Received: 28 September 2021 | Accepted: 2 August 2022 | Published on: 2 November 2022

DOI: 10.1111/rssb.12554

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY | Series B Statistical Methodology

Linear regression and its inference on noisy network-linked data

Can M. Le[1] | **Tianxi Li**[2]

[1]Department of Statistics, University of California, Davis, Davis, California, USA

[2]Department of Statistics, University of Virginia, Charlottesville, Virginia, USA

Correspondence
Tianxi Li, Department of Statistics, University of Virginia, Charlottesville, VA, USA.
Email: tianxili@virginia.edu

Funding information
College and Graduate School of Arts and Sciences, Grant/Award Number: 3Caverliers Award; National Science Foundation, Grant/Award Numbers: DMS-2015134, DMS-2015298

Abstract

Linear regression on network-linked observations has been an essential tool in modelling the relationship between response and covariates with additional network structures. Previous methods either lack inference tools or rely on restrictive assumptions on social effects and usually assume that networks are observed without errors. This paper proposes a regression model with non-parametric network effects. The model does not assume that the relational data or network structure is exactly observed and can be provably robust to network perturbations. Asymptotic inference framework is established under a general requirement of the network observational errors, and the robustness of this method is studied in the specific setting when the errors come from random network models. We discover a phase-transition phenomenon of the inference validity concerning the network density when no prior knowledge of the network model is available while also showing a significant improvement achieved by knowing the network model. Simulation studies are conducted to verify these theoretical results and demonstrate the advantage of the proposed method over existing work in

Can M. Le and Tianxi Li contributed equally to this study.

terms of accuracy and computational efficiency under different data-generating models. The method is then applied to middle school students' network data to study the effectiveness of educational workshops in reducing school conflicts.

KEYWORDS

linear regression, network-linked data, network modelling, network perturbation, random networks

1 | INTRODUCTION

Nowadays, networks appear frequently in many areas, including social sciences, transportation and biology. In most cases, networks are used to represent relationships or interactions between units of a complex (social, physical or biological) system, so analysing network data may render crucial insights into the dynamics and/or interaction mechanism of the system. One particular, yet commonly encountered, situation is when a group of units is observed connected by a network and a set of attributes for each of these units is available. Such data sets are sometimes called network-linked data (Li et al., 2019; Li et al., 2020c) or multi-view network data (Gao et al., 2019). Network-linked data are widely available in almost all fields involving network analysis about social effects (Michell & West, 1996; Pearson & West, 2003), collaborations (Ji & Jin, 2016; Su et al., 2020), and causal experiments (Basse & Airoldi, 2018a; Basse & Airoldi, 2018b). In network-linked data, rich information is available from the perspective of both the individual attributes and the network, and the challenge is to find proper statistical methods to incorporate both. Suppose one single network is available. Consider the situation where, for each node i of the network, we observe (x_i, y_i), in which $x_i \in \mathbb{R}^p$ is a vector of covariates while $y_i \in \mathbb{R}$ is a scalar response. In particular, we aim for a regression model of y_i against x_i that also takes the network information into account. Such a model arises naturally in any problem when a prediction model or inference of a specific attribute is of interest.

Although the systematic study of regression on network-linked data has only recently begun to attract interest in statistics (Li et al., 2019; Su et al., 2020; Zhu et al., 2017), it has been studied in econometrics by many authors, who mostly focused on multiple networks (Bramoullé et al., 2009; Lee, 2007; Manski, 1993) or longitudinal data (Jackson & Rogers, 2007; Manresa, 2013). Social effects are typically observed in the form that connected units share similar behaviours or properties. The similarity or correlation may be due to either *homophily*, where social connections are established because of similarity, or *contagion*, where individuals become similar through the influence of their social ties. In general, one cannot distinguish homophily from contagion in a single snapshot of observational data (Shalizi & Thomas, 2011), as in our setting. Therefore, the two directions of causality will not be distinguished, and this type of generic similarity between connected nodes will be called 'network cohesion', as in (Li et al., 2019; Li et al., 2020c). A significant class of models for this type of network regression problems is the class of autoregressive models, which is based on ideas in spatial statistics. The spatial autoregressive models have been widely used in econometrics, such as in (Bramoullé et al., 2009; Hsieh & Lee, 2016; Lee, 2007; Manski, 1993; Zhu et al., 2017), to name a few. A common form for such a model is

$$y_i = \gamma\left(\frac{1}{n_i}\sum_{i'\sim i} y_{i'}\right) + X\beta + \frac{1}{n_i}\left(\sum_{i'\sim i} x_{i'}\right)^T \eta + \epsilon,$$

where y_i is the response, x_i is the covariate vector and n_i is the number of neighbours of i (the notation $i \sim i'$ indicates that i and i' are connected). In this model, the neighbourhood averages of the response and covariates are used to model the social effect. The parameters γ and η are typically called 'endogenous' and 'exogenous' effects, respectively, while β is the standard regression slope. This class of models is also called the linear in means model or social interaction model (SIM). When multiple networks for different populations are available, the intercept can be treated as a group effect, called the external effect (Lee, 2007; Lee et al., 2010; Manski, 1993). The SIM framework can also combined with network formulation models to study the correlation, homophily and contagion effects, if rather than a single network, one could observe multiple network snapshots over time (Goldsmith-Pinkham & Imbens, 2013; McFowland III & Shalizi, 2021). The SIM framework, though it has been a popular set-up for regression on network-linked data, suffers from two crucial drawbacks. The first comes from its restrictive parametric form of the social effect. Assuming social effect in the form of autoregressive neighbourhood average (or summation) is far from realistic, and this stringent assumption significantly limits the usefulness of the framework. As shown in Li et al. (2019) and also in our empirical study, the restrictive assumption of social effect leads to poor prediction performance. The second limitation comes from the assumption that the network structure is precisely observed from the data. In practice, it is well known that most network data are subject to errors, due to missingness (Butts, 2003; Clauset & Moore, 2005; Lakhina et al., 2003), observational errors (Handcock & Gile, 2010; Khabbazian et al., 2017; Le et al., 2018; Lunagómez et al., 2018; Newman, 2018; Rolland et al., 2014) or data collection method (Rohe, 2019; Wu et al., 2018). If the imprecise network is used in SIM, the model is ill-defined and the inference would also be problematic, as shown by Chandrasekhar and Lewis (2011).

In this paper, a new regression model to fit the network-linked data is proposed that addresses both drawbacks mentioned above. The new model is based on a flexible network effect assumption and is robust to network observational errors. It also allows us to perform tests and construct confidence intervals for the model parameters. Our theoretical analysis provides the support for model estimation and inference and quantifies the magnitude of network observational errors under which the inference remains robust. Moreover, in the random network perturbation scenario, the tradeoff between the available information of the network structure and the level of robustness of the statistical inference is characterised. In its most difficult setting, when no prior knowledge about the network model is available, the result reveals a phase-transition phenomenon at the network average degree of $\sqrt{n}$, above which the inference is asymptotically correct, and below which the inference becomes invalid. The inference could remain valid for much sparser networks if more information about the network structure is available. For example, when the network model is known and an effective parametric estimation can be applied, the sparsity requirement can be relaxed to $\log n$. To the best of our knowledge, this paper is the first work that addresses the inference of the network-linked regression models and accounts for the network observational errors.

Related to the network-linked model proposed in this paper is the semi-parametric model called regression with network cohesion (RNC), introduced by Li et al. (2019). In their model, the network effect is represented by individual parameters that are assumed to be 'smooth' over the network, and a similar idea is used in a few other statistical estimation settings (Fan & Guan, 2018;

TABLE 1 Comparison between the proposed method and two other popular methods for network-linked data

Models	Social effect flexibility	Inference	Network robustness
SIM (Manski, 1993)	×	✓	×
RNC (Li et al., 2019)	✓	×	×
Current method	✓	✓	✓

Abbreviations: RNC, regression with network cohesion; SIM, social interaction model.

Wang et al., 2016; Zhao & Shojaie, 2016). Despite its flexibility of social effect assumption and excellent predictive performance, the RNC model lacks a valid inference framework and cannot be applied in many modern applications where statistical inference is needed (Ogburn, 2018; Su et al., 2020). The result of Li et al. (2020c) indicates that the RNC estimator fails to guarantee valid inference under reasonable assumptions unless additional assumptions such as sparsity are made (Zhao & Shojaie, 2016). Moreover, little is known about the robustness of the RNC method to network observational errors, although preliminary results have been obtained in a particular scenario of network sparsification (Li et al., 2019; Sadhanala et al., 2016). In contrast, our model does not assume the smoothness of network effects as in the RNC method. Instead, we use a general relational subspace to define the social effects. As can be seen later, the proposed model overcomes both of the two aforementioned limitations of RNC and is computationally more efficient.

Table 1 gives a high-level comparison between the proposed model and two popular benchmark regression models (discussed previously) in three aspects: the flexibility of modelling social effects, the availability of an inference method, and the provable robustness to the network perturbation. The model we introduced in this paper is the only one of the three which renders all of the desired properties.

The rest of the paper is organised as follows. Section 2 introduces our model and the corresponding statistical inference algorithm. Sections 3 and 4 are devoted to our main theoretical results; the generic theory for model estimation consistency and the asymptotic inference is given in Section 3; then, under the random network modelling framework, detailed discussions of the technical requirement are introduced in Section 4. Extensive simulation experiments are given in Section 5 to verify our theory and compare our method with a few benchmark methods mentioned above. Section 6 demonstrates the usefulness of the proposed method in analysing a study about the effects of educational workshops on reducing school conflicts. Concluding remarks and future directions are discussed in Section 7. Extended theoretical results, proofs, additional experiments and data analysis are given as Appendix S1.

2 | NETWORK REGRESSION MODEL

Throughout the paper, $0_{k\times\ell}$ and I_k are used to denote the zero matrix of size $k \times \ell$ and the identity matrix of size $k \times k$, respectively. The subscripts may be dropped when the dimensions are clear from the context. We use e_i to denote the vector whose ith coordinate is 1 and the remaining coordinates are zero; the dimension of e_i may vary according to the context. Given a matrix $M \in \mathbb{R}^{k\times\ell}$ and $1 \le i \le j \le \ell$, $M_{i:j} \in \mathbb{R}^{k\times(j-i+1)}$ is used to denote the matrix whose columns are those of M with indices $i, i+1, \ldots, j$; when $i = j$, M_i is used instead of $M_{i:i}$ for the simplicity of the notation. Throughout the paper, $\|.\|$ denotes the spectral norm, which is the largest singular

value, for matrices and the Euclidean norm for vectors. The number of nodes in the network is denoted by n. We say that an event E occurs with high probability if $\mathbb{P}(E) \geq 1 - n^{-c}$ for some constant $c > 2$. We use $a_n = o(1)$ and $b_n = O(1)$ to indicate that $\lim_{n\to\infty} a_n = 0$ and b_n is a bounded sequence, respectively.

2.1 | A semi-parametric regression model with network effects

The unobserved true network is captured by the relational matrix $P \in \mathbb{R}^{n\times n}$, where P_{ij} describes the strength of the relation between nodes i and j. For each node i of the network (x_i, y_i) is observed, where $x_i \in \mathbb{R}^p$ is a vector of covariates while $y_i \in \mathbb{R}$ is a scalar response. Denote by $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ the vector of responses and by $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n\times p}$ the design matrix, which sometimes has the first column as the all-one vector. Our goal is to construct a model that describes the dependence of Y on X and P.

The observed network is represented by an $n \times n$ adjacency matrix A, where $0 \leq A_{ij} \leq 1$ is the weight of the edge between i and j. In the special case of an unweighted network, A is a binary matrix and $A_{ij} = 1$ if and only if node i and node j are connected. We view A as a perturbed version of P and focus on undirected networks, for which A and P are symmetric matricies. Moreover, we only consider the *fixed design* setting, to avoid unnecessary complication of jointly modelling covariates and relational data. That means X and P are always treated as fixed. The adjacency matrix A is also treated as fixed for the moment as the regression model and its generic inference framework are described. Later on, in only Section 4 where we study the robustness of our framework under deviation of A from P, we will assume A as a perturbation of P following random network models.

Intuitively, the structural assumption is that y_i's tend to be similar for individuals having strong connections. This is called the assortative mixing or network cohesion property (Kolaczyk, 2009; Li et al., 2019). To incorporate this intuition, consider the model

$$\mathbb{E}Y = X\zeta + \mu, \tag{1}$$

where $\zeta \in \mathbb{R}^p$ is the vector of coefficients for the covariates X and $\mu \in \mathbb{R}^n$ is the vector of individual effects reflecting the network cohesion property. This model was studied in Li et al. (2019). Instead of modelling the network effect by using a specific auto-regressive dependence as in Manski (1993), (1) treats the network effect as a non-parametric component. As shown by Li et al. (2019) and the experiments later on in this paper, this approach is more flexible and gives significantly better predictive performance than the SIM framework of Manski (1993). In Li et al. (2019), μ is assumed to have small sum of squared differences $\sum_{i,j:A_{ij}=1} (\mu_i - \mu_j)^2$. In contrast, we rely on a different form of the network cohesion that proves to be more general and stable. Specifically, the cohesion requirement of μ is formulated by assuming that

$$\mu \in S_K(P), \tag{2}$$

where $S_K(P)$ is the subspace spanned by the K leading eigenvectors of P. Depending on specific problems it may be possible to replace P in (2) with other appropriate matrix-valued functions of P for which the inference framework remains valid. One such example is provided in Section A of Appendix S1. The idea of using the eigenspace of P to encode the cohesive pattern over the network is motivated by many previous studies. It is related to the standard

spectral embedding (Belkin & Niyogi, 2003; Ng et al., 2001; Shi & Malik, 2000; Tang et al., 2013), which maps the nodes of the network to a set of points in a Euclidean space so that the geometric relations between these points are similar to the topological relations of the nodes in the original network. Moreover, under random network models, Li et al. (2020a) and Lei et al. (2020) show that the eigenspaces of the expected adjacency matrix and Laplacian matrix encode varying resolutions of node similarity. Intuitively, since the spectral space combines both the network's macro-scale and micro-scale patterns, it is expected to be reasonably robust to perturbations on local connections. This gives a significant advantage compared with the SIM model. Figure 3 (Section 6) includes such an example. In summary, we assume the following mean structure:

Definition 1. The mean structure of the regression model with network effects is defined to be

$$\mathbb{E}Y \in \text{span}\{\text{col}(X), S_K(P)\}, \tag{3}$$

where $\text{col}(X)$ is the column space of X and span denotes the linear subspace jointly spanned by $\text{col}(X)$ and $S_K(P)$. This is equivalent to (1) with the cohesion property (2).

The advantage of our mean structure (1) lies in its generality. In particular, one distinction between our model and the more commonly assumed linear mean structure is that our model includes the situation when the two subspaces $\text{col}(X)$ and $S_K P$ have non-trivial intersection. For example, one of the covariates may be perfectly cohesive over the network, and this situation is allowed in our model. In this case, one could not uniquely determine $X\zeta$ and μ. However, such a tricky situation is unavoidable for the level of generality and robustness we want to achieve. As a matter of fact, this type of ambiguity is an instance of the general conclusion from Shalizi and Thomas (2011) that contagion, and different types of homophily effects from a single snapshot of observational data are not distinguishable. Nevertheless, the mean-structure, as well as an interpretable decomposition of the effects, can be uniquely identified with the following parameterisation.

Definition 2. Let $\mathcal{R} = \text{col}(X) \cap S_K(P)$ be the intersection of $\text{col}(X)$ and $S_K(P)$. Model (3) can reparametrised as

$$\mathbb{E}Y = X\beta + \xi + \alpha, \tag{4}$$

where

$$\xi \in \mathcal{R}, \quad X\beta \perp \mathcal{R}, \quad \alpha \in S_K(P), \quad \alpha \perp \mathcal{R}, \tag{5}$$

and $\beta \in \mathbb{R}^p$, $\xi, \alpha \in \mathbb{R}^n$. We call it the regression model with network effects.

Although there are several ways to reparameterise the mean structure (3), the one in (5) is preferable because it combines the two sources of information (the covariates X and the relational information P) in a natural way so that the interpretation of the corresponding parameters is straightforward. Specifically, we observe that

- When $\beta = 0$, we have $\mathbb{E}Y \in S_K(P)$ so the model can be specified as a cohesive network effects without using the covariates X. Therefore, β represents the conditional covariate effects of X given P.

- When $\alpha = 0$, we have $\mathbb{E}Y \in \text{col}(X)$ so the model can be specified by the covariates without using the relational information P. Therefore, α represents the conditional network effect of P given X.

Thus, β and α reflect the conditional effects similar to the covariate effects in the standard linear regression setting. In addition, ξ is the overlap of the two sources. The following result confirms the identifiability of our parameters.

Proposition 1 (Parameter identifiability). *The parameterisation in Definition 2 is identifiable. That is, if there exist* (β, α, ξ) *and* (β', α', ξ') *satisfying* (4) *and* (5) *simultaneously then* $\beta = \beta'$, $\alpha = \alpha'$, *and* $\xi = \xi'$.

Remark 1. A seemingly easier way to combine the two sources of information is to treat the eigenvectors of P as another set of covariates and fit Y using both X and the eigenvectors. However, this approach has to assume $\mathcal{R} = \{0\}$ for identifiability. This model is a special case of our model, and as can be seen later, our model estimation method would adapt to this reduced case. Our model does not enforce this assumption because it is restrictive and also causes difficulties in interpretations (see Section H in in Appendix S1 for details). Note also that (5) does not require $\alpha \perp X\beta$, although this strong restriction would significantly simplify the fitting procedure and its analysis.

In the present setting, even when p is much smaller than n, the model inference is a high-dimensional problem because of the non-parametric individual effects α. Throughout this paper, we assume that the noise in the regression model of Definition 2 is a multivariate Gaussian vector, as in many other inference methods of high-dimensional regression model (Javanmard & Montanari, 2014; Van de Geer et al., 2014; Zhang & Zhang, 2014). Specifically,

$$Y = \mathbb{E}Y + \epsilon \quad \text{and} \quad \epsilon \sim N(0, \sigma^2 I), \tag{6}$$

where $I \in \mathbb{R}^{n \times n}$ is the identity matrix and σ^2 is the variance of the noise. We leave the study of other noise distributions for future work.

Remark 2. As to be clear in our theory later, the reasonable interpretation of K in (2) is the index where the spectral space of P has a large eigen gap. The problem of estimating such a K has been extensively studied in network literature and many efficient methods are now available (Chatterjee, 2015; Chen & Lei, 2018; Jin et al., 2020; Le & Levina, 2022; Li, Levina, & Zhu, 2020). These methods all provide theoretical guarantees for the recovery of K with overwhelming probability and can be applied to our setting. Since the task of estimating K neither makes our problem conceptually more challenging nor brings more insight about it, for simplicity of presentation, we assume that K is known throughout the paper.

2.2 | Statistical inference method

Our generic inference framework requires access to a certain approximation of P, denoted by $\hat{P}$. The specific $\hat{P}$ is determined by the user according to the understanding of the data problem. For example, one may assume that the adjacency matrix A is a perturbed version of P. In this situation, without additional information, a natural option is to set $\hat{P} = A$ and we refer to the corresponding estimation and inference procedure as the 'model-free' version of our method, highlighting that no specific model for the network structure is assumed. Section 4 discusses some other cases

for which additional network model assumptions are considered and more accurate parametric estimates of P are available.

The identifiability condition (5) suggests a natural procedure for estimating parameters in model (4) via subspace projections. However, the fact that α and $X\beta$ need not be orthogonal complicates the estimation procedure and its analysis. To highlight the main idea, we first describe the population-level estimation, assuming that both P and $\mathbb{E}Y$ are known.

Let $Z \in \mathbb{R}^{n\times p}$ be a matrix whose columns form an orthonormal basis of the covariate subspace $\mathrm{col}(X)$. Similarly, let $W \in \mathbb{R}^{n\times K}$ be the matrix whose columns are eigenvectors of P that span the subspace $S_K(P)$. Define the singular value decomposition (SVD) of matrix Z^TW to be $Z^TW = U\Sigma V^T$. Here, $U \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{K\times K}$ are orthonormal matrices of singular vectors while $\Sigma \in \mathbb{R}^{p\times K}$ is the matrix with the following singular values on the main diagonal:

$$\sigma_1 = \sigma_2 = \cdots = \sigma_r = 1 > \sigma_{r+1} \geq \cdots \geq \sigma_{r+s} > 0 = \sigma_{r+s+1} = \cdots = 0. \tag{7}$$

Thus, r is the dimension of $\mathcal{R} = \mathrm{col}(X) \cap S_K(P)$ and Z^TW has $r+s$ non-zero singular values. Column vectors of the matrices

$$\tilde{Z} = ZU \text{ and } \tilde{W} = WV, \tag{8}$$

also form a basis of $\mathrm{col}(X)$ and $S_K(P)$, respectively. In particular, the first r column vectors of $\tilde{Z}$ and $\tilde{W}$ coincide and form a basis of $\mathcal{R}$. Moreover, the last $p-r$ column vectors of $\tilde{Z}$ form a basis of the subspace of $\mathrm{col}(X)$ that is perpendicular to $\mathcal{R}$; similarly, the last $K-r$ column vectors of $\tilde{W}$ form a basis of the subspace of $S_K(P)$ that is perpendicular to $\mathcal{R}$. The model assumption (5) then indicates

$$\xi \in \mathrm{col}(\tilde{Z}_{1:r}) = \mathrm{col}(\tilde{W}_{1:r}) = \mathcal{R}, \quad X\beta \in \mathrm{col}(\tilde{Z}_{(r+1):p}) \perp \mathcal{R}, \quad \alpha \in \mathrm{col}(\tilde{W}_{(r+1):K}) \perp \mathcal{R}. \tag{9}$$

Therefore, ξ can be recovered by $\xi = \mathcal{P}_{\mathcal{R}}\mathbb{E}Y$, where $\mathcal{P}_{\mathcal{R}}$ is the orthogonal projection onto $\mathcal{R}$ written as

$$\mathcal{P}_{\mathcal{R}} = \tilde{Z}_{1:r}\tilde{Z}_{1:r}^T = \tilde{W}_{1:r}\tilde{W}_{1:r}^T. \tag{10}$$

For the convenience of theoretical analysis later, we also introduce the projection coefficient $\theta \in \mathbb{R}^p$

$$\theta = (X^TX)^{-1}X^T\xi.$$

Recall that $\xi \in \mathcal{R} \subset \mathrm{col}(X)$, so we have $\xi = X\theta$, and we have the one-to-one correspondence between θ and ξ. However, notice that θ cannot be attributed to conditional effects of covariates, as explained in the previous section.

For estimating α and β, we project $\mathbb{E}Y$ on $\mathrm{col}(\tilde{Z}_{(r+1):p})$ and $\mathrm{col}(\tilde{W}_{(r+1):K})$, respectively. Since these subspaces need not be orthogonal to each other (the principle angles between the two subspaces are the singular values $\sigma_{r+1}, \ldots, \sigma_{r+s}$), the corresponding projections are not necessarily orthogonal projections. Instead, they admit the following forms:

$$\mathcal{P}_{\mathcal{C}} = \left(\tilde{Z}_{(r+1):p}, 0_{n\times(K-r)}\right)(M^TM)^{-1}M^T, \tag{11}$$

$$\mathcal{P}_{\mathcal{N}} = \left(0_{n\times(p-r)}, \tilde{W}_{(r+1):K}\right)(M^TM)^{-1}M^T, \tag{12}$$

where $M = (\tilde{Z}_{(r+1):p}, \tilde{W}_{(r+1):K})$. Both α and β can then be recovered by

$$\alpha = \mathcal{P}_{\mathcal{N}}\mathbb{E}Y \quad \text{and} \quad \beta = (X^TX)^{-1}X^T\mathcal{P}_{\mathcal{C}}\mathbb{E}Y. \tag{13}$$

Finally, to test the hypothesis $H_0 : \alpha = 0$, we can use $\|\alpha\|^2$ as the statistic. We have $\alpha = \sum_{i=1}^{K-r}\tilde{W}_{r+i}\gamma_i = \tilde{W}_{(r+1):K}\gamma$, such that $\|\alpha\| = \|\gamma\|$. Although γ cannot be uniquely determined because $\tilde{W}_{(r+1):K}$ is only unique up to an orthogonal transformation, we can still uniquely identify its magnitude $\|\gamma\|^2$ to perform a chi-squared test against the null hypothesis $H_0 : \alpha = 0$.

In practice, when P and $\mathbb{E}Y$ are not observed, it is natural to replace them everywhere by the available approximations $\hat{P}$ and Y in the procedure above. There is, however, one crucial issue that requires special attention: the plugin estimate $\mathrm{col}(X) \cap S_K(\hat{P})$ for $\mathcal{R} = \mathrm{col}(X) \cap S_K(P)$ is often a bad approximation of $\mathcal{R}$. This is because the intersection of subspaces is not robust with respect to small perturbations to the subspaces. For example, it is easy to see that in low-dimensional settings, when $\mathrm{col}(X) \cap S_K(P)$ is non-trivial, a small perturbation of P can easily make $\mathrm{col}(X) \cap S_K(\hat{P})$ a null space. Therefore, rather than using $\mathrm{col}(X) \cap S_K(\hat{P})$ to approximate $\mathcal{R}$, the projections in (10), (11) and (12) are directly approximated using the eigenvectors of $\hat{P}$ (this partially explains the detailed discussion of the population level estimation above). The whole estimation procedure is summarised in Algorithm 1. From here through Section 3, we will assume that the dimension r of $\mathcal{R}$ is known for simplicity. This is because, so far, A has been treated as fixed while a discussion of selecting r naturally involves a detailed analysis of the perturbation mechanism from P to A. In Section 4 when the perturbation model for A is introduced, we provide a simple method to select r with a theoretical guarantee (see Corollary 3).

Algorithm 1. (Spectral projection estimation). Given X, Y, $\hat{P}$, K and r.

1. Calculate an orthonormal basis of $\mathrm{col}(X)$ and forms $Z \in \mathbb{R}^{n\times p}$. Similarly, calculate K eigenvectors of $\hat{P}$ and form $\hat{W} \in \mathbb{R}^{n\times K}$.
2. Calculate the SVD

$$Z^T\hat{W} = \hat{U}\hat{\Sigma}\hat{V}^T, \tag{14}$$

and denote

$$\hat{Z} = Z\hat{U}, \quad \breve{W} = \hat{W}\hat{V}. \tag{15}$$

3. Let $\hat{M} = (\hat{Z}_{(r+1):p}, \breve{W}_{(r+1):K})$. Estimate $\mathcal{P}_R$, $\mathcal{P}_{\mathcal{C}}$ and $\mathcal{P}_{\mathcal{N}}$ by

$$\hat{\mathcal{P}}_{\mathcal{R}} = \hat{Z}_{1:r}\hat{Z}_{1:r}^T, \tag{16}$$

$$\hat{\mathcal{P}}_{\mathcal{C}} = \left(\hat{Z}_{(r+1):p}, 0_{n\times(K-r)}\right)(\hat{M}^T\hat{M})^{-1}\hat{M}^T, \tag{17}$$

$$\hat{\mathcal{P}}_{\mathcal{N}} = \left(0_{n\times(p-r)}, \breve{W}_{(r+1):K}\right)(\hat{M}^T\hat{M})^{-1}\hat{M}^T. \tag{18}$$

4. Estimate ξ, β and α by

$$\hat{\xi} = \hat{\mathcal{P}}_{\mathcal{R}}Y, \quad \hat{\beta} = (X^TX)^{-1}X^T\hat{\mathcal{P}}_{\mathcal{C}}Y, \quad \hat{\alpha} = \hat{\mathcal{P}}_{\mathcal{N}}Y. \tag{19}$$

And recover the projection coefficient of ξ to $\mathrm{col}(X)$ as $\hat{\theta} = (X^TX)^{-1}X^T\hat{\xi}$.

5. Let $\hat{H} = \hat{\mathcal{P}}_{\mathcal{R}} + \hat{\mathcal{P}}_C + \hat{\mathcal{P}}_{\mathcal{N}}$. Estimate the variance σ^2 of ϵ in (4) by

$$\hat{\sigma}^2 = \left\| Y - \hat{H}Y \right\|^2 / (n - p - K + r). \tag{20}$$

6. Estimate the covariance of $\hat{\beta}$ by

$$\hat{\sigma}^2 (X^T X)^{-1} X^T \hat{\mathcal{P}}_C \hat{\mathcal{P}}_C^T X (X^T X)^{-1}. \tag{21}$$

   The inference of β can be done according to Theorem 1.
7. To test the significance of α, estimate γ by

$$\hat{\gamma} = \breve{W}_{(r+1):K} \hat{\alpha}, \tag{22}$$

   and estimate the covariance matrix of $\hat{\gamma}$ by $\hat{\Sigma}_{\hat{\gamma}} = \hat{\sigma}^2 \breve{W}_{(r+1):K} \hat{\mathcal{P}}_{\mathcal{N}} \hat{\mathcal{P}}_{\mathcal{N}}^T \breve{W}_{(r+1):K}^T$. Normalise $\hat{\gamma}$ to obtain $\hat{\gamma}_0 = \hat{\Sigma}_{\hat{\gamma}}^{-1/2} \hat{\gamma}$. Use $\|\hat{\gamma}_0\|^2$ for a chi-squared test (with K degrees of freedom) against the null hypothesis $H_0 : \alpha = 0$ according to Theorem 2.

3 | GENERIC INFERENCE THEORY OF MODEL PARAMETERS

For valid inference of parameters, we make the following standard assumptions about the data.

Assumption 1 (Standardised scale). All columns of X satisfy $\|X_i\| = \sqrt{n}$. Moreover,

$$\|\mathbb{E}Y\| \le C\sqrt{np},$$

for some constant $C > 0$.

Assumption 2 (Weak dependence of X). Denote $\Theta = (X^T X/n)^{-1}$. Assume that Θ is well-conditioned, that is, there exists a constant $\rho > 0$ such that

$$\rho \le \lambda_{\min}(\Theta) \le \lambda_{\max}(\Theta) \le 1/\rho.$$

Due to the deviation of $\hat{P}$ from the true signal P, the estimators would be biassed. A common approach to ensure the valid inference is to control the bias levels of the parameter estimates (Javanmard & Montanari, 2014; Van de Geer et al., 2014; Zhang & Zhang, 2014). In the current context, we need to guarantee that the biases are negligible compared to the standard deviations of the estimators. This requires $\hat{P} - P$ to be controlled to some extent. In particular, the biases will depend on the magnitude of the perturbation of $S_K(P)$ associated with col(X), defined by

$$\tau_n = \|Z^T \hat{W}\hat{W}^T - Z^T W W^T\| = \|(\hat{W}\hat{W}^T - WW^T)Z\|. \tag{23}$$

As a warm-up, it is not difficult to get a bound on $\|\hat{\Sigma} - \Sigma\|$ in terms of τ_n. Indeed, from the previous discussion, the singular decompositions of $Z^T W W^T$ and $Z^T \hat{W}\hat{W}^T$ are

$$Z^T W W^T = U \Sigma \tilde{W}^T, \quad Z^T \hat{W}\hat{W}^T = \hat{U}\hat{\Sigma}\breve{W}^T. \tag{24}$$

Therefore, by Weyl's inequality,

$$\|\hat{\Sigma} - \Sigma\| \leq \tau_n. \tag{25}$$

Since the estimates depend crucially on the SVD of $Z^T W W^T$ and $Z^T \hat{W} \hat{W}^T$, (23), (24) and (25) play a central role in establishing the theoretical results. We now introduce our assumption on the error level that ensures the validity of the inference. It can be seen as a way to characterise the level of perturbation our framework could tolerate. This validity and theoretical insights about this assumption will be studied in more detail later (see Section 4). As we show there, the condition is mild in many commonly studied network settings.

Assumption 3 (Small projection perturbation). Let Z and W be the matrices formed by the bases of $\mathrm{col}(X)$ and $S_K(P)$, respectively. Let $Z^T W = U \Sigma V^T$ be the SVD with singular values specified in (7). Assume $\hat{P}$ is a small projection perturbation of P with respect to X, by which we mean

$$\frac{\tau_n}{\min\left\{(1-\sigma_{r+1})^3, \sigma_{r+s}^3\right\}} = o\left(1/\sqrt{np}\right).$$

The first property to be introduced is the estimation consistency of $\hat{\sigma}^2$, which serves as the critical building block for later inference.

Proposition 2 (Consistency of variance estimation). *If Assumption 1 holds and*

$$40\tau_n \leq \min\{(1-\sigma_{r+1})^2, \sigma_{r+s}^2\}. \tag{26}$$

Then with high probability, the estimate $\hat{\sigma}^2$ defined by (20) satisfies

$$|\hat{\sigma}^2 - \sigma^2| \leq \frac{C\tau_n}{\min\{(1-\sigma_{r+1})^3, \sigma_{r+s}^3\}} \cdot \frac{n(\sigma^2+p)}{n-p-K+r},$$

for some constant $C > 0$. In particular, if Assumption 3 holds and $p + K = o(n)$ then $\hat{\sigma}^2$ is consistent.

The covariate effects β will be the central target for inference. The first result is the following bound on the bias of $\hat{\beta}$, defined to be $\left\|\mathbb{E}(\hat{\beta}) - \beta\right\|_\infty$.

Proposition 3 (The bias of $\hat{\beta}$). *If condition (26) holds then there exists a constant $C > 0$ such that*

$$\left\|\mathbb{E}(\hat{\beta}) - \beta\right\| \leq \frac{C\tau_n}{\min\{(1-\sigma_{r+1})^3, \sigma_{r+s}^3\}} \cdot \|(X^T X)^{-1}\| \cdot \|X\| \cdot \|X\beta + X\theta + \alpha\|.$$

In particular, if Assumptions 1, 2 and 3 hold then the bias of $\hat{\beta}$ is of order $o(1/\sqrt{n})$.

In general, one may be interested in inferring the contrast $\omega^T \beta$ for a given unit vector ω, such as in comparing covariate effects or making predictions. Let $\tilde{X} = X/\sqrt{n}$, then from (19),

$$\mathrm{Var}(\omega^T \hat{\beta}) = \frac{\sigma^2}{n} \omega^T \Theta \tilde{X}^T \hat{P}_C \hat{P}_C^T \tilde{X} \Theta \omega, \tag{27}$$

with $\hat{\mathcal{P}}_C$ defined by (17). According to Proposition 3, a sufficient condition for valid inference of $\omega^T\beta$ is $\omega^T\Theta\tilde{X}^T\hat{\mathcal{P}}_C\hat{\mathcal{P}}_C^T\tilde{X}\Theta\omega \geq c$ for some constant $c > 0$. Note that this quantity is directly computable in practice. However, we will focus on the population condition with respect to the true matrix $\mathcal{P}_C$ in the following result.

Theorem 1 (Asymptotic distribution of $\omega^T\hat{\beta}$). *If Assumptions 1, 2, 3 hold and*

$$\sqrt{np} \cdot \min\{1 - \sigma_{r+1}, \sigma_{r+s}\} \geq 1. \tag{28}$$

For a given unit vector $\omega \in \mathbb{R}^p$, assume that

$$\|\tilde{Z}_{(r+1):p}^T\tilde{X}\Theta\omega\| \geq c, \tag{29}$$

for some constant $c > 0$ and sufficiently large n. Then as $n \to \infty$,

$$\mathbb{P}\left(\frac{\omega^T\hat{\beta} - \omega^T\beta}{\sqrt{\frac{\hat{\sigma}^2}{n}\omega^T\Theta\tilde{X}^T\hat{\mathcal{P}}_C\hat{\mathcal{P}}_C^T\tilde{X}\Theta\omega}} \leq t\right) \to \Phi(t),$$

where Φ is the cumulative distribution function of the standard normal distribution.

Next, we discuss the special case when $\omega = e_j$, which corresponds to making inference for the individual parameter $\beta_j = e_j^T\beta$ for some $1 \leq j \leq p$. In this case, condition (29) has a simple interpretation.

Corollary 1 (Valid inference of individual regression coefficient). *Let η_j be the partial residual from regressing $\tilde{X}_j$ against all other columns of $\tilde{X}$ and R_j be the corresponding partial R^2 for this regression. If Assumption 1, 2, 3 hold and there exists a constant $c > 0$ such that*

$$\|\tilde{Z}_{(r+1):p}^T\eta_j\|/(1 - R_j^2) \geq c. \tag{30}$$

Then, as $n \to \infty$,

$$\mathbb{P}\left(\frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{\hat{\sigma}^2}{n}e_j^T\Theta\tilde{X}^T\hat{\mathcal{P}}_C\hat{\mathcal{P}}_C^T\tilde{X}\Theta e_j}} \leq t\right) \to \Phi(t),$$

where Φ is the cumulative distribution function of the standard normal distribution.

To understand condition (30), notice that η_j is the signal from X_j, after adjusting for the other covariates, while $1 - R_j^2$ is the proportion of this additional information from $\tilde{X}_j$. Since $\|\tilde{Z}_{(r+1):p}^T\eta_j\|$ is the magnitude of η_j's projection on the subspace $\mathrm{col}(\tilde{Z}_{(r+1):p})$, (30) essentially requires that the additional information contributed by X_j after conditioning on other covariates should nontrivially align with $\mathrm{col}(\tilde{Z}_{(r+1):p})$, the subspace on which the effect of β_j lies. This type of condition is needed because if $\tilde{Z}_{(r+1):p}^T\eta_j = 0$, the parameter space of β_j degenerates to $\{0\}$ and the inference is not meaningful. Notice that requirement (30) is completely different from assuming that $|\beta_j|$ is large, and it does allow $|\beta_j|$ to be zero or small.

The next result shows the consistency of estimating α. It also provides bounds for the bias and variance of $\hat{\gamma}$, which are later used for testing the hypothesis $H_0 : \alpha = 0$.

Proposition 4 (Consistency of $\hat{\alpha}$ and $\hat{\gamma}$). *If Assumption 1 and (26) hold. Then there exists a constant $C > 0$ such that with high probability, the following hold.*

1. *Let $\|W\|_{2\to\infty}$ be the maximum of the Euclidean norms of the row vectors of W, then*

$$\|\hat{\alpha} - \alpha\|_\infty \leq \frac{C}{\min\{(1-\sigma_{r+1})^3, \sigma_{r+s}^3\}} \cdot \left(\|W\|_{2\to\infty} \sqrt{K\sigma^2} \log n + \tau_n \sqrt{n(p+\sigma^2)} \right).$$

2. *There exists an orthogonal matrix $O \in \mathbb{R}^{(K-r)\times(K-r)}$ such that*

$$\|\mathbb{E}\hat{\gamma} - O\gamma\| \leq \frac{C\tau_n\sqrt{np}}{\min\{(1-\sigma_{r+1})^3, \sigma_{r+s}^3\}}.$$

3. *Let $\Sigma_{\hat{\gamma}}$ be the covariance matrix of $\hat{\gamma}$, then*

$$\left\| \Sigma_{\hat{\gamma}} - \sigma^2 \begin{pmatrix} (I_s - \Gamma^2)^{-1} & 0 \\ 0 & I_{K-r-s} \end{pmatrix} \right\| \leq \frac{C\tau_n}{\min\{(1-\sigma_{r+1})^2, \sigma_{r+s}^2\}},$$

where $\Gamma = \mathrm{diag}(\sigma_{r+1}, \cdots, \sigma_{r+s})$.

Proposition 4 shows that $\hat{\alpha}$ is an element-wise consistent estimate of α if Assumption 3 holds, $\min\{(1-\sigma_{r+1})^3, \sigma_{r+s}^3\}$ is bounded away from zero, and $\|W\|_{2\to\infty}\sqrt{K\sigma^2}\log n = o(1)$. The last requirement holds when P is an incoherent matrix — the rows of W are in similar magnitudes — a commonly observed property introduced by (Candès & Recht, 2009; Candès & Tao, 2010). The other two inequalities of Proposition 4 show that the bounds for both bias and variance of $\hat{\gamma}$ are of order $o(1)$ if Assumption 3 holds. These properties directly lead to the validity of the chi-squared test for network effects.

Theorem 2 (Chi-squared test for α). *If Assumptions 1 and 3 hold, K is a fixed integer, and $p = o(n)$. Then as $n \to \infty$, the statistic $\|\hat{\gamma}_0\|^2$ constructed in Algorithm 1 satisfies*

$$\|\hat{\gamma}_0\|^2 \xrightarrow{d} \chi_K^2,$$

where χ_K^2 is a random variable that follows the chi-squared distribution with K degrees of freedom and $\xrightarrow{d}$ denotes the convergence in distribution.

The next theorem provides the theoretical result for $\hat{\theta}$.

Theorem 3 (Inference of θ and ξ). *The estimator $\hat{\theta}$ in (19) follows a Gaussian distribution with covariance matrix $\mathrm{Cov}(\hat{\theta}) = \frac{\sigma^2}{n}\Theta\tilde{X}^T\hat{\mathcal{P}}_R\tilde{X}\Theta$. Furthermore, under Assumptions 1, 2 and 3, $\left\|\mathbb{E}(\hat{\theta}) - \theta\right\| = o(1/\sqrt{n})$. If in addition, there exists a constant $c > 0$ such that*

$$\|\tilde{Z}_{1:r}^T\eta_j\|/(1-R_j^2) \geq c, \tag{31}$$

where η_j and R_j are defined in Corollary 1, then as $n \to \infty$,

$$\mathbb{P}\left(\frac{\hat{\theta}_j - \theta_j}{\sqrt{\frac{\hat{\sigma}^2}{n} e_j^T \Theta \tilde{X}^T \hat{\mathcal{P}}_{\mathcal{R}} \tilde{X} \Theta e_j}} \leq t \right) \to \Phi(t).$$

The inference of each $\xi_i, i \in [n]$ can be done by noticing $\xi_i = X_{i\cdot}\theta$ where $X_{i\cdot}$ is the ith row of X.

Condition (31) is in parallel of (30); see the discussion following Corollary 1 for its interpretation.

4 | SMALL PROJECTION PERTURBATION UNDER RANDOM NETWORK MODELS

We have introduced the small projection perturbation condition as a general way to characterise the structural perturbation level our framework could tolerate. To provide more insights about this characterisation, in this section, we study the small projection perturbation assumption in a special setting when the perturbation of P comes from random network models. Specifically, we assume that the edge between each pair of nodes (i, j) is generated independently from a Bernoulli distribution with $\mathbb{P}(A_{ij} = 1) = P_{ij}, 1 \leq i < j \leq n$. In particular, $\mathbb{E}A = P$. This model is known as the inhomogeneous Erdős-Rényi model (Bollobas et al., 2007). To clarify, the randomness discussed in this section comes from the random network model above, which is assumed to be independent of the randomness from the noise ϵ in the regression model.

Section 4.1 investigates the most difficult setting when no prior information is available about the network model, while Section 4.2 presents the results in the arguably easiest setting when the true underlying network can be accurately estimated. Section A further extends the results in another commonly seen setting when the Laplacian matrix represents the relational information.

4.1 | The non-informative situation: small projection perturbation for adjacency matrices

In this section, we investigate the situation when no prior information about the network model is available. This can be seen as the most difficult, yet the most general, situation. Arguably, the only reasonable approximation of P is $\hat{P} = A$. We will study Assumption 3 for this version of $\hat{P}$.

Recall that the main requirement of Assumption 3 is $\tau_n = o(1/\sqrt{n})$, where τ_n measures the level of perturbation of the network subspace $S_K(P)$ associated with the covariate subspace col(X), defined in (23). Existing results relevant for controlling τ_n, such as (Abbe et al., 2017; Cape et al., 2019; Lei, 2019; Mao et al., 2020), do not give sufficiently tight bounds to support the inference, even for dense networks. The recent work of Xia (2021) contains useful tools for obtaining such error bound that allows sparsity when P is a low-rank matrix. However, since in general problems, P may be of full rank, a new tool is needed for theoretical analysis. The following assumption is made about P.

Assumption 4 (Eigenvalue gap of the expected adjacency matrix). Let A be the adjacency matrix of a random network generated from the inhomogeneous Erdős-Rényi model with the edge

probability matrix $P = \mathbb{E}A$. Assume that the K largest eigenvalues of P are well separated from the remaining eigenvalues and their range is not too large:

$$\min_{i\le K,\ i'>K} |\lambda_i - \lambda_{i'}| \ge \rho' d, \qquad \max_{i,i'\le K} |\lambda_i - \lambda_{i'}| \le d/\rho',$$

where $\rho' > 0$ is a constant and $d = n \cdot \max_{ij} P_{ij}$.

The quantity d can be seen as an upper bound of the network node degree, which has been widely used to measure network density (e.g., (Le et al., 2017; Lei & Rinaldo, 2014)). Assumption 4 is very general in the sense that it only assumes that P has a sufficiently large eigenvalue gap, but P need not be low-rank or even approximately low-rank. It appears that this is already sufficient to guarantee the small projection perturbation property of A, as stated in the next theorem. We believe the theorem itself can be used as a general tool for statistical analysis of independent interest.

Theorem 4 (Concentration of perturbed projection for adjacency matrix). *Let $w_1, \dots, w_n$ and $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ be eigenvectors and corresponding eigenvalues of P and similarly, let $\hat{w}_1, \dots, \hat{w}_n$ and $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_n$ be the eigenvectors and eigenvalues of A. Denote $W = (w_1, \dots, w_K)$ and $\hat{W} = (\hat{w}_1, \dots, \hat{w}_K)$. Suppose that Assumption 4 holds and $d \ge C \log n$ for a sufficiently large constant $C > 0$. Then for any fixed unit vector v, with high probability,*

$$\|(\hat{W}\hat{W}^T - WW^T)v\| \le \frac{2\sqrt{K \log n}}{d}. \tag{32}$$

Notice that the bound is for a given deterministic vector v instead of all unit vectors. The latter would be equivalent to a bound on $\left\|\hat{W}\hat{W}^T - WW^T\right\|$ and is too large for our inference purpose. By restricting the scope of applicability, our result trades off for the tighter bound (32), which is crucial for our inference to work in relatively sparse network settings.

Corollary 2 (Small projection perturbation, adjacency matrix case). *Assume that $\hat{P} = A$ is used in Algorithm 1. If Assumption 4 holds and d satisfies*

$$\frac{d \cdot \min\left\{(1-\sigma_{r+1})^3, \sigma_{r+s}^3\right\}}{\left(Knp \log n\right)^{1/2}} \to \infty,$$

then Assumption 3 holds with high probability for sufficiently large n.

Corollary 2 provides a sufficient condition for valid inference. If $\min\{1-\sigma_{r+1}, \sigma_{r+s}\}$ is bounded away from zero while K and p are fixed then d needs to grow faster than $\sqrt{n \log n}$. This is arguably a strong assumption, although it does allow for moderately sparse networks. A natural question is whether it can be relaxed. Unfortunately, the answer to this is negative. We now show that under the current assumptions, the bound (32) is rate optimal up to a logarithm factor and we cannot guarantee valid inference if d is of the order $\sqrt{n}$.

Theorem 5 (Tightness of concentration and degree requirements). *Assume $C\sqrt{n/\log n} \le d \le n^{1-\xi}$ for some constant $\xi \in (0, 1/2]$ and some sufficiently large constant $C > 0$. The following statements hold.*

1. *There exists a configuration of* (K, v, P) *satisfying the condition of Theorem 4 with* $K = O(1)$, *under which, for a sufficiently large* n,

$$\left\|(\hat{W}\hat{W}^T - WW^T)v\right\| \geq c/d,$$

holds with high probability and a constant $c > 0$.

2. *There exists a configuration of* $(K, p, X, P, \beta, \theta, \alpha)$ *satisfying Assumptions 1,2 and the conditions of Theorem 4 with* $K, p = O(1)$ *and* $\min\{1 - \sigma_{r+1}, \sigma_{r+s}\} \geq 1/4$, *under which, for a sufficiently large* n, *we have*

$$\left\|\mathbb{E}\hat{\beta} - \beta\right\|_\infty \geq c'/d \text{ and } \max_j Var(\hat{\beta}_j) \leq C'/n,$$

with high probability for some constant $C', c' > 0$.

Notice that the two statements of Theorem 5 are different and, in general, neither implies the other. The first statement is about the general concentration bound of Theorem 4. It indicates that for the given range of d, the bound of (32) is rate optimal (up to a logarithm order). The second statement is about the necessary condition of the inference problem, under the model-free setting. When $d = \sqrt{n}$, the bias is at least in the order of $1/\sqrt{n}$. In contrast, the SD is at most of order $1/\sqrt{n}$. Therefore, we will not be able to give asymptotically correct testing or confidence intervals under these circumstances. Meanwhile, under the same condition, Corollary 2 shows that if d grows faster than $\sqrt{n \log n}$, valid inference could be achieved. In combination, we observe a phase transition at the network degree of $\sqrt{n}$.

To conclude this section, we address the problem of selecting the correct dimension r, which is used in Algorithm 1, as another application of Theorem 4. Specifically, denote $\hat{d} = \frac{1}{n}\sum_{i,j=1}^n A_{ij}$ and $\bar{d} = \frac{1}{n}\sum_{i,j=1}^n P_{ij}$. The following rule is used to select r:

$$\hat{r} = \max\left\{ i : \hat{\sigma}_i \geq 1 - \frac{4\sqrt{pK \log n}}{\hat{d}} \right\}. \tag{33}$$

This estimate can be shown to give the correct dimension with high probability.

Corollary 3 (Consistency of estimating the dimension of $\mathcal{R}$). *Assume that the conditions of Theorem 4 hold and*

$$\bar{d} \geq \frac{16\sqrt{pK \log n}}{1 - \sigma_{r+1}}. \tag{34}$$

Then for a sufficiently large n, *the dimension estimate* $\hat{r}$ *defined in* (33) *satisfies* $\hat{r} = r$ *with high probability*.

4.2 | The informative situation: small projection perturbation for parametric models

In the previous section, it is shown that if A is used as $\hat{P}$ in Algorithm 1, the network degree needs to grow faster than $\sqrt{n}$ for the small perturbation assumption to hold. In many cases, it may be

reasonable to assume that P satisfies certain structural conditions. Leveraging such additional information can provide a more accurate estimate of P and, in turn, ensure the validity of inference for potentially much sparser networks. In this section, we investigate a few special cases for which the edge density requirement can be substantially relaxed, highlighting the benefit of knowing the correct model.

The discussion begins with the stochastic block model (SBM) as the true network generating model. Specifically, assume that there exists a vector $g \in [K]^n$ specifying the community node labels and a symmetric matrix $B \in [0,1]^{K\times K}$ such that $P_{ij} = B_{g_i g_j}$. The SBM has been widely used to model community structures (Holland et al., 1983), and its theoretical properties have been well-understood thanks to intensive research in recent years. For more detail, refer to the review paper of Abbe (2018) and the references therein. Since the SBM is used in this section only to illustrate how the degree requirement can be relaxed, we will focus on the following special configuration.

Assumption 5 (Stochastic block model). Let n_k be the number of nodes in the k th community. Assume that $(1-t)n/K \le n_k \le (1+t)n/K$ for some constant t and all $1 \le k \le K$. Moreover, assume that $B = \kappa_n B_0$, where $B_0 \in [0,1]^{K\times K}$ is a fixed matrix and κ_n controls the dependence of network density on n and $K = o(n^{1/4})$.

Another example is the degree-corrected block model (DCBM) proposed by Karrer and Newman (2011), which generalises the SBM by allowing the degree heterogeneity of the nodes. Specifically, in addition to the SBM parameters, the model has a degree parameter $\nu \in \mathbb{R}^n$ such that $P_{ij} = \nu_i \nu_j B_{g_i g_j}$. Notice that ν_i is only identifiable up to a scale so additional constraints are needed. The following simplified assumption is made about it.

Assumption 6 (Identifiability of degree parameter). Assume $\sum_{g_i=k} \nu_i = n_k$ and there exists a constant ζ such that $1/\zeta \le \nu_i \le \zeta$.

Notice that under Assumptions 5 and 6, both the SBM and DCBM satisfy Assumption 4 with $d = n\kappa_n$. In particular, under the SBM, the subspace of the K leading eigenvectors of P has the same component for all nodes within the same community. Therefore, the model under the SBM reduces to a linear regression model with fixed group effects according to communities. The following assumption is further made for the two models.

Assumption 7 (Exact recovery of community labels). Assume the community g is known with $n\kappa_n / \log n \to \infty$.

Although g is seldom known in practice, this assumption is based on the fact that in many cases g can be exactly recovered from the network. Indeed, there exist many polynomial-time algorithms that ensure the exact recovery of g such as (Abbe et al., 2017; Gao et al., 2017; Lei et al., 2020; Lei & Zhu, 2017; Li et al., 2020a) for the SBM and (Chen et al., 2018; Gao et al., 2018; Lei & Zhu, 2017) for the DCBM, to name a few. The corresponding regularity condition needed for such strong consistency of community detection is already reflected in the degree requirement in Assumption 7. Given the community labels, the commonly used estimator of P would be the MLE (see Section G of Appendix S1 for details). When such parametric estimates are used as $\hat{P}$ in Algorithm 1, the corresponding estimation and inference procedure are referred to as the *parametric version* of our method, in contrast to the model-free version when A is used as $\hat{P}$. The following theorem shows that the parametric version delivers valid inference under much weaker network density assumptions compared to its nonparametric counterpart.

Theorem 6 (Small projection perturbation under block models). *Let $d = n\kappa_n$. Let $\hat{P}$ be the parametric estimator of P introduced in Section G. Under either of the following settings*:

1. *A is generated from P according to the SBM satisfying Assumptions 5 and 7 with*

$$\min\left\{(1-\sigma_{r+1})^6, \sigma_{r+s}^6\right\} \cdot \frac{d}{p \log n} \to \infty;$$

2. *A is generated from P according to the DCBM satisfying Assumptions 5, 6 and 7 with*

$$\min\left\{(1-\sigma_{r+1})^6, \sigma_{r+s}^6\right\} \cdot \frac{d}{K^2 p \log n} \to \infty;$$

the estimator $\hat{P}$ satisfies Assumption 3 with high probability.

Again, if p, K and $\min\{(1-\sigma_{r+1}), \sigma_{r+s}\}$ are fixed, then the degree requirement of the parametric version for the small projection perturbation to hold is $d/\log n \to \infty$ for either of two block models. This is much weaker than the degree requirement for the non-parametric estimation, indicating the benefit of knowing the underlying model for P. Although the above result is only about two special classes of models for the network, it clearly shows the tradeoff between the model assumption and statistical efficiency: without any information about the true model, the model-free version of the proposed method requires the average degree to grow faster than $\sqrt{n}$ to be effective; in contrast, if the class that the true model belongs to is known then the parametric version of our method performs well for much sparser networks.

5 | SIMULATION

In this section, we present a simulation study to evaluate the proposed method and compare it with other benchmark methods for linear regression on networks. First, we evaluate the validity of the inference framework and demonstrate the predicted phase transition of the model-free method, as well as the advantage of the parametric version in the situation when the true network model is known. After that, we introduce comparisons with other benchmark methods under several network effect models.

5.1 | Inference validation

We will evaluate our theoretical claims about the inference and phase transition in this subsection. For this purpose, we will again assume K to be known to exactly match our theoretical framework in this subsection. However, this does not lose generality because in all of our experiment settings, because the methods we mentioned before (Chen & Lei, 2018; Le & Levina, 2022; Li et al., 2020b) can accurately identify the K almost perfectly in our settings. We fix $p = K = 4$ in all configurations. The network is generated by the SBM with $K = 4$ communities. In Section J of Appendix S1, we provide additional results and also a study when the network is generated by the more general DCBM. Using the parameterisation in Section 4.1, we set $B = 0.2\mathbf{1}_4\mathbf{1}_4^T + 0.8I_4$, where $\mathbf{1}_k$ denotes the all-one vector of length k. Three levels of sample size $n = 300, 500, 1000, 2000, 4000$, and three levels of average expected degree $\varphi_n = 2\log n, \sqrt{n}$, and $n^{2/3}$ for each level of n are

compared. Given the eigenvectors $w_1, ..., w_n$ from P, we generate $X \in \mathbb{R}^{n\times p}$ in the following way: Set $X_1/\sqrt{n} = w_1$; Set $X_j/\sqrt{n} = \sqrt{\frac{1}{25}}w_j + \sqrt{\frac{24}{25}}w_{j+3}, j = 2, 3, 4$. This configuration gives a design with $r = 1, s = 3$ and $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (1, 0.2, 0.2, 0.2)$. In particular, the setup with $\beta = (0, 1, 1, 1)^T$ and $\theta = (1, 0, 0, 0)^T$ satisfies the proposed model and is fixed in all settings. Similarly, the γ can be any vector with the first coordinate being zero. We consider two cases: $\gamma = (0, 1, 1, 1)^T$ and $\gamma = (0, 0, 0, 0)^T$. As discussed, the case of $\gamma = (0, 0, 0, 0)^T$ indicates that there is no network effect. The two settings do not give any difference on inference of β while the latter is used to verify the validity of the χ^2 test. Also, notice that X_1 violates assumption (30) because it is completely orthogonal to the space of col($\hat{Z}_{r+1:p}$). Therefore, valid inference cannot be made for β_1. This set-up highlights that having a degenerate parameter space on β_1 does not impact the validity of our inference on the other parameters $\beta_2, \beta_3, \beta_4$, as shown by both the theoretical and numerical analyses. Both the model-free and parametric versions of the subspace projection (SP) methods are used to demonstrate the empirical difference, denoted by 'SP' and 'SP-SBM', respectively. All results are taken as the average of the 50 independent replications.

Table 2 shows the ratio between the absolute bias and standard deviation (SD) averaged across $\hat{\beta}_2, \hat{\beta}_3$ and $\hat{\beta}_4$ for different n and network densities, when the network is generated from the SBM. For average degree of order $\sqrt{n}$ or below, the ratio does not vanish with increasing n for the SP, indicating that valid statistical inference for the parameters cannot be achieved. For denser networks, the bias becomes vanishing; thus, the asymptotic inference becomes valid. These observations coincide with the prediction of the phase transition at degree $\sqrt{n}$. In contrast, the SP-SBM results in much smaller bias-SD ratios and achieves vanishing ratio even at degree $\sqrt{n}$. At the borderline case $2\log n$, such vanishing pattern is not clearly observed, even for the SP-SBM.

Correspondingly, Table 2 also shows the resulting average coverage probabilities p_{cov} of the 95% confidence intervals for β_2, β_3 and β_4. Notice that, differently from Table 2, the coverage probability is calculated as the average of 500 Monte Carlo runs, thereby taking the errors of $\hat{\sigma}$ into account. It can be seen that the confidence interval for the SP is not good enough for degree of

TABLE 2 Average bias-SD ratios for β_2, β_3 and β_4 when the network is generated from the SBM

		Bias-SD ratio Average expected degree			**Coverage probability** Average expected degree		
Method	n	$2\log n$	$\sqrt{n}$	$n^{2/3}$	$2\log n$	$\sqrt{n}$	$n^{2/3}$
SP	300	1.497	0.796	0.393	0.813	0.908	0.941
	500	1.310	0.720	0.255	0.845	0.910	0.949
	1000	1.929	0.791	0.232	0.717	0.888	0.948
	2000	2.273	0.723	0.254	0.585	0.891	0.943
	4000	2.818	0.790	0.213	0.554	0.884	0.949
SP-SBM	300	1.257	0.281	$<10^{-4}$	0.844	0.945	0.950
	500	0.826	0.157	$<10^{-4}$	0.911	0.949	0.949
	1000	0.887	0.012	$<10^{-4}$	0.895	0.949	0.950
	2000	0.934	$<10^{-4}$	$<10^{-4}$	0.856	0.950	0.950
	4000	0.898	$<10^{-4}$	$<10^{-4}$	0.838	0.950	0.951

Abbreviations: SP, subspace projection; SP-SBM, SP-stochastic block model.

$\sqrt{n}$ or lower, but does give valid results for denser graphs. The parametric version is always more accurate than the model-free version, and in particular, remains valid for the order of $\sqrt{n}$.

5.2 | Comparison with other benchmarks

In this subsection, we will evaluate the practical performance of our method in different settings of network effects and compare it with other benchmark methods. The first benchmark in the comparison is the ordinary least squares method (OLS). The other two are the SIM of Manski (1993) and the RNC of Li et al. (2019) introduced before. It has been shown that the RNC is a better modelling option than both OLS and the SIM (Li et al., 2019). However, the RNC method cannot make inference of the parameters, and it is computationally more intensive than our method. In this experiment, the RNC is always tuned by 10-fold cross-validation. For our method, the number K is determined by the method of Le and Levina (2022). The theory-driven tuning of r introduced in (33) is based on large-sample properties and may be conservative for small samples. We also propose a bootstrapping method in Section I of Appendix S1. This method is used to select r for the numerical and data examples from now on and works very well. Lastly, a systematic model fitting procedure is used for SP to take advantage of its available inference framework. Specifically, the SP methods would check the χ^2 test result of γ. If the p-value for the χ^2 test is more significant than 0.05 (any other reasonable level can be chosen), the method would instead return OLS fit, which is equivalent to constraining $\gamma = 0$ in our estimation procedure.

As before, the network is generated by the SBM with $K = 4$ with the same configurations. We take $n = 300$ and $n = 1000$, as representative cases for small-sample and large-sample settings. The results for the more general situation of DCBM can be found in Section J of Appendix S1. However, to avoid overly tailoring the setup for our methods, X's are generated differently. Specifically, X_2, X_3, X_4 are randomly generated from Gaussian distribution $N(0, 1)$, uniform distribution $U(0, 1)$ and exponential distribution $Exp(1)$ before standardisation is applied. X_1 is directly set to be $\sqrt{n}w_1$ as before, where w_1 is the first eigenvector of P. Meanwhile, three different schemes are used to generate the individual effect α, so that the model misspecification situations can also be tested. In particular, the first scenario is exactly the same setting as in the last section, which corresponds to the assumed model. This individual effect setting is called 'eigenspace'. In the second scenario, we set $\gamma = 0$, thus the model becomes the standard linear regression model for which OLS is designed. In the last scenario, we set α to be the average of the three eigenvectors corresponding to the three smallest non-zero eigenvalues of the observed Laplacian matrix L. This α gives a small value of $\sum_{i\sim j}(\alpha_i - \alpha_j)^2$, and thus matches the assumption of the RNC framework (Li et al., 2019). As discussed in Appendix S1 (Section A), our method should still work by using $\hat{P} = L$ when α is defined in this way. Therefore, for the evaluation in this section, the Laplacian version of the SP estimator is also included and labelled as 'SP-L'.

Since the data generating models have different parameterisations, we directly compare the relative mean squared error (MSE) of the expected response $\mathbb{E}Y$. All of the results are averaged over 50 independent replications and are shown in Table 3. The overall pattern remains the same for small-sample and large-sample problems. Under the model with eigenspace individual effects (the proposed model), it can be seen that OLS never renders competitive performance while the SIM is only slightly better than OLS. The RNC gives much better results than OLS and the SIM because it effectively incorporates the network information. However, since the network is noisy, its performance is far from adequate. Both versions of our method (SP and SP-SBM) significantly outperform the other methods, with the parametric version performing better than

TABLE 3 Mean squared error (MSE) $\times 10^2$ of $\mathbb{E}Y$ when the network is generated from the SBM with $p = 4, K = 4$ with three types of individual effects, in small sample case ($n = 300$) and large sample case ($n = 1000$), respectively

n	Individual effects	Average degree	SP	SP-SBM	OLS	SIM	RNC	SP-L	SP-SBM-L
300	Eigenspace	$2\log n$	18.20	13.59	40.65	246.05	12.11	40.40	27.13
		$\sqrt{n}$	11.13	2.36	40.65	40.16	11.57	39.00	21.50
		$n^{2/3}$	3.66	0.30	40.65	36.35	9.53	22.93	20.47
	$\gamma = 0$	$2\log n$	0.44	0.43	0.37	0.70	1.11	0.43	0.42
		$\sqrt{n}$	0.42	0.42	0.37	0.68	1.73	0.38	0.44
		$n^{2/3}$	0.42	0.41	0.37	0.69	2.31	0.40	0.42
	Smooth	$2\log n$	25.39	25.40	25.36	189.83	17.83	16.98	25.38
		$\sqrt{n}$	15.67	14.81	27.09	25.82	23.16	13.87	20.94
		$n^{2/3}$	2.55	2.97	24.07	23.24	16.61	12.30	13.52
1000	Eigenspace	$2\log n$	13.60	8.67	38.47	38.68	12.07	39.38	23.57
		$\sqrt{n}$	5.24	0.24	38.47	37.90	10.52	23.13	19.35
		$n^{2/3}$	1.44	0.11	38.47	35.82	7.83	20.16	19.29
	$\gamma = 0$	$2\log n$	0.14	0.14	0.13	0.43	0.56	0.13	0.14
		$\sqrt{n}$	0.14	0.14	0.13	0.22	0.80	0.14	0.13
		$n^{2/3}$	0.14	0.14	0.13	0.24	2.19	0.13	0.14
	Smooth	$2\log n$	25.34	25.26	25.33	26.40	21.73	12.75	25.29
		$\sqrt{n}$	14.69	14.75	17.61	16.54	18.79	8.90	16.18
		$n^{2/3}$	1.16	1.23	3.30	2.55	11.95	1.74	2.26

Abbreviations: OLS, ordinary least squares method; RNC, regression with network cohesion; SIM, social interaction model; SP, subspace projection; SP-SBM, SP-stochastic block model.

the model-free version, as expected. When there are no individual effects ($\gamma = 0$), OLS becomes the correct model and gives optimal results, as expected. The RNC and our methods all are correctly adaptive to this setting, but the SP methods are still better than the RNC. The performance under the model with smooth individual effects (the RNC model) is noisier as it depends on the perturbation of the network. Overall, neither OLS nor the SIM works. However, under the SBM with average degree $n^{2/3}$, the network is too dense, such that the smooth individual effects become almost identical everywhere, and OLS becomes effective. In this setting, using the Laplacian matrix in our method still models the correct form of individual effects, so SP-L still works and is more effective than the RNC. Note that the parametric versions (SP-SBM-L) are no longer better than the SP in this situation because the individual effects depend on the observed network instead of the population signal; thus, the parametric methods are not using the correct model.

In conclusion, when there are network effects, our method outperforms the other benchmark methods in the experiments. When there are no network effects, it is adaptive to the simpler model and gives a similar result to OLS. Compared to the model-free SP, the parametric SP method is less robust to model misspecification.

5.3 | Timing evaluation

The proposed method is generally computationally efficient. We include the timing evaluation in this section. All the computations are on a Linux system Intel(R) Xeon(R) CPU E5-2630 v3 2.40GHz CPU and 2G memory. Our implementation of the SP method is completely done in R by taking advantage of sparse matrices and the efficient partial eigen-decomposition algorithm in Qiu and Mei (2019). The SIM fitting is based on the two-stage least square method in the R package `spatialreg` (Bivand et al., 2013) and the RNC method is implemented in the R package `netcoh` (Li et al., 2016). In Table 4 we include the average computational time for model fitting in the same settings of Section 5.1.

The computational efficiency of our method depends on the network density. The computation is slightly faster on a sparse network, but the difference is marginal. This is because while sparse networks result in more efficient spectrum decomposition, the other part of model fitting involving X and the eigenvectors do not benefit from sparsity. Overall, the model fitting procedure takes about 1.3 s for a network with 4000 nodes. All the other methods are also reasonably fast. The SIM is similar to our method on sparse networks, but it becomes much slower when the network is denser. The RNC is generally the slowest, taking about 2.5 times that of the SP method.

Notice that the timing only accounts for modelling fitting without prior tuning because these prior procedures can be flexibly specified by users. Among the methods we recommend for selecting K in the SP method, the USVT of Chatterjee (2015) is the fastest while the Beth–Hession method of Le and Levina (2022) is slightly slower; both of them are cheaper than our model fitting

TABLE 4 Average timing (in seconds) of model fitting procedures in the settings of Table 2

n	Average degree	SP	OLS	SIM	RNC
300	$2\log n$	0.01	<0.01	0.01	0.01
	$\sqrt{n}$	0.01	<0.01	0.02	0.01
	$n^{2/3}$	0.01	<0.01	0.02	0.01
500	$2\log n$	0.01	<0.01	0.03	0.02
	$\sqrt{n}$	0.01	<0.01	0.03	0.02
	$n^{2/3}$	0.01	<0.01	0.05	0.02
1000	$2\log n$	0.04	<0.01	0.12	0.11
	$\sqrt{n}$	0.04	<0.01	0.09	0.10
	$n^{2/3}$	0.04	<0.01	0.20	0.09
2000	$2\log n$	0.19	<0.01	0.24	0.52
	$\sqrt{n}$	0.18	<0.01	0.29	0.51
	$n^{2/3}$	0.18	<0.01	0.79	0.51
4000	$2\log n$	1.30	<0.01	1.22	3.15
	$\sqrt{n}$	1.33	<0.01	1.43	3.10
	$n^{2/3}$	1.34	<0.01	3.94	3.12

Note: The SD is below 3% of the timing in all settings.

Abbreviations: OLS, ordinary least squares method; RNC, regression with network cohesion; SIM, social interaction model; SP, subspace projection.

procedure itself. The cross-validation method of Li et al. (2020b) is much more computationally intensive, but still remains feasible for moderately large networks. The r selection based on (33) does not introduce an additional cost. Alternatively, if the bootstrap is used to select r, the computational cost is upper bounded by that of the single model fitting in Table 4 times the bootstrap cost, if no distributed computing is involved. Similarly, for the RNC method, the cross-validation tuning time is not included, which depends on how many tuning parameter values are examined and the number of data-splitting for each value. Empirically, RNC needs cross-validation for a wide range of tuning parameter values so its computational disadvantage will be even more significant than Table 4.

6 | SCHOOL CONFLICT REDUCTION STUDY

In Paluck et al. (2016), an experiment is conducted to study the impacts of educational workshops on reducing conflicts in schools. The experimenters randomly selected 26 middle schools in New Jersey in which they held educational workshops. In each school, the experimenters first determined a group of eligible students. A small proportion of them was selected (randomly after gender and race blocking) to participate in bi-monthly educational workshops about school conflicts. Students were also asked to name their friends (with whom they spend time) in school. The friend nomination may not be symmetric in this case. Following the standard approach in previous studies (Bramoullé et al., 2009; Goldsmith-Pinkham & Imbens, 2013; Paluck et al., 2016), we ignore the directions of edges and treat two students as connected as long as either one identifies the other as a connection. This approach is widely applied with the assumption that as long as one side nominates the association, it reliably indicates that connections reasonably exist.

One interesting aspect of the data set is that the social network information is collected in two waves of surveys, one at the early stage of the school year and one towards the end of the year. The nominated edge sets are very different from the two surveys in all schools. On average, across all schools, 66% of edges only appear in one survey. Such a mismatch is commonly observed in social studies (Bramoullé et al., 2009; Goldsmith-Pinkham & Imbens, 2013). It also highlights the fact that the observed social network relation is noisy. First, it is questionable whether one can use a 'true network'. Second, it is not immediately clear how to construct a social network for further analysis. These difficulties reveal that a suitable model in practice should be robust to network variations, a property that our proposed model has been designed for. In Sections 6.1 and 6.2, we will use the weighted network based on both surveys. If an edge is nominated in both surveys, the edge weight is 1; if the edge only appears in one survey, the edge weight is 0.5; otherwise, there is no connection. In Section 6.3, we will study the impacts of different ways to construct the network and show that our proposed SP model gives consistent conclusions under these different constructions.

At the end of the school year, students answered the 13 questions about their rating of the school's 'friendliness' atmosphere, such as 'How many students at this school think it is good to be friendly and nice with all students no matter who'. For each of such questions, the student would give a score ranging from 0 to 5, representing their estimate of the proportion of students satisfying the condition in the question, ranging from 'Almost nobody' (0) to 'Almost everyone' (5). Overall, a higher rating indicates a more friendly environment.[1] We use the average of the

[1]Two of the questions were about negative atmosphere. So we transform then by the 5−the original scores before combining with other questions to ensure all scores are measuring positive atmosphere.

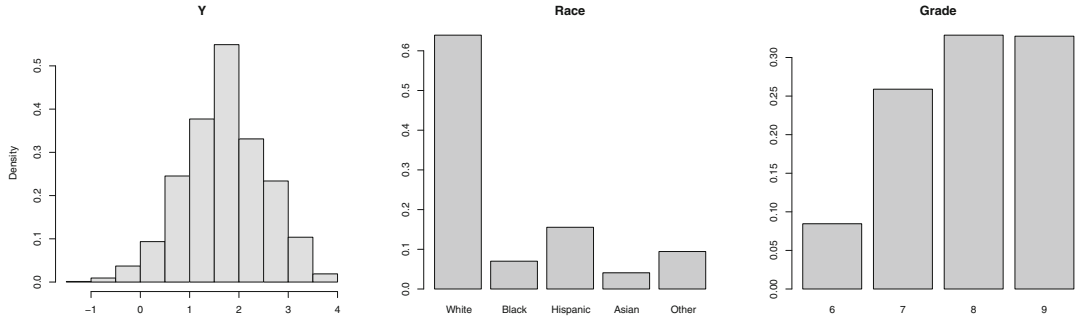

FIGURE 1 Marginal distributions of the response variable, race and grade in the data set

13 scores as the overall friendliness rating of each student. The score is further raised by a power of 1.2 to make the data more Gaussian based on the Box-Cox test. In this example, our target is to learn the workshop's impact on students' perception of the school atmosphere and conflict situations, and which of the other demographic attributes are also strongly related to the response. The demographic attributes include race (white, black, Hispanic, Asian, other), gender (M, F), grade, whether the student is a returning student from the previous year, and whether the student lives with both parents. The involvement of the workshop is indicated by a single binary vector (treatment). For each school, after removing observations with missing values in the response and predictors, we take the largest connected component of the social network as the final data. In total, 9026 students from 26 schools are included in our analysis. Each school will be assigned an individual effect parameter to account for potential school-level differences. Figure 1 shows distributions of non-binary variables in the data set. The distributions of the binary variables are included in Appendix S1 (Section K).

Notice that the treatment is randomised by design, uncorrelated with other covariates. We expect all reasonable estimation procedures to give a similar treatment effect estimate. However, the variance of the estimation (and the validity of the inference) would depend on modelling effectiveness. This will be reflected in our results to follow.

6.1 | Full models

We first use all predictors to fit a regression model based on our proposed SP and the three benchmark methods (OLS, SIM and RNC). Due to the redundancy of insignificant variables, these may not serve as final models. However, we can have an evaluation of four methods on common ground.

6.1.1 | Network effects and model significance

In the SP estimation, r is detected to be 0, so there is no confounding observed between the covariates and network space. The χ^2 test of our SP method gives a very small p-value ($< 10^{-6}$), representing strong evidence of network effects. Section K of Appendix S1 includes visualisations of our model fitting residuals, and the residuals follow a Gaussian pattern reasonably well. The fitted α values are shown in Figure 2a. Recall that our model includes the standard linear

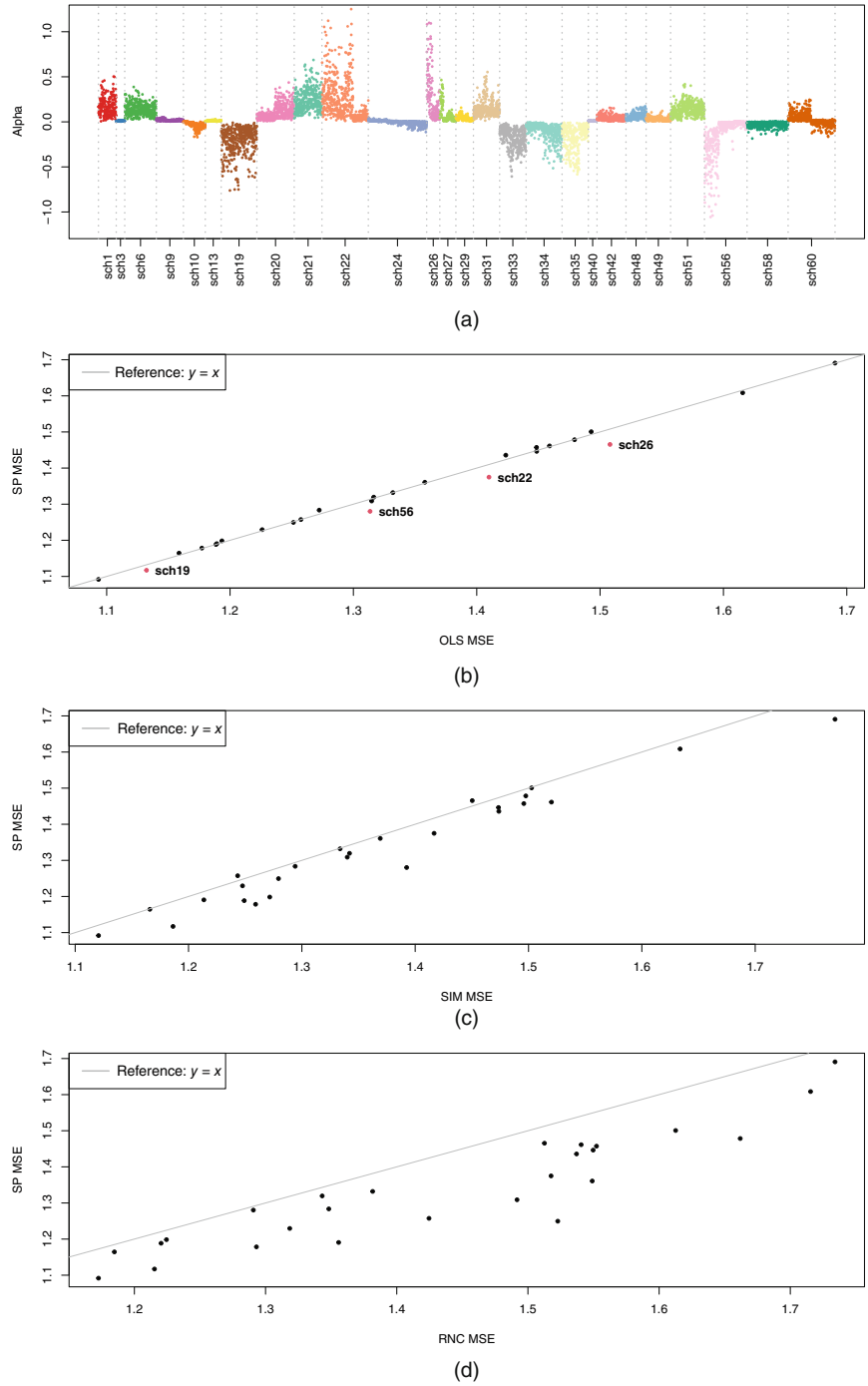

FIGURE 2 Fitted α_i's in the subspace projection method (SP) and the mean square prediction error comparison for each school. (a) Fitted α by the SP method; (b) Mean squared prediction error at each school: SP versus ordinary least squares method; (c) Mean squared prediction error at each school: SP versus social interaction model; (d) Mean squared prediction error at each school: SP versus regression with network cohesion [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 5 Estimates and *p*-values of the common covariates (excluding the intercept and schools)

	SP		OLS		SIM		RNC	
	Coefficient	***p*-Value**	**Coefficient**	***p*-Value**	**Coefficient**	***p*-Value**	**Coefficient**	***p*-Value**
Race: White	0.0260	0.2432	0.0261	0.4831	0.0152	0.6941	0.2131	–
Race: Black	−0.1193	0.0121	−0.1288	0.0149	−0.1079	0.0564	0.194	–
Race: Hispanic	−0.0313	0.2261	−0.0244	0.5580	−0.0207	0.6272	0.1102	–
Race: Asian	0.0433	0.2477	0.0394	0.5358	0.0297	0.6449	0.1637	–
Grade	−0.1663	$<10^{-4}$	−0.1784	$<10^{-4}$	0.0148	0.8574	0.3672	–
Gender	−0.0091	0.3557	−0.0099	0.6842	0.0842	0.1458	0.1207	–
Return student	−0.1429	0.0001	−0.1296	0.0002	−0.0816	0.0859	−0.1379	–
Live w/ both parents	0.0715	0.0066	0.0767	0.0078	0.0610	0.0376	0.1710	–
Treatment	−0.0839	0.0442	−0.0889	0.0709	−0.0895	0.0702	−0.0768	–
Local avg.: White	–	–	–	–	0.0719	0.5769	–	–
Local avg.: Black	–	–	–	–	0.0144	0.9321	–	–
Local avg.: Hispanic	–	–	–	–	−0.0023	0.987	–	–
Local avg.: Asian	–	–	–	–	−0.0152	0.9426	–	–
Local avg.: grade	–	–	–	–	−0.1776	0.2977	–	–
Local avg.: gender	–	–	–	–	−0.1389	0.1267	–	–
Local avg.: return	–	–	–	–	−0.1041	0.2623	–	–
Local avg.: w. parents	–	–	–	–	0.2654	0.0064	–	–
Local avg.: treatment	–	–	–	–	0.1982	0.2734	–	–

Notes: For the SIM, exogenous coefficients (for local averages of covariates) are also included. '–' indicates that the quantity is not available in the setting.
Abbreviations: OLS, ordinary least squares method; RNC, regression with network cohesion; SIM, social interaction model; SP, subspace projection.

regression model as a special case. That means, if the OLS inference is correct, our inference should also be correct. Therefore, the rejection in the χ^2 test also indicates that the OLS model-fitting without network effects is not proper for the data. The test of network autoregressive effect in the SIM model gives a *p*-value of 0.92, suggesting no strong evidence of the endogenous correlation. Furthermore, the exogenous coefficients are not significantly different from zero either, as shown in Table 5. Together, these indicate that the SIM autoregressive pattern is supported. The RNC model does not come with an inference framework. However, the cross-validation procedure in the model fitting selects a penalty that gives a very different fitting from the OLS, which implicitly indicates the potential network effects.

6.1.2 | Parameter estimation

Table 5 displays the estimated coefficients (excluding schools and intercepts). The SP fitting and OLS fitting agree on the overall magnitude in most coefficients with a 5%–10% difference in the parameters with small *p*-values. In particular, on the treatment effect, the SP estimates a negative effect of −0.084 while the OLS gives an estimate of −0.089. The similarity is expected, as discussed

before. However, the effectiveness of incorporating other factors would influence the variance for the inference. In this example, while the SP estimates a slightly smaller treatment effect, the corresponding p-value turns out to be smaller, indicating a much smaller estimation variance. The SIM model has very different estimates due to incorporating all local averages, except for the treatment effect due to the randomised design. The coefficients are not directly comparable with SP or OLS, but most of them come with larger p-values, especially considering the number of tests. Overall, we consider the current SIM fitting as non-informative. RNC method does not allow for the intercept and the school effects. Therefore, the estimated parameters are not meaningfully comparable to the other methods. We treat this as a limitation because it eliminates the straightforward interpretation of school-level fixed effects. The drawback of lacking an inference framework becomes more critical in this case. The table shows that the RNC renders larger estimated effects, but it is unclear how significant they are.

6.1.3 | Predictive performance

We can also compare the methods by their predictive performance. The predictive performance is evaluated in a cross-validation manner. We randomly split the 9026 students into 200 folds. One fold of students' responses is held out each time, and we estimate the models using the remaining 199 folds of responses with the full set of predictors and the network. This procedure is repeated for all 200 folds. As schools exhibit significant variations in social effects, we calculate the mean squared prediction errors within each school $\sum_{i\in \mathrm{sch}_k}(y_i-\hat{y}_i)^2/|\mathrm{sch}_k|$. The comparisons between the SP method with the other three are shown in Figure 2. Recall that when the α_i's are uniformly close to zero, our model would be similar to the standard linear regression, while when the α_i's exhibit large deviations from zero, the SP and OLP tend to have very different results. This phenomenon is verified in the example. Figure 2b shows that the SP and OLS have similar performance for most schools, but the SP has more accurate predictions for schools sch19, sch22, sch26 and sch56. This observation matches the fitted α values in Figure 2a, as these four schools have substantial variations in α's. In most of the other schools, the estimated α's have a smaller and more uniform magnitude and the difference between the two methods is small. Compared with the SIM and RNC, the SP renders better predictive performance in most schools. These results show that the SP model is more proper for the data than the others.

6.2 | Interpretation with refined models

In the full models of the previous section, many schools do not have significantly different effects from the reference level, and Table 5 also suggests that many covariates are not significant. To better understand and interpret the data, we will resort to statistical inference for model selection. The RNC is not suitable for this model refinement procedure due to the lack of an inference framework. We will focus only on SP, OLS and SIM.

Starting from the full model, we first check the significance of the categorical variables, School (26 categories) and Race (five categories). Bonferroni correction is applied to the multiple-level tests. Insignificant levels (adjusted p-value ≥ 0.05) are merged into the reference level (School: sch1, Race: other). After this step, the model selection follows backward elimination (Halinski & Feldt, 1970). The variable with the largest p-value that exceeds 0.05 (after Bonferroni correction) is removed at each step until no further elimination is possible. Throughout this

TABLE 6 Estimates and *p*-values of variables in the reduced models

	SP		OLS		SIM	
	Coefficient	***p*-Value**	**Coefficient**	***p*-Value**	**Coefficient**	***p*-Value**
(Intercept)	4.1221	$<10^{-4}$	4.3117	$<10^{-4}$	3.4427	$<10^{-4}$
Treatment	−0.0856	0.0411	−0.0812	0.0993	−0.079	0.1204
Race: Black	−0.136	0.0033	−0.1786	0.0003		
Grade	−0.1828	$<10^{-4}$	−0.1948	$<10^{-4}$		
Return	−0.1162	0.0004	−0.0987	0.0016	−0.1967	0.0001
Live w/ both parents	0.0767	0.0038			0.0927	0.0019
School 3	0.5421	$<10^{-4}$	0.4437	0.0001	0.3904	0.001
School 19			−0.2901	$<10^{-4}$	−0.4269	$<10^{-4}$
School 20	0.3503	$<10^{-4}$	0.3281	$<10^{-4}$	0.4512	$<10^{-4}$
School 27	−0.3371	0.0005	−0.3954	$<10^{-4}$		
School 34			−0.3017	$<10^{-4}$		
School 35	0.8085	$<10^{-4}$	0.5303	$<10^{-4}$		
School 40	0.4912	$<10^{-4}$	0.3802	0.0008	0.4674	0.0002
School 42	0.4274	$<10^{-4}$	0.3678	$<10^{-4}$		
School 48			−0.3634	$<10^{-4}$		
School 49			−0.3869	$<10^{-4}$		
School 51	−0.5481	$<10^{-4}$	−0.5234	$<10^{-4}$	−0.5032	$<10^{-4}$
School 56	−0.3192	0.0002	−0.5437	$<10^{-4}$	−0.606	$<10^{-4}$
Local avg.: live w/ both parents	–	–	–	–	0.6465	$<10^{-4}$

Notes: '–' indicates that the quantity is not available in the setting. A blank means the variable is excluded by model selection.
Abbreviations: OLS, ordinary least squares method; RNC, regression with network cohesion; SIM, social interaction model; SP, subspace projection.

procedure, we always keep the treatment variable from elimination. For the SIM, before the backward elimination step, the elimination of the exogenous parameters is applied. The final models from the three methods are given in Table 6. In the SP model, the χ^2 test for the network effect gives a *p*-value smaller than 10^{-10}, indicating strong evidence of network effects. However, the SIM model results in a *p*-value of 0.168, showing no firm evidence of network autoregressive correlation. The SIM model does have the local average of 'live with both parents' as a significant covariate effect. However, it fails to identify the impacts of the Race:black and the Returning Student, different from SP and OLS.

Comparing the OLS and SP, the high-level messages from the two models coincide in many aspects. Both models identify that black students tend to have more negative ratings of the school atmosphere, suggesting potential racial effects in school conflicts. Return students and students from higher grades also tend to have more negative perceptions. The two models also have differences in their conclusions. Based on the SP model, students living with both parents tend to have a more positive perception of the school atmosphere, while the OLS fails to detect this phenomenon. Though the ground truth about the data set is unknown, the SP's finding agrees with a few previous social studies on experiments (Anderson, 2014; Musick & Meier, 2010). When

it comes to the treatment effect, as expected, all three methods give similar estimates but due to the more effective modelling, the SP delivers a smaller p-value of 0.04, suggesting potentially weak effects. Note that the treatment effect is negative. This is reasonable since the education workshops help introduce more information about school conflicts, so students may be aware of many conflicts they did not know about before. Our conclusion is implicitly supported by the original analysis of experimenters (Paluck et al., 2016), who showed that students involved in the workshops were more likely to wear orange waistbands as a public sign that they stand against school conflicts. However, the analysis of Paluck et al. (2016) considers the experiment design with additional assumptions about the spill-over effects of the educational workshops and is based on complicated causal inference methods. Therefore, their conclusion is about the stronger causal relation. In contrast, our method does not take advantage of the design, nor do we know whether the network cohesion assumption explicitly corresponds to the widely used spill-over effects assumptions. So our conclusion is not causal. We leave the development of causal inference under the current framework for future work.

6.3 | Robustness to network perturbation

In the previous sections, we use the information of both nomination surveys to construct a weighted network for the regression analysis. In practice, researchers seldom know whether such a construction is the best one or if an optimal construction exists. It may also be more natural for some people to use the Wave II survey (or the Wave I survey) as the primary information. However, a valuable analysis that can reflect the true nature of the data should deliver consistent results as long as one uses some reasonable construction of the network. In this section, we examine how different ways of constructing the network would change the resulting models from the previous analysis. OLS does not use the network information, so it would not be affected. We will focus on comparing SP and the SIM.

The first alternative construction is the undirected network by only the Wave II survey. The same model fitting procedures of the previous section are applied, where model selection is done by back elimination with p-values and multiple comparison correction. The fitted SP and SIM with the Wave II network are given in Table 7. The SP gives a very small p-value ($< 10^{-10}$) for the χ^2 test, indicating strong network effects. Moreover, comparing Tables 7 and 6, we can see that the high-level message of the SP model remains consistent: the Race:Black, Grade, and Return Student variable have negative effects while 'live with parents' has a positive effect. The parameter estimates are slightly different but the changes are marginal. The school effects are also very similar, with the only difference on the school sch49. So overall, the change of the network construction method does not lead to material changes in the SP inference results. The SIM, in contrast, delivers very different messages from Tables 6 and 7. For example, on the Wave II network, the SIM model identifies a strong network autoregressive correlation (p-value $= 0.0003$). However, the local average of 'live with parents' is no longer included in the model. Both Gender and Race:Black are now identified as strong effects in the model. These are all different from the previous result.

Naturally, the second alternative network is the unweighted network based on the Wave I survey. The model fitting results are shown in Table 8. The conclusion remains the same. Though the estimates of the SP model change numerically, the significance and the selected model remain the same. The χ^2 test indicates a strong network effect in all three cases. The SIM again selects different results compared to either Table 6 or Table 7. It also fails to identify the autoregressive

TABLE 7 Estimates and *p*-values of reduced models based on the Wave II network

	SP		SIM	
	Coefficient	***p*-Value**	**Coefficient**	***p*-Value**
(Intercept)	4.0926	$<10^{-4}$	4.9619	$<10^{-4}$
Treatment	−0.0838	0.0439	−0.0904	0.1673
Race: Black	−0.1362	0.0032	−0.2084	0.0022
Grade	−0.1795	$<10^{-4}$		
Return	−0.0761	0.0175	−0.1965	0.0005
Live with both parents	0.0765	0.0038	0.1392	0.0004
Gender			0.0850	0.0451
sch3	0.5213	$<10^{-4}$		
sch20	0.3621	$<10^{-4}$		
sch19			−0.6291	$<10^{-4}$
sch22			−0.3637	$<10^{-4}$
sch24			−0.4440	$<10^{-4}$
sch26			−0.4029	0.0015
sch27	−0.3483	0.0004	−0.6313	$<10^{-4}$
sch29			−0.5946	$<10^{-4}$
sch34			−0.3376	$<10^{-4}$
sch35	0.7134	$<10^{-4}$	0.2783	0.0019
sch40	0.4733	$<10^{-4}$		
sch42	0.4121	$<10^{-4}$		
sch48			−0.7667	$<10^{-4}$
sch49	−0.3490	0.0003	−0.6358	$<10^{-4}$
sch51	−0.4448	0.0001	−1.1455	$<10^{-4}$
sch56	−0.3561	$<10^{-4}$	−0.9281	$<10^{-4}$
sch58			−0.4807	$<10^{-4}$
sch60			−0.7074	$<10^{-4}$

Notes: '–' indicates that the quantity is not available in the setting. A blank means the variable is excluded in the selection procedure.
Abbreviations: SIM, social interaction model; SP, subspace projection.

correlation. In summary, it is evident that the SIM heavily relies on how one constructs the network. Since it is unclear which is the best way to construct the network, the SIM fails provide a reliable analysis in this case.

The above difference in robustness highlights the crucial advantage of our model over the SIM framework. Figure 3 displays the 'sch40' networks from the two surveys. The two networks have only about 60% overlapping edges. So a statistical model focusing on local edges may deliver very different results, as reflected in the SIM case. Our model, in contrast, relies on the subspace assumption that is more robust to these changes. In other words, the eigenspaces remain stable

TABLE 8 Estimates and *p*-values of reduced models based on the Wave I network

	SP		SIM	
	Coefficient	***p*-Value**	**Coefficient**	***p*-Value**
(Intercept)	4.1216	$<10^{-4}$	3.5797	$<10^{-4}$
Treatment	−0.0860	0.0401	−0.0691	0.1687
Race: Black	−0.1333	0.0037	−0.1409	0.0063
Grade	−0.1797	$<10^{-4}$		
Return	−0.0966	0.0017	−0.2237	$<10^{-4}$
Live with both parents	0.0868	0.0012	0.0928	0.0019
sch3	0.5135	$<10^{-4}$		
sch20	0.4070	$<10^{-4}$	0.5619	$<10^{-4}$
sch27	−0.3530	0.0002		
sch33	0.3088	0.0032		
sch35	0.7650	$<10^{-4}$		
sch40	0.4672	$<10^{-4}$		
sch42	0.3836	0.0001	0.2750	0.0384
sch49	−0.3465	0.0004		
sch51	−0.5468	$<10^{-4}$	−0.4526	$<10^{-4}$
sch56	−0.3861	$<10^{-4}$	−0.5451	$<10^{-4}$
sch19			−0.3956	$<10^{-4}$
sch31			0.3844	$<10^{-4}$
Local average: live with parents			0.3617	$<10^{-4}$

Notes: '–' indicates that the quantity is not available in the setting. A blank means the variable is excluded in the selection procedure.

Abbreviations: SIM, social interaction model; SP, subspace projection.

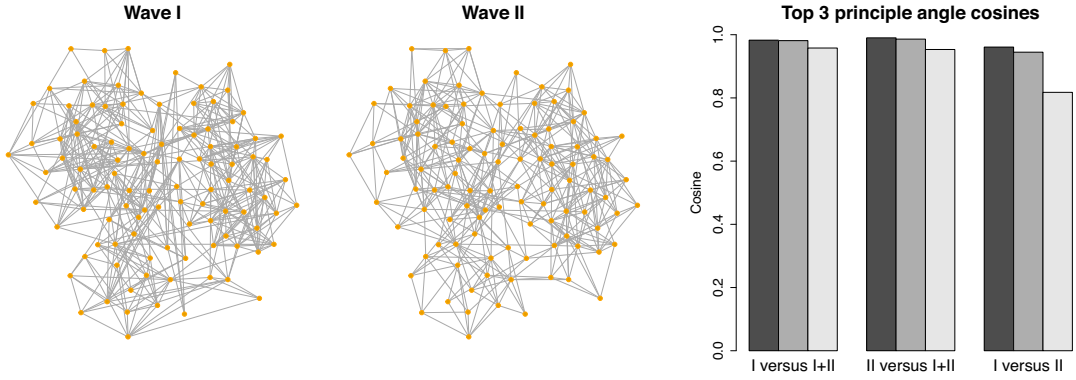

FIGURE 3 Demonstration of network perturbation in two waves of surveys in one school. About 40% edges in one network disappear in the other one. However, the leading three-dimensional eigenspaces of the networks still align well, indicated by the large cosine values of the principle angles. [Colour figure can be viewed at wileyonlinelibrary.com]

under a certain amount of perturbations. Recall that the alignment of two subspaces is measured by the cosine values of their principal angles. A perfect alignment would result in cosine values of 1. Figure 3 shows the principle angle cosine values for the pairwise alignment between the three-dimensional leading eigenspaces of Wave I network, combined (and weighted) Wave I+II network, and Wave II network. The eigenspace alignment is close to perfect (> 0.95) for the leading two dimensions and remains high for the third dimension. Therefore, even though the two networks are very different in their edge sets, the eigenspace remains stable. In Figure K.3 of Appendix S1, we include the principle angle plots for the other 25 schools in the data set. Such stable alignment of the eigenspace holds in most schools. This observation explains the robustness of our model.

7 | CONCLUSION

We have introduced a linear regression model on observations linked by a network. The model comes with a computationally efficient inference algorithm that can tolerate network observational errors. The study in this paper focuses on using Assumption 3 to control the estimation bias and deliver valid inference. It makes important progress towards the inference problem for the network-linked model.

The next step is to explore whether a certain bias correction can be applied so that the small perturbation assumption can be further relaxed. Developing such a technique will require an accurate estimate of the perturbation $\hat{P} - P$ and, in turn, new tools for characterising such random matrices. Meanwhile, the current focus is on the fixed design problem of X and P; the framework remains valid if we condition on X and P while assuming that they are independent. Exploring the model with dependence between X and P is another promising research direction. One such situation is when network evolution is observed over time. In Goldsmith-Pinkham and Imbens (2013) and McFowland III and Shalizi (2021), assuming the network is perfectly observed, the generative models between X and P lead to valid estimations of homophily effects. It would be very interesting to study if embedding our subspace regression strategies in such settings would lead to robust inference framework of homophily.

ACKNOWLEDGEMENTS

Can M. Le is supported in part by the NSF grant DMS-2015134. Tianxi Li is supported in part by the NSF grant DMS-2015298 and the 3-Caverliers Award from the University of Virginia.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://doi.org/10.1073/pnas.1514483113.

ORCID

Tianxi Li https://orcid.org/0000-0003-4595-1777

REFERENCES

Abbe, E. (2018) Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18, 1–86.

Abbe, E., Fan, J., Wang, K. & Zhong, Y. (2017) Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*.

Anderson, J. (2014) The impact of family structure on the health of children: effects of divorce. *The Linacre Quarterly*, 81, 378–387.
Basse, G.W. & Airoldi, E.M. (2018a) Limitations of design-based causal inference and a/b testing under arbitrary and network interference. *Sociological Methodology*, 48, 136–151.
Basse, G.W. & Airoldi, E.M. (2018b) Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, 105, 849–858.
Belkin, M. & Niyogi, P. (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15, 1373–1396.
Bivand, R.S., Pebesma, E. & Gomez-Rubio, V. (2013) *Applied spatial data analysis with R*, 2nd edition. New York, NY: Springer. Available at: https://asdar-book.org/
Bollobas, B., Janson, S. & Riordan, O. (2007) The phase transition in inhomogeneous random graphs. *Random Structures and Algorithms*, 31, 3–122.
Bramoullé, Y., Djebbari, H. & Fortin, B. (2009) Identification of peer effects through social networks. *Journal of Econometrics*, 150, 41–55.
Butts, C.T. (2003) Network inference, error, and informant (in) accuracy: a Bayesian approach. *Social Networks*, 25, 103–140.
Candès, E.J. & Recht, B. (2009) Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9, 717–772.
Candès, E.J. & Tao, T. (2010) The power of convex relaxation: near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56, 2053–2080.
Cape, J., Tang, M. & Priebe, C.E. (2019) The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47, 2405–2439.
Chandrasekhar, A. & Lewis, R. (2011) Econometrics of sampled networks. *Unpublished manuscript, MIT [422]*.
Chatterjee, S. (2015) Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43, 177–214.
Chen, K. & Lei, J. (2018) Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113, 241–251.
Chen, Y., Li, X. & Xu, J. (2018) Convexified modularity maximization for degree-corrected stochastic block models. *The Annals of Statistics*, 46, 1573–1602.
Clauset, A. & Moore, C. (2005) Accuracy and scaling phenomena in internet mapping. *Physical Review Letters*, 94, 018701.
Fan, Z. & Guan, L. (2018) Approximate ℓ_0-penalized estimation of piecewise-constant signals on graphs. *The Annals of Statistics*, 46, 3217–3245.
Gao, C., Ma, Z., Zhang, A.Y. & Zhou, H.H. (2017) Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18, 1980–2024.
Gao, C., Ma, Z., Zhang, A.Y. & Zhou, H.H. (2018) Community detection in degree-corrected block models. *The Annals of Statistics*, 46, 2153–2185.
Gao, L.L., Witten, D. & Bien, J. (2019) Testing for association in multi-view network data. *arXiv preprint arXiv:1909.11640*.
Goldsmith-Pinkham, P. & Imbens, G.W. (2013) Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31, 253–264.
Halinski, R.S. & Feldt, L.S. (1970) The selection of variables in multiple regression analysis. *Journal of Educational Measurement*, 7, 151–157.
Handcock, M.S. & Gile, K.J. (2010) Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4, 5.
Holland, P.W., Laskey, K.B. & Leinhardt, S. (1983) Stochastic blockmodels: first steps. *Social Networks*, 5, 109–137.
Hsieh, C.S. & Lee, L.F. (2016) A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics*, 31, 301–319.
Jackson, M.O. & Rogers, B.W. (2007) Relating network structure to diffusion properties through stochastic dominance. *The BE Journal of Theoretical Economics*, 7, 1–13.
Javanmard, A. & Montanari, A. (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15, 2869–2909.

Ji, P. & Jin, J. (2016) Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10, 1779–1812.

Jin, J., Ke, Z.T., Luo, S. & Wang, M. (2020) Estimating the number of communities by stepwise goodness-of-fit. *arXiv preprint arXiv:2009.09177*.

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

Khabbazian, M., Hanlon, B., Russek, Z. & Rohe, K. (2017) Novel sampling design for respondent-driven sampling. *Electronic Journal of Statistics*, 11, 4769–4812.

Kolaczyk, E.D. (2009) *Statistical analysis of network data: methods and models*, 1st edition. New York: Springer Publishing Company, Incorporated.

Lakhina, A., Byers, J.W., Crovella, M. & Xie, P. (2003) Sampling biases in ip topology measurements. *Proceedings of the IEEE INFOCOM 2003. 22nd annual joint conference of the ieee computer and communications societies (IEEE Cat. No. 03CH37428)*, vol. 1. IEEE, pp. 332–341.

Le, C.M., Levin, K. & Levina, E. (2018) Estimating a network from multiple noisy realizations. *The Electronic Journal of Statistics*, 12, 4697–4740.

Le, C.M. & Levina, E. (2022) Estimating the number of communities in networks by spectral methods. *The Electronic Journal of Statistics*, 16, 3315–3342.

Le, C.M., Levina, E. & Vershynin, R. (2017) Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51, 538–561.

Lee, L.F. (2007) Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics*, 140, 333–374.

Lee, L.F., Liu, X. & Lin, X. (2010) Specification and estimation of social interaction models with network structures. *The Econometrics Journal*, 13, 145–176.

Lei, J. & Rinaldo, A. (2014) Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43, 215–237.

Lei, J. & Zhu, L. (2017) Generic sample splitting for refined community recovery in degree corrected stochastic block models. *Statistica Sinica*, 27, 1639–1659.

Lei, L. (2019) Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. *arXiv preprint arXiv:1909.04798*.

Lei, L., Li, X. & Lou, X. (2020) Consistency of spectral clustering on hierarchical stochastic block models. *arXiv preprint arXiv:2004.14531*.

Li, T., Lei, L., Bhattacharyya, S., Van den Berge, K., Sarkar, P., Bickel, P.J. et al. (2020) Hierarchical community detection by recursive partitioning. *Journal of the American Statistical Association*, 117, 951–968.

Li, T., Levina, E. & Zhu, J. (2016) *netcoh: statistical modeling with network cohesion*. R package version 0.11. Available at: http://CRAN.R-project.org/package=netcoh.

Li, T., Levina, E. & Zhu, J. (2019) Prediction models for network-linked data. *The Annals of Applied Statistics*, 13, 132–164.

Li, T., Levina, E. & Zhu, J. (2020) Network cross-validation by edge sampling. *Biometrika*, 107, 257–276.

Li, T., Qian, C., Levina, E. & Zhu, J. (2020) High-dimensional Gaussian graphical models on network-linked data. *Journal of Machine Learning Research*, 21, 1–45. Available at: http://jmlr.org/papers/v21/19-563.html

Lunagómez, S., Papamichalis, M., Wolfe, P.J. & Airoldi, E.M. (2018) Evaluating and optimizing network sampling designs: decision theory and information theory perspectives. *arXiv preprint arXiv:1811.07829*.

Manresa, E. (2013) *Estimating the structure of social interactions using panel data*. Unpublished Manuscript. CEMFI, Madrid.

Manski, C.F. (1993) Identification of endogenous social effects: the reflection problem. *The Review of Economic Studies*, 60, 531–542.

Mao, X., Sarkar, P. & Chakrabarti, D. (2020) Estimating mixed memberships with sharp eigenvector deviations. *Journal of the American Statistical Association*, 116, 1928–1940.

McFowland, E., III & Shalizi, C.R. (2021) Estimating causal peer influence in homophilous social networks by inferring latent locations. *Journal of the American Statistical Association*, 1–12.

Michell, L. & West, P. (1996) Peer pressure to smoke: the meaning depends on the method. *Health Education Research*, 11, 39–49.

Musick, K. & Meier, A. (2010) Are both parents always better than one? Parental conflict and young adult well-being. *Social Science Research*, 39, 814–830.
Newman, M. (2018) Estimating network structure from unreliable measurements. *Physical Review E*, 98, 062321.
Ng, A.Y., Zheng, A.X. & Jordan, M.I. (2001) Link analysis, eigenvectors and stability. *Proceedings of the international joint conference on artificial intelligence*, vol. 17. Lawrence Erlbaum Associates Ltd, pp. 903–910.
Ogburn, E.L. (2018) Challenges to estimating contagion effects from observational data. In: *Complex spreading phenomena in social systems*. Cham: Springer, pp. 47–64.
Paluck, E.L., Shepherd, H. & Aronow, P.M. (2016) Changing climates of conflict: a social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, 113, 566–571.
Pearson, M. & West, P. (2003) Drifting smoke rings. *Connections*, 25, 59–76.
Qiu, Y. & Mei, J. (2019) *RSpectra: solvers for large-scale eigenvalue and SVD problems*. R package version 0.16-0. Available at: https://CRAN.R-project.org/package=RSpectra
Rohe, K. (2019) A critical threshold for design effects in network sampling. *The Annals of Statistics*, 47, 556–582.
Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N. et al. (2014) A proteome-scale map of the human interactome network. *Cell*, 159, 1212–1226.
Sadhanala, V., Wang, Y.X. & Tibshirani, R.J. (2016) Graph sparsification approaches for Laplacian smoothing. *Proceedings of the 19th international conference on artificial intelligence and statistics*, pp. 1250–1259.
Shalizi, C.R. & Thomas, A.C. (2011) Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, 40, 211–239.
Shi, J. & Malik, J. (2000) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 888–905.
Su, L., Lu, W., Song, R. & Huang, D. (2020) Testing and estimation of social network dependence with time to event data. *Journal of the American Statistical Association*, 115, 1–28. Available from:. https://doi.org/10.1080/01621459.2019.1617153
Tang, M., Sussman, D.L. & Priebe, C.E. (2013) Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41, 1406–1430.
Van de Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42, 1166–1202.
Wang, Y.X., Sharpnack, J., Smola, A. & Tibshirani, R.J. (2016) Trend filtering on graphs. *Journal of Machine Learning Research*, 17, 1–41.
Wu, Y.J., Levina, E. & Zhu, J. (2018) Link prediction for egocentrically sampled networks. *arXiv preprint arXiv:1803.04084*.
Xia, D. (2021) Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics*, 15, 3798–3851.
Zhang, C.H. & Zhang, S.S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 217–242.
Zhao, S. & Shojaie, A. (2016) A significance test for graph-constrained estimation. *Biometrics*, 72, 484–493.
Zhu, X., Pan, R., Li, G., Liu, Y. & Wang, H. (2017) Network vector autoregression. *The Annals of Statistics*, 45, 1096–1123.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Le, C.M. & Li, T. (2022) Linear regression and its inference on noisy network-linked data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(5), 1851–1885. Available from: https://doi.org/10.1111/rssb.12554