OXFORD UNIVERSITY PRESS | Biometrika

Dimension Reduction for Network Data

Journal:	*Biometrika*
Manuscript ID	BIOMTRKA-19-515
Manuscript Type:	Article
Date Submitted by the Author:	30-Oct-2019
Complete List of Authors:	Zhao, Junlong; Beijing Normal University, Department of statistics Liu, Xiumin; Beijing Normal University, School of Ststistics; Wang, Hansheng; Peking University, Guanghua School of Management; Leng, Chenlei; The University of Warwick, Department of Statistics;
Keywords:	Clustering, Community detection, Dimension reduction, Graph, Network

Biometrika (2019), **xx**, x, *pp*. 1–18
Printed in Great Britain

Dimension Reduction for 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 we formulate as finding the smallest number to partition a graph, in a way similar to that in 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 $\mathcal{V} = \{1, 2, \cdots, n\}$ 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 adjacent 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,$w_{ij} \geq 0$ is valued in that the magnitude of w_{ij} indicates 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 $\mathbf{X}_{ij} = (X_{ij,1}, \cdots, X_{ij,p})^T$, 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 $\mathbf{X}_{ij} = g(X_i, X_j)$ for some prespecified function $g(\cdot, \cdot)$, where $X_i \in \mathbb{R}^p$ is the nodal covariate of the ith node. Popular choices widely used in the literature include $\mathbf{X}_{ij,k} = X_{i,k} - X_{j,k}, 1 \leq k \leq p$ if the kth covariate is continuous, and $\mathbf{X}_{ij,k} = I(X_{i,k} = X_{j,k})$ if it is categorical, where $I(\cdot)$ is the indicator function. 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. 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.

In this paper, we propose a novel approach named Network-supervised Dimension Reduction (NDR) 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 $\|B^T\mathbf{X}_{ij}\|$ 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 $\|B^T\mathbf{X}_{ij}\|$ 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 $\mathbf{X}_{ij} = X_i - X_j$ will correspond to a large value of w_{ij} intuitively. If a large value of $X_{ij,k}$ corresponds to a large w_{ij}, we can re-define $X_{ij,k}$ as $c - X_{ij,k}$ for some constant c such that small values of $X_{ij,k}$ correspond to large values of w_{ij}. In practice, we can also work with $\mathbf{S} = (s_{ij}) \in \mathbb{R}^{n\times n}$ with $s_{ij} = f(w_{ij})$, where f is a monotonic one-to-one function from $\mathbb{R}$ to $\mathbb{R}$. In the simplest case, $s_{ij} = w_{ij}$. The use of f allows a more general notion of similarity between the nodes in a network. The interpretation of s is similar to that of w in that a larger value of s 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 $\{s_{ij}, \mathbf{X}_{ij}\}$ for $1 \le i \ne j \le n$, we seek to find a matrix $B \in \mathbb{R}^{p\times r}$ to project $\mathbf{X}$ such that the value of $\|B^T\mathbf{X}_{ij}\|$ reflects the similarity of the nodes in terms of s_{ij}. More precisely, the projection is such that the smaller $\|B^T\mathbf{X}_{ij}\|$ is, the larger 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 $B^T\mathbf{X}$ and s. 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 NDR and principal component analysis (PCA), as well as between NDR and linear discriminant analysis (LDA). The connection to the latter enables us to leverage covariate information for better community detection, which we illustrate via simulations showing that an NDR-based clustering algorithm outperforms K-means clustering based only on covariate information and community detection based only on the adjacency matrix information.

The main content of the paper is organized as follows. In Section 2, we propose NDR, establish its connection to PCA and LDA, and illustrate its applications in community detection. Asymptotic properties of the estimator are established in Section 3. Simulation results and a real data analysis are presented in Section 4 and 5, respectively. A short discussion on future work is found in Section 6. Technical details of the proofs are relegated to the Supplementary Materials.

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 $\|A\|_{\max} = \max_{i,j}|a_{ij}|$. 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. We use I_n to denote the $n \times n$ identity matrix. For any set S, we use $|S|$ to denote its cardinality.

2. Network-supervised Dimension Reduction

2.1. *Notation and background*

Recall that our data consists of network-covariate tuple $\{s_{ij}, \mathbf{X}_{ij}\}, 1 \le i \ne j \le n$. Our goal is to find $B \in \mathbb{R}^{p\times r}$ such that a small value of $\|B^T\mathbf{X}_{ij}\|$ corresponds to a large value of s_{ij}. In the danger of causing confusion, we refer to B as the projection matrix and its columns as the NDR directions of the projection. To partially ensure identifiability of B, we constrain $B \in \Theta_{r,A}$, where $\Theta_{r,A} \subset \mathbb{R}^{p\times r}$ satisfies

$$\Theta_{r,A} = \{U \in \mathbb{R}^{p\times r} : U^TAU = I_r\},$$

for a symmetric positive definite matrix $A \in \mathbb{R}^{p\times p}$ whose eigenvalues are bounded away from 0 and ∞ uniformly for all p. An obvious example is $A = I_p$. Since a small value of $\|B^T\mathbf{X}_{ij}\|$ corresponds to a large value of s_{ij}, a natural estimator of B is found as

$$\hat{B}_{r,A} = (\hat{\beta}_{A,1}, \cdots, \hat{\beta}_{A,r}) = \arg\min_{U\in\Theta_{r,A}} H(U),$$

where

$$H(U) = \frac{1}{n(n-1)}\sum_{i\ne j} s_{ij}\|\mathbf{X}_{ij}^TU\|^2 = tr\left(U^T\hat{G}U\right), \quad (1)$$

by denoting

$$\hat{G} = \frac{1}{n(n-1)} \sum_{i \neq j} s_{ij} \mathbf{X}_{ij} \mathbf{X}_{ij}^T := \frac{1}{n(n-1)} \sum_{i \neq j} Z_{ij},$$

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

PROPOSITION 1. *Let B be the matrix consisting of the eigenvectors associated with the r smallest eigenvalues of $A^{-1/2} \hat{G} A^{-1/2}$. Then $\hat{B}_{r,A} = A^{-1/2} B$.*

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

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

which is the population version of $\hat{B}_{r,A}$. Similar to Proposition 1, if we denote B as the matrix consisting of the eigenvectors associated with the r smallest eigenvalues of $A^{-1/2} G_0 A^{-1/2}$, then we also have $B_{r,A} = A^{-1/2} B$.

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

PROPOSITION 2. *Assume that the following conditions hold: (i) $E(s_{ij} | \mathbf{X}_{ij}) = f(\boldsymbol{\beta}^T \mathbf{X}_{ij})$ where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_r) \in \mathbb{R}^{p \times r}$ satisfies $\boldsymbol{\beta} \in \Theta_{r,A}$, and f is monotonically increasing but unspecified; (ii) $\mathrm{cov}(s_{ij}, (\beta_m^T \mathbf{X}_{ij})^2) < \mathrm{cov}(s_{ij}, (v^T \mathbf{X}_{ij})^2)$, for any v such that $v^T A \boldsymbol{\beta} = 0$, $1 \leq i \neq j \leq n, 1 \leq m \leq r$. Then, it holds that* $\mathrm{span}(B_{r,A}) = \mathrm{span}(\boldsymbol{\beta})$.

This proposition requires that the conditional mean of s_{ij} depends on $\mathbf{X}_{ij}$ only through the linear combination $\boldsymbol{\beta}^T \mathbf{X}_{ij}$. 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 network structure. The unknown link function f is left unspecified. The estimation procedure in (1) does not offer an estimator of f. As such, our estimation procedure is model free. To understand assumption (ii), consider the case when the covariates are defined as $\mathbf{X}_{ij} = X_i - X_j$ with $\mathbf{X}_i \sim N(\mu, \Sigma)$. Then this assumption becomes $\mathrm{cov}(s_{ij}, (\mathbf{X}_{ij}^T \beta_m)^2) < 0$ as shown after the proof of this proposition in the Supplementary Materials. Assumption (ii) is reasonable since we expect that a large s corresponds to a small value of $\|\beta^T \mathbf{X}_{ij}\|$ and subsequently a small value of $(\mathbf{X} \beta_m)^2$.

2.2. *Connections to PCA and LDA*

In the context of the so-called stochastic block model (SBM), we establish novel connections between NDR and PCA and between NDR and LDA in this subsection. PCA and LDA are two widely used statistical methods for reducing the dimensionality of data both by finding the best linear combinations of covariates. PCA is an unsupervised method that projects observations onto the so-called principal component directions such that the variance of the projected data is maximized. On the other hand, LDA 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 $C_i \in \{1, \cdots, k\}$ for $1 \le i \le n$. The SBM assumes that these community labels are $i.i.d.$ random variables such that $P(C_i = t) = \pi_t$, $1 \le t \le k$, 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

$$P(w_{ij} = 1 | C_i, C_j) = P_{C_i C_j}, \quad 1 \le i \ne j \le n,$$

independent of all other pairs, where $P_{C_i C_j}$ is a parameter depending only on C_i and C_j. We look at a simplified SBM where

$$P_{C_i C_j} = \begin{cases} a_t & \text{for } C_i = C_j = t \\ b & \text{for } C_i \ne C_j \end{cases}.$$

That is, all the probabilities of inter-communities connections are the same. For the covariates, we take $\mathbf{X}_{ij} = X_i - X_j$, where the covariate vector for the ith node satisfies

$$X_i = \mu_{C_i} + \epsilon_i, \qquad 1 \le i \le n. \tag{3}$$

for $i.i.d.$ 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 SBM above. That is, the covariates follow a multivariate normal 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 $E(s_{ij} | C_i = t, C_j = t')$ is a constant (depending on b) for any $1 \le t \ne t' \le k$, which will be denoted as γ_0 hereafter. We denote $E(s_{ij} | C_i = C_j = t) = \gamma_t$ for $1 \le t \le k$ for ease of notation.

If we apply PCA 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 LDA to the labelled data $\{C_i, X_i\}_{i=1}^n$ assuming that the latent community labels are known in model (3), the LDA directions at the population level are the leading $k-1$ eigenvectors of the generalized eigenvalue problem that solves $\Sigma_b U = \lambda \Sigma_\epsilon U$ for $U \in \mathbb{R}^{p \times (k-1)}$, where

$$\Sigma_b = \frac{1}{k} \sum_{i=1}^{k} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T$$

with $\bar{\mu} = (\sum_{i=1}^k \mu_i)/k$, and $\Sigma_\epsilon = \mathrm{cov}(\epsilon)$ is the common covariance matrix across the classes.

PROPOSITION 3. *Assume that $W = (w_{ij})$ is generated from the stochastic block model above and X_i's are from model* (3)*. The following conclusions hold.*

(1) If we set $A = I_p$ in $\Theta_{A,r}$, NDR is equivalent to PCA conducted as eigenvalue decomposition of $\mathrm{cov}(X_i)$ at the population level, if and only if

$$\gamma_0 = \sum_{t=1}^{k} \pi_t^2 \gamma_t / \sum_{t=1}^{k} \pi_t^2 < 0.$$

(2) If $\gamma_0 < 0$, $\sum_{t=1}^{k} \pi_t^2 \gamma_t + \gamma_0 \sum_{t_1 \ne t_2} \pi_{t_1} \pi_{t_2} > 0$, and we choose $A = \Sigma_\epsilon$ in $\Theta_{A,r}$, then NDR is equivalent to LDA for the model in (3)*.*

This proposition shows that NDR can be equivalent to unsupervised PCA or supervised LDA, depending on the data generating process. We explain what we mean by this further by examining the special case of two communities when $k = 2$ and $s_{ij} = 2w_{ij} - 1 \in \{1, -1\}$. Recall the definition of $B_{r,A}$ in (2).

COROLLARY 1. *Assume that* $\mathbf{X}_{ij}$ *and* $W = (w_{ij})$ *are generated as in Proposition* 3. *The following conclusions hold.*

(1) If $A = I_p$ *and* $b = (\pi_1^2 a_1 + \pi_2^2 a_2)/(\pi_1^2 + \pi_2^2) < 1/2$, *then NDR is equivalent to PCA conducted as eigenvalue decomposition of* $\mathrm{cov}(X_i)$ *at the population level.*

(2) If $A = \Sigma_\epsilon$ *and* $(1 - 2\pi_1^2 a_1 - 2\pi_2^2 a_2)/(4\pi_1\pi_2) < b < 1/2$, *then the first direction* $\beta_{A,1}$ *of NDR is equivalent to that of LDA for the model in* (3).

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 that NDR is equivalent to PCA if and only if $b = (a_1 + a_2)/2 < 1/2$. That is, when $a_1 + a_2 < 1$, 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 within the same communities is small, we do not expect the adjacency matrix to be useful. On the other hand, (2) states that NDR is equivalent to LDA when $1 - (a_1 + a_2)/2 < b < 1/2$. This is the case when $a_1 + a_2 > 1$, implying that when the link probabilities of the nodes in the same communities are large and those between different communities are small, NDR aims to find the projected space in a way similar to LDA.

2.3. *Application in community detection*

Motivated by the covariate model in (3), we propose using NDR for community detection via the following algorithm when nodal covariates are present.

Step 0. Initiate $\hat{\Sigma}_\epsilon = I_p$. For a given number of communities k and a given number of NDR directions r, repeat the following two steps until convergence is declared.

Step 1. Given $\hat{\Sigma}_\epsilon$, apply NDR to obtain the directions $\hat{B}_{r,A}$, and estimate the latent community labels by applying a clustering algorithm on the projected observations $\{o_i : o_i = \hat{B}_{r,A}^T X_i, i = 1, 2, \cdots, n\}$.

Step 2. Given the estimated labels (clusters) in Step 1, estimate Σ_ϵ as the average of the intra-cluster sample covariance matrices.

We stop the iteration when $R = T_{in}/T_{total}$ is small, where T_{in} and T_{total} are the within-class sum of squares and the total sum of squares, respectively.

We now present the result of a small numerical experiment to evaluate the performance of this NDR-based community detection method. The data is generated such that W follows the simplified SBM for the edges as in Section 2.2 and X follows the covariate model in (3). We set $p = 5$, $\mu_1 = (u, 0, \cdots, 0)^T \in \mathbb{R}^p$ where $u = 1.0, 1.2, 1.4, 1.6, 1.8$ or 2.0, $\mu_2 = -\mu_1$ 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. By varying the value of u, we want to assess how our approach performs with respect to the informativeness of the covariates in clustering. In the SBM, we set $\pi_1 = \pi_2 = 1/2$, $a_1 = 0.8$, $b = 0.3$ and evaluate three values of a_2 in that $a_2 = 0.7, 0.4$, or 0.3. It is understood that a small a_2 gives a weaker community in the second group. By varying the magnitude of a_2, we want to evaluate how the proposed method performs with respect to the strength of the community structure. The response variable s in NDR is taken as $s_{ij} = 2w_{ij} - 1$ and the number of NDR 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

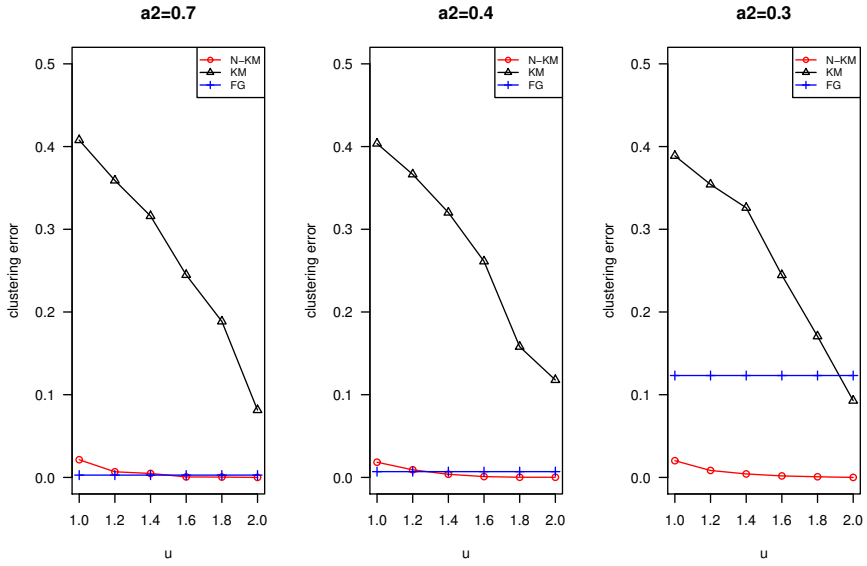


Fig. 1. The average clustering errors: K-Means clustering that only uses covariates information (KM); the NDR-based K-Means clustering that uses information both in the covariates and in the adjacency matrix (N-KM); the fast greedy community detection method in Clauset et al. (2004) that only uses the information in the adjacency matrix (blue solid lines).

approach is evaluated by calculating its clustering errors defined as the proportions of the nodes that are misclassified.

We apply K-means clustering as the clustering method in our algorithm. The performance of our algorithm for community detection is compared to the standard K-means clustering that only uses covariate information and to the community detection method in Clauset et al. (2004) that only uses the information in the adjacency matrix W via a hierarchical agglomeration algorithm. For the latter, we use the R function "cluster_fast_greedy" in package "igraph" that implements the method in Clauset et al. (2004). The clustering errors for these three methods are presented in Figure 1. It is seen that for K-means clustering, the clustering error decreases as u increases, while the fast greedy method of Clauset et al. (2004) becomes worse as a_2 decreases. Overall, the NDR-based K-means clustering method performs much better than the other two competitors except when u is small, and is rather insensitive to a_2 and u. This implies that our approach can exploit the information in the covariates as well as the network structure. Note that when $a_2 = 0.3$, there is only one community in the generated network. However, with the help of the information from the covariates, the NDR-based K-means clustering method still estimates the community structure well.

In addition, we present the projected data points along the NDR direction in one simulation in Figure 2 where $a_2 = 0.4$, and $u = 1$ or 1.4. It is seen that the projected values of the two communities are well separated. This enables the K-means algorithm to work well, and consequently leads to a good estimate of the latent labels.

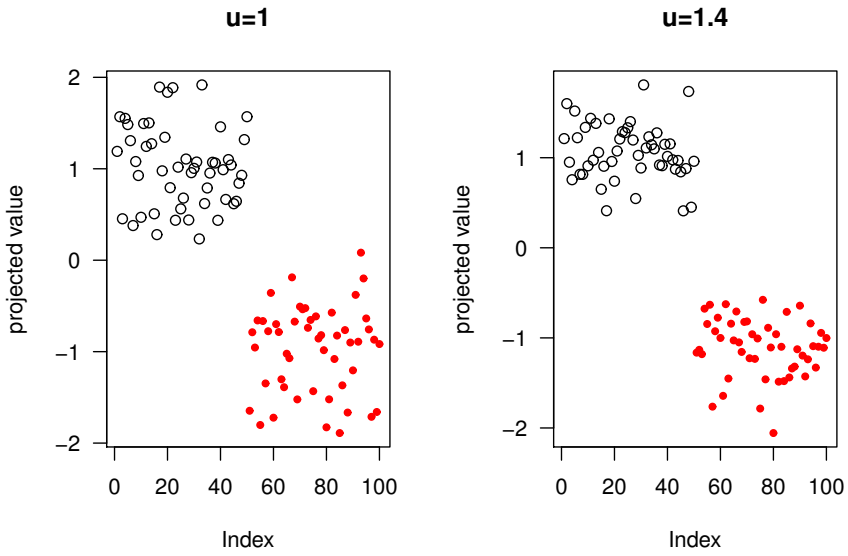


Fig. 2. The observations projected along the NDR direction in one simulation. The two subgroups are presented in different colors. The numbers in x-axis are the indices of the observations.

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. For example, if all the entries of the resulting S matrix are strongly correlated, $\hat{G}$ may not converge to its expectation at all. A simple case is when $s_{ij} \equiv s$ for all $i \neq j$, where s is a non-degenerate random variable. In this case,

$$\hat{G} = s \cdot [n(n-1)]^{-1} \sum_{i \neq j} \mathbf{X}_{ij} \mathbf{X}_{ij}^T := s \cdot U_{\mathbb{X}}.$$

Since $U_{\mathbb{X}}$ is a U-statistic, under suitable regularity conditions, it is expected that $U_{\mathbb{X}}$ will converge to its expectation as n increases. Then in the limit, $\hat{G}$ as the product of a random variable s and a constant remains a random variable. To overcome this issue of non-existence of a deterministic limit of $\hat{G}$, we impose assumptions to rule out similar 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

$$\theta_{ij} = f(\xi_i, \xi_j), \qquad 1 \leq i \neq j \leq n.$$

The sequence $\{\xi_i\}$ are the $i.i.d.$ latent random variables that are from the uniform distribution on $[0, 1]$, and given $\{\xi_i\}$, w_{ij}'s are independent for $1 \leq i < j \leq n$. 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 $\xi_i, 1 \le i \le n$ and $\zeta_j, 1 \le j \le n$ are $i.i.d.$ latent random variables. Denote $\Xi = (\xi_1, ..., \xi_n)^T \in \mathbb{R}^n$. 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 $N_i := \{j : \xi_j \text{ is associated with node } i\}$. Note that in the graphon model, $i \in N_i$. We then assume the edge random variable $w_{ij} \sim \text{Bernoulli}(\theta_{ij})$, where

$$\theta_{ij} = f_{ij}(\Xi_{N_i}, \zeta_i, \Xi_{N_j}, \zeta_j, \mathbf{X}_{ij}), \qquad 1 \le i \ne j \le n. \tag{4}$$

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 $f_{ij}(\Xi_{N_i}, \zeta_i, \Xi_{N_j}, \zeta_j, \mathbf{X}_{ij}) = f(\xi_i, \xi_j)$.

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

$$\mathbb{V} = \Big\{\{(i,j),(k,t)\} : N_{ij} \cap N_{kt} = \emptyset, 1 \le i \ne j \ne k \ne t \le n\Big\}$$

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 $\mathbf{X}_{ij}$ and $\mathbf{X}_{kt}$. The cardinality of $\mathbb{V}$ provides a rough characterization of the dependence structure of a network intuitively and is seen to be bounded as $|\mathbb{V}| \le \binom{n}{4}$. 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 $\{1, ..., n\}$ and an edge exists between the ith and jth nodes if $N_i \cap N_j \ne \emptyset$, then it forms a cycle graph. For this example, it is not difficult to see that $|\mathbb{V}| = 24[\sum_{i=1}^4 \binom{n-8}{i} + \sum_{i=1}^3 \binom{n-7}{i}] + 8\sum_{i=1}^2 \binom{n-6}{i} = O(n^4)$ 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.

To obtain the properties of $\|\hat{G} - G_0\|_{op}$, we first establish a concentration inequality taking the form of $\alpha^T(\hat{G} - G_0)\alpha$ for any $\|\alpha\| = 1$. To this end, the key step is to upper bound $E\left(\exp\{g_{(\sigma(1),\cdots,\sigma(n))}\}\right)$ where

$$g_{(\sigma(1),\cdots,\sigma(n))} = \frac{2\lambda}{n}\left[\xi_{\alpha,\sigma(1)\sigma(2)} + \xi_{\alpha,\sigma(3)\sigma(4)} + \cdots + \xi_{\alpha,\sigma(n-1)\sigma(n)}\right],$$

for a constant $\lambda > 0$. Here $(\sigma(1), \cdots, \sigma(n))$ can be any permutation of $\{1, \cdots, n\}$ and $\xi_{\alpha,ij} = s_{ij}(\alpha^T X_{ij})^2 - E(s_{ij}(\alpha^T X_{ij})^2)$. Suppose that we split the index pairs $\{\boldsymbol{\sigma}(i) = (\sigma(2i-1), \sigma(2i)), i = 1, \cdots, n/2\}$ into m groups $G_1, \cdots, G_m$ such that any two pairs $\boldsymbol{\sigma}(i)$ and $\boldsymbol{\sigma}(j)$ within the same groups satisfy $\{\boldsymbol{\sigma}(i), \boldsymbol{\sigma}(j)\} \in \mathbb{V}$. That is, given $\{\mathbf{X}_{ij}\}$, $\xi_{\alpha,ij}$'s with (i,j)'s in the same group are independent, which will be referred to as the *conditional independence property*

hereafter. Then

$$g_{(\sigma(1),\cdots,\sigma(n))} = \frac{2\lambda}{n}\sum_{s=1}^{m}\sum_{(i,j)\in G_s}\xi_{\alpha,ij} := \frac{1}{m}\sum_{s=1}^{m} Y_s,$$

where $Y_s = \frac{2m\lambda}{n}\sum_{(i,j)\in G_s}\xi_{\alpha,ij}$. Based on the conditional independence property and if $\mathbf{X}_{ij}$ is sub-Gaussian, one can show that

$$E\left(\exp\{Y_s\}\right) \le \exp\left(\frac{4(m\lambda K)^2}{n}\right), \quad 1 \le s \le m.$$

Then by Jensen's inequality, we have

$$E\left(\exp\{g_{(\sigma(1),\cdots,\sigma(n))}\}\right) \le \exp\left(\frac{4(m\lambda K)^2}{n}\right). \tag{5}$$

Obviously, a smaller m is desired as it leads to a tighter upper bound on $E\left(\exp\{g_{(\sigma(1),\cdots,\sigma(n))}\}\right)$. Finding the smallest m associated with permutation $(\sigma(1),\cdots,\sigma(n))$ 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

$$m_{\text{net}} = \max_{(\sigma(1),\cdots,\sigma(n))} m_\sigma,$$

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

$$G_0 = \sum_{i=1}^{p}\lambda_i v_i v_i^T, \qquad \hat{G} = \sum_{i=1}^{p}\hat{\lambda}_i \hat{v}_i \hat{v}_i^T,$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$ and $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_p$ 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

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

where $\phi_1^A \ge \phi_2^A \ge \cdots \ge \phi_p^A$ and $\hat{\phi}_1^A \ge \hat{\phi}_2^A \ge \cdots \ge \hat{\phi}_p^A$ 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

$$\hat{B}_{r,A} = (\hat{\beta}_{A,1},\cdots,\hat{\beta}_{A,r}) = A^{-1/2}(\hat{\varphi}_p^A,\cdots,\hat{\varphi}_{p-r+1}^A),$$

$$B_{r,A} = (\beta_{A,1},\cdots,\beta_{A,r}) = A^{-1/2}(\varphi_p^A,\cdots,\varphi_{p-r+1}^A). \tag{6}$$

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

$$\hat{B}_{r,\hat{A}} = (\hat{\beta}_{\hat{A},1}, \cdots, \hat{\beta}_{\hat{A},r}) = \hat{A}^{-1/2}(\hat{\varphi}_p^{\hat{A}}, \cdots, \hat{\varphi}_{p-r+1}^{\hat{A}}).$$

To study the properties of $\hat{B}_{r,A}$ and $\hat{B}_{r,\hat{A}}$, we make the following assumptions. 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 $\mathbf{Z} \in \mathbb{R}^p$, define $\|\mathbf{Z}\|_{\psi_2} = \sup_{x \in S^{p-1}} \|\langle \mathbf{Z}, x \rangle\|_{\psi_2}$, where S^{p-1} is the unit sphere in $\mathbb{R}^p$.

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

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

When $\mathbf{X}_{ij} = X_i - X_j$ where X_i's are $i.i.d.$ 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

$$\|\hat{G} - G_0\|_{op} = O_p\left(\delta_n^{op} + \sqrt{\frac{pm_{\text{net}}^2}{n}}\right),$$

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* $i = 1, \cdots, p$, *it holds that*

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

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 $\{\sigma(1), \cdots, \sigma(n)\}$, 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 $\boldsymbol{\sigma}(i)$ and $\boldsymbol{\sigma}(j)$ in G_s with $1 \leq s \leq \tilde{m}_{\text{net}} - 1$, we have $\{\boldsymbol{\sigma}(j), \boldsymbol{\sigma}(j)\} \in \mathbb{V}$. Combining the conditional independence property for the first $\tilde{m}_{\text{net}} - 1$ groups with Assumption (A3) below, we still have the bound in (5) but with m replaced by $\tilde{m}_{\text{net}}$.

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

It is easy to see that $d_{\max} \leq \max_{1\leq i\leq n} |N_i| \cdot \max_{1\leq i\leq n} |\{j : \xi_i \text{ is assoicated with node } j\}|$ 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 $\sqrt{pm_{\text{net}}^2/n}$ (or $\sqrt{p\tilde{m}_{\text{net}}^2/n}$) 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 $O_p(\sqrt{p/n}\log n)$. 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 $\tilde{m}_{\text{net}} = O((\log n)^2)$ and the convergence rate of $\hat{G}$ becomes $O_p(\sqrt{p/n}(\log n)^2)$. Following the proof of this theorem, it can be seen that if s_{ij}'s are independent, then $\|\hat{G} - G_0\|_{op} = O_p\left(\delta_n^{op} + \sqrt{p/n}\right)$. 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 $\{(\sigma(2i-1), \sigma(2i)), i = 1, \cdots, n/2\}$ can be split to groups such that s_{ij}'s with (i,j)'s in the same group are conditionally independent given $\{\mathbf{X}_{ij}\}$. 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 $0 < C^{-1} < \lambda_{\min}(A) \leq \lambda_{\max}(A) < C < \infty$ 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 for any given $1 \leq r \leq p$,

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

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 $1 \le r \le p$,

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

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

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 $\tilde{m}_{\text{net}}$.*

4. SIMULATION

To verify the effectiveness of NDR for reducing the dimension of the covariates, we conduct extensive simulation. The performance of our method is examined by computing the error measure defined as $\|P_\beta - P_{\hat{B}_{r,\hat{A}}}\|_F$, where β is the true parameter to be estimated, $\hat{B}_{r,\hat{A}}$ is the estimator of β using the method developed in this paper, and $P_B = B(B^TB)^{-1}B^T$ for any matrix B is the projection matrix onto the space spanned by the columns of B. Here for simplicity, we take $\hat{A}$ as the estimator of $A = \Sigma = \text{Cov}(X_i)$ that is computed by penalizing the Gaussian likelihood with a lasso penalty on the entries of the covariance matrix using the method in Bien & Tibshirani (2011). In all simulations, we take $s_{ij} = w_{ij}$ 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.

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 $P(C_i = 1) = P(C_i = 2) = 0.5$.
(ii) Generate covariates $X_i \sim N(0, \Sigma)$ where $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = 0.4^{|i-j|}I(\{|i - j| < 5\})$.
(iii) Given $(C_i, C_j, \mathbf{X}_{ij})$ with $\mathbf{X}_{ij} = X_i - X_j$, generate $w_{ij} \in \{0, 1\}$ according to the following model

$$P(w_{ij}|C_i, C_j, \mathbf{X}_{ij}) = P(w_{ij}|C_i, C_j)\frac{\exp(1 - c_{\text{coef}}|\beta^T\mathbf{X}_{ij}|)}{1 + \exp(1 - c_{\text{coef}}|\beta^T\mathbf{X}_{ij}|)}, \tag{7}$$

where $P(w_{ij}|C_i, C_j)$ is set as $P(w_{ij} = 1|C_i = C_j) = a$ and $P(w_{ij} = 1|C_i \ne C_j) = b$.

In model (7), 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 $\beta = (1, 1, 0, \cdots, 0)^T \in \mathbb{R}^p$, $a = 0.8$, $b = c_{\text{com}}a$ 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 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 simulation results are found in Figure 3.

Example 2. Consider an example where each node i is affected by its K neighbors and denote the set of their indices as N_i. The data is generated as follows.

(i) First, generate N_i for node i. Let $\mu_1, \cdots, \mu_n$ be $i.i.d.$ random variables from $U(0, 1)$, and let $d_{ij} = |\mu_i - \mu_j|, 1 \le i, j \le n$. For each node i, compute its K-nearest neighbors,

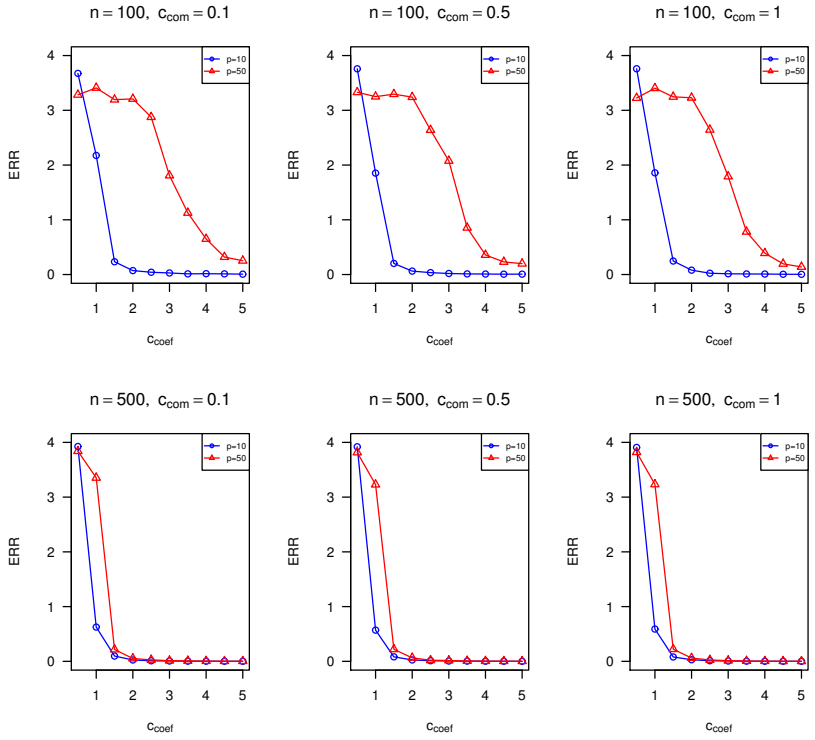


Fig. 3. Average errors for Example 1 for $p = 10$ and 50, respectively, where $r = 1$.

according to the distance d_{ij}. Define 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, $i \notin N_i$.

(*ii*) Let $Z_1, \cdots, Z_n$ be independent random variables generated as $Z_i \sim N(\mu_i, 0.1)$ and $Z_{ij} = Z_i - Z_j$. Generate X_i as in Example 1 and define $\mathbf{X}_{ij} = X_i - X_j$.

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

$$P(w_{ij} = 1 | \{Z_{ij}\}, \mathbf{X}_{ij}) = \exp\Big(-10\Big\{|Z_{ij}| \wedge \sum_{k \in N_i, k' \in N_j} |Z_{kk'}|/K^2\Big\}\Big) \frac{\exp(1 - c_{\text{coef}}\|\beta^T \mathbf{X}_{ij}\|_2)}{1 + \exp(1 - c_{\text{coef}}\|\beta^T \mathbf{X}_{ij}\|_2)},$$

where c_{coef} is specified as in Example 1, and $a \wedge b = \min\{a, b\}$.

We set $\beta = (v_1, v_2) \in \mathbb{R}^{p \times 2}$ where $v_1 = (1, 1, 0, \cdots, 0)^T \in \mathbb{R}^p$ and $v_2 = (1, -1, 0, \cdots, 0)^T \in \mathbb{R}^p$, and set $K = 2$ or $K = 4$. For this model, the probability of $w_{ij} = 1$ depends on latent variables in $\{Z_k : Z_k \in N_i \cup N_j\}$ 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 4.

We briefly discuss these simulation results. It is easy to see that the influence of the covariate $\mathbf{X}_{ij}$ decreases in both examples when c_{coef} decreases. Particularly, $\mathbf{X}_{ij}$ has no effect when $c_{\text{coef}} = 0$. We can see from Figure 3 and 4 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 3 that the errors are similar for different c_{com} in Example 1. This is due to the fact that the com-

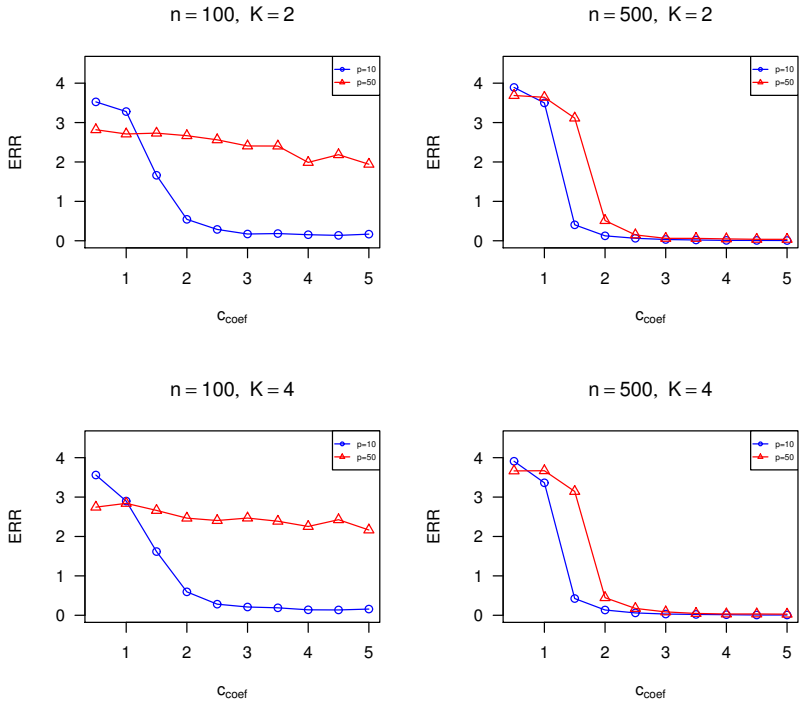


Fig. 4. Average errors for Example 2 for $p = 10$ and 50, respectively, where $r = 2$.

munity label C_i is independent of the covariate $\mathbf{X}_{ij}$ in the data generating process. Additional results on the computational time and the sparsity of the generated networks can be found in the Supplementary Materials, demonstrating that our procedure is computationally very fast.

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 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. 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.

Table 1. *The mean of clustering errors and sum of squares ratios over 100 replicates.*

	K-means	NDR			
Network density		70%	50%	30%	10%
Misclassification error	0.378	0.078	0.073	0.069	0.075
Sum of squares ratio	0.531	0.152	0.153	0.155	0.155

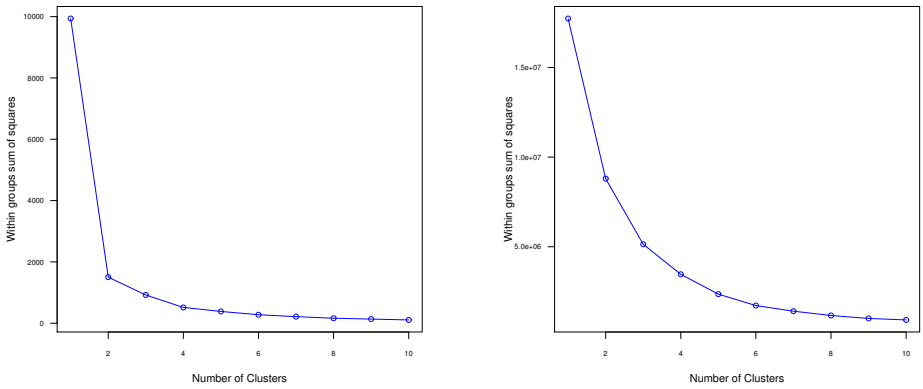
We randomly select 1000 observations from 16259 spurious examples and 600 observations from 1639 real pulsar examples to construct a graph. For these 1600 nodes, we say that two nodes are connected if their difference in the first variable (the mean of the integrated profile) is small. We choose a threshold such that the network density, defined as the ratio of edges over the maximum possible number of edges is $70\%, 50\%, 30\%$, or 10%. The rest of the eight variables are used as nodal covariates. Note that in defining the graph, we do not use the information on the labels of these observations. Since it is known that there are two clusters, we set $K = 2$. For the number of NDR directions, we set $r = 1$. This data generating process is repeated 100 times.

We apply the NDR-based community detection as described in Section 2.3 and compare it to the classical K-means clustering algorithm based only on covariates. Since the true community membership of each node is known, we report the average of the proportions of the nodes that are misclassified, and that of the ratios of the sum of squares within classes over the total sum of squares. It is obvious that the smaller these quantities are, the best an approach is. The results averaged over 100 random datasets are found in Table 1. It is clearly seen that NDR-based community detection approach outperforms the K-means clustering based approach by a large margin in terms of the two measures that have been examined.

To further illustrate the benefit of NDR, we present the scree plot of within sum of squares versus the number K of clusters in Figure 5, and the observations projected along the first direction of NDR in Figure 6 in one simulation when the network density is 10%. As shown in Figure 5, the within sum of squares decreases rapidly and becomes stable as soon as $K \geq 2$, indicating that a sensible choice of K is 2. On the other hand, it is not easy to determine the optimal cluster number for the K-means method, as the optimal number of clusters appear to be 4 or 5, which is larger than the true number of clusters. The left panel of Figure 6 uses different colors to represent the true cluster membership, while the right panel panel presents the estimated ones. It can be seen that only a small proportion of the observations are assigned with incorrect membership, implying that NDR is rather successful at reducing the dimensionality of the covariate for community detection.

6. CONCLUSION

We have proposed a novel approach for reducing the dimension of the covariate for explaining the strength of linkage in a network without assuming a model. The resulting estimator is necessarily dense nevertheless in that all of its entries are nonzero. There are important scenarios where sparse estimators are desired, for example, when the dimension of the covariate is large and not all the covariates are useful in explaining connections. To address this, we can impose constraints on the estimator, in a way similar to sparse PCA (Zou et al., 2006). Another problem we have not fully explored is the issue of choosing r, the dimension of the reduced space. One way to proceed is to eyeball the scree plot as we have demonstrated in the data analysis. The other approach is to examine the estimated eigenvalues as in Proposition 1. For the latter, we



(a) NDR (b) K-means

Fig. 5. The within sum of squares versus the number of clusters in one random sample.

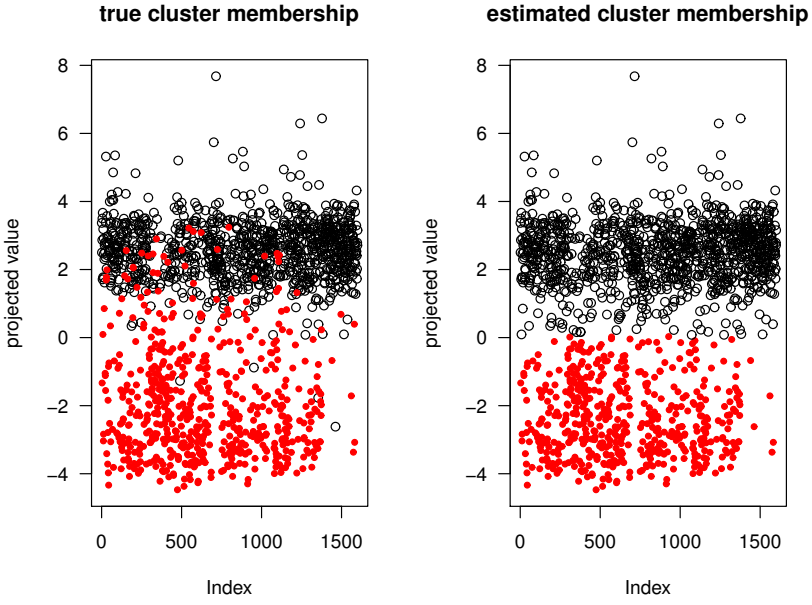


Fig. 6. The observations projected along the estimated NDR direction, where the x-axis is the indices. In the left panel the true cluster memberships are shown in different colors, while in the right panel, the estimated cluster memberships are shown in these colors.

may study the ratios of consecutive eigenvalues, in a way similar to the ratio estimator in Lam & Yao (2012). These problems are beyond the scope of this paper and will be explored elsewhere.

REFERENCES

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.

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., Raftery, 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. & Straten, W. v. (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.

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.

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, 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.

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 *October* 2019]