Statistica Sinica Preprint No: SS-2025-0362	
Title	Nonlinear Analysis of Nodal Covariates in Network: Dimension Reduction and Clustering
Manuscript ID	SS-2025-0362
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0362
Complete List of Authors	Zhonghan Wang, Shenbin Zheng and Junlong Zhao
Corresponding Authors	Junlong Zhao
E-mails	zhaojunlong928@126.com
Notice: Accepted author version.	

NONLINEAR ANALYSIS OF NODAL COVARIATES IN NETWORKS: DIMENSION REDUCTION AND CLUSTERING

Zhonghan Wang[1], Shengbin Zheng[1], Junlong Zhao[1,*]

[1]*School of Statistics, Beijing Normal University*

Abstract: Research on network data with nodal covariates has received increasing attention, yet few studies have focused on nonlinear patterns among nodal covariates. In this work, we propose a model-free framework that leverages network information to achieve nonlinear dimension reduction of nodal covariates. An efficient regularization-based estimation procedure is proposed and the asymptotic properties of estimated projection directions are studied. For the downstream task of community detection, we propose a two-step algorithm along with theoretical guarantees. Besides, we draw connections between our method and three existing kernel methods. Extensive simulations and a real data analysis support the advantages of the proposed method.

Key words and phrases: Nonlinear dimension reduction, Community detection, Degree-corrected stochastic block-model.

1. Introduction

Network data integrated with node covariates are pervasive across a wide range of fields and have garnered considerable research attention (Segal et al., 2003; Bansal et al., 2006;

*Corresponding author.

Hunter et al., 2008; Gao et al., 2023; Ogburn et al., 2024). Such integration facilitates network analyses by complementing relational ties among nodes with node-level attribute information. For example, identifying communities in a social network requires considering both interpersonal connections and individual covariates – such as occupation – to fully capture the similarity among people.

Incorporating nodal covariates via their linear combinations is widely adopted in network analyses, which can be viewed as implementing linear dimension reduction on the covariates (Sweet, 2015; Zhao et al., 2022; Xu and Wang, 2023; Huang et al., 2024). For example, Sweet (2015) assumed that the connection probability between two nodes is influenced by a linear combination of covariates. Zhao et al. (2022) performed community detection on the covariates after linear dimensionality reduction. A more comprehensive review of related works can be found in Section 1.1.

However, the aforementioned studies suffer from two major limitations. First, linear combinations of nodal covariates have limitations in capturing complex patterns of how nodal covariates influence edge formation. For example, the formation of friendship ties within social networks relates to an interaction effect between age and family income. Specifically, compared with teenagers, elementary school children are less sensitive to disparities in family income when forming friendships (Rhodes, 2018). Consequently, simply applying linear dimension reduction to node covariates can be misleading: it implicitly assumes that the effect of the family income disparity on the probability of friendship formation is invariant when the age gap is fixed, which is inconsistent with the observations.

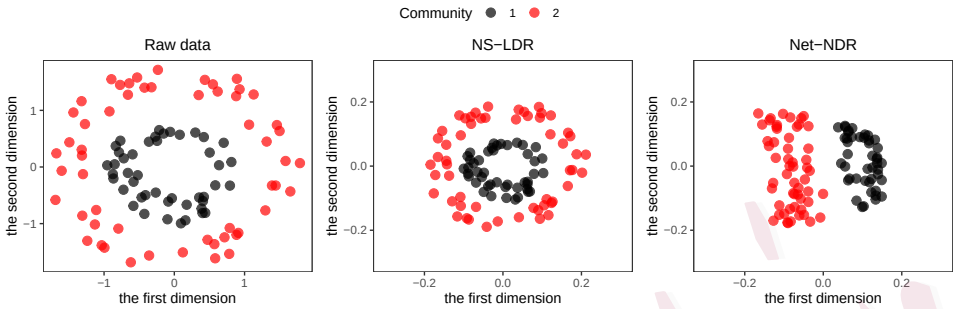

Figure 1: Scatter plots of raw data and projected data via NS-LDR in Zhao et al. (2022) and our method Net-NDR. The data is generated under the default setting of Case 1 with $n = 100$; see Subsection 5.1 for details.

Second, most existing works are based on specific models. For example, two recent studies on community detection (Huang et al., 2024; Fan et al., 2025) are both based on the stochastic block-model or its variants, suffering from the risk of model misspecification.

To address the first limitation, we consider nonlinear dimension reduction for nodal covariates in network data. Nonlinear dimension reduction is advantageous for uncovering complex relationships among variables, and has been proven effective in other tasks such as classification (Baudat and Anouar, 2000). To illustrate the superiority of nonlinear dimensionality reduction on nodal covariates, we simulate concentric ring data – a classic example of linear inseparability – and compare the performance of the linear dimension reduction method proposed by Zhao et al. (2022) against the nonlinear method proposed herein on this dataset. Notably, Figure 1 demonstrates that the linear method proposed in Zhao et al. (2022) fails to linearly separate covariates from two communities, while our nonlinear method delivers satisfactory separation results.

To solve the second issue of model misspecification, we formulate our model-free dimension reduction framework within homophily principle (McPherson et al., 2001), which has been widely adopted in network analyses (Elkabani and Khachfeh, 2015; Xu et al., 2023; Bomiriya et al., 2023; Jackson et al., 2023). This principle posits that the nodes with similar covariates are more likely to form edges. More formally, with the notations defined in Section 2.1, this principle means that for nodes i and j, the edge weight w_{ij} exhibits a negative dependence on the pairwise distance between nodal covariates X_i and X_j. Building upon it, we formulate the nonlinear dimension reduction process as a constrained optimization problem in a model-free manner to obtain a low-dimensional representation of the covariates, which can be used for downstream tasks such as community detection.

The main contributions of this paper are as follows. First, we propose a model-free network-supervised nonlinear dimension reduction method (termed Net-NDR) for nodal covariates, grounded in the well-known homophily principle in network analyses. Our second contribution is the development of a community detection algorithm with strong consistency. This algorithm does not rely on a specific model on the relationship between the network and community labels (e.g., the stochastic block-model), and thus alleviates the model misspecification problem. Moreover, we establish the connection between our method Net-NDR and kernel discriminant analysis – a method that assumes community labels are known. This connection implies that Net-NDR leverages label information effectively even when labels are unknown.

1.1 Related works

Linear combinations of covariates in network models. The integration of linear combinations of nodal covariates into network models has garnered increasing attention, as demonstrated by recent advances in the stochastic block-model and its variants (Sweet, 2015; Roy et al., 2019; Xu and Wang, 2023; Huang et al., 2024; Fan et al., 2025), the sparse β-model (Stein and Leng, 2023, 2025), and the latent space model (Ma et al., 2020).

Community detection with nodal covariates. Many methods leverage covariate information in community detection, including the methods based on the stochastic block-model and its variants mentioned above, as well as other approaches. For example, Binkiewicz et al. (2017) enhanced spectral clustering by adding a covariate kernel matrix to a Laplacian matrix. Zhao et al. (2022) performed community detection on the covariates after network-supervised linear dimension reduction. Xu et al. (2023) constructed an augmented adjacency tensor using a covariate kernel matrix. Hu and Wang (2024) applied spectral clustering on a network-adjusted covariate matrix.

Most of the studies mentioned above focused on linear patterns among nodal covariates, employing either linear combinations or a linear kernel. However, as discussed earlier, the linearity assumption is inefficient to capture the complexity of real-world scenarios. Note that Yan and Sarkar (2021) proposed an optimization framework on community detection, using information from both the adjacent matrix and a covariate Gaussian kernel matrix. However, their method fails when community-specific means of covariates are not well-separated linearly. Detailed discussions and comprehensive com-

parisons with Yan and Sarkar (2021) can be found in Section S3.1 of the Supplementary Materials.

1.2 Organization and notations

The rest of the paper is organized as follows. In Section 2, we first introduce some basic assumptions on the data, and then present our methods for both dimension reduction and community detection. In Section 3, asymptotic properties of estimated projection functions and strong consistency of our community detection algorithm are studied. In Section 4, we present the relationship between our method and some nonlinear dimension reduction methods. Simulations and the analysis of a real-world dataset are presented in Sections 5 and 6 respectively.

Notations. Let $\mathbb{I}(\cdot)$ denote the indicator function. For $x, y \in \mathbb{R}^p$, let $\|x\|$ denote the l_2-norm and $\langle x, y\rangle$ denote the inner product in $\mathbb{R}^p$. For $A \in \mathbb{R}^{p\times p}$, we denote its trace by $\mathrm{Tr}(A)$. We say that $x \in \mathbb{R}^p$ is the k-th eigenvector of A if x corresponds to the k-th largest eigenvalue of A. For a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with a Mercer kernel $\mathcal{K}(\cdot,\cdot)$, we let $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ denote the inner product and induced norm on $\mathcal{H}$ respectively. Let $\otimes$ denote the tensor product which satisfies $(u\otimes v)w = \langle v, w\rangle_{\mathcal{H}} u$ for any $u, v, w \in \mathcal{H}$. For an operator $T : \mathcal{H} \to \mathcal{H}$, we let $\|T\|_{\mathrm{HS}}$, $\|T\|_{\mathrm{OP}}$, $\mathrm{Tr}(T)$ and $\mathrm{span}(T)$ denote its Hilbert-Schmidt norm, operator norm, trace and range respectively. Let $\mathbf{0}$ denote the zero element in $\mathcal{H}$. For $h_1, \cdots, h_r \in \mathcal{H}$, we let $\boldsymbol{h}_r(\cdot) = (\langle h_1, \cdot\rangle_{\mathcal{H}}, \cdots, \langle h_r, \cdot\rangle_{\mathcal{H}})^\top$ denote a map from $\mathcal{H}$ to $\mathbb{R}^r$ and abbreviate it as $\boldsymbol{h}_r = (h_1, \cdots, h_r)$. Define $\mathrm{tr}(\boldsymbol{h}_r, T) =$

$\sum_{k=1}^{r}\langle h_k, Th_k\rangle_{\mathcal{H}}$.

2. Methodology

2.1 Data and assumptions

In this paper, we focus on an undirected network of n nodes, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \cdots, n\}$ is the node set and $\mathcal{E}$ is the edge set. The connectivity among n nodes can be represented mathematically by an $n \times n$ symmetric matrix with zero diagonal entries $W = (w_{ij})_{n\times n}$, known as the adjacency matrix. The element w_{ij} represents the edge weight or connection strength between nodes i and j. For instance, in an unweighted graph, $w_{ij} = 1$ if there is an edge between nodes i and j, and $w_{ij} = 0$ otherwise. Alongside the network, we observe nodal covariates, denoted by $X_1, \cdots, X_n \in \mathbb{R}^p$. The data assumptions are listed below.

Assumption 1. *We make the following assumptions:*

- $X_1, \cdots, X_n$ *are independent and identically distributed (i.i.d.) random vectors.*
- *The edge weights* w_{ij}*'s for* $i < j$ *are dependent. Except in Section 2.2, we allow* w_{ij}*'s* $(i < j)$ *to be differently distributed.*

Remark 1. In Section 2.2, w_{ij}'s $(i < j)$ are assumed to be identically distributed solely to simplify the presentation and facilitate interpretations. Specifically, for the infinite-dimensional operator $\hat{G}$ introduced in (2.3), this assumption allows us to represent $\lim_n E(\hat{G})$ – the limit of an average over $n(n-1)$ expectations – as a single one,

which simplifies the presentation of the population level objective function (2.1).

2.2 Network-supervised nonlinear dimension reduction under the homophily principle

In this subsection, we adopt the identical-distribution assumption for w_{ij}'s $(i < j)$, which is not a prerequisite for our method but solely serves to simplify the presentation, as argued in Remark 1. Denote the population counterpart of the data $\{(w_{ij}, X_i, X_j) : i < j\}$ by (w, X, X'), where w and w_{ij}'s $(i < j)$ are identically distributed, X_i's are identically distributed as X and X', and X' is independent of X. To realize nonlinear dimension reduction, we first map the covariates X and X' into a RKHS $\mathcal{H}$ by $\phi : \mathbb{R}^p \to \mathcal{H}$ to get $\phi(X)$ and $\phi(X')$. Then we seek r projection functions or directions in $\mathcal{H}$, denoted by $\boldsymbol{f}_r = (f_1, \cdots, f_r)$, to project $\phi(X)$ and $\phi(X')$ onto $\mathbb{R}^r$; $\boldsymbol{f}_r(\phi(X))$ and $\boldsymbol{f}_r(\phi(X'))$ are the projected features.

The idea of finding projection functions is grounded in the homophily principle. This principle states that two nodes with more similar values in covariates tend to form stronger connections, implying w is negatively associated with the distance between $\boldsymbol{f}_r(\phi(X))$ and $\boldsymbol{f}_r(\phi(X'))$, denoted as $d_{\boldsymbol{f}_r}(X, X')$. In this paper, we take $d_{\boldsymbol{f}_r}(X, X') = \|\boldsymbol{f}_r(\phi(X)) - \boldsymbol{f}_r(\phi(X'))\|^2$. Given this principle, an intuitive choice for $\boldsymbol{f}_r$ is the one that minimizes the nonnegative quantity $E[w \cdot d_{\boldsymbol{f}_r}(X, X')]$, subject to some constraints on $\boldsymbol{f}_r$. However, this formulation has a potential limitation. Note that $\phi(X)$ typically lies in an approximately low-dimensional subspace (denoted by $\mathcal{H}_\phi$). Then the $\boldsymbol{f}_r$ that minimizes $E[w \cdot d_{\boldsymbol{f}_r}(X, X')]$

likely lies in the orthogonal complement of $\mathcal{H}_\phi$, resulting in $\boldsymbol{f}_r(\phi(X)) \approx \mathbf{0}$ for any $X \in \mathbb{R}^p$. Consequently, the minimum value of $E[w \cdot d_{\boldsymbol{f}_r}(X, X')]$ is approximately zero, which is meaningless and trivial. To tackle this, we introduce a monotonic, one-to-one decreasing function of w, denoted $s := s(w)$, a transformation also employed in Zhao et al. (2022). Our objective then becomes to find a mapping $\boldsymbol{f}_r$ that maximizes $E[s \cdot d_{\boldsymbol{f}_r}(X, X')]$ under appropriate constraints, avoiding the trivial case. We discuss the selection of s in Remark 2 below.

Remark 2. Since Net-NDR is built upon the homophily principle, we do not assume a specific model between s and $(\phi(X), \phi(X'))$, and the selection of s is inherently flexible. Following Zhao et al. (2022), we use the linear function s due to its good interpretability. In practice, we recommend using $s_{ij} = 1 - 2w_{ij}$ and apply this throughout Sections 5 and 6. More discussions on s are given in Section S3.4 of the Supplementary Materials.

Let the mean function and covariance operator of $\phi(X)$ be

$$\mu^\phi = E[\phi(X)],\ \Sigma_\phi = E\left\{[\phi(X) - \mu^\phi] \otimes [\phi(X) - \mu^\phi]\right\}$$

respectively. Additionally, $\boldsymbol{f}_r$ is constrained in the following set to ensure the identifiability and prevent the maximum from diverging to infinity

$$\Theta_r = \{\boldsymbol{g}_r : \mathcal{H} \to \mathbb{R}^r \mid \boldsymbol{g}_r = (g_1, \cdots, g_r) \text{ for some } g_1, \cdots, g_r \in \mathcal{H}$$
$$\text{such that } \langle g_k, \Sigma_\phi g_l \rangle_{\mathcal{H}} = \mathbb{I}(k = l) \text{ for } k, l = 1, \cdots, r\},$$

where the condition $\langle g_k, \Sigma_\phi g_l \rangle_{\mathcal{H}} = \mathbb{I}(k = l)$ means that g_k and g_l are orthonormal after being adjusted by Σ_ϕ. With the above notations, we introduce our network-supervised

nonlinear dimension reduction method (briefly Net-NDR), of which the optimization problem at the population level can be formulated as

$$\boldsymbol{f}_r = \arg\max_{\boldsymbol{g}_r \in \Theta_r} E\left[s \cdot d_{\boldsymbol{g}_r}(X, X')\right] = \arg\max_{\boldsymbol{g}_r \in \Theta_r} \operatorname{tr}\left(\boldsymbol{g}_r, G_0\right), \tag{2.1}$$

where $G_0 = E\left\{s\left[\phi(X) - \phi(X')\right] \otimes \left[\phi(X) - \phi(X')\right]\right\}$. Throughout this paper, we assume $\operatorname{span}(G_0) \subseteq \operatorname{span}(\Sigma_\phi)$ to ensure that the operator $\Sigma_\phi^{-1} G_0$ is well-defined. The space $\operatorname{span}(\boldsymbol{f}_r)$ can be recovered by solving the generalized eigen-decomposition problems successively

$$f_k = \arg\max_{f \in \mathcal{C}_k(\Sigma_\phi)} \langle f, G_0 f \rangle_{\mathcal{H}}, \tag{2.2}$$

where $\mathcal{C}_k(\Sigma_\phi) = \{f \in \mathcal{H} : f \perp \mathcal{L}_{k-1}, \langle f, \Sigma_\phi f \rangle_{\mathcal{H}} = 1\}$ with $\mathcal{L}_0 = \mathbf{0}$ and $\mathcal{L}_{k-1} = \operatorname{span}(\Sigma_\phi f_1, \cdots, \Sigma_\phi f_{k-1})$ for $k = 2, \cdots, r$ (the notation $\perp$ denotes orthogonality). We now provide an interpretation of Equation (2.1) in Proposition 1.

Proposition 1. *The function $\boldsymbol{f}_r$ in (2.1) satisfies that*

$$\boldsymbol{f}_r = \arg\max_{\boldsymbol{g}_r \in \Theta_r} \sum_{k=1}^{r} \operatorname{cov}(s, \langle g_k, \phi(X) - \phi(X') \rangle_{\mathcal{H}}^2) \quad \textit{and}$$

$$\operatorname{tr}(\boldsymbol{f}_r, G_0) = \max_{\boldsymbol{g}_r \in \Theta_r} \sum_{k=1}^{r} \operatorname{cov}(s, \langle g_k, \phi(X) - \phi(X') \rangle_{\mathcal{H}}^2) + 2rE(s).$$

Proposition 1 shows that $\boldsymbol{f}_r$ maximizes the sum of covariances between s and $\langle g_k, \phi(X) - \phi(X') \rangle_{\mathcal{H}}^2$ for $k = 1, \cdots, r$, which confirms that the projection directions we found maximize homophily. Besides, the term $\operatorname{tr}(\boldsymbol{f}_r, G_0)$ is exactly the maximal sum plus a constant. When ϕ is a nonlinear map, a larger value of $\operatorname{tr}(\boldsymbol{f}_r, G_0)$ implies the stronger nonlinear dependence between the edge weight and the pairwise distance of the covariates.

In fact, $\boldsymbol{f}_r$ can also be explained from the perspective of subspace recovery. Suppose that s depends on (X, X') through the function $h(\boldsymbol{f}_{r,0}(\phi(X)) - \boldsymbol{f}_{r,0}(\phi(X')))$, for some $\boldsymbol{f}_{r,0} \in \Theta_r$ and an unspecified $h(\cdot)$. Then it can be shown that $\text{span}(\boldsymbol{f}_r) = \text{span}(\boldsymbol{f}_{r,0})$ under some conditions. Consequently, the invariance of this subspace to the specific form of h validates the model-free nature of Net-NDR. We leave this discussion in Section S3.3 of the Supplementary Materials.

Last, it is worth noting that the projected covariates $\boldsymbol{f}_r(\phi(X))$ can be integrated into a wide range of existing models as nodal covariates, such as Ma et al. (2020), Huang et al. (2024), and Stein and Leng (2025). Moreover, $\boldsymbol{f}_r(\phi(X))$ is applicable to many downstream tasks such as link prediction and anomaly detection. We provide a discussion on this point in Section S3.2 of the Supplementary Materials.

2.3 Estimation of network-supervised projection directions

We now give an estimator of $\boldsymbol{f}_r$; the selection of r is discussed later in Remark 4. With the notation $s_{ij} := s(w_{ij})$, G_0 in (2.1) can be estimated by

$$\hat{G} = [n(n-1)]^{-1} \sum_{i \neq j} s_{ij} \left[\phi(X_i) - \phi(X_j)\right] \otimes \left[\phi(X_i) - \phi(X_j)\right]. \tag{2.3}$$

In addition, Σ_ϕ can be estimated by $\hat{\Sigma}_\phi = n^{-1} \sum_{i=1}^n [\phi(X_i) - \hat{\mu}^\phi] \otimes [\phi(X_i) - \hat{\mu}^\phi]$, where $\hat{\mu}^\phi = n^{-1} \sum_{i=1}^n \phi(X_i)$. Consequently, Θ_r can be estimated by

$$\hat{\Theta}_r = \{\boldsymbol{g}_r : \mathcal{H} \to \mathbb{R}^r \mid \boldsymbol{g}_r = (g_1, \cdots, g_r) \text{ for some } g_1, \cdots, g_r \in \mathcal{H}$$
$$\text{such that } \langle g_k, \hat{\Sigma}_\phi g_l \rangle_{\mathcal{H}} = \mathbb{I}(k = l) \text{ for } k, l = 1, \cdots, r\}.$$

Then the sample-version of the optimization problem (2.1) is formulated as

$$\hat{\boldsymbol{f}}_r = (\hat{f}_1, \cdots, \hat{f}_r) = \arg\max_{\boldsymbol{g}_r \in \hat{\Theta}_r} \mathrm{tr}(\boldsymbol{g}_r, \hat{G}), \tag{2.4}$$

and the space $\mathrm{span}(\hat{\boldsymbol{f}}_r)$ can be recovered by solving the following generalized eigen-decomposition problems successively

$$\hat{f}_k = \arg\max_{f \in \mathcal{C}_k(\hat{\Sigma}_\phi)} \langle f, \hat{G} f \rangle_{\mathcal{H}}, \tag{2.5}$$

where $\mathcal{C}_k(\hat{\Sigma}_\phi) = \{f \in \mathcal{H} : f \perp \hat{\mathcal{L}}_{k-1}, \langle f, \hat{\Sigma}_\phi f \rangle_{\mathcal{H}} = 1\}$ with $\hat{\mathcal{L}}_0 = \mathbf{0}$ and $\hat{\mathcal{L}}_{k-1} = \mathrm{span}(\hat{\Sigma}_\phi \hat{f}_1, \cdots, \hat{\Sigma}_\phi \hat{f}_{k-1})$ $(k = 2, \cdots, r)$.

By the well-known kernel trick (Berlinet and Thomas-Agnan, 2011), $\hat{f}_k$'s can be written as linear combinations of $\phi(X_i) - \hat{\mu}^\phi$ $(i = 1, \cdots, n)$, that is,

$$\hat{f}_k = \sum_{i=1}^{n} \hat{\beta}_{k,i} (\phi(X_i) - \hat{\mu}^\phi)$$

for some constants $\hat{\beta}_{k,1}, \cdots, \hat{\beta}_{k,n}$ to be estimated. Then the projection of any $X \in \mathbb{R}^p$ in the k-th direction can be calculated by

$$\langle \hat{f}_k, \phi(X) \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \hat{\beta}_{k,i} \left[\mathcal{K}(X_i, X) - n^{-1} \sum_{j=1}^{n} \mathcal{K}(X_j, X) \right],$$

where $\mathcal{K}$ is the kernel associated with the reproducing kernel Hilbert space $\mathcal{H}$. To obtain $\hat{f}_k$'s, it is sufficient to compute $\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_{k,1}, \cdots, \hat{\beta}_{k,n})^\top$. Let I_n be the $n \times n$ identity matrix. Let J_n be an $n \times n$ matrix with all entries being one. Given the kernel matrix $K = (\mathcal{K}(X_i, X_j))_{n \times n}$, we define

$$\bar{K} = (I_n - J_n/n) K (I_n - J_n/n).$$

Let $S = (s_{ij})_{n\times n}$ and define D_S as an $n \times n$ diagonal matrix where the (i,i)-th entry is equal to $\sum_{j=1}^{n} s_{ij}$. Denote $L_S = D_S - S \in \mathbb{R}^{n\times n}$. The following proposition shows that for $k = 1, \cdots, r$, $\hat{\boldsymbol{\beta}}_k$ can be calculated by solving a spectral decomposition problem.

Proposition 2. *For $k = 1, \cdots, r$, $\hat{\boldsymbol{\beta}}_k$ satisfies the equation $\bar{K}\hat{\boldsymbol{\beta}}_k = n^{1/2}\nu_k$, where ν_k is the k-th eigenvector of L_S.*

To avoid overfitting when estimating $\hat{\boldsymbol{\beta}}_k$'s, similar to Ying and Yu (2022) and Zhang et al. (2024), we let $\tilde{\boldsymbol{\beta}}_k = (\tilde{\beta}_{k,1}, \cdots, \tilde{\beta}_{k,n})^\top = n^{1/2}(\bar{K} + \epsilon_n I_n)^{-1}\tilde{\nu}_k$ be an approximation to $\hat{\boldsymbol{\beta}}_k$ for $k = 1, \cdots, r$, where $\tilde{\nu}_k$ is the k-th eigenvector of

$$K_L = (\bar{K} + \epsilon_n I_n)^{-1}\bar{K}L_S\bar{K}(\bar{K} + \epsilon_n I_n)^{-1}$$

with a tuning parameter $\epsilon_n > 0$. Let $\tilde{f}_k = \sum_{i=1}^{n} \tilde{\beta}_{k,i}(\phi(X_i) - \hat{\mu}^\phi)$ with $\tilde{\boldsymbol{f}}_r = (\tilde{f}_1, \cdots, \tilde{f}_r)$. Then the projected covariates can be calculated by

$$(\tilde{\boldsymbol{f}}_r(\phi(X_1)), \cdots, \tilde{\boldsymbol{f}}_r(\phi(X_n)))^\top = K(I_n - J_n/n)(\tilde{\boldsymbol{\beta}}_1, \cdots, \tilde{\boldsymbol{\beta}}_r) \in \mathbb{R}^{n\times r}. \tag{2.6}$$

Remark 3. In Section 3, we will focus on the asymptotic properties of $\hat{f}_{\epsilon_n,k}$'s defined through the following successive optimization problems (Ying and Yu, 2022)

$$\hat{f}_{\epsilon_n,k} = \arg \max_{f \in \mathcal{C}_k(\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}})} \langle f, \hat{G}f\rangle_{\mathcal{H}}, \tag{2.7}$$

where $I_{\mathcal{H}}$ denotes the identity mapping in $\mathcal{H}$ which maps any $v \in \mathcal{H}$ to itself, $\mathcal{C}_k(\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}}) = \{f \in \mathcal{H} : f \perp \hat{\mathcal{L}}_{\epsilon_n,k-1}, \langle f, (\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}})f\rangle_{\mathcal{H}} = 1\}$ with $\hat{\mathcal{L}}_{\epsilon_n,0} = \mathbf{0}$ and $\hat{\mathcal{L}}_{\epsilon_n,k-1} = \text{span}((\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}})\hat{f}_{\epsilon_n,1}, \cdots, (\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}})\hat{f}_{\epsilon_n,k-1})$ $(k = 2, \cdots, r)$.

Remark 4 (Selection of ϵ_n and r)**.** The value of ϵ_n is selected via grid search to maximize the Calinski-Harabasz (CH) index (Caliński and Harabasz, 1974; Xu and Wunsch II, 2008) that measures the strength of the homophily structure. A larger CH index implies that the strongly connected nodes are embedded closer together, while the weakly connected ones are farther apa1rt. Let $\hat{\lambda}_1(\epsilon_n) \geq \hat{\lambda}_2(\epsilon_n) \geq \cdots \geq \hat{\lambda}_n(\epsilon_n)$ denote the eigenvalues of K_L which are associated with the value of ϵ_n. Specifically, for each candidate value of ϵ_n, we calculate $\hat{r}(\epsilon_n) = \arg\max_{i=1,\cdots,M}[\hat{\lambda}_i(\epsilon_n) - \hat{\lambda}_{i+1}(\epsilon_n)]/[\hat{\lambda}_i(\epsilon_n) + \hat{\lambda}_{i+1}(\epsilon_n)]$ with a pre-specified constant M, referring to Lam and Yao (2012). We select the values of ϵ_n and $\hat{r}(\epsilon_n)$ associated with the largest CH index.

Remark 5 (Computational cost)**.** According to Remark 4, only the first $M+1$ dominant eigenvalues and eigenvectors of the $n \times n$ matrix K_L need to be computed, which can be done efficiently via fast algorithms such as the function `eigs_sym` in R package `RSpectra` (Qiu, 2024). Simulations on the computation cost and memory usage are provided in Section S2.2 of the Supplementary Materials.

2.4 Application to community detection

The method Net-NDR developed in previous sections lays a groundwork for downstream tasks, in particular community detection. Denote the community label of node i as $C_i \in \{1, \cdots, N_c\}$ for $i = 1, \cdots, n$, where N_c is the number of communities. Assume that C_i's are *i.i.d.* random variables with $P(C_i = t) = \pi_t$ $(t = 1, \cdots, N_c)$ satisfying $\sum_{t=1}^{N_c} \pi_t = 1$. We consider a common setting in the literature where C_i affects both the

covariates X_i and the edge weight w_{ij} (Binkiewicz et al., 2017; Yan and Sarkar, 2021; Hu and Wang, 2024). However, different from existing works that usually assumed the stochastic block-model or its variants, we do not impose a specific model on how C_i's affects w_{ij}'s, thereby avoiding model misspecification. Specifically, given C_i, we assume that $\phi(X_i)$ is generated from the following nonlinear mixture model

$$\phi(X_i) = \mu^{\phi}_{C_i} + \varsigma_i, \tag{2.8}$$

where $\mu^{\phi}_{C_i}$ is the community-specific mean function in $\mathcal{H}$ and ς_i's (independent of C_i's) are $i.i.d.$ zero-mean random elements in $\mathcal{H}$ with a covariance operator Σ_ς. We propose a two-step community detection algorithm as described below.

Algorithm 1: A two-step community detection algorithm based on Net-NDR

Input: $\{(s_{ij}, X_i, X_j) : i \neq j\}$ with $s_{ij} = 1 - 2w_{ij}$ (recommended) and the pre-specified number of communities $\hat{N}_c$.

Output: The estimated labels $\hat{C}_i$ for $i = 1, \cdots, n$.

Step 1: Calculate the projected covariates $\tilde{\boldsymbol{f}}_r(\phi(X_i))$ for $i = 1, \cdots, n$, according to (2.6).

Step 2: Perform K-means on $\tilde{\boldsymbol{f}}_r(\phi(X_i))$'s to identify $\hat{N}_c$ communities, and obtain $\hat{C}_i$ for $i = 1, \cdots, n$.

3. Asymptotic analyses

In this section, we study the asymptotic properties of the sample projection functions and establish strong consistency of the community detection algorithm. Note that we allow

s_{ij}'s $(i < j)$ to have different distributions and to be dependent.

3.1 Consistency of sample projection functions

The population counterpart of $\hat{G}$ in (2.3) is now defined as the limit of $E(\hat{G})$ – denoted by $\lim_n E(\hat{G})$ – which, provided this limit exists, is a q-dimensional operator with q being the dimension of $\mathcal{H}$. Obviously, $\lim_n E(\hat{G})$ generalizes the G_0 defined in (2.1), without requiring w_{ij}'s to be identically distributed. For simplicity of notation, we still denote this limit as G_0. Furthermore, in this subsection we use the shorthand $\phi_{ij} := \phi(X_i) - \phi(X_j)$. A major challenge in theoretical analyses arises from the dependence among s_{ij}'s. To address this problem, we adopt the following conditional independence property (Zhao et al., 2022).

Definition 1 (Conditional independence property)**.** For any permutation of $\{1, \cdots, n\}$, denoted by $\{\sigma(1), \cdots, \sigma(n)\}$, the node pairs $\{\tilde{\sigma}(i) = (\sigma(2i-1), \sigma(2i)), i = 1, \cdots, n/2\}$ can be divided into groups such that s_{ij}'s with (i, j)'s in the same group are independent given $\{\phi_{ij}, i \neq j\}$.

This property allows s_{ij}'s with (i, j)'s in different groups to be dependent. For any permutation σ, the minimum number of groups satisfying this property is denoted as m_σ. Let $m_{\text{net}} = \max_\sigma m_\sigma$, which can be seen as the network effect. Loosely speaking, a larger value of m_{net} corresponds to stronger dependence among edges, making it harder for the estimated projection functions to converge to the true functions. The following generalized graphon model (Zhao et al., 2022) is a typical example satisfying the conditional

independence property, which shares some similarities with Davezies et al. (2021) and Menzel (2021).

Example 1 (Generalized graphon model)**.** Let ξ_i $(i = 1, \cdots, n)$ be *i.i.d.* latent random variables. The dependence structure is introduced by assigning each node with a subvector of $\Xi = (\xi_1, \cdots, \xi_n)^\top$, denoted by $\Xi_{\mathcal{N}_i}$ with $\mathcal{N}_i \subset \{1, \cdots, n\}$. Let $\zeta_1, \cdots, \zeta_n$ be *i.i.d.* variables denoting the node-specific effects. Assume that w_{ij} is generated from a Bernoulli distribution $\text{Ber}(\theta_{ij})$ with $\theta_{ij} = h_{ij}(\Xi_{\mathcal{N}_i}, \Xi_{\mathcal{N}_j}, \zeta_i, \zeta_j, \phi_{ij})$. Let $\mathcal{N}_{ij} = \mathcal{N}_i \cup \mathcal{N}_j$ and $\mathbb{V} = \{\{(i,j),(k,t)\} : \mathcal{N}_{ij} \cap \mathcal{N}_{kt} = \varnothing, i \neq j \neq k \neq t\}$. It is easy to see that for any $\{(i,j),(k,t)\} \in \mathbb{V}$, the two node pairs (i,j) and (k,t) do not share common latent random variables, implying that the corresponding edges are independent given the nodal covariates.

To establish the convergence of $\|\hat{G} - G_0\|_{\text{HS}}$, we introduce the following assumptions.

Assumption 2. *For any* $i \neq j$, $E\|s_{ij}^{1/2}\phi_{ij}\|_{\mathcal{H}}^4 < \infty$.

Assumption 3. *For any* $i \neq j$, s_{ij} *is independent of* $\{\phi_{kt} : (k,t) \neq (i,j)\}$ *conditioning on* ϕ_{ij}.

Assumption 2 is a regularity condition and one sufficient condition of it is that both $\max_{i<j} s_{ij}$ and $E\{[\mathcal{K}(X,X)]^4\}$ are bounded. Assumption 3 means that s_{ij} depends on $\{\phi_{kt}, k < t\}$ only through ϕ_{ij}.

Theorem 1. *Under the conditional independence property and Assumptions 1-3, it holds that* $\|\hat{G} - G_0\|_{\text{HS}} = O_p(m_{\text{net}}/\sqrt{n} + e_n)$, *where* $e_n = \|E(\hat{G}) - G_0\|_{\text{HS}}$.

Theorem 1 shows that the convergence rate of $\|\hat{G} - G_0\|_{\text{HS}}$ consists of two parts. The first term $m_{\text{net}}/\sqrt{n}$ comes from the estimation error $\|\hat{G} - E(\hat{G})\|_{\text{HS}}$ with m_{net} representing the network effect. The stronger dependence among w_{ij}'s leads to a larger m_{net} and, consequently, a slower convergence rate. The second term e_n is the approximation error, which becomes zero when w_{ij}'s $(i < j)$ are identically distributed.

Next, we establish the convergence rate of sample projection functions. Recall that the population projection functions f_k's are calculated via (2.2) and the sample projection functions $\hat{f}_{\epsilon_n,k}$'s are calculated approximately via (2.7). The following assumptions are required.

Assumption 4. *The covariates satisfy that* $E\{[\mathcal{K}(X,X)]^2\} < \infty$.

Assumption 5. *The operator* G_0 *satisfies that* $\text{span}(G_0) \subseteq \text{span}(\Sigma_\phi^2)$.

Assumption 6. $\Sigma_\phi^{-1} G_0$ *is a Hilbert-Schmidt operator, that is,* $\|\Sigma_\phi^{-1} G_0\|_{\text{HS}} < \infty$.

Assumption 4 is a regularity condition for reproducing kernel Hilbert spaces. Assumption 5 ensures there exists some bounded operator $\mathcal{Z}$ such that $G_0 = \Sigma_\phi^2 \mathcal{Z}$.

Theorem 2. *Under the conditional independence property and Assumptions 1-6, if* ϵ_n *tends to zero, then for* $k = 1, \cdots, r$, *it holds that*

$$\|(\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}})^{-1} \hat{G} - \Sigma_\phi^{-1} G_0\|_{\text{HS}} = O_p\left(\epsilon_n + \frac{m_{\text{net}}}{\epsilon_n n^{1/2}} + \frac{e_n}{\epsilon_n}\right) \tag{3.9}$$

and

$$\|\hat{f}_{\epsilon_n,k} - c_0 f_k\|_{\mathcal{H}} = O_p\left({\epsilon_n}^{1/2} + \frac{1}{n^{1/2}\epsilon_n^{3/2}} + \frac{m_{\text{net}}}{\epsilon_n n^{1/2}} + \frac{e_n}{\epsilon_n}\right), \tag{3.10}$$

where $c_0 \in \{-1, 1\}$ *such that* $c_0\langle \hat{f}_{\epsilon_n,k}, f_k\rangle_{\mathcal{H}} > 0$.

For the convergence rate of the operator, the first term ϵ_n of (3.9) comes from the error $\|(\Sigma_\phi + \epsilon_n I_{\mathcal{H}})^{-1}G_0 - \Sigma_\phi^{-1}G_0\|_{\text{HS}}$ and the remaining terms come from $\|(\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}})^{-1}(\hat{G} - G_0)\|_{\text{HS}}$. To facilitate discussions on the convergence rate of the functions in (3.10), we introduce two notations. Denote the unit eigenfunctions of $\Sigma_\phi^{-1}G_0$ and $(\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}})^{-1}\hat{G}$ by ψ_k's and $\hat{\psi}_k$'s respectively. Then it is easy to see that $f_k = \psi_k/\|\Sigma_\phi^{1/2}\psi_k\|_{\mathcal{H}}$ and $\hat{f}_{\epsilon_n,k} = \hat{\psi}_k/\|(\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}})^{1/2}\hat{\psi}_k\|_{\mathcal{H}}$ for $k = 1, \cdots, r$. The first two terms of (3.10) arise from $\|(\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}})^{1/2}\hat{\psi}_k\|_{\mathcal{H}} - \|\Sigma_\phi^{1/2}\psi_k\|_{\mathcal{H}}$ and the last two originate from $\|\hat{\psi}_k - \psi_k\|_{\mathcal{H}}$.

Under the generalized graphon model, Zhao et al. (2022) proved that for any permutation $\{\sigma(1), \cdots, \sigma(n)\}$, the index pairs $\{\tilde{\sigma}(i) = (\sigma(2i-1), \sigma(2i)), i = 1, \cdots, n/2\}$ can always be split into $\tilde{m}_\sigma$ groups such that the conditional independence property holds for the first $\tilde{m}_\sigma - 1$ groups, where $\tilde{m}_\sigma \le \tilde{m}_{\text{net}} = \log(n/4)/\log\{4d_{\max}/(4d_{\max}-1)\} + 1$ and

$$d_{\max} = \max_{i=1,\ldots,n} \text{Card}(\{j : \mathcal{N}_j \cap \mathcal{N}_i \neq \varnothing\})$$

with Card$(\cdot)$ denoting the cardinality of a set, where $\mathcal{N}_i$'s are defined in Example 1. We show that if the additional assumption below holds, the conclusions of Theorems 1 and 2 still hold with m_{net} replaced by $\tilde{m}_{\text{net}}$.

Assumption 7. *The dependence structure among* w_{ij}*'s* $(i < j)$ *satisfies that* $d_{\max} < n^{1/2}$.

Theorem 3. *Assume that Assumption 7 and the conditions of Theorems 1 and 2 hold. Under the generalized graphon model, the conclusions of Theorems 1 and 2 hold with* m_{net} *replaced by* $\tilde{m}_{\text{net}}$.

To facilitate the discussion on the convergence rate under the generalized graphon model, we assume that the approximation error is zero ($e_n = 0$) and present the following corollary.

Corollary 1. *Assume that the conditions of Theorem 3 hold and* $e_n = 0$. *If* $d_{\max} = O(1)$, *then* $\tilde{m}_{\text{net}} = O(\log n)$ *and* $\|(\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}})^{-1}\hat{G} - \Sigma_\phi^{-1} G_0\|_{\text{HS}} = O_p(\epsilon_n + \epsilon_n^{-1} n^{-1/2} \log n)$. *If* $d_{\max} = O(\log n)$, *then* $\tilde{m}_{\text{net}} = O(\log^2 n)$ *and* $\|(\hat{\Sigma}_\phi + \epsilon_n I_{\mathcal{H}})^{-1}\hat{G} - \Sigma_\phi^{-1} G_0\|_{\text{HS}} = O_p(\epsilon_n + \epsilon_n^{-1} n^{-1/2} \log^2 n)$.

When $d_{\max} = O(1)$, the optimal convergence rate of the operator is $O_p(n^{-1/4}\log^{1/2} n)$ when $\epsilon_n = O(n^{-1/4}\log^{1/2} n)$. When $d_{\max} = O(\log n)$, the optimal convergence rate of the operator is $O_p(n^{-1/4} \log n)$ when $\epsilon_n = O(n^{-1/4} \log n)$. The rates here, up to a factor $\log^{1/2} n$ or $\log n$, are the same as the rates obtained in some studies on nonlinear sufficient dimension reduction (Li and Song, 2017; Virta et al., 2022).

3.2 Consistency of community detection

In this subsection, we demonstrate the strong consistency of the community detection algorithm given the true number of communities, that is, the label of each node can be exactly recovered with high probability. Recall that the covariates are assumed to follow the model $\phi(X_i) = \mu_{C_i}^{\phi} + \varsigma_i$ $(i = 1, \cdots, n)$. Denote $\hat{\boldsymbol{f}}_{r,\epsilon_n} = (\hat{f}_{\epsilon_n,1}, \cdots, \hat{f}_{\epsilon_n,r})$ where $\hat{f}_{\epsilon_n,k}$'s are defined in (2.7). The following lemma gives an upper bound on the distance between the projection $\hat{\boldsymbol{f}}_{r,\epsilon_n}(\phi(X_i))$ and its oracle center $E[\boldsymbol{f}_r(\phi(X_i))] = \boldsymbol{f}_r(\mu_{C_i}^{\phi})$ for $i = 1, \cdots, n$.

Lemma 1. *Suppose that* $\max_{i=1,\cdots,n}\mathcal{K}(X_i,X_i)=O_p(\delta_\phi^2)$ *and* $\max_{t=1,\cdots,N_c}\|\mu_t^\phi\|_{\mathcal{H}}=O(\delta_\mu)$. *Under assumptions in Theorem 2, it holds for each* $i=1,\cdots,n$ *that*

$$\|\hat{\boldsymbol{f}}_{r,\epsilon_n}(\phi(X_i))-\boldsymbol{f}_r(\mu_{C_i}^\phi)\|=O_p\left(r^{1/2}\delta_\mu+r^{1/2}\delta_\phi\left[1+\frac{m_{\text{net}}}{\epsilon_n n^{1/2}}+\frac{e_n}{\epsilon_n}+\frac{1}{n^{1/2}\epsilon_n^{3/2}}\right]\right).$$

Denote the rate obtained in Lemma 1 by R_{ϵ_n}. Define $L_{\text{dev}}=\sum_{i=1}^n\|\hat{\boldsymbol{f}}_{r,\epsilon_n}(\phi(X_i))-\boldsymbol{f}_r(\mu_{C_i}^\phi)\|^2$. Let n_t denote the number of nodes in community t for $t=1,\cdots,N_c$ satisfying $\sum_{t=1}^{N_c}n_t=n$. To prove strong consistency, the following assumptions are needed.

Assumption 8. $\min_{t=1,\cdots,N_c}n_k/n\geq c_b$ *for some constant* $c_b>0$.

Assumption 9. *For any* $t_1\neq t_2\in\{1,\cdots,N_c\}$, $\|\boldsymbol{f}_r(\mu_{t_1}^\phi)-\boldsymbol{f}_r(\mu_{t_2}^\phi)\|\geq\max\{4(L_{\text{dev}}/c_bn)^{1/2},D_n^{1/2}R_{\epsilon_n}(4+4/c_b^{1/2})\}$ *where* $D_n=O(1)$.

Assumption 8 requires that the sizes of communities are comparable, which is commonly used in existing literature (Yan and Sarkar, 2021; Hu and Wang, 2024; Huang et al., 2024). Assumption 9 gives a lower bound of the minimum distance between pairwise true community centers. When D_n is lower bounded from 0, Assumption 9 implies that $\|\boldsymbol{f}_r(\mu_{t_1}^\phi)-\boldsymbol{f}_r(\mu_{t_2}^\phi)\|\geq c_1R_{\epsilon_n}$ with some constant $c_1>0$. Combined with Lemma 1, it tells us that the order of magnitude of $\|\boldsymbol{f}_r(\mu_{t_1}^\phi)-\boldsymbol{f}_r(\mu_{t_2}^\phi)\|$ is larger than that of $\|\hat{\boldsymbol{f}}_{r,\epsilon_n}(\phi(X_i))-\boldsymbol{f}_r(\mu_{C_i}^\phi)\|$ for each $i=1,\cdots,n$. In other words, the distance between pairwise community centers dominates the random errors. The following theorem shows that the community labels can be exactly recovered with probability approaching to one.

Theorem 4 (Strong consistency of the community detection algorithm)**.** *Under Assump-*

tions the1-6 and 8-9, there exists a permutation ρ such that as n tends to infinity,

$$P\left(\rho(\hat{C}_i) = C_i, i = 1, \cdots, n\right) \to 1.$$

Three theoretical advantages of our community detection method are listed as follows. First, different from existing works that relied on specific network models to characterize how C_i's influence w_{ij}'s (e.g., the stochastic block-model and its variants in Binkiewicz et al. (2017), Yan and Sarkar (2021) and Hu and Wang (2024)), we make no such restrictive assumptions. Second, some prior studies assumed covariates follow a mixture distribution with strictly linearly separable community-specific means (e.g., the three literature mentioned above), while we adopt a more general framework: covariates after a nonlinear transformation follow a mixture model. This allows us to handle linearly inseparable cases; see simulations in Section 5.1. Third, we avoid other stringent assumptions required by existing works, such as bounded covariates (Binkiewicz et al., 2017; Hu and Wang, 2024) and independence among the w_{ij}'s.

4. Connection with other methods

4.1 Relation to spectral clustering and KSPCA

Relation to spectral clustering. Recall the definition of L_S in Section 2.3. Consider $s_{ij} = -\alpha_1 w_{ij}$ with $\alpha_1 > 0$. Then $-{\alpha_1}^{-1} L_S$ is exactly the Laplacian matrix (Merris, 1994), leading to a connection between Net-NDR and spectral clustering (Luxburg, 2007). Specifically, the scaled projected covariate vector ν_k, which is defined in Proposition 2 as

the k-th eigenvector of L_S, is proportional to the k-th eigenvector of the Laplacian matrix used for K-means in spectral clustering.

Relation to KSPCA. Kernel supervised principal component analysis (KSPCA) is a nonlinear extension of supervised linear principal component analysis (Barshan et al., 2011), which aims at finding projection directions to maximize the dependence between projected covariates and the response variables. Recall that $\bar{K} = (I_n - J_n/n)K(I_n - J_n/n)$. The optimization problem of KSPCA can be formulated as finding n projected covariates in $\mathbb{R}^r$:

$$V_{\text{KSPCA}} = \arg \max_{V \in \mathbb{R}^{n \times r}: V^\top \bar{K} V = I_r} \text{Tr}(V^\top \bar{K} K_{\text{Y}} \bar{K} V),$$

where $K_{\text{Y}} \in \mathbb{R}^{n \times n}$ is a kernel matrix associated with the responses. We now turn to Net-NDR. To establish the relation to KSPCA, we consider orthonormal projection functions, that is, $\langle \hat{f}_k, \hat{f}_l \rangle_{\mathcal{H}} = \mathbb{I}(k = l)$ for all $1 \leq k, l \leq r$. Then the optimization problem of Net-NDR can be formulated as

$$V_{\text{Net}} = \arg \max_{V \in \mathbb{R}^{n \times r}: V^\top \bar{K} V = I_r} \text{Tr}(V^\top \bar{K} L_S \bar{K} V).$$

Furthermore, the equation (S3.5) in the Supplementary Materials shows that L_S is positive semi-definite provided that $s_{ij} \geq 0$ for all $i \neq j$, implying that L_S can be interpreted as a kernel matrix. When $K_{\text{Y}} = L_S$, Net-NDR with orthonormal projections reduces to KSPCA.

4.2 Connection with kernel discriminant analysis

In this subsection, we establish the connection between Net-NDR and kernel discriminant analysis (KDA) under the degree-corrected stochastic block-model (DC-SBM) (Karrer and Newman, 2011). Under DC-SBM, node i is assigned a degree heterogeneity parameter θ_i for $i = 1, \cdots, n$, and the connection probability depends on both the community labels and the degree heterogeneity parameters, that is,

$$P\left(w_{ij} = 1 \mid C_i, C_j\right) = \theta_i \theta_j \, \mathrm{pr}_{C_i C_j} \quad \text{for} \quad i \neq j,$$

where $\mathrm{pr}_{C_i C_j}$ is the block-level link probability that depends only on C_i and C_j. For simplicity, we consider a special case in which $\mathrm{pr}_{C_i C_j} = a_t$ for $C_i = C_j = t$ and $\mathrm{pr}_{C_i C_j} = b$ for any $C_i \neq C_j$. Denote $\eta_{t,ij} = E(s_{ij} \mid C_i = C_j = t)$ for $t = 1, \cdots, N_c$ and $\nu_{ij} = E(s_{ij} \mid C_i \neq C_j)$. Note that s_{ij}'s are not identically distributed due to the θ_i's. For the covariates, given C_i, we assume that $\phi(X_i)$ is from the model in (2.8) for $i = 1, \cdots, n$.

The following proposition establishes the connection between Net-NDR and KDA (Baudat and Anouar, 2000).

Proposition 3. *Suppose that* $W = (w_{ij})_{n \times n}$ *is generated from the DC-SBM outlined above and that* X_i*'s follow the model in (2.8). Denote* $\bar{\nu} = \lim_n [n(n-1)]^{-1} \sum_{i \neq j} \nu_{ij}$, $\bar{\eta} = \lim_n [n(n-1)]^{-1} \sum_{i \neq j} \sum_{t=1}^{N_c} \pi_t^2 \eta_{t,ij}$ *and* $\bar{M} = \bar{\eta} + \bar{\nu} \sum_{t_1 \neq t_2} \pi_{t_1} \pi_{t_2}$. *Assume that*

(i) both $\bar{\eta}$ *and* $\bar{\nu}$ *exist and are bounded, and* $\bar{\nu} > \bar{M}$*;*

(ii) the r *largest eigenvalues of* $\Sigma_\phi^{-1} G_0$ *are distinct.*

Then Net-NDR is equivalent to KDA at the population level in the sense that for any given $r \leq n$, f_k in (2.2) is proportional to the k-th direction of KDA for $k = 1, \cdots, r$.

Remark 6. Consider a special case where $N_c = 2$, $\pi_1 = \pi_2 = 0.5$ and $s_{ij} = \alpha_0 - \alpha_1 w_{ij}$ with $\alpha_1 > 0$. Under this case, a sufficient condition for $\bar{\nu} > \bar{M}$ is that $b < (a_1 + a_2)/2$, that is, the link probability across communities is smaller than the largest within-community link probability. This requirement is lenient, weaker than the condition $b < \min\{a_1, a_2\}$ in Amini and Levina (2018).

Proposition 3 indicates that by utilizing information from the network, Net-NDR leverages the information of community labels well even if the labels are unknown. Finally, Net-NDR also has a connection with kernel (unsupervised) principal component analysis (Schölkopf et al., 1997); see Proposition B in the Supplementary Materials.

5. Numerical experiments

5.1 Community detection performance

In this subsection, we evaluate the performance of our two-step Net-NDR based community detection method (briefly denoted as Net-NDR) by comparing with the following four methods:

- NS-LDR (network-supervised linear dimension reduction in Zhao et al. (2022));
- CASC (covariate-assisted spectral clustering in Binkiewicz et al. (2017));
- NAC (network-adjusted covariates in Hu and Wang (2024));

- SDP (semidefinite programming in Yan and Sarkar (2021)).

The first three competitors NS-LDR, CASC, and NAC consider linear patterns among nodal covariates, whereas Net-NDR and SDP capture nonlinear ones. For fairness, both Net-NDR and SDP use the Gaussian kernel, with the Gaussian parameter determined by the heuristic rule in Gretton et al. (2012).

The performance is evaluated by the community detection accuracy (ACC) (proportion of correctly clustered nodes after the best permutation), and the adjusted Rand index (ARI) (Hubert and Arabie, 1985). Since the results of these two metrics yield similar conclusions, we present only the ACC results in the main text, and report the ARI results in Section S2.1 of the Supplementary Materials.

Recall that N_c denotes the number of communities. We consider two cases with two and four communities respectively (i.e. $N_c = 2$ and 4). The latent community labels C_i's are independently generated with probabilities $\{P(C_i = t), t = 1, \cdots, N_c\}$, taking values $\{\kappa, 1-\kappa\}$ for $N_c = 2$ and $\{\kappa, \kappa, 1/2-\kappa, 1/2-\kappa\}$ for $N_c = 4$; the parameter κ controls the balance of sizes of communities. Given C_i's, we consider a degree-corrected stochastic block-model for generating the network, that is,

$$P(w_{ij} = 1 | C_i = C_j) = \frac{1}{2}\gamma\theta_i\theta_j \text{ and } P(w_{ij} = 1 | C_i \neq C_j) = b\gamma\theta_i\theta_j,$$

where γ controls the network sparsity, b regulates the between-community link probability, and $\theta_1, \cdots, \theta_n \overset{i.i.d.}{\sim} U(0.5, 1)$ denote the degree heterogeneity. Specifically, a smaller γ results in a sparser network, and a smaller b implies a clearer community structure. Given C_i's, the covariates are generated as follows.

Case 1 (Concentric Rings) In this case, we set $N_c = 2$. For $i = 1, \cdots, 100$, the covariates are independently generated by

$$X_i = (R_{C_i} \cos(2\pi u_i), R_{C_i} \sin(2\pi u_i))^\top,$$

where u_i's $\overset{i.i.d.}{\sim} U(0,1)$, R_{C_i}'s $\sim U(0.5, 1)$ when $C_i = 1$ and R_{C_i}'s $\sim U(0.5+\delta, 1+\delta)$ when $C_i = 2$ with $\delta \geq 0$. The parameter δ serves to account for the differences in covariates from different communities. We set the default setting $(\delta, \kappa, \gamma, b) = (1, 0.5, 0.5, 0.2)$ and assess performance of each method by varying one of the parameters: (a) $\delta \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, (b) $\kappa \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, (c) $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and (d) $b \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$. For each setting, we carry out 300 repetitions and report the average ACCs along with the standard deviations in Figure 2.

Case 2. In this case, we set the sample size $n = 500$, the dimension $p = 10$, and the number of communities $N_c = 4$. For $i = 1, \cdots, 500$, the covariates are generated by

$$X_i = \left(\sin(2\pi e_{i,1}) \exp(-e_{i,2}^2), \exp(-e_i^\top e_i), l_{i,1}, l_{i,2}, u_{i,1}, \cdots, u_{i,p-4}\right)^\top,$$

where $e_i = (e_{i,1}, e_{i,2}, e_{i,3})^\top$ with $e_{i,j}$'s $\overset{i.i.d.}{\sim} U(0,1)$, $l_{i,j}$'s $\overset{i.i.d.}{\sim} U((k-1)\delta, (k-1)\delta + 1)$ when $C_i = k$ for $k = 1, \cdots, N_c$, and $u_{i,j}$'s $\overset{i.i.d.}{\sim} N(0,1)$. Similar to Case 1, a larger value of δ corresponds to a larger distance between covariates of different communities. We set the default setting of Case 2 as $(\delta, \kappa, \gamma, b) = (1, 0.25, 0.3, 0.2)$ and assess the performance of each method by varying one of the parameters: (a) $\delta \in \{1, 1.5, 2, 2.5, 3\}$, (b) $\kappa \in \{0.05, 0.1, 0.15, 0.2, 0.25\}$, (c) $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and (d) $b \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$. For each setting, we carry out 300 repetitions and report the average ACCs along with

the standard deviations in Figure 3.

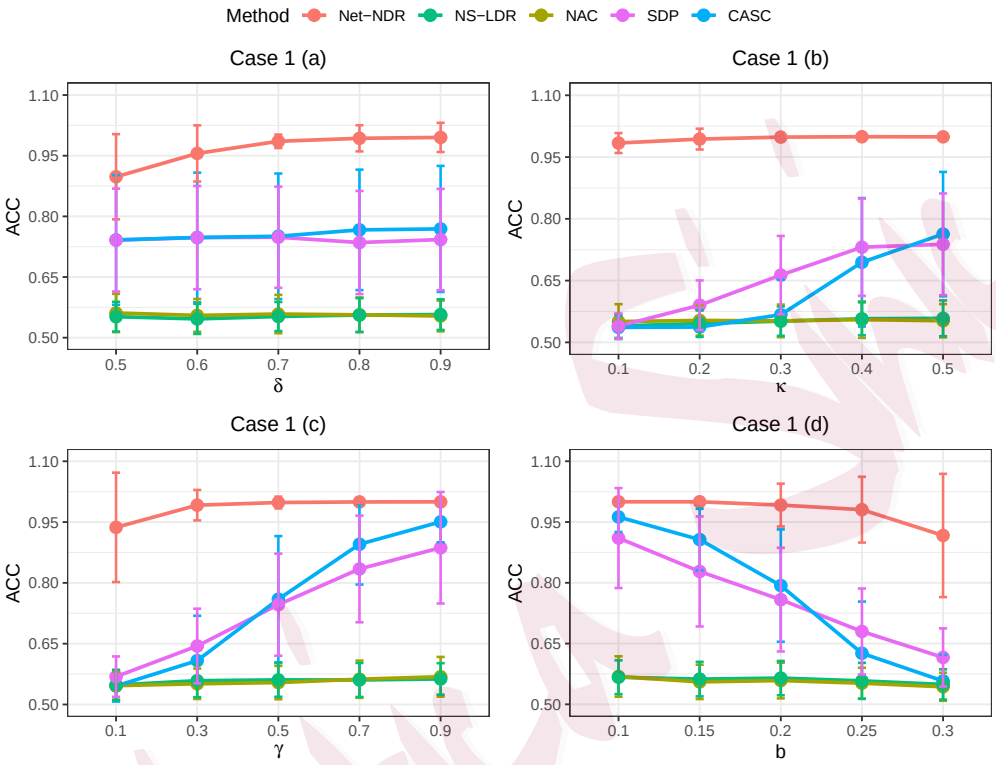

Figure 2: The community detection accuracy (ACC) of different methods for Case 1. Each dot represents the mean and the vertical bar represents the mean ± standard deviation.

As observed in Figures 2 and 3, Net-NDR outperforms the other four methods in most cases, largely due to its ability to effectively capture the nonlinear patterns among nodal covariates. The advantages of Net-NDR are summarized in three key aspects.

Effectiveness of Net-NDR in detecting changes in the between-community covariate difference. The ACC of Net-NDR increases as the between-community co-

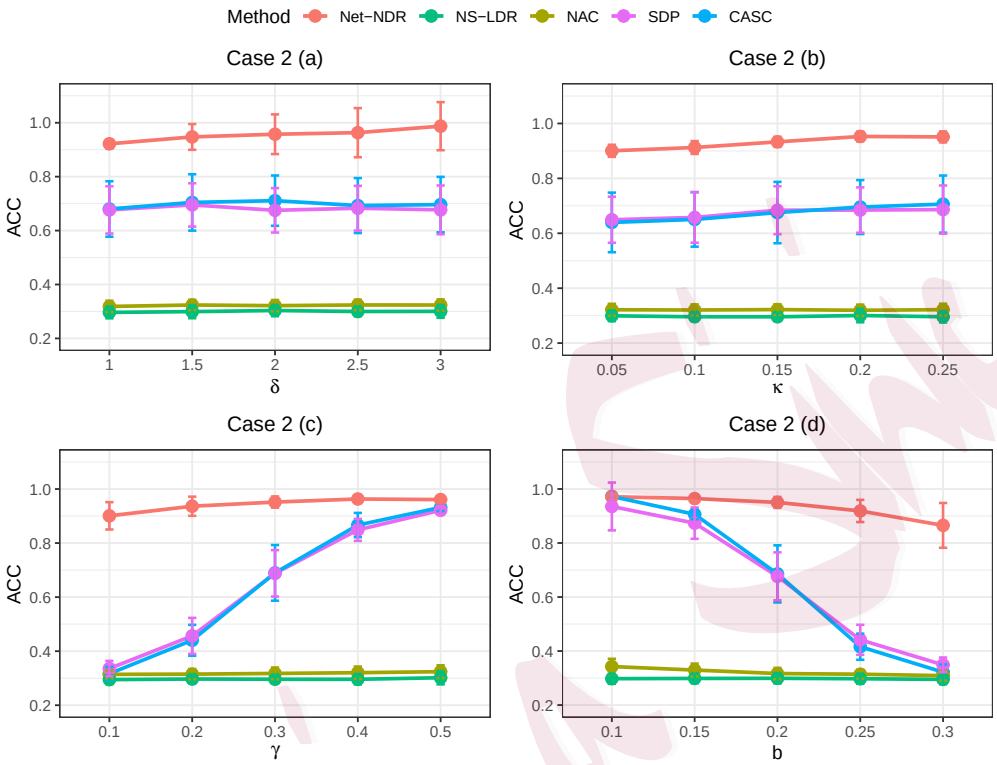

Figure 3: The community detection accuracy (ACC) of different methods for Case 2. Each dot represents the mean and the vertical bar represents the mean ± standard deviation.

variate difference δ grows, as shown in panel (a) of Figures 2 and 3. In contrast, the ACCs of four competitive methods remain at a level lower than that of Net-NDR, regardless of how δ changes, indicating that they can not effectively detect covariate changes in these two non-linear cases.

Robustness of Net-NDR to the imbalance of community sizes. Note that community sizes become more balanced when κ tends to $1/N_c$ in both two cases; particularly, the sizes are balanced when $\kappa = 0.5$ for Case 1 and $\kappa = 0.25$ for Case 2. Results in panel

(b) of Figures 2 and 3 indicate that Net-NDR has superior performance with strong robustness to the imbalance of community sizes. In contrast, the ACCs of SDP and CASC drop as the community sizes become more unbalanced in Case 1, and remain at a lower level than the ACC of Net-NDR in Case 2. The ACCs of NS-LDR and NAC remain at a low level, indicating their ineffectiveness in nonlinear cases.

Superiority of Net-NDR for sparse networks and unclear community structures. In panels (c) and (d) of Figures 2 and 3, when the network becomes sparser (i.e., γ decreases) and the community structure becomes less clear (i.e., b increases), the ACCs of SDP and CASC exhibit a sharp downward trend. In contrast, the ACC of Net-NDR decreases much slower.

5.2 Performance under varying n, p and N_c

We further investigate the effects of the sample size (n), the dimension (p), and the number of communities (N_c) on the accuracy of community detection. We take Case 2 with $(\delta, \gamma, b) = (1, 0.3, 0.2)$ in Section 5.1 as an example, and set the balance parameter $\kappa = 1/N_c$. For each setting, we carry out 300 repetitions and report the average ACCs along with the standard deviations in Figure 4. It shows that Net-NDR outperforms the other four methods in most cases. As n grows, the ACCs of Net-NDR, SDP and CASC increase, and the standard deviation of the ACC of Net-NDR becomes smaller. As p grows, the ACC of Net-NDR slightly decreases but remains higher than those of other methods. Finally, all the methods exhibit a decrease in ACC with the increasing number

of communities N_c, while Net-NDR shows a more gentle trend.

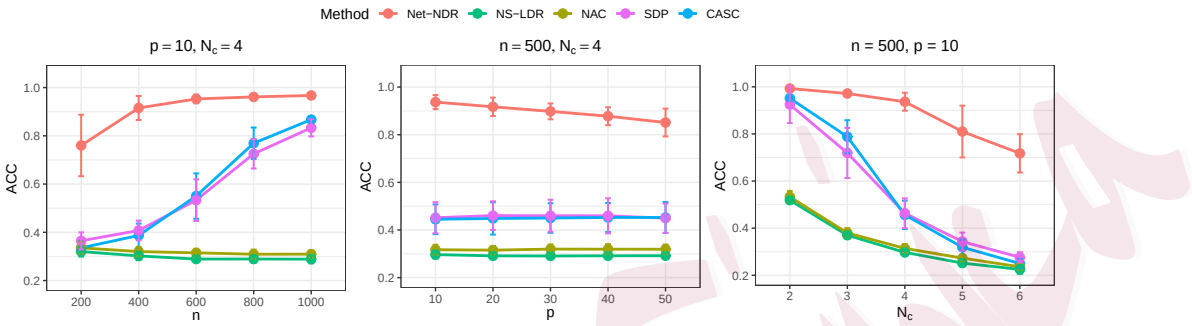

Figure 4: The community detection accuracy (ACC) for Case 2 with the varying sample size n (left), dimension p (middle), and the number of communities N_c (right). Each dot represents the mean and the vertical bar represents the mean $\pm$ standard deviation.

5.3 Robustness to the number of projection directions

In this subsection, we examine the robustness of Net-NDR to the number of projection directions. Taking Case 1 with $(\delta, \kappa, \gamma, b) = (0.5, 0.5, 0.8, 0.2)$ and Case 2 with $(\delta, \kappa, \gamma, b) = (1, 0.25, 0.8, 0.2)$ in Section 5.1 as examples, we vary the number of projection directions in {1,2,3,4,5}. For each setting, we carry out 300 repetitions and report the boxplot of ACC in Figure 5. It shows that the performance of Net-NDR is robust to the number of projection directions. We also conducted simulations to evaluate the consistency of rank selection; see Section S2.4 in the Supplementary Materials.

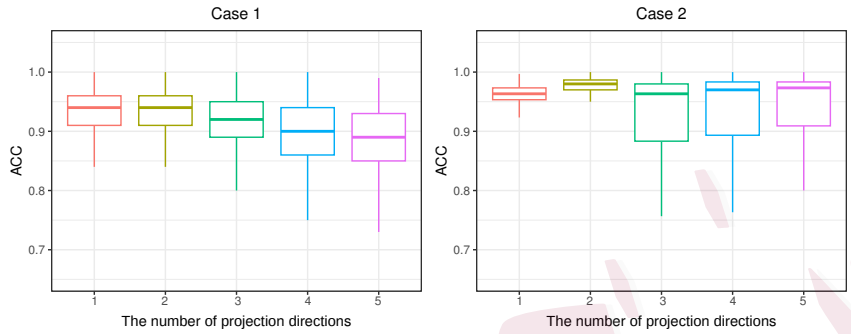

Figure 5: The community detection accuracy (ACC) of Net-NDR with different numbers of projection directions in Case 1 and Case 2.

5.4 Robustness to kernel functions

We examine the performance of Net-NDR using different kernel functions under two cases in Section 5.1: i) Case 1 with $(\delta, \kappa, \gamma, b) = (0.6, 0.5, 0.2, 0.2)$ and ii) Case 2 with $(\delta, \kappa, \gamma, b) = (1, 0.25, 0.3, 0.2)$. In addition to the Gaussian kernel function, we also consider the Laplacian kernel $\mathcal{K}(X_i, X_j) = \exp(\|X_i - X_j\|_1/\sigma_{\text{lap}})$, where σ_{lap} is determined by the heuristic rule in Gretton et al. (2012), and the polynomial kernel $\mathcal{K}(X_i, X_j) = (c_0 X_i^\top X_j + c_1)^{d_0}$ with parameters (c_0, c_1, d_0). The c_0 and c_1 are selected by grid search to maximize the Calinski-Harabasz index, with ranges $c_0 \in \{0.3, 0.6, 0.9, 1.2, 1.5\}$ and $c_1 \in \{0, 0.5, 1, 1.5, 2\}$; d_0 is set to 3 for Case 1 and 2 for Case 2. For each setting, we carry out 300 repetitions and report the boxplot of ACC in Figure 6.

As observed in Figure 6, the performance of Net-NDR remains largely consistent across different kernel functions. In practice, we recommend using the Gaussian kernel,

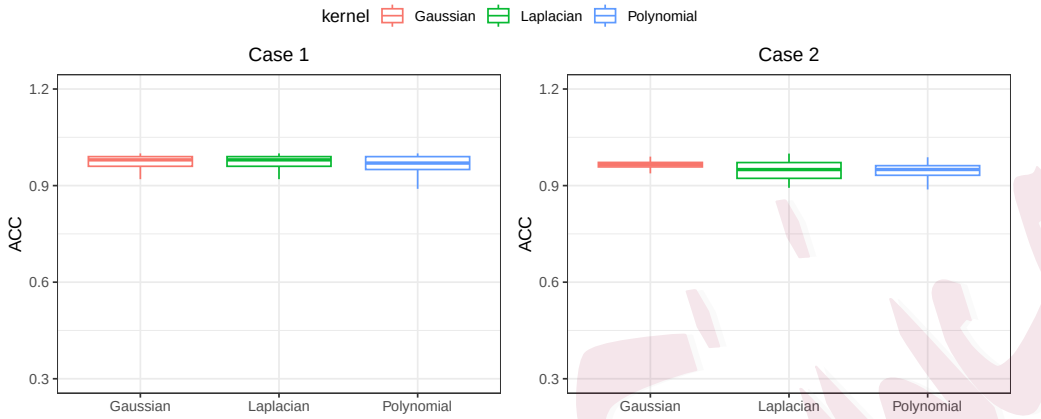

Figure 6: The community detection accuracy (ACC) of Net-NDR with different kernel functions in Case 1 and Case 2.

given its strong nonlinear mapping capability and robust generalization performance.

6. Real data analysis

In this section, we demonstrate the application of Net-NDR to a pulsar candidate dataset, available at http://archive.ics.uci.edu/ml/datasets/HTRU2. Pulsars are a rare class of neutron stars that emit radio emissions detectable from Earth. Their scientific value lies in their use as natural probes for studying spacetime, the interstellar medium, and extreme states of matter. Accurate classification of pulsar signals can thus advance our knowledge in multiple areas of physics, including particle behavior in ultra-strong magnetic fields and tests of gravitational theories under strong-field conditions.

The dataset contains 16,259 spurious (negative) instances caused by radio frequency interference or noise, and 1,639 real pulsar (positive) instances verified by human anno-

tators. Each observation is described by a binary class variable (i.e., $N_c = 2$) and eight continuous features. The first four features are statistical measures (mean, standard deviation, excess kurtosis, and skewness) of the integrated pulse profile, while the last four correspond to the same statistics calculated from the dispersion-measure-signal-to-noise ratio (DM-SNR) curve. For our analysis, we randomly select 200 observations from the 16,259 negative instances and 100 observations from the 1,639 positive instances.

To construct a network among these 300 nodes, we consider two scenarios for generating edges by thresholding the covariate difference:

Scenario 1: For nodes i and j, we set $w_{ij} = \mathbb{I}[d_1(X_i, X_j) \geq d_\tau]$, where $d_1(X_i, X_j) = \exp(-|X_{i,1} - X_{j,1}|)$ denotes the difference in the first variable and d_τ denotes the τ quantile of $\{d_1(X_i, X_j), i < j\}$. The remaining seven features are used as nodal covariates.

Scenario 2: To make the covariates less likely to be linearly separated, we perform principal component analysis on the covariates, discard the first $p_c \in \{1, 2, 3\}$ principal components, and project covariates onto the remaining $8 - p_c$ components to form the design matrix $\tilde{\mathbb{X}} = (\tilde{X}_1^\top, \cdots, \tilde{X}_n^\top)^\top \in \mathbb{R}^{n \times (8-p_c)}$. We set $P(w_{ij} = 1 \mid \tilde{X}_i, \tilde{X}_j) = 0.9$ when $d_2(\tilde{X}_i, \tilde{X}_j) \geq d_\tau$ and $P(w_{ij} = 1 \mid \tilde{X}_i, \tilde{X}_j) = 0.1$ otherwise, where $d_2(\tilde{X}_i, \tilde{X}_j) = \exp\left(-\|\tilde{X}_i - \tilde{X}_j\|^2/8\right)$ and d_τ denotes the τ quantile of $\{d_2(\tilde{X}_i, \tilde{X}_j), i < j\}$.

For both two scenarios, we set $\tau \in \{0.5, 0.7, 0.8, 0.9, 0.95\}$; a larger τ (i.e., a smaller $1-\tau$) leads to a sparser network. For fairness, the two methods Net-NDR and SDP, which capture nonlinear patterns among nodal covariates, both use the Gaussian kernel function. The generation process is repeated 200 times, and community detection accuracy results

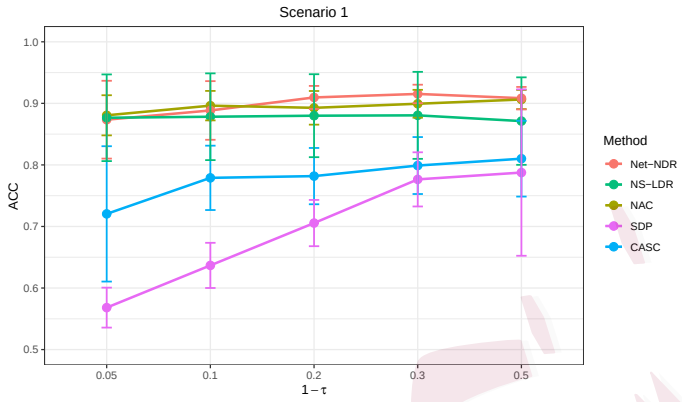

Figure 7: Community detection accuracy (ACC) of different methods on the pulsar dataset in Scenario 1. Each dot represents the mean, and vertical bars represent mean ± standard deviation.

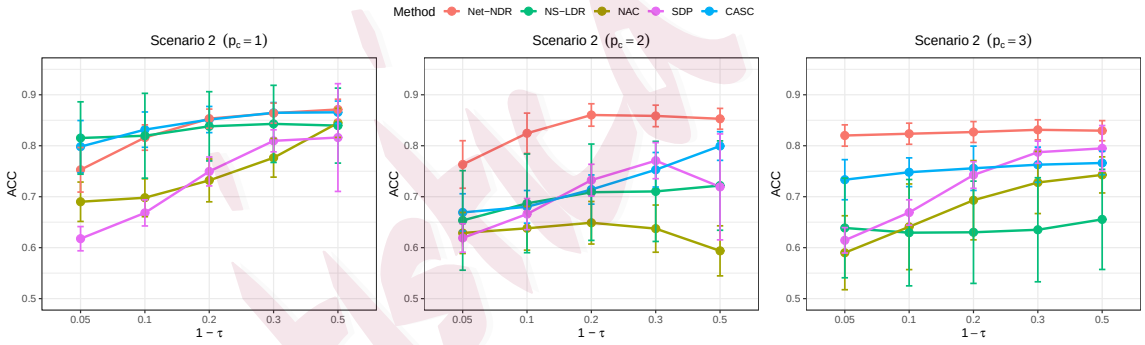

Figure 8: Community detection accuracy (ACC) of different methods on the pulsar dataset in Scenario 2 for $p_c = 1, 2, 3$ (from left to right). Each dot represents the mean, and vertical bars represent mean ± standard deviation.

are shown in Figures 7 and 8. Figure 7 shows that the linear dimension reduction method NS-LDR exhibits strong performance, indicating that Scenario 1 probably presents linear patterns among nodal covariates. Notably, Net-NDR retains competitive performance

even under this linear case. In Figure 8, as p_c becomes larger, the performance of the linear method NS-LDR deteriorates, suggesting that removing the leading principal components likely enhances nonlinear patterns in the nodal covariates. It can be seen that when $p_c = 2$ and 3, Net-NDR outperforms other methods across different network sparsity levels, highlighting its advantages in the nonlinear setting. Additionally, Net-NDR exhibits smaller standard deviations than other methods, further demonstrating its robustness.

Supplementary Materials

The Supplementary Materials include proofs of all the theoretical results in the main text, additional simulations, discussions and theoretical results.

Acknowledgments

Junlong Zhao's research was supported in part by National Natural Science Foundation of China grants No.12371288 and 12131006, and the Fundamental Research Funds for the Central Universities.

References

Amini, A. and E. Levina (2018). On semidefinite relaxations for the block model. The Annals of Statistics 46(1), 149–179.

Bansal, M., G. D. Gatta, and D. Di Bernardo (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics 22(7), 815–822.

Barshan, E., A. Ghodsi, Z. Azimifar, and M. Z. Jahromi (2011). Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. Pattern Recognition 44(7), 1357–1371.

Baudat, G. and F. Anouar (2000). Generalized discriminant analysis using a kernel approach. Neural Computation 12(10), 2385–2404.

Berlinet, A. and C. Thomas-Agnan (2011). Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer Science & Business Media.

Binkiewicz, N., J. T. Vogelstein, and K. Rohe (2017). Covariate-assisted spectral clustering. Biometrika 104(2), 361–377.

Bomiriya, R. P., A. R. Kuvelkar, D. R. Hunter, and S. Triebel (2023). Modeling homophily in exponential-family random graph models for bipartite networks. arXiv:2312.05673.

Caliński, T. and J. Harabasz (1974). A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods 3(1), 1–27.

Davezies, L., X. D'haultfœuille, and Y. Guyonvarch (2021). Empirical process results for exchangeable arrays. The Annals of Statistics 49(2), 845–862.

Elkabani, I. and R. A. A. Khachfeh (2015). Homophily-based link prediction in the facebook online social network: A rough sets approach. Journal of Intelligent Systems 24(4), 491–503.

Fan, J., J. Ge, and J. Hou (2025). Covariates-adjusted mixed-membership estimation: A novel network model with optimal guarantees. arXiv:2502.06671.

Gao, T., Y. Zhang, R. Pan, and H. Wang (2023). Large-scale multi-layer academic networks derived from

statistical publications. arXiv:2308.11287.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). A kernel two-sample test. The Journal of Machine Learning Research 13(25), 723–773.

Hu, Y. and W. Wang (2024). Network-adjusted covariates for community detection. Biometrika 111(4), 1221–1240.

Huang, S., J. Sun, and Y. Feng (2024). PCABM: Pairwise covariates-adjusted block model for community detection. Journal of the American Statistical Association 119(547), 2092–2104.

Hubert, L. and P. Arabie (1985). Comparing partitions. Journal of Classification 2, 193–218.

Hunter, D. R., S. M. Goodreau, and M. S. Handcock (2008). Goodness of fit of social network models. Journal of the American Statistical Association 103(481), 248–258.

Jackson, M., S. M. Nei, E. Snowberg, and L. Yariv (2023). The dynamics of networks and homophily. SSRN Electronic Journal.

Karrer, B. and M. E. Newman (2011). Stochastic blockmodels and community structure in networks. Physical Review E 83(1), 016107.

Lam, C. and Q. Yao (2012). Factor modeling for high-dimensional time series: inference for the number of factors. The Annals of Statistics 40(2), 694–726.

Li, B. and J. Song (2017). Nonlinear sufficient dimension reduction for functional data. The Annals of Statistics 45(3), 1059–1095.

Luxburg, U. V. (2007). A tutorial on spectral clustering. Statistics and Computing 17(4), 395–416.

Ma, Z., Z. Ma, and H. Yuan (2020). Universal latent space model fitting for large networks with edge covariates.

Journal of Machine Learning Research 21(4), 1–67.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. Annual Review of Sociology 27, 415–444.

Menzel, K. (2021). Bootstrap with cluster-dependence in two or more dimensions. Econometrica 89(5), 2143–2188.

Merris, R. (1994). Laplacian matrices of graphs: a survey. Linear Algebra and its Applications 197-198, 143–176.

Ogburn, E. L., O. Sofrygin, I. Diaz, and M. J. Van der Laan (2024). Causal inference for social network data. Journal of the American Statistical Association 119(545), 597–611.

Qiu, Y. (2024). Large-scale eigenvalue decomposition and svd with RSpectra. Website. `https://cran.r-project.org/web/packages/RSpectra/vignettes/introduction.html`.

Rhodes, A. (2018). The age of belonging: friendship formation after residential mobility. Social Forces 97(2), 583–606.

Roy, S., Y. Atchadé, and G. Michailidis (2019). Likelihood inference for large scale stochastic blockmodels with covariates based on a divide-and-conquer parallelizable algorithm with communication. Journal of Computational and Graphical Statistics 28(3), 609–619.

Schölkopf, B., A. Smola, and K.-R. Müller (1997). Kernel principal component analysis. In Artificial Neural Networks — ICANN'97, pp. 583–588. Springer Berlin Heidelberg.

Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics 34(2), 166–176.

Stein, S. and C. Leng (2023). An annotated graph model with differential degree heterogeneity for directed

networks. Journal of Machine Learning Research 24(119), 1–69.

Stein, S. and C. Leng (2025). A sparse beta regression model for network analysis. Journal of the American Statistical Association 120(550), 1281–1293.

Sweet, T. M. (2015). Incorporating covariates into stochastic blockmodels. Journal of Educational and Behavioral Statistics 40(6), 635–664.

Virta, J., K.-Y. Lee, and L. Li (2022). Sliced inverse regression in metric spaces. Statistica Sinica 32, 2315–2337.

Xu, M. and Q. Wang (2023). A network poisson model for weighted directed networks with covariates. Communications in Statistics - Theory and Methods 52(15), 5274–5293.

Xu, R. and D. C. Wunsch II (2008). Clustering. John Wiley & Sons.

Xu, S., Y. Zhen, and J. Wang (2023). Covariate-assisted community detection in multi-layer networks. Journal of Business & Economic Statistics 41(3), 915–926.

Yan, B. and P. Sarkar (2021). Covariate regularized community detection in sparse graphs. Journal of the American Statistical Association 116(534), 734–745.

Ying, C. and Z. Yu (2022). Fréchet sufficient dimension reduction for random objects. Biometrika 109(4), 975–992.

Zhang, Q., B. Li, and L. Xue (2024). Nonlinear sufficient dimension reduction for distribution-on-distribution regression. Journal of Multivariate Analysis 202, 105302.

Zhao, J., X. Liu, H. Wang, and C. Leng (2022). Dimension reduction for covariates in network data. Biometrika 109(1), 85–102.