Biometrika (2021), **xx**, x, *pp*. 1–17
Printed in Great Britain

Dimension Reduction for Covariates in Network Data

BY JUNLONG ZHAO
School of Statistics, Beijing Normal University, Beijing 100875, China
zhaojunlong928@126.com

XIUMIN LIU
School of Statistics, Beijing Normal University, Beijing100875, China
liuxiumin880407@yeah.net

HANSHENG WANG
Guanghua School of Management, Peking University, Beijing 100871, China
hansheng@gsm.pku.edu.cn

AND CHENLEI LENG
Department of Statistics, University of Warwick, Coventry, CV47AL, U.K.
C.Leng@warwick.ac.uk

SUMMARY

A problem of major interest in network data analysis is to explain the strength of connections using context information. To achieve this, we introduce a novel approach named network-supervised dimension reduction by projecting covariates onto low-dimensional spaces for revealing the linkage pattern, without assuming a model. We propose a new loss function for estimating the parameters in the resulting linear projection, based on the notion that closer proximity in the low-dimension projection renders stronger connections. Interestingly, the convergence rate of our estimator is shown to depend on a network effect factor which is the smallest number that can partition a graph in a way similar to the graph coloring problem. Our methodology has interesting connections to principal component analysis and linear discriminant analysis, which we exploit for clustering and community detection. The methodology developed is further illustrated by numerical experiments and the analysis of a pulsar candidates data in astronomy.

Some key words: Clustering; Community detection; Dimension reduction; Graph; Network.

1. INTRODUCTION

Network data that include multiple objects with measurements on interaction between pairs of objects are becoming increasingly common in a wide variety of fields (Holland & Leinhardt, 1981; Wolfe, 1997; Jin et al., 2001; Newman et al., 2002; Watts et al., 2002; Newman & Park, 2003; Newman, 2006; Sarkar & Moore, 2005; Hunter et al., 2008; Kolaczyk, 2009; Goldenberg et al., 2010; Fienberg, 2012; Scott, 2017). The topology of a network is often represented as a graph denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \ldots, n\}$ is the set of n nodes, and $\mathcal{E}$ is the set of edges among nodes. The relationships among nodes can be described by an adjacency matrix $W = (w_{ij}) \in \mathbb{R}^{n \times n}$, where w_{ij} is some measure of the connection strength between node i and

j. For an unweighted graph, w_{ij} is binary in that $w_{ij} = 1$ indicates the existence of connection and $w_{ij} = 0$ indicates otherwise. For a weighted graph,
colorblack$w_{ij} \geq 0$ represents the strength of connection. The methodology developed in this paper works for undirected and directed graphs. As a reminder, for a directed graph, $w_{ij} > 0$ if there is a directed edge from i to j, and $w_{ij} = 0$ otherwise. For an undirected graph, W is symmetric in that $w_{ij} = w_{ji}$ for any $i \neq j$.

A distinctive feature of many network datasets is that they often come with covariate information collected at the node or edge level. For example, a participant in an online social network can be contextualized by its gender, social status, education and so on, while edge variables measured on pairs of participants, such as whether two participants share common interest or attend the same school, may be present. One of the main purposes of network analysis is to explain the linking pattern w_{ij} by using information in $X_{ij} = (X_{ij,1}, \ldots, X_{ij,p})^T$, a p-dimensional covariate vector between node i and j. In practice, p, the dimension of the covariates, can be large. When only nodal covariates are available, a general way of defining these edge covariates is to construct X_{ij} as a bivariate function of $X_i = (X_{i,1}, \ldots, X_{i,p})^T$ and $X_j = (X_{j,1}, \ldots, X_{j,p})^T$, the node covariates of the ith and jth node. Popular choices in the literature include
colorblack$X_{ij,t} = X_{i,t} - X_{j,t}$ $(t = 1, \ldots, p)$ if the tth covariate is continuous, and
colorblack$X_{ij,t} = I(X_{i,t} \neq X_{j,t})$ if it is categorical, where $I(\cdot)$ is the indicator function. Our approach can incorporate edge covariates as well. The incorporation of covariate information into a network model has attracted increasing attention in network data analysis in recent years. We refer to Hoff et al. (2002) for using Markov chain Monte Carlo procedures for inference within maximum likelihood and Bayesian frameworks, Zhang et al. (2016), Weng & Feng (2016) and Huang & Feng (2018) for conducting community detection in the stochastic block model, Wu et al. (2017) for using the generalized linear model with low-rank effects, Graham (2017) for the β-model that assigns individual merit parameter to each node, Ma & Ma (2017) for using nuclear norm penalization and projected gradient descent to fit a latent space model with covariates, and Yan et al. (2019) for how to conduct statistical inference for the parameters in a directed version of the β-model. Deshpande et al. (2018) provided an information theoretical analysis for inference of latent community structure given a sparse graph along with high dimensional node covariates. These papers typically assume a known link function to associate the probability of the existence of an edge to covariates and possibly other latent variables, sometimes with an additional independence assumption on the edges as random variables. In a different direction, Binkiewicz et al. (2017) proposed a method to uncover latent communities in a graph, using a modification of spectral clustering. Yan & Sarkar (2020) proposed a community detection method for sparse network with node information.

In this paper, we propose a novel approach named network-supervised dimension reduction that seeks to project the covariates onto a low-dimensional space for best explaining the strength of connection in a network in light of the contextual information. This is achieved by formulating a new loss function to estimate a linear projection matrix $B \in \mathbb{R}^{p \times r}$ with $r \leq p$, such that the magnitude of $\|B^T X_{ij}\|$ informs the strength of connection in terms of w_{ij}, where $\|\cdot\|$ is the ℓ_2 norm. Without loss of generality, we assume that a smaller value of $\|B^T X_{ij}\|$ corresponds to a stronger connection, that is, a larger value of w_{ij}. As a concrete example, when nodal information is available and B is an identity matrix, a small value of
colorblack$\|X_{ij}\|$ with $X_{ij} = X_i - X_j$ will correspond to a large value of w_{ij} intuitively. For ease of presentation, we work with $S = (s_{ij}) \in \mathbb{R}^{n \times n}$ where s_{ij} is a monotonic one-to-one decreasing function of w_{ij}. In the simplest case, $s_{ij} = 1 - w_{ij}$. The interpretation of s_{ij} is that a smaller value of s_{ij} implies a stronger relationship between the two corresponding nodes.

Thus, we can state our problem as follows. Given data represented as a collections of tuples $\{s_{ij}, X_{ij}\}$ for $i \neq j$, we seek to find a matrix $B \in \mathbb{R}^{p\times r}$ to project X_{ij} such that the value of $\|B^T X_{ij}\|$ reflects the similarity of the nodes in terms of s_{ij}. More precisely, the projection is such that the smaller $\|B^T X_{ij}\|$ is, the smaller s_{ij} is. Toward this, we propose a novel estimator of B based on a new loss function and study its rate of convergence for approximating the columns of B in terms of ℓ_2 distance. These are achieved without the restrictive independence assumption on w_{ij}'s or the need to assume a link function between $B^T X_{ij}$ and s_{ij}. We show that the convergence rate of the projection depends, among other things, critically on a factor referred to as the network effect of a graph closely related to the graph coloring problem. Proposing such an estimator and characterizing its properties can be seen as the first contribution of this work. Our second contribution is to establish a natural connection between our method and existing methods such as principal component analysis and linear discriminant analysis. The connection to the latter enables us to leverage covariate information for better community detection, which we illustrate via simulations showing that a clustering algorithm based on network-supervised dimension reduction outperforms the competitors.

The following notations are used throughout this paper. For any matrix $A = (a_{ij}) \in \mathbb{R}^{p\times p}$, $\|A\|_{op}$ and $\|A\|_F$ denote its operator norm and Frobenius norm, respectively, and $\|A\|_{\max} = \max_{i,j} |a_{ij}|$. Let $\{a_{ij}\}$ be a set with all items in A. For any symmetric matrix A, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ stand for the maximum and minimum eigenvalues of A, respectively, where $tr(A)$ is its trace. For a vector $v \in \mathbb{R}^p$, $\|v\|$ denotes its ℓ_2 norm. For any variable $Z \in \mathbb{R}$, define $\|Z\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}\{E(|Z|^p)\}^{1/p}$, and for any $\mathrm{Z} \in \mathbb{R}^p$, define $\|\mathrm{Z}\|_{\psi_2} = \sup_{x\in S^{p-1}} \|\langle \mathrm{Z}, x\rangle\|_{\psi_2}$, where S^{p-1} is the unit sphere in $\mathbb{R}^p$. We use I_n to denote the $n \times n$ identity matrix. For any set V, we use $|V|$ to denote its cardinality. For any matrix B, we denote $\mathrm{span}(B)$ as the space spanned by the columns of B, and let P_B be the projection matrix onto the space $\mathrm{span}(B)$. Denote $a \wedge b = \min\{a, b\}$.

2. Network-supervised Dimension Reduction

2.1. *Notation and background*

Recall that our data consists of network-covariate tuples $\{s_{ij}, X_{ij}\}$ $(i \neq j)$. Our goal is to find $B \in \mathbb{R}^{p\times r}$ such that a small value of $\|B^T X_{ij}\|$ corresponds to a smaller value of s_{ij}. In the danger of causing confusion, we refer to B as the projection matrix and its columns as the projection directions. To partially ensure identifiability of B, we constrain $B \in \Theta_{r,A}$, where $\Theta_{r,A} \subset \mathbb{R}^{p\times r}$ satisfies $\Theta_{r,A} = \{B \in \mathbb{R}^{p\times r} : B^T A B = I_r\}$, for a symmetric positive definite matrix $A \in \mathbb{R}^{p\times p}$ with eigenvalues uniformly bounded away from 0 and ∞. An obvious example is $A = I_p$. Since a small value of $\|B^T X_{ij}\|$ corresponds to a small value of s_{ij}, our proposed network-supervised dimension reduction estimates B as $\hat{B}_{r,A} = (\hat{\beta}_{A,1}, \ldots, \hat{\beta}_{A,r}) = \arg\max_{B\in\Theta_{r,A}} H(B)$, where $H(B) = \{n(n-1)\}^{-1} \sum_{i\neq j} s_{ij} \|X_{ij}^T B\|^2 = tr\left(B^T \hat{G} B\right)$, by denoting

$$\hat{G} = \frac{1}{n(n-1)} \sum_{i\neq j} s_{ij} X_{ij} X_{ij}^T = \frac{1}{n(n-1)} \sum_{i\neq j} Z_{ij}, \tag{1}$$

with $Z_{ij} = s_{ij} X_{ij} X_{ij}^T \in \mathbb{R}^{p\times p}$. This optimization problem for estimating the projection directions only requires a standard eigenvalue decomposition as shown in the following proposition.

PROPOSITION 1. *Suppose that all the eigenvalues of $A^{-1/2}\hat{G}A^{1/2}$ are distinct. Let $\hat{\Psi}_r$ be the matrix consisting of the eigenvectors associated with the first r largest eigenvalues of $A^{-1/2}\hat{G}A^{-1/2}$. Then* $\mathrm{span}(\hat{B}_{r,A}) = \mathrm{span}(A^{-1/2}\hat{\Psi}_r)$.

While $\mathrm{span}(\hat{B}_{r,A})$ is unique but $\hat{B}_{r,A}$ is not, Proposition 1 suggests that we can take $\hat{B}_{r,A} = A^{-1/2}\hat{\Psi}_r$. We next provide analogous results at the population level. Let $G_{0n} = E(\hat{G})$ be the expectation of $\hat{G}$ which may depend on the size of the network n, and assume that $G_0 = \lim_n G_{0n}$ for some $G_0 \in \mathbb{R}^{p\times p}$. When Z_{ij}'s have the same distribution but not necessarily independent, it is seen that $G_0 = G_{0n} = E(Z_{ij})$. Denote

$$B_{r,A} = (\beta_{A,1}, \ldots, \beta_{A,r}) = \arg\max_{B\in\Theta_{r,A}} tr(B^T G_0 B), \tag{2}$$

which is the population version of $\hat{B}_{r,A}$. Similar to Proposition 1, supposing that the eigenvalues of $A^{-1/2}G_0A^{-1/2}$ are distinct, if we denote Ψ_r as the matrix consisting of the eigenvectors associated with the first r largest eigenvalues of $A^{-1/2}G_0A^{-1/2}$, then we also have $\mathrm{span}(B_{r,A}) = \mathrm{span}(A^{-1/2}\Psi_r)$. By the argument similar to $\hat{B}_{r,A}$, we simply set $B_{r,A} = A^{-1/2}\Psi_r$.

We now provide sufficient conditions that guarantee the population maximizer of $H(B)$ spans the same column space spanned by the true projection directions. Letting the matrix A in $\Theta_{r,A}$ be $A = E(X_{ij}X_{ij}^T)$, which equals $\mathrm{cov}(X_{ij})$ when $E(X_{ij}) = 0$, we have the following result.

PROPOSITION 2. *Suppose that $\{(s_{ij}, X_{ij}), i \neq j\}$ are identically distributed. Assume that the following conditions hold: (i) s_{ij} satisfies $E(s_{ij} \mid X_{ij}) = h(B_0^T X_{ij})$ where $B_0 = (\beta_1, \ldots, \beta_r) \in \mathbb{R}^{p\times r} \in \Theta_{r,A}$ and h is left unspecified; (ii) the eigenvalues of $A^{-1/2}G_0A^{-1/2}$ are distinct; (iii) $\mathrm{cov}\{s_{ij}, (\beta_m^T X_{ij})^2\} > \mathrm{cov}\{s_{ij}, (v^T X_{ij})^2\}$ $(i \neq j;\ m = 1, \ldots, r)$ for any $v \in \mathbb{R}^p$ satisfying $v^T AB_0 = 0$ and $v^T Av = 1$. Then, it holds that $\mathrm{span}(B_{r,A}) = \mathrm{span}(B_0)$.*

This proposition requires that the conditional mean of s_{ij} depends on X_{ij} only through the linear combination $B_0^T X_{ij}$, with an unknown link function h left unspecified. This is reminiscent of the assumption made in the literature of sufficient dimension reduction (Li, 1991), especially for inferring about the conditional mean of the response given the predictors (Cook & Li, 2002). The key difference is that the responses in our setup are typically correlated due to the existence of the network structure. Condition (i) can be seen as the true model. Particularly, when $s_{ij} \in \{0, 1\}$ (e.g. $s_{ij} = 1 - w_{ij}$), this condition states $\mathrm{pr}(s_{ij} = 1) = h(B_0^T X_{ij})$. The estimation procedure in (1) does not offer an estimator of h. As such, our estimation procedure is model free. To understand Assumption (iii) in this proposition, consider the case when the covariates are defined as $X_{ij} = X_i - X_j$ with $X_i \sim N(\mu, \Sigma)$. Then this assumption becomes $\mathrm{cov}\{s_{ij}, (X_{ij}^T\beta_m)^2\} > 0$ as shown after the proof of this proposition in the Supplementary Materials. This assumption is intuitive since we expect that a smaller s_{ij} corresponds to a small value of $\|B_0^T X_{ij}\|$ and subsequently a small value of $(X_{ij}^T\beta_m)^2$.

2.2. *Connections to other methods*

In the context of the so-called stochastic block model, we establish novel connections between our dimension reduction method and principal component analysis, as well as linear discriminant analysis. The latter two methods are widely used statistical tools for reducing the dimensionality of data, both by finding the best linear combinations of covariates. Principal component analysis is an unsupervised method that projects observations onto the so-called principal component directions such that the variance of the projected data is maximized. Linear discriminant analysis is a supervised learning algorithm that finds the so-called linear discriminant directions for projecting data to maximize the separation between observations belonging to different groups (Johnson & Wichern, 1998).

Recall that for a stochastic block model with k communities, each node belongs to a latent community (Holland et al., 1983). Notationally, denote the latent community label of

the ith node as C_i, where $C_i \in \{1, \ldots, k\}$ for $i = 1, \ldots, n$. The stochastic block model assumes that these community labels are independent and identically distributed random variables such that $\mathrm{pr}(C_i = t) = \pi_t$ $(t = 1, \ldots, k)$, where π_t's are unknown parameters satisfying $\sum_{t=1}^{k} \pi_t = 1$. Given their respective communities, node i and j make a connection with probability $\mathrm{pr}(w_{ij} = 1 \mid C_i, C_j) = \mathrm{pr}_{C_i C_j}$ $(i \neq j)$, independent of all other pairs, where $\mathrm{pr}_{C_i C_j}$ is a parameter depending only on C_i and C_j. We look at a simplified stochastic block model where $\mathrm{pr}_{C_i C_j} = a_t$ for $C_i = C_j = t$ and $\mathrm{pr}_{C_i C_j} = b$ for any $C_i \neq C_j$. That is, all the probabilities of inter-communities connections are the same. For the covariates, we take $X_{ij} = X_i - X_j$, where the covariate vector for the ith node satisfies

$$X_i = \mu_{C_i} + \epsilon_i \quad (i = 1, \ldots, n), \tag{3}$$

for independent and identically distributed random variables ϵ_i with $E(\epsilon_i) = 0$ and $\mathrm{cov}(\epsilon_i) = \Sigma_\epsilon$. Here it is assumed that ϵ_i is independent of C_i, the latent community label of node i in the stochastic block model above. That is, the covariates follow a distribution with a common covariance matrix and a community-specific mean. Under these setups, if s_{ij} is a one-to-one mapping of w_{ij}, it is easily seen that $E(s_{ij} \mid C_i = t, C_j = t')$ is a constant (depending on b) for any $t \neq t'$, which will be denoted as γ_b hereafter. Denote $E(s_{ij} \mid C_i = C_j = t) = \gamma_t$ for $t = 1, \ldots, k$ for ease of notation. We point out that the w_{ij}'s in the above model depend only on the labels C_i's, which is different from the model assumed in (i) of Proposition 2.

If we apply principal component analysis to the nodal feature X_i, at the population level, the principal component directions are the leading eigenvectors of $\mathrm{cov}(X_i)$ corresponding to its largest eigenvalues. If we apply linear discriminant analysis to the labelled data $\{C_i, X_i\}_{i=1}^n$ assuming that the latent community labels are known in model (3), the linear discrimant directions at the population level are the leading $k-1$ eigenvectors of the generalized eigenvalue problem that solves $\Sigma_{\mathrm{bt}} U = \lambda \Sigma_\epsilon U$ for $U \in \mathbb{R}^{p \times (k-1)}$, where
colorblack$\Sigma_{\mathrm{bt}} = k^{-1} \sum_{t=1}^{k} (\mu_t - \bar{\mu})(\mu_t - \bar{\mu})^T$ with $\bar{\mu} = (\sum_{t=1}^{k} \mu_t)/k$, and Σ_ϵ is the covariance matrix of ϵ_i defined above. We have the following proposition that connects our approach with principal component analysis and linear discriminant analysis.

PROPOSITION 3. *Assume that $W = (w_{ij})$ is generated from the simple stochastic block model outlined above and that X_i's are generated from model* (3)*. If all the eigenvalues of $A^{-1/2} G_0 A^{-1/2}$ are distinct for A defined below, the following conclusions hold.*

(1) Specify $A = I_p$ in $\Theta_{A,r}$. Our approach is equivalent to principal component analysis conducted as eigenvalue decomposition of $\mathrm{cov}(X_i)$ at the population level in the sense that $B_{r,A}$ is exactly the eigenvectors associated with the first r largest eigenvalues of $\mathrm{cov}(X_i)$, if and only if $\gamma_b = \sum_{t=1}^{k} \pi_t^2 \gamma_t / \sum_{t=1}^{k} \pi_t^2 > 0$.

(2) If $\sum_{t=1}^{k} \pi_t^2 (\gamma_b - \gamma_t) > 0$ and we choose $A = \mathrm{cov}(X)$ in $\Theta_{A,r}$, then our approcah is equivalent to linear discriminant analysis for the model in (3) *at the population level in the sense that*
colorblack$\beta_{A,m}$ in (2) *is proportional to the ith direction of linear discriminant analysis, colorblack$m = 1, \ldots, r$.*

(3) If $\gamma_b > 0$ and we choose $A = \Sigma_\epsilon$ in $\Theta_{A,r}$, then our approach is equivalent to linear discriminant analysis for the model in (3) *at the population level in the sense that*
colorblack$\beta_{A,m}$ in (2) *is proportional to the ith direction of linear discriminant analysis, colorblack$m = 1, \ldots, r$.*

This proposition shows that network-supervised dimension reduction can be equivalent to unsupervised principal component analysis or supervised linear discriminant analysis, depending on

the data generating process. Note that in Proposition 3, $\mathrm{pr}_{C_iC_j} = b$ for any $C_i \neq C_j$ is assumed. By checking the proof, it is seen that the proposed method may not be equivalent to principle component or linear discriminate analysis in a general case when this assumption does not hold. In this case, the associated objective function for our method can be seen as a generalized version of those in the latter two. For community detection, we explain what we mean by further examining the special case of two communities when $k = 2$ and s_{ij} is a linear decreasing function of w_{ij}. Recall the definition of $B_{r,A}$ in (2).

COROLLARY 1. *Assume that X_{ij} and $W = (w_{ij})$ are generated as in Proposition* 3. *Let $s_{ij} = \alpha_0 - \alpha_1 w_{ij}$ with $\alpha_1 > 0$ and $\alpha_0 \in \mathbb{R}$ be a linear decreasing function of w_{ij}. Suppose that the eigenvalues of $A^{-1/2}G_0A^{-1/2}$ for A involved below are distinct. The following conclusions hold.*

(1) Let $A = I_p$. If $\alpha_0 > 0$ and $b = (\pi_1^2 a_1 + \pi_2^2 a_2)/(\pi_1^2 + \pi_2^2) < \alpha_0/\alpha_1$, then network-supervised dimension reduction is equivalent to principal component analysis conducted as eigenvalue decomposition of $\mathrm{cov}(X_i)$ at the population level.

(2) Let $A = \mathrm{cov}(X)$. If $b < (\pi_1^2 a_1 + \pi_2^2 a_2)/(\pi_1^2 + \pi_2^2)$, then the first direction $\beta_{A,1}$ of network-supervised dimension reduction is equivalent to that of linear discriminant analysis for the model in (3) *at the population level.*

(3) Let $A = \Sigma_\epsilon$. If $b < 1$ and $\alpha_0 \geq \alpha_1 > 0$, the first direction $\beta_{A,1}$ of network-supervised dimension reduction is equivalent to that of linear discriminant analysis for the model in (3) *at the population level.*

To understand this corollary, assume for simplicity that the two communities are equally sized in that $\pi_1 = \pi_2 = 1/2$. In this case, (1) states the equivalence of the proposed approach and principal component analysis if and only if $b = (a_1 + a_2)/2 < \alpha_0/\alpha_1$, which is simplified as $b = (a_1 + a_2)/2$ when $\alpha_0 > \alpha_1$, due to the fact that $b \leq 1$. That is, when $b = (a_1 + a_2)/2$, the network information in terms of the adjacency matrix W do not contribute to the identification of the projections. This is reasonable, since when the probabilities of making connections between different communities is not small. When $\pi_1 = \pi_2 = 1/2$, the condition in (2) becomes $(a_1 + a_2)/2 > b$, and (2) states the equivalence of the proposed approach and linear discriminant analysis when $A = \mathrm{cov}(X)$ and the connection probabilities between different communities are small. The assumption $(a_1 + a_2)/2 > b$ is weaker than strong and weak assortativity (Amini & Levina, 2018) that require $\min\{a_1, a_2\} > b$ in the setting above. In (3) when $A = \Sigma_\epsilon$, our approach is equivalent to linear discriminant analysis for any linear decreasing function when $\alpha_0 \geq \alpha_1 > 0$ and $b < 1$. When our method is applied to community detection in Section 2.3, Proposition S2 in the Supplementary Material shows that the misclassification error depends on the connection probabilities mainly through the projected directions at population level. Notably, a_1, a_2 and b satisfying the constraints in (2) and (3) can be small, implying our approach is applicable to sparse networks. Relevant simulation results are presented in Section 2.3.

The results in Proposition 3 and Corollary 1 suggest that the choice of A matters in order to connect to principal component analysis and linear discriminant analysis. In practice when A is not pre-specified, we suggest taking $A = \mathrm{cov}(X)$. In practice, when A is estimated by $\hat{A}$ from data, there will be an error between $B_{r,\hat{A}}$ and its population version $B_{r,A}$. We give a bound on the error, presented in Proposition S1 of the Supplementary Material.

2.3. *Application in community detection*

Motivated by the covariate model in (3), we may use network-supervised dimension reduction for community detection. To proceed, we first estimate the projection directions denoted as $\hat{B}_{r,A}$ if A is given or $\hat{B}_{r,\hat{A}}$ if A is estimated as $\hat{A}$. Then we can apply a clustering method based on

the projected observations $\hat{B}_{r,A}^T X_i$ or $\hat{B}_{r,\hat{A}}^T X_i$ $(i = 1, \ldots, n)$. For illustration, we apply K-means clustering in the second step. In practice, to check whether our method is applicable, one may examine the scree plot of the cluster algorithm by plotting the ratio of within variance over total variance versus the number of clusters. If there is a clear gap in the plot, one may infer that there are community/cluster structures and our method is applicable.

We now present the result of a small numerical experiment to evaluate the performance of our approach based community detection method. The data is generated such that W follows the simplified stochastic block model for the edges with two communities as in Section 2.2 and X follows the covariate model in (3) with ϵ_i being a multivariate normal random vector. We set $p = 5$, $\mu_1 = (u, 0, \ldots, 0)^T \in \mathbb{R}^p$, $\mu_2 = -\mu_1$ and $\Sigma_\epsilon = (\sigma_{ij})$ with $\sigma_{ij} = 0.7^{|i-j|/3}$ in model (3). It is understood that when u increases, the data in the two communities are better separated by the covariates. In the stochastic block model, we set $\pi_1 = \pi_2 = 1/2$, and $(a_1, a_2, b) = \delta_0(0.5, \tau, 0.1)$, where δ_0 and τ are constants in $[0, 1]$. It is understood that smaller δ_0 gives a sparser network, and smaller value of τ gives weaker community in the second group. When $\tau \leq 0.1$, the signal of the second group is weak and detecting it is difficult.

By varying the magnitude of u, δ_0 and τ, we want to evaluate how the proposed method performs with respect to the informativeness of the covariates, the sparsity of the network, and the strength of the community structure. Specially, we consider the following three cases:

(a) fix $\delta_0 = 0.05$ and $\tau = 0.1$ and vary u in $\{1.0, 1.2, 1.4, 1.6, 1.8, 2.0\}$;
(b) fix $\delta_0 = 0.05$ and $u = 1.6$ and vary τ in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$;
(c) fix $u = 1.6$ and $\tau = 0.1$ and vary δ_0 in $\{0.05, 0.1, 0.3, 0.5, 0.7, 1\}$.

We examine two choices of A for estimating $B_{r,A}$. The first is $A = \mathrm{cov}(X)$ which is estimated by the sample covariance matrix of X. The second is $A = \Sigma_\epsilon$ with the corresponding algorithm and simulation results given in the Supplementary Materials. These two choices of A yield similar results. To apply network-supervised dimension reduction, the response variable s_{ij} is taken as $s_{ij} = 1 - w_{ij}$ and the number of the directions is taken as $r = 1$. Each time, we generate a dataset with $n = 100$ and repeat the process 100 times. The performance of an approach is evaluated by calculating its clustering errors defined as the proportions of the nodes that are misclassified. The performance of K-means clustering after applying our approach is compared to the standard K-means clustering that only uses covariate information, and to several competing methods that use information from both the network and covariates, including those in Binkiewicz et al. (2017), Zhang et al. (2016) and Yan & Sarkar (2020). Moreover, we also consider the cases where the network is dense in that δ_0 is relatively large and the signal of the second community is strong in the Supplementary Materials, where we also report the performance of the method in Huang & Feng (2018) and the spectral clustering method of Rohe et al. (2011). The clustering errors for these methods are presented in Figure 1 and the first figure in the Supplementary Materials.

It is seen that for K-means clustering, the clustering error decreases as u increases. The performance of the methods in Zhang et al. (2016) and Binkiewicz et al. (2017) improve when the network becomes dense, but since detecting the second group is difficulty when $\tau = 0.1$ (i.e. $a_2 = b$), both methods perform worse than the method in Yan & Sarkar (2020) and our method, as shown in plot (c). In addition, we see that the method in Yan & Sarkar (2020) is better than that in Zhang et al. (2016) and Binkiewicz et al. (2017) in most of the cases, but worse than our method. Overall, our approach based clustering performs much better than the other competitors, and it is rather insensitive to the parameters u, δ_0 and τ. This implies that our approach can exploit the information in the covariates as well as the network structure. Especially, when $\tau = 0.1$, it is quite difficult to detect the second group. However, with the help of the information

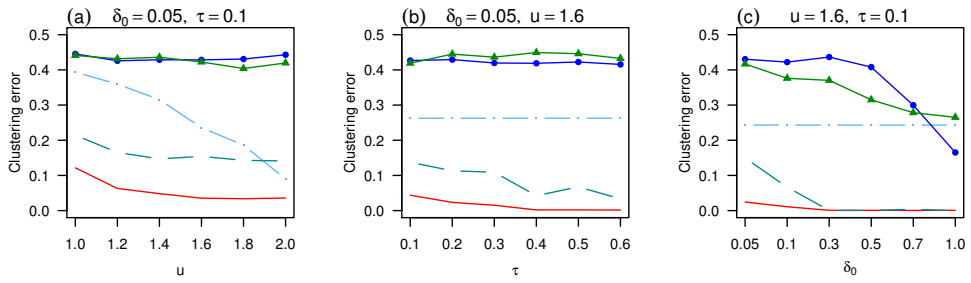

Fig. 1: The average clustering errors. We compare K-means clustering (long dash dot), our proposed method (solid line), and that in Binkiewicz et al. (2017) (solid line with circle), Zhang et al.(2016) (solid line with triangle), Yan & Sarkar (2020) (long dash).

from the covariates, our approach based K-means clustering method still estimates the community structure well. To gain more insight, we consider a simple case when $r = 1$, $C_i \in \{1, 2\}$, and ϵ_i follows a normal distribution, and give an explicit bound on the classification error rate in Proposition S2 of the Supplementary Material. Finally, we also observe that for a network with a community structure with between community probabilities dominating the within ones or admitting a core-periphery one, the performance of our method can deteriorate.

3. ASYMPTOTICS

We study the statistical properties of $\hat{G}$ defined in (1) as an estimator of its population version $G_0 = \lim_n E(\hat{G})$. Due to the network structure, each w_{ij} in the adjacency matrix W may be affected by the other off-diagonal entries of W in complex ways, which raises great challenges for theoretical analysis. we impose assumptions to rule out the cases where all its entries can be strongly dependent, without explicitly modelling the dependence structure of the edges of a network.

We motivate our assumptions by generalizing a notion for inducing edge dependence widely used in the graphon model (Lovász & Szegedy, 2006; Diaconis & Janson, 2008; Bickel & Chen, 2009), for which we will follow the notations in Gao et al. (2015). For an undirected graph, the graphon model assumes the edge random variables $w_{ij} = w_{ji} \sim \text{Bernoulli}(\theta_{ij})$, where $\theta_{ij} = f(\xi_i, \xi_j)$ $(i \neq j)$. The sequence $\{\xi_i\}$ are the independent and identically distributed latent random variables that are from the uniform distribution on $[0, 1]$, and given $\{\xi_i\}$, w_{ij}'s are independent for $i < j$. The function f, a bivariate function symmetric in its arguments, is called graphon. In the graphon model, because the ith latent variable ξ_i is assumed to be associated with the ith node, two edge random variables w_{ij} and w_{kl} are independent as long as they do not share a common node index.

We now introduce what we call the generalized graphon model that is useful for characterizing the dependence structure in our setup. Assume that ξ_i $(i = 1, \ldots, n)$ and ζ_j $(j = 1, \ldots, n)$ are independent and identically distributed latent random variables. Denote $\Xi = (\xi_1, \ldots, \xi_n)^T \in \mathbb{R}^n$. Instead of associating a single element ξ_i of Ξ with node i as in the graphon model, we associate the ith node with a subset of Ξ for introducing dependence, as well as an independent ζ_i for node-specific effect. Denote the subset for node i as $N_i = \{j : \xi_j \text{ is associated with node } i\}$. In the graphon model, $i \in N_i$. We then assume the edge random variable $w_{ij} \sim \text{Bernoulli}(\theta_{ij})$,

where

$$\theta_{ij} = f_{ij}(\Xi_{N_i}, \zeta_i, \Xi_{N_j}, \zeta_j, X_{ij}) \quad (i \neq j). \tag{4}$$

Here Ξ_{N_i} is the sub-vector of Ξ with indices in N_i. In our construction, we have purposely left unspecified the exact distributions of the random variables $\{\xi_i\}$ and $\{\zeta_j\}$, as well as the functions $\{f_{ij}\}$, as we only need this general construction for relating the edge random variables. In the special case of the graphon model, $N_i = \{i\}$ and $f_{ij}(\Xi_{N_i}, \zeta_i, \Xi_{N_j}, \zeta_j, X_{ij}) = f(\xi_i, \xi_j)$.

Denote $N_{ij} = N_i \cup N_j$ and let

$$\mathbb{V} = \Big\{\{(i,j),(k,t)\} : N_{ij} \cap N_{kt} = \emptyset \quad (i \neq j \neq k \neq t)\Big\}$$

be the set in which any two pairs of nodes do not share common latent random variables. It is clear by construction that for any $\{(i,j),(k,t)\} \in \mathbb{V}$, w_{ij} is independent of w_{kt} given X_{ij} and X_{kt}. The cardinality of $\mathbb{V}$ provides a rough characterization of the dependence structure of a network intuitively and is seen to be bounded as $|\mathbb{V}| \leq \binom{n}{4}$. The graphon model achieves the upper bound.

We now present another example where $N_i = \{i, i+1\}$ for $i < n$ and $N_n = \{n, 1\}$. That is, we associate each node with two latent random variables in Ξ. If we represent this example via a graph in which nodes are $\{1, \ldots, n\}$ and an edge exists between the ith and jth nodes if $N_i \cap N_j \neq \emptyset$, then it forms a cycle graph. For this example, it is not difficult to see that $|\mathbb{V}| = n(n-5)(n^2 - 9n + 22) = O(n^4)$ which is of the same order as the maximum possible cardinality of $|\mathbb{V}|$.

Next we study $\|\hat{G} - G_0\|_{op}$. Establishing the rate of convergence of $\hat{G}$ in the operator norm is challenging, due to the dependence among the nodes. In the generalized graphon model above for example, node i is correlated with node j for any $j \in N_i$ which will complicate theoretical analysis. We overcome the dependency challenge by splitting all the node pairs into groups such that any two node pairs in the same group are conditionally independent given covariates.

Let $(\sigma(1), \ldots, \sigma(n))$ can be any permutation of $\{1, \ldots, n\}$ and $\xi_{\alpha,ij} = s_{ij}(\alpha^T X_{ij})^2 - E\{s_{ij}(\alpha^T X_{ij})^2\}$ for any given $\alpha \in \mathbb{R}^p$ satisfying $\|\alpha\| = 1$. Suppose that we split the index pairs $\{\tilde{\sigma}(i) = (\sigma(2i-1), \sigma(2i)) \ (i = 1, \ldots, n/2)\}$ into m groups $G_1, \ldots, G_m$ such that any two pairs $\tilde{\sigma}(i)$ and $\tilde{\sigma}(j)$ within the same groups satisfy $\{\tilde{\sigma}(i), \tilde{\sigma}(j)\} \in \mathbb{V}$. That is, given $\{X_{ij}\}$, $\xi_{\alpha,ij}$'s with (i,j)'s in the same group are independent, which will be referred to as the *conditional independence property* hereafter. It is shown in the Supplementary Material that a smaller m is desired as it leads to a tighter upper bound. Finding the smallest m associated with permutation $\{\sigma(1), \ldots, \sigma(n)\}$ is very challenging and can be viewed as a graph coloring problem where the interest is often to find the chromatic number of a graph, defined as the minimum number of colours required for a vertex colouring scheme with any two adjacent vertices coloured differently (see Supplementary Materials for further discussion). Denote this number as m_σ and define $m_{\text{net}} = \max_{(\sigma(1),\ldots,\sigma(n))} m_\sigma$, which can be loosely seen as the network effect. The asymptotic property of $\|\hat{G} - G_0\|_{op}$ is presented in Theorem 1 below.

Moreover, we study the asymptotic properties of the eigenvalues and eigenvectors of $\hat{G}$. Towards this, denote the eigenvalue decompositions of G_0 and $\hat{G}$, respectively, as $G_0 = \sum_{i=1}^p \lambda_i v_i v_i^T$ and $\hat{G} = \sum_{i=1}^p \hat{\lambda}_i \hat{v}_i \hat{v}_i^T$, where $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ are the eigenvalues, and v_i's and $\hat{v}_i$'s are the associated eigenvectors. The eigenvalues and eigenvectors depend on p but we omit p hereafter for simplicity. Similarly, denote

$$G_{0A} = A^{-1/2} G_0 A^{-1/2} = \sum_{i=1}^p \phi_i^A \varphi_i^A (\varphi_i^A)^T, \quad \hat{G}_A = A^{-1/2} \hat{G} A^{-1/2} = \sum_{i=1}^p \hat{\phi}_i^A \hat{\varphi}_i^A (\hat{\varphi}_i^A)^T,$$

where $\phi_1^A \geq \cdots \geq \phi_p^A$ and $\hat{\phi}_1^A \geq \cdots \geq \hat{\phi}_p^A$ are the eigenvalues, and φ_i^A's and $\hat{\varphi}_i^A$'s are the associated eigenvectors. Recall the definitions of $B_{r,A}$ and $\hat{B}_{r,A}$ in Section 2. By Proposition 1, we see that

$$\hat{B}_{r,A} = (\hat{\beta}_{A,1}, \ldots, \hat{\beta}_{A,r}) = A^{-1/2}(\hat{\varphi}_1^A, \ldots, \hat{\varphi}_r^A), \tag{5}$$

and that $B_{r,A} = (\beta_{A,1}, \ldots, \beta_{A,r}) = A^{-1/2}(\varphi_1^A, \ldots, \varphi_r^A)$. When A is unknown and estimated as $\hat{A}$, we can define $\hat{G}_{\hat{A}}$ and $\hat{\varphi}_i^{\hat{A}}$ analogously, and estimate $B_{r,A}$ by $\hat{B}_{r,\hat{A}} = (\hat{\beta}_{\hat{A},1}, \ldots, \hat{\beta}_{\hat{A},r}) = \hat{A}^{-1/2}(\hat{\varphi}_1^{\hat{A}}, \ldots, \hat{\varphi}_r^{\hat{A}})$.

To study the properties of $\hat{B}_{r,A}$ and $\hat{B}_{r,\hat{A}}$, we make the following assumptions.

(A1) (i) For any integer $l > 0$ and any subset $I = \{(i_t, j_t) \ (t = 1, \ldots, l)\}$ satisfying $\{(i_t, j_t), (i_{t'}, j_{t'})\} \in \mathbb{V}$ for any $t \neq t'$, $\{X_{i_t j_t} \ (t = 1, \ldots, l)\}$ are independent variables, following sub-Gaussian distributions with $\max_{i \neq j} \|X_{ij}\|_{\psi_2} < K_0 < \infty$ for some constant $K_0 > 0$. (ii) The conditional distribution of $s_{ij} \mid \{X_{ij}\}$ is the same as that of $s_{ij} \mid X_{ij}$.

(A2) Assume that $\delta = \inf_{i=1,\ldots,p-1}(\lambda_i - \lambda_{i+1}) > 0$ and $\delta_A = \inf_{i=1,\ldots,p-1}(\phi_i^A - \phi_{i+1}^A) > 0$ uniformly over p.

When $X_{ij} = X_i - X_j$ where X_i's are independent and identically distributed random variables following a sub-Gaussian distribution, (i) of (A1) holds. (A2) assumes that all the eigenvalues of G_0 and G_{0A} are distinct with positive gaps. We have the following convergence results.

THEOREM 1. *Assume that* $\max_{i \neq j} |s_{ij}| < c_0$ *almost surely and that (A1) and (A2) hold.*

(1) It holds that $\|\hat{G} - G_0\|_{op} = O_p\left[\delta_n^{op} + \left(pm_{\text{net}}^2/n\right)^{1/2}\right]$, *where* $\delta_n^{op} = \|G_{0n} - G_0\|_{op}$.

(2) Assume further $\|G_0\|_{op} < C_0$ *for some constant* C_0 *independent of* p. *Then for* $i = 1, \ldots, p$, *it holds that*

$$|\hat{\lambda}_i - \lambda_i| = O_p\left[\delta_n^{op} + \left(pm_{\text{net}}^2/n\right)^{1/2}\right], \ \ \|\hat{v}_i - cv_i\| = O_p\left[\delta_n^{op} + \left(pm_{\text{net}}^2/n\right)^{1/2}\right],$$

where $c \in \{-1, 1\}$ *is a sign scalar to ensure* $c\hat{v}_i^T v_i > 0$.

Next, we provide an approximation to m_{net} when an additional assumption on the largest degree as in Assumption (A3) below is imposed. Specifically, we only require the conditional independence property to hold for all but one groups. For $\tilde{m}_{\text{net}}$ defined in Theorem 2 below, we show in the Supplementary Materials that for any permutation $\{\sigma(1), \ldots, \sigma(n)\}$, one can always split the index pairs into $\tilde{m}_{\text{net}}$ groups such that the conditional independence property holds for the first $\tilde{m}_{\text{net}} - 1$ groups. In other words, for any $\tilde{\sigma}(i)$ and $\tilde{\sigma}(j)$ in G_s with $s = 1, \ldots, \tilde{m}_{\text{net}} - 1$, we have $\{\tilde{\sigma}(i), \tilde{\sigma}(j)\} \in \mathbb{V}$. Combining the conditional independence property for the first $\tilde{m}_{\text{net}} - 1$ groups with Assumption (A3) below, we will show the conclusions of Theorem 1 still hold but with m_{net} replaced by $\tilde{m}_{\text{net}}$.

(A3) We assume that $d_{\max} < \sqrt{n}$, where $d_{\max} = \max_{i=1,\ldots,n} |\{j : N_j \cap N_i \neq \emptyset\}|$.

It is easy to see that $d_{\max} \leq \max_i |N_i| \ \max_i |\{j : \xi_i \text{ is assoiciated with node } j\}|$ for the generalized graphon model. Condition (A3) enables us to control the $\tilde{m}_{\text{net}}$th group, where the conditional independence property may fail to hold and an upper bound on the number of correlated nodes is then necessary.

THEOREM 2. *Assume additionally that (A3) holds in Theorem* 1. *Let* $\tilde{m}_{\text{net}} = \log(n/4)/\log\{4d_{\max}/(4d_{\max} - 1)\} + 1$. *Then all the conclusions of Theorem* 1 *hold if* m_{net} *is replaced by* $\tilde{m}_{\text{net}}$.

Theorems 1 and 2 present the asymptotic properties of $\hat{G}$. The term δ_n^{op} in these theorems can be seen as the approximation error, and is zero when Z_{ij}'s have the same distribution. The term $\left(pm_{\text{net}}^2/n\right)^{1/2}$ (or $\left(p\tilde{m}_{\text{net}}^2/n\right)^{1/2}$) can be seen as the estimation error in which m_{net} (or $\tilde{m}_{\text{net}}$) can be loosely understood as the effect of a network. If $d_{\max}$ is bounded by a constant, then $\tilde{m}_{\text{net}} = O(\log n)$ and the convergence rate of $\hat{G}$ is $O_p[(p/n)^{1/2}\log n]$. If $d_{\max} = O(\log n)$, then by noting that $1/\log\{4d_{\max}/(4d_{\max}-1)\} = 1/\log\{1+(1/(4d_{\max}-1))\} \approx 4d_{\max}-1 = O(\log n)$, we have $\tilde{m}_{\text{net}} = O(\log^2 n)$ and the convergence rate of $\hat{G}$ becomes $O_p[(p/n)^{1/2}(\log n)^2]$. Following the proof of this theorem, it can be seen that if s_{ij}'s are independent, then $\|\hat{G} - G_0\|_{op} = O_p\left[\delta_n^{op} + (p/n)^{1/2}\right]$. Theorem 1 indicates that, if $d_{\max}$ is small (e.g. $d_{\max} = O(\log n)$), the convergence rate of $\hat{G}$ is similar to the independent case up to a factor of a power function of $\log n$.

In Theorems 1 and 2, the convergence rate of the estimator is established under the generalized graphon model by exploiting its latent variable representation. In fact, as shown in the proof of Theorem 1, the conclusions of Theorem 1 still hold without the generalized graphon model assumption, as long as the following conditional independence property holds. Specifically, a sufficient condition for these theorems to hold is that the node pairs $\{(\sigma(2i-1), \sigma(2i))\ (i = 1, \ldots, n/2)\}$ can be split to groups such that s_{ij}'s with (i,j)'s in the same group are conditionally independent given $\{X_{ij}\}$. Here the s_{ij}'s with (i,j)'s in different groups can still be correlated.

By the relationship between $\hat{G}_A$ and $\hat{G}$, one can establish the asymptotic properties of $\hat{G}_A$ and its eigenvectors. Consequently, the convergence of $\hat{B}_{r,A}$ can be established, by noting that $\hat{B}_{r,A}$ is a function of A and the eigenvectors of $\hat{G}_A$. The same argument is applicable for $\hat{B}_{r,\hat{A}}$, when A is unknown and estimated as $\hat{A}$. We make the following assumptions on the estimator of A.

(A4) Assume that $0 < C^{-1} < \lambda_{\min}(A) \le \lambda_{\max}(A) < C < \infty$ uniformly over p.
(A5) Assume that the estimator $\hat{A}$ of A satisfies $\|\hat{A}^{-1/2} - A^{-1/2}\|_{op} = O_p(\tau_n)$.

Assumption (A4) is standard and in (A5), τ_n is a function of n and p, with p omitted for simplicity. The following theorem shows the convergence rate of the estimator when A is known or estimated as $\hat{A}$. For simplicity, we assume that r is known.

THEOREM 3. *Assume that* $\max_{i\neq j}|s_{ij}| < c_0$ *almost surely and that (A1),(A2) and (A4) hold. The following conclusions hold.*

(1) Assume that A is known. Then $\max_{i=1,\ldots,r}\|\hat{\beta}_{A,i} - c\beta_{A,i}\| = O_p\left[\delta_n^{op} + \left(pm_{\text{net}}^2/n\right)^{1/2}\right]$, *for any given $r = 1, \ldots, p$, where $c \in \{-1, 1\}$ such that $c\hat{\beta}_{A,i}^T\beta_{A,i} > 0$.*

(2) When A is unknown, assume further that (A5) holds. Then for any given $r = 1, \ldots, p$,

$$\max_{i=1,\ldots,r}\|\hat{\beta}_{\hat{A},i} - c\beta_{A,i}\| = O_p\left[\tau_n + \delta_n^{op} + \left(pm_{\text{net}}^2/n\right)^{1/2}\right],$$

where $c \in \{-1, 1\}$ such that $c\hat{\beta}_{\hat{A},i}^T\beta_{A,i} > 0$.

We give a concrete example to show the values of δ_n^{op} and τ_n.

COROLLARY 2. *Suppose that s_{ij}'s are identically distributed (but are dependent), and that $X_{ij} = X_i - X_j$ with X_i's i.i.d. from $N(\mu, \Sigma)$, where the eigenvalues of Σ are bounded away from 0 and ∞ uniformly over p. Assume that $\hat{A}$ is taken as the sample covariance matrix. Then it holds that $\tau_n = (p/n)^{1/2}$ and $\delta_n^{op} = 0$.*

Theorem 3 shows that the convergence rate is determined by the approximation error δ_n^{op}, the dimension of the covariates p, the network effect m_{net}, and the convergence rate τ_n of $\hat{A}$, if A is unknown. Similar to Theorem 2, we can replace the unknown m_{net} with $\tilde{m}_{net}$ as shown in the following Theorem 4, of which the proof is the same as that of Theorem 3 and is omitted.

THEOREM 4. *Suppose additionally that (A3) holds in Theorem* 3. *The conclusions of Theorem* 3 *hold if* m_{net} *is replaced by* $\tilde{m}_{\text{net}}$.

When w_{ij} and consequently s_{ij} depend on n, by replacing the condition $\max_{i\neq j} s_{ij} < c_0$ above by $\max_n \max_{i\neq j} s_{ij,n} < c_0$, we can see that the above conclusions still hold. Thus, Theorems 1-4 continue to hold for sparse networks. Finally, we briefly discuss the selection of r motivated by a similar procedure in Lam & Yao (2012), among others. Recall that $\hat{\phi}_1^A \geq \cdots \geq \hat{\phi}_p^A$ are eigenvalues of $\hat{G}_A$. We select r as

$$\hat{r} = \arg \max_{i=1,\ldots,M} (\hat{\phi}_i^A - \hat{\phi}_{i+1}^A)/(\hat{\phi}_i^A + \hat{\phi}_{i+1}^A),$$

where M is a fixed number. When A is unknown and estimated as $\hat{A}$, we use the eigenvalue of $\hat{G}_{\hat{A}}$ instead.

4. SIMULATION

For the model in Proposition 2, it is assumed that $E(s_{ij}|X_{ij}) = h(B_0^T X_{ij})$. Proposition 2 shows that our method can be used to recover span(B_0) in this setting. To verify the effectiveness of the proposed network-supervised dimension reduction method in recovering span(B_0), we conduct extensive simulation. The performance of our method is examined by computing the error measure defined as $\|P_{B_0} - P_{\hat{B}_{r,\hat{A}}}\|_F$, where B_0 is the true parameter to be estimated, $\hat{B}_{r,\hat{A}}$ is the estimator of B_0 using the method developed in this paper, and P_B for any matrix B is the projection matrix onto the space spanned by the columns of B. Here we take colorblack$A = \text{cov}(X_1)$ and the sample covariance matrix as its estimator. In all simulations, we take $s_{ij} = 1 - w_{ij}$ and assume that r is known. we set $n = 100$ or 500 and dimension $p = 10$ or 50. For each example, 100 datasets are generated. Additional simulation for selection of r using the method in Section 3 is presented in the Supplementary Materials.

Example 1. We generate data according to the following procedure inspired by a similar setup in Weng & Feng (2016).

(*i*) Let $C_i \in \{1, 2\}$ be the latent community label. Generate C_i from a Bernoulli distribution such that $\text{pr}(C_i = 1) = \text{pr}(C_i = 2) = 0.5$.

(*ii*) Generate covariates $X_i \sim N(0, \Sigma)$ where colorblack$\Sigma = (\sigma_{t_1 t_2})$ with $\sigma_{t_1 t_2} = 0.4^{|t_1 - t_2|} I(\{|t_1 - t_2| < 5\})$.

(*iii*) Given (C_i, C_j, X_{ij}) with $X_{ij} = X_i - X_j$, generate $w_{ij} \in \{0, 1\}$ according to the following model

$$\text{pr}(w_{ij} \mid C_i, C_j, X_{ij}) = \text{pr}(w_{ij} \mid C_i, C_j) \frac{\exp(1 - c_{\text{coef}}|B_0^T X_{ij}|)}{1 + \exp(1 - c_{\text{coef}}|B_0^T X_{ij}|)}, \tag{6}$$

where $\text{pr}(w_{ij} \mid C_i, C_j)$ is set as $\text{pr}(w_{ij} = 1 \mid C_i = C_j) = a$ and $\text{pr}(w_{ij} = 1 \mid C_i \neq C_j) = b$.

In model (6), the first part can be seen as the community effect and the second part is a logistic model representing the nodal effect. In the simulation, we set $r = 1$ and $B_0 = (1, 1, 0, \ldots, 0)^T \in$

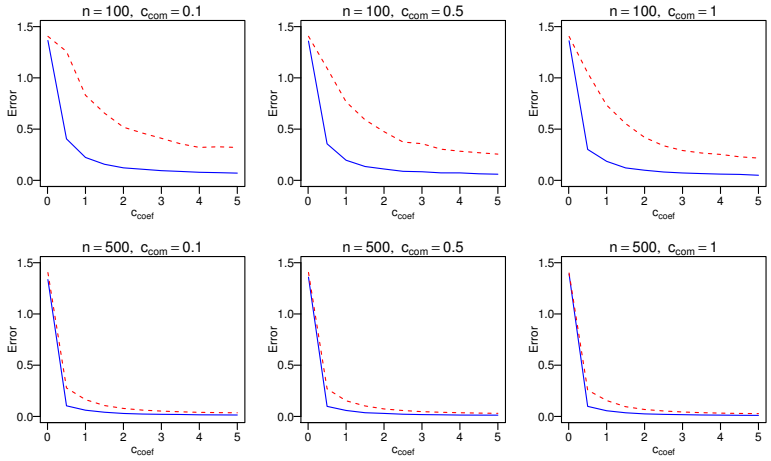

Fig. 2: Average errors for Example 1 for $p = 10$ (solid line) and 50 (dotted line), respectively, where $r = 1$.

$\mathbb{R}^p$, $a = 0.8$, $b = c_{\text{com}}a$ with $c_{\text{com}} = 1, 0.5$, or 0.1, and $c_{\text{coef}} = [0.5 : 0.5 : 5]$, the grid points in the interval $[0.5, 5]$ with step length 0.5. Obviously, $c_{\text{com}} = 1$ corresponds to no community effect, while $c_{\text{com}} = 0.1$ corresponds to strong community effect. A larger c_{coef} implies a larger nodal effect, and when $c_{\text{coef}} = 0$ there is no nodal effect. The generated networks have a wide range of densities, ranging from 3.8% when $c_{\text{com}} = 0.1$ and $c_{\text{coef}} = 0.1$, and 41.5% when $c_{\text{com}} = 1.0$ and $c_{\text{coef}} = 0.5$. The simulation results are found in Figure 2.

Example 2. Consider an example where each node i is affected by its K neighbors and denote the set of their indices as $\bar{N}_i$. The data is generated as follows.

(i) First, generate $\bar{N}_i$ for node i. Let $\mu_1, \ldots, \mu_n$ be independent and identically distributed random variables from $U(0, 1)$, and let $d_{ij} = |\mu_i - \mu_j|$ $(i, j = 1, \ldots, n)$. For each node i, compute its K-nearest neighbors, according to the distance d_{ij}. Define $\bar{N}_i$ as the set that contains those indices j $(j \neq i)$ such that node j is one of node i's K-nearest neighbors. By construction, $i \notin \bar{N}_i$.

(ii) Let $Y_1, \ldots, Y_n$ be independent random variables generated as $Y_i \sim N(\mu_i, 0.1)$ and $Y_{ij} = Y_i - Y_j$. Generate X_i as in Example 1 and define $X_{ij} = X_i - X_j$.

(iii) Generate $w_{ij} \in \{0, 1\}$ according to the following model

$$\text{pr}(w_{ij} = 1 \mid \{Y_{ij}\}, X_{ij}) = \exp\{-10g(\{Y_{ij}\})\} \frac{\exp(1 - c_{\text{coef}}\|B_0^T X_{ij}\|_2)}{1 + \exp(1 - c_{\text{coef}}\|B_0^T X_{ij}\|_2)},$$

where c_{coef} is specified as in Example 1, $g(\{Y_{ij}\}) = |Y_{ij}| \wedge \sum_{k \in \bar{N}_i, k' \in \bar{N}_j} |Y_{kk'}|/K^2$, and $a \wedge b = \min\{a, b\}$.

In this model, we have $N_i = \{i\} \cup \bar{N}_i$ with $i \notin \bar{N}_i$, where $\bar{N}_i$ may be seen as the latent neighbor of node i. When $\bar{N}_i = \emptyset$, we see that the node i is only affected by its latent variable Y_i, where Y_i's are independent of each other. When $\bar{N}_i \neq \emptyset$, we actually have an underlying network introduced by the latent neighbor sets $\bar{N}_i$'s. This network can be seen as underlying truth whereas the network generated by w_{ij} is an observed one. The added dependence can lead to a flexible model with better interpretation. For example, in the study of genetic data, one might view the

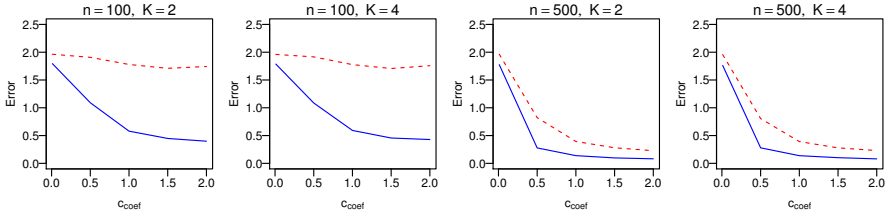

Fig. 3: Average errors for Example 2 for $p = 10$ (solid line) and 50 (dotted line), respectively, where $r = 2$.

underlying network as the true network among genes, and the observed network represented by w_{ij} as a noisy one, contaminated by measurement errors and affected by environment factors.

We set $B_0 = (\beta_1, \beta_2) \in \mathbb{R}^{p\times 2}$ where $\beta_1 = (1, 1, 0, \ldots, 0)^T \in \mathbb{R}^p$ and $\beta_2 = (1, -1, 0, \ldots, 0)^T \in \mathbb{R}^p$, and set $K = 2$ or $K = 4$. For this model, the probability of $w_{ij} = 1$ depends on latent variables in $\{Y_k : Y_k \in N_i \cup N_j\}$ and the covariates X_i and X_j. Clearly, this model is a generalized graphon model defined in Section 3. The results of this simulation can be found in Figure 3. We briefly discuss these simulation results. It is easy to see that the influence of the covariate X_{ij} decreases in both examples when c_{coef} decreases. Particularly, X_{ij} has no effect when $c_{\text{coef}} = 0$. We can see from Figure 2 and 3 that the average errors decreases when c_{coef} increases. This is reasonable because the covariates contribute more and more information with an increasing c_{coef}. Overall, it is seen that the errors decrease as n increases in both examples, which is expected from the theoretical results on the convergence rate. Interestingly, it is seen from Figure 2 that the errors are similar for different c_{com} in Example 1. This is due to the fact that the community label C_i is independent of the covariate X_{ij} in the data generating process.

Finally, to illustrate another application of the proposed dimension reduction method, we briefly outline how to select important covariates. Since our method is similar to the principal component analysis that aims to find the eigenvectors of a matrix, we have developed a procedure similar to sparse principle components analysis for obtaining a sparse estimator of the projections as in Zou et al. (2006). More specifically, as in (5), our estimator is $\hat{B}_{r,A} = (\hat{\beta}_{A,1}, \ldots, \hat{\beta}_{A,r})$ with $\hat{\beta}_{A,j} = A^{-1/2}\hat{\varphi}_j^A$, where $\hat{\varphi}_j^A$ is an eigenvector of $\hat{G}_A = A^{-1/2}\hat{G}A^{-1/2}$. To implement our procedure, we can use the truncated power algorithm in Yuan & Zhang (2013) and we illustrate this algorithm as follows when we want to find a sparse estimate of $\beta_{A,1}$. Given an initial value $v_0 \in \mathbb{R}^p$, for $t = 1, 2, \ldots$, let $v_t' = \hat{G}_A v_{t-1}/\|\hat{G}_A v_{t-1}\|$, truncate $A^{-1/2}v_t'$ by keeping only the largest m_0 entries in absolute values, denote the resulting vector as ϑ_t, and set $v_t = A^{1/2}\vartheta_t$. Repeat the procedure until ϑ_t converges. The final ϑ_t obtained is the sparse estimate of $\beta_{A,1}$. In Table 1, we report some preliminary results on variable selection using Example 1 with $p = 10$ and $n = 100$ for illustration, where the experiments are run 100 times under each setting. Since the first two variables are significant in this example, we set $m_0 = 2$ in the algorithm. We can see that the results are all satisfactory especially when $c_{\text{coef}} > 0.5$.

5. REAL DATA ANALYSIS

We apply the method in this paper to a pulsar candidates data collected by the High Time Resolution Universe (HTRU) survey (Keith et al., 2010), which is available on `http://archive.ics.uci.edu/ml/datasets/HTRU2`. Pulsars are a rare type of Neutron star that produces

Table 1: True positive rate and false positive rate for variable selection when $(p, n) = (10, 100)$ for Example 1

	$c_{\text{com}} = 0.1$		$c_{\text{com}} = 0.5$		$c_{\text{com}} = 1.0$	
c_{coef}	TP	FP	TP	FP	TP	FP
0.5	0.49	0.13	0.50	0.13	0.51	0.12
1.0	1.00	0.00	0.99	0.00	1.00	0.00
1.5	1.00	0.00	1.00	0.00	1.00	0.00
2.0	1.00	0.00	1.00	0.00	1.00	0.00

TP, true positive rate; FP, false positive rate.

radio emission detectable here on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter. Their study yields a better understanding of many physics problems, ranging from acceleration of particles in the ultra-strong magnetic field, to tests of gravity in the strong field regime. Since some pulsars are binary system (Lyne & Smith, 2012), the signals detected for a star are mixed ones from this star and its neighbors, and some noise, implying that our generalized graphon model is applicable for this data. In this dataset, each pulsar is described by eight continuous variables, and a single class variable including 16259 spurious examples caused by radio frequency interference or noise and 1639 real pulsar examples which have been checked by human annotators. The continuous variables are the mean of the integrated profile, the standard deviation of the integrated profile, the excess kurtosis of the integrated profile, the skewness of the integrated profile, the mean of the dispersion measurement–signal-to-noise Ratio curve, the standard deviation of the ratio curve, the excess kurtosis of the ratio curve, and the skewness of the ratio curve. That is, the first four variables are simple statistics obtained from the integrated pulse profile, while the remaining four variables are similarly obtained from the ratio curve. In addition, we observe that the sample covariance matrices of these two groups are different.

We randomly select 200 observations from 16259 spurious examples and 100 observations from 1639 real pulsar examples to construct a graph. For these 300 nodes, we say that two nodes are connected if their difference in the first variable (the mean of the integrated profile) is small. We choose a threshold such that the network density, defined as the ratio of edges over the maximum possible number of edges is $50\%, 30\%, 10\%, 5\%, 3\%$, or 1%. The rest of the eight variables are used as nodal covariates. In defining the graph, we do not use the information on the labels of these observations. This data generating process is repeated 100 times.

For our approach, we take $A = \text{cov}(X)$ and estimate it by the sample covariance matrix. The rank r is chosen by the method outlined at the end of Section 3 by setting $M = 4$. Our approach is compared to the methods in Binkiewicz et al. (2017), Zhang et al. (2016), Yan & Sarkar (2020), Huang & Feng (2018) and the spectral clustering method of Rohe et al. (2011). In addition, we include our approach by estimating $A = \Sigma_\epsilon$ via the algorithm in the Supplementary Materials with $r = 1$. Since the true community membership of each node is known, we report the average of the proportions of the nodes that are misclassified. It is obvious that the smaller this quantity is, the better an approach is. The results averaged over 100 random datasets are found in Figure 4. It is observed that our network-supervised dimension reduction approach based on the two choices of A, are better than the other methods in most of the cases, especially when the network is sparse. In addition, our approach is insensitive to the sparsity of the network, while the methods in Binkiewicz et al. (2017), Rohe et al. (2011), and Zhang et al. (2016) work only when the network is dense.

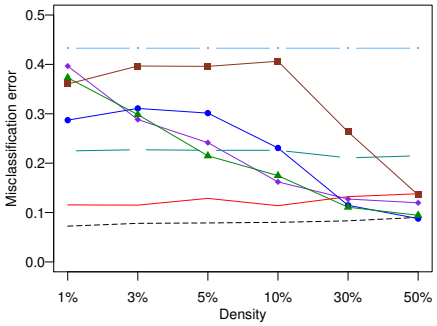

Fig. 4: Comparison of different algorithms for the real data. Our method with $A = \mathrm{cov}(X)$ (solid line) and with $A = \Sigma_\epsilon$ (dash line), K-means (long dash dot), and the method in Binkiewicz et al. (2017) (solid line with circle), Zhang et al.(2016) (solid line with triangle), Yan & Sarkar (2020) (long dash), Rohe et al. (2011) (solid line with diamond) and Huang & Feng (2018) (solid line with square).

ACKNOWLEDGEMENTS

We thank three anonymous reviewers, the associate editor, and the editor for their very constructive comments that have led to a much improved paper. The research of Zhao is supported by National Science Foundation of China. Wang's research is partially supported by National Natural Science Foundation of China and China's National Key Research Special Program. The research of Leng is supported by a Turing Fellowship.

SUPPLEMENTARY MATERIAL

Supplementary material available at Biometrika online includes proofs of the theoretical properties and additional theoretical and simulation results.

colorblackThe code for community detection is available on https://github.com/DR-Network/community-detection.

REFERENCES

AMINI, A. A. & Levina, E. (2018). On semidefinite relaxations for the block model. *Ann. Statist.* **46**, 149–179.

BICKEL, P. J. & CHEN, A. (2009). A nonparametric view of network models and newman-girvan and other modularities. *P. Natl. Acad. Sci. U.S.A.* **106**, 21068–21073.

BIEN, J. & TIBSHIRANI, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98**, 807–820.

BINKIEWICZ, N., VOGELSTEIN, J. T. & ROHE, K. (2017). Covariate-assisted spectral clustering. *Biometrika* **104**, 361–377.

CLAUSET, A., NEWMAN, M. E. & MOORE, C. (2004). Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111.

COOK, R.D. & LI, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455–474.

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

DESHPANDE, Y., SEN, S., MONTANARI, A., & MOSSEL, E. (2018). Contextual Stochastic Block Models. neural information processing systems.

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

GAO, C., LU, Y. & ZHOU, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43**, 2624–2652.

GOLDENBERG, A., ZHENG, A.X., FIENBERG, S.E. & AIROLDI, E.M. (2010). A survey of statistical network models. *Found. Trends. Mach. Learn.* **2**, 129–233.

GRAHAM, B. S. (2017), An econometric model of network formation with degree heterogeneity. *Econometrica* **85**, 1033–1063.
HOFF, P.D., RAFTER, A.E. & HANDCOCK, M.S. (2002). Latent space approaches to social network analysis. *J. Am. Statist. Assoc.* **97**, 1090–1098.
HOLLAND, P. W., LASKEY, K. B. & LEINHARDT, S. (1983). Stochastic block models: First steps. *Social Networks* **5**, 109–137.
HOLLAND, P. W. & LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Am. Statist. Assoc.* **76**, 33–50.
HUANG, S. & FENG, Y. (2018). Pairwise covariates-adjusted block model for community detection. *arXiv*: 1807.03469.
HUNTER, D. R., GOODREAU, S. M. & HANDCOCK, M. S. (2008). Goodness of fit of social network models. *J. Am. Statist. Assoc.* **103**, 248–258.
JIN, E. M., GIRVAN, M. & NEWMAN, M. E. (2001). Structure of growing social networks. *Phys. Rev. E* **64**, 046132.
JOHNSON, R. A. & WICHERN, D. W. (1988). *Applied Multivariate Statistical Analysis*. Upper Saddle River: Prentice Hall.
KEITH, M. J., JAMESON, A., VAN STRATEN, W., BAILES, M., JOHNSTON, S., KRAMER, M., POSSENTI, A., BATES, S. D., BHAT, N. D. R., BURGAY, M., BURKE-SPOLAOR, D'AMICO, N., LEVIN, L., MCMAHON, P.L., MILIA, S. & STAPPERS, B. W. (2010). The high time resolution universe pulsar survey-i system configuration and initial discoveries. *Mon. Not. R. Astron. Soc.* **409**, 619–627.
KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. New York: Springer.
LAM, C. & YAO, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *Ann. Statist.***40**, 694–726.
LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Am. Statist. Assoc.* **86**, 316–327.
LOVÁSZ, L. & SZEGEDY, B. (2006). Limits of dense graph sequences. *J. Comb. Theory B* **96**, 933–957.
LYNE, A. G. & SMITH, F. G. Pulsar Astronomy. Cambridge: Cambridge University Press, 2012: 64.
MA, Z. & MA, Z. (2017). Exploration of large networks via fast and universal latent space model fitting. *arXiv*: 1705.02372.
NEWMAN, M. E. (2006). Modularity and community structure in networks. *P. Natl. Acad. Sci. U.S.A.* **103**, 8577–8582.
NEWMAN, M. E. & PARK, J. (2003). Why social networks are different from other types of networks. *Phys. Rev. E* **68**, 036122.
NEWMAN, M. E., WATTS, D. J. & Strogatz, S. H. (2002). Random graph models of social networks. *P. Natl. Acad. Sci. U.S.A.* **99**, 2566–2572.
ROHE, K., CHATTERJEE, S. & YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics* **39**, 1878–1915.
SARKAR, P. & MOORE, A. W. (2005). Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter* **7**, 31–40.
SCOTT, J. (2017). *Social network analysis*. London: Sage.
WATTS, D. J., DODDS, P. S. & NEWMAN, M. E. (2002). Identity and search in social networks. *Science* **296**, 1302–1305.
WENG, H. & FENG, Y. (2016). Community detection with nodal information. *arXiv*: 1610.09735.
WOLFE, A. W. (1997). Social network analysis: Methods and applications. *American Ethnologist* **24**, 219–220.
WU, Y.J., LEVINA, E. & ZHU, J. (2017). Generalized linear models with low rank effects for network data. *arXiv*: 1705.06772.
YAN, B. & SARKAR, P. (2020). Covariate regularized community detection in sparse graphs. *J. Am. Statist. Assoc.*, 10.1080/01621459.2019.1706541.
YAN, T., JIANG, B., FIENBERG, S. E. & LENG, C. (2019), Statistical inference in a directed network model with covariates. *J. Am. Statist. Assoc.* **114**, 857–868.
YUAN, X., & ZHANG, T. (2013). Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.*, **14**, 899–925.
ZHANG, Y., LEVINA, E. & ZHU, J. (2016). Community detection in networks with node features. *Electron. J. Statist.* **10**, 3153–3178.
ZOU, H., HASTIE, T. & TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15**, 265–286.

[*Received on* 30 *August* 2020]