*Biometrika* (2021), **00**, 0, *pp.* 1–14
*Printed in Great Britain*
doi: 10.1093/biomet/asab057



# Local linear graphon estimation using covariates

BY S. CHANDNA

*Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, U.K.*
s.chandna@bbk.ac.uk

S. C. OLHEDE

*Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne, Station 8, 1015 Lausanne, Switzerland*
sofia.olhede@epfl.ch

AND P. J. WOLFE

*Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, Indiana 47907, U.S.A.*
patrick@purdue.edu

## SUMMARY

We consider local linear estimation of the graphon function, which determines probabilities of pairwise edges between nodes in an unlabelled network. Real-world networks are typically characterized by node heterogeneity, with different nodes exhibiting different degrees of interaction. Existing approaches to graphon estimation are limited to local constant approximations, and are not designed to estimate heterogeneity across the full network. In this paper, we show how continuous node covariates can be employed to estimate heterogeneity in the network via a local linear graphon estimator. We derive the bias and variance of an oracle-based local linear graphon estimator, and thus obtain the mean integrated squared error optimal bandwidth rule. We also provide a plug-in bandwidth selection procedure that makes local linear estimation for unlabelled networks practically feasible. The finite-sample performance of our approach is investigated in a simulation study, and the method is applied to a school friendship network and an email network to illustrate its advantages over existing methods.

*Some key words*: Exchangeable network; Graph limit; Graphon estimation; Nonparametric regression.

## 1. INTRODUCTION

The easy availability of network data has fundamentally changed data analysis in a variety of fields, ranging from social to biological applications. This has led to burgeoning interest in novel statistical methods that take network dependence into account. This paper concerns estimation of the graphon, the generative mechanism for unlabelled networks, formulated using the framework of exchangeability (e.g., Diaconis & Janson, 2007). Recently, there has been growing interest in the problem of graphon estimation (e.g., Chan & Airoldi, 2014; Olhede & Wolfe, 2014; Cai et al., 2015; Gao et al., 2015; Klopp et al., 2017; Pensky, 2019) and its applications to,

for example, bootstrapping network data (Green & Shalizi, 2017), testing for equivalence of network distribution using subgraph counts (Maugis et al., 2020), and estimating missing links (Borgs & Chayes, 2017). The existing literature on graphon function estimation focuses entirely on local constant approximations, providing the extremely useful histogram tool for analysing complex networks. Although histogram estimators yield a meaningful clustering of nodes, they are not designed to allow estimation of the full degree of heterogeneity across the network. As networks grow in size and complexity, they typically exhibit heterogeneous behaviour, along with the absence of homogeneous communities (Leskovec et al., 2009), suggesting that the use of histogram approximations may be limited only to exploratory analysis. Motivated by these issues, we propose a novel local linear graphon estimator that uses covariates to account for node heterogeneity, and enables improved graphon estimation.

We consider the setting where a single undirected network without self-loops is observed along with continuous covariates at each node. For a network comprising $n$ nodes, pairwise interactions between nodes are represented by an $n \times n$ adjacency matrix $A = (A_{ij})$, where $A_{ij} = 1$ represents the presence, and $A_{ij} = 0$ the absence, of interaction between nodes $i$ and $j$. Further, $A_{ij} = A_{ji}$ for $1 \leqslant i,j \leqslant n$, and $A_{ii} = 0$ because there are no self-loops. Using the notion of exchangeable arrays (e.g., Hoover, 1979; Aldous, 1981), the $A_{ij}$ $(i \leqslant j)$ are modelled as independent Ber$(p_{ij})$ trials for binary networks, where the interaction or edge probabilities $p_{ij}$ are determined by the graphon $f : [0,1]^2 \to [0,\infty)$, a bivariate symmetric measurable function. Specifically, $p_{ij} \propto f(\xi_i, \xi_j)$ where $\xi_1, \ldots, \xi_n$ denote independent latent node variables. Since the graphon provides a generative mechanism for unlabelled networks, as in the existing literature, we study local linear estimation of the graphon function as an element of an equivalence class that is invariant under symmetric rearrangement of its axes.

From the model described above, it is clear that graphon estimation is a nonparametric regression problem with the pairwise interactions $A_{ij}$ as the observed response corresponding to latent design points $\xi_i$ and $\xi_j$. Given this latency of design points, it is natural to construct histogram approximations by clustering pairwise interactions into bins of a suitable size. In the case of a single network observation, this has been achieved, for instance, by Chan & Airoldi (2014) and Yang et al. (2014) under the restrictive assumption of strict monotonicity of the degree sequence, by Olhede & Wolfe (2014) through a combinatorial likelihood approach, by Cai et al. (2015) via iterative clustering based on the cut-metric, and by Pensky (2019) through penalized least squares. The aforementioned approaches, as well as related works on the special problem of probability matrix estimation (e.g., Chatterjee, 2015; Zhang et al., 2017; Padilla, 2019), with the exception of Su et al. (2020), are based on the adjacency matrix alone and ignore the commonly available covariates. In numerous applications, covariates are known to provide information complementary to pairwise interactions; therefore, the integration of covariates is pivotal and is a topic of growing research interest (e.g., Hoff et al., 2002; Perry & Wolfe, 2013; Zhang et al., 2016; Binkiewicz et al., 2017; Latouche et al., 2018; Yan et al., 2019).

Since $p_{ij} \propto f(\xi_i, \xi_j)$, node heterogeneity in the nonparametric framework described above is clearly determined by the node-specific design points $\{\xi_i\}_{i=1}^n$. In practice, these are never observed; however, node covariates are commonly recorded. Empirically observed heterogeneity in the network can often be attributed to heterogeneity in the nodes expressed through covariates. For example, in a contact network of employees, differences in interactions between employees can be attributed to employee-specific features such as seniority in the profession or popularity. Likewise, in a student friendship network, differences in friendship-forming behaviour can be attributed to school grades. With this in mind, we employ node covariates via an error-in-variable model to inform local linear modelling within disjoint neighbourhoods. Our algorithm for local linear estimation is based on profile least squares with iterative updates for the neighbourhood

assignment of nodes and covariate-informed local linear model-fitting. Our approach may be viewed as analogous to classical scatter plot smoothing (e.g., Härdle & Marron, 1995), but is designed to deal with the additional challenges arising from the latency of design points. Key to our graphon estimation procedure is the size of the neighbourhoods, or the bandwidth, which must be determined appropriately from the observed network. Given the absence of design points, we do this by studying the theoretical properties of an oracle-informed local linear graphon estimator. Our results show how local linear modelling leads to reduced bias relative to local constant fitting, and results in a mean integrated squared error that decays much faster than that of the network histogram of Olhede & Wolfe (2014). The mean integrated squared error optimal bandwidth choice is derived and found to be determined by, the number of nodes in the network, global sparsity, and a measure of network variability defined in terms of the Laplacian of the graphon. We provide a procedure based on spectral decomposition for estimating this measure of network variability from the adjacency matrix, yielding a simple plug-in bandwidth selection rule.

The finite-sample performance of our method assessed via synthetic graphons with different properties, under both dense and sparse regimes, suggests that our approach can lead to significantly improved estimation, even when covariates do not provide ideal proxies for latent design points. Application of our method to a student friendship network from the Add Health dataset and to a network of email interactions from the Enron database reveals key structural features that are not observed when estimation is performed using existing approaches, with our method showing improved prediction of edge probabilities.

## 2. Preliminaries

A simple stochastic model for an undirected, unlabelled binary network of size $n$, represented by an $n \times n$ adjacency matrix $A = (A_{ij})$, is given in Definition 1 (Hoover, 1979; Aldous, 1981). For convenience, we use $[n]$ to denote $\{1, \ldots, n\}$ for any positive integer $n$.

Definition 1. *Let $\{\xi_i\}_{i=1}^n$ be a sequence of independently distributed* $\mathrm{Un}[0,1]$ *random variables. Then $A_{ij} = A_{ji}$ for $(i,j) \in [n] \times [n]$ are modelled as conditionally independent Bernoulli trials, $A_{ij} \mid \xi_i, \xi_j \sim \mathrm{Ber}\{\rho_n f(\xi_i, \xi_j)\}$, where the graphon function $f : [0,1]^2 \to [0,\infty)$ is a symmetric measurable function, and the scaling factor $\rho_n > 0$, assumed to be nonincreasing in $n$, allows for globally sparse networks (e.g., Bollobás & Riordan, 2011).*

For statistical identifiability of $\rho_n$, it is assumed that $f$ integrates to unity, i.e., $\iint_{(0,1)^2} f(u,v)\,\mathrm{d}u\,\mathrm{d}v = 1$. Then, noting that $E(A_{ij}) = \mathrm{pr}(A_{ij} = 1) = \rho_n$, we estimate it as the proportion of nonzero edges in the network, i.e., $\hat{\rho}_n = \binom{n}{2}^{-1} \sum_{i<j} A_{ij}$ (e.g., Olhede & Wolfe, 2014). We let $p_{ij} = \rho_n f(\xi_i, \xi_j)$ denote the probability of an edge between nodes $i$ and $j$, and use $p(\xi_i, \xi_j) = p_{ij}$ to denote the corresponding probability function.

## 3. Methodology

### 3.1. *Local linear modelling*

Because of the lack of traditional geometry on the space of unlabelled networks, a neighbourhood is commonly defined in terms of disjoint blocks comprising a fixed number of nodes, similar to communities (e.g., Kolaczyk & Csárdi, 2014), rather than in terms of distance from a focal point as in classical nonparametric regression. Given a suitable bandwidth $1 < h \leqslant n$, representing the number of nodes in each block, we divide the total number of nodes $n$ into

$k = \lfloor n/h \rfloor$ blocks, writing $n = hk + r$ where $r = n \bmod h$ is the remainder lying between 0 and $h - 1$. Given $n$ and $h$ such that $n = hk + r$, let $\mathcal{Z}_{n,h} \subseteq \{1, \ldots, k\}^n$ contain all block assignment vectors $z = (z_1, \ldots, z_n)^{\mathrm{T}}$ with $h$ components equal to each of the integers from 1 to $k - 1$, up to relabelling, and $h + r$ components equal to $k$, again up to relabelling. Then for node pair $(i,j) \in [n] \times [n]$ such that $z_i = a$ and $z_j = b$, where $(a,b) \in [k] \times [k]$, we consider a local linear model (e.g., Fan & Gijbels, 1996) for the unknown probability function $p$,

$$p(\xi_i, \xi_j) = \tilde{\kappa}_{ab,0} + \tilde{\kappa}_{ab,1}\xi_i + \tilde{\kappa}_{ab,2}\xi_j, \tag{1}$$

where $\tilde{\kappa}_{ab,0}, \tilde{\kappa}_{ab,1}$ and $\tilde{\kappa}_{ab,2}$ denote the local linear coefficients for block pair $(a,b)$. As mentioned in the introduction, the node-specific design points $\{\xi_i\}_{i=1}^n$ are latent, but node covariates $\{x_i\}_{i=1}^n$ are commonly observed and may be used to explain heterogeneity in interactions (e.g., Fosdick & Hoff, 2015). We integrate covariate information via an error-in-variable model where node covariates are modelled as proxies for unobserved node positions, i.e.,

$$x_i = \xi_i + \epsilon_i, \tag{2}$$

where $\{\epsilon_i\}_{i=1}^n$ denotes the error term with mean zero and variance $\sigma^2$, such that $\{\xi_i, \epsilon_i\}_i$ are mutually independent. Then, upon replacing the latent variable $\xi_i$ by $x_i$ and likewise $\xi_j$ by $x_j$ in (1), we get $p(\xi_i, \xi_j) = \kappa_{ab,0} + \kappa_{ab,1}x_i + \kappa_{ab,2}x_j$, where the change in coefficients from $\tilde{\kappa}_{ab,l}$ to $\kappa_{ab,l}$ $(l = 0, 1, 2)$ takes into account the change in regressors from $\xi$ to $x$. We use (2) to employ node covariates $x_i$ as proxies, or measurements, for $\xi_i$ to explain heterogeneity in pairwise interactions within blocks via linear associations. Thus, estimation of the probability function $p$ in our framework is a nonparametric regression problem with unknown neighbourhoods in the form of disjoint blocks, and with node covariates $\{x_i\}$ as regressors to explain local structure. For simplicity, we proceed with the case of a single covariate at each node; integration of multiple covariates follows similarly, as described in the Supplementary Material. For further discussion on integration of covariates via (2), see the Supplementary Material.

### 3.2. *Least-squares estimator*

The least-squares approach to local constant graphon estimation has been studied, for example, by Gao et al. (2015) and Klopp et al. (2017). In contrast, we are concerned with local linear least-squares estimation, and hence are not restricted to the class of block-constant matrices. Given our local linear model, let $\kappa_{ab} = (\kappa_{ab,0}, \kappa_{ab,1}, \kappa_{ab,2})^{\mathrm{T}}$ denote the vector of local linear coefficients and $z \in \mathcal{Z}_{n,h}$ the block assignment vector. Consider the residual sum of squares,

$$L(\kappa, z; A, X) = \sum_{\{(a,b)\in[k]\times[k]\}} \sum_{\{(i,j)\in z^{-1}(a)\times z^{-1}(b),\, i<j\}} l_{ij}(\kappa_{ab}, A, X), \tag{3}$$

where $l_{ij}(\kappa_{ab}, A, X) = (A_{ij} - \kappa_{ab}^{\mathrm{T}} X_{ij})^2$, $X_{ij} = (1, x_i, x_j)^{\mathrm{T}}$ is the vector of regressors, and $\kappa = (\kappa_{ab})_{(a,b)\in[k]\times[k]} \in \mathbb{R}^{3\times k\times k}$ denotes the full set of $k^2$ local linear coefficient vectors. Thus, the least-squares estimator of $(\kappa, z)$ is

$$(\hat{\kappa}, \hat{z}) \in \underset{\kappa\in\mathbb{R}^{3\times k\times k},\, z\in\mathcal{Z}_{n,h}}{\arg\min} L(\kappa, z; A, X).$$

Given the dependence of local linear coefficients $\kappa_{ab}$ on the unknown block assignment vector $z$, we solve this optimization problem via profile least-squares estimation with iterative updates,

as outlined in the Supplementary Material. Then, for any $(i,j) \in [n] \times [n]$ such that $\hat{z}_i = a$ and $\hat{z}_j = b$, the local linear edge probability estimate is $\hat{p}_{ij} = \hat{\kappa}_{ab,0} + \hat{\kappa}_{ab,1}x_i + \hat{\kappa}_{ab,2}x_j$, which implies the corresponding graphon estimate $\hat{f}_{ij} = \hat{\rho}_n^{-1}\hat{p}_{ij}$. In general, for any given point $(u,v) \in (0,1)^2$, setting $a = \min(\lceil nu/h \rceil, k)$ and $b = \min(\lceil nv/h \rceil, k)$ we define $\hat{f}(u,v) = \hat{\rho}_n^{-1}\hat{p}(u,v)$ where $\hat{p}(u,v) = \hat{\kappa}_{ab,0} + \hat{\kappa}_{ab,1}u + \hat{\kappa}_{ab,2}v$. Our local linear graphon estimator depends on the size of neighbourhoods, or the bandwidth $h$, which must be selected to control the bias-variance trade-off, one of the key issues in nonparametric regression. This is the topic of the next section.

## 4. Bandwidth selection

### 4.1. *Oracle-based local linear graphon estimator*

Given the latency of the design points $\{\xi_i\}$, we employ an oracle (e.g., Donoho & Johnstone, 1994; Olhede & Wolfe, 2014) that supplies information about the true neighbourhoods to construct a theoretically tractable local linear graphon estimator. Consider the oracle that provides order statistics of the unobserved node positions $\{\xi_i\}_{i=1}^n$, i.e., for each $i \in [n]$ we are given $(i)$ such that $\xi_{(1)} \leqslant \xi_{(2)} \leqslant \cdots \leqslant \xi_{(n)}$. We use this information to construct the oracle block assignment vector $\bar{z}^* = (\bar{z}_1^*, \ldots, \bar{z}_n^*)^{\mathrm{T}}$, where $\bar{z}_i^* = \min\{\lceil (i)^{-1}/h \rceil, k\}$ with $(i)^{-1}$ denoting the rank, ordered from smallest to largest, of the $i$th element $\xi_i$ (Olhede & Wolfe, 2014). A fixed block neighbourhood design as given by $\bar{z}^*$ is, in general, asymmetric about a point of estimation and leads to a theoretically intractable local linear estimator. We therefore use the order statistics of $\{\xi_i\}_{i=1}^n$ to construct a symmetric neighbourhood around a given point of estimation (e.g., Fan & Gijbels, 1996), as detailed below; see the Supplementary Material for further discussion.

Define $i_n = i/(n+1)$. Since $E(\xi_{(i)}) = i_n$ for $\xi_i \sim \mathrm{Un}[0,1]$ (e.g., Davison & Hinkley, 2007), for a given $u$ in the interior of $[0,1]$, i.e., $h/2n \leqslant u \leqslant 1 - h/2n$ (e.g., Ruppert & Wand, 1994), we define the oracle neighbourhood indicator vector $z^*(u;h) = \{z_1^*(u), \ldots, z_n^*(u)\}^{\mathrm{T}}$, where $z_i^*(u) = 1$ if $(i)_n^{-1} \in [u - h/2n,\ u + h/2n]$ and $z_i^*(u) = 0$ otherwise. Given (2), the oracle must also inform the choice of ideal design points. Using the fact that $E(\xi_{(i)}) = i_n$, we define $x_i^* = i_n$ $(i \in [n])$ as oracle measurements for unobserved $\xi_{(i)}$. Let $X_{ij}^* = (1, x_i^*, x_j^*)^{\mathrm{T}}$, and let $A_{ij}^* = A_{(i)(j)}$ denote the adjacency with nodes rearranged to correspond to increasing node positions $\{\xi_{(i)}\}$. Then, given $(u,v) \in (0,1)^2$, following (3), the residual sum of squares in the oracle setting is $L^*(\gamma_{uv}) = \sum_{\{(i,j):\, i<j\}} l_{ij}(\gamma_{uv}, A^*, X^*) z_i^*(u) z_j^*(v)$, where $\gamma_{uv} = (\gamma_{uv,0}, \gamma_{uv,1}, \gamma_{uv,2})^{\mathrm{T}} \in \mathbb{R}^{3\times 1}$ denotes the vector of local linear coefficients. There are implicit assumptions on the parameter space, namely that $\gamma_{uv}^{\mathrm{T}} X_{ij}^* \in [0,1]$ for all $(i,j) \in [n] \times [n]$ where $z_i^*(u) z_j^*(v) = 1$ (e.g., Battey et al., 2019). The least-squares estimator $\hat{\gamma}_{uv} = \hat{\gamma}_{uv}(z^*)$ is $\hat{\gamma}_{uv} = \arg\min_{\gamma_{uv}} L^*(\gamma_{uv})$. This leads to the oracle local linear graphon estimator $\hat{f}^*(u,v;h) = \hat{\rho}_n^{-1}\hat{p}^*(u,v)$, where $\hat{p}^*(u,v) = \hat{\gamma}_{uv,0} + \hat{\gamma}_{uv,1}u + \hat{\gamma}_{uv,2}v$. Expressions for the bias and variance of this oracle local linear graphon estimator for a bandwidth $h > 1$ are given in the following theorem.

Theorem 1. *Given a bandwidth $h > 1$, assume that $h/n \to 0$ as $n \to \infty$, $h = \omega(\sqrt{n})$, and the graphon $f : [0,1]^2 \to [0,\infty)$ is twice differentiable with continuous second-order partial derivatives. Let $f_u$ and $f_v$ denote the first-order partial derivatives with respect to the first and second variables, respectively. Then, as $n \to \infty$, for any interior point $(u,v) \in (0,1)^2$,*

$$\mathrm{bias}\{\hat{f}^*(u,v;h)\} = \frac{\mu_2(\mathcal{K})}{2}\frac{h^2}{n^2}\left\{\frac{\partial^2 f}{\partial u^2}(u,v) + \frac{\partial^2 f}{\partial v^2}(u,v)\right\}\{1 + o(1)\}$$

*and*

$$\mathrm{var}\{\hat{f}^*(u,v;h)\} = \left[\frac{\bar{f}_\omega - \rho_n \bar{f}_\omega^2}{\rho_n h^2} + 2\bar{f}_{u;\omega}\bar{f}_{v;\omega}\frac{(u+v)}{(n+2)}\left\{\frac{2n}{n+1} - (u+v)\right\}\right]\{1+o(1)\}, \quad (4)$$

*where* $\mu_2(\mathcal{K})$ *denotes the second moment with respect to the kernel* $\mathcal{K} \equiv \mathrm{Un}[-1/2, 1/2]$, *implying* $\mu_2(\mathcal{K}) = 1/12$. *Further, we denote by* $\bar{f}_\omega = |\omega_{uv}|^{-1}\int\int_{\omega_{uv}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y$, $\bar{f}_\omega^2 = |\omega_{uv}|^{-1}\int\int_{\omega_{uv}} f^2(x,y)\,\mathrm{d}x\,\mathrm{d}y$, $\bar{f}_{u;\omega} = |\omega_{uv}|^{-1}\int\int_{\omega_{uv}} f_u(x,y)\,\mathrm{d}x\,\mathrm{d}y$ *and* $\bar{f}_{v;\omega} = |\omega_{uv}|^{-1}\int\int_{\omega_{uv}} f_v(x,y)\,\mathrm{d}x\,\mathrm{d}y$ *the corresponding local averages over the size-h oracle region* $\omega_{uv} = [u - h/2n, u + h/2n] \times [v - h/2n, v + h/2n]$.

The leading bias term captures how the curvature of graphon $f$ controls the bias of the local linear graphon estimator, with higher bias at points corresponding to higher curvature. The product of the unmixed second-order partial derivative(s) of $f$ with bandwidth $n^{-2}h^2$ indicates the amount of smoothing that must be performed to reduce bias. The variance given by the first term in (4) scales as the inverse of the effective degrees of freedom, i.e., $(\rho_n h^2)^{-1}$ in each neighbourhood. The second term in (4) is the covariance term arising from the order statistics of $\{\xi_i\}_{i=1}^n$ and is determined by local averages of first-order partial derivatives of $f$, specifically $\bar{f}_{u;\omega}$ and $\bar{f}_{v;\omega}$.

*Remark* 1. Writing $H_n = \mathrm{diag}(n^{-2}h^2, n^{-2}h^2)$, the bias term in Theorem 1 is expressed as $\mathrm{bias}\{\hat{f}^*(u,v;h)\} = 0.5\mu_2(\mathcal{K})\,\mathrm{tr}\{H_n\mathcal{H}_f(u,v)\}\{1+o(1)\}$, where tr denotes trace and $\mathcal{H}_f(u,v)$ denotes the $2\times 2$ Hessian matrix of graphon $f$ at $(u,v)$. This is similar in form to the asymptotic bias obtained in classical local linear regression (Ruppert & Wand, 1994, Theorem 2.1). Likewise, the variance of our estimator scales as $(\rho_n h^2)^{-1}$, which can be expressed as $(n^{-2}h^2 \times n^2\rho_n)^{-1} = (|H_n|^{1/2} \times \text{effective sample size})^{-1}$ as in the classical setting.

Using the properties of our local linear estimator given in Theorem 1, we obtain the mean integrated squared error optimal bandwidth rule.

PROPOSITION 1. *Under the conditions of Theorem 1, as* $n \to \infty$, *the mean integrated squared error* $\mathrm{MISE}(\hat{f}^*)$ *of the oracle local linear estimator satisfies*

$$\mathrm{MISE}(\hat{f}^*) \leqslant \left\{\frac{h^4}{4n^4}\mu_2^2(\mathcal{K})\psi_{2,f} + (\rho_n h^2)^{-1} + \frac{5n}{3(n+1)(n+2)}\max_{(u,v)\in[0,1]^2}\bar{f}_{u;\omega}\bar{f}_{v;\omega}\right\}\{1+o(1)\},$$

*where* $\psi_{2,f} = \int\int_{(0,1)^2}\Delta f(u,v)^2\,\mathrm{d}u\,\mathrm{d}v$, *with* $\Delta f(u,v)$ *denoting the Laplacian of* $f$ *at* $(u,v)$. *This leads to the mean integrated squared error optimal bandwidth*

$$h^* = \left(\frac{2}{\rho_n\mu_2^2(\mathcal{K})\psi_{2,f}}\right)^{1/6} n^{2/3}. \quad (5)$$

With $n$, $\rho_n$ and $\psi_{2,f}$ specified, Proposition 1 provides a bandwidth selection rule $h^*$ that respects the global properties of the network. Specifically, an increase in the bandwidth with an increase in network size $n$ is dictated by a trade-off between the sparsity of the network and its structural variability as measured by $\psi_{2,f}$. The rule in (5) governs how the bandwidth should increase with decreasing network density specified via $\rho_n$, and explains how a high overall variability in the

network must imply a smaller bandwidth. With $h = h^*$, it follows that

$$\mathrm{MISE}(\hat{f}^*)\Big|_{h=h^*} \leqslant \left\{ \frac{3\mu_2^{2/3}(\mathcal{K})\psi_{2,f}^{1/3}}{2^{4/3}\rho_n^{2/3}n^{4/3}} + \frac{5n}{3(n+1)(n+2)} \max_{(u,v)\in[0,1]^2} \bar{f}_{u;\omega}\bar{f}_{v;\omega} \right\} \{1 + o(1)\}.$$

Thus, the mean integrated squared error of the local linear estimator decays at a rate of $(n^{4/3}\rho_n^{2/3})^{-1}$, which is much faster than the $(n\rho_n{}^{1/2})^{-1}$ rate at which the mean integrated squared error of the histogram decays, for $\rho_n > n^{-2}$. This is what we would expect, as using a local linear fit leads to a reduced bias, with the contribution from the bias term being proportional to $h^4n^{-4}$, in comparison with $h^2n^{-2}$ for the histogram of Olhede & Wolfe (2014).

### 4.2. *Data-dependent bandwidth selection*

To provide a plug-in bandwidth selection rule based on $h^*$ in (5), the unknown quantity $\psi_{2,f}$ involving derivatives of the unknown function $f$ must be estimated. In classical nonparametric regression, this is typically achieved via a simple parametric fit (e.g., Fan & Gijbels, 1996) or a piecewise-parametric fit (e.g., Ruppert et al., 1995) to the unknown function. However, these approaches are not directly applicable in our setting because of the latency of design points. To overcome this problem, we use the spectral decomposition of $A$ to estimate a rank-$\tau$ representation of $f$ (e.g., Lovász, 2012), which is then used to obtain the partial derivatives of $f$ via global quadratic fitting. This results in an approximation to $\psi_{2,f}$ as given by Algorithm 1.

*Algorithm* 1. Plug-in bandwidth selection.

Input: the $n \times n$ adjacency matrix $A$
Output: $\hat{\psi}_{2,f}$ for plug-in bandwidth selection rule $\hat{h}^* = \lceil h^*(\hat{\rho}_n, \hat{\psi}_{2,f}, n)\rceil$

*Step* 1. Let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of $A$ indexed such that $|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$, with corresponding eigenvectors $v_1, \ldots, v_n$ where $v_l = (v_{l,1}, \ldots, v_{l,n})^{\mathrm{T}}$ for $l \in [n]$.

*Step* 2. Define

$$\hat{\alpha}_l = \left( \frac{\sum_{i=1}^n v_{l,i}}{n} \right)^2 \frac{\lambda_l}{\hat{\rho}_n}, \quad \hat{\kappa}_{l,i} = \frac{nv_{l,(i)}}{\sum_{m=1}^n v_{l,m}},$$

where $\{v_{l,(j)}\}_{j\in[n]}$ are obtained by sorting $\{v_{l,j}\}_{j\in[n]}$ in increasing order.

*Step* 3. Obtain the quadratic fit $\hat{\kappa}_{l,i} = b_{l,0}i_n^2 + b_{l,1}i_n + b_{l,2}$ where $i_n = i/(n+1)$, for instance via least squares.

*Step* 4. Let $\tilde{\tau}$ denote the estimated rank obtained as the smallest positive integer such that

$$\frac{\sum_{l=1}^{\tilde{\tau}+1} |\hat{\alpha}_l|}{\sum_{l=1}^n |\hat{\alpha}_l|} - \frac{\sum_{l=1}^{\tilde{\tau}} |\hat{\alpha}_l|}{\sum_{l=1}^n |\hat{\alpha}_l|} \leqslant c',$$

with $c' \propto 10^{-t}$ for an integer $t \geqslant 3$.

Table 1. *Graphons for numerical experiments with corresponding properties; the theoretical mean integrated squared error optimal bandwidth $h^*$ is reported for $n = 2000$ and $\rho_n = 1/3$*

| Index | $f_{\text{Index}}(u, v)$ | Rank | Monotone degree | $\psi_{2,f}$ | $h^*(\psi_{2,f}; n, \rho_n)$ |
|---|---|---|---|---|---|
| 1 | $\exp(u+v)/[0.7238\{1+\exp(u+v)\}]$ | Full | Yes | 0.05 | 808 |
| 2 | $\frac{3}{4} + \frac{9}{4}u^2v^2$ | Low (2) | Yes | 12.60 | 322 |
| 3 | $\exp(-0.5\lvert u-v\rvert)/0.8522$ | Full | No | 0.25 | 618 |

*Step* 5. Set

$$\hat{\psi}_{2,f} = \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\sum_{l=1}^{\tilde{\tau}}\hat{\alpha}_l b_{l,0}(\hat{\kappa}_{l,j} + \hat{\kappa}_{l,i})\right\}^2.$$

Let $\hat{h}^*$ denote the plug-in bandwidth selection rule obtained by replacing $\psi_{2,f}$ with $\hat{\psi}_{2,f}$ and $\rho_n$ with $\hat{\rho}_n$ in (5). Comparisons of our plug-in bandwidth $\hat{h}^*$ with the true bandwidth $h^*$, for synthetic graphons as in Table 1, are included in the Supplementary Material. From these comparisons, it is evident that our method performs well in both the dense and the sparse regimes; see the Supplementary Material for further details and discussion.

## 5. FINITE-SAMPLE PERFORMANCE

The finite-sample performance of the proposed local linear estimator is investigated using synthetic networks under two density regimes. We report results of comparisons with the network histogram of Olhede & Wolfe (2014), the sort-and-smooth method of Chan & Airoldi (2014), the universal singular-value thresholding of Chatterjee (2015), and the neighbourhood smoothing of Zhang et al. (2017). Graphons used to generate synthetic networks, along with their rank, degree structure, network variability as measured by $\psi_{2,f}$, and corresponding bandwidth $h^*$ for $n = 2000$ and $\rho_n = 1/3$ are listed in Table 1. The true structures implied by these graphons are visualized using heatmaps in the first column of Fig. 1.

We consider a dense regime with $\rho_n \propto 1$ and a sparse regime with $\rho_n \propto n^{-1}\log^3(n)$, dictating how the density of edges decreases with increasing $n$. For each $f$, $\rho_n$ is chosen to coincide in the two regimes for $n = 200$. This leads to a significantly different $\rho_n$ in the sparse regime for the largest value of $n = 1600$, results for which are reported. To study the performance of our local linear estimator in situations where the observed node covariates $x_i$ may not provide ideal measurements for the latent node positions $\xi_i$, we use node covariates generated following (2) where $\epsilon_i \sim N(0, \sigma^2)$, with (i) $\sigma^2 = 0.0001$, implying a high signal-to-noise ratio of 28, referred to as local linear high, and (ii) $\sigma^2 = 0.02$, corresponding to a much lower signal-to-noise ratio of 2, referred to as local linear low.

Table 2 shows mean squared error comparisons averaged over 100 replications; visual comparisons of graphon matrix estimates via heatmaps from a single network of size $n = 1600$ are displayed in Fig. 1. Due to space constraints, we have displayed heatmaps only for $f_3$ in the sparse regime; see the Supplementary Material for $f_1$ and $f_2$. From the mean squared error comparisons in the dense case, we see that for $f_1$ our method yields the best results, with an average mean

Table 2. *Comparisons of mean squared error ($\times 10^3$, with standard deviation in parentheses) averaged over* 100 *replications*

| Graphon | $n$ | $\rho_n$ | Proposed (high) | Proposed (low) | NH | SAS | USVT | NBS |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho_n \propto 1$ | | | | |
| $f_1$ | 800 | 0.70 | 0.60 (0.13) | 0.91 (0.13) | 3.70 (1.40) | 1.00 (0.06) | 1.30 (0.07) | 5.10 (0.19) |
| | 1600 | 0.70 | 0.34 (0.06) | 0.64 (0.10) | 1.40 (0.27) | 0.52 (0.04) | 0.78 (0.03) | 3.10 (0.11) |
| $f_2$ | 800 | 0.28 | 6.60 (2.97) | 8.90 (1.40) | 18.70 (1.80) | 9.50 (0.69) | 13.60 (0.73) | 29.90 (0.66) |
| | 1600 | 0.28 | 3.64 (0.84) | 6.32 (0.98) | 9.00 (0.97) | 5.00 (0.27) | 6.50 (0.28) | 17.80 (0.22) |
| $f_3$ | 800 | 0.71 | 4.00 (1.40) | 5.00 (1.56) | 1.90 (0.45) | 11.10 (0.47) | 3.20 (0.09) | 6.50 (0.11) |
| | 1600 | 0.71 | 2.30 (1.32) | 2.75 (1.70) | 1.20 (0.91) | 10.50 (0.40) | 2.00 (0.04) | 3.70 (0.05) |
| | | | | $\rho_n \propto n^{-1}\log^3(n)$ | | | | |
| $f_1$ | 1600 | 0.24 | 3.40 (0.75) | 3.60 (0.63) | 13.10 (1.90) | 3.50 (0.13) | 4.40 (0.14) | 27.60 (0.16) |
| $f_2$ | 1600 | 0.10 | 11.80 (1.80) | 12.10 (1.60) | 33.10 (1.40) | 16.60 (0.80) | 33.30 (1.00) | 89.10 (0.95) |
| $f_3$ | 1600 | 0.24 | 3.80 (0.50) | 4.40 (0.32) | 10.30 (1.30) | 13.70 (0.38) | 10.10 (0.24) | 34.50 (0.24) |

NH, network histogram of Olhede & Wolfe (2014); SAS, sort-and-smooth method of Chan & Airoldi (2014); USVT, universal singular-value thresholding of Chatterjee (2015); NBS, neighbourhood smoothing of Zhang et al. (2017).

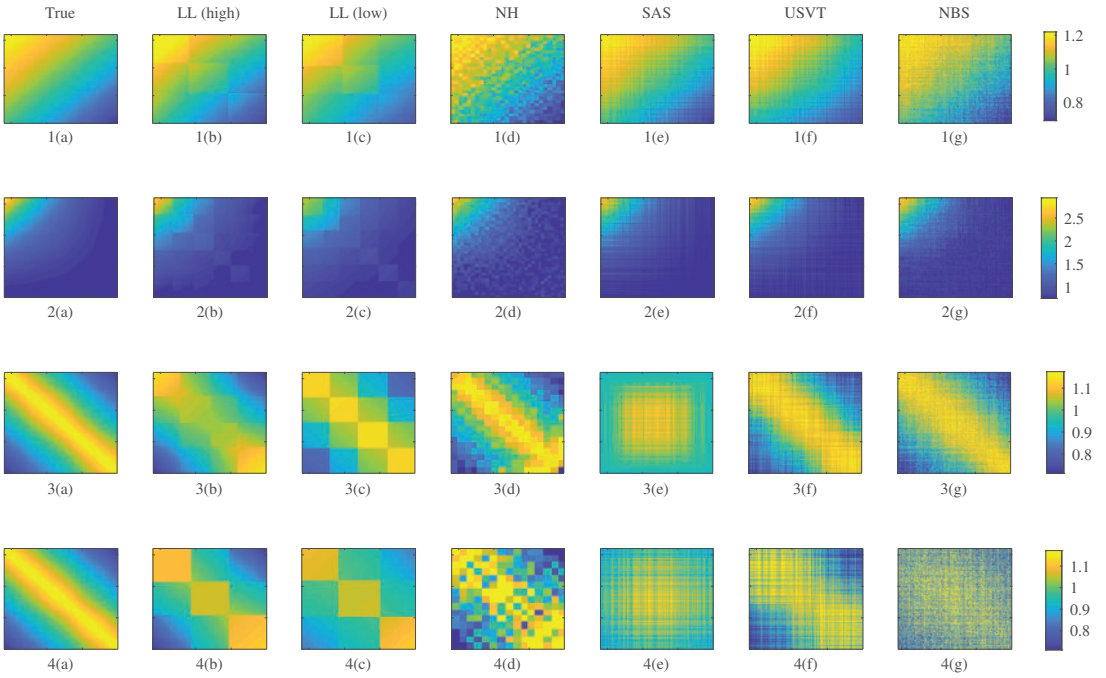


Fig. 1. Graphon estimates from networks of size $n = 1600$, with $f_1$ and $\rho_n \propto 1$ in the first row, $f_2$ and $\rho_n \propto 1$ in the second row, $f_3$ and $\rho_n \propto 1$ in the third row, and $f_3$ and $\rho_n \propto n^{-1}\log^3(n)$ in the fourth row, using the proposed local linear method (LL) under high and low signal-to-noise ratio settings in (b) and (c) and using the network histogram (NH), sort-and-smooth method (SAS), universal singular-value thresholding (USVT) and neighbourhood smoothing (NBS) in (d)–(g).

squared error of 7.75 for $n = 800$ and 4.98 for $n = 1600$ over the two signal-to-noise ratio settings; the sort-and-smooth and singular-value thresholding methods give comparable results. In contrast, the network histogram and neighbourhood smoothing do not perform well. For $f_2$, the best results are obtained from our approach, which gives an average mean squared error of 7.75 for $n = 800$ and 4.98 for $n = 1600$, and from the sort-and-smooth method for each $n$. The other

methods have relatively high mean squared errors. For $f_3$, the sort-and-smooth approach clearly has the worst performance, which can also be seen in Fig. 1. This is what one would expect given the nonmonotonic degree structure of $f_3$. From Table 2 it is evident that for $f_3$ the best-performing methods are the network histogram and singular-value thresholding. The proposed local linear method, with an average mean squared error of 4.50 for $n = 800$ and 2.53 for $n = 1600$, and neighbourhood smoothing yield the second-best results.

From the mean squared error comparisons in the sparse regime, we see that for $f_1$, as in the dense case, local linear estimation continues to perform the best, with the mean squared errors from sort-and-smooth and singular-value thresholding being comparable. For $f_2$, our method leads to the smallest mean squared error, with sort-and-smooth producing the second-best results; however, the mean squared error of all other methods is at least three times as high as the mean squared error of our approach. Interestingly, for $f_3$ in the sparse regime, our approach leads to the smallest mean squared error, with the mean squared errors of all other methods being at least twice as high. This is in contrast to the dense setting, where the network histogram and singular-value thresholding give the best results, and is also apparent from the last row of Fig. 1, where bands with different intensities along the diagonal are clearly visible in the local linear estimates; neighbourhood smoothing and sort-and-smooth estimates, on the other hand, are unable to identify these key structural components.

To summarize, our method consistently performs well, significantly outperforming existing approaches in the sparse regime of greater practical interest, and giving the best or second-best results in the dense regime. On average, the results in Table 2 imply a 27% reduction in mean squared error in the high signal-to-noise ratio case, and a 10% reduction in the low signal-to-noise ratio case from using the proposed method, in comparison to the best-performing of the network histogram, sort-and-smooth, singular-value thresholding and neighbourhood smoothing methods.

## 6. DATA ANALYSIS

### 6.1. *School friendship network*

We illustrate local linear graphon estimation by applying it to a student friendship network comprising 994 students, collected as part of the U.S. National Longitudinal Study of Adolescent Health, called Add Health for short (Resnick et al., 1997). One of the objectives of Add Health is to examine friendship patterns among high school students. Students participating in the survey were asked to report their school grade, coded 7–12; race, coded 1–5, with 0 representing a missing response; and sex, coded female or male. They were then asked to nominate up to five friends of each sex. For our example we select School 75, one of the more heterogeneous schools, with 62% of participants reporting their race as 'White only', coded as 1, 28% as 'Black only', coded as 2, and 10% from the remaining three categories, 3–5. Using the data on friendship nominations, an undirected network was constructed with $A_{ij} = 1$ if either student $i$ or student $j$ nominated the other as a friend. We illustrate our approach with the school grade covariate observed for each student. As a pre-processing step, nodes with missing grade information were removed, leaving a network of size $n = 987$ and an edge density of $\hat{\rho}_n = 0.0089$. Using Algorithm 1 with $c'$ varying between $10^{-3}$ and $10^{-7}$, we obtained bandwidths in the range of 138–145; we select $\hat{h}^* = 141$, as it implies seven equally sized blocks, or bins.

The reported race and grade attributes of nodes in the estimated local linear and network histogram bins are displayed in Fig. 2. Observing the attributes of students in histogram bins, it is apparent that each bin has a clear majority of its nodal attributes in one of two reported

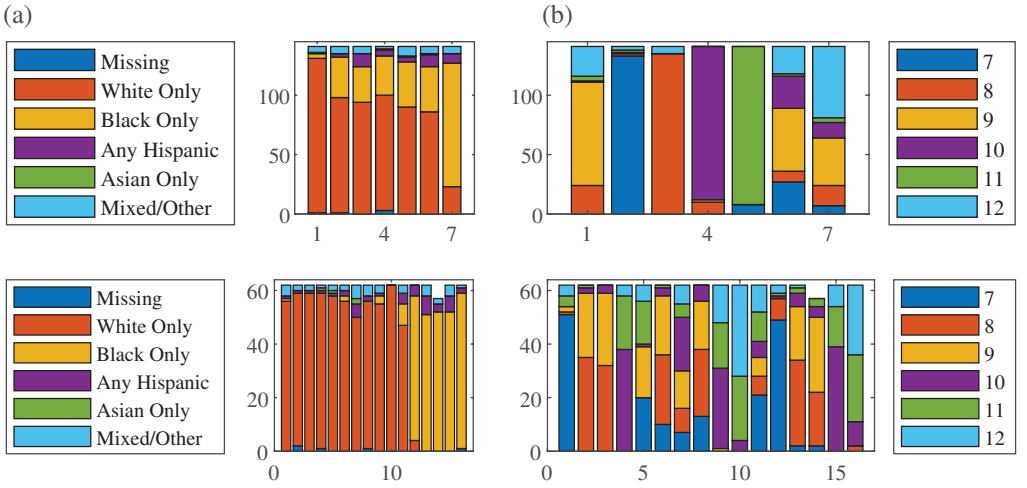


Fig. 2. (a) Race and (b) corresponding grade attributes reported by students, for bins identified via our local linear approach (top row) and via the network histogram (bottom row). The bin index is marked along the horizontal axis, and the counts of students with different reported race attributes and school grades (7–12) are marked on the vertical axis (see legends).

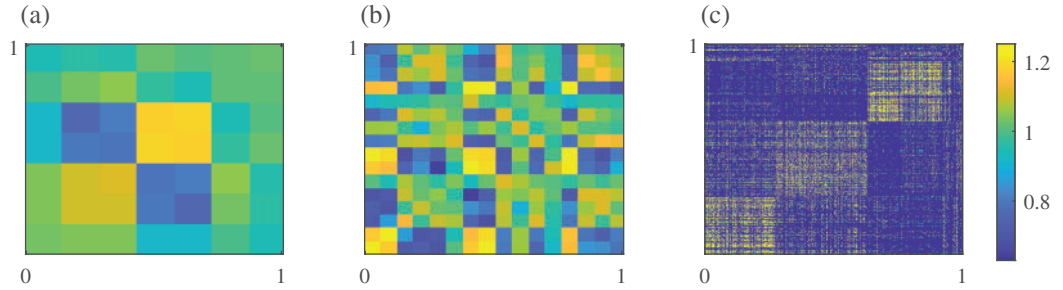


Fig. 3. Graphon estimates $\hat{f}^{1/4}$ by race code for the school network, using (a) the proposed local linear method, (b) the network histogram, and (c) neighbourhood smoothing.

race categories,'White only', and 'Black only', and this corresponds to a weak majority across most grade categories, 7–12, as is visible from the rightmost subplot in the second row. This is not surprising as each grade includes students from these two reported race categories. Our method, on the other hand, yields a balance between the race and grade attributes of bins; for example, there is a clear majority in grades 7–11 across bins, and bins 1–6 contain mostly students reporting their race as 'White only', while bin 7 contains mostly students reporting their race as 'Black only', which evidently includes students from each of the six different grades. These differences in clustering are reflected in the corresponding graphon estimates displayed in Fig. 3, which also includes the neighbourhood smoothing method of Zhang et al. (2017) for comparison. Here the histogram and local linear bins as labelled in Fig. 2, i.e., by race code, are positioned on the diagonal from $(0, 0)$ to $(1, 1)$; likewise, the graphon matrix estimate from neighbourhood smoothing has nodes arranged by race code. Our local linear graphon estimate in Fig. 3(a) clearly identifies a division of the friendship network into three key components: (i) bins 2 and 3, (ii) bins 4 and 5, and (iii) bins 1, 6 and 7, with an assortative community-like structure displayed jointly by the first two components, and with low-intensity hub-like interactions of students in the third component. The network histogram is clearly unable to identify structure stemming from nodes

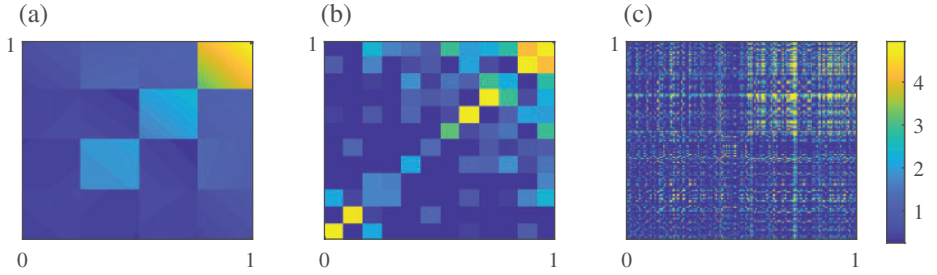


Fig. 4. Graphon estimates $\hat{f}$ by seniority for the Enron network, using (a) our proposed local linear method, (b) the network histogram, and (c) neighbourhood smoothing.

belonging to the same grade, as expected. Further comparisons, reported in the Supplementary Material, show that (i) unlike the histogram method of Olhede & Wolfe (2014), our local linear estimate is able to identify structure irrespective of the choice of bin labelling, i.e., by race code or grade, and (ii) our local linear estimate leads to the lowest root mean squared prediction error and log loss when used for prediction. From comparisons reported here and in the Supplementary Material, it is also evident that matrix estimation techniques such as neighbourhood smoothing and singular-value thresholding are harder to interpret owing to the absence of a clustering of nodes.

### 6.2. *Enron email network*

The Enron email corpus comprises a subset of the email messages exchanged within the Enron corporation between 1998 and 2002. We analyse the dataset compiled by Zhou et al. (2007), which consists of 21 635 messages sent between 156 employees from 13 November 1998 to 21 June 2002. In order to exclude messages sent en masse to large groups, emails with more than five recipients were removed from the analysis. We consider the undirected binary network of $n = 156$ nodes, i.e., employees, where $A_{ij} = 1$ if node $i$ either sent an email to or received an email from node $j$ and $A_{ij} = 0$ otherwise. This implies $\hat{\rho}_n = 0.1524$. The dataset also includes information on the directed count of 'send' and 'receive' interactions; node covariates such as seniority, classified as 'junior' or 'senior'; department, i.e., 'legal', 'trading' or 'other'; and the sex of employees. Motivated by Perry & Wolfe (2013), we illustrate our local linear graphon estimation using a receive-senior covariate, $x_i$, defined as the proportion of emails received by node $i$ from senior nodes. Following our bandwidth selection procedure, we obtain $\hat{h}^* = 39$, implying $r = 0$, and hence four bins of the same size. A comparison of graphon estimates is displayed in Fig. 4, where each local linear and histogram bin is identified as 'junior' or 'senior' via a majority vote based on the seniority of the employees making up the bin. Subsequently, bins labelled $1, 2, \ldots$ are arranged on the diagonal with the highest junior majority at $(0, 0)$ and the highest senior majority at $(1, 1)$. Likewise, the graphon estimate of Zhang et al. (2017) is displayed in Fig. 4(c), with nodes arranged by seniority; see the Supplementary Material for comparisons with the sort-and-smooth method and singular-value thresholding.

At a coarse level, these estimates appear similar to each other, with high edge intensities near $(1, 1)$ and significantly lower edge intensities on the off-diagonal; however, a feature unique to our local linear estimate is the separation of the network into four weakly connected components. The employees in bin 1, containing a majority of junior employees, are almost disconnected from the rest of the network because of low estimated edge intensities; bins 2 and 3, also with a majority of junior employees, display an assortative community-like behaviour, whereas bin 4, with a majority

of senior employees, displays a high degree of heterogeneity and low-intensity interactions with bins 2 and 3. Regarding the attributes, we find that bin 4 comprises senior employees with pivotal positions such as 'President and CEO' or 'VP of Regulatory Affairs', whereas bin 1 consists of junior employees from all three departments with roles such as 'Attorney', 'Associate' and 'Analyst'. Further comparisons, reported in the Supplementary Material, confirm that our local linear estimation method, not only provides a meaningful visualization of the underlying structure, but also leads to the lowest root mean squared prediction error and log loss.

## Acknowledgement

The authors thank an associate editor and two referees for very helpful suggestions. Wolfe is also affiliated with the Department of Computer Science and the Department of Electrical & Computer Engineering at Purdue University.

## Supplementary material

Supplementary Material available at *Biometrika* online includes proofs of Theorem 1 and Proposition 1, the derivation of Algorithm 1, and further results and discussion.

## References

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Mult. Anal.* **11**, 581–98.
Battey, H. S., Cox, D. R. & Jackson, M. V. (2019). On the linear in probability model for binary data. *R. Soc. Open Sci.* **6**, article no. 190067, DOI: 10.1098/rsos.190067.
Binkiewicz, N., Vogelstein, J. T. & Rohe, K. (2017). Covariate-assisted spectral clustering. *Biometrika* **104**, 361–77.
Bollobás, B. & Riordan, O. (2011). Sparse graphs: Metrics and random models. *Random Struct. Algor.* **39**, 1–38.
Borgs, C. & Chayes, J. (2017). Graphons: A nonparametric method to model, estimate, and design algorithms for massive networks. In *Proc. 2017 ACM Conf. Economics and Computation*. New York: Association for Computing Machinery, pp. 665–72.
Cai, D., Ackerman, N. & Freer, C. (2015). An iterative step-function estimator for graphons. *arXiv:* 1412.2129v2.
Chan, S. H. & Airoldi, E. (2014). A consistent histogram estimator for exchangeable graph models. In *Proc. 31st Int. Conf. Machine Learning (ICML'14)*. JMLR, pp. 208–16.
Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43**, 177–214.
Davison, A. C. & Hinkley, D. V. (2007). *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.
Diaconis, P. & Janson, S. (2007). Graph limits and exchangeable random graphs. *arXiv:* 0712.2749.
Donoho, D. L. & Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–55.
Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, vol. 66 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall/CRC.
Fosdick, B. K. & Hoff, P. D. (2015). Testing and modeling dependencies between a network and nodal attributes. *J. Am. Statist. Assoc.* **110**, 1047–56.
Gao, C., Lu, Y. & Zhou, H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43**, 2624–52.
Green, A. & Shalizi, C. R. (2017). Bootstrapping exchangeable random graphs. *arXiv:* 1711.00813.
Härdle, W. & Marron, J. S. (1995). Fast and simple scatterplot smoothing. *Comp. Statist. Data Anal.* **20**, 1–17.
Hoff, P. D., Raftery, A. E. & Handcock, M. S. (2002). Latent space approaches to social network analysis. *J. Am. Statist. Assoc.* **97**, 1090–8.
Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. Preprint, Institute for Advanced Study, Princeton, New Jersey.
Klopp, O., Tsybakov, A. B. & Verzelen, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* **45**, 316–54.
Kolaczyk, E. D. & Csárdi, G. (2014). *Statistical Analysis of Network Data with R*. New York: Springer.
Latouche, P., Robin, S. & Ouadah, S. (2018). Goodness of fit of logistic regression models for random graphs. *J. Comp. Graph. Statist.* **27**, 98–109.
Leskovec, J., Lang, K. J., Dasgupta, A. & Mahoney, M. W. (2009). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Math.* **6**, 29–123.

LOVÁSZ, L. (2012). *Large Networks and Graph Limits*, vol. 60 of *Colloquium Publications*. Providence, Rhode Island: American Mathematical Society.

MAUGIS, P. A. G., OLHEDE, S. C., PRIEBE, C. E. & WOLFE, P. J. (2020). Testing for equivalence of network distribution using subgraph counts. *J. Comp. Graph. Statist.* **29**, 455–65.

OLHEDE, S. C. & WOLFE, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proc. Nat. Acad. Sci.* **111**, 14722–7.

PADILLA, O. H. M. (2019). Graphon estimation via nearest neighbor algorithm and 2D fused lasso denoising. *arXiv:* 1805.07042v4.

PENSKY, M. (2019). Dynamic network models and graphon estimation. *Ann. Statist.* **47**, 2378–403.

PERRY, P. O. & WOLFE, P. J. (2013). Point process modelling for directed interaction networks. *J. R. Statist. Soc.* B **75**, 821–49.

RESNICK, M. D., BEARMAN, P. S., BLUM, R. W., BAUMAN, K. E., HARRIS, K. M., JONES, J., TABOR, J., BEUHRING, T., SIEVING, R. E., SHEW, M. et al. (1997). Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health. *J. Am. Med. Assoc.* **278**, 823–32.

RUPPERT, D., SHEATHER, S. J. & WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Am. Statist. Assoc.* **90**, 1257–70.

RUPPERT, D. & WAND, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–70.

SU, Y., WONG, R. K. W. & LEE, T. C. M. (2020). Network estimation via graphon with node features. *IEEE Trans. Network Sci. Eng.* **7**, DOI: 10.1109/TNSE.2020.2973994.

YAN, T., JIANG, B., FIENBERG, S. E. & LENG, C. (2019). Statistical inference in a directed network model with covariates. *J. Am. Statist. Assoc.* **114**, 857–68.

YANG, J., HAN, C. & AIROLDI, E. (2014). Nonparametric estimation and testing of exchangeable graph models. In *Proc. 17th Int. Conf. Artificial Intelligence and Statistics (AISTATS 2014)*. JMLR, pp. 1060–7.

ZHANG, Y., LEVINA, E. & ZHU, J. (2016). Community detection in networks with node features. *Electron. J. Statist.* **10**, 3153–78.

ZHANG, Y., LEVINA, E. & ZHU, J. (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika* **104**, 771–83.

ZHOU, Y., GOLDBERG, M., MAGDON-ISMAIL, M. & WALLACE, A. (2007). Strategies for cleaning organizational emails with an application to Enron email dataset. In *5th Conf. North American Association for Computational Social and Organizational Science*.

[*Received on* 26 *March* 2020. *Editorial decision on* 18 *October* 2021]