2008) Goodness of fit of social network models. *J Am Stat Assoc* 103(481):248–258.
25. van Eeden C (1985) Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous. *Ann Inst Stat Math* 37(1): 461–472.