Dimension reduction for covariates in network data

BY JUNLONG ZHAO, XIUMIN LIU
School of Statistics, Beijing Normal University,
19 Xinjekouwai Street, Haidian District, Beijing 100875, China
zhaojunlong928@126.com liuxiumin880407@yeah.net

HANSHEG WANG
Guanghua School of Management, Peking University,
5 Yiheyuan Road, Haidian District, Beijing 100871, China
hansheng@gsm.pku.edu.cn

AND CHENLEI LENG
Department of Statistics, University of Warwick, Coventry CV4 7AL, 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, called network-supervised dimension reduction, in which covariates are projected onto low-dimensional spaces to reveal 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 corresponds to stronger connections. Interestingly, the convergence rate of our estimator is found to depend on a network effect factor, which is the smallest number that can partition a graph in a manner similar to the graph colouring problem. Our method has interesting connections to principal component analysis and linear discriminant analysis, which we exploit for clustering and community detection. The proposed approach is further illustrated by numerical experiments and analysis of a pulsar candidates dataset from astronomy.

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

1. Introduction
Network data that include multiple objects with measurements on the interactions 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; Sarkar & Moore, 2005; Newman, 2006; 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 by $G = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of $n$ nodes and $E$ is the set of edges between the nodes. The relationships between 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 nodes.
i and j. For an unweighted graph, $w_{ij}$ is binary in that $w_{ij} = 1$ indicates the existence of a connection and $w_{ij} = 0$ indicates no connection. For a weighted graph, $w_{ij} \geq 0$ represents the strength of connection. The method developed in this paper works for both undirected and directed graphs.

As a reminder, for a directed graph, $w_{ij} > 0$ if there is a directed edge from node i to node j and $w_{ij} = 0$ otherwise. For an undirected graph, $W$ is symmetric so that $w_{ij} = w_{ji}$ for any $i \neq j$.

A distinctive feature of 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 gender, social status, education and so on, while edge variables measured on pairs of participants, such as whether the two participants share a 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 node $j$. 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 nodes $i$ and $j$. Popular choices in the literature include $X_{ij,t} = X_{i,t} - X_{j,t}$ ($t = 1, \ldots, p$) if the $r$th covariate is continuous, and $X_{ij,t} = I(X_{i,t} \neq X_{j,t})$ if the $r$th covariate is categorical, where $I(\cdot)$ denotes the indicator function. Our approach can take into account 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 the use of Markov chain Monte Carlo procedures for inference within maximum likelihood and Bayesian frameworks; Weng & Feng (2016), Zhang et al. (2016) and Huang & Feng (2020) for community detection in the stochastic block model; Wu et al. (2017) for use of the generalized linear model with low-rank effects; Graham (2017) for the $\beta$-model that assigns an 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 $\beta$-model. Deshpande et al. (2018) provided an information-theoretic analysis for inference of latent community structure given a sparse graph along with high-dimensional node covariates. These works typically assume a known link function that associates the probability of existence of an edge with 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 for uncovering latent communities in a graph, using a modification of spectral clustering. Yan & Sarkar (2020) proposed a community detection method for sparse networks with nodal information.

In this paper we propose a novel approach, referred to as network-supervised dimension reduction, which seeks to project the covariates onto a low-dimensional space to best explain the strength of connections 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 $\ell_2$-norm. Without loss of generality, we assume that a smaller value of $\|B^T X_{ij}\|$ corresponds to a stronger connection, i.e., a larger value of $w_{ij}$. As a concrete example, when nodal information is available and $B$ is the identity matrix, a small value of $\|X_{ij}\|$ 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 nodes.

Thus, we can state our problem as follows. Given data represented as a collection of tuples $\{s_{ij}, X_{ij}\}$ for $i \neq j$, our goal is 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 $\|B^TX_j\|$ is the smaller $s_{ij}$ is. To this end, 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 $\ell_2$ distance. This is achieved without imposing a restrictive independence assumption on the $w_{ij}$, or needing to assume a link function between $B^TX_j$ and $s_{ij}$. We show that the convergence rate of the projection depends critically on a factor referred to as the network effect of a graph, which is closely related to the graph colouring 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 notation is used throughout this paper. For any matrix $A = (a_{ij}) \in \mathbb{R}^{p \times p}$, $\|A\|_2$ and $\|A\|_F$ denote its operator norm and Frobenius norm, respectively, and $\|A\|_{\text{sup}} = \max_{i,j}|a_{ij}|$. Let $\{a_{ij}\}$ be a set whose items are all in $A$. For any symmetric matrix $A$, $\lambda_{\text{max}}(A)$ and $\lambda_{\text{min}}(A)$ stand for the maximum and minimum eigenvalues of $A$, respectively, and $\text{tr}(A)$ denotes the trace of $A$. For a vector $v \in \mathbb{R}^p$, $\|v\|$ denotes its $\ell_2$-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 $Z \in \mathbb{R}^p$, define $\|Z\|_{\psi_2} = \sup_{x \in S^{p-1}} \|\langle Z, x \rangle\|_\psi_2$, 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 $V$, $|V|$ denotes its cardinality. For any matrix $B$, we denote by $\text{span}(B)$ the space spanned by the columns of $B$ and let $P_B$ be the projection matrix onto the space span($B$). We write $a \wedge b = \min\{a, b\}$.

2. NETWORK-SUPERVISED DIMENSION REDUCTION

2.1. Notation and background

Recall that our data consist 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 $\|B^TX_j\|$ corresponds to a small value of $s_{ij}$. 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 $\Theta_{r,A} = \{B \in \mathbb{R}^{p \times r} : B^TABA = I_r\}$ for a symmetric positive-definite matrix $A \in \mathbb{R}^{p \times p}$ with eigenvalues uniformly bounded away from 0 and $\infty$. An obvious example is $A = I_p$. Since a small value of $\|B^TX_j\|$ 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 $H(B) = \{n(n-1)^{-1} \sum_{i \neq j} s_{ij} \|X_{ij}B\|^2 = \text{tr}(B^T \hat{G}B)$ with

$$\hat{G} = \frac{1}{n(n-1)} \sum_{i \neq j} s_{ij}X_{ij}X_{ij}^T = \frac{1}{n(n-1)} \sum_{i \neq j} Z_{ij}, \quad Z_{ij} = s_{ij}X_{ij}X_{ij}^T \in \mathbb{R}^{p \times p}. \quad (1)$$

This optimization problem for estimating the projection directions requires only 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 $r$ largest eigenvalues of $A^{-1/2} \hat{G}A^{-1/2}$. Then $\text{span}(\hat{B}_{r,A}) = \text{span}(A^{-1/2} \hat{\Psi}_r)$.

While $\text{span}(\hat{B}_{r,A})$ 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
Then when \( \hat{\phi} \), 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 the \( Z_{ij} \) have the same distribution, but are not necessarily independent, it can be seen that \( G_0 = G_{0n} = E(Z_{ij}) \). Let

\[
B_{r,A} = (\beta_{A,1}, \ldots, \beta_{A,r}) = \arg \max_{B \in \Theta_{r,A}} \text{tr}(B^T G_0 B),
\]

which is the population version of \( \hat{B}_{r,A} \). As in Proposition 1, assuming that the eigenvalues of \( A^{-1/2} G_0 A^{-1/2} \) are distinct, if we write \( \Psi_r \) for the matrix consisting of the eigenvectors associated with the \( r \) largest eigenvalues of \( A^{-1/2} G_0 A^{-1/2} \), then we also have \( \text{span}(B_{r,A}) = \text{span}(A^{-1/2} \Psi_r) \).

By an argument similar to that used for \( \hat{B}_{r,A} \), we simply set \( B_{r,A} = A^{-1/2} \Psi_r \).

We now state sufficient conditions which guarantee that the population maximizer of \( H(B) \) spans the same column space as spanned by the true projection directions. Letting the matrix \( A \) in \( \Theta_{r,A} \) be \( A = E(X_{ij} X_{ij}^T) \), which equals \( \text{cov}(X_{ij}) \) 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 \( A^{-1/2} G_0 A^{-1/2} \) are distinct;

(iii) \( \text{cov}(s_{ij}, (\beta_{mr}^T X_{ij})^2) > \text{cov}(s_{ij}, (v^T X_{ij})^2) \) (i \( \neq j; m = 1, \ldots, r \)) for any \( v \in \mathbb{R}^p \) satisfying \( v^T A B_0 = 0 \) and \( v^T A v = 1 \).

Then \( \text{span}(B_{r,A}) = \text{span}(B_0) \).

This proposition requires that the conditional mean of \( s_{ij} \) depend on \( X_{ij} \) only through the linear combination \( B_{0}^T X_{ij} \), with an unknown link function \( h \) that is left unspecified. This is reminiscent of the assumption in the literature of sufficient dimension reduction (Li, 1991), especially for making inference about the conditional mean of the response given the predictors (Cook & Li, 2002). The key difference is that the responses in our set-up are typically correlated owing to the existence of the network structure. Condition (i) can be seen as the true model. In particular, when \( s_{ij} \in \{0, 1\} \), for example \( s_{ij} = 1 - w_{ij} \), this condition says that \( \text{pr}(s_{ij} = 1) = h(B_{0}^T X_{ij}) \). The estimation procedure in (1) does not provide an estimator of \( h \). As such, our estimation procedure is model-free. To understand the condition (iii), consider the case where the covariates are defined as \( X_{ij} = X_i - X_j \) with \( X_i \sim N(\mu, \Sigma) \). Then this assumption becomes \( \text{cov}(s_{ij}, (X_{ij}^T \beta_m)^2) > 0 \), as shown in the Supplementary Material; it is intuitive since we expect a smaller \( s_{ij} \) to correspond to a small value of \( \|B_{0}^T X_{ij}\| \) and therefore 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 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, and both work 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 in such a way 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, 2019).
Recall that for a simplified stochastic block model with \( k \) communities, each node belongs to a latent community (Holland et al., 1983). Denote the latent community label of the \( i \)th node by \( C_i \), 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 \( \Pr(C_i = t) = \pi_t \) \((t = 1, \ldots, k)\), where the \( \pi_t \)s are unknown parameters satisfying \( \sum_{t=1}^{k} \pi_t = 1 \). Given their respective communities, nodes \( i \) and \( j \) make a connection with probability \( \Pr(w_{ij} = 1 | C_i, C_j) = \Pr_{CiCj} (i \neq j) \), independently of all other pairs, where \( \Pr_{CiCj} \) is a parameter that depends only on \( C_i \) and \( C_j \).

We look at a simplified stochastic block model where \( \Pr_{CiCj} = a_t \) for \( C_i = C_j = t \) and \( \Pr_{CiCj} = b \) for any \( C_i \neq C_j \); that is, all the probabilities of intercommunity connections are the same. For the covariates, we take \( X_{ij} = X_i - X_j \), where the covariate vector for node \( i \) satisfies

\[
X_i = \mu_{Ci} + \epsilon_i \quad (i = 1, \ldots, n)
\]  

for independent and identically distributed random variables \( \epsilon_i \) with \( E(\epsilon_i) = 0 \) and \( \text{cov}(\epsilon_i) = \Sigma_\epsilon \).

Here it is assumed that \( \epsilon_i \) 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. In such a setting, if \( s_{ij} \) is a one-to-one mapping of \( w_{ij} \), it is easily seen that \( E(s_{ij} | C_i = t, C_j = t') \) is a constant, depending on \( b \), for any \( t \neq t' \), which will be denoted by \( \gamma_t \) hereafter. Write \( E(s_{ij} | C_i = C_j = t) = \gamma_t \) \((t = 1, \ldots, k)\) for ease of notation. We point out that the \( w_{ij} \) in the above model depend only on the labels \( C_i \), which differs from the model assumed in condition (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 \( \text{cov}(X_i) \) 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 discriminant directions at the population level are the leading \( k - 1 \) eigenvectors of the generalized eigenvalue problem that solves \( \Sigma_{bt} U = \lambda \Sigma_{\epsilon} U \) for \( U \in \mathbb{R}^{p \times (k-1)} \), where \( \Sigma_{bt} = k^{-1} \sum_{t=1}^{k} (\mu_t - \bar{\mu})(\mu_t - \bar{\mu})^T \) with \( \bar{\mu} = (\sum_{t=1}^{k} \mu_t)/k \) and \( \Sigma_\epsilon \) is the covariance matrix of \( \epsilon_i \) defined above. We have the following result 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 the \( X_i \) are generated from model (3). If all the eigenvalues of \( A^{-1/2} G_A A^{-1/2} \) are distinct for \( A \) as specified below, the following conclusions hold.

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

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

(iii) 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 (3) at the population level in the sense that \( \beta_{A,m} \) in (2) is proportional to the \( m \)th direction of linear discriminant analysis, for \( m = 1, \ldots, r \).

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. In Proposition 3, it is assumed that \( \Pr(C_i, C_j) = b \) for any \( C_i \neq C_j \). By checking the proof, one can see that the proposed method may not be equivalent to principal component or linear discriminant analysis in general when this assumption does not hold. In this case, the associated objective function for our method can be viewed as a generalized version of those in principal component analysis and linear discriminant analysis. For community detection, we explain what we mean by further examining the special case of two communities in the situation where \( k = 2 \) 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 let \( \alpha_0 \in \mathbb{R} \) be a linear decreasing function of \( w_{ij} \). Suppose that the eigenvalues of \( A^{-1/2}G_0A^{-1/2} \) for \( A \) as specified below are distinct. The following conclusions hold.

(i) 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 \( \text{cov}(X_i) \) at the population level.

(ii) Let \( A = \text{cov}(X) \). 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 (3) at the population level.

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

To understand this corollary, assume for simplicity that the two communities are of equal size so that \( \pi_1 = \pi_2 = 1/2 \). In this case, (i) 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 simplifies to \( b = (a_1 + a_2)/2 \) when \( \alpha_0 > \alpha_1 \), because \( b \leq 1 \). In other words, when \( b = (a_1 + a_2)/2 \), the network information in terms of the adjacency matrix \( W \) does not contribute to identification of the projections. This is reasonable, since in that case the probabilities of connections between different communities are not small. When \( \pi_1 = \pi_2 = 1/2 \), the condition in (ii) becomes \( (a_1 + a_2)/2 > b \), and (ii) asserts the equivalence of the proposed approach and linear discriminant analysis when \( A = \text{cov}(X) \) 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), which require \( \min\{a_1, a_2\} > b \) in the setting above. In (iii), when \( A = \Sigma_c \), our approach is equivalent to linear discriminant analysis for any linear decreasing function when \( \alpha_0 \geq \alpha_1 > 0 \) and \( b < 1 \). When we apply our method to community detection in the Supplementary Material we show 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 (ii) and (iii) can be small, implying that our approach is applicable to sparse networks. Relevant simulation results are presented in §2.3.

The results in Proposition 3 and Corollary 1 suggest that the choice of \( A \) matters in making connections to principal component analysis and linear discriminant analysis. When \( A \) is not prespecified, we suggest taking \( A = \text{cov}(X) \). In practice, when \( A \) is estimated by \( \hat{A} \) from data, there will be an error between \( B_{r,A} \) and its population version \( B_{r,\.} \). We give a bound on this error in the Supplementary Material.
2.3. Application to community detection

Motivated by the covariate model in (3), we use network-supervised dimension reduction for community detection. To proceed, we first estimate the projection directions, denoted by $\hat{B}_{r,A}$ if $A$ is given or $\hat{B}_{r,\hat{A}}$ if $A$ is estimated as $\hat{A}$. We can then 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 can 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, it can be inferred 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 the community detection method based on our approach. The data are generated such that $W$ follows the simplified stochastic block model for the edges with two communities as in § 2.2 and $X$ follows the covariate model in (3), with $\epsilon_i$ 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 $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = 0.7|\delta - 1/3|$ in model (3). It is understood that as $u$ 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 $(a_1, a_2, b) = (\delta_0, \tau, 0.1)$, where $\delta_0$ and $\tau$ are constants in $[0, 1]$. It is known that a smaller $\delta_0$ gives a sparser network, and a smaller value of $\tau$ 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 magnitudes of $u$, $\delta_0$ and $\tau$, we 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. In particular, we consider the following three scenarios:

(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 $\tau$ in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$;
(c) fix $u = 1.6$ and $\tau = 0.1$ and vary $\delta_0$ 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 $A = \text{cov}(X)$, which is estimated by the sample covariance matrix of $X$. The second is $A = \Sigma_{x}$, the corresponding algorithm and simulation results for which are given in the Supplementary Material. These two choices of $A$ yield similar results. To apply network-supervised dimension reduction, the response variable $s_{ij}$ is taken to be $s_{ij} = 1 - w_{ij}$ and the number of the directions is set to $r = 1$. 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 nodes that are misclassified. The performance of K-means clustering after applying our approach is compared with that of the standard K-means clustering which uses only covariate information, as well as several competing methods that use information from both the network and the covariates, including the methods of Zhang et al. (2016), Binkiewicz et al. (2017) and Yan & Sarkar (2020). Moreover, in the Supplementary Material we consider cases where the network is dense in that $\delta_0$ is relatively large and the signal of the second community is strong, and we also report the performance of the method of Huang & Feng (2020) and the spectral clustering method of Rohe et al. (2011). The clustering errors for these methods are presented in Fig. 1 and in the Supplementary Material.

It is seen that for K-means clustering, the clustering error decreases as $u$ increases. The performance of the methods of Zhang et al. (2016) and Binkiewicz et al. (2017) improves when the network becomes dense, but since detecting the second group is difficult when $\tau = 0.1$, i.e., $a_2 = b$, both methods perform worse than the method of Yan & Sarkar (2020) and our method,
Fig. 1. Average clustering errors, comparing K-means clustering (dash-dot line), the proposed method (solid line), and the methods of Binkiewicz et al. (2017) (solid line with circles), Zhang et al. (2016) (solid line with triangles) and Yan & Sarkar (2020) (dashed line).

as shown in Fig. 1(c). In addition, we see that the method of Yan & Sarkar (2020) performs better than the methods of Zhang et al. (2016) and Binkiewicz et al. (2017) in most of the cases, but is worse than our method. Overall, clustering based on our approach performs much better than the competing methods, and it is rather insensitive to the parameters \( u, \delta_0 \) and \( \tau \). This implies that our approach can exploit the information in the covariates as well as the network structure. In particular, when \( \tau = 0.1 \) it is quite difficult to detect the second group; but with the help of the information from the covariates, K-means clustering based on our approach still estimates the community structure well. To gain more insight, we consider a simple case where \( r = 1 \), \( C_i \in \{1,2\} \) and \( \epsilon_i \) follows a normal distribution, and give an explicit bound on the classification error rate in the Supplementary Material. Finally, we also observe that for a network that has community structure with between-community probabilities dominating the within-community 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 \( G_0 = \lim_n E(\hat{G}) \). Because of 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 poses great challenges for theoretical analysis. We impose assumptions to rule out the cases where all entries of \( W \) can be strongly dependent, without explicitly modelling the dependence structure of the edges of the network.

We motivate our assumptions by generalizing an idea for inducing edge dependence that is widely used in the graphon model (Lovász & Szegedy, 2006; Diaconis & Janson, 2008; Bickel & Chen, 2009), and we will follow the notation in Gao et al. (2015). For an undirected graph, the graphon model assumes the edge random variables \( w_{ij} = w_{ji} \sim \text{Ber}(\theta_{ij}) \), where \( \theta_{ij} = f(\xi_i, \xi_j) \) (\( i \neq j \)). The sequence \( \{\xi_i\} \) consists of independent and identically distributed latent random variables from the uniform distribution on \([0,1]\), and given \( \{\xi_i\} \), the \( w_{ij} \) are independent for \( i < j \). The function \( f \), a bivariate function that is symmetric in its arguments, is called a graphon. In the graphon model, because the \( i \)th latent variable \( \xi_i \) is assumed to be associated with the \( i \)th 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, which is useful for characterizing the dependence structure in our set-up. Assume that \( \xi_i \) (\( i = 1, \ldots, n \)) and \( \xi_j \) (\( j = 1, \ldots, n \))
are independent and identically distributed latent random variables. Let $\Xi = (\xi_1, \ldots, \xi_n)^T \in \mathbb{R}^n$. Instead of associating a single element $\xi_i$ of $\Xi$ with node $i$ as in the graphon model, we associate with node $i$ a subset of $\Xi$ for introducing dependence, as well as an independent $\zeta_i$ for a node-specific effect. Denote the subset for node $i$ by $N_i = \{j : \xi_j$ is associated with node $i\}$. In the graphon model, $i \in N_j$. We then assume the edge random variable $w_{ij} \sim \text{Ber}(\theta_{ij})$, where

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

Here $\Xi_{N_i}$ is the subvector of $\Xi$ with indices in $N_i$. In our construction, we have purposely left unspecified the exact distributions of the random variables $\{\xi_i\}$ and $\{\zeta_i\}$, 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 $f_{ij}(\Xi_{N_i}, \xi_i, \Xi_{N_j}, \zeta_j, X_{ij}) = f(\xi_i, \zeta_j)$.

Write $N_j = N_i \cup N_j$ and let

$$\mathbb{V} = \{(i,j), (k,t) : N_j \cap N_k = \emptyset, i \neq j \neq k \neq t\}$$

be the set of pairs of nodes such that any two pairs 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 $|\mathbb{V}| \leq (\binom{n}{2})^2$. 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 with each node two latent random variables in $\Xi$. If we represent this example by a graph in which the nodes are $\{1, \ldots, n\}$ and an edge exists between nodes $i$ and $j$ if $N_i \cap N_j \neq \emptyset$, then it forms a cycle graph. In this example, it is not difficult to see that $|\mathbb{V}| = n(n-5)(2n^2 - 9n + 22) = O(n^4)$, which is of the same order as the maximum possible cardinality of $|\mathbb{V}|$.

Next we study $\|\hat{G} - G_0\|_{\text{op}}$. Establishing the rate of convergence of $\hat{G}$ in the operator norm is challenging because of the dependence between 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 the theoretical analysis. We overcome the dependency challenge by splitting the node pairs into groups such that any two node pairs in the same group are conditionally independent given covariates.

Let $\{\sigma(1), \ldots, \sigma(n)\}$ be any permutation of $\{1, \ldots, n\}$, and let $\xi_{\alpha,ij} = s_{ij}(\alpha^T X_{ij})^2 - E(s_{ij}^2) = E(s_{ij}(\alpha^T X_{ij}))^2$ for any given $\alpha \in \mathbb{R}^p$ satisfying $\|\alpha\| = 1$. Suppose that we split the index pairs $\{\sigma(i) = (\sigma(2i - 1), \sigma(2i))\}, i = 1, \ldots, n/2$ into $m$ groups $G_1, \ldots, G_m$ such that any two pairs $\sigma(i)$ and $\sigma(j)$ within the same group satisfy $\{\sigma(i), \sigma(j)\} \in \mathbb{V}$. That is, given $\{X_{ij}\}$, different $\xi_{\alpha,ij}$ with their $(i,j)$ in the same group are independent; this 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 the permutation $\{\sigma(1), \ldots, \sigma(n)\}$ is very challenging and can be viewed as a graph colouring problem where interest often lies in finding the chromatic number of the graph, defined as the minimum number of colours required for a vertex colouring scheme in which any two adjacent vertices are coloured differently; see the Supplementary Material for further discussion. Denote this number by $m_\sigma$, 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 $\|\hat{G} - G_0\|_{\text{op}}$ is presented in Theorem 1 below.

Moreover, we study the asymptotic properties of the eigenvalues and eigenvectors of $\hat{G}$. For this, we write the eigenvalue decompositions of $G_0$ and $\hat{G}$ as $G_0 = \sum_{i=1}^p \lambda_i v_i v_i^T$ and $\hat{G} = \sum_{i=1}^p \hat{\lambda}_i \hat{v}_i \hat{v}_i^T$, respectively, where $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ are the eigenvalues, and $v_i$
and $\hat{v}_i$ the associated eigenvectors. The eigenvalues and eigenvectors depend on $p$, but we omit $p$ hereafter for simplicity. Similarly, let

$$G_{0A} = A^{-1/2}G_0A^{-1/2} = \sum_{i=1}^{p} \phi_i^A \psi_i^A (\psi_i^A)^T, \quad \hat{G}_A = A^{-1/2}\hat{G}A^{-1/2} = \sum_{i=1}^{p} \hat{\phi}_i^A \hat{\psi}_i^A (\hat{\psi}_i^A)^T,$$

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 $\psi_i^A$ and $\hat{\psi}_i^A$ the associated eigenvectors. Recall the definitions of $B_{r,A}$ and $\hat{B}_{r,A}$ in §2. By Proposition 1, we see that

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

and that $B_{r,A} = (\beta_{A,1}, \ldots, \beta_{A,r}) = A^{-1/2}(\phi_1^A, \ldots, \phi_r^A)$. When $A$ is unknown and estimated as $\hat{A}$, we can define $\hat{G}_A$ and $\hat{\psi}_i^A$ analogously and estimate $B_{r,A}$ by $\hat{B}_{r,\hat{A}} = (\hat{\beta}_{\hat{A},1}, \ldots, \hat{\beta}_{\hat{A},r}) = \hat{A}^{-1/2}(\hat{\phi}_1^A, \ldots, \hat{\phi}_r^A)$.

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

**Assumption 1.** (i) For any integer $l > 0$ and any subset $I = \{(i_t,j_t), (t = 1, \ldots, l)\}$ satisfying $\{(i_t,j_t), (i_t',j_t')\} \in \mathcal{V}$ for any $t \neq t'$, $\{X_{i_t,j_t}, t = 1, \ldots, l\}$ are independent variables that follow sub-Gaussian distributions with $\max_{i \neq j} \|X_{ij}\|_2 < K_0 < \infty$ for some constant $K_0 > 0$. (ii) The conditional distribution of $s_{ij} | \{X_{ij}\}$ is the same as that of $s_{ij} | X_{ij}$.

**Assumption 2.** We have that $\delta = \inf_{i=1,\ldots,p} \lambda_i > 0$ and $\delta_A = \inf_{i=1,\ldots,p} (\phi_i^A - \phi_{i+1}^A) > 0$ uniformly over $p$.

When $X_{ij} = X_i - X_j$ where the $X_i$ are independent and identically distributed random variables following a sub-Gaussian distribution, (i) of Assumption 1 holds. Assumption 2 requires that all the eigenvalues of $G_0$ and $G_{0A}$ be distinct with positive gaps between them. We have the following convergence results.

**Theorem 1.** Assume $\max_{i \neq j} |s_{ij}| < c_0$ almost surely and that Assumptions 1 and 2 hold.

(i) We have that $\|\hat{G} - G_0\|_{op} = O_p(\delta_0^{op} + (pm_0^2/n)^{1/2})$, where $\delta_0^{op} = \|G_0 - G_0\|_{op}$.

(ii) If we assume further that $\|G_0\|_{op} < C_0$ for some constant $C_0$ independent of $p$, then for $i = 1, \ldots, p$,

$$|\hat{\lambda}_i - \lambda_i| = O_p\{\delta_0^{op} + (pm_0^2/n)^{1/2}\}, \quad \|\hat{v}_i - cv_i\| = O_p\{\delta_0^{op} + (pm_0^2/n)^{1/2}\},$$

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

Next, we provide an approximation to $m_{\text{net}}$ when an additional condition on the largest degree as in Assumption 3 below is imposed. Specifically, we only require the conditional independence property to hold for all but one of the groups. For $\hat{m}_{\text{net}}$ defined in Theorem 2 below, we show in the Supplementary Material that for any permutation $\{\sigma(1), \ldots, \sigma(n)\}$, one can always split the index pairs into $\hat{m}_{\text{net}}$ groups such that the conditional independence property holds for the first $\hat{m}_{\text{net}} - 1$ groups. In other words, for any $\sigma(i)$ and $\sigma(j)$ in $G_s$ with $s = 1, \ldots, \hat{m}_{\text{net}} - 1$, we have $\{\sigma(i), \sigma(j)\} \in \mathcal{V}$. By combining the conditional independence property for the first $\hat{m}_{\text{net}} - 1$ groups with Assumption 3, we show that the conclusions of Theorem 1 still hold, but with $m_{\text{net}}$ replaced by $\hat{m}_{\text{net}}$. 
Assumption 3. We have that \( d_{\text{max}} < \sqrt{n} \), where \( d_{\text{max}} = \max_{i=1,\ldots,n} |\{j : N_j \cap N_i \neq \emptyset\}| \).

It is easy to see that for the generalized graphon model, \( d_{\text{max}} \leq \max_i |N_i| \max_i |\{j : \xi_i \text{ is associated with node } j\}| \). Assumption 3 enables us to control the \( m_{\text{net}} \)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 Assumption 3 holds in Theorem 1. Let \( \tilde{m}_{\text{net}} = \log(n/4) / \log(4d_{\text{max}}/(4d_{\text{max}} - 1)) + 1 \). Then all the conclusions of Theorem 1 hold with \( m_{\text{net}} \) replaced by \( \tilde{m}_{\text{net}} \).

Theorems 1 and 2 establish the asymptotic properties of \( \hat{G} \). The term \( \delta_{\text{op}} \) in these theorems can be seen as the approximation error, and it is zero when the \( Z_j \) have the same distribution. The term \( (pm_{\text{net}}^2/n)^{1/2} \), or \( (\tilde{p}m_{\text{net}}^2/n)^{1/2} \), can be seen as the estimation error in which \( m_{\text{net}} \), or \( \tilde{m}_{\text{net}} \), can be loosely understood as the effect of a network. If \( d_{\text{max}} \) is bounded by a constant, then \( \tilde{m}_{\text{net}} = O(\log n) \) and the convergence rate of \( \hat{G} \) is \( O((p/n)^{1/2}\log n) \). If \( d_{\text{max}} = O(\log n) \), then upon noting that \( 1/\log(4d_{\text{max}}/(4d_{\text{max}} - 1)) = 1/\log[1 + (1/(4d_{\text{max}} - 1))] \approx 4d_{\text{max}} - 1 = O(\log n) \), we have \( \tilde{m}_{\text{net}} = O(n^{2}\log n) \) and the convergence rate of \( \hat{G} \) becomes \( O((p/n)^{1/2}(\log n)^2) \).

Following the proof of this theorem, it can be seen that if the \( s_{ij} \) are independent, then \( \|\hat{G} - G_0\|_{\text{op}} = O_p(\kappa_{\text{op}} + (p/n)^{1/2}) \). Theorem 1 indicates that if \( d_{\text{max}} \) is small, such as \( d_{\text{max}} = O(\log n) \), the convergence rate of \( \hat{G} \) is similar to that in 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 the proof of Theorem 1 shows, the conclusions of the theorem 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 into groups such that the \( s_{ij} \) in the same group are conditionally independent given \( \{X_{ij}\} \). Here the \( s_{ij} \) with \( (i,j) \) in different groups can still be correlated.

Using the relationship between \( \hat{G}_A \) and \( G \), one can derive 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 to \( \hat{B}_{r,A}^t \), when \( A \) is unknown and estimated as \( \hat{A} \). We make the following assumptions on the estimator of \( A \).

**Assumption 4.** We have that \( 0 < C^{-1} < \lambda_{\min}(A) \leq \lambda_{\max}(A) < C < \infty \) uniformly over \( p \).

**Assumption 5.** The estimator \( \hat{A} \) of \( A \) satisfies \( \|\hat{A}^{-1/2} - A^{-1/2}\|_{\text{op}} = O_p(\tau_n) \).

Assumption 4 is standard, and \( \tau_n \) in Assumption 5 is a function of \( n \) and \( p \), with \( p \) omitted for simplicity. The following theorem gives 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 Assumptions 1, 2 and 4 hold. Then the following conclusions hold.

(i) When \( A \) is known, \( \max_{i=1,\ldots,r} \|\hat{p}_{\delta_A,i} - c\beta_{\delta_A,i}\| = O_p(\delta_{\text{op}} + (pm_{\text{net}}^2/n)^{1/2}) \) for any given \( r = 1,\ldots,p \), where \( c \in (-1,1) \) such that \( c\beta_{\delta_A,i} > 0 \).
Theorem 3 shows that the convergence rate is determined by the approximation error $\delta_{A}^{\text{op}}$, the dimension $p$ of the covariates, the network effect $m_{\text{net}}$, and the convergence rate $\tau_{n}$ of $\hat{A}$, if $A$ is unknown. As in Theorem 2, we can replace the unknown $m_{\text{net}}$ with $\hat{m}_{\text{net}}$, as shown in the following theorem, the proof of which is the same as that of Theorem 3 and hence omitted.

**Theorem 4.** Suppose additionally that Assumption 3 holds in Theorem 3. The conclusions of Theorem 3 hold with $m_{\text{net}}$ replaced by $\hat{m}_{\text{net}}$.

When $wij$ and hence $s_{ij}$ depend on $n$, upon 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 $\phi_{1}^{A} \geq \cdots \geq \phi_{p}^{A}$ 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. Simulations

For the model in Proposition 2, it is assumed that $E(s_{ij} \mid X_{ij}) = h(B_{0}^{T}X_{ij})$. Proposition 2 shows that our method can be used to recover $\text{span}(B_{0})$ in this setting. To verify the effectiveness of the proposed network-supervised dimension reduction method in recovering $\text{span}(B_{0})$, we conduct extensive simulations. The performance of our method is examined by computing the error measure $\|P_{\hat{B}_{0}} - P_{\hat{B}_{r,\hat{A}}}\|_{F}$, 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 $A = \text{cov}(X_{1})$ and take the sample covariance matrix as its estimator. In all simulations, we take $s_{ij} = 1 - wij$ and assume that $r$ is known. We set $n = 100$ or 500 and take the dimension to be $p = 10$ or 50. For each example, 100 datasets are generated. Additional simulations for the selection of $r$ using the method in § 3 are presented in the Supplementary Material.
**Example 1.** We generate data according to the following procedure inspired by a similar set-up 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 $\Sigma = (\sigma_{t_1t_2})$ with $\sigma_{t_1t_2} = 0.4|t_1 - t_2|I(|t_1 - t_2| < 5)$.

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

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

where $\text{pr}(w_{ij} \mid C_i, C_j)$ is defined 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 (5), 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 $B_0 = (1, 1, 0, \ldots, 0)^T \in \mathbb{R}^p$, $a = 0.8$, $b = c_{\text{com}}a$ 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 size 0.5. Obviously, $c_{\text{com}} = 1$ corresponds to no community effect, while $c_{\text{com}} = 0.1$ corresponds to a strong community effect. A larger $c_{\text{coef}}$ 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, from 3.8% when $c_{\text{com}} = 0.1$ and $c_{\text{coef}} = 0.1$ to 41.5% when $c_{\text{com}} = 1.0$ and $c_{\text{coef}} = 0.5$. The simulation results are plotted in Fig. 2.

**Example 2.** Consider a setting where each node $i$ is affected by its $K$ neighbours, and denote the set of their indices by $\bar{N}_i$. The data are generated as follows.
We briefly discuss these results. It is easy to see that the influence of the covariate are similar for different because the covariates contribute more and more information with increasing $R$ are independent of each other. When $n$ increases, we see from Figures 2 and 3 that the average errors decrease as $n$ increases in both examples, which is expected from the theoretical results on the convergence rate. Interestingly, it can be seen from Fig. 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.

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

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

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

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 neighbour of node $i$. When $\bar{N}_i = \emptyset$, we see that node $i$ is affected only by its latent variable $Y_i$, where the $Y_i$ are independent of each other. When $\bar{N}_i \neq \emptyset$, we actually have an underlying network introduced by the latent neighbour sets $\bar{N}_i$. This network can be seen as the 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 underlying network as the true network between genes, and view the observed network represented by $w_{ij}$ as a noisy one contaminated by measurement errors and affected by environmental 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 we consider $K = 2$ and $K = 4$. 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 as defined in §3. The results of this simulation are shown in Fig. 3. We briefly discuss these results. It is easy to see that the influence of the covariate $X_{ij}$ decreases in both examples when $c_{\text{coef}}$ decreases. In particular, $X_{ij}$ has no effect when $c_{\text{coef}} = 0$. We can see from Figures 2 and 3 that the average errors decrease as $c_{\text{coef}}$ increases. This is reasonable because the covariates contribute more and more information with increasing $c_{\text{coef}}$. 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 can be seen from Fig. 2 that the errors are similar for different $c_{\text{com}}$ 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.
Table 1. True positive rate and false positive rate for variable selection when \((p, n) = (10, 100)\) in Example 1

| \(c_{\text{coef}}\) | \(c_{\text{com}} = 0.1\) | \(c_{\text{com}} = 0.5\) | \(c_{\text{com}} = 1.0\) |
|----------------------|-------------------------|-------------------------|-------------------------|
| \(0.5\)          | 0.49                    | 0.50                    | 0.51                    |
| \(1.0\)          | 1.00                    | 0.99                    | 1.00                    |
| \(1.5\)          | 1.00                    | 1.00                    | 1.00                    |
| \(2.0\)          | 1.00                    | 1.00                    | 1.00                    |

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

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 principal component analysis, which aims to find the eigenvectors of a matrix, we have developed a procedure similar to sparse principal component analysis for obtaining a sparse estimator of the projections as in Zou et al. (2006). More specifically, as in (4), our estimator is \(\hat{\beta}_{r,A} = (\hat{\beta}_{A,1}, \ldots, \hat{\beta}_{A,r})\) with \(\hat{\beta}_{A,j} = A^{-1/2} \hat{\phi}_j^A\), where \(\hat{\phi}_j^A\) is an eigenvector of \(\hat{G}_A = A^{-1/2} \hat{G} A^{-1/2}\). To implement our procedure, we use the truncated power algorithm in Yuan & Zhang (2013), and illustrate this algorithm as follows in the case where a sparse estimate of \(\beta_{A,1}\) is sought. 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 value, denote the resulting vector by \(\vartheta_t\), and set \(v_t = A^{1/2} \vartheta_t\). Repeat the procedure until \(\vartheta_t\) converges. The final \(\vartheta_t\) 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. It can be seen that the results are all satisfactory, especially when \(c_{\text{coef}} > 0.5\).

5. REAL DATA ANALYSIS

We apply the proposed method to a pulsar candidates dataset collected by the High Time Resolution Universe survey (Keith et al., 2010), available at http://archive.ics.uci.edu/ml/datasets/HTRU2. Pulsars are a rare type of neutron star that produces radio emissions detectable on Earth. They are of considerable scientific interest as probes of space-time, the interstellar medium and states of matter. Their study can lead to a better understanding of many physics problems, ranging from acceleration of particles in the ultrastrong magnetic field to tests of gravity in the strong field regime. Since some pulsars are binary systems (Lyne & Smith, 2012), the signals detected for a star may be mixed signals from that star and its neighbours, together with some noise, implying that our generalized graphon model is applicable to such data. In the dataset we analyse, each pulsar is described by eight continuous variables and a single class variable; the data include 16 259 spurious examples caused by radio frequency interference or noise and 1639 real pulsar examples that 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 versus 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
Fig. 4. Comparison of different algorithms on the real data: our method with $A = \text{cov}(X)$ (solid line) and with $A = \Sigma_e$ (short-dash line), the K-means method (dash-dot line), and the methods of Binkiewicz et al. (2017) (solid line with circles), Zhang et al. (2016) (solid line with triangles), Yan & Sarkar (2020) (long-dash line), Rohe et al. (2011) (solid line with diamonds) and Huang & Feng (2020) (solid line with squares).

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 the 16 259 spurious examples and 100 observations from the 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, i.e., the mean of the integrated profile, is small. We choose a threshold such that the network density, defined as the ratio of edges to the maximum possible number of edges, is 50%, 30%, 10%, 5%, 3% or 1%. The remaining 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 §3 by setting $M = 4$. Our approach is compared with the methods of Zhang et al. (2016), Binkiewicz et al. (2017), Huang & Feng (2020) and Yan & Sarkar (2020), as well as the spectral clustering method of Rohe et al. (2011). In addition, we include a version of our approach that estimates $A = \Sigma_e$ via the algorithm in the Supplementary Material with $r = 1$. Since the true community membership of each node is known, we report the average of the proportions of nodes that are misclassified. It is obvious that the smaller this quantity is, the better a method is. The results averaged over 100 random datasets are plotted in Fig. 4. It is observed that our network-supervised dimension reduction approach with the two choices of $A$ performs better than the other methods in most of the cases, especially when the network is sparse. Moreover, our approach is insensitive to the sparsity of the network, while the methods of Binkiewicz et al. (2017), Rohe et al. (2011) and Zhang et al. (2016) work only when the network is dense.

ACKNOWLEDGEMENT

We thank three reviewers, an associate editor and the editor for constructive comments that have led to a much improved paper. The research of Zhao was supported by the National Science Foundation of China. Wang was partially supported by the National Natural Science Foundation
of China and China’s National Key Research Special Program. Leng was 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. The code for community detection is available at https://github.com/DR-Network/community-detection.

**References**

Amini, A. A. & Levina, E. (2018). On semidefinite relaxations for the block model. *Ann. Statist.* **46**, 149–79.

Bickel, P. J. & Chen, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Nat. Acad. Sci.* **106**, 21068–73.

Binkiewicz, N., Vogelstein, J. T. & Rohe, K. (2017). Covariate-assisted spectral clustering. *Biometrika* **104**, 361–77.

Cook, R. D. & Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455–74.

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. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*. New York: Curran Associates, pp. 8581–93.

Fienberg, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *J. Comp. Graph. Statist.* **21**, 825–39.

Gao, C., Lu, Y. & Zhou, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43**, 2624–52.

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**, 1033–63.

Hoff, P. D., Rafter, A. E. & Handcock, M. S. (2002). Latent space approaches to social network analysis. *J. Am. Statist. Assoc.* **97**, 1090–8.

Holland, P. W., Laskey, K. B. & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Netw.* **5**, 109–37.

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. (2020). Pairwise covariates-adjusted block model for community detection. *arXiv*: 1807.03469v2.

Hunter, D. R., Goodreau, S. M. & Handcock, M. S. (2008). Goodness of fit of social network models. *J. Am. Statist. Assoc.* **103**, 248–58.

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. (2019). *Applied Multivariate Statistical Analysis*. Upper Saddle River, New Jersey: Prentice Hall, 6th ed.

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. et al. (2010). The high time resolution universe pulsar survey – I. System configuration and initial discoveries. *Mon. Not. R. Astron. Soc.* **409**, 619–27.

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**, 316–27.

Lovász, L. & Szegedy, B. (2006). Limits of dense graph sequences. *J. Comb. Theory B* **96**, 933–57.

Lyne, A. G. & Smith, F. G. (2012). *Pulsar Astronomy*. Cambridge: Cambridge University Press.

Ma, Z. & Ma, Z. (2017). Exploration of large networks via fast and universal latent space model fitting. *arXiv*: 1705.02372v2.

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Nat. Acad. Sci.* **103**, 8577–82.

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. *Proc. Nat. Acad. Sci.* **99**, 2566–72.

Rohe, K., Chatterjee, S. & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39**, 878–915.

Sarkar, P. & Moore, A. W. (2005). Dynamic social network analysis using latent space models. *ACM SIGKDD Explor. 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, 1302–5.
WENG, H. & FENG, Y. (2016). Community detection with nodal information. arXiv: 1610.09735v2.
WOLFE, A. W. (1997). Social network analysis: Methods and applications. Am. Ethnologist 24, 219–20.
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., DOI: 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, 857–68.
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, 3153–78.
ZOU, H., HASTIE, T. & TIBSHIRANI, R. (2006). Sparse principal component analysis. J. Comp. Graph. Statist. 15, 265–86.

[Received on 30 October 2019. Editorial decision on 25 January 2021]