Estimating Network Spillovers Under Dense Measurement Error

Yingxing Li, Áureo de Paula and Weining Wang

Discussion Paper 26/841
22 July 2026

School of Economics
University of Bristol
Priory Road Complex
Bristol
BS8 1TU
United Kingdom

University of BRISTOL
School of Economics

ESTIMATING NETWORK SPILLOVERS UNDER DENSE MEASUREMENT ERROR

YINGXING LI[A], ÁUREO DE PAULA[B], AND WEINING WANG[C]

Abstract. This paper analyzes spillover effects in spatial (network) models when the neighborhood (adjacency) matrix is contaminated by measurement error from reporting, aggregation, or disclosure imperfections, leading to inconsistent estimation of network effects. We introduce a regularization framework for the latent network that allows for sparse and/or low-rank structure and accommodates potential correlation between measurement errors and outcomes. We propose two estimators: (i) a two-stage procedure that first denoises the adjacency matrix and then incorporates the purified network into a regression analysis, and (ii) a Generalized Method of Moments (GMM) estimator that jointly estimates regression parameters and refines the network structure. We then establish strictly improved consistency rates for the spillover effect estimator relative to naive estimation ignoring measurement error. Simulations demonstrate that, in the presence of noisy networks, our approach reduces the root mean squared error of spillover estimates relative to conventional methods by approximately $50-80\%$. We apply our framework to examine the international spillover of economic growth, and the tax competition across U.S. states, illustrating that denoising might restore Leontief stability and yields improved estimates of spillovers.

JEL Classification: C21, C23, D57

Keywords: network analysis, measurement error, LASSO, nuclear norm penalty, penalized GMM

Date: July 22, 2026.

Author names are listed in alphabetical order.

We thank Santiago Montoya-Blandon for comments and preliminary data analysis. We thank Debopam Bhattacharya, Tom Boot, Guido Kuersteiner, Oliver Linton, Elena Manresa, Whitney Newey, Alexey Onatski, Bernard Salanie, Shuyang Sheng, Xun Tang as well as participants at the Gothenburg Econometrics Workshop, the Oxford Econometrics Workshop, the VTSS Seminar, the $(EC)^2$ Conference, the 2025 Econometric Society World Congress and SoFiE 2026 for their thoughtful comments. Áureo de Paula gratefully acknowledges financial support from the Economic and Social Research Council through the ESRC Institute for the Microeconomic Analysis of Public Policy (Grant Ref: ES/T014334/1), and from UK Research and Innovation (UKRI) under the UK Government's Horizon Europe funding guarantee (Grant Ref: EP/X02931X/1). Any remaining errors are our own.

[a]: School of Business, Sun Yat-sen University.

[b]: Department of Economics, University College London, Institute for Fiscal Studies and CeMMAP.

[c]: Department of Economics, University of Bristol. Corresponding author: Weining Wang, email: weining.wang@bristol.ac.uk.

1. Introduction

Network analysis provides powerful tools to quantify policy spillovers and identify pivotal agents for targeted interventions, particularly when policymakers face binding resource constraints (Ballester et al., 2006; Zenou, 2016; Galeotti et al., 2020). A large literature shows that economic outcomes often depend on interactions among interconnected units, giving rise to amplification, diffusion, and multiplier effects that are central to policy design (Acemoglu et al., 2012; de Giorgi et al., 2020; Araujo et al., 2023). These insights have motivated the widespread use of spatial and social interaction models to study peer effects, diffusion dynamics, and network-based policy interventions across a broad range of economic settings.

Methodologically, the literature on spatial and social interaction models differs mainly in how the network structure is handled. A dominant strand assumes that the adjacency matrix is correctly specified and directly observed, and studies spillover effects conditional on this object (Lee, 2007b; Bramoullé et al., 2009; Lee et al., 2010; Banerjee et al., 2013; Elhorst, 2014; Lee and Yu, 2014; Yang and Lee, 2017; Zhu et al., 2020; Kuersteiner and Prucha, 2020), (Shin and Thornton, 2021; Chernozhukov et al., 2021). An alternative and growing line of research relaxes this assumption by estimating the network jointly with model parameters, typically imposing dimension reduction constraints such as sparsity to address the curse of dimensionality in high-dimensional settings (Manresa, 2016; Lam and Souza, 2020; de Paula et al., 2025).

Sparsity-based methods work well when the latent network is dominated by a few strong links, but they fail in the dense networks that arise in production, trade, and financial settings, where the observed adjacency matrix is itself contaminated by measurement error (Fisman and Wei, 2004; Upper, 2011; Anand et al., 2015). In such settings, two distinct forms of latent structure coexist beyond idiosyncratic noise. First, outcomes may be driven by a small number of pervasive (though individually weak) latent forces such as common shocks, shared technologies, or institutional linkages, naturally represented by a low-rank component (Chudik and Pesaran, 2013; Kapetanios et al., 2021). Second, networks may contain a small number of strong bilateral or idiosyncratic interactions, naturally represented by a sparse component (Lam and Souza, 2020; de Paula et al., 2025). Imposing sparsity alone discards the pervasive signal, while ignoring measurement error produces non-vanishing bias as elementwise small errors accumulate along network paths (Lewbel et al., 2024). The stakes are practical: in our cross-state tax-competition application, the standard spatial GMM estimator with the raw inverse-distance weight matrix yields $\widehat{\lambda} > 1$, violating the Leontief stability condition and implying explosive multipliers, while our estimator returns $\widehat{\lambda} < 1$ within the admissible region. We therefore propose a regularization approach for the observed adjacency matrix that filters measurement error while preserving low-rank or sparse signal, developed formally in Section 3.

Throughout, by *regularization* we mean imposing restrictions: low-rank, sparse, or low-rank-plus-sparse, directly on the latent adjacency matrix W_0 rather than on the regression coefficients, as identifying restrictions that make denoising of the observed network feasible.

The problem is non-trivial because it combines matrix-valued measurement error with a spillover estimator that depends nonlinearly on the network matrix through the Leontief inverse $(I_n - \lambda W)^{-1}$, so errors in W propagate through the estimator in ways classical errors-in-variables corrections do not handle. Low-rank and sparse structures have deep foundations in statistics and econometrics, appearing in approximate factor models, latent-variable graphical models, and high-dimensional covariance estimation (Bai and Ng, 2002; Chandrasekaran et al., 2010; Fan et al., 2011). A large methodological literature develops penalized procedures to recover low-rank or sparse patterns in covariance and precision matrices (Cai et al., 2011; Candès et al., 2011; Agarwal et al., 2012; Fan et al., 2018). Despite their success in these settings, regularization of this kind has not, to our knowledge, been adopted in the econometrics literature to address measurement error in observed network adjacency matrices, or to improve the estimation of spillover effects and other policy-relevant network statistics. We build on this literature by developing two complementary estimation approaches. The first is a two-stage procedure that denoises the observed adjacency matrix by separating low-rank or sparse signal from measurement error before estimating spillover effects. The second is a regularized Generalized Method of Moments (GMM) estimator that jointly estimates the network and the model parameters in a single step. We establish that our estimator is consistent, in contrast to naive estimation that ignores measurement error and remains inconsistent in dense regimes. Simulation results show that the proposed method achieves a 50–80% reduction in the root mean squared error of spillover estimates under noisy weight matrices, with measurement noise ranging from moderate to high.

Our main theoretical contribution is to characterize how the bias of standard spillover estimators depends on the structure of the latent network and the measurement error, and to show that this bias does not vanish in the dense regime that characterizes economic networks. Lewbel et al. (2024) show that modest link misclassification bias in the estimation of linear social-interaction models may be negligible as long as misclassification is not pervasive; we extend their analysis to settings in which the latent network has many but individually weak connections, and in which the measurement error itself need not be sparse. We further allow the measurement error to be correlated with unobservables in the outcome equation (Candelaria and Ura, 2023; Lewbel et al., 2023), which can introduce non-vanishing bias if ignored. We derive convergence rates for both the spillover-parameter estimator and our estimated adjacency matrix. Beyond these formal results, the framework also delivers an interpretable decomposition of network connections (a low-rank component capturing pervasive relationships and a sparse component isolating localized interactions, with identification following standard strategies in the literature (Chandrasekaran et al., 2011)) and we use this decomposition to provide a novel expression of the network eigencentrality that quantifies the contributions of each component.

We illustrate the practical relevance of our framework through two empirical applications: international spillovers in GDP growth across 23 economies and tax competition among 48 U.S. states. In both settings, our estimated network produces spillover estimates that differ markedly from those obtained using the raw adjacency matrix, highlighting the empirical importance of network denoising. In the cross-economy GDP-growth application, regularization affects not only the scale of spillovers but also the inferred structure of interdependence. The network decomposition further reveals distinct regional and global components of cross-economy interdependence. These empirical findings demonstrate how our framework can uncover the structure of interdependence in economic networks and support more reliable inference in spatial econometric analysis. In the U.S. tax-competition application, the raw inverse-distance weight matrices yield explosive estimates $\widehat{\lambda} > 1$ that violate the Leontief stability condition, while our estimator returns stable estimates.

The remainder of our paper is organized as follows. Section 2 motivates the framework by explaining why measurement error matters in dense networks, illustrating three possible network structures (sparse, low-rank, low-rank-plus-sparse), and demonstrating with the U.S. Input-Output Table that real-world dense networks exhibit exploitable regularization structure. Section 3 presents the spatial interaction model and introduces the plug-in and regularized GMM estimators under low-rank, sparse, or low-rank-plus-sparse structure. Section 4 provides assumptions and establishes consistency and asymptotic normality results for both estimators. Section 5 presents simulation evidence. Section 6 applies our methods to estimate spillover effects using noisy network data. Section 7 concludes. Proofs, algorithms, and additional technical details are provided in the appendices.

2. Motivation

This section sets up the notation and previews the regularization framework used throughout the paper. Let $\mathcal{N} \coloneqq \{1, \ldots, n\}$ denote the set of n units or nodes and $\mathcal{E}$ the set of edges between them, so that $(\mathcal{N}, \mathcal{E})$ represents the graph. Let W_0 be the $n \times n$ adjacency matrix encoding the true connections, with (i, j)-th element $W_{0,ij}$; we write $W_{0,ij} \neq 0$ when unit i is connected to unit j and $W_{0,ij} = 0$ otherwise.[1] We further assume no self-loops, so that $W_{0,ii} = 0$ for all $i = 1, \ldots, n$.

The econometrician observes a noisy version of W_0,

$$W = W_0 + E,$$

where E is a measurement-error matrix whose sources are discussed in Section 2.1. Before formally motivating the regularization framework, we briefly preview the structural decomposition of W_0 that plays a central role throughout the paper.

To reduce estimation complexity for large n, it is natural to impose structural assumptions on W_0. The most common choice (sparsity) is plausible for social networks dominated by a few strong links, but fails to describe networks in which a large number of units share

[1] We allow the connection from i to j to differ from that from j to i, so the graph may be directed and W_0 need not be symmetric.

common interests or characteristics: a highly sparse network typically splits into several disconnected components. We therefore propose to model W_0 as the sum of a low-rank and a sparse component,

$$W = \underbrace{L_0 + S_0}_{\equiv W_0} + E, \tag{1}$$

where L_0 captures pervasive, common connections and S_0 captures sporadic or idiosyncratic connections.

The components L_0 and S_0 need not individually satisfy the zero-diagonal constraint. When W_0 has a zero diagonal, we can define $L_0^* = L_0 - \text{diag}(L_0)$ and $S_0^* = S_0 + \text{diag}(L_0)$, which absorb the diagonal corrections so that both represent networks without self-loops:

$$W = \underbrace{L_0^* + S_0^*}_{\equiv W_0} + E. \tag{2}$$

While L_0^* is not exactly low-rank, it inherits the low-rank structure of L_0 and differs only on the diagonal. Section 2.3 illustrates the usefulness of this decomposition through concrete examples. We further develop our theoretical analysis (Section 4) in terms of L_0 and S_0 and note that the results extend directly to L_0^* and S_0^*. Both theoretical and empirical evidence in the following sections show that this decomposition captures several meaningful network structures encountered in practice (Diebold and Yılmaz, 2014; Barranca et al., 2015; Gamble et al., 2016; Newman, 2010).

2.1. **Why dense measurement errors matter.** A central feature of the applications mentioned in the introduction is that the observed network is dense and measured with error. We document four such sources of error in Section 2.2; here we develop the consequences for spillover estimation. The existing literature, such as Lewbel et al. (2024), assumes that measurement error in the network is sufficiently sparse that it does not affect the consistency of the spillover estimator. In particular, consistency of $\widehat{\lambda}$ can be preserved when the aggregate magnitude of the error vanishes asymptotically, for example when

$$\frac{1}{n}\sum_{i,j} |E_{ij}| = o_p(1).$$

This condition is naturally satisfied when the total error budget grows slowly relative to the size of the network, as in settings where the underlying adjacency matrix itself is sparse. Intuitively, when only a small fraction of links are contaminated *and the magnitude of each contamination is small*, the resulting perturbation to the network propagation mechanism becomes negligible in large samples.

The situation is fundamentally different when measurement errors are *dense*. Even when each entry of E is elementwise small, the errors accumulate across the $O(n^2)$ entries of the matrix and across the propagation paths generated by the spatial multiplier. The aggregate distortion may not vanish asymptotically, and the estimator $\widehat{\lambda}_{\text{GMM}}$ may fail to be consistent.

To recover W_0 from W in the dense regime, some structural restriction on W_0 is needed. Depending on the application, the appropriate structure is:

i) **Sparse:** $W_0 = S_0^*$, where most links are absent. Suitable for social networks.

ii) **Low-rank:** $W_0 = L_0^*$, where connections arise from a small number of common factors or shared institutional features. The primary case for dense networks such as input–output tables and financial exposure matrices.

iii) **Low-rank plus sparse:** $W_0 = L_0^* + S_0^*$, the general case combining pervasive common connections with idiosyncratic strong links.

Misspecifying this structure leaves residual bias: assuming $W_0 = S_0^*$ when a non-negligible L_0^* is present discards the pervasive systematic connections that drive spillovers; assuming $W_0 = L_0^*$ when strong idiosyncratic links are present absorbs those links into the residual. Moreover, the low-rank and sparse decomposition introduced above is the device we use to operationalize this regularization. We emphasize that it is a flexible modelling device, not an economic interpretation. The interactions-model estimator $\widehat{\theta}_p$ depends only on $\widehat{W} = \widehat{L}^* + \widehat{S}^*$, not on the individual components. We thus do not rely on any economic interpretation of L_0^* and S_0^* as separate network objects. Nevertheless, Appendix C discusses the economic interpretation of L_0^* and provides identification conditions for the $L_0^* + S_0^*$ decomposition.

The practical importance of regularizing W_0 rests on three observations. First, measurement error in dense networks is *non-negligible*: it may not vanish with sample size, so standard asymptotic consistency arguments fail. Second, it may be *endogenous*: measurement errors may share common sources with the outcome disturbances or covariates, so that E can be statistically dependent on ε_t or X_t, even conditional on the fixed matrix W_0. This violates the exogeneity conditions underlying standard instrumental-variable corrections. Third, it is *economically consequential*: in our empirical illustration with the U.S. Input–Output Table (Section 2.4), ignoring measurement error entirely distorts the estimated propagation matrix by 9%, and misspecifying the structure of the error (treating it as purely sparse) compounds the distortion to 68%. A method that filters dense, structured noise while preserving the signal in W_0 is therefore not a refinement but a prerequisite for credible inference about network spillovers.

2.2. **Sources of measurement error in economic networks.** The economic literature documents several mechanisms through which the observed adjacency matrix W may depart from the latent network W_0. Many policy-relevant production, trade, and reconstructed financial networks contain numerous weak links and can be dense relative to conventional social networks. Consequently, measurement error may affect a large fraction of their entries. A related source of potentially dense measurement error arises when the adjacency matrix is constructed from geographical or economic distance (LeSage and Pace, 2009; Getis and Aldstadt, 2004): errors in the underlying distance measurem (arising, for example, from centroid approximations, omitted transport links, or bandwidth choices) can propagate across entire rows or columns of W. We discuss several additional sources of measurement error below.

(i) Aggregation and reporting error in input–output networks. A widely studied dense economic network is the input–output table, where the (i, j) entry records the share of commodity i used as an intermediate input in producing commodity j. Acemoglu et al. (2012) establish that propagation multipliers are highly sensitive to the precise pattern of bilateral input shares across all n^2 entries of this matrix. The observed Bureau of Economic Analysis table aggregates transactions across hundreds of heterogeneous firms and products into sector-level flows, and is further subject to reporting inaccuracies, imputation of missing entries, periodic revisions, and reconciliation between the supply and use tables. Given these, the analyst observes a sector-level W that departs from any meaningful sector-level W_0, and the resulting errors propagate along multi-step Leontief chains.[2] Our method takes the sector-level network as primitive and denoises around a sector-level truth. We show in Section 6 that even at the sector level, denoising restores Leontief stability and materially changes estimated cross-sector multipliers.

(ii) Mirror-statistic discrepancies in bilateral trade networks. International trade flows provide a second canonical example of a dense network where measurement error affects every bilateral entry. Trade flows are recorded twice: by the exporting country as exports and by the importing country as imports. These two independent records of the same physical shipment should agree, but in practice they differ substantially and systematically. Fisman and Wei (2004) exploit this bilateral mirror-statistic structure to measure the *evasion gap*, i.e., the log difference between Hong Kong's reported exports to China and China's reported imports from Hong Kong at the six-digit product level and show that a one-percentage-point increase in the tariff rate is associated with a three-percent increase in the evasion gap. This tariff-related underreporting may affect many product categories and can be amplified by product misclassification, as importers reclassify high-tariff goods into lower-tariff categories. The resulting measurement error need not be confined to a small number of trade links and may affect a substantial fraction of the bilateral trade network in a manner correlated with tariff rates. Because tariff schedules are systematically related to trade volumes and revealed comparative advantage, the measurement error E_{ij} tends to correlate with the true $W_{0,ij}$, violating classical errors-in-variables assumptions and biasing estimates of trade-network spillovers.

(iii) Entropy-based reconstruction error in interbank exposure networks. Bilateral interbank exposures form a naturally dense network for studying systemic risk and financial contagion. Each bank's balance sheet records its total interbank lending and borrowing, but the bilateral breakdown, who lent to whom and in what amount, is typically confidential. Regulatory data under the Basel II framework require reporting of exposures only above a threshold of 10% of regulatory capital, so the majority of bilateral links are unobserved (Anand et al., 2015). Researchers and regulators therefore reconstruct the full $n \times n$ bilateral exposure matrix from the observable marginals (total lending and borrowing of each

[2]The literature on aggregation in production networks treats this as a separate methodological problem; see e.g. Barrot and Sauvagnat (2016) and Boehm et al. (2019). Our framework is complementary: it addresses reporting and revision error at whatever level the network is observed, and is silent on the firm-to-sector aggregation step.

bank) using maximum-entropy algorithms. As Upper (2011) and the systematic evaluation of Anand et al. (2015) document, the maximum-entropy solution fills in the matrix as uniformly as possible, implicitly treating the network as if every bank lent to every other bank in proportion to its total activity. This assumption yields a fully dense reconstructed matrix, but one that systematically underestimates concentration and overestimates diversification: in reality, interbank lending is relationship-based and concentrated among a small set of counterparties. The reconstruction error $E = W - W_0$ may therefore be dense. It affects every entry of the matrix, and is potentially correlated with the true bilateral exposures, since banks with large total lending are disproportionately misrepresented. Stress-testing exercises that rely on the reconstructed network underestimate contagion risk precisely because the entropy assumption smooths away the concentrated exposures that drive cascading failure.

(iv) Recall error and endogeneity in reported network links. Even when network data are collected via direct reports rather than algorithmic reconstruction, the observed adjacency matrix is subject to recall error, misunderstanding, and measurement timing mismatches. Lewbel et al. (2024) study economic network models in which some reported binary links are misclassified: true connections are recorded as absent (false negatives) and absent connections are recorded as present (false positives). They establish that this misclassification introduces new sources of endogeneity beyond the standard simultaneity problem in spatial models: the mismeasured adjacency matrix contaminates the spatial lag WY, creating correlation between regressors and structural errors, and invalidating standard instrumental variables based on powers of W. The result is that conventional Two-Stage Least Squares (2SLS) estimators that ignore misclassification are inconsistent even when the fraction of misclassified links is small relative to n, because in a dense network a small misclassification rate still corresponds to a large absolute number of erroneous entries. Panel data settings compound this problem: if the network is observed at lower frequency than the outcome variable, links created or dissolved between survey waves appear as persistent misclassifications correlated with individual fixed effects that cannot be differenced away (Lewbel et al., 2023).

The four mechanisms above are heterogeneous in origin but share the three features highlighted at the end of Section 2.1: the measurement error is dense, correlated across entries through shared sources (aggregation, tariff incentives, reconstruction algorithm, or recall bias), and asymptotically non-negligible. These properties together explain why classical errors-in-variables corrections may be insufficient: they presume sparse or vanishing noise that can be handled by instrumental variable (IV) or measurement-error bias corrections, whereas the dense, correlated errors documented here require a different approach. Imposing a regularization on W_0 (as a low-rank matrix, a sparse matrix, or their combination) provides the identifying restrictions needed to disentangle the systematic signal from the noise, motivating the framework developed in the remainder of the paper.

2.3. **Examples of network structures.** To build intuition, we describe four common economic network structures and show how each fits into the low-rank, sparse, or low-rank plus

sparse framework. For simplicity, throughout this subsection, we assume the network is observed without measurement error, so $W = W_0$.[3]

2.3.1. *Complete network.* We begin by considering one of the simplest examples: a fully connected network. In this setup, the link matrix W_0 is an $n \times n$ matrix of nonzero off-diagonal values, where it is common to set $W_{0,ii} = 0$ to avoid self-loops. The link matrix corresponding to a complete network generated by nonzero constants $\boldsymbol{c} = (c_1, \cdots, c_n)^\top$ is defined as

$$W_0 = \begin{bmatrix} 0 & c_1c_2 & c_1c_3 & \cdots & c_1c_n \\ c_2c_1 & 0 & c_2c_3 & \cdots & c_2c_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_nc_1 & c_nc_2 & c_nc_3 & \cdots & 0 \end{bmatrix},$$

where c_i's are positive constants within the bounded support $[c_{\min}, c_{\max}]$, and they could be viewed as individual characteristics that vary across i. Let $L_0 = \boldsymbol{c}\boldsymbol{c}^\top$, which is a symmetric low-rank matrix with rank 1. Its singular value is $\|\boldsymbol{c}\|_2^2$, and the associated singular vector is $\boldsymbol{c}/\|\boldsymbol{c}\|_2$. Then we could also express $W_0 = L_0 + S_0 = L_0^*$, where the sparse component S_0 is a diagonal matrix whose (i,i)th element is $-c_i^2$ to guarantee no self-loops in W_0. Such a network could be used to capture the common interaction structure driven by individuals or agents who interact through a common market or platform.

2.3.2. *Low-rank network.* Motivated by the structure of the latent factor, we consider the network in which the strength of the connection is determined by $W_{0,ij} = a_i + z_iz_j$ if $i \neq j$, where a_i's are individual characteristics and z_i's are some latent variables.[4] Denote $\boldsymbol{a}$ as a vector consisting of a_i. In this example, $W_0 = L_0 + S_0 = L_0^*$, where $L_0 = \boldsymbol{a}\mathbf{1}_n^\top + \boldsymbol{z}\boldsymbol{z}^\top$ is a matrix of rank no more than 2 with $\boldsymbol{a} = (a_1, \cdots, a_n)$ and $\boldsymbol{z} = (z_1, \cdots, z_n)$, and S_0 is a diagonal matrix whose (i,i)th element is $-L_{0,ii}$ to remove self-loops. Networks of this kind arise in production networks and financial systems, where common shocks or institutional linkages generate pervasive, dense interactions.

2.3.3. *Group network.* Another example of interest is the pure group network, where there are two or more communities, potentially of distinct sizes to avoid eigenvalue multiplicity in the link matrix. Moreover, individuals within the same community are assumed to be connected with each other, while individuals do not connect to others outside their community. For example, suppose $n = 20$, and 6 individuals belong to community 1 and the rest belong to community 2. We define

$$W_0 = \begin{bmatrix} L_1 & \mathbf{0}_{6\times 14} \\ \mathbf{0}_{14\times 6} & L_2 \end{bmatrix} + S_0 = L_0^*,$$

[3]Note that the link matrix or adjacency matrix could be used to describe the connection strength between individuals. If it does not satisfy the bounded row sum assumption, it could be further scaled by its ℓ_∞ norm.

[4]In a more general term, the low rank component can be understood as network effects related to a principal agent, see Galeotti et al. (2020) . The principal agent is formed by connections with respect to many nodes within the network, and the effects are dense.

where $L_1 = (c_1, c_2, \cdots, c_6)^\top (c_1, c_2, \cdots c_6)$ and $L_2 = (c_7, \cdots, c_{20})^\top (c_7, \cdots, c_{20})$, and S_0 is a diagonal matrix such that $S_{0,ii} = -c_i^2$. Note that W_0 still admits the low-rank and sparse decomposition. The low-rank component admits a block structure and its rank equals the number of communities. One of its two singular vectors is proportional to $(0, \cdots, 0, c_7, \cdots, c_{20})$, while the other is proportional to $(c_1, \cdots, c_6, 0, \cdots, 0)$. See Figure 1 for the network graph and the heatmap of W_0. This setting captures trade blocs, regional banking clusters, and sectoral linkages, and the rank of L_0 equals the number of communities. We could also extend it to allow sporadic between-group connections, which are described by few nonzero off-diagonal elements in S_0. Then the general model could be written as $W_0 = L_0 + S_0 = L_0^* + S_0^*$, where the low-rank component captures the within-group connections and the sparse component represents the between-group connections.

2.3.4. *Dominant units.* An interesting example comes from considering networks containing a set of dominant units also known as key players, see Calvó-Armengol et al. (2009); Zenou (2016). One definition of dominant units used in the network literature identifies them as the set of units whose actions have large and persistent effects on the other agents.[5] One way to measure the effect of the unit j on others is to use its weighted out-degree d_j, defined simply by the j-th column sum of W_0: $d_j = \sum_{i=1}^n W_{0,ij}$. We then call the unit j a dominant unit if its weighted out-degree is large or grows as the sample size does. For example, suppose $n = 20$ and there is only one dominant unit, namely the first individual, who is connected to 13 individuals. We assume that each nondominant unit is also connected to its immediate neighbors. Then the link matrix W_0 is specified as below:

$$W_0 = \begin{bmatrix} 0 & w_{1,2} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ w_{2,1} & 0 & w_{2,3} & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ w_{14,1} & 0 & \cdots & w_{14,13} & 0 & w_{14,15} & \cdots & 0 & 0 \\ 0 & \cdots & \cdots & 0 & w_{15,14} & 0 & w_{15,16} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & w_{20,19} & 0 \end{bmatrix}.$$

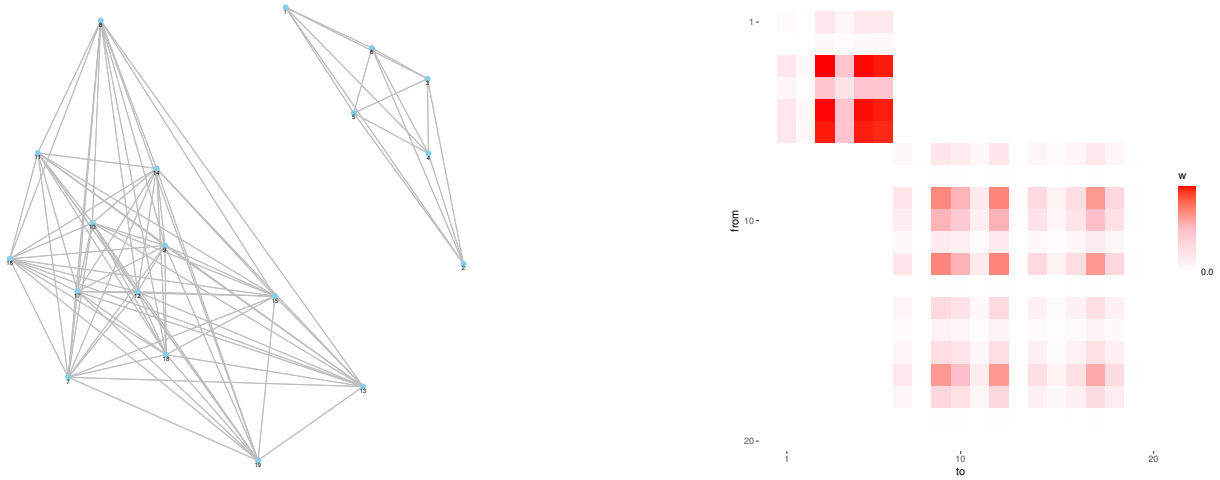
See Figure 2 for the network graph and the heat map of W_0.

This network is naturally sparse. The number of nonzero entries in each row is bounded, although the column corresponding to the dominant unit may contain a growing number of nonzero entries. While this column can be represented as a sparse rank-one matrix, the full adjacency matrix need not be low-rank because it also contains links among neighboring units. We therefore treat W_0 as a sparse matrix. Appendix D provides further discussion of alternative low-rank-plus-sparse representations and their identification.

2.4. **Empirical illustration: the US Input-Output Table.** To demonstrate that measurement error in dense economic networks has economically meaningful consequences, and that the $L + S$ decomposition provides an effective device for recovering the underlying true

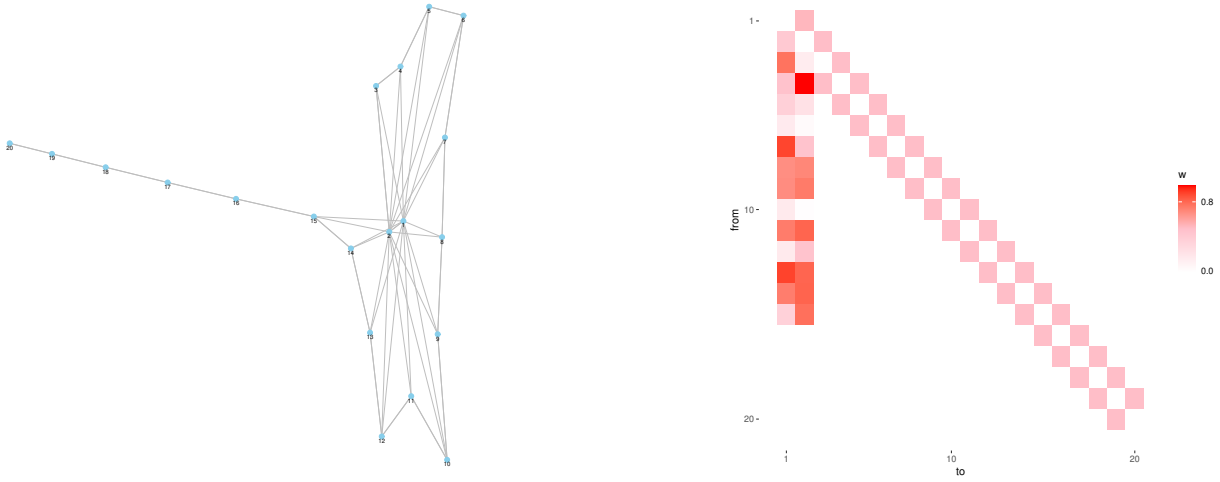
[5]We follow the definition of dominant units as in Pesaran and Yang (2020, 2021); Lee et al. (2023).

FIGURE 1. Graphical representations of simulated adjacency matrices associated with different groups



Notes: Examples for the weight matrix formed by individuals who belong to different groups. The left one is the network graph and the right one is the heat map of the adjacency matrix. There are two groups, one of which has 6 members and the other has 14 members. Only members within the same group have connections, and the connection strength is randomly simulated.

FIGURE 2. Graphical representations of the simulated adjacency matrix with dominant units.



Notes: Examples for the weight matrix with dominant units. The left one is the network graph and the right one is the heat map of the adjacency matrix. There are 20 individuals and each individual connects to its neighbors with strength set by 0.25. The first one is a key player that influences 13 other individuals, and the influence strengths are generated by Uniform$(0, 1)$ distributed random variables.

network structure, we revisit the 2002 US Input-Output table from Graham and de Paula (2020), compiled by the Bureau of Economic Analysis and studied by Acemoglu et al. (2012). It contains an $n \times n$ network of $n = 417$ sectors (after dropping rows and columns with zero direct input requirements), where entry (i, j) indicates the share of sector i's output used in sector's j production. This network is dense by construction(nearly all sectors interact) making it a canonical example where sparsity cannot be assumed and measurement error is pervasive. We treat the observed W as a noisy version of the true network W_0 and apply our Algorithm 1 (Appendix A) to recover W_0.

To assess the quality of the denoising, we compare three candidate representations of W_0, a purely sparse estimate $\widehat{S}$, a purely low-rank estimate $\widehat{L}$, and the combined version $\widehat{W} = \widehat{L}+\widehat{S}$. The purely sparse estimate satisfies $\|\widehat{S} - W\|_{\max} = 0.025$ and $\|\widehat{S} - W\|_2 = 0.470$, indicating that the tiny elements might have non-negligible aggregation effects. The purely low-rank estimate satisfies $\|\widehat{L} - W\|_{\max} = 0.999$ and $\|\widehat{L} - W\|_2 = 1.214$, indicating that the low-rank structure alone might fail to capture occasionally large idiosyncratic links. The combined estimate achieves $\|\widehat{W} - W\|_{\max} = 0.024$ and $\|\widehat{W} - W\|_2 = 0.183$, materially outperforming either alternative. This superiority is further confirmed by eigenvector alignment. The inner product between the leading eigenvectors of W and $\widehat{W}$ is 0.981, compared with 0.526 for $\widehat{S}$ and 0.222 for $\widehat{L}$, indicating that the low-rank and sparse decomposition might best preserve the spectral structure of the observed network. From Figure 3, we could see that the sparse component $\widehat{S}$ captures large idiosyncratic bilateral links, while the low-rank component $\widehat{L}$ exhibits a distinctive stripe pattern, revealing that pervasive but individually weak linkages aggregate into systematic structure that a purely sparse representation would miss.

To quantify the econometric consequences of measurement error, we compare the Leontief diffusion matrices $(I_n - \lambda\widehat{W})^{-1}$, $(I_n - \lambda W)^{-1}$, and $(I_n - \lambda\widehat{S})^{-1}$ following Acemoglu et al. (2016) with $\lambda = 0.7$. Leaving measurement errors uncorrected produces a 9% deviation in the spectral norm relative to our estimators:

$$\frac{\|(I_n - 0.7\widehat{W})^{-1} - (I_n - 0.7W)^{-1}\|_2}{\|(I_n - 0.7\widehat{W})^{-1}\|_2} = 9\%.$$

Misspecifying the structure of measurement errors by treating it as a purely sparse discarding the low-rank component compounds the distortion substantially, producing a 68% deviation:

$$\frac{\|(I_n - 0.7\widehat{W})^{-1} - (I_n - 0.7\widehat{S})^{-1}\|_2}{\|(I_n - 0.7\widehat{W})^{-1}\|_2} = 68\%.$$

These results demonstrate that measurement error in dense networks is not merely a theoretical concern: it materially distorts estimated propagation channels. Hence it is essential for reliable inference to correctly specify the error structure accounting for both its sparse idiosyncratic component and its low-rank systematic component. Figure 4 illustrates this point visually: the stripe pattern present in the full diffusion matrix $(I_n - 0.7\widehat{W})^{-1}$ is eliminated when the low-rank component is omitted, confirming that misspecification of network structure biases spillover estimates in economically meaningful ways.

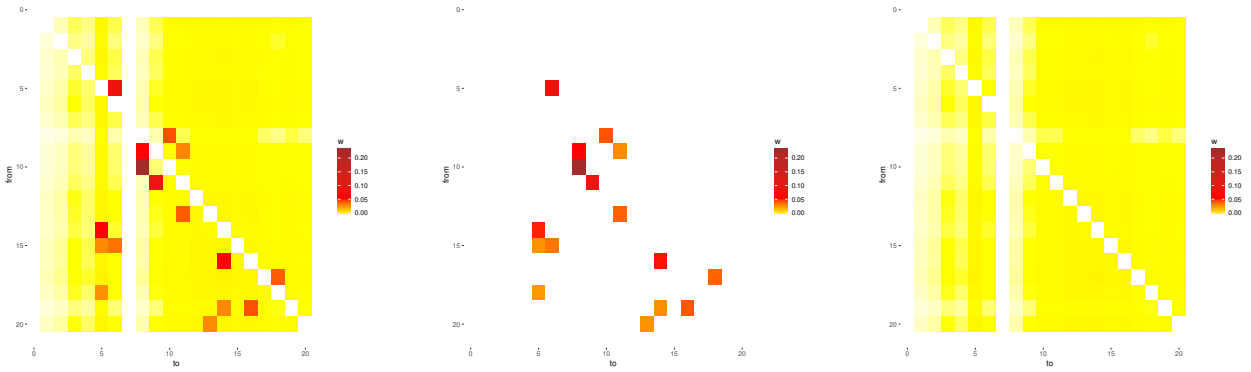


FIGURE 3. Heatmaps of 20 sectors from the 2002 Input-Output Table W (left), estimated sparse component $\widehat{S}$ (centre), and estimated low-rank component $\widehat{L}$ (right). The stripe pattern in $\widehat{L}$ reveals pervasive systematic structure missed by a purely sparse representation.

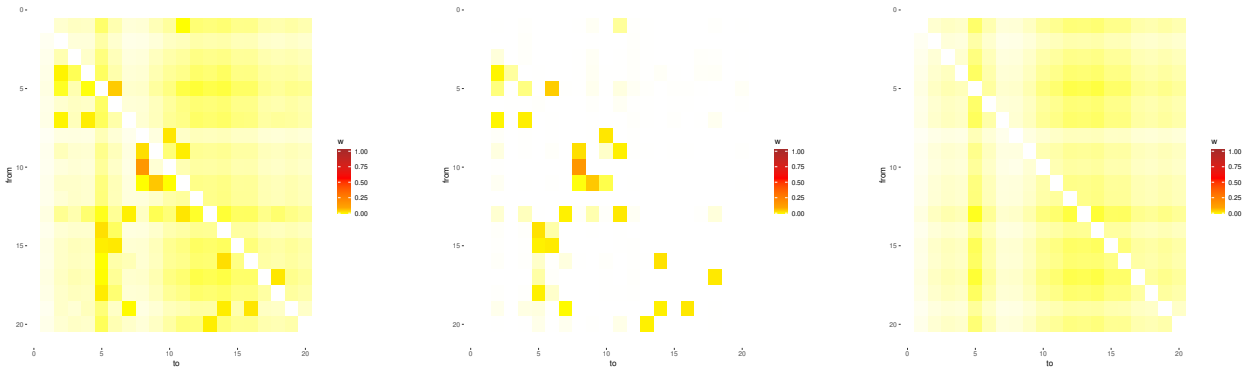


FIGURE 4. Heatmaps of 20 sectors from the diffusion matrix $(I_n - 0.7\widehat{W})^{-1}$ (left), $(I_n - 0.7\widehat{S})^{-1}$ (centre), and the difference (right). Omitting the low-rank component eliminates the stripe pattern and distorts estimated spillover propagation by 68%.

3. METHODOLOGY

Without additional restrictions, the decomposition need not be unique or economically interpretable. Assumption 1 imposes the conditions required for its mathematical identification. The joint regularization on L_0 and S_0 then provides the structure needed to separate the latent network W_0 from the noise E. We provide the modeling definitions and estimation procedures we propose to tackle this problem. We offer additional assumptions and asymptotic results in Section 4.

3.1. **Estimation.**

We formally describe here the social/spatial interaction model of interest. We consider a panel data scenario, but assume the true underlying network does not vary with time. Denote data points as $\{Y_t, X_t\}_{t=1,\cdots,T}$, where Y_t is the $n \times 1$ outcome vector and X_t is the $n \times K$ design matrix of exogenous covariates. We assume that the true model is generated by the unobserved $n \times n$ adjacency matrix W_0 encoding the true connection, and

that we have a noisy or contaminated observation, such that we can write

$$\begin{aligned} Y_t &= \lambda_0 W_0 Y_t + \alpha + \iota_t \mathbf{1}_{n\times 1} + X_t\beta_0 + W_0 X_t \gamma_0 + \varepsilon_t \\ W &= W_0 + E, \end{aligned} \tag{3}$$

where λ_0 represents the scalar (social / spatial) interaction or spillover effect, $\beta_0 \in \mathbb{R}^K$ are standard slope coefficients, $\gamma_0 \in \mathbb{R}^K$ are contextual effects (spillover effects of the observable characteristics X_t), $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$ is the vector of individual fixed effects, ι_t is a time fixed effect at period t, $\mathbf{1}_{n\times 1}$ is an $n \times 1$ vector of ones, and ε_t is an n-dimensional vector of idiosyncratic disturbances, assumed for now to be independent and identically distributed over units i and time t. We assume the covariates X_t to be exogenous to the outcome unobservables and network formation process such that $\mathbb{E}[\varepsilon_{it} \mid X_t, W_0X_t, \cdots, W_0^d X_t] = 0$, for a small integer $d > 0$, $i = 1, \cdots, n$ and $t = 1, \cdots, T$. Crucially, we allow for the measurement errors to be endogenous in the sense that $\mathbb{E}[\varepsilon_{it} \mid E] \neq 0$, which introduces additional bias into conventional estimates when W_0 is replaced by W (Boucher and Houndetoungan, 2025; Hardy et al., 2019; Lewbel et al., 2024).

Collect all parameters of interest into $\theta_0 \coloneqq (\lambda_0, \beta_0^\top, \gamma_0^\top)^\top \in \mathbb{R}^{2K+1}$. When covariates X_t are exogenous and the adjacency matrix W_0 is correctly observed, the literature proposes to select m linearly independent columns from $[X_t, W_0X_t, \ldots, W_0^dX_t]$, for a small positive integer d, as instruments for the regressors $\bar{X}_t(W_0) = [W_0Y_t, X_t, W_0X_t]$ in estimating θ_0 (Kelejian and Prucha, 1998; Lee, 2003, 2007a).

The degree d should be chosen such that the number of retained instruments is at least $2K+1$ to satisfy the order condition, i.e. $2K+1 \leq m \leq (d+1)K$ (see Assumption 6). For an arbitrarily pre-specified $n \times n$ adjacency matrix M, let $Z_t(M)$ be the $n \times m$ matrix that collects the linearly independent columns from $[X_t, MX_t, \ldots, M^dX_t]$ and let $X_t(W)$ be the $n \times (2K+1)$ matrix that collects the regressors $[WY_t, X_t, WX_t]$. Note that $X_t(W)$ uses the observed adjacency matrix W as opposed to the true W_0 showing up in model (3).

By defining $\varepsilon_t(\theta, W) \coloneqq Y_t - X_t(W)\theta$ and thus $\varepsilon_t = \varepsilon_t(\theta_0, W_0)$, we obtain the population moment condition $\mathbb{E}[Z_t^\top(M)\varepsilon_t] = \mathbf{0}_{m\times 1}$ for any exogenous M. Thus we have the empirical moment conditions:

$$\bar{g}_{nT}(\theta, W, M) \coloneqq \frac{1}{nT}\sum_{t=1}^{T} Z_t(M)^\top \varepsilon_t(\theta, W)\,. \tag{4}$$

The distinction made between the matrix used to construct the regressors $X_t(W)$ and the instruments $Z_t(M)$ is necessary to define our supervised estimator below. When $M = W$ such that no distinction is necessary, we simply write $\bar{g}_{nT}(\theta, W) = \bar{g}_{nT}(\theta, W, W)$. For a given $m \times m$ moment-weighting matrix $\widehat{\Lambda}_{nT}^{-1}$, we define the sample GMM objective function as

$$J_{nT}(\theta, W, M; \widehat{\Lambda}_{nT}^{-1}) \coloneqq (nT)\bar{g}_{nT}(\theta, W, M)^\top \widehat{\Lambda}_{nT}^{-1}\, \bar{g}_{nT}(\theta, W, M).$$

When fixed effects are present, within transformation [6] is applied to Y_t, $X_t(W)$, and $Z_t(M)$ before constructing the GMM objective; we suppress this transformation in the notation that follows. The extension to two-way fixed effects is discussed in Remark 3.

Finally, we stack all time series observations into the nT-dimensional outcome vector $Y \coloneqq [Y_1^\top, \ldots, Y_T^\top]^\top$, the $nT \times (2K+1)$ design matrix $X(W) \coloneqq [X_1(W)^\top, \ldots, X_T(W)^\top]^\top$, and the $nT \times m$ matrix of instruments $Z(M) \coloneqq [Z_1(M)^\top, \ldots, Z_T(M)^\top]^\top$. Using this notation, we can express the GMM estimator as

$$\begin{aligned} \widehat{\theta}_{\mathrm{GMM}}(W, M) &\coloneqq \arg\min_{\theta \in \Theta} J_{nT}(\theta, W, M; \widehat{\Lambda}_{nT}^{-1}) \\ &= \left[X(W)^\top Z(M) \widehat{\Lambda}_{nT}^{-1} Z(M)^\top X(W) \right]^{-1} X(W)^\top Z(M) \widehat{\Lambda}_{nT}^{-1} Z(M)^\top Y. \end{aligned} \tag{5}$$

We define $\widehat{\theta}_{\mathrm{GMM}} = \widehat{\theta}_{\mathrm{GMM}}(W, W)$. In practice, we can obtain a two-step GMM estimator by using the weight matrix $\widehat{\Lambda}_{nT}^{-1}$. [7] representing an initial consistent estimator obtained by using the identity $\widehat{\Lambda}_{nT}^{-1} = I_m$ as the first-step GMM weight matrix.

The rates of our estimators along with all formal assumptions and theoretical results, are presented in Section 4. If W_0 were observed without error, Kelejian and Prucha (1998) and Lee (2007a), for example, guarantee that the standard GMM estimator $\widehat{\theta}_{\mathrm{GMM}}$ is consistent for θ_0 in (3). However, when measurement error is present, small elementwise noises in E can accumulate, leading to inconsistent estimation of the regression coefficients and spillover effects.

Therefore, when measurement error exists, we need a way to correct for this additional noise in our estimation. We make use of the information afforded by the low-rank plus sparse plus error decomposition as in (1) to obtain an estimate $\widehat{W}$. Theoretically, we quantify the estimation accuracy of $\widehat{W}$ and its asymptotic influence when used as an input for estimating θ_0. We now describe the details leading to our two proposed estimators. Our first option is the "plug-in estimator," and the other is the "supervised estimator," both defined in the following subsection.

3.2. **The plug-in estimator and the supervised estimator.** Our proposed estimators make use of the structure for the adjacency W in (3) by using penalized convex low-rank and sparse estimation. Our first proposal is simply a plug-in estimator that first produces a denoised estimate for W_0 and uses this estimate in place of the observed version when

[6] The within transformation eliminates the individual effects α by subtracting unit-specific time averages: writing $Y_{\cdot} \coloneqq \frac{1}{T}\sum_{t=1}^{T} Y_t$, the transformed outcome is $\ddot{Y}_t \coloneqq Y_t - Y_{\cdot}$, and $X_t(W)$ and $Z_t(M)$ are transformed analogously. Since α is constant over t, it is differenced out by this demeaning; the time effects ι_t are removed by the additional cross-sectional demeaning discussed in Remark 3. The moment conditions (4) constructed from the demeaned variables remain valid.

[7] In practice, we obtain a feasible two-step GMM estimator by first computing $\widetilde{\theta}$ using the identity weighting matrix and then using $\widehat{\Lambda}_{nT}^{-1}$, where

$$\widehat{\Lambda}_{nT} = \frac{1}{nT} \sum_{t=1}^{T} Z_t(W)^\top \operatorname{diag}\left(\widehat{\varepsilon}_{1t}^2, \ldots, \widehat{\varepsilon}_{nt}^2\right) Z_t(W), \qquad \widehat{\varepsilon}_t = Y_t - X_t(W)\widetilde{\theta}.$$

calculating the GMM estimator. Recall that A is an $n \times n$ matrix and A_{ij} denotes its (i,j)th element. We first minimize the following loss function to denoise the adjacency matrix by taking advantage of its low-rank plus sparse structure:

$$(\widehat{L}, \widehat{S}) \coloneqq \arg \min_{L \in \mathbb{R}^{n\times n}, S \in \mathbb{R}^{n\times n}, L_{ii}+S_{ii}=0, 1\leq i\leq n} \frac{1}{2}\|W - L - S\|_F^2 + \nu_n \|L\|_* + \tau_n \sum_{i\neq j} |S_{ij}|, \tag{6}$$

where $\|A\|_F$ is the Frobenius norm of A, $\|L\|_*$ is the nuclear norm of L, ν_n and τ_n are positive tuning parameters. Similar to (Candès et al., 2011; Cao et al., 2017), we propose to use Algorithm 1 to solve the optimization problem in (6). Defining $\widehat{W} \coloneqq \widehat{L} + \widehat{S}$, our plug-in estimator is then simply given as in (5) to replace W with $\widehat{W}$.

$$\begin{aligned}\widehat{\theta}_p(\widehat{W}) &\coloneqq \arg\min_{\theta\in\Theta} J_{nT}(\theta, \widehat{W}, \widehat{W}; \widehat{\Lambda}_{nT}^{-1}) \\ &= \left[X(\widehat{W})^\top Z(\widehat{W}) \widehat{\Lambda}_{nT}^{-1} Z(\widehat{W})^\top X(\widehat{W})\right]^{-1} X(\widehat{W})^\top Z(\widehat{W}) \widehat{\Lambda}_{nT}^{-1} Z(\widehat{W})^\top Y. \end{aligned} \tag{7}$$

Our second proposal, the supervised estimator, minimizes the combination of the GMM objective function and the penalized square loss based on the low-rank plus sparse decomposition. The estimator $(\widehat{\theta}_s, \widehat{L}_s, \widehat{S}_s)$ is defined as

$$\begin{aligned}&\arg \min_{\theta\in\Theta, L\in\mathbb{R}^{n\times n}, S\in\mathbb{R}^{n\times n}, L_{ii}+S_{ii}=0, 1\leq i\leq n} \xi_{nT} J_{nT}(\theta, L+S, M; \widehat{\Lambda}_{nT}^{-1}) \\ &+\frac{1}{2}\|W - L - S\|_F^2 + \nu_{nT}\|L\|_* + \tau_{nT}\sum_{i\neq j}|S_{ij}|, \end{aligned} \tag{8}$$

where ξ_{nT} is an additional nonnegative tuning parameter necessary to balance the variations contained in the observed adjacency and the outcome variables. It is worth noting that Cai et al. (2021) also study a supervised method based on a purely low-rank structure for analyzing network centrality. Their approach, which focuses on the leading singular vectors, differs from ours in that it does not incorporate a spatial regression.

A full procedure for solving (8) is given in Algorithm 2 in Appendix B. Note that, as a by-product of the algorithm, the estimator produces supervised low-rank and sparse decompositions that incorporate information from the outcome and covariates.

4. Asymptotic theory

In this section, we discuss the theoretical properties of our estimation procedures. Our main interest is to quantify the impact of using an estimated adjacency matrix on recovering both the true adjacency and the spillover parameters of interest. We shall use the following notation: Let A and B be matrices of dimensions $n \times n$. For $i = 1, \ldots, n$, we use $\lambda_i(A)$ and $\sigma_i(A)$ denote the i-th largest eigenvalue and the i-th largest singular value of the matrix A. Let $\lambda_{\min}(A)$ denote the minimum eigenvalue of A. Let $\|.\|_2$ denote the Euclidean-norm of a vector. We use $\|A\|_1$, $\|A\|_2$, $\|A\|_\infty$, $\|A\|_{\max}$ to denote the l_1-norm, l_2-norm, l_∞ norm, max norm respectively. $|A|_{1,1} = \sum_{i=1}^n \sum_{j=1}^n |A_{ij}|$. Let $1 \leq r \leq n$ be the rank of A. For a matrix

$A \in \mathbb{R}^{n\times n}$, let:

$$m_r(A) := \max_{1\le i\le n} \sum_{j=1}^{n} \mathbb{I}(A_{ij} \neq 0), \qquad m_c(A) := \max_{1\le j\le n} \sum_{i=1}^{n} \mathbb{I}(A_{ij} \neq 0),$$

$$m_s(A) := \sum_{i,j} \mathbb{I}(A_{ij} \neq 0), \qquad \deg_{\max}(A) := \max\{m_r(A), m_c(A)\}.$$

We use e_i to denote the standard basis vector, i.e., the ith column of the identity matrix, and $\mathrm{vec}(A)$ to denote the column vectorization of matrix A. For sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we denote $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n/b_n \le C$ for all n, and denote $a_n = o(b_n)$ (resp. $a_n \asymp b_n$) if $a_n/b_n \to 0$ or $a_n \ll b_n$ (resp. $a_n \lesssim b_n$ and $b_n \lesssim a_n$). We use $\xrightarrow{d}$ and $\to_p$ to indicate convergence in distribution and in probability, respectively. We use $X_n = O_p(a_n)(\lesssim_p a_n)$ to indicate X_n/a_n is bounded in probability, i.e. for any $\varepsilon > 0$, there exists $M(\varepsilon) > 0$ such that $\mathbb{P}(|X_n/a_n| > M(\varepsilon)) < \varepsilon$ for all $n \in \mathbb{N}$. We assume $\min(n, T) \to \infty$ or $n \to \infty$. In what follows, Y_t, X_t, $Z_t(M)$, and ε_t denote the within-transformed versions of the model objects, with individual fixed effects removed. The asymptotic results carry through under the two-way transformation that additionally removes time fixed effects; see Remark 3.

Assumption 1 (True weight)**.** (i) (Low-rank component) L_0 is an $n \times n$ low-rank matrix with $\mathrm{rank}(L_0) = r$, $\min(m_r(L_0), m_c(L_0)) \to \infty$, and $\|L_0\|_{\max} = O(1/\sqrt{m_r(L_0)m_c(L_0)})$.

(ii) (Sparse component) S_0 is an $n \times n$ sparse matrix with $\|S_0\|_{\max} = O(1)$.

(iii) (Joint identification) $\frac{m_s(S_0)}{m_r(L_0)m_c(L_0)} \to 0$ and $\frac{\deg_{\max}(S_0)^2}{\min\{m_r(L_0), m_c(L_0)\}} \to 0$.

Assumption 1 collects conditions on W_0. Part (i) requires the low-rank signal to be pervasive but individually weak: the condition $\|L_0\|_{\max} = O(1/\sqrt{m_r(L_0)m_c(L_0)})$ allows L_0 to be distinguished from S_0; it fails when L_0 is spiked (as in the dominant-units example), in which case the low-rank part is better absorbed into S_0. The rank r may diverge with n provided the relevant rates still vanish. Part (ii) requires S_0 to be bounded and Part (iii) allows S_0 to have relatively few nonzero links compared with the dense part.

Hence the decomposition $W_0 = L_0 + S_0$ is uniquely determined when n is large, making the rank parameters r and $m_s(S_0)$ well-defined. We stress this is a *mathematical* requirement only: $\widehat{\theta}_p$ uses $\widehat{W} = \widehat{L} + \widehat{S}$ as a single object, so we rely on no economic interpretation of the components, and inference about θ is robust to weak identification of the decomposition. Part (iii) holds trivially when $L_0 = \mathbf{0}_{n\times n}$ or $S_0 = \mathbf{0}_{n\times n}$, and as a side implication limits the sum over individual columns of W_0, yielding $\|W_0\|_1 = o(\sqrt{n})$, which is used in the rate analysis of $\widehat{\theta}_p(\widehat{W})$. The following assumption imposes conditions on the measurement error E in the observed matrix W, required for the large-sample analysis as $n \to \infty$.

Assumption 2 (Noise)**.** (i) E is an $n \times n$ noise matrix such that $E_{ii} = 0$ for $i = 1, \cdots, n$.

(ii) E satisfies $\|E\|_2 \le w_{2n}$ with probability greater than $1 - p_n$, where $p_n = o(1)$.

(iii) E satisfies $\|E\|_{\max} \leq w_{\max n}$ with probability greater than $1 - p_n$, where $p_n = o(1)$.

Assumption 2 is mild. It does not require the entries of E to be independent or identically distributed (a useful feature in spatial and social network contexts), where measurement errors often exhibit dependence. It also does not require w_{2n} or $w_{\max n}$ to vanish; these are sequences that bound the spectral and entrywise norms of E, and may grow with n. Whether the bounds delivered by Theorems 1–2 are informative depends on joint conditions involving $(w_{2n}, w_{\max n}, r, m_s(S_0), n)$, discussed after each theorem.

The roles of w_{2n} and $w_{\max n}$ are complementary. The spectral norm bound w_{2n} controls the aggregate noise level and governs how well the low-rank component L_0 can be separated from E. The entrywise maximum $w_{\max n}$ controls the largest individual perturbation and governs the recovery of S_0. We illustrate how E looks like in the following examples.

Example 1 (A sample-covariance error matrix). Suppose the error arises from estimating a covariance matrix: let $z_1, \ldots, z_T$ be independent mean-zero Gaussian vectors in $\mathbb{R}^n$ with covariance W_z, whose eigenvalues are positive and bounded away from 0 and ∞. Let $\widetilde{E} = T^{-1}\sum_{t=1}^{T} z_t z_t^\top - W_z$ be the sampling error. To respect the no-self-loop convention we take $E = \widetilde{E} - \mathrm{diag}(\widetilde{E})$, so $E_{ii} = 0$ and Assumption 2(i) holds. Since $\|\mathrm{diag}(\widetilde{E})\|_2 = \max_i |\widetilde{E}_{ii}| = O_p(\sqrt{\log n/T})$ is dominated by $\|\widetilde{E}\|_2$, removing the diagonal leaves the rates unchanged. By Theorem 6.5 of Wainwright (2019), taking $\eta = \sqrt{n/T}$ gives $\|E\|_2 \leq C\|W_z\|_2\{\sqrt{n/T} + n/T\} =: w_{2n}$ with probability $\geq 1 - c_1\exp(-c_2 n)$; taking $\eta = \sqrt{\log(n)/T}$ and applying a union bound gives $\|E\|_{\max} \leq C(\max_i W_{z,ii})\{\sqrt{\log(n)/T} + \log(n)/T\} =: w_{\max n}$ with probability $\geq 1 - 2n^{-c_3}$. Thus, Assumption 2 holds with $w_{2n} \asymp \sqrt{n/T} + n/T$ and $w_{\max n} \asymp \sqrt{\log(n)/T} + \log(n)/T$. If $n/T \to 0$, these rates simplify to $w_{2n} \asymp \sqrt{n/T}$ and $w_{\max n} \asymp \sqrt{\log(n)/T}$, and both vanish, with the spectral bound dominating the entrywise one — the aggregate noise exceeds the worst individual entry, as is typical in dense networks. ■

Example 2 (Dense Gaussian noise: comparison with Lewbel et al. (2024)). Suppose E_{ij} are i.i.d. $\mathbb{N}(0, \sigma_n^2)$ for $i \neq j$ and $E_{ii} = 0$, with σ_n possibly diminishing in n. As in Example 1, applying the operator-norm bound of Wainwright (2019) and a union bound over the entries gives, with probability approaching 1, $\|E\|_2 \leq 4\sigma_n\sqrt{n} =: w_{2n}$ and $\|E\|_{\max} \leq 8\sigma_n\sqrt{\log n} =: w_{\max n}$. Assumption 2 holds with $w_{\max n} = 8\sigma_n\sqrt{\log n}$ and $w_{\max n} \to 0$ whenever $\sigma_n = o(1/\sqrt{\log n})$. By contrast, $\mathbb{E}|E|_{1,1} \asymp n^2\sigma_n$. The condition $|E|_{1,1} = o_p(n)$ used by Lewbel et al. (2024) requires $\sigma_n = o(1/n)$. For $1/n \ll \sigma_n \ll 1/\sqrt{n}$, the spectral-norm condition $w_{2n} \to 0$ ensures the GMM estimator $\widehat{\theta}_{GMM}(W)$ is consistent (cf. Theorem 2 b), while their $|E|_{1,1} = o_p(n)$ condition fails. Our estimator $\widehat{\theta}_p(\widehat{W})$ further sharpens the rate via dimension discounting. For $\sigma_n \geq c > 0$, the available bound for the GMM estimator does not guarantee consistency. More generally, the plug-in estimator is consistent whenever the dimension-discounted rate condition stated in Theorem 2(a) holds. ■

We next show the properties $\widehat{L}$ and $\widehat{S}$ corresponding to (6). Define

$$a_n \coloneqq \begin{cases} \{m_r(L_0)m_c(L_0)\}^{-1/2}, & L_0 \neq \mathbf{0}_{n\times n}, \\ 0, & L_0 = \mathbf{0}_{n\times n}. \end{cases}$$

Define the rates

$$\begin{aligned} \mathcal{R}_F(W_0, E) &= \max\{r, 1\}w_{2n}^2 + m_s(S_0)w_{\max n}^2 + m_s(S_0)a_n^2, \\ \mathcal{R}_2(W_0, E) &= w_{2n}^2 + m_s(S_0)w_{\max n}^2 + m_s(S_0)a_n^2, \\ \mathcal{R}_s(W_0, E) &= (w_{\max n} + a_n)\sqrt{m_s(S_0)}. \end{aligned}$$

Let

$$\begin{aligned} B_0 &= (I_n - \lambda_0 W_0)^{-1}, & \rho_n &= w_{2n} + \mathcal{R}_s(W_0, E), \\ s_n &= \mathcal{R}_s(W_0, E)\sqrt{m_s(S_0)}, & \kappa_n &= 1 + \|W_0\|_2 + \rho_n, \end{aligned}$$

and define the dimension-discounted rate

$$\mathcal{R}_{nT}^* \coloneqq (1 + \|B_0\|_2)\kappa_n^{d+1}\frac{(r \vee 1)\rho_n + s_n}{n}.$$

Theorem 1. *Suppose Assumptions 1–2 hold. Let the tuning parameters satisfy*

$$\nu_n = C_\nu w_{2n}, \qquad \tau_n = C_\tau(w_{\max n} + a_n),$$

for sufficiently large constants $C_\nu, C_\tau > 0$.

(a) Frobenius bound. *The two-step estimator satisfies*

$$\|\widehat{L} - L_0\|_F^2 + \|\widehat{S} - S_0\|_F^2 \lesssim_p \mathcal{R}_F(W_0, E), \tag{9}$$

$$|\widehat{S} - S_0|_{1,1} \lesssim_p \mathcal{R}_s(W_0, E)\sqrt{m_s(S_0)} + n\{w_{2n} + \mathcal{R}_s(W_0, E)\}. \tag{10}$$

(b) Sharpened spectral-norm bound. *The two-step estimator satisfies*

$$\|\widehat{L} - L_0\|_2^2 + \|\widehat{S} - S_0\|_2^2 \lesssim_p \mathcal{R}_2(W_0, E).$$

The rates $\mathcal{R}_F$, $\mathcal{R}_2$, and $\mathcal{R}_s$ share a common structure with three error sources: a term in w_{2n} about low-rank recovery; a term $m_s(S_0)\, w_{\max n}^2$ related to $m_s(S_0)$ nonzero sparse entries, and a term involving $\|L_0\|_{\max}$. The factor r in $\mathcal{R}_F$ is replaced by unity in $\mathcal{R}_2$, reflecting the rank discount of Theorem 9.24 in Wainwright (2019). Both rates parallel Corollary 1 of Agarwal et al. (2012) and Theorem 9.19 of Wainwright (2019). Note that estimator of L_0^* and S_0^* inherits the theoretical properties of the corresponding estimators of L_0 and S_0. For instance, since $\widehat{L}^* = \widehat{L} - \mathrm{diag}(\widehat{L})$, we have $\|\widehat{L}^* - L_0^*\|_2 = O_p(\|\widehat{L} - L_0\|_2)$, which implies that estimation accuracy is preserved under the diagonal adjustment.

The conditions required for first-step and second-step consistency differ in an essential way. For the L+S decomposition $(\widehat{L}, \widehat{S})$ to be consistent in operator norm using the bounds in Assumption 2, i.e., $\|\widehat{L} - L_0\|_2 \to_p 0$ and $\|\widehat{S} - S_0\|_2 \to_p 0$, we require

$$w_{2n}^2 \to 0, \qquad m_s(S_0)w_{\max n}^2 \to 0, \qquad m_s(S_0)a_n^2 \to 0.$$

in keeping with the standard $L+S$ literature. Frobenius consistency, $\|\widehat{L}-L_0\|_F \to_p 0$ and $\|\widehat{S}-S_0\|_F \to_p 0$, requires the stronger condition $\max\{r,1\}w_{2n}^2 \to 0$ in place of $w_{2n}^2 \to 0$, while the other two conditions remain unchanged. Consistency of the spillover estimator $\widehat{\theta}_p(\widehat{W})$, $\|\widehat{\theta}_p(\widehat{W})-\theta_0\|_2 \to_p 0$, requires only weaker dimension-discounted conditions, given in Theorem 2 below.

Next, define $\widehat{W} \coloneqq \widehat{L}+\widehat{S}$, where $(\widehat{L},\widehat{S})$ is the estimator in Theorem 1. We use $\widehat{W}$ as the working matrix for estimating the model in (3). We now impose the conditions used to analyze the spillover-parameter estimators.

Assumption 3 (True parameters)**.** (i) The parameter space $\Theta = \mathcal{C}\times\mathcal{B}\times\Gamma$ is a compact subset of $\mathbb{R}\times\mathbb{R}^K\times\mathbb{R}^K$ for some finite dimension K.

(ii) The true value, denoted as $\theta_0 = (\lambda_0, \beta_0^\top, \gamma_0^\top)^\top$, lies in the interior of the parameter space Θ. Moreover, $\lambda_0\beta_0+\gamma_0 \neq \mathbf{0}_{K\times 1}$.

Assumption 4 (Adjacency matrix)**.** W_0 is a non-stochastic spatial weight matrix satisfying

(i) The diagonal elements of W_0 are 0, such that $W_{0,ii}=0$ for all $i=1,\dots,n$.

(ii) I_n, W_0, W_0^2 are linearly independent.

(iii) Suppose that J is fixed and that the absolute column sums of W_0 are uniformly bounded in n, except possibly for those of its first J columns. Let $W_{0,u}$ contain the first J columns of W_0 and zeros elsewhere, and let $W_{0,b}=W_0-W_{0,u}$. When $J=0$, set $W_{0,u}=\mathbf{0}_{n\times n}$ and $W_{0,b}=W_0$. Assume that there exist constants $c_\infty, c_1>0$, independent of n, such that

$$\sup_{\lambda\in\mathcal{C}}|\lambda|\|W_0\|_\infty \le 1-c_\infty, \qquad \sup_{\lambda\in\mathcal{C}}|\lambda|\|W_{0,b}\|_1 \le 1-c_1.$$

Assumption 3 imposes standard conditions on the true parameters. The assumption $\lambda_0\beta_0+\gamma_0\neq\mathbf{0}_{K\times1}$ is imposed to guarantee the identification of all the parameters in the model (see, e.g., Bramoullé et al. (2009)). Assumption 4 (i) requires that the true adjacency matrix W_0 have zero diagonal, which rules out self-loops. Regarding Assumption 4 (ii), the matrices I_n, W_0, and W_0^2 are linearly independent, i.e. there do not exist any nontrivial scalars $\alpha_0,\alpha_1,\alpha_2$ such that

$$\alpha_0 I_n+\alpha_1 W_0+\alpha_2 W_0^2 = \mathbf{0}_{n\times n}.$$

Assumption 4(iii) restricts the number of dominant units to a fixed finite value and imposes uniform stability. The first stability condition guarantees that $I_n-\lambda W_0$ is invertible uniformly over $\lambda\in\mathcal{C}$.

Assumption 5 (Spatial errors)**.** The primitive disturbances ε_{it} are independently and identically distributed across i and t with mean 0, variance $\sigma_0^2>0$ and $\mathbb{E}\,|\varepsilon_{it}|^{2+\delta}<\infty$ for a constant $\delta>2$.

Assumption 6 (Instrumental variables)**.** (i) Define the covariates $x_{t,i}=(x_{ti1},\dots,x_{tiK})^\top$. Suppose $\sup_{n,t,i,\,1\le l\le K}\mathbb{E}|x_{til}|^{2+\delta}<\infty$ for some $\delta>2$. For each i, $\{x_{t,i}: t\in\mathbb{Z}\}$ is strictly

stationary and strongly mixing over t. The mixing coefficients are uniform in n and i: there exist constants $C < \infty$ and $\rho \in (0,1)$ such that

$$\sup_{n,i} \alpha_{n,i}(h) \leq C\rho^h, \qquad h \geq 1.$$

For each t, the cross-sectional processes $\{x_{t,i} : i \in \mathbb{Z}\}$ are independent across i. Dependence between E and $x_{t,i}$'s is allowed, subject to the conditional moment restrictions stated in Lemma 10(c) in the Appendix.

(ii) M is a pre-specified adjacency matrix. The instrumental variable matrix $Z_t(M)$ has columns chosen from $M^\ell X_{jt}$ for $\ell = 0, \ldots, d$ and $j = 1, \ldots, K$. We require $\mathbb{E}[\varepsilon_t \mid Z_t(M)] = \mathbf{0}_{n\times 1}$. Suppose $Z_t(M)$ has m linearly independent columns, where m is a fixed constant satisfying $m \geq 2K+1$.

$$\sup_{n,t,i,\, 1\leq l\leq m} \mathbb{E}|z_{t,il}|^{2+\delta} < \infty, \qquad \delta > 2.$$

(iii) $\lim_{\min(n,T)\to\infty} (nT)^{-1} \sum_{t=1}^T \mathbb{E}\big[Z_t(M)^\top Z_t(M)\big]$ exists and is nonsingular, denoted as $\Sigma_{Z_0 Z_0}$.

(iv) Let $Q_{0t} \coloneqq \big[W_0(I_n - \lambda_0 W_0)^{-1}(X_t\beta_0 + W_0 X_t \gamma_0),\, X_t,\, W_0 X_t\big]$. We assume the $m \times (2K+1)$ matrix $\lim_{\min(n,T)\to\infty} (nT)^{-1} \sum_{t=1}^T \mathbb{E}\big[Z_t(M)^\top Q_{0t}\big]$ exists, denoted as G_0, and is of full column rank.

Assumption 7 (Moment condition and GMM weight matrix)**.** The GMM weight matrix $\widehat{\Lambda}_{nT}^{-1} \in \mathbb{R}^{m\times m}$ converges in probability to some symmetric positive-definite matrix $\Lambda^{-1} \in \mathbb{R}^{m\times m}$, whose eigenvalues are bounded away from 0 and ∞.

Assumption 5 and 6 include standard moment assumptions. The independence requirement across i in Assumption 5 and Assumption 6(i) might be strong in network settings, but it can be relaxed to weak cross-sectional dependence with all rates in Theorems 2 and 3 unchanged in order. We maintain independence for expositional simplicity. Assumptions 6 (ii) are regularity conditions on the instrumental variables. Among other things, the requirement that $\mathbb{E}[\varepsilon_t \mid Z_t(M)] = \mathbf{0}_{n\times 1}$ precludes $M = W$ if the measurement error E is endogenous. Assumptions 6 (iii) and (iv) impose nonsingularity on the variance covariance matrix of the instrumental variables, and rank conditions of gradient of the moment functions. The full column rank condition ensures that we only consider linearly independent columns of the instrument matrix, a common practice in the literature (Lee, 2003). Take $M = W_0$, this is linked to the rank conditions of W_0, the variance covariance structure of $Z_t(W_0)$ and the correlation between X_t and $Z_t(W_0)$. To further understand the implications of the above assumption, let $Z_t(W_0)$ be the $n \times m$ matrix. Note that $(nT)^{-1} \sum_{t=1}^T Z_t(W_0)^\top Z_t(W_0)$ is a partitioned matrix whose elements have the form $(nT)^{-1} \sum_{t=1}^T X_t^\top [W_0^{l_1}]^\top W_0^{l_2} X_t$, for any l_1 and l_2 between 0 and d. This condition can be implied by $\lambda_{\min}(\lim_{n,T\to\infty}\{nT\}^{-1} \sum_t \mathbb{E}\{Z_t^\top Z_t\}) > 0$, and $\lambda_{2K+1}(W_0 W_0^\top) > 0$. Then $\Sigma_{Z_0 Q_0}$ is of full column rank could be implied by a condition that $\lambda_{2K+1}(W_0 W_0^\top) > 0$, $\lim_{T\to\infty} T^{-1} \sum_t [X_t, W_0 X_t, \cdots, W_0^d X_t]$ has column rank greater than or equal to $2K+1$, and all eigenvalues of $I_n - \lambda_0 W_0$ are bounded away from 0 and ∞. Assumption 7 is a standard assumption imposed on the GMM weight matrix. Because

M is fixed and time invariant, $Z_t(M)$ is a measurable function of X_t. Therefore, $\{Z_t(M)\}$ inherits the strong-mixing property, with $\alpha_{Z(M),n}(h) \leq \alpha_{X,n}(h)$.

Before establishing the consistency of our spillover-parameter estimator, we first examine the bias introduced by E when using the classical GMM estimator. We provide an intuitive argument illustrating how the measurement error matrix E biases the moment functions and the importance of its removal, particularly when the instrument matrix $Z_t(M)$ also depends on W, as in standard spatial regression. For simplicity here, assume no contextual effects are present.When $Z_t(M)$ consists of exogenous instruments satisfying $\mathbb{E}[Z_t(M)^\top \varepsilon_t] = \mathbf{0}_{m\times 1}$, the moment function is given by

$$\bar{g}_{nT}(\theta, W, M) = \frac{1}{nT}\sum_{t=1}^{T} Z_t(M)^\top (Y_t - \lambda W Y_t - X_t\beta).$$

Substituting $W = W_0 + E$,

$$\bar{g}_{nT}(\theta, W, M) = \frac{1}{nT}\sum_{t=1}^{T} Z_t(M)^\top (Y_t - \lambda W_0 Y_t - X_t\beta) - \frac{\lambda}{nT}\sum_{t=1}^{T} Z_t(M)^\top E Y_t.$$

Thus, if E is independent of $(Z_t(M), Y_t)$ and has mean zero, then

$$\mathbb{E}\big[Z_t(M)^\top E Y_t\big] = \mathbf{0}_{m\times 1},$$

and hence

$$\mathbb{E}[\bar{g}_{nT}(\theta_0, W, M)] = \mathbf{0}_{m\times 1}.$$

However, when $Z_t(M)$ is a plug-in version of the instrument with $M = W$, i.e.,

$$Z_t(W) = [X_t, WX_t] = [X_t, (W_0 + E)X_t],$$

the moment function becomes

$$\bar{g}_{nT}(\theta, W) = \frac{1}{nT}\sum_{t=1}^{T} [X_t, (W_0 + E)X_t]^\top \big[(Y_t - \lambda W_0 Y_t - X_t\beta) - \lambda E Y_t\big].$$

Expanding the moment function $\bar{g}_{nT}(\theta, W)$ with the instrument matrix $[X_t, (W_0+E)X_t]$, we get:

$$\left[\frac{1}{nT}\sum_{t=1}^{T}(-\lambda X_t^\top E Y_t)^\top, \frac{1}{nT}\sum_{t=1}^{T}[X_t^\top E^\top (Y_t - \lambda W_0 Y_t - X_t\beta) - \lambda X_t^\top W_0^\top E Y_t - \lambda X_t^\top E^\top E Y_t]^\top\right]^\top.$$

Even if E_{ij}'s are i.i.d. mean zero, with bounded second moment and independent of (X_t, ε_t), there will be moment condition misspecification. Specifically,

$$\mathbb{E}[\bar{g}_{nT}(\theta_0, W)] \neq \mathbf{0}_{m\times 1},$$

since the expectation of the last term yields: $n^{-1}\mathbb{E}[X_t^\top E^\top E Y_t] = n^{-1}\mathbb{E}\big[X_t^\top \mathbb{E}[E^\top E]\,(I - \lambda_0 W_0)^{-1} X_t\beta_0\big] \neq \mathbf{0}_{K\times 1}$, which might be non-negligible under our assumptions. Recall that $\widehat{\theta}_{GMM}$ denotes the efficient two-step GMM estimator based on a given adjacency matrix W,

obtained in closed form in (5). From the preceding results, it follows that $\widehat{\theta}_{GMM}$ is likely to be inconsistent. We now analyze the convergence rate of $\widehat{\theta}_p(\widehat{W})$.

Since W_0 is not directly available for constructing instruments, one approach is to use $Z_t(\widehat{W})$ as plug-in instruments, where $\widehat{W}$ is obtained from Algorithms 1 and satisfies Theorem 1. When W_0 is correctly observed, the minimizer $\widehat{\theta}_p(W_0)$ satisfies the standard panel-data estimator convergence rate:

$$\|\widehat{\theta}_p(W_0) - \theta_0\|_2 = O_p(1/\sqrt{nT}).$$

However, when W_0 is replaced by a denoised estimate $\widehat{W}$, the associated minimizer $\widehat{\theta}_p(\widehat{W})$ depends on the deviation $\widehat{W} - W_0$. The corresponding convergence properties are detailed in the following results.

Theorem 2 (Performance of the plug-in estimator)**.** *Recall the definitions of B_0, ρ_n, s_n, κ_n, and $\mathcal{R}^*_{nT}$ given above. Consider the model described in* (3)*. Suppose Assumptions 1–7 hold.*
(a) The plug-in estimator. *If $\mathcal{R}^*_{nT} = o(1)$, then*

$$\|\widehat{\theta}_p(\widehat{W}) - \theta_0\|_2 = O_p(\mathcal{R}^*_{nT}) + O_p\left(\frac{1}{\sqrt{nT}}\right).$$

(b)The GMM estimator. *Using the noisy matrix $W = W_0 + E$ directly,*

$$\|\widehat{\theta}_{GMM} - \theta_0\|_2 = O_p(w_{2n} + w_{2n}^2) + O_p\left(\frac{1}{\sqrt{nT}}\right). \tag{11}$$

(c) Asymptotic normality. *If*

$$\sqrt{nT}\,\mathcal{R}^*_{nT} = o(1),$$

then

$$\sqrt{nT}\left(\widehat{\theta}_p(\widehat{W}) - \theta_0\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_{GMM}).$$

where

$$\Sigma_{GMM} = \left(G_0^\top \Lambda^{-1} G_0\right)^{-1} G_0^\top \Lambda^{-1} \Omega_0 \Lambda^{-1} G_0 \left(G_0^\top \Lambda^{-1} G_0\right)^{-1},$$

and

$$\Omega_0 = \frac{1}{n}\mathbb{E}(Z_t(W_0)^\top \varepsilon_t \varepsilon_t^\top Z_t(W_0)).$$

Under the assumptions of Theorem 2(a), the plug-in estimator $\widehat{\theta}_p(\widehat{W})$ is consistent provided that

$$\mathcal{R}^*_{nT} = o(1).$$

The factor $1/n$ discounts the noise contribution by the network size, so consistency holds even when w_{2n} and $w_{\max n}$ do not vanish individually, for example, when r and $m_s(S_0)$ remain bounded and w_{2n} is of constant order. By contrast, the available bound for the GMM estimator based on the noisy matrix does not guarantee consistency when w_{2n} is bounded away from zero; the moment-misspecification example above shows that inconsistency can

occur in this case. The noise-induced component of the plug-in estimator's rate is strictly smaller whenever

$$\mathcal{R}^*_{nT} = o(w_{2n} + w^2_{2n}).$$

The improvement reflects how the $L + S$ decomposition enters the analysis. Denote $\Delta L := \widehat{L} - L_0$ and $\Delta S := \widehat{S} - S_0$ as the recovery errors from Theorem 1. The recovery error $\widehat{W} - W_0 = \Delta L + \Delta S$ affects the spillover estimator only through bilinear forms (averages over the network) with the instruments and residual design, and these averages discount the error according to its structure. The low-rank error enters through its nuclear norm, controlled by $\sqrt{r}\,\|\Delta L\|_F$, contributing $O(r)$ effective directions rather than all n; the sparse error enters through $|\Delta S|_{1,1}$, controlled by its $m_s(S_0)$ nonzero entries rather than all n^2. The measurement error can therefore be large in aggregate while the spillover estimator still converges.

Example 2 constructs regimes in which $|E|_{1,1}/n$ does not vanish, while the dimension-discounted condition $\mathcal{R}^*_{nT} = o(1)$ may still hold. This condition is sharp for sparse measurement error such as link misclassification in 0–1 networks, but fails in the dense case that motivates our analysis: Example 2 constructs the case in which $|E|_{1,1}$ is of order n^2 yet the plug-in estimator remains consistent. The two analyses cover disjoint measurement-error regimes: sparse errors with Lewbel et al. (2024), dense errors with structured latent adjacency in our framework. When $L_0 = \mathbf{0}_{n\times n}$, the rate specializes to $O_p(m_s(S_0)w_{\max n}/n)$ (Corollary 1).

The following corollary specializes Theorem 2 to the pure-sparse case $L_0 = \mathbf{0}_{n\times n}$. Within Assumption 1, only part (ii) on S_0 is needed, and the coherence terms drop out. The rate is driven entirely by the sparse recovery error $n^{-1}|\widehat{S} - S_0|_{1,1}$, the same ℓ_1-type quantity as the $|E|_{1,1}/n$ condition of Lewbel et al. (2024), but applied to the recovery error after denoising rather than to the raw measurement error.

Corollary 1. *Suppose $L_0 = \mathbf{0}_{n\times n}$ and let $\widehat{S}$ denote the pure-sparse estimator obtained by fixing $\widehat{L} = \mathbf{0}_{n\times n}$. Suppose Assumption 1(ii), Assumption 2(i),(iii), and Assumptions 3–7 hold. Assume additionally that*

$$\|S_0\|_1 = O(1), \qquad \|E\|_\infty = O_p(1),$$

and

$$\max_{1\le j\le K} \frac{1}{T}\sum_{t=1}^{T} \|X_{jt}\|^2_{\max} = O_p(1), \qquad \frac{1}{T}\sum_{t=1}^{T} \|\varepsilon_t\|^2_{\max} = O_p(1).$$

Then

$$\begin{aligned}\|\widehat{\theta}_p(\widehat{S}) - \theta_0\|_2 &= O_p\left(\frac{|\widehat{S} - S_0|_{1,1}}{n}\right) + O_p\left(\frac{1}{\sqrt{nT}}\right) \\ &= O_p\left(\frac{m_s(S_0)w_{\max n}}{n}\right) + O_p\left(\frac{1}{\sqrt{nT}}\right).\end{aligned}$$

The rate in Corollary 1 follows from Lemma 10(b): the sparse recovery error enters the spillover estimator only through $n^{-1}|\widehat{S} - S_0|_{1,1}$. In the pure-sparse case, the sparse-recovery argument used in the proof of Theorem 1 gives

$$\frac{1}{n}|\widehat{S} - S_0|_{1,1} = O_p\left(\frac{m_s(S_0)w_{\max n}}{n}\right).$$

Although Corollary 1 covers the pure-sparse case, the gain from our general analysis is most pronounced when the low-rank component L_0 is genuinely present and r and $m_s(S_0)$ remain controlled relative to n.

Remark 1 (Asymptotic collinearity). It should be noted that variation in peer connections plays an important role in identification and estimation as noted, for example, in Manski (1993) (p.535) and de Paula (2017). In a *single*, *dense* and *large* network, the leading eigenvector of a row-normalized and positive adjacency matrix may become (asymptotically) proportional to a vector of ones. In this case, peer averages become increasingly similar across individuals, complicating identification and estimation of spillover effects. A recent formulation of this issue in the context of peer effects is provided by Hayes and Levin (2024), who show that this collinearity arises when W_0 is row-normalized, positive, and exhibits unbounded minimum degrees. Our setting differs in two respects. For a nonnegative row-normalized matrix, $W_0\mathbf{1}_n = \mathbf{1}_n$ holds exactly. When nodal covariates are independent of the network and the minimum degree diverges, peer averages may concentrate around a common value, producing asymptotic collinearity; see Hayes and Levin (2024). Our assumptions do not generally impose row normalization or diverging minimum degree, so this conclusion does not follow automatically in our setting. The presence of a sparse component S_0 does not by itself eliminate this possibility. Without row normalization, $\mathbf{1}_n$ need not be an eigenvector of W_0, and the argument requires separate verification. ■

As mentioned previously, we emphasize that our results are also suitable for the case when E is endogenous with respect to ε_t. As an example, in social networks this can arise when misclassification errors are not generated at random, instead being correlated with unobservables driving the network formation process, such as the framework considered by Candelaria and Ura (2023). To see this, we provide a rate analysis to compare the rate when E is endogenous or not of both our estimator and the GMM estimator. For simplicity, we adopt the following assumption on the relationship between these two sources of error, see similar assumptions as in (Lee and Yu, 2014):

Assumption 8 (Endogenous measurement error)**.** Suppose Model (3) holds and W_0 is symmetric and positive semidefinite. For each i, let

$$\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})^\top \in \mathbb{R}^T,$$

and let

$$E_{i,-i} = (E_{ij} : j \neq i)^\top \in \mathbb{R}^{n-1}$$

collect the off-diagonal elements of the ith row of E, with $E_{ii} = 0$.

Suppose that the vectors $(\varepsilon_i^\top, E_{i,-i}^\top)^\top$ are independent across i and jointly Gaussian:

$$\begin{pmatrix} \varepsilon_i \\ E_{i,-i} \end{pmatrix} \sim \mathcal{N}(\mathbf{0},\, \Sigma_{\varepsilon E}),$$

where

$$\Sigma_{\varepsilon E} = \begin{bmatrix} \sigma_0^2 I_T & \sigma_0 \sigma_E n^{-s} \rho_{\varepsilon E} \\ \sigma_0 \sigma_E n^{-s} \rho_{\varepsilon E}^\top & \sigma_E^2 n^{-2s} \Sigma_E \end{bmatrix}, \qquad s > \frac{1}{2}.$$

Here, $\rho_{\varepsilon E} \in \mathbb{R}^{T \times (n-1)}$ governs the dependence between the outcome disturbances and the measurement error, and $\Sigma_E \in \mathbb{R}^{(n-1)\times(n-1)}$ is a correlation matrix. Assume that there exist constants $c_E, C_E, C_\rho > 0$, independent of n and T, such that

$$c_E \le \lambda_{\min}(\Sigma_E) \le \lambda_{\max}(\Sigma_E) \le C_E, \qquad \|\rho_{\varepsilon E}\|_2 \le C_\rho,$$

and that $\Sigma_{\varepsilon E}$ is positive semidefinite.

Assumption 8 is imposed together with Assumptions 2 and 5. Thus, the disturbances remain marginally i.i.d. across i and t, while dependence between E and ε_t is allowed. The Gaussian specification is imposed solely to illustrate the rate in the presence of endogenous measurement error.

Denote $\widehat{\theta}_{GMM}(W, M)$ and $\widehat{\theta}_p(\widehat{W}, M)$ the GMM and plug-in estimator with instrument $Z_t(M)$. In comparison to the exogenous measurement error case, the bias of $\widehat{\theta}_{GMM}(W, M) = \widehat{\theta}_p(W, M)$ could be larger than our proposed estimators. It is because the deviation between sample moments $\bar{g}_{nT}(\theta, W, M)$ and $\bar{g}_{nT}(\theta, W_0, M)$ might be more severe. This is reflected in our simulation exercises, see Section 5. For simplicity, we shall illustrate without the contextual-effects regressor $W_0 X_t$ (i.e., setting $\gamma_0 = \mathbf{0}_K$), so the regressors are $[W_0 Y_t, X_t]$. Define $G(M) = -n^{-1}\mathbb{E}[Z_t^\top(M)[W_0 Y_t, X_t]]$. To explain this, take M as exogenous; the leading term for the moment estimator $\widehat{\theta}_p(\widehat{W}, M) - \theta_0$ is $-(G^\top(M)\Lambda^{-1}G(M))^{-1}G^\top(M)\Lambda^{-1}\bar{g}_{nT}(\theta_0, W_0, M)$, according to Theorem 2. We shall suppress the dependence of M on W_0 and E. The following proposition characterizes a rate bound for both the GMM estimator and our estimator.

Recall that $\bar{g}_{nT}(\theta, W, M) = \frac{1}{nT}\sum_{t=1}^T Z_t(M)^\top (Y_t - \lambda W Y_t - X_t\beta - W X_t \gamma)$.

Proposition 1 (Endogenous error bias)**.** *Suppose that M is fixed or exogenous and that the conditional-moment conditions in Lemma 10(c) hold. Under Assumptions 1–8, with $\|B_0\|_2(1 + \|W_0\|_2) = O(1)$ and $\mathcal{R}_{nT}^* = o(1)$, we have*

$$\|\widehat{\theta}_p(\widehat{W}, M) - \theta_0\|_2 \lesssim_p \mathcal{R}_{nT}^* + \frac{1}{\sqrt{nT}}, \tag{12}$$

$$\|\widehat{\theta}_{GMM}(W, M) - \theta_0\|_2 \lesssim_p n^{1/2-s}\|\rho_{\varepsilon E}\|_2 + w_{2n} + \frac{1}{\sqrt{nT}}. \tag{13}$$

Three features of this comparison are worth highlighting.

First, the endogeneity bias $n^{1/2-s}\|\rho_{\varepsilon E}\|_2$ appears only in the GMM rate (13). It originates from a bilinear form that is quadratic in E: the correlation between E and ε_t contributes a non-vanishing mean to $Z_t^\top E B_0 \varepsilon_t$, whose magnitude is governed by $\rho_{\varepsilon E}$. In the plug-in case,

the analogous bilinear form involves $(\Delta L + \Delta S)B_0\varepsilon_t$. Its low-rank component is controlled by

$$\|\Delta L\|_* = O_p((r \vee 1)(w_{2n} + \mathcal{R}_s)),$$

whereas its off-diagonal sparse component is controlled by

$$|\Delta S|_1 = O_p\Big(\mathcal{R}_s\sqrt{m_s(S_0)}\Big).$$

Under the conditional-moment conditions, the plug-in bound therefore contains no additional explicit term involving $\rho_{\varepsilon E}$. *Second*, even in the absence of endogeneity ($\rho_{\varepsilon E} = 0$), the denoised rate improves upon the GMM rate: the noise-level term w_{2n} in (13) is replaced by the rate $\mathcal{R}^*_{nT}$ in (12), which is strictly smaller whenever $\mathcal{R}^*_{nT} = o(w_{2n})$. *Third*, the denoised rate *attenuates* the dependence on the noise magnitude w_{2n} rather than eliminating it: w_{2n} enters through the $L+S$ coupling $\|\Delta L\|_2 = O_p(w_{2n} + \mathcal{R}_s)$, but its effect is multiplied by the factor r/n rather than by 1. As a result, the rate scales not with the raw noise level w_{2n} but with the ratio of the structural complexity $(r,\, m_s(S_0))$ to the sample size n. This reflects the benefit of our approach. Finally, we consider the rate of the supervised estimator as defined in Equation (8). Denote $\widehat{W}_s = \widehat{L}_s + \widehat{S}_s$.

Theorem 3 (Performance of the supervised estimator)**.** *Suppose Assumptions 1–7 hold. Let M be fixed, exogenous, or $\sigma(E)$-measurable and satisfy the stated instrument and design conditions. Let $\widehat{\theta}_s$ denote the supervised estimator defined in* (8)*.*

$$\nu_{nT} = C_\nu w_{2n}, \qquad \tau_{nT} = C_\tau(w_{\max n} + a_n),$$

for sufficiently large constants $C_\nu, C_\tau > 0$.

Suppose

$$\mathcal{R}^*_{nT} = o(1) \qquad \textit{and} \qquad \xi_{nT} = O(1).$$

Then

$$\|\widehat{\theta}_s - \theta_0\|_2 = O_p\left(\frac{1}{\sqrt{nT}} + \mathcal{R}^*_{nT}\right). \tag{14}$$

If, in addition,

$$\sqrt{nT}\,\mathcal{R}^*_{nT} = o(1) \qquad \textit{and} \qquad \xi_{nT} = o(1),$$

then

$$\sqrt{nT}(\widehat{\theta}_s - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_{GMM}).$$

The supervised estimator achieves the same rate as the plug-in estimator in Theorem 2 and it is strictly smaller than that of the noisy-W GMM estimator whenever $\mathcal{R}^*_{nT} = o(w_{2n} + w_{2n}^2)$. The rate still depends on w_{2n}, but only through the dimension-discounted recovery rate $\mathcal{R}^*_{nT}$. It is worth noting that the rate in (14) depends on the parameters $(r, m_s(S_0), n, T)$ and on the spectral characteristics of the underlying network $(\|B_0\|_2, \|W_0\|_2)$, rather than directly on the noise level w_{2n} compared to the rate of the GMM estimator. The condition on ξ_{nT} ensures that the generated error from the first-iteration estimator $\widehat{\theta}^{[1]}$ does not inflate either the sparse recovery in Step b) or the low-rank recovery in Step c) beyond the unsupervised noise level, so that the supervised decomposition matches the plug-in decomposition rates.

The following corollary specializes Theorem 3 to the pure-sparse case $L_0 = \mathbf{0}_{n\times n}$, paralleling Corollary 1: only Assumption 1(ii) on S_0 is needed, the coherence terms drop out, and the rate is driven entirely by the sparse recovery error $n^{-1}|\widehat{S}_s - S_0|_{1,1}$, making the connection to the misclassification framework of Lewbel et al. (2024) explicit.

Corollary 2. *Suppose $L_0 = \mathbf{0}_{n\times n}$ and consider the pure-sparse supervised estimator obtained by imposing $L = \mathbf{0}_{n\times n}$, so that $\widehat{W}_s = \widehat{S}_s$. Suppose Assumption 1(ii), Assumption 2(i),(iii), and Assumptions 3–7 hold. Assume, in addition, that*

$$\|W_0\|_1 = O(1), \qquad \|E\|_\infty = O_p(1), \qquad \xi_{nT} = O(1),$$

and

$$\max_{1\le j\le K} \frac{1}{T}\sum_{t=1}^{T} \|X_{jt}\|_{\max}^2 = O_p(1), \qquad \frac{1}{T}\sum_{t=1}^{T} \|\varepsilon_t\|_{\max}^2 = O_p(1).$$

Then

$$\|\widehat{\theta}_s - \theta_0\|_2 = O_p\left(\frac{1}{n}|\widehat{S}_s - S_0|_{1,1}\right) + O_p\left(\frac{1}{\sqrt{nT}}\right). \tag{15}$$

Moreover, Lemma 10(b), together with the sparse-recovery bound for $\widehat{S}_s$, gives

$$\|\widehat{\theta}_s - \theta_0\|_2 = O_p\left(\frac{m_s(S_0) w_{\max n}}{n}\right) + O_p\left(\frac{1}{\sqrt{nT}}\right). \tag{16}$$

The theoretical performance of the supervised estimator is similar to that of the plug-in estimator, as confirmed by the simulation and application results in Sections 5 and 6. We shall also note that both Theorems 2 and 3 do not require E to be exogenous, as long as $Z_t(M)$ does not directly involve W. Furthermore, both theorems yield rates in which the noise level w_{2n} enters only after dimension-discounting by the factor r/n (rather than at the raw spectral rate w_{2n} delivered by the noisy-W baseline), reflecting the fundamental advantage of the $L + S$ decomposition combined with the within transformation.

5. Simulations

In this section, we present simulation exercises to evaluate the performance of our proposed estimators. We first introduce four data-generating processes (DGPs) for the network adjacency matrix W_0. We then describe how it is contaminated with noise, and discuss the Monte Carlo simulation results.

5.1. Generating $\mathbf{W}_0$.

The diagonal elements of W_0 are assumed to be 0. Hence we set all diagonal elements of E to be 0 and do not penalize S_0 on its diagonal. The other entries of E are generated as explained in Section 5.2.

i) **DGP 1**(Low rank) In this setup, L_0 is generated by a pure low-rank matrix with $r = 2$,

$$L_0 = UDV^\top, \quad U, V \in \mathbb{R}^{n\times r}, \text{ and } \quad U^\top U = I_r, \quad V^\top V = I_r, \tag{17}$$

where D is the $r \times r$ diagonal matrix $D = \text{diag}\{0.8^0, 0.8^1, \ldots, 0.8^{r-1}\}$, and U and V —the left and right singular vectors of L_0, respectively— are the first r columns of two random $n \times n$ orthonormal matrices generated according to standard algorithms (Stewart, 1980; Mezzadri, 2007). We then set $W_0 = L_0^*$, where L_0^* has zero diagonal elements and its off-diagonal elements are the same as L_0.

ii) **DGP 2** (Low rank+ sparse) Our second DGP still uses the same L_0 as in DGP 1, but adds a sparse $n \times n$ component S_0, where we randomly pick $m_r = 2$ elements in each row and draw their values from a Uniform$(0.5, 1)$ distribution. Note that W_0 could be written as

$$W_0 = L_0^* + S_0^*, \tag{18}$$

where L_0^* is defined as in DGP 1 using U, V and D generated as in (17), and S_0^* has zero diagonal elements and its off-diagonal elements are the same as S_0.

iii) **DGP 3** (Dominant units) Our third DGP generates W_0 as a network with dominant units as presented in Subsection 2.3.4. Suppose the number of dominant units is $J = 2$. The first J units are dominant players in the network, whereas the remaining $n - J$ units are ordinary players. We could partition W_0 as

$$W_0 = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}. \tag{19}$$

Let $W_D = [W_{11}^\top, W_{21}^\top]^\top$ be the $n \times J$ dominant block of the adjacency matrix corresponding to the dominant units, and $W_R = [W_{12}^\top, W_{22}^\top]^\top$, the right block of the adjacency matrix corresponding to ordinary units.

For each column of W_D, the first $\lfloor n^{0.9} \rfloor$ entries, except the diagonal elements, are drawn independently from a Uniform$(0, 1)$ distribution, and the remaining entries are set equal to 0. For W_R, we induce sparsity by imposing that each individual is connected with equal weights to the immediate predecessor and successor. Note that W_0 could be viewed as a purely sparse matrix S_0^* as each of its row has at most $2 + J$ nonzero elements, and we treat L_0^* as a zero matrix.

iv) **DGP 4** (Group) Our fourth DGP divides the individuals into two groups. The first $n/4$ members form group 1, the remaining $3n/4$ form group 2, and only members within the same group have connections. For each group, the connection strength between any two different individuals i and j within the same group is defined by $d_g u_{g,i} u_{g,j} / \sum_i u_{g,i}^2$, where $d_{1,1} = 1$ and $d_{2,2} = 0.9$, $u_{g,i}$'s are latent individual-specific characteristic generated independently from standard normal random variables. Note that DGP 4 admits the low-rank form such that $L_0 = UDU^\top$, where $U = [U_1, U_2]$, D is a 2×2 diagonal matrix with $d_{1,1} = 1$ and $d_{2,2} = 0.9$, $U_1 = [u_{1,1}, \cdots, u_{1,n/4}, 0, \cdots, 0] / \sqrt{\sum_i u_{1,i}^2}$ and $U_2 = [0, \cdots, 0, u_{2,1}, \cdots, u_{2,3n/4}] / \sqrt{\sum_i u_{2,i}^2}$, and $W_0 = L_0^*$, where L_0^* has zero diagonal entries and its off diagonal elements are the same as L_0.

5.2. **Monte Carlo setup.** For our simulation exercises, we set $\lambda_0 = 0.25$, $\beta_0 = (-1, 2)$. We generate two covariates for X_t independently from a $\mathbb{N}(0, 1)$ and a $\mathbb{N}(5, 2)$, respectively. We set sample sizes $n \in \{40, 80, 120\}$ and time dimensions $T \in \{1, 5, 15, 50\}$, considering specific combinations of panel sizes due to computational constraints. We take as given a true network W_0 generated using one of the DGPs introduced in the previous subsection, and keep it constant across 100 Monte Carlo simulations.

Our exercises additionally consider the possibility that measurement errors are endogenous. That is, we allow for correlated disturbances between the outcome equation (ε_t) and the errors that contaminate the adjacency matrix (E), as outlined in our Assumption 8. Let σ_E and σ_ε be two positive constants. Specifically, for $i, j = 1, \cdots, n$, we generate $E_{ij} \sim$ *i.i.d.* $\mathbb{N}(0, \sigma_E^2)$ for $i \neq j$, and set $E_{ii} = 0$. We generate ε_t such that its ith component ε_{it} satisfies $\varepsilon_{it} = \rho\sigma_\varepsilon \sum_{j=1}^n E_{ij}/\sqrt{n} + v_{it}$, and v_{it} are i.i.d. $\mathbb{N}(0, (1-\rho^2)\sigma_\varepsilon^2)$ that are independent of E_{ij}. Note that $\varepsilon_t|E$ is normally distributed and we could simply obtain its conditional mean and variance. We define $\text{Vec}(E)$ as the vector obtained by stacking the off-diagonal elements of E.

$$\begin{pmatrix} \varepsilon_t \\ \text{Vec}(E) \end{pmatrix} \sim \mathbb{N}\left(\mathbf{0}_{n^2}, \begin{bmatrix} \sigma_\varepsilon^2 \cdot I_n & \Sigma_{\varepsilon E} \\ \Sigma_{\varepsilon E}^\top & \sigma_E^2 \cdot I_{n^2-n} \end{bmatrix}\right), \quad t = 1, \ldots, T.$$

where $\Sigma_{\varepsilon E}$ is the $n \times (n^2 - n)$ matrix that captures the correlation between ε_{it} and $\text{Vec}(E)$ stacks all E_{ij} for $i \neq j$ into a long vector. Note that it is implied from the previous sections that $\text{cov}(\varepsilon_{it}, E_{ij}) = \rho\sigma_E^2\sigma_\varepsilon/\sqrt{n}$ for $i \neq j$ and $\text{cov}(\varepsilon_{it}, E_{kj}) = 0$ if $k \neq i$. In the simulation, we set $\sigma_\varepsilon = 0.15$, $\sigma_E = 0.3/n^{0.7}$ and $\rho = 0$ or 0.7. By setting $\rho = 0$, we recover the exogenous measurement error case from Assumption 2.

Given values for the adjacency W_0, covariates X_t, parameters θ_0, and disturbances ε_t, we use the reduced-form representation of model (3) to generate our outcome Y_t at every time period:

$$Y_t = (I_n - \lambda_0 W_0)^{-1}(X_t\beta_0 + \varepsilon_t),\ t = 1, \ldots, T\,. \tag{20}$$

Throughout the simulation we set $W = W_0 + E$, which is the observed noisy version of the adjacency matrix, to construct estimates of parameters θ_0. Once the adjacency matrix is generated, we use it to construct a robust scale proxy of the errors and set the tuning parameters for estimation as below. We fix $\xi_{nT} = 1$ and denote $\text{Int}(W)$ as the interquartile range of W_{ij}. Then we choose the sparse penalty as $\tau_n = \tau_{nT} = 2\text{Int}(W)\log(n)$ and $\nu_n = \nu_{nT} = \text{Int}(W)\sqrt{n}$.

5.3. **Simulation results.** Tables 1 - 3 present the simulation results using the setting outlined in the previous subsections. We highlight several key takeaways in this setup. First, Tables 1 reports relative root mean squared errors (RMSE) for estimates of the spillover scalar parameter λ_0, where the benchmark is the standard GMM estimator. Observe that both our plug-in and supervised estimators achieve lower RMSEs than the GMM benchmark uniformly across all DGPs and sample sizes. Moreover, the GMM estimates tend to perform worse in the endogenous case and hence there is more improvement in the endogenous case.

TABLE 1. Relative RMSE for estimators of λ_0

Exogenous	Plug-in	Supervised	Plug-in	Supervised	Plug-in	Supervised	Plug-in	Supervised
L (Low-rank)	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.499	0.409	0.262	0.258	0.251	0.245	0.239	0.235
n=80	0.279	0.279	0.203	0.203	0.200	0.196	0.197	0.188
n=120	0.361	0.353	0.146	0.140	0.131	0.124	0.120	0.135
L+Sparse	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.444	0.438	0.285	0.280	0.231	0.221	0.211	0.207
n=80	0.360	0.359	0.237	0.235	0.201	0.198	0.181	0.181
n=120	0.297	0.296	0.146	0.145	0.101	0.102	0.071	0.071
Dominant	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.449	0.450	0.328	0.331	0.248	0.252	0.260	0.282
n=80	0.341	0.341	0.218	0.218	0.158	0.160	0.125	0.131
n=120	0.285	0.285	0.185	0.186	0.164	0.166	0.128	0.133
Group	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.364	0.356	0.243	0.228	0.214	0.200	0.207	0.204
n=80	0.239	0.230	0.154	0.155	0.140	0.137	0.135	0.145
n=120	0.206	0.205	0.148	0.147	0.130	0.126	0.132	0.122
Endogenous	Plug-in	Supervised	Plug-in	Supervised	Plug-in	Supervised	Plug-in	Supervised
L (Low-rank)	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.376	0.317	0.261	0.260	0.256	0.261	0.249	0.260
n=80	0.223	0.224	0.207	0.206	0.207	0.201	0.204	0.196
n=120	0.276	0.269	0.135	0.124	0.128	0.120	0.123	0.145
L+Sparse	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.371	0.364	0.274	0.266	0.242	0.235	0.235	0.234
n=80	0.327	0.325	0.218	0.214	0.190	0.190	0.181	0.183
n=120	0.238	0.238	0.136	0.135	0.105	0.105	0.092	0.093
Dominant	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.383	0.384	0.314	0.318	0.290	0.299	0.280	0.305
n=80	0.274	0.274	0.203	0.204	0.167	0.169	0.162	0.171
n=120	0.211	0.211	0.144	0.144	0.119	0.120	0.114	0.119
Group	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.311	0.305	0.237	0.218	0.219	0.202	0.217	0.224
n=80	0.183	0.183	0.152	0.153	0.144	0.141	0.142	0.152
n=120	0.174	0.172	0.140	0.141	0.135	0.132	0.136	0.135

Denote $\widehat{\lambda}_W$ and $\widehat{\lambda}_p$ as the baseline estimate and the plugin estimate calculated using equation (7) with weight matrices $W = W_0 + E$ and $\widehat{W}$ respectively. The supervised estimate is obtained by equation (8) with $M = \widehat{W}$ to construct the instruments. All entries represent RMSE values normalized by the RMSE of the baseline $\widehat{\lambda}_W$. The endogenous measurement error is simulated with a correlation structure parameter $\rho = 0.7$ as described in (20). The listed four DGPs correspond to those introduced in Subsection 5.1.

TABLE 2. Relative recovery accuracy of $\widehat{W}$

Exogenous	$\widehat{W}$	Supervised $\widehat{W}$	$\widehat{W}$	Supervised $\widehat{W}$	$\widehat{W}$	Supervised	$\widehat{W}$	Supervised $\widehat{W}$
L (Low-rank)		$T=1$		$T=5$		$T=15$		$T=50$
n=40	0.320	0.320	0.320	0.321	0.320	0.323	0.320	0.328
n=80	0.226	0.226	0.226	0.227	0.226	0.227	0.226	0.228
n=120	0.185	0.185	0.185	0.186	0.185	0.186	0.185	0.188
L+Sparse		$T=1$		$T=5$		$T=15$		$T=50$
n=40	0.392	0.391	0.392	0.391	0.392	0.391	0.392	0.392
n=80	0.275	0.274	0.275	0.274	0.275	0.274	0.275	0.274
n=120	0.225	0.225	0.225	0.225	0.225	0.225	0.225	0.225
Dominant		$T=1$		$T=5$		$T=15$		$T=50$
n=40	0.556	0.556	0.556	0.556	0.556	0.556	0.556	0.556
n=80	0.489	0.489	0.489	0.489	0.489	0.489	0.489	0.489
n=120	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333
Group		$T=1$		$T=5$		$T=15$		$T=50$
n=40	0.340	0.362	0.340	0.363	0.340	0.364	0.340	0.365
n=80	0.224	0.224	0.224	0.225	0.224	0.226	0.224	0.227
n=120	0.185	0.187	0.185	0.187	0.185	0.187	0.185	0.188
Endogenous	$\widehat{W}$	Supervised $\widehat{W}$	$\widehat{W}$	Supervised $\widehat{W}$	$\widehat{W}$	Supervised	$\widehat{W}$	Supervised $\widehat{W}$
L (Low-rank)		$T=1$		$T=5$		$T=15$		$T=50$
n=40	0.320	0.320	0.320	0.323	0.320	0.328	0.320	0.346
n=80	0.226	0.226	0.226	0.227	0.226	0.228	0.226	0.232
n=120	0.185	0.186	0.185	0.186	0.185	0.188	0.185	0.194
L+Sparse		$T=1$		$T=5$		$T=15$		$T=50$
n=40	0.392	0.391	0.392	0.391	0.392	0.391	0.392	0.392
n=80	0.275	0.274	0.275	0.274	0.275	0.274	0.275	0.274
n=120	0.225	0.225	0.225	0.225	0.225	0.225	0.225	0.225
Dominant		$T=1$		$T=5$		$T=15$		$T=50$
n=40	0.556	0.556	0.556	0.556	0.556	0.556	0.556	0.556
n=80	0.489	0.489	0.489	0.489	0.489	0.489	0.489	0.489
n=120	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333
Group		$T=1$		$T=5$		$T=15$		$T=50$
n=40	0.340	0.362	0.340	0.365	0.340	0.367	0.340	0.374
n=80	0.224	0.225	0.224	0.226	0.224	0.228	0.224	0.233
n=120	0.185	0.187	0.185	0.187	0.185	0.188	0.185	0.190

The relative recovery accuracy is defined as $\|\widehat{W} - W_0\|_F / \|W - W_0\|_F$, with baseline weight matrix $W = W_0 + E$. Endogenous measurement error is simulated with a correlation parameter $\rho = 0.7$, as specified in (20). The listed four DGPs correspond to those introduced in Subsection 5.1.

TABLE 3. Relative recovery accuracy of $(I_n - \widehat{\lambda}_{\widehat{W}}\widehat{W})^{-1}$

Exogenous	Plug-in	Supervised	Plug-in	Supervised	Plug-in	Supervised	Plug-in	Supervised
L (Low-rank)	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.541	0.478	0.420	0.418	0.415	0.411	0.414	0.407
n=80	0.315	0.315	0.297	0.296	0.296	0.294	0.296	0.292
n=120	0.447	0.440	0.288	0.281	0.277	0.269	0.271	0.277
L+Sparse	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.417	0.416	0.398	0.397	0.394	0.392	0.392	0.390
n=80	0.295	0.294	0.278	0.278	0.274	0.274	0.273	0.272
n=120	0.243	0.243	0.227	0.226	0.223	0.223	0.222	0.222
Dominant	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.545	0.545	0.543	0.543	0.542	0.542	0.542	0.542
n=80	0.507	0.507	0.504	0.505	0.505	0.505	0.504	0.505
n=120	0.332	0.332	0.324	0.324	0.324	0.324	0.324	0.323
Group	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.539	0.556	0.475	0.486	0.468	0.480	0.469	0.480
n=80	0.387	0.383	0.332	0.332	0.328	0.325	0.328	0.326
n=120	0.329	0.331	0.300	0.304	0.295	0.297	0.296	0.295
Endogenous	Plug-in	Supervised	Plug-in	Supervised	Plug-in	Supervised	Plug-in	Supervised
L (Low-rank)	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.474	0.421	0.409	0.408	0.406	0.407	0.405	0.404
n=80	0.324	0.324	0.325	0.324	0.328	0.323	0.328	0.320
n=120	0.340	0.334	0.224	0.215	0.217	0.210	0.213	0.230
L+Sparse	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.409	0.407	0.393	0.391	0.388	0.386	0.387	0.385
n=80	0.297	0.296	0.276	0.275	0.273	0.272	0.272	0.271
n=120	0.241	0.241	0.223	0.223	0.219	0.219	0.218	0.218
Dominant	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.533	0.532	0.535	0.535	0.535	0.535	0.536	0.537
n=80	0.492	0.492	0.494	0.494	0.496	0.496	0.496	0.496
n=120	0.320	0.320	0.321	0.321	0.321	0.321	0.321	0.321
Group	$T=1$		$T=5$		$T=15$		$T=50$	
n=40	0.492	0.507	0.442	0.442	0.440	0.441	0.444	0.450
n=80	0.327	0.331	0.301	0.301	0.298	0.293	0.297	0.295
n=120	0.315	0.317	0.297	0.301	0.295	0.295	0.296	0.294

The recovery accuracy of the Leontief inverse is defined as $\|[I_n - \widehat{\lambda}_p\widehat{W}]^{-1} - (I_n - \lambda_0 W_0)^{-1}\|_F$ and $\|[I_n - \widehat{\lambda}_s(\widehat{L}_s + \widehat{S}_s)]^{-1} - (I_n - \lambda_0 W_0)^{-1}\|_F$ for the plug-in and supervised estimates solved from equations (7) and (8) respectively. Both are then normalized by the baseline $\|(I_n - \widehat{\lambda}_W W)^{-1} - (I_n - \lambda_0 W_0)^{-1}\|_F$ with $W = W_0 + E$ to obatin the relative recovery accuracy. Endogenous measurement error is simulated with a correlation parameter $\rho = 0.7$, as specified in (20). The DGPs correspond to those introduced in Subsection 5.1.

Our proposed methods perform better than GMM by approximately 50–80% for small n and T. Our methods achieve at least a 75% improvement in RMSE for $n = 120$, $T = 50$. The relative performance of our methods generally improves as the number of nodes in the network n and the number of time periods T increase. This behavior is as predicted by the theory in the cases where the adjacency matrix is assumed to be constant across time.

In terms of the recovery performance of the denoised adjacency matrix, as evidenced in Table 2, the plug-in method and the supervised method perform similarly. Moreover, there is little difference between the exogenous and endogenous cases because we use the same observed adjacency matrix.

Finally, we explore how our methods perform in recovering not just the scalar spillover effect λ_0, but the full diffusion multiplier matrix given by $(I_n - \lambda_0 W_0)^{-1}$. This quantity is highly important as a policy tool, as it summarizes the cumulative spillover effects of a given economic shock. As shown in Table 3, by providing reliable estimates for both λ_0 and W_0, our proposed methods achieve better performance in terms of Frobenius norm across DGPs and sample sizes. Although the estimates of W_0 might be similar in both the exogenous and the endogenous cases, our method could obtain more reliable estimates of λ_0 as well as better recovery accuracy in the endogenous case.

6. Empirical Applications

This section illustrates the use of our estimators for social/spatial interactions. We consider two examples, applying our methods and examining the spillover effects. In each example, we first construct a raw spatial weight matrix and then apply Algorithms 1 and 2 described in the Appendix to obtain our estimates. To select the penalties, we use the interquartile range (IQR) of the raw weight-matrix entries as a robust scale estimate, denoted by $\widehat{\sigma}$. The sparse and low-rank penalties are parameterized as $C_S\widehat{\sigma}\log(n)$ and $C_L\widehat{\sigma}\sqrt{n}$, respectively, with the constants chosen by the Bayesian Information Criterion (BIC). We estimate the spillover coefficient $\widehat{\lambda}$ and refer the column sum of the Leontief inverse $(I_n - \widehat{\lambda}\widehat{W})^{-1}$ as the total network influence.

6.1. **Example 1.** In our first example, we examine an application to annual Gross Domestic Product (GDP) growth as our outcome of interest. Following Mankiw et al. (1992), the explanatory variables include the annual growth rate of the working-age population defined as individuals aged 15 to 64, denoted as X_1, and the log investment share, X_2, which serves as a proxy for the saving rate. We also include individual and time fixed effects in our specifications. The dataset is downloaded from the World Development Indicators (WDI) and covers a balanced panel of 23 economies from 1970 to 2023.[8]

[8]These economies are: Australia (AU), Canada (CA), Mainland China (CN), Denmark (DK), Finland (FI), France (FR), Germany (DE), Hong Kong SAR China (HK), Indonesia (ID), Ireland (IE), Italy (IT), Japan (JP), Korea (KR), Malaysia (MY), Mexico (MX), Netherlands (NL), New Zealand (NZ), Singapore (SG), Spain (ES), Sweden (SE), Thailand (TH), United Kingdom (GB), and United States (US).

Economic activity is well known to exhibit substantial spatial dependence (see, e.g., Krugman, 1991). Neighboring economies tend to trade more, share regional demand and supply shocks, and compete for mobile factors of production. For this reason, we use a geographic-proximity weight matrix as a natural and predetermined benchmark to study spillovers. The raw spatial weight matrix is constructed as the row-normalized inverse squared distance matrix based on Haversine distances and denoted by W_2.[9] Note that row-normalization on the weight matrix is standard in the spatial econometrics literature, as it facilitates a clear interpretation of the spatial autoregressive coefficient as the strength of the average influence. All off-diagonal entries of W_2 are strictly positive, which contrasts sharply with the sparsity assumptions sometimes invoked in the literature. This dense structure makes W_2 a natural candidate for our estimating approach, since many of the small off-diagonal weights likely reflect weak, noise-contaminated links rather than economically meaningful connections.

We estimate Model (3) with two-way fixed effects under two scenarios. Panel A includes only X_1 and X_2 as covariates. We assume that those variables are contemporaneously exogenous and use the spatial lags WX_1 and WX_2 as instruments. Panel B includes both the original covariates and the contextual variables WX_1 and WX_2 as regressors, using the second-order spatial lags W^2X_1 and W^2X_2 as instruments. Table 4 reports the deviations of the estimated weight matrices from the raw matrix W_2. We consider three alternative specifications: a purely sparse structure with $\tau = 0.0443$, a purely low-rank structure with $\nu = 0.2709$, and a combination of both with $\tau = 0.0443$ and $\nu = 0.2709$, where the penalties are selected according to BIC respectively. We first examine the differences between estimated adjacency matrices and the raw matrix W_2. We find that the purely low-rank estimator $\widehat{L}$ exhibits substantial departure from W_2, particularly in terms of the matrix maximum norm. Most of the elements in $\widehat{L}$ are small but not exactly 0, suggesting that a low-rank approximation alone may fail to preserve several important links in the original network. By contrast, the purely sparse estimator $\widehat{S}$ retains only 96 nonzero entries. Although this produces a more parsimonious network representation, it has the largest deviation from W_2 under l_2 norm. The low-rank-plus-sparse estimators provide a more balanced representation and both $\|\widehat{W}_2 - W_2\|_{\max}$ and $\|\widehat{W}_2 - W_2\|_2$ remain small relative to the corresponding deviations of the competitors. A further calculation shows that $\|\widehat{W}_2 - \widehat{S}\|_{\max} = 0.02$ and $\|\widehat{W}_2 - \widehat{S}\|_2 = 0.11$, indicating that it may capture diffuse connections that may be discarded by pure sparsification. Notably, the supervised low-rank and sparse estimates based on two panel specifications remain very similar to the unsupervised one, which is consistent with our theoretical findings. A further comparison of the heat maps of the raw W_2, the estimated $\widehat{W}_2$, and $\widehat{S}$ is also shown in Figure 5. We observe that the low-rank plus sparse estimate $\widehat{W}_2$ appears to preserve the dominant network structure while allowing for pervasive weak link effects.

[9]For economies $i \neq j$, we compute the Haversine great-circle distance D_{ij} between capital cities (in millions of meters) and then form the raw weight matrix by setting the (i, j) entry as $1/D_{ij}^2$ for $i \neq j$, and the diagonal entries as 0, and then apply row-normalization.

TABLE 4. Comparisons of Estimated Weight Matrices

Estimator	$\|\widehat{W} - W_2\|_{\max}$	$\|\widehat{W} - W_2\|_2$	Rank($\widehat{L}_2$)	Nonzero entries($\widehat{S}_2$)
$\widehat{S}$	0.044	0.385	–	96
$\widehat{L}$	0.269	0.349	14	–
$\widehat{W}_2$	0.044	0.289	2	78
$\widehat{W}_s$ (Panel A)	0.047	0.301	2	115
$\widehat{W}_s$ (Panel B)	0.045	0.299	2	116

Note: We calculate the differences between each estimated weight matrix and the raw weight matrix W_2. We consider the purely sparse matrix estimate $\widehat{S}$, the purely low-rank matrix estimate $\widehat{L}$, the low-rank plus sparse matrix estimate $\widehat{W}_2$ and its supervised estimates $\widehat{W}_s$ under model (3) in Panel A (without contextual effects) or Panel B (with contextual effects). Rank ($\widehat{L}$) reports the rank of the low-rank component, and nonzero entries ($\widehat{S}$) refer to the number of nonzero entries in the sparse component when the corresponding decomposition is available.

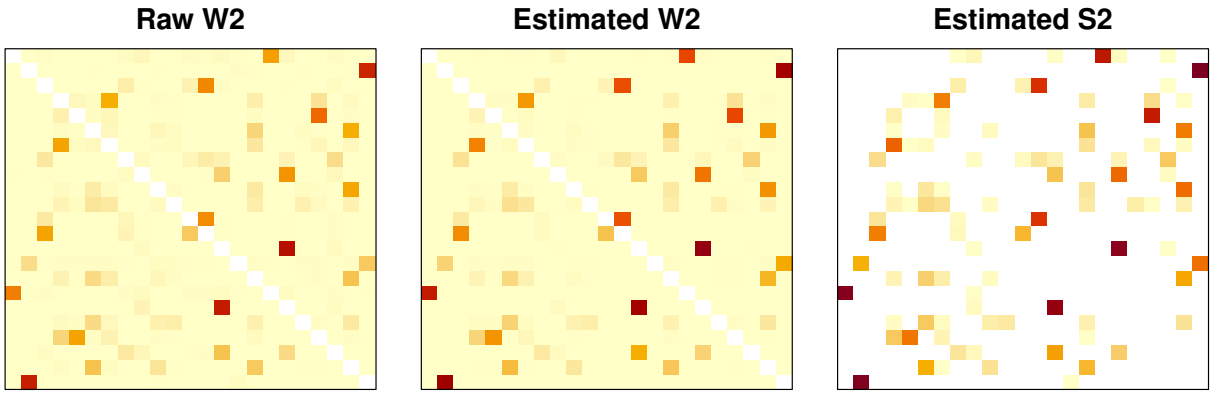


FIGURE 5. Heatmaps of different spatial weight matrices associated with 23 economies, including the raw weight matrix W_2 (left), the estimated low-rank and sparse matrix $\widehat{W}_2$ (center), and the estimated purely sparse matrix $\widehat{S}$ (right).

In the regression analysis, we evaluate the impact of the investment rate and working-age population growth, and report the estimated spillover coefficient $\widehat{\lambda}$ based on different weight matrices in Table 5. For each candidate weight matrix W, we also compute the column sums of the Leontief inverse $(I_n - \widehat{\lambda}\widehat{W})^{-1}$ and designate the economy with the largest column sum as the key player. Relative changes compared to the raw W_2 benchmark are also reported.

In Panel A (without contextual effects), the purely sparse estimator $\widehat{S}_2$ only retains the strong connections and reduces $\widehat{\lambda}$ by 16.7% and total influence (i.e., the corresponding column sum from the Leontief matrix, see above) by 11.3% compared to those obtained from the raw W_2. Conversely, the purely low-rank estimator $\widehat{L}_2$, amplifies network connections via aggregating pervasive weak links and increases $\widehat{\lambda}$ by 11.3% and total influence by 12.0%

compared to those obtained from the raw W_2. These results indicate that the sparse and low-rank components might capture distinct aspects of the network. The sparse component filters out scattered weak links, whereas the low-rank component summarizes pervasive dependence. Correspondingly, the low-rank-plus-sparse estimator $\widehat{W}_2$ produced more moderate deviations from the raw benchmark.

The benefits of our approach become more pronounced in Panel B with contextual effects. The raw W_2 produces a negative estimate $\widehat{\lambda}$ that is much lower than the our estimates. The sparse estimator and the low-rank-plus-sparse estimator produce moderate positive values that might be more reasonable. As higher-order spatial lags can magnify noise in the raw matrix, this reversal suggests that our approach can improve the plausibility and stability of the estimated spillover effects.

TABLE 5. GDP-Growth GMM Spillover Estimates and Influence Analysis

Weight matrix	$\widehat{\lambda}$	$\Delta\widehat{\lambda}$	Key player	Total influence	$\Delta influence$
Panel A: without contextual effects					
Raw W_2	0.377	0.0%	SGP	2.211	0.0%
$\widehat{S}_2$	0.314	-16.7%	SGP	1.961	-11.3%
$\widehat{L}_2$	0.420	11.3%	SGP	2.475	12.0%
$\widehat{W}_2$	0.354	-6.2%	SGP	2.147	-2.9%
$\widehat{W}_s$	0.356	-5.6%	SGP	2.178	-1.5%
Panel B: with contextual effects					
Raw W_2	-0.109	0.0%	MEX	0.985	0.0%
$\widehat{S}_2$	0.124	214.2%	SGP	1.290	31.0%
$\widehat{L}_2$	-0.052	52.3%	MEX	0.993	0.8%
$\widehat{W}_2$	0.095	187.6%	SGP	1.212	23.0%
$\widehat{W}_s$	0.095	186.9%	SGP	1.211	22.9%

We report the estimated spillover coefficient $\widehat{\lambda}$ obtained using different weight matrices. The raw W_2 matrix is the row-normalized inverse squared distance weight matrix. We consider its sparse estimator $\widehat{S}$, its low-rank estimator $\widehat{L}$, and its low-rank plus sparse estimator $\widehat{W}_2$ and its supervised version $\widehat{W}_s$. Panels A and B correspond to the scenarios without and with contextual effects respectively. We record the maximum column sum of the Leontief inverse $(I_n - \widehat{\lambda}\widehat{W})^{-1}$ and refer to it as the total influence of key player. The relative changes of the spillover effect and the total influence are also computed using the raw matrix W_2 as the baseline.

Moreover, the decomposition also yields interesting practical insights. Compared with $\widehat{S}_2$, we find that including the low-rank component in $\widehat{W}_2$ primarily rescales the overall magnitude of the spillover influence vector rather than altering the key player. This pattern indicates that,

while the identity of the key player remains robust to the decomposition strategy, the estimated strength of network spillovers depends materially on whether the low-rank component is retained, underscoring the practical relevance of the full low-rank plus sparse decomposition over the purely sparse estimator. Figure 6 presents the graphical representations of $\widehat{W}_2$, $\widehat{L}_2$, and $\widehat{S}_2$, respectively. $\widehat{W}_2$ admits the decomposition $\widehat{W}_2 = \widehat{S}_{\widehat{W}_2} + \widehat{L}_{\widehat{W}_2}$ and comparing the leading eigenvectors of $\widehat{W}_2$, $\widehat{L}_{\widehat{W}_2}$ and $\widehat{S}_{\widehat{W}_2}$, we find that $\mathbf{v}_{\widehat{W}_2} \approx 0.56\,\mathbf{v}_{\widehat{L}_{\widehat{W}_2}} + 0.48\,\mathbf{v}_{\widehat{S}_{\widehat{W}_2}}$ from a regression perspective as discussed in equation (C.1). Both the low-rank and the sparse components contribute substantially in this application, indicating that the identification of influential nodes depends not only on a few strong connections, but also on the cumulative effect of widespread weak links. Notably, as shown in Figure 7, $\mathbf{v}_{\widehat{S}_{\widehat{W}_2}}$ has many zero entries, and its nonzero entries concentrate on the Asian economies, whereas $\mathbf{v}_{\widehat{L}_{\widehat{W}_2}}$ loads on all economies but assigns slightly smaller weights to Western economies. One possible interpretation is that the sparse eigenvector captures localized regional clustering centered on Asia, while the low-rank eigenvector reflects a global common factor that is slightly attenuated for Western economies in accounting for the broader geographic diffusion inherent in the low-rank structure.

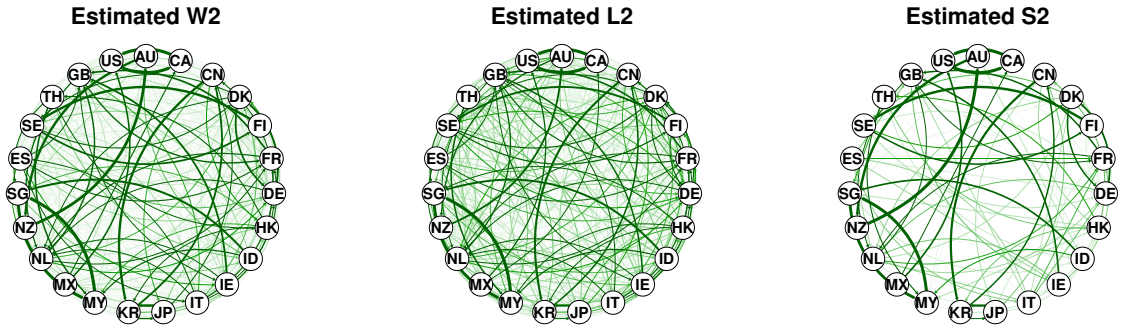


FIGURE 6. Graphical representation of different spatial weight matrix associated with 23 economies, including the estimated low-rank and sparse one $\widehat{W}_2$ (left), the estimated low-rank one $\widehat{L}$ (center) and the estimated purely sparse one $\widehat{S}$ (right).

6.2. **Example 2.** In our second example, we apply our methods to examine tax competition among the 48 contiguous U.S. states. This application was originally studied by Besley and Case (1995) using data from 1962 to 1988, with a pre-specified geographic weight matrix W_g defined as the row-normalized 0-1 neighborhood adjacency matrix setting 1 to state pairs that are geographically adjacent and zero, otherwise. Besley and Case (1995) emphasize a political-economy channel for tax-setting spillovers, as policymakers respond to voters who benchmark their policies against those of peer states. Consequently, states may face both political and economic incentives to align their fiscal policies with those of their neighbors. This dataset has recently been extended to cover 1962–2015 ($T = 53$) and analyzed by de Paula et al. (2025). We obtain the spatial weight matrix W_2, which is a row-normalized inverse squared distance matrix based on Haversine distances between state

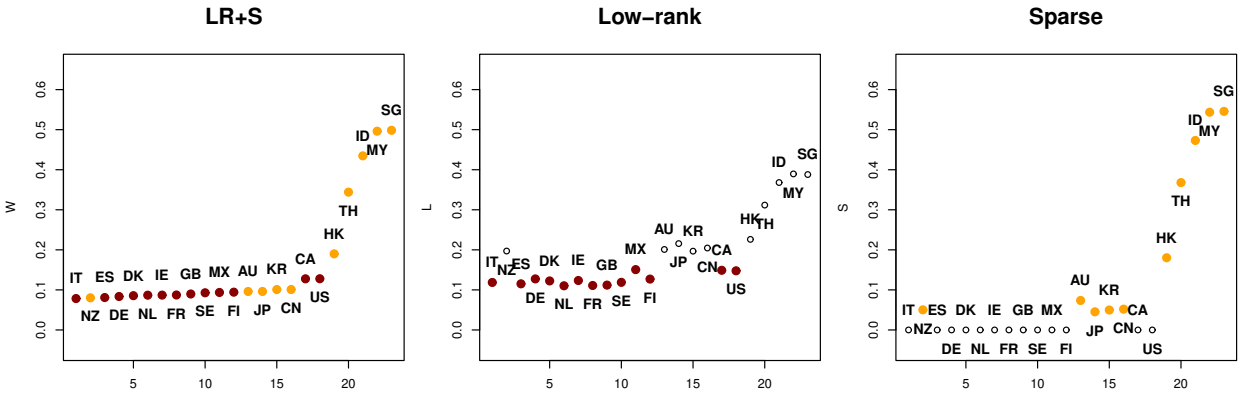


FIGURE 7. Plots of eigenvector centrality associated with different weight matrices. Denote the low-rank and sparse components of our estimator $\widehat{W}_2$ as $\widehat{L}$ and $\widehat{S}$, i.e. $\widehat{W}_2 = \widehat{L}_{\widehat{W}_2} + \widehat{S}_{\widehat{W}_2}$. From left to right, we provide the eigenvector centrality plots of $\widehat{W}_2$, $\widehat{L}_{\widehat{W}_2}$ and $\widehat{S}_{\widehat{W}_2}$ respectively. Orange dots mark economies whose entries in the leading eigenvector of $\widehat{S}_{\widehat{W}_2}$ are nonzero, while red dots mark economies whose entries in the leading eigenvector of $\widehat{L}_{\widehat{W}_2}$ are smaller than 0.2. The economies corresponding to red dots are Canada, Denmark, Finland, France, Germany, Ireland, Italy, Mexico, Netherlands, Spain, Sweden, United Kingdom, and United States, and those corresponding to orange dots are Australia, China, Hong Kong SAR China, Indonesia, Japan, Korea Rep., Malaysia, New Zealand, Singapore, Thailand.

capitals, as the observed weight matrix. In addition to endogenous spatial effects captured by WY, the covariates include state-level characteristics such as per-capita income, the unemployment rate, and the proportions of young and elderly populations, without and with their contextual covariates WX in Panel A and B respectively. We also include individual and time fixed effects in our specifications. Lagged neighbor income and lagged neighbor unemployment are used as instruments as in the original paper by Besley and Case (1995).

We consider three specifications, which yield a purely sparse matrix estimate $\widehat{S}_2$, a purely low-rank matrix estimate $\widehat{L}_2$, and a low-rank plus sparse estimate $\widehat{W}_2$. The purely sparse estimate $\widehat{S}_2$ captures the strong links and retains 90 nonzero entries. The estimate $\widehat{L}_2$ is a zero matrix under the selected penalty, suggesting that the data do not provide enough evidence of a low-rank component. The combined estimator $\widehat{W}_2$ coincides with the sparse-only estimator $\widehat{S}_2$, and hence we only report results about $\widehat{W}_2$ in the subsequent analysis. The supervised variants W_s in Panels A and B are very close to those of the unsupervised estimate $\widehat{W}_2$.

Figures 8 and 9 provide heatmaps and graphical representations of the raw geographic weight matrices W_g, W_2 and the estimate $\widehat{W}_2$, respectively. We find that the raw inverse geographic distance matrix W_2 is much denser than the raw geographic adjacency matrix W_g. It may be more vulnerable to measurement errors, and thus exaggerate the overall degree of network dependence and obscure the dominant transmission channels. By contrast, the estimated network discards many weak, noise-dominated connections, thus providing a clearer visualization of the significant links and improving interpretability.

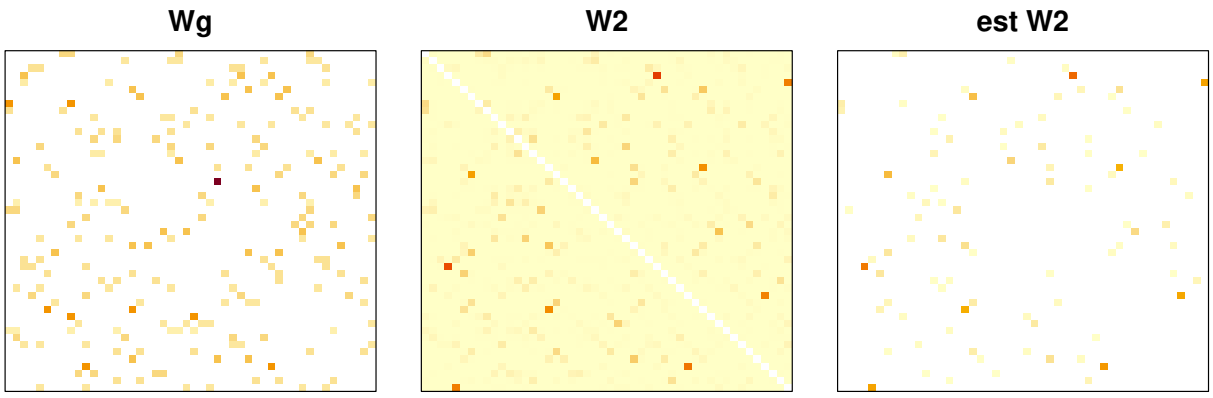


FIGURE 8. Heatmaps of adjacency matrices associated with 48 states. From left to right, this figure provides the heatmaps associated with the row-normalized 0-1 geographic adjacency matrix W_g, the row-normalized inverse squared distance matrix W_2, the estimated low-rank and sparse matrix $\widehat{W}_2$.

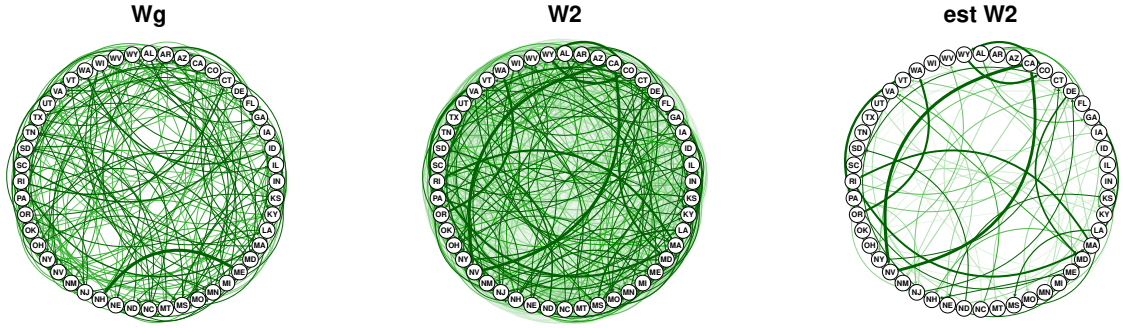


FIGURE 9. Graphical representations of different weight matrices. From left to right, this figure provides the network representation for the row-normalized neighborhood weight matrix W_g, the row-normalized inverse squared distance matrix W_2 and the estimated low-rank and sparse weight matrix $\widehat{W}_2$.

Table 6 provides a more detailed characterization of the state-level network structure. States are sorted according to their column sums, which measure the aggregate strength of their outward connections. Massachusetts is identified as the most influential state under all three weight matrices. Its out-degree is 1.70 under the row-normalized adjacency matrix W_g, 1.61 under the row normalized inverse squared distance matrix W_2, and 0.72 under the estimate $\widehat{W}_2$. The estimated network therefore preserves the leading position of Massachusetts while substantially reducing the magnitude of its estimated outward influence. This suggests that our estimate does not fundamentally alter the ranking of the major network hubs, but removes potentially noisy links that may inflate the overall strength of network connections rather than incorporating them into the core neighborhood topology. By contrast, the ranking based on W_g differs more noticeably because the geographical adjacency matrix only

captures direct neighboring relationships, whereas W_2 allows for distance-decaying connections between all states.

TABLE 6. Top 4 States Ranked by Out-degree

	Raw W_g		Raw W_2		$\widehat{W}_2$	
Rank	State	Out-degree	State	Out-degree	State	Out-degree
1	MA	1.70	MA	1.61	MA	0.72
2	GA	1.62	MD	1.57	MD	0.64
3	TN	1.58	RI	1.53	RI	0.63
4	ID, NH	1.53	DE	1.52	DE	0.63

Notes: We consider three choices of the weight matrix, including the raw row-normalized adjacency matrix W_g, the raw row-normalized inverse squared distance matrix W_2, and the estimate $\widehat{W}_2$. For each matrix, we compute the out-degree (column sums) and report the four states with the largest values in descending order. The abbreviations MA, GA, TN, ID, NH, MD, RI, DE stand for Massachusetts, Georgia, Tennessee, Idaho, New Hampshire, Maryland, Rhode Island and Delaware, respectively.

TABLE 7. Spillover Estimates and Influence Analysis

Weight matrix	$\widehat{\lambda}$	$\Delta\widehat{\lambda}$	Key region	Region impact	Δ Region impact
Panel A: without contextual effects					
Raw W_g	0.658	30.3%	Mountain	0.181	3.9%
Raw W_2	0.505	0.0%	SA	0.174	0.0%
$\widehat{W}_2$	0.220	-56.5%	SA	0.168	-3.3%
$\widehat{W}_s$	0.220	-56.5%	SA	0.168	-3.3%
Panel B: with contextual effects					
Raw W_g	1.006	-11.0%	NA	NA	NA
Raw W_2	1.130	0.0%	NA	NA	NA
$\widehat{W}_2$	0.811	-28.3%	SA	0.177	NA
$\widehat{W}_s$	0.811	-28.2%	SA	0.177	NA

Notes: We report the estimates of the spillover coefficients obtained using different spatial weight matrices. W_g and W_2 denote the row-normalized 0-1 adjacency matrix and the row-normalized inverse squared distance weight matrix respectively. $\widehat{W}_2$ and $\widehat{W}_s$ are the estimator and the supervised estimator constructed from the raw weight matrix W_2. If $|\widehat{\lambda}| < 1$, region impact is computed by first obtaining the column sums of the Leontief inverse, and then aggregating them according to the nine geographical divisions. The divisions with the largest total impact are reported as the key region. The relative changes of the spillover and the region impact are reported using the results associated with the raw W_2 matrix as the baseline when available.

Table 7 reports the estimated spillover coefficient $\widehat{\lambda}$ and the regional influence analysis using different weight matrices. We consider two scenarios, Panel A (without WX) and Panel B (with WX), respectively. In both panels, two raw weight matrices W_g and W_2 produce quite similar results, and they both exhibit much larger or even explosive values of $\widehat{\lambda}$ in Panel B. Such large values are likely driven by measurement error or noise in the raw adjacency matrices, yielding exaggerated spillover estimates that could be unreliable and potentially misleading. In contrast, the estimators, including the low-rank-plus-sparse estimator $\widehat{W}$, and its supervised variants, produce substantially smaller and more stable $\widehat{\lambda}$ values, and consistently identify the South Atlantic as the key region. Recall that each weight matrix is row-normalized prior to estimation. By the Perron–Frobenius theorem, the spectral radius of a row-normalized non-negative matrix is 1, and hence the standard Leontief-stability condition reduces to $|\widehat{\lambda}| < 1$ (LeSage and Pace, 2009). Explosive spillover estimates can cause some elements of the Leontief inverse to become negative and hard to interpret. These results indicate that our approach might yield more plausible and interpretable spillover effects, avoiding the inflated influence suggested by raw network matrices.

TABLE 8. Top Four States Most Strongly Influenced by Massachusetts

Weight matrix	1st	2nd	3rd	4th
Panel A: Without contextual effects				
Raw W_g	MA (1.30)	RI (0.59)	CT (0.48)	VT (0.45)
Raw W_2	MA (1.13)	RI (0.34)	NH (0.24)	CT (0.17)
$\widehat{W}_2$	MA (1.04)	RI (0.19)	NH (0.17)	ME (0.06)
$\widehat{W}_s$	MA (1.04)	RI (0.19)	NH (0.17)	ME (0.06)
Panel B: With contextual effects				
$\widehat{W}_2$	MA (2.51)	RI (1.88)	NH (1.82)	ME (1.48)
$\widehat{W}_s$	MA (2.52)	RI (1.88)	NH (1.82)	ME (1.48)

Notes: We evaluate the column corresponding to Massachusetts in the Leontief inverse and interpret its entries as the total influence that a shock originating from Massachusetts exerts. We report the four largest values in parentheses along with the corresponding states. Four choices of the weight matrix are considered: the raw adjacency matrix W_g, the raw row-normalized inverse squared distance matrix W_2, the estimator $\widehat{W}_2$ and the supervised estimator $\widehat{W}_s$. The abbreviations MA, RI, CT, VT, NH, ME stand for Massachusetts, Rhode Island, Connecticut, Vermont, New Hampshire and Maine, respectively. We do not include the results for Raw W_2 and Raw W_g in Panel B because the estimated spillover effects do not satisfy the Leontief-stable condition.

Table 8 further examines the propagation of shocks originating from Massachusetts. We calculate the Leontief Inverse and examine the four largest elements in the column that corresponding to Massachusetts. The two states most strongly influenced by a shock originating from Massachusetts are Massachusetts itself and Rhode Island. However, the magnitude of

the estimated network influence is attenuated. When contextual effects WX are incorporated, the Leontief influence values increase substantially. Under the estimate $\widehat{W}_2$, the influence measure for Massachusetts rises from 1.04 to 2.51, while the corresponding values for Rhode Island, New Hampshire, and Maine increase to 1.88, 1.82, and 1.48, respectively. Overall, the results consistently identify Massachusetts as an important regional hub and indicate that contextual effects could considerably amplify the propagation of state-level shocks. Our approach retains stronger local transmission channels while suppressing weaker long-range associations that may be sensitive to measurement error.

7. Conclusions

This paper introduces a robust framework for studying interaction effects in social and spatial networks under noisy adjacency matrices. By leveraging the low-rank and sparse structure of real-world networks, our de-noising approach, combining LASSO and nuclear norm penalization, significantly improves estimation accuracy. We propose two estimation procedures i.e. a two-step and a one-step supervised GMM estimator that outperform standard GMM by $50-80\%$ in RMSE under noise and endogeneity.

Applying our proposed methodology to examine the international spillover of economic growth and the tax competition across U.S. states, we find pronounced discrepancy in estimating the spillover effects, which highlights the essential role of accurate network denoising. Failing to distinguish genuine connectivity from measurement noise can systematically distort spillover coefficients as well as the strength and direction of cross-unit interactions. Moreover, our decomposition framework might provide a new perspective on complex spatial dependencies by disentangling global and local components. This decomposition not only sharpens the economic interpretation but also improve econometric inference, thereby contributing to credible spatial econometric analysis.

8. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used AI to suggest edits to some lengthy passages for concision and clarity. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Acemoglu, D., Carvalho, V. M., Ozdaglar, A., and Tahbaz-Salehi, A. (2012). The network origins of aggregate fluctuations. *Econometrica*, 80(5):1977–2016.

Acemoglu, D., Ozdaglar, A., and Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608.

Acemoglu, D., Ozdaglar, A., and Tahbaz-Salehi, A. (2016). Networks, shocks, and systemic risk. In *The Oxford Handbook of the Economics of Networks*. Oxford University Press.

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Annals of Statistics*, 40:1171–1197.

Anand, K., Craig, B., and von Peter, G. (2015). Filling in the blanks: Network structure and interbank contagion. *Quantitative Finance*, 15(4):625–636.

Araujo, R., Assunção, J., Hirota, M., and Scheinkman, J. A. (2023). Estimating the spatial amplification of damage caused by degradation in the Amazon. *Proceedings of the National Academy of Sciences*, 120(46):e2312451120.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70:191–221.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. Wanted: The key player. *Econometrica*, 74(5):1403–1417.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2013). The diffusion of microfinance. *Science*, 341(6144):1236498.

Barranca, V. J., Zhou, D., and Cai, D. (2015). Low-rank network decomposition reveals structural characteristics of small-world networks. *Physical Review E*, 92(6):062822.

Barrot, J.-N. and Sauvagnat, J. (2016). Input specificity and the propagation of idiosyncratic shocks in production networks. *The Quarterly Journal of Economics*, 131(3):1543–1592.

Battiston, S., Puliga, M., Kaushik, R., Tasca, P., and Caldarelli, G. (2012). Debtrank: Too central to fail? financial networks, the fed, and systemic risk. 2:541.

Besley, T. and Case, A. (1995). Incumbent behavior: Vote-seeking, tax-setting, and yardstick competition. *American Economic Review*, 85:25–45.

Boehm, C. E., Flaaen, A., and Pandalai-Nayar, N. (2019). Input linkages and the transmission of shocks: Firm-level evidence from the 2011 tōhoku earthquake. *Review of Economics and Statistics*, 101(1):60–75.

Boucher, V. and Houndetoungan, A. (2025). Estimating peer effects using partial network data. *Review of Economics and Statistics*, pages 1–48.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55.

Cai, J., Yang, D., Zhu, W., Shen, H., and Zhao, L. (2021). Network regressions and supervised centrality estimation-supplemental materials of proof. *Available at SSRN 3963526*.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Calvó-Armengol, A., Patacchini, E., and Zenou, Y. (2009). Peer effects and social networks in education. *Review of Economic Studies*, 76(4):1239–1267.

Candelaria, L. E. and Ura, T. (2023). Identification and inference of network formation games with misclassified links. *Journal of Econometrics*, 235(2):862–891.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM*, 58(3):1–37.

Cao, F., Chen, J., Ye, H., Zhao, J., and Zhou, Z. (2017). Recovering low-rank and sparse matrix based on the truncated nuclear norm. *Neural Networks*, 85:10–20.

Carvalho, V. M. (2014). From micro to macro via production networks. *Journal of Economic Perspectives*, 28(4):23–48.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2010). Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2009). Sparse and low-rank matrix decompositions. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 962–967.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*.

Chernozhukov, V., Huang, C., and Wang, W. (2021). Uniform inference on high-dimensional spatial panel networks. *arXiv preprint arXiv:2105.07424*.

Chudik, A. and Pesaran, M. H. (2013). Econometric analysis of high dimensional VARs featuring a dominant unit. *Econometric Reviews*, 32(5-6):592–649.

de Giorgi, G., Frederiksen, A., and Pistaferri, L. (2020). Consumption network effects. *Review of Economic Studies*, 87(1):130–163.

de Paula, A. (2017). Econometrics of network models. In *Advances in Economics and Econometrics: Theory and Applications, Eleventh World Congress*, pages 268–323.

de Paula, A., Rasul, I., and Souza, P. (2025). Identifying network ties from panel data: Theory and an application to tax competition. *Review of Economic Studies*, 92:2691–2729.

Degryse, H. and Nguyen, G. (2007). Interbank exposures: an empirical investigation of systemic risk in the belgian banking system. 3:123–172.

Diebold, F. X. and Yılmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 182(1):119–134.

Elhorst, J. P. (2014). *Spatial Panel Data Models*, pages 37–93. Springer.

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics*, 39(6):3320–3356.

Fan, J., Wang, W., and Zhong, Y. (2018). An l_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18:7608–7649.

Fisman, R. and Wei, S.-J. (2004). Tax rates and tax evasion: Evidence from "missing imports" in China. *Journal of Political Economy*, 112(2):471–500.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica*, 79(3):733–772.

Galeotti, A., Golub, B., and Goyal, S. (2020). Targeting interventions in networks. *Econometrica*, 88(6):2445–2471.

Gamble, J., Chintakunta, H., Wilkerson, A., and Krim, H. (2016). Node dominance: Revealing community and core-periphery structure in social networks. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):186–199.

Getis, A. and Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, 36(2):90–104.

Graham, B. and de Paula, Á. (2020). *The econometric analysis of network data*. Academic Press.

Greenwood, R., Landier, A., and Thesmar, D. (2015). Vulnerable banks. *Journal of Financial Economics*, 115(3):471–485.

Hardy, M., Heath, R. M., Lee, W., and McCormick, T. H. (2019). Estimating spillovers using imprecisely measured networks. *arXiv preprint arXiv:1904.00136*.

Hayes, A. and Levin, K. (2024). Minimax rates for the linear-in-means model reveal an identifiability-estimability gap. *arXiv preprint arXiv:2410.10772*.

Jackson, M. O. (2008). *Social and economic networks*, volume 3. Princeton university press Princeton.

Kapetanios, G., Pesaran, M. H., and Reese, S. (2021). Detection of units with pervasive effects in large panel data models. *Journal of Econometrics*, 221(2):510–541.

Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17(1):99–121.

Krugman, P. (1991). Increasing returns and economic geography. *Journal of Political Economy*, 99(3):483–499.

Kuersteiner, G. M. and Prucha, I. R. (2020). Dynamic spatial panel models: Networks, common shocks, and sequential exogeneity. *Econometrica*, 88(5):2109–2146.

Lam, C. and Souza, P. C. L. (2020). Estimation and selection of spatial weight matrix in a spatial lag model. *Journal of Business & Economic Statistics*, 38(3):693–710.

Lee, L.-F. (2003). Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews*, 22(4):307–335.

Lee, L.-F. (2007a). GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, 137(2):489–514.

Lee, L.-F. (2007b). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics*, 140(2):333–374.

Lee, L.-F., Liu, X., and Lin, X. (2010). Specification and estimation of social interaction models with network structures. *The Econometrics Journal*, 13(2):145–176.

Lee, L.-F., Yang, C., and Yu, J. (2023). QML and efficient GMM estimation of spatial autoregressive models with dominant (popular) units. *Journal of Business & Economic*

Statistics, 41(2):550–562.

Lee, L.-F. and Yu, J. (2014). Efficient GMM estimation of spatial dynamic panel data models with fixed effects. *Journal of Econometrics*, 180(2):174–197.

LeSage, J. P. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. CRC Press.

Lewbel, A., Qu, X., and Tang, X. (2023). Social networks with unobserved links. *Journal of Political Economy*, 131(4):898–946.

Lewbel, A., Qu, X., and Tang, X. (2024). Ignoring measurement errors in social networks. *Econometrics Journal*, 27(2):171–187.

Mankiw, N., Romer, D., and Weil, D. (1992). A contribution to the empirics of economic growth. *Quarterly Journal of Economics*, 107(407):437.

Manresa, E. (2016). Estimating the structure of social interactions using panel data. Technical report.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60(3):531–542.

Mezzadri, F. (2007). How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 54(5):592–604.

Newman, M. (2010). *Networks: An introduction*. Oxford University Press.

Pesaran, M. H. and Yang, C. F. (2020). Econometric analysis of production networks with dominant units. *Journal of Econometrics*, 219(2):507–541.

Pesaran, M. H. and Yang, C. F. (2021). Estimation and inference in spatial models with dominant units. *Journal of Econometrics*, 221(2):591–615.

Shin, Y. and Thornton, M. A. (2021). Dynamic network analysis via diffusion multipliers. Technical report, University of York.

Stewart, G. W. (1980). The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, 17(3):403–409.

Upper, C. (2011). Simulation methods to assess the danger of contagion in interbank markets. *Journal of Financial Stability*, 7(3):111–125.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Woodbury, M. (1950). Inverting modified matrices.

Yang, K. and Lee, L.-f. (2017). Identification and QML estimation of multivariate and simultaneous equations spatial autoregressive models. *Journal of Econometrics*, 196(1):196–214.

Zenou, Y. (2016). Key Players. In *The Oxford Handbook of the Economics of Networks*. Oxford University Press.

Zhu, X., Huang, D., Pan, R., and Wang, H. (2020). Multivariate spatial autoregressive model for large scale social networks. *Journal of Econometrics*, 215(2):591–606.

A. Algorithms

B. Additional Numerical Results

Algorithm 1 Alternating minimization algorithm to obtain $\widehat{L}$ and $\widehat{S}$

Require: Observed matrix W, penalty parameters $\nu_n > 0$ and $\tau_n > 0$, and a positive constant tol

1: Initialize $\widehat{L}_0$ as $\mathbf{0}_{n\times n}$, and $\widehat{W}_0 = W$
2: **while** not converged **do**
3: $\widehat{S}_{k+1} \leftarrow \mathbf{S}_{\tau_n}(W - \widehat{L}_k)$, where $\mathbf{S}_{\tau_n}(A)$ is denoted as the element-wise soft thresholding estimator proposed by Friedman et al. (2010), i.e., $\max((A_{ij}-\tau_n),0)+\min((A_{ij}+\tau_n),0)$ for $i \neq j$.(set diagonal entries to zero.)
4: $\widehat{L}_{k+1} \leftarrow \mathrm{SVT}_{\nu_n}(W-\widehat{S}_{k+1})$, where $\mathrm{SVT}_{\nu_n}(A)$ denotes the singular value decomposition with thresholding proposed by Cai et al. (2010). Specifically, let matrix A admit the singular value decomposition $A = UDV^\top$. Then $\mathrm{SVT}_{\nu_n}(A) = UD_{\nu_n}V^\top$, where D and D_{ν_n} are both diagonal matrices, whose ith diagonal element satisfies $(D_{\nu_n})_i = \max(D_i - \nu_n, 0)$.
5: Set $\widehat{W}_{k+1} = \widehat{L}_{k+1} + \widehat{S}_{k+1}$. Check convergence condition: $\|\widehat{W}_{k+1} - \widehat{W}_k\|_F \leq tol$
6: **end while**
7: **return** $\widehat{L} \leftarrow \widehat{L}_{k+1}$ and $\widehat{S} \leftarrow \widehat{S}_{k+1}$, $\widehat{W} \leftarrow \widehat{W}_{k+1}$

Algorithm 2 Supervised estimation of spatial autoregressive model

Require: Observed matrix W, outcome $nT \times 1$ vector Y, regressors $nT \times K$ matrix X, GMM weight matrix $\widehat{\Lambda}_n^{-1}$, penalty parameters ν_{nT}, τ_{nT} and ξ_{nT}, tolerances $tol_0 > 0$, M and $tol_1 > 0$.

1: Set $\check{Z}_t = (X_t, MX_t, \cdots, M^dX_t)$. Let Z_t be the m linearly independent columns of $\check{Z}_t$. Define $nT \times m$ matrix $Z = (Z_1^\top, \cdots, Z_T^\top)^\top$.
2: Set θ_0 as the GMM estimate using W, initialize $\widehat{W}_0$ and $\widehat{L}_0$ as W and $0_{n\times n}$ respectively.
3: **while** not converged **do**
4: Obtain $\widetilde{W}_{k+1}$ by minimizing $\xi_{nT}J_{nT}(\theta_k, \widetilde{W}, M; \widehat{\Lambda}_n^{-1}) + \frac{1}{2}\|W - \widetilde{W}\|_F^2$ over $\widetilde{W}$. Recall the definition of $J_{nT}(.)$ and ξ_{nT} in Equation (8).
5: Obtain $\widehat{L}_{k+1}$ and $\widehat{S}_{k+1}$ using Algorithm 1 with W replaced by $\widetilde{W}_{k+1}$. Set $\widehat{W}_{k+1} = \widehat{L}_{k+1} + \widehat{S}_{k+1}$.
6: Update $\widehat{\theta}_{k+1}$ as the GMM estimate using $\widehat{W}_{k+1}$.
7: Check convergence conditions $\|\widehat{\theta}_{k+1} - \widehat{\theta}_k\|_2 \leq tol_0$, $\|\widehat{W}_{k+1} - \widehat{W}_k\|_F \leq tol_1$, for positive constants $tol_0, tol_1 > 0$.
8: **end while**
9: **return** $\hat{\theta} \leftarrow \widehat{\theta}_{k+1}$, $\widehat{L} \leftarrow \widehat{L}_{k+1}$, $\widehat{S} \leftarrow \widehat{S}_{k+1}$, and $\widehat{W} \leftarrow \widehat{W}_{k+1}$

C. Low-rank plus sparse decomposition

C.1. **Motivation for Low-Rank Plus Sparse Structure.** The low-rank plus sparse decomposition is useful in network analysis because it can improve interpretability and computational efficiency across various applications. Additionally, it is closely related to eigenvector centrality and the Leontief inverse, both of which play crucial roles in policy targeting by identifying key structural dependencies and intervention points (Jackson, 2008; Newman, 2010).

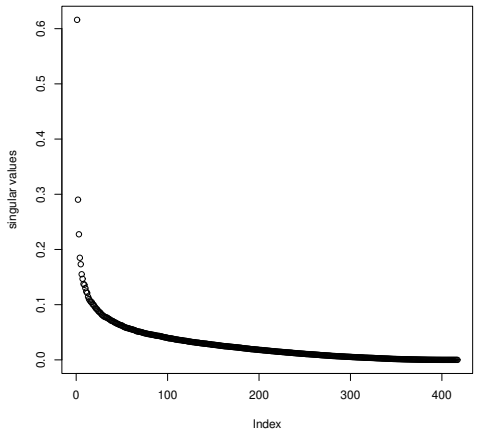


FIGURE B.1. Plots of singular values associated with the tiny weight matrix which is calculated by subtracting the sparse network using a threshold 0.05 from the input-output weight table 2002.

C.1.1. *Application to eigenvector centrality.* Our decomposition can be useful in interpreting and understanding centrality as such measure can be shown to be dominated by either the sparse or low-rank components in certain potentially relevant circumstances. The (eigenvector) centrality of each node corresponds to the eigenvector entries attached to the highest eigenvalue of the adjacency matrix representing the network.[10] The low-rank plus sparse decomposition separates the adjacency matrix W_0 into a low-rank structure L_0^* (or L_0, before diagonal adjustment), representing the core economic interactions, and a sparse component S_0^* (or S_0), capturing highly influential entities or individual heterogeneity (Gabaix, 2011; Carvalho, 2014). The eigenvector centrality under this decomposition expands to:

$$(L_0^* + S_0^*)\mathbf{v}_{W_0} = \lambda_{W_0}\mathbf{v}_{W_0},$$

where λ_{W_0} and $\mathbf{v}_{W_0}$ are the leading eigenvalue and eigenvector associated with W_0. The entries of $\mathbf{v}_{W_0}$ are related to the leading eigenvectors of L_0^* and S_0^*, denoted by $\mathbf{v}_{L_0^*}$ and $\mathbf{v}_{S_0^*}$, respectively, which capture different aspects of the underlying structure. If the leading eigenvector of $\mathbf{v}_{W_0}$, is closer to $\mathbf{v}_{L_0^*}$ or $\mathbf{v}_{S_0^*}$, this indicates that the low-rank or the sparse structure, respectively, is more dominant.

We now formalize how $\mathbf{v}_{W_0}$ is informed by $\mathbf{v}_{L_0^*}$ and $\mathbf{v}_{S_0^*}$. Let $\rho_{L_0^*S_0^*}$ denote the inner product between $\mathbf{v}_{L_0^*}$ and $\mathbf{v}_{S_0^*}$, i.e., $\rho_{L_0^*S_0^*} = \mathbf{v}_{L_0^*}^\top \mathbf{v}_{S_0^*}$. Similarly define $\rho_{W_0L_0^*} = \mathbf{v}_{W_0}^\top \mathbf{v}_{L_0^*}$ and $\rho_{W_0S_0^*} = \mathbf{v}_{W_0}^\top \mathbf{v}_{S_0^*}$ and let $k_{L_0^*}\mathbf{v}_{L_0^*} + k_{S_0^*}\mathbf{v}_{S_0^*}$ denote the best linear approximation of $\mathbf{v}_{W_0}$ using $\mathbf{v}_{L_0^*}$ and $\mathbf{v}_{S_0^*}$. Then from a regression perspective, it satisfies:

$$k_{L_0^*} = \frac{\rho_{W_0L_0^*} - \rho_{L_0^*S_0^*}\rho_{W_0S_0^*}}{1 - \rho^2_{L_0^*S_0^*}}, \qquad k_{S_0^*} = \frac{\rho_{W_0S_0^*} - \rho_{L_0^*S_0^*}\rho_{W_0L_0^*}}{1 - \rho^2_{L_0^*S_0^*}} \tag{C.1}$$

[10]For a connected network W_0 whose entries are all non-negative, the Perron-Frobenius theorem implies that its maximum eigenvalue (in terms of radius) is real and non-negative.

(see Appendix E.2). Further calculations show that the coefficients $k_{L_0^*}$ and $k_{S_0^*}$ should be related to the maximum eigenvalues of L_0^* and S_0^* respectively, i.e. $\lambda_{L_0^*}$ and $\lambda_{S_0^*}$. If L_0^* and S_0^* are symmetric, we have

$$\frac{k_{L_0^*}}{k_{S_0^*}} \approx \frac{\rho_{L_0^*S_0^*}(\lambda_{L_0^*} - \mathbf{v}_{S_0^*}^\top L_0^* \mathbf{v}_{S_0^*})}{\lambda_{S_0^*} + \mathbf{v}_{S_0^*}^\top L_0^* \mathbf{v}_{S_0^*} - \lambda_{L_0^*} - \mathbf{v}_{L_0^*}^\top S_0^* \mathbf{v}_{L_0^*}}, \text{ if } \lambda_{L_0^*} + \mathbf{v}_{L_0^*}^\top S_0^* \mathbf{v}_{L_0^*} < \lambda_{S_0^*} + \mathbf{v}_{S_0^*}^\top L_0^* \mathbf{v}_{S_0^*}, \quad \text{(C.2)}$$

$$\frac{k_{S_0^*}}{k_{L_0^*}} \approx \frac{\rho_{L_0^*S_0^*}(\lambda_{S_0^*} - \mathbf{v}_{L_0^*}^\top S_0^* \mathbf{v}_{L_0^*})}{\lambda_{L_0^*} + \mathbf{v}_{L_0^*}^\top S_0^* \mathbf{v}_{L_0^*} - \lambda_{S_0^*} - \mathbf{v}_{S_0^*}^\top L_0^* \mathbf{v}_{S_0^*}}, \text{ if } \lambda_{L_0^*} + \mathbf{v}_{L_0^*}^\top S_0^* \mathbf{v}_{L_0^*} > \lambda_{S_0^*} + \mathbf{v}_{S_0^*}^\top L_0^* \mathbf{v}_{S_0^*}, \quad \text{(C.3)}$$

where the approximation error diminishes when $\rho_{L_0^*S_0^*} \to 0$. (We also derive the expression when L_0^* or S_0^* is asymmetric in Appendix E.2.) In practical scenarios involving large networks, $\mathbf{v}_{L_0^*}$ is typically dense, with few zero entries, while $\mathbf{v}_{S_0^*}$ tends to be sparse, featuring many zero entries, thus implying that the correlation $\rho_{L_0^*S_0^*}$ is small. Consequently, if $\rho_{L_0^*S_0^*}$ is close to 0, then $k_{L_0^*}/k_{S_0^*} \approx 0$ if (C.2) holds or $k_{S_0^*}/k_{L_0^*} \approx 0$ if (C.3) holds. The inequality $\lambda_{L_0^*} + \mathbf{v}_{L_0^*}^\top S_0^* \mathbf{v}_{L_0^*} \lessgtr \lambda_{S_0^*} + \mathbf{v}_{S_0^*}^\top L_0^* \mathbf{v}_{S_0^*}$ determines which component L_0^* or S_0^* plays a more dominant role in shaping the eigenvector centrality of W_0. Specifically, the left-hand side approximates the effective contribution of the low-rank component L_0^* (and the part of S_0^* aligned with it), whereas the right-hand side captures the influence of the sparse component S_0^*. In contrast, the inner product $\mathbf{v}_{L_0^*}^\top \mathbf{v}_{S_0^*}$ measures the degree of alignment between the leading eigenvectors of L_0^* and S_0^*, indicating whether the two structures reinforce or counteract each other in determining the dominant direction of W_0. For the detailed calculations above, please refer to Section E.2 in the Appendix.

Ultimately, this means that when $\rho_{L_0^*S_0^*} \to 0$ we have:

$$\mathbf{v}_{W_0} \approx k_{L_0^*} \mathbf{v}_{L_0^*} \quad \text{or} \quad \mathbf{v}_{W_0} \approx k_{S_0^*} \mathbf{v}_{S_0^*}.$$

Therefore, we obtain a measure that either emphasizes the systemic structure of the network L_0^* or the eigenvector centrality of S_0^* which can be attributed to the individual shocks or dominant units (Battiston et al., 2012). For example, in interbank networks, systemically important banks often dominate centrality rankings when using the full adjacency matrix (Degryse and Nguyen, 2007). Applying the low-rank plus sparse decomposition allows us to disentangle the influence of connections with some important institutions (captured in S_0^*) from the broader systemic structure (captured in L_0^*). This approach thus potentially provides a more nuanced understanding of systemic risk, as it distinguishes between the centrality of highly connected hubs and the underlying economic network (Acemoglu et al., 2015; Greenwood et al., 2015).

To illustrate this, we now present two numerical examples to examine the underlying structure of the dominant eigenvector for a network with $n = 100$. In both examples, the network $W_0 = L_0^* + S_0^*$, where $L_0 = \mathbf{v}_{L_0}\mathbf{v}_{L_0}^\top$ has the leading eigenvector $\mathbf{v}_{L_0}$ whose elements are all $1/10$, and $L_0^* = L_0 - 1/100 I_{100}$ with $\mathbf{v}_{L_0^*}$=$\mathbf{v}_{L_0}$.The sparse matrix has the form $S_0^* = kS_A^*$ for some pre-specified sparse matrix S_A^* with zero diagonals. The constant k affects the relative strength between the low-rank component L_0^* and the sparse component S_0^*, and we consider $k = 0.4$ and $k = 4$. In Case 1, S_A^* is chosen such that it only has two nonzero elements

$S^*_{A,12} = S^*_{A,21} = 1$. In this case, the maximum eigenvalues of L^*_0 and S^*_0 are 0.99 and k respectively. If $k = 0.4$, the calculated leading eigenvectors show that $\mathbf{v}_{W_0} \approx 0.983\mathbf{v}_{L^*_0} + 0.091\mathbf{v}_{S^*_0}$, signifying a predominantly uniform structure with minor local adjustments. However, when $k = 4$, the eigenvalue of S^*_0 dominates that of L^*_0, and we find that $\mathbf{v}_{W_0} \approx 0.046\mathbf{v}_{L^*_0} + 0.992\mathbf{v}_{S^*_0}$, reflecting a reduced influence of L^*_0 due to the increased eigenvalue of S^*_0.

In Case 2, we set S^*_A to be an asymmetric sparse matrix with $S^*_{A,12} = 1$ and $S^*_{A,i1} = 1$ for $2 \leq i \leq 30$. This represents the case when the first individual is a key player that affects 29 individuals. Because S^*_A is asymmetric, its nonzero eigenvalues are ± 1; we take $\mathbf{v}_{S^*_0}$ to be the eigenvector associated with the eigenvalue $+1$. This $\mathbf{v}_{S^*_0}$ has more nonzero entries than in Case 1, so $\rho_{L^*_0 S^*_0} = \mathbf{v}^\top_{L^*_0}\mathbf{v}_{S^*_0}$ is larger. When $W_0 = L^*_0 + kS^*_A$ with $k = 0.4$, $\mathbf{v}_{W_0} \approx 0.845\,\mathbf{v}_{L^*_0} + 0.244\,\mathbf{v}_{S^*_0}$; when $k = 4$, $\mathbf{v}_{W_0} \approx 0.148\,\mathbf{v}_{L^*_0} + 0.911\,\mathbf{v}_{S^*_0}$, implying a negligible correction from L^*_0.

C.1.2. *Application to the Leontief inverse.* We also highlight how our decomposition is reflected in the Leontief inverse for W_0. Defining I_n as the $n \times n$ identity matrix, the Leontief inverse is given by $(I_n - \lambda_0 W_0)^{-1}$. For notational simplicity, throughout this subsection we set $\lambda_0 = 1$ and write the Leontief inverse as $(I_n - W_0)^{-1}$; the general case follows by replacing W_0 with $\lambda_0 W_0$ throughout. Assume that the low-rank component L_0 admits the singular value decomposition as $L_0 = UDV^\top$, where D is a square diagonal matrix of dimension $r \times r$, and U and V are the matrices that collect the left and right singular vectors respectively. When $r < n$, this allows dimensionality reduction while retaining key structural information. We seek to calculate the Leontief matrix: $(I_n - W_0)^{-1} = (I_n - (UDV^\top + S_0))^{-1}$. Applying the Woodbury matrix identity (Woodbury, 1950):

$$(A + UDV^\top)^{-1} = A^{-1} - A^{-1}U(D^{-1} + V^\top A^{-1}U)^{-1}V^\top A^{-1},$$

where $A = I_n - S_0$ is assumed to be invertible, we have

$$(I_n - UDV^\top - S_0)^{-1} = (I_n - S_0)^{-1} + (I_n - S_0)^{-1}U(D^{-1} - V^\top(I_n - S_0)^{-1}U)^{-1}V^\top(I_n - S_0)^{-1}. \tag{C.4}$$

Denote $\|.\|_{\max}$ and $\|.\|_2$ as the maximum norm and two-norm of a matrix respectively. Suppose $\|\mathrm{diag}(L_0)\|_{\max} \to 0$. By matrix algebra[11] and $S^*_0 - S_0 = \mathrm{diag}(L_0)$, we have

$$\|(I_n - S^*_0)^{-1} - (I_n - S_0)^{-1}\|_2 = \|(I_n - S^*_0)^{-1}\mathrm{diag}(L_0)(I_n - S_0)^{-1}\|_2 \to 0.$$

If we substitute $(I_n - S_0)^{-1}$ with $(I_n - S^*_0)^{-1}$ in Equation (C.4) and use Woodbury's formula, we have

$$(I_n - W_0)^{-1} \approx (I_n - S^*_0)^{-1} + (I_n - S^*_0)^{-1}U(D^{-1} - V^\top(I_n - S^*_0)^{-1}U)^{-1}V^\top(I_n - S^*_0)^{-1}. \tag{C.5}$$

This decomposition reveals two key effects in the network. The first term, $(I_n - S^*_0)^{-1}$, represents the Leontief matrix for the sparse component, capturing localized effects such as idiosyncratic shocks or sporadic interactions. For example, in economic networks, this can model firm-specific disruptions or localized financial shocks Gabaix (2011); Acemoglu et al.

[11] Take $A = I_n - S^*_0$ and $B = I_n - S_0$ and note that $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$.

(2012). In supply chains, it reflects disruptions in specific suppliers without affecting the entire network Barrot and Sauvagnat (2016); Boehm et al. (2019).

The second term in the Woodbury expansion of $(I_n - W_0)^{-1}$ captures the influence of dense network structures, exploiting the low-rank form of U and V. When $L_0 = k\mathbf{v}_{L_0}\mathbf{v}_{L_0}^\top$, the second term in Equation (C.5) yields, for symmetric S_0^*, a rank-one matrix $\tilde{k}\,\tilde{v}\tilde{v}^\top$, where $\tilde{v} = (I_n - S_0^*)^{-1}\mathbf{v}_{L_0}$ and $\tilde{k} = \left(1/k - \mathbf{v}_{L_0}^\top (I_n - S_0^*)^{-1}\mathbf{v}_{L_0}\right)^{-1}$, which reflects the difference between $(I_n - W_0)^{-1}$ and $(I_n - S_0^*)^{-1}$ due to the presence of L_0. Note that $\tilde{k}\tilde{v}\tilde{v}^\top$ lies along the direction of $\tilde{v}$, with both $\mathbf{v}_{L_0}$ and S_0^* contributing to it. Its effect can be substantial, particularly when L_0 is dominant. In economic networks, this balances uniform information spread, as in Jackson (2008) where dense ties drive equilibrium outcomes in coordination games, against localized dynamics. Thus, the correction reconciles global connectivity, akin to Acemoglu et al. (2012)'s analysis of aggregate economic fluctuations via network linkages.

When studying the network effect, it is often assumed W_0 has bounded eigenvalues even when the dimension of the network grows. If the network is highly connected, columns of U and V are dense vectors with many non-zero entries whose magnitude must converge to 0 due to the bounded row-sum constraint. Unlike factor models, the singular values in D may not diverge as n increases, which makes the elements of L_0 corresponding to a dense network being very small. However, our calculation above implies that in a dense network, naive thresholding methods are insufficient, as this might result in missing the second term in Equation (C.5) and inconsistent evaluation of network effects that overlook systemic dependencies. We empirically illustrate this point in our empirical application Section 6.

Moreover, utilizing the low-rank and sparse structure, we only need $O(nr + s)$ parameters to store the matrix information, where $r = \text{rank}(L_0)$ and $s = \|S_0\|_0$, which highlights the computational advantages we refer to earlier in the section. Note that it costs less time to compute the inverse of a sparse matrix compared to a dense matrix. With the help of Equation (C.4) or (C.5), we could simplify the calculation of the Leontief matrix.

D. Identification

In this section, we discuss the identification issue, i.e., given a weight matrix W_0 that admits the low-rank and sparse decomposition, it could be uniquely written as $W_0 = L_0 + S_0$ under some regularity conditions. As discussed earlier, we focus on L_0 and S_0 in the theoretical derivations, since the properties of L_0^* and S_0^* can be derived from them. Subsection D.1 presents conditions regarding the identification of L_0 and S_0, and Subsection D.2 discusses the conditions using examples.

Before proceeding, we recall from Section 4 the sparsity measures $m_r(A)$, $m_c(A)$, $m_s(A)$, and $\deg_{\max}(A) = \max\{m_r(A), m_c(A)\}$, and additionally define the *incoherence index* of a nonzero matrix A as $\text{inc}(A) = \sqrt{\max(\mu_u(A), \mu_v(A))}$, where

$$\mu_u(A) \coloneqq \max_{1\le i\le n} \sum_{j=1}^{r} u_{ij}^2(A), \qquad \mu_v(A) \coloneqq \max_{1\le i\le n} \sum_{j=1}^{r} v_{ij}^2(A),$$

r is the rank of A, and $u_{ij}(A)$, $v_{ij}(A)$ are the j-th elements of the i-th left and right singular vectors. The measure is scale-invariant and reflects whether A concentrates its mass in few entries: for a diagonal matrix with distinct diagonal entries $\mathrm{inc}(A) = 1$, whereas for a fully connected matrix with equal entries $\mathrm{inc}(A) = n^{-1/2}$.

The incoherence index controls the maximum entry via $\|L_0\|_{\max} \leq \sigma_1(L_0)\,\mathrm{inc}(L_0)^2$,[12] so when $\sigma_1(L_0) = O(1)$, the quantity $\mathrm{inc}(L_0)^2$ plays the same role as $1/\sqrt{m_r(L_0)m_c(L_0)}$ in the coherence bound of Assumption 1(i): both equal $1/n$ for a fully dense L_0 and are of constant order for a spiked L_0. We use $\mathrm{inc}(L_0)$ here because it yields the sharp identification condition $\mathrm{inc}(L_0)\ \deg_{\max}(S_0) \leq 1/16$.

D.1. **Identification conditions.** In the following, we provide a theoretical result that establishes the conditions under which L_0 and S_0 can be disentangled.

Assumption 9 (Low rank plus sparse components)**.** Suppose W_0 is an $n \times n$ matrix which can be decomposed as $W_0 = L_0 + S_0$, where L_0 is a low-rank matrix of a rank $0 \leq r < n$ and $S_0 \neq \mathbf{0}_{n\times n}$. Moreover,

$$\mathrm{inc}(L_0)\deg_{\max}(S_0) \leq 1/16.$$

Lemma 1. *Under Assumption 9, for any given decomposition $W_0 = L_0 + S_0$ with $S_0 \neq \mathbf{0}_{n\times n}$, the pair (L_0, S_0) can be uniquely recovered.*[13]

Assumption 9 is a direct consequence of Corollary 3 in Chandrasekaran et al. (2009). This assumption imposes that the sparse matrix S_0 is not excessively dense, while the low-rank matrix L_0 must not be overly sparse. The incoherence measure provides a quantitative assessment of the dense level of L_0. Specifically, if L_0 is sparse, its incoherence measure is of constant order, which might preclude the above inequality from holding when $S_0 \neq \mathbf{0}_{n\times n}$. Conversely, if L_0 is dense and contains infinitely many small entries, its incoherence measure tends toward zero. For instance, if L_0 possesses a singular vector with entries of order $1/\sqrt{n}$, the incoherence measure attains an order of $1/\sqrt{n}$, which is the smallest order achievable. In such a case, it could be separated from S_0 if $\deg_{\max}(S_0) = o(\sqrt{n})$, which includes the sparse network with bounded degree, or the dominant network with finite keyplayers. It also implies that $m_s(S_0) = o(n^{\frac{3}{2}})$ as $m_r(S_0) \vee m_c(S_0) = o(\sqrt{n})$.

D.2. **Examples.** We now discuss how the identification assumption applies to the four examples of network structures mentioned in Section 2.

Complete network: In this example, $W_0 = L_0 + S_0$, where $L_0 = \boldsymbol{c}\boldsymbol{c}^\top$, $\boldsymbol{c} = (c_1, \cdots, c_n)^\top$, and the sparse component S_0 is a diagonal matrix to guarantee that the diagonal elements of W_0 are zero. Note that $\deg_{\max}(S_0) = 1$ and $\mathrm{inc}(L_0) = \sqrt{\max_{1\leq i\leq n} c_i^2/\|\boldsymbol{c}\|_2^2} \leq c_{\max}/(\sqrt{n}\,c_{\min})$.

[12] For $L_0 = UDV^\top$, $L_{0,ij} = (U^\top e_i)^\top D(V^\top e_j)$, so by Cauchy–Schwarz $|L_{0,ij}| \leq \sigma_1(L_0)\|U^\top e_i\|_2\|V^\top e_j\|_2 \leq \sigma_1(L_0)\,\mathrm{inc}(L_0)^2$, using $\|U^\top e_i\|_2^2 = \sum_{k=1}^r u_{ik}^2 \leq \mu_u(L_0)$.

[13] Uniqueness here means that there exists an optimization scheme that pins down the pair (L_0, S_0) uniquely; see Chandrasekaran et al. (2009). Note that the identification is local, in the sense that what we can identify is a discrete set. Local perturbations of the solution can also be identified; however, at the global level, multiple identified sets may exist.

Hence, a sufficient condition for the identification condition $\mathrm{inc}(L_0)\deg_{\max}(S_0) \le 1/16$ of Assumption 9 is $c_{\max}/c_{\min} \le \sqrt{n}/16$.

Low-rank network: In this example, $W_0 = L_0 + S_0$, where $L_0 = \boldsymbol{a}\mathbf{1}_n^\top + \boldsymbol{z}\boldsymbol{z}^\top$ is a matrix of rank no more than 2 with $\boldsymbol{a} = (a_1, \cdots, a_n)$ and $\boldsymbol{z} = (z_1, \cdots, z_n)$, and S_0 is a diagonal matrix whose (i,i)th element is $-L_{0,ii}$ to remove self-loops. For simplicity, we assume $\boldsymbol{z}$ is orthogonal to both $\boldsymbol{a}$ and $\mathbf{1}_n$, and hence the left singular vectors are proportional to $\boldsymbol{a}$ and $\boldsymbol{z}$, while the right singular vectors are proportional to $\mathbf{1}_n$ and $\boldsymbol{z}$. Since $\deg_{\max}(S_0) = 1$, the identification assumption could be satisfied if $\mathrm{inc}(L_0) = \sqrt{\max\left\{\max_{1\le i\le n}\left(\frac{a_i^2}{\|\boldsymbol{a}\|_2^2} + \frac{z_i^2}{\|\boldsymbol{z}\|_2^2}\right),\ \max_{1\le i\le n}\left(\frac{1}{n} + \frac{z_i^2}{\|\boldsymbol{z}\|_2^2}\right)\right\}} \le 1/16$.

Group network: In this example, $W_0 = L_0 + S_0$, where $L_0 = \boldsymbol{c}_a\boldsymbol{c}_a^\top + \boldsymbol{c}_b\boldsymbol{c}_b^\top$, $\boldsymbol{c}_a = (c_1, \cdots, c_6, 0, \cdots, 0)^\top$, and $\boldsymbol{c}_b = (0, \cdots, 0, c_7, \cdots, c_{20})^\top$, and S_0 is a diagonal matrix to ensure that the diagonal elements of W_0 are 0. Since $\deg_{\max}(S_0) = 1$, the identification assumption could be satisfied if $\mathrm{inc}(L_0) = \sqrt{\max(\max_{1\le i\le 6} c_i^2/\sum_{i=1}^6 c_i^2, \max_{7\le i\le 20} c_i^2/\sum_{i=7}^{20} c_i^2)} \le 1/16$. In the general case when S_0 has nonzero off-diagonal elements, the identification assumption then requires that $\deg_{\max}(S_0)\mathrm{inc}(L_0) = \deg_{\max}(S_0)\sqrt{\max(\max_{1\le i\le 6} c_i^2/\sum_{i=1}^6 c_i^2, \max_{7\le i\le 20} c_i^2/\sum_{i=7}^{20} c_i^2)} \le 1/16$.

Dominant units: In this example,

$$W_0 = \begin{bmatrix} 0 & w_{1,2} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ w_{2,1} & 0 & w_{2,3} & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ w_{14,1} & 0 & \cdots & w_{14,13} & 0 & w_{14,15} & \cdots & 0 & 0 \\ 0 & \cdots & \cdots & 0 & w_{15,14} & 0 & w_{15,16} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & w_{20,19} & 0 \end{bmatrix}.$$

When we treat it as a pure sparse matrix, the identification condition is automatically satisfied as $\mathrm{inc}(L_0) = 0$. Though it seems that one could write $W_0 = \mathbf{w}_1\boldsymbol{e}_1^\top + W_r$, where $\mathbf{w}_1$ is the first column of W_0, $\boldsymbol{e}_1$ is the first column of the identity matrix I_n, and W_r is the $n \times n$ partitioned matrix satisfying

$$\begin{bmatrix} 0 & \boldsymbol{s}_1^\top \\ \mathbf{0}_{(n-1)\times 1} & \boldsymbol{S}_r \end{bmatrix}$$

with $\boldsymbol{s}_1 = [w_{1,2}, \mathbf{0}_{1\times(n-2)}]^\top$, and the (i,j)th element in $\boldsymbol{S}_r$ is identical to the $(i+1, j+1)$th element of W_0. In such a decomposition, the right singular vector of the low-rank part is $[1, \mathbf{0}_{1\times(n-1)}]^\top$ and its incoherence index is $\mathrm{inc}(\boldsymbol{w}_1\boldsymbol{e}_1^\top) = 1$. Since $\deg_{\max}(W_r) = 2$, we have $\mathrm{inc}(\boldsymbol{w}_1\boldsymbol{e}_1^\top)\deg_{\max}(W_r) > 1/16$. Therefore, we have to treat W_0 as a pure sparse matrix so that the identification Assumption holds.

E. Proofs of main results

E.1. **Notation.** $\|A\|_1 = \max_{1\le j\le n}\sum_{i=1}^n |A_{ij}|$, $\|A\|_2 = \sigma_1(A)$, $\|A\|_\infty = \max_{1\le i\le n}\sum_{j=1}^n |A_{ij}|$, $\|A\|_{\max} = \max_{1\le i,j\le n}|A_{i,j}|$, $\|A\|_F = \sqrt{\sum A_{ij}^2}$, and $\|A\|_* = \sum_{i=1}^n \sigma_i(A)$. Define $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$ as the Frobenius inner product between A and B.

E.2. **Proof of Perturbation Lemma.** Justification of Equation (C.1):
Assumption (two-dimensional spectral concentration and separation). For each n, let L_0^* and S_0^* be real symmetric matrices with normalized leading eigenvectors $\mathbf{v}_{L_0^*}$ and $\mathbf{v}_{S_0^*}$ and corresponding simple leading eigenvalues $\lambda_{L_0^*}$ and $\lambda_{S_0^*}$. Let

$$\rho_n \coloneqq \mathbf{v}_{L_0^*}^\top \mathbf{v}_{S_0^*}, \qquad |\rho_n| < 1, \qquad \rho_n \to 0.$$

Let $\mathbf{v}_{W_0}$ be the normalized leading eigenvector of $W_0 = L_0^* + S_0^*$ and write

$$\mathbf{v}_{W_0} = k_1\mathbf{v}_{L_0^*} + k_2\mathbf{v}_{S_0^*} + k_3 e, \qquad e \perp \mathrm{span}\{\mathbf{v}_{L_0^*}, \mathbf{v}_{S_0^*}\}, \qquad \|e\|_2 = 1.$$

Assume

$$\|L_0^*\|_2 + \|S_0^*\|_2 \le C, \qquad |k_1| + |k_2| \asymp 1, \qquad |k_3| = o(|\rho_n|).$$

Define

$$\widetilde{a}_n \coloneqq \lambda_{L_0^*} - \mathbf{v}_{S_0^*}^\top L_0^* \mathbf{v}_{S_0^*}, \qquad \widetilde{c}_n \coloneqq \lambda_{S_0^*} - \mathbf{v}_{L_0^*}^\top S_0^* \mathbf{v}_{L_0^*}.$$

Assume that, for some constants $g_0, \kappa > 0$, one of the following two cases holds eventually:

(i) $\widetilde{a}_n - \widetilde{c}_n \ge g_0$ and $|k_1| \ge \kappa$;

(ii) $\widetilde{c}_n - \widetilde{a}_n \ge g_0$ and $|k_2| \ge \kappa$.

Lemma 2 (Perturbation of the leading eigenvector)**.** *Under the two-dimensional spectral concentration and separation conditions stated above, Equations* (C.1)*,* (C.2) *and C.3 hold; in case (i),* $|\mathbf{v}_{W_0}^\top \mathbf{v}_{L_0^*}| > |\mathbf{v}_{W_0}^\top \mathbf{v}_{S_0^*}|$ *for all sufficiently large n, and in case (ii) the reverse inequality holds.*

Proof. Define

$$\rho_{W_0 L_0^*} \coloneqq \mathbf{v}_{W_0}^\top \mathbf{v}_{L_0^*} = k_1 + k_2\rho_n, \qquad \rho_{W_0 S_0^*} \coloneqq \mathbf{v}_{W_0}^\top \mathbf{v}_{S_0^*} = k_1\rho_n + k_2.$$

Note that

$$\rho_n = \rho_{L_0^* S_0^*} = \mathbf{v}_{L_0^*}^\top \mathbf{v}_{S_0^*}.$$

Moreover, $\mathbf{v}_{L_0^*}$ and $\mathbf{v}_{S_0^*}$ are normalized eigenvectors. Hence

$$\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} 1 & \rho_n \\ \rho_n & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho_{W_0 L_0^*} \\ \rho_{W_0 S_0^*} \end{pmatrix} = \begin{pmatrix} \dfrac{\rho_{W_0 L_0^*} - \rho_n \rho_{W_0 S_0^*}}{1-\rho_n^2} \\ \dfrac{\rho_{W_0 S_0^*} - \rho_n \rho_{W_0 L_0^*}}{1-\rho_n^2} \end{pmatrix},$$

and Equation (C.1) holds.

We now prove Equation (C.2) in the main text.

Recall the decomposition

$$\mathbf{v}_{W_0} = k_1 \mathbf{v}_{L_0^*} + k_2 \mathbf{v}_{S_0^*} + k_3 e,$$

where $\|e\|_2 = 1$ and e is orthogonal to both $\mathbf{v}_{L_0^*}$ and $\mathbf{v}_{S_0^*}$. Then

$$(L_0^* + S_0^*)\mathbf{v}_{W_0} = k_1 \lambda_{L_0^*} \mathbf{v}_{L_0^*} + k_2 L_0^* \mathbf{v}_{S_0^*} + k_1 S_0^* \mathbf{v}_{L_0^*} + k_2 \lambda_{S_0^*} \mathbf{v}_{S_0^*} + r_n, \tag{E.1}$$

where

$$r_n \coloneqq k_3 (L_0^* + S_0^*) e.$$

Because $\|L_0^*\|_2 + \|S_0^*\|_2 \leq C$,

$$\|r_n\|_2 \leq C|k_3|.$$

Let

$$r_{Ln} \coloneqq \mathbf{v}_{L_0^*}^\top r_n, \qquad r_{Sn} \coloneqq \mathbf{v}_{S_0^*}^\top r_n.$$

Then

$$|r_{Ln}| + |r_{Sn}| = O(|k_3|).$$

Multiplying Equation (E.1) by $\mathbf{v}_{L_0^*}^\top$ and using the symmetry of L_0^* gives

$$\begin{aligned} k_1 \lambda_{L_0^*} + k_2 \lambda_{L_0^*} \rho_n + k_1 \mathbf{v}_{L_0^*}^\top S_0^* \mathbf{v}_{L_0^*} + k_2 \lambda_{S_0^*} \rho_n + r_{Ln} \\ = \lambda_{W_0} (k_1 + k_2 \rho_n). \end{aligned}$$

Similarly, multiplying by $\mathbf{v}_{S_0^*}^\top$ and using the symmetry of S_0^* gives

$$\begin{aligned} k_1 \lambda_{L_0^*} \rho_n + k_2 \mathbf{v}_{S_0^*}^\top L_0^* \mathbf{v}_{S_0^*} + k_1 \lambda_{S_0^*} \rho_n + k_2 \lambda_{S_0^*} + r_{Sn} \\ = \lambda_{W_0} (k_1 \rho_n + k_2). \end{aligned}$$

Recall that

$$\widetilde{a}_n = \lambda_{L_0^*} - \mathbf{v}_{S_0^*}^\top L_0^* \mathbf{v}_{S_0^*}, \qquad \widetilde{c}_n = \lambda_{S_0^*} - \mathbf{v}_{L_0^*}^\top S_0^* \mathbf{v}_{L_0^*}.$$

Since $\lambda_{L_0^*}$ and $\lambda_{S_0^*}$ are the largest eigenvalues of the symmetric matrices L_0^* and S_0^*, respectively,

$$\widetilde{a}_n \geq 0, \qquad \widetilde{c}_n \geq 0.$$

Consider case (i), so that

$$\widetilde{a}_n - \widetilde{c}_n \geq g_0 > 0, \qquad |k_1| \geq \kappa > 0.$$

Set

$$q_n \coloneqq \frac{k_2}{k_1}.$$

Cross-multiplying the two projected eigenvalue equations and retaining the remainder gives

$$\widetilde{a}_n \rho_n q_n^2 + (\widetilde{a}_n - \widetilde{c}_n) q_n - \widetilde{c}_n \rho_n = O(|k_3|).$$

Let

$$g_n \coloneqq \widetilde{a}_n - \widetilde{c}_n.$$

The sequence q_n is bounded because $|k_1| \geq \kappa$ and $|k_1| + |k_2| \asymp 1$. Moreover,

$$2\widetilde{a}_n \rho_n q_n + g_n = g_n + o(1) \geq \frac{g_0}{2}$$

for all sufficiently large n. Hence the $O(|k_3|)$ perturbation of the quadratic produces an $O(|k_3|)$ perturbation of its bounded root. Therefore,

$$\begin{aligned}\frac{k_2}{k_1} &= \frac{2\widetilde{c}_n\rho_n}{g_n + \sqrt{g_n^2 + 4\widetilde{a}_n\widetilde{c}_n\rho_n^2}} + O(|k_3|)\\ &= \frac{\widetilde{c}_n\rho_n}{g_n} + O(\rho_n^3) + O(|k_3|)\\ &= \frac{(\lambda_{S_0^*} - \mathbf{v}_{L_0^*}^\top S_0^* \mathbf{v}_{L_0^*})\rho_n}{\lambda_{L_0^*} - \mathbf{v}_{S_0^*}^\top L_0^* \mathbf{v}_{S_0^*} - (\lambda_{S_0^*} - \mathbf{v}_{L_0^*}^\top S_0^* \mathbf{v}_{L_0^*})} + o(|\rho_n|),\end{aligned}$$

where the last equality uses $\rho_n \to 0$ and $|k_3| = o(|\rho_n|)$.

Since $q_n = O(\rho_n)$,

$$|\mathbf{v}_{W_0}^\top \mathbf{v}_{L_0^*}| = |k_1 + k_2\rho_n| = |k_1|\{1 + o(1)\},$$

whereas

$$|\mathbf{v}_{W_0}^\top \mathbf{v}_{S_0^*}| = |k_1\rho_n + k_2| = O(|\rho_n|).$$

Thus, for all sufficiently large n,

$$|\mathbf{v}_{W_0}^\top \mathbf{v}_{L_0^*}| > |\mathbf{v}_{W_0}^\top \mathbf{v}_{S_0^*}|.$$

Now consider case (ii), so that

$$\widetilde{c}_n - \widetilde{a}_n \geq g_0 > 0, \qquad |k_2| \geq \kappa > 0.$$

Set

$$p_n \coloneqq \frac{k_1}{k_2}.$$

Cross-multiplying the projected eigenvalue equations gives

$$\widetilde{c}_n\rho_n p_n^2 + (\widetilde{c}_n - \widetilde{a}_n)p_n - \widetilde{a}_n\rho_n = O(|k_3|).$$

Let

$$h_n \coloneqq \widetilde{c}_n - \widetilde{a}_n.$$

As in case (i), p_n is bounded and

$$2\widetilde{c}_n\rho_n p_n + h_n = h_n + o(1) \geq \frac{g_0}{2}$$

for all sufficiently large n. Therefore,

$$\begin{aligned}\frac{k_1}{k_2} &= \frac{2\widetilde{a}_n\rho_n}{h_n + \sqrt{h_n^2 + 4\widetilde{a}_n\widetilde{c}_n\rho_n^2}} + O(|k_3|)\\ &= \frac{\widetilde{a}_n\rho_n}{\widetilde{c}_n - \widetilde{a}_n} + O(\rho_n^3) + O(|k_3|)\\ &= \frac{\widetilde{a}_n\rho_n}{\widetilde{c}_n - \widetilde{a}_n} + o(|\rho_n|).\end{aligned}$$

Since $p_n = O(\rho_n)$,

$$|\mathbf{v}_{W_0}^\top \mathbf{v}_{S_0^*}| = |k_2 + k_1\rho_n| = |k_2|\{1 + o(1)\},$$

whereas

$$|\mathbf{v}_{W_0}^\top \mathbf{v}_{L_0^*}| = |k_2\rho_n + k_1| = O(|\rho_n|).$$

Consequently, for all sufficiently large n,

$$|\mathbf{v}_{W_0}^\top \mathbf{v}_{S_0^*}| > |\mathbf{v}_{W_0}^\top \mathbf{v}_{L_0^*}|.$$

□

Remark 2. The preceding perturbation result applies to symmetric matrices L_0^* and S_0^*. When S_0^* is asymmetric, the argument must be reformulated using the corresponding left and right eigenvectors. We leave this extension for future work.

E.3. **Some Useful Lemmas.** We first present some useful lemmas for the proof of the theorem. The notation within this section does not necessary align the main text and the proofs of the theorem since they are independent results. Assume that the rank of the matrix A is r.

To summarise the steps we shall first give a few definitions. Let $L_0 \in \mathbb{R}^{n\times n}$ be a low rank matrix with rank r. Let $\mathcal{T}(L_0)$ denote the span of all matrices whose row space is contained in the row space of L_0 or whose column space is contained in the column space of L_0. Writing the singular value decomposition $L_0 = U\Sigma V^\top$ with $U, V \in \mathbb{R}^{n\times r}$ and $\text{rank}(L_0) = r$, this space is

$$\mathcal{T}(L_0) = \left\{UX^\top + YV^\top \mid X, Y \in \mathbb{R}^{n\times r}\right\}.$$

Analogously to $\deg_{\max}(A)$, define $\deg_{\min}(A) := \min\Big\{\min_{1\le i\le n}\sum_{j=1}^n \mathbb{I}(A_{ij}\neq 0),\ \min_{1\le j\le n}\sum_{i=1}^n \mathbb{I}(A_{ij}\neq 0)\Big\}$, the minimum number of nonzero entries in any row or column of A. For any matrix $S_0 \in \mathbb{R}^{n\times n}$, the space $\Omega(S_0)$ is defined by,

$$\Omega(S_0) = \left\{N \in \mathbb{R}^{n\times n} \mid \text{support}(N) \subseteq \text{support}(S_0)\right\}.$$

Given $\mathcal{T}(L_0)$ and $\Omega(S_0)$, we define two measures that quantify the denseness and sparsity levels of a matrix. $\xi(L_0) = \max_{N\in\mathcal{T}(L_0),\|N\|_2\le 1}\|N\|_{\max}$, $\mu(S_0) = \max_{N\in\Omega(S_0),\|N\|_{\max}\le 1}\|N\|_2$.

Lemma 3. *(Chandrasekaran et al. (2009)) Let $A \in \mathbb{R}^{n\times n}$ be any matrix. We have that*

$$\text{inc}(A) \le \xi(A) \le 2\,\text{inc}(A),$$

$$\deg_{\min}(A) \le \mu(A) \le \deg_{\max}(A),$$

$$\xi(A)\mu(A) \ge 1.$$

Let Ω denote the parameter space for θ, and let $\mathcal{L}$ denote the population cost function. Define the target parameter by

$$\theta^* \in \arg\min_{\theta\in\Omega}\mathcal{L}(\theta).$$

Based on the observed sample

$$Z_1^n = \{Z_1, \ldots, Z_n\},$$

we estimate θ^* by solving

$$\widehat{\theta} \in \arg\min_{\theta\in\Omega} \left\{\mathcal{L}_n(\theta; Z_1^n) + \lambda_n \Phi(\theta)\right\},$$

where $\mathcal{L}_n$ is the empirical cost function, $\lambda_n > 0$ is a regularization parameter, and $\Phi : \Omega \to \mathbb{R}_+$ is a norm. Let Φ^* denote the dual norm of Φ.

Let $\mathbb{M}$ and $\overline{\mathbb{M}}$ be subspaces satisfying

$$\mathbb{M} \subseteq \overline{\mathbb{M}},$$

and denote their orthogonal complements by $\mathbb{M}^\perp$ and $\overline{\mathbb{M}}^\perp$, respectively.

For any subspace $\mathbb{S}$, let $\theta_{\mathbb{S}}$ denote the orthogonal projection of θ onto $\mathbb{S}$. In particular, $\theta^*_{\mathbb{M}^\perp}$ denotes the component of θ^* in $\mathbb{M}^\perp$.

Our bound involves the quantity

$$\varepsilon_n^2\left(\overline{\mathbb{M}}, \mathbb{M}^\perp\right) := \underbrace{9\frac{\lambda_n^2}{\kappa^2}\Psi^2\left(\overline{\mathbb{M}}\right)}_{\text{estimation error}} + \underbrace{\frac{8}{\kappa}\left\{\lambda_n \Phi\left(\theta^*_{\mathbb{M}^\perp}\right) + 16\tau_n^2 \Phi^2\left(\theta^*_{\mathbb{M}^\perp}\right)\right\}}_{\text{approximation error}},$$

which depends on the choice of the subspace pair $\left(\overline{\mathbb{M}}, \mathbb{M}^\perp\right)$.

Assumption 10. i) The cost function $\mathcal{L}_n$ is convex and satisfies the local restricted strong convexity condition with curvature $\kappa > 0$, radius $R > 0$, and tolerance $\tau_n^2 \geq 0$. That is,

$$\mathcal{E}_n(\Delta) := \mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \langle \nabla \mathcal{L}_n(\theta^*), \Delta\rangle \geq \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2 \Phi^2(\Delta)$$

for every

$$\Delta \in \mathbb{B}(R) := \{\Delta : \theta^* + \Delta \in \Omega,\ \|\Delta\| \leq R\}.$$

ii) There exists a pair of subspaces

$$\left(\mathbb{M}, \overline{\mathbb{M}}^\perp\right), \qquad \mathbb{M} \subseteq \overline{\mathbb{M}},$$

such that the regularizer Φ is decomposable with respect to $\left(\mathbb{M}, \overline{\mathbb{M}}^\perp\right)$; that is,

$$\Phi(u + v) = \Phi(u) + \Phi(v)$$

for every $u \in \mathbb{M}$ and $v \in \overline{\mathbb{M}}^\perp$.

iii) Define the good event

$$\mathbb{G}(\lambda_n) := \left\{\Phi^*\left(\nabla \mathcal{L}_n(\theta^*)\right) \leq \frac{\lambda_n}{2}\right\}, \qquad \lambda_n \geq 0.$$

iv) For a given norm $\|\cdot\|$ and subspace $\mathbb{S}$, define the subspace Lipschitz constant by

$$\Psi(\mathbb{S}) \coloneqq \sup_{u \in \mathbb{S}\setminus\{\mathbf{0}\}} \frac{\Phi(u)}{\|u\|}.$$

The following are two important lemmas that we use to prove Theorem 1.

Lemma 4. (Theorem 9.19 in Wainwright (2019)). *Suppose Assumption 10 holds. Conditional on the event $\mathbb{G}(\lambda_n)$, the following statements hold:*

(a) Any optimal solution satisfies

$$\Phi\left(\widehat{\theta} - \theta^*\right) \leq 4\left\{\Psi\left(\overline{\mathbb{M}}\right)\left\|\widehat{\theta} - \theta^*\right\|_2 + \Phi\left(\theta^*_{\mathbb{M}^\perp}\right)\right\}.$$

(b) For any subspace pair $\left(\overline{\mathbb{M}}, \mathbb{M}^\perp\right)$ satisfying

$$\tau_n^2 \Psi^2\left(\overline{\mathbb{M}}\right) \leq \frac{\kappa}{64} \quad \textit{and} \quad \varepsilon_n\left(\overline{\mathbb{M}}, \mathbb{M}^\perp\right) \leq R,$$

we have

$$\left\|\widehat{\theta} - \theta^*\right\|_2^2 \leq \varepsilon_n^2\left(\overline{\mathbb{M}}, \mathbb{M}^\perp\right).$$

Assumption 11. The cost function satisfies the Φ^*-norm curvature condition with curvature $\kappa_\Phi > 0$, tolerance $\tau_{n,\Phi} \geq 0$, and radius $R_\Phi > 0$:

$$\Phi^*\left(\nabla\mathcal{L}_n(\theta^* + \Delta) - \nabla\mathcal{L}_n(\theta^*)\right) \geq \kappa_\Phi \Phi^*(\Delta) - \tau_{n,\Phi}\Phi(\Delta)$$

for every

$$\Delta \in \mathbb{B}_{\Phi^*}(R_\Phi) \coloneqq \left\{\Delta \in \Omega : \Phi^*(\Delta) \leq R_\Phi\right\}.$$

Lemma 5 (Theorem 9.24 in Wainwright (2019))**.** *Suppose $\theta^* \in \mathbb{M}$, the decomposability condition in Assumption 10(ii) holds, and Assumption 11 holds. Suppose also that*

$$\tau_{n,\Phi}\Psi^2(\overline{\mathbb{M}}) < \frac{\kappa_\Phi}{32}.$$

Conditional on

$$\mathbb{G}(\lambda_n) \cap \left\{\Phi^*(\widehat{\theta} - \theta^*) \leq R_\Phi\right\},$$

any optimal solution satisfies

$$\Phi^*(\widehat{\theta} - \theta^*) \leq \frac{3\lambda_n}{\kappa_\Phi}.$$

Lemma 6 (Lindeberg–Feller central limit theorem)**.** *For each n, let $Y_{n,1}, \ldots, Y_{n,k_n}$ be independent, mean-zero random vectors with finite covariance matrices. Suppose that, for every $\varepsilon > 0$,*

$$\sum_{i=1}^{k_n} \mathbb{E}\left[\|Y_{n,i}\|_2^2 \mathbf{1}\{\|Y_{n,i}\|_2 > \varepsilon\}\right] \longrightarrow 0$$

and

$$\sum_{i=1}^{k_n} \mathrm{Cov}(Y_{n,i}) \longrightarrow \Sigma.$$

Then

$$\sum_{i=1}^{k_n} Y_{n,i} \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Lemma 7. *Let $A, B \in \mathbb{R}^{n\times n}$. Let $\mathcal{C}_1 \subseteq \{1, \ldots, n\}$ satisfy $|\mathcal{C}_1| \leq K$, where K is fixed, and define*

$$\mathcal{C}_2 = \{1, \ldots, n\} \setminus \mathcal{C}_1.$$

Let

$$c_n = \max_{k\in\mathcal{C}_1} \sum_{j=1}^{n} |B_{kj}|, \qquad c = \max_{k\in\mathcal{C}_2} \sum_{j=1}^{n} |B_{kj}|.$$

Then

$$\|AB\|_\infty \leq K\|A\|_{\max} c_n + \|A\|_\infty c.$$

Here, the maximum over an empty set is defined to be zero.

Lemma 8. *Suppose Assumptions 1 and 4(iii) hold, J is fixed, and*

$$\|W_0\|_1 = O(\sqrt{n}).$$

Assume that there exist constants $c_\infty, c_1 > 0$, independent of n, such that

$$\sup_{\lambda\in\mathcal{C}} |\lambda| \|W_0\|_\infty \leq 1 - c_\infty, \qquad \sup_{\lambda\in\mathcal{C}} |\lambda| \|W_{0,b}\|_1 \leq 1 - c_1.$$

Then, uniformly over $\lambda \in \mathcal{C}$,

$$\left\|(I_n - \lambda W_0)^{-1}\right\|_2 = O(n^{1/4}).$$

Lemma 9. *Suppose Assumptions 4(iii) and 6(i) hold. Let X_{jt} be the jth column of X_t. Then, for $\ell = 0, 1, 2$,*

$$\|W_0^\ell X_{jt}\|_2 = O_p(\sqrt{n}), \qquad \left\|W_0^\ell (I_n - \lambda_0 W_0)^{-1} X_{jt}\right\|_2 = O_p(\sqrt{n}).$$

E.4. **Proof of Lemma 1, 7-9.**

Proof. It follows from Corollary 3 in Chandrasekaran et al. (2009).

$\mu(S_0)$ can be bounded by $\deg_{\max}(S_0)$ and $\xi(L_0)$ can be bounded by its incoherence measure $2\,\mathrm{inc}(L_0)$. The identification condition is $\xi(L_0)\mu(S_0) < 1/8$. A stronger condition can be $\deg_{\max}(S_0)\,\mathrm{inc}(L_0) \leq 1/16$, due to Lemma 3. □

Proof of Lemma 7.

Proof. Recall that

$$\|AB\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |(AB)_{ij}|, \qquad \|A\|_\infty = \max_{1\leq i\leq n} \sum_{k=1}^{n} |A_{ik}|,$$

and

$$\|A\|_{\max} = \max_{1\leq i,k\leq n} |A_{ik}|.$$

For every i,

$$\begin{aligned}\sum_j |(AB)_{ij}| &\le \sum_j \sum_{k\in\mathcal{C}_1} |A_{ik}B_{kj}| + \sum_j \sum_{k\in\mathcal{C}_2} |A_{ik}B_{kj}| \\ &\le \max_{k\in\mathcal{C}_1} |A_{ik}| \sum_{k\in\mathcal{C}_1}\sum_j |B_{kj}| + \sum_{k\in\mathcal{C}_2} |A_{ik}| \max_{k\in\mathcal{C}_2} \sum_j |B_{kj}| \\ &\le K\|A\|_{\max} c_n + \|A\|_\infty c.\end{aligned}$$

Taking the maximum over i gives

$$\|AB\|_\infty \le K\|A\|_{\max} c_n + \|A\|_\infty c,$$

which is the stated result. □

Proof of Lemma 8.

Proof. Let

$$B_n(\lambda) = (I_n - \lambda W_{0,b})^{-1}.$$

Since zeroing columns cannot increase the maximum absolute row sum,

$$\|W_{0,b}\|_\infty \le \|W_0\|_\infty.$$

Therefore, the two stability conditions and the Neumann-series argument give

$$\sup_{\lambda\in\mathcal{C}} \|B_n(\lambda)\|_1 \le c_1^{-1}, \qquad \sup_{\lambda\in\mathcal{C}} \|B_n(\lambda)\|_\infty \le c_\infty^{-1}.$$

Consequently,

$$\sup_{\lambda\in\mathcal{C}} \|B_n(\lambda)\|_2 \le \sup_{\lambda\in\mathcal{C}} \sqrt{\|B_n(\lambda)\|_1 \|B_n(\lambda)\|_\infty} = O(1).$$

If $J = 0$, then $W_{0,u} = 0$ and the result follows immediately. Hence, suppose $J \ge 1$. Let

$$\Pi_J = (e_1, \dots, e_J) \in \mathbb{R}^{n\times J}$$

be the matrix selecting the first J coordinates, and define

$$A_J := W_{0,u}\Pi_J \in \mathbb{R}^{n\times J}.$$

Since $W_{0,u}$ vanishes outside its first J columns,

$$W_{0,u} = A_J \Pi_J^\top.$$

Set

$$U = B_n(\lambda)A_J, \qquad V = \Pi_J, \qquad Q_n(\lambda) = I_n - \lambda UV^\top.$$

Then

$$I_n - \lambda W_0 = (I_n - \lambda W_{0,b})Q_n(\lambda).$$

Let

$$R_n(\lambda) = (I_n - \lambda W_0)^{-1}.$$

The first stability condition and the Neumann-series argument give

$$\sup_{\lambda\in\mathcal{C}} \|R_n(\lambda)\|_\infty \le c_\infty^{-1}.$$

Both $I_n - \lambda W_0$ and $I_n - \lambda W_{0,b}$ are invertible. Hence, $Q_n(\lambda)$ is invertible and

$$Q_n(\lambda)^{-1} = R_n(\lambda)(I_n - \lambda W_{0,b}).$$

Therefore,

$$\begin{aligned}\sup_{\lambda\in\mathcal{C}} \|Q_n(\lambda)^{-1}\|_\infty &\le \sup_{\lambda\in\mathcal{C}} \|R_n(\lambda)\|_\infty \big(1 + |\lambda|\|W_{0,b}\|_\infty\big) \\ &\le \frac{2 - c_\infty}{c_\infty} = O(1).\end{aligned}$$

Define

$$M_J(\lambda) = I_J - \lambda V^\top U.$$

Since

$$V^\top Q_n(\lambda) = M_J(\lambda) V^\top,$$

right-multiplication by $Q_n(\lambda)^{-1}V$ gives

$$M_J(\lambda)\big[V^\top Q_n(\lambda)^{-1} V\big] = I_J.$$

Thus, $M_J(\lambda)$ is invertible and

$$M_J(\lambda)^{-1} = V^\top Q_n(\lambda)^{-1} V.$$

Because J is fixed,

$$\sup_{\lambda\in\mathcal{C}} \|M_J(\lambda)^{-1}\|_2 \le J \sup_{\lambda\in\mathcal{C}} \|Q_n(\lambda)^{-1}\|_\infty = O(1).$$

The Woodbury formula may therefore be applied:

$$Q_n(\lambda)^{-1} = I_n + \lambda U M_J(\lambda)^{-1} V^\top.$$

Since $\mathcal{C}$ is compact, $\sup_{\lambda\in\mathcal{C}} |\lambda| < \infty$. Moreover,

$$\begin{aligned}\|U\|_2 \|V^\top\|_2 &\le \|B_n(\lambda)\|_2 \|W_{0,u}\|_2 \\ &\le O(1)\sqrt{\|W_{0,u}\|_1 \|W_{0,u}\|_\infty} \\ &= O(n^{1/4}),\end{aligned}$$

because

$$\|W_{0,u}\|_1 \le \|W_0\|_1 = O(\sqrt{n}), \qquad \|W_{0,u}\|_\infty \le \|W_0\|_\infty = O(1).$$

Therefore, the Woodbury representation gives

$$\sup_{\lambda\in\mathcal{C}} \|Q_n(\lambda)^{-1}\|_2 = O(n^{1/4}).$$

Finally,

$$R_n(\lambda) = Q_n(\lambda)^{-1} B_n(\lambda),$$

and hence

$$\sup_{\lambda \in \mathcal{C}} \|(I_n - \lambda W_0)^{-1}\|_2 = O(n^{1/4}),$$

as required. □

Proof of Lemma 9.

Proof. Fix j and write $X = X_{jt}$. For $\ell = 0, 1, 2$, Assumption 4(iii) gives

$$\|W_0^\ell\|_\infty \leq \|W_0\|_\infty^\ell = O(1).$$

Let $\mu = \mathbb{E}X$. By Assumption 6(i), the components of X are independent across i and have means and variances that are uniformly bounded in n, t, i, and j. Hence,

$$\mathbb{E}\|W_0^\ell (X - \mu)\|_2^2 \leq Cn\|W_0^\ell\|_\infty^2 = O(n).$$

Moreover,

$$\|W_0^\ell \mu\|_2^2 \leq n\|W_0^\ell\|_\infty^2 \|\mu\|_{\max}^2 = O(n).$$

It follows that

$$\mathbb{E}\|W_0^\ell X\|_2^2 = O(n),$$

and therefore, by Markov's inequality,

$$\|W_0^\ell X\|_2 = O_p(\sqrt{n}).$$

Let

$$R_{0n} = (I_n - \lambda_0 W_0)^{-1}.$$

By the Neumann-series bound under Assumption 4(iii),

$$\|R_{0n}\|_\infty \leq c_\infty^{-1}.$$

Consequently,

$$\|W_0^\ell R_{0n}\|_\infty = O(1).$$

Applying the same second-moment argument with $A_n = W_0^\ell R_{0n}$ gives

$$\mathbb{E}\|W_0^\ell R_{0n} X\|_2^2 = O(n).$$

Hence, by Markov's inequality,

$$\|W_0^\ell (I_n - \lambda_0 W_0)^{-1} X\|_2 = O_p(\sqrt{n}),$$

for $\ell = 0, 1, 2$. □

E.5. **Proof of Main Theorems.**

Remark 3 (Proof convention: within transformation and two-way fixed effects). Throughout the proofs that follow, Y_t, X_t, $Z_t(M)$, and ε_t denote the within-transformed quantities obtained by removing the individual fixed effects α from (3).

The extension to specifications containing time fixed effects $\iota_t \mathbf{1}_n$ uses the two-way within transformation:

$$\ddot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot t} + \bar{\varepsilon}_{\cdot\cdot}.$$

Under Assumption 5, for $i \neq j$,

$$\operatorname{Cov}\left(\ddot{\varepsilon}_{it}, \ddot{\varepsilon}_{jt}\right) = -\frac{\sigma_0^2}{n}\left(1 - \frac{1}{T}\right).$$

Let

$$M_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top, \qquad M_T = I_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^\top.$$

Because the errors, instruments, and regressors are transformed using the same symmetric and idempotent projection matrices,

$$\sum_{t=1}^{T} \ddot{Z}_t(M)^\top \ddot{\varepsilon}_t = \sum_{t=1}^{T} \ddot{Z}_t(M)^\top \varepsilon_t.$$

Under the strict-exogeneity condition imposed in the main text, these moments have conditional mean zero. Consequently, the same rate, variance, and central-limit-theorem arguments continue to apply, with the instruments and regressors interpreted as two-way within transformed.

Proof of Theorem 1. First, note that the rate of the estimator can diminish with respect to large n if both $w_{2n}, w_{\max n} \to 0$. This theorem indicates that $\|\widehat{W} - W_0\|_F^2$ is of the order $\nu_n^2 r + \tau_n^2 n$, which is a direct result from Theorem 1 in Agarwal et al. (2012). Hence, the proof boils down to formally verifying the following assumptions (note that the italic text maintain the same notations in Agarwal et al. (2012)).
Verification of conditions in Theorem 1:

i) For RSC, it trivially holds, as the operator $\mathcal{X}(.)$ becomes the identity operator. Therefore in our case $\gamma = 1$ and tolerance parameter $\tau_n = 0$ in their RSC condition.

ii) Equation (25) in the theorem does not need to be verified due to $\tau_n = 0$.

iii) Equation (26) has two conditions. One poses restriction on the ℓ_2 norm of the error W and the second poses restriction on the maximum norm of W as well as the low rank component Θ.

Moreover the choice of two tuning parameter satisfies that

$$\lambda_d \geq 4\left\|\mathfrak{X}^*(W)\right\|_{\mathrm{op}},$$

$$\mu_d \geq 4\mathcal{R}^*\left(\mathfrak{X}^*(W)\right) + \frac{4\gamma\alpha}{\kappa_d},$$

where W in our context is our error and $\kappa_d = \sqrt{m_r(L_0)m_c(L_0)}$ in our context.

Their spikiness parameter α is chosen such that

$$\frac{\alpha}{\kappa_d} \asymp \frac{1}{\sqrt{m_r(L_0)m_c(L_0)}},$$

as implied by Assumption 1(i).

Define a large enough positive constant α. Therefore for us it suffices to just verify (iii) for our context. Assumption 1 and 2 and conditions of theorem yield that with probability greater than $1 - p_n$ ($p_n \to 0$), we have $\nu_n > 4\|E\|_2$ and

$$\tau_n \geq 4\|E\|_{\max} + 4\frac{\alpha}{\sqrt{m_r(L_0)m_c(L_0)}} \geq 4(\|E\|_{\max} + \|L_0\|_{\max}). \tag{E.2}$$

Thus, assumption (iii) is also satisfied. *In addition, for the conclusion of Theorem 1, we have $\mathcal{K}_{\tau_n} = 0$ as τ_n equals 0 therein. Also we do not have misspecification in the low rank and sparse component, then the second term in $\mathcal{K}^*_\Theta$ and $\mathcal{K}^*_\Gamma$ are all zero.*

Hence there exists a positive constant C_1 such that, with the same probability event,

$$\begin{aligned}\|\widehat{L} - L_0\|_F^2 + \|\widehat{S} - S_0\|_F^2 \quad &\leq r\nu_n^2 + 4\tau_n^2 m_s(S_0) \\ &\leq C_1 r w_{2n}^2 + C_1 m_s(S_0) w_{\max n}^2 + C_1 \frac{\alpha^2 m_s(S_0)}{m_r(L_0)m_c(L_0)},\end{aligned}$$

where the second inequality is due to Assumption 2. Next, we present the results for $\|.\|_2$, corresponding to the second statement of Theorem 1. According to Agarwal et al. (2012), in case of identity design, we can cast the problem to a two step one. Define $|S|_1 = \sum_{i,j\in\{1,\cdots,n\}:i\neq j} |S_{i,j}|$. Namely:

Step 1 For a given $L = \mathbf{0}_{n\times n}$, we shall optimize:

$$\widehat{S} \coloneqq \arg\min_{S\in\mathbb{R}^{n\times n}} \frac{1}{2}\|W - L - S\|_F^2 + \tau_n|S|_1, \tag{E.3}$$

Step 2 For the estimator in the previous step $\widehat{S}$, we shall optimize:

$$\widehat{L} \coloneqq \arg\min_{L\in\mathbb{R}^{n\times n},} \frac{1}{2}\|W - L - \widehat{S}\|_F^2 + \nu_n\|L\|_*. \tag{E.4}$$

Then set $\widehat{S}_{ii} = -\widehat{L}_{ii}$, so that $\widehat{W} = \widehat{L} + \widehat{S}$ has zero diagonals.

Then we can apply Lemma 4(Theorem 9.19 b) in Wainwright (2019).) *We shall verify respectively for those two steps the following three conditions:*

i) The term

$$\Psi(\mathbb{S}) := \sup_{u\in\mathbb{S}\setminus(0)} \frac{\Phi(u)}{\|u\|}.$$

ii) The gradient condition to select λ_n:

$$\mathbb{G}(\lambda_n) := \left\{\Phi^*(\nabla\mathcal{L}_n(\theta^*)) \leq \frac{\lambda_n}{2}\right\}.$$

iii) Identification: exists a positive c_0 and κ such that,

$$\frac{\|\Delta\|_F^2}{2} \geq \frac{\kappa}{2}\|\Delta\|_F^2 - c_0\Phi(\Delta)^2 \quad \textit{for all} \quad \|\Delta\|_F \leq 1.$$

Thus following the conclusion of the theorem we should have the bound:

$$9\frac{\lambda_n^2}{\kappa^2}\Psi^2(\overline{\mathbb{M}}).$$

Because $\frac{8}{\kappa}\left\{\lambda_n\Phi\left(\theta^*_{\mathbb{M}^\perp}\right) + 16\tau_n^2\Phi^2\left(\theta^*_{\mathbb{M}^\perp}\right)\right\}$ *is zero.*

$[A]_{i,j}$ denotes a matrix with element A_{ij}. For the optimization in the Step 1: τ_n is of order $\|E\|_{\max} + \|L_0\|_{\max} \lesssim_p w_{\max n} + \frac{\alpha}{\sqrt{m_r(L_0)m_c(L_0)}}$ due to the gradient evaluated at the true value S_0 is just $[-(W_{ij} - S_{0,ij})]_{i,j}$ and the dual norm $\Phi^*(.)$ is the $\|.\|_{\max}$. $\kappa = 1$, and $\Psi(\mathbb{S}) \lesssim \sqrt{m_s(S_0)}$. This results to a bound of the order:

$$\|\widehat{S} - S_0\|_2 \lesssim_p (w_{\max n} + \frac{1}{\sqrt{m_r(L_0)m_c(L_0)}})\sqrt{m_s(S_0)} = \mathcal{R}_s. \tag{E.5}$$

Then we can apply Lemma 4 (Theorem 9.19 in Wainwright (2019)) to Step 1. For Step 2, we use the singular-value-thresholding representation; see also the related operator-norm result in Lemma 5 (Theorem 9.24 in Wainwright (2019)). The relevant conditions are verified as follows:

i) For both steps, the loss is quadratic. Therefore,

$$\mathcal{E}_n(\Delta) = \frac{1}{2}\|\Delta\|_F^2,$$

so the curvature condition holds globally with $\kappa = 1$ and zero tolerance.

ii) For Step 1, the entrywise ℓ_1 norm is decomposable with respect to the support space of S_0, and

$$\Psi(\overline{\mathbb{M}}) \leq \sqrt{m_s(S_0)}.$$

iii) The gradient of the Step 1 loss evaluated at S_0 is

$$\nabla\mathcal{L}_n(S_0) = S_0 - W = -(L_0 + E).$$

Since the dual of the entrywise ℓ_1 norm is the maximum norm,

$$\Phi^*(\nabla\mathcal{L}_n(S_0)) \leq \|L_0\|_{\max} + \|E\|_{\max} \lesssim_p \frac{1}{\sqrt{m_r(L_0)m_c(L_0)}} + w_{\max n}.$$

Thus, the gradient condition holds for the stated choice of τ_n.

iv) For Step 2, the nuclear norm is decomposable with respect to the tangent-space pair generated by L_0, and

$$\Psi(\overline{\mathbb{M}}) \lesssim \sqrt{r}.$$

v) The gradient of the Step 2 loss evaluated at L_0 is

$$\nabla \mathcal{L}_n(L_0) = L_0 + \widehat{S} - W = \widehat{S} - S_0 - E,$$

and hence

$$\Phi^*(\nabla \mathcal{L}_n(L_0)) = \|\widehat{S} - S_0 - E\|_2 \le \|\widehat{S} - S_0\|_2 + \|E\|_2 \lesssim_p \mathcal{R}_s + w_{2n}.$$

vi) Because S_0 is exactly sparse and L_0 has rank r, the approximation-error terms for both steps are zero.

According to Lemma 4 and the Step 1 Frobenius bound, we have

$$|\widehat{S} - S_0|_1 \lesssim_p \left(w_{\max n} + \frac{1}{\sqrt{m_r(L_0) m_c(L_0)}} \right) m_s(S_0) = \mathcal{R}_s \sqrt{m_s(S_0)}. \tag{E.6}$$

Denote

$$\mathrm{SVT}_{\nu_n}(A_n) \coloneqq U \,\mathrm{diag}\left\{ \big(\sigma_j(A_n) - \nu_n\big)_+ \right\} V^\top, \qquad (x)_+ \coloneqq \max\{x, 0\}.$$

For the optimization in Step 2, let

$$A_n = W - \widehat{S}.$$

The solution is obtained by singular-value thresholding:

$$\widehat{L} = \mathrm{SVT}_{\nu_n}(A_n), \qquad \|A_n - \widehat{L}\|_2 \le \nu_n.$$

Moreover,

$$A_n - L_0 = E - (\widehat{S} - S_0).$$

Therefore,

$$\begin{aligned} \|\widehat{L} - L_0\|_2 &\le \|\widehat{L} - A_n\|_2 + \|A_n - L_0\|_2 \\ &\le \nu_n + \|E\|_2 + \|\widehat{S} - S_0\|_2 \\ &\lesssim_p w_{2n} + \mathcal{R}_s, \end{aligned} \tag{E.7}$$

where the last inequality uses $\nu_n \asymp w_{2n}$ and $\|\widehat{S} - S_0\|_2 \lesssim_p \mathcal{R}_s$ from Step 1. Combining the two steps and applying the diagonal adjustment gives

$$\|\widehat{S} - S_0\|_2 \lesssim_p w_{2n} + \mathcal{R}_s.$$

Therefore,

$$\|\widehat{L} - L_0\|_2^2 + \|\widehat{S} - S_0\|_2^2 \lesssim_p w_{2n}^2 + \mathcal{R}_s^2 \lesssim_p \mathcal{R}_2(W_0, E),$$

which proves the second conclusion of the theorem.

Finally, since

$$\widehat{S}_{ii} - S_{0,ii} = -(\widehat{L}_{ii} - L_{0,ii}),$$

we have

$$\begin{aligned}|\widehat{S} - S_0|_{1,1} &\le |\widehat{S} - S_0|_1 + |\operatorname{diag}(\widehat{L} - L_0)|_{1,1} \\ &\le |\widehat{S} - S_0|_1 + n\|\widehat{L} - L_0\|_2 \\ &\lesssim_p \mathcal{R}_s\sqrt{m_s(S_0)} + n(\mathcal{R}_s + w_{2n}).\end{aligned}$$

□

Assumption 12 (Dependence between the measurement error and data)**.** Let

$$\mathcal{E} = \sigma(E), \qquad U_t = \lambda_0 Y_t + X_t\gamma_0,$$

and define the finite collection

$$\mathcal{P} = \{(X_{jt}, X_{kt}), (X_{jt}, Y_t), (X_{jt}, U_t), (X_{jt}, \varepsilon_t) : 1 \le j, k \le K\}.$$

Assume

$$\sup_{(a_t,b_t)\in\mathcal{P}} \sup_{1\le t\le T} \left\{\left\|\mathbb{E}[a_t b_t^\top \mid \mathcal{E}]\right\|_2 + \left\|\mathbb{E}[a_t b_t^\top \mid \mathcal{E}]\right\|_{\max}\right\} = O_p(1).$$

Moreover, uniformly over $(a_t, b_t) \in \mathcal{P}$ and all $\mathcal{E}$-measurable matrices $A \in \mathbb{R}^{n\times n}$,

$$\operatorname{Var}\left(\sum_{t=1}^{T} a_t^\top A b_t \,\middle|\, \mathcal{E}\right) \lesssim_p T\|A\|_F^2.$$

Lemma 10. *Define the gradient matrices*

$$\begin{aligned}\widehat{G}(\widehat{W}) &:= \frac{\partial \bar{g}_{nT}(\theta, \widehat{W})}{\partial\theta^\top} = -\frac{1}{nT}\sum_{t=1}^{T} Z_t^\top(\widehat{W})\left[\widehat{W}Y_t \quad X_t \quad \widehat{W}X_t\right], \\ \widehat{G}(W_0) &:= \frac{\partial \bar{g}_{nT}(\theta, W_0)}{\partial\theta^\top} = -\frac{1}{nT}\sum_{t=1}^{T} Z_t^\top(W_0)\left[W_0Y_t \quad X_t \quad W_0X_t\right],\end{aligned}$$

which do not depend on the evaluation point θ.

Suppose Assumptions 1–7 hold. Let $\Delta L := \widehat{L} - L_0$, $\Delta S := \widehat{S} - S_0$, $B_0 := (I_n - \lambda_0 W_0)^{-1}$, and $\mathcal{R}_s := \left(w_{\max n} + 1/\sqrt{m_r(L_0)m_c(L_0)}\right)\sqrt{m_s(S_0)}$. Throughout, Y_t, Z_t, X_t, ε_t denote the within-transformed quantities (see Remark 3).

(a) Baseline rate using the noisy matrix W. *Assume*

$$\|W_0\|_2 = O(1), \qquad w_{2n} = O(1).$$

Using $W = W_0 + E$ directly,

$$\begin{aligned}\|\widehat{G}(W) - \widehat{G}(W_0)\|_2 &= O_p(w_{2n} + w_{2n}^2), \\ \|\bar{g}_{nT}(\theta_0, W) - \bar{g}_{nT}(\theta_0, W_0)\|_2 &= O_p(w_{2n} + w_{2n}^2).\end{aligned}$$

(b) Alternative bound via the $|\cdot|_{1,1}$ norm. *Suppose, in addition, that*

$$\|W_0\|_1 = O(1), \qquad \|E\|_\infty = O_p(1),$$

and

$$\max_{1\le j\le K}\frac{1}{T}\sum_{t=1}^{T}\|X_{jt}\|_{\max}^2=O_p(1),\qquad \frac{1}{T}\sum_{t=1}^{T}\|\varepsilon_t\|_{\max}^2=O_p(1),$$

where X_{jt} is the jth column of X_t, $\|X_{jt}\|_{\max}=\max_{1\le i\le n}|x_{t,ij}|$, and $\|\varepsilon_t\|_{\max}=\max_{1\le i\le n}|\varepsilon_{it}|$. Then

$$\|\widehat{G}(W)-\widehat{G}(W_0)\|_2=O_p\big(n^{-1}|E|_{1,1}\big),$$
$$\|\bar{g}_{nT}(\theta_0,W)-\bar{g}_{nT}(\theta_0,W_0)\|_2=O_p\big(n^{-1}|E|_{1,1}\big).$$

(c) Improved rate using the denoised matrix $\widehat{W}$. *Let*

$$\mathcal{E}:=\sigma(E),\qquad U_t:=\lambda_0 Y_t+X_t\gamma_0,$$

and consider the finite collection of primitive pairs

$$\mathcal{P}=\{(X_{jt},X_{kt}),(X_{jt},Y_t),(X_{jt},U_t),(X_{jt},\varepsilon_t):1\le j,k\le K\}.$$

Suppose that

$$\sup_{(a_t,b_t)\in\mathcal{P}}\sup_{1\le t\le T}\big\|\mathbb{E}[a_tb_t^\top\mid\mathcal{E}]\big\|_2=O_p(1),\tag{E.8}$$

and

$$\sup_{(a_t,b_t)\in\mathcal{P}}\sup_{1\le t\le T}\big\|\mathbb{E}[a_tb_t^\top\mid\mathcal{E}]\big\|_{\max}=O_p(1).\tag{E.9}$$

Assume also that, uniformly over the pairs in $\mathcal{P}$ and over all $\mathcal{E}$-measurable matrices A,

$$\mathrm{Var}\left(\sum_{t=1}^{T}a_t^\top Ab_t\,\middle|\,\mathcal{E}\right)\lesssim_p T\|A\|_F^2.\tag{E.10}$$

Independence between E and the regressors or disturbances is not required; the conditional-moment conditions above are the operative restriction. Under independence, the two conditional-moment bounds reduce to their corresponding unconditional counterparts in the main assumptions, whereas under dependence they restrict the strength of that dependence.

Define

$$\rho_n:=w_{2n}+\mathcal{R}_s,\qquad s_n:=\mathcal{R}_s\sqrt{m_s(S_0)},\qquad \kappa_n:=1+\|W_0\|_2+\rho_n,$$

and

$$\mathcal{R}_{nT}^*:=(1+\|B_0\|_2)\kappa_n^{d+1}\frac{(r\vee 1)\rho_n+s_n}{n}.\tag{E.11}$$

If $\mathcal{R}_{nT}^*=o(1)$, then

$$\|\widehat{G}(\widehat{W})-\widehat{G}(W_0)\|_2=O_p(\mathcal{R}_{nT}^*),$$
$$\|\bar{g}_{nT}(\theta_0,\widehat{W})-\bar{g}_{nT}(\theta_0,W_0)\|_2=O_p(\mathcal{R}_{nT}^*).$$

The bound in part (b) corresponds to the rate available under sparsity-style arguments for measurement error in the adjacency matrix, of the kind used by Lewbel et al. (2024). When E is sparse, $|E|_{1,1}=o(n)$ and part (b) is faster than the baseline of part (a). When E is dense, however, $|E|_{1,1}$ is generically of order n or larger, and part (b) delivers no improvement over

part (a). This is precisely the regime in which the L+S-denoising rate of part (c) is needed: by exploiting the assumed L+S structure of the latent adjacency matrix $W_0 = L_0 + S_0$, the denoising step of Theorem 1 produces a recovery error $\widehat{W} - W_0 = \Delta L + \Delta S$ with controlled spectral behavior, allowing part (c) to achieve a rate that the elementwise-sparse bound of part (b) cannot reach.

The denoised rate of Lemma 10(c) is strictly faster than the baseline of Lemma 10(a) whenever

$$\frac{\|B_0\|_2(1 + \|W_0\|_2)\, r(w_{2n} + \mathcal{R}_s) + \mathcal{R}_s\sqrt{m_s(S_0)}}{n} = o(w_{2n} + w_{2n}^2).$$

This holds in particular for fixed or slowly growing rank r in the dense-network regime where $\|W_0\|_2$ remains bounded, so that $\|B_0\|_2 = O(1)$ when there are no dominant units ($J = 0$). Under finitely many dominant units, Lemma 8 gives $\|B_0\|_2 = o(n^{1/4})$, and the improvement holds under correspondingly stronger conditions on r and $m_s(S_0)$.

Proof of Lemma 10. **Part (a): $\|\cdot\|_2$ based bound.**

Throughout this part, $d = 2$, so the columns of $Z_t(W)$ are chosen from $W^l X_{jt}$, where $l = 0, 1, 2$ and $j = 1, \ldots, K$. Since K and the number of instruments are fixed, it suffices to bound each entry of the relevant matrices.

Recall that

$$B_0 = (I_n - \lambda_0 W_0)^{-1}.$$

The stability condition on W_0 implies

$$\|W_0\|_\infty = O(1), \qquad \|B_0\|_\infty = O(1), \qquad \|B_0 W_0\|_\infty = O(1).$$

The uniform moment bounds in Assumptions 6 and 5, together with these row-sum bounds, imply

$$\max_{1\le j\le K} \frac{1}{nT}\sum_{t=1}^{T} \|X_{jt}\|_2^2 = O_p(1), \tag{E.12}$$

$$\frac{1}{nT}\sum_{t=1}^{T} \|\varepsilon_t\|_2^2 = O_p(1), \tag{E.13}$$

$$\max_{\substack{1\le j\le K \\ H\in\{I_n, W_0\}}} \frac{1}{nT}\sum_{t=1}^{T} \|B_0 H X_{jt}\|_2^2 + \frac{1}{nT}\sum_{t=1}^{T} \|B_0\varepsilon_t\|_2^2 = O_p(1). \tag{E.14}$$

Indeed, uniformly in t, j, and $H \in \{I_n, W_0\}$, the uniform componentwise second-moment bounds and the bounded row sums of $B_0 H$ imply

$$\mathbb{E}\|B_0 H X_{jt}\|_2^2 = O(n).$$

Moreover, because ε_t denotes the within-transformed disturbance and the within transformation is a contraction,

$$\lambda_{\max}\big(\mathbb{E}[\varepsilon_t\varepsilon_t^\top]\big) \le \sigma_0^2.$$

Consequently,

$$\mathbb{E}\|B_0\varepsilon_t\|_2^2 = \operatorname{tr}\big(B_0^\top B_0\mathbb{E}[\varepsilon_t\varepsilon_t^\top]\big) \le \sigma_0^2\|B_0\|_F^2 \le \sigma_0^2 n\|B_0\|_\infty^2 = O(n),$$

where

$$\|B_0\|_F^2 = \sum_{i=1}^n \|B_{0,i\cdot}\|_2^2 \le \sum_{i=1}^n \|B_{0,i\cdot}\|_1^2 \le n\|B_0\|_\infty^2.$$

Therefore, the expectations of the normalized time averages are uniformly bounded, and Markov's inequality gives (E.12)–(E.14). Since

$$Y_t = B_0(X_t\beta_0 + W_0X_t\gamma_0 + \varepsilon_t),$$

the compactness of the parameter space, the fact that K is fixed, and (E.14) implies

$$\frac{1}{nT}\sum_{t=1}^T \|Y_t\|_2^2 = O_p(1). \tag{E.15}$$

Consequently, defining

$$U_t := \lambda_0 Y_t + X_t\gamma_0,$$

we also have

$$\frac{1}{nT}\sum_{t=1}^T \|U_t\|_2^2 = O_p(1). \tag{E.16}$$

Define

$$A_l := W^l - W_0^l, \qquad l = 0, 1, 2.$$

Since $W = W_0 + E$,

$$A_0 = 0, \qquad A_1 = E,$$

and

$$A_2 = W^2 - W_0^2 = W_0E + EW_0 + E^2.$$

Therefore, using $\|W_0\|_2 = O(1)$ and $\|E\|_2 = O_p(w_{2n})$,

$$\max_{0\le l\le 2}\|A_l\|_2 = O_p(w_{2n} + w_{2n}^2). \tag{E.17}$$

Moreover,

$$\|W\|_2 \le \|W_0\|_2 + \|E\|_2 = O_p(1),$$

where the final equality uses $w_{2n} = O(1)$.

We first consider the gradient. By its definition,

$$\begin{aligned}\|\widehat{G}(W)-\widehat{G}(W_0)\|_2 \lesssim & \left\|\frac{1}{nT}\sum_{t=1}^{T}\left[Z_t^\top(W)-Z_t^\top(W_0)\right]X_t\right\|_2 \\ & + \left\|\frac{1}{nT}\sum_{t=1}^{T}\left[Z_t^\top(W)W-Z_t^\top(W_0)W_0\right]Y_t\right\|_2 \\ & + \left\|\frac{1}{nT}\sum_{t=1}^{T}\left[Z_t^\top(W)W-Z_t^\top(W_0)W_0\right]X_t\right\|_2 \\ =: & \|\mathcal{R}_{n1}\|_2+\|\mathcal{R}_{n2}\|_2+\|\mathcal{R}_{n3}\|_2.\end{aligned}$$

For $\mathcal{R}_{n1}$, every (j,k,l) entry satisfies

$$\begin{aligned}&\left|\frac{1}{nT}\sum_{t=1}^{T}X_{jt}^\top A_l^\top X_{kt}\right| \\ &\quad\le \|A_l\|_2\left(\frac{1}{nT}\sum_{t=1}^{T}\|X_{jt}\|_2^2\right)^{1/2}\left(\frac{1}{nT}\sum_{t=1}^{T}\|X_{kt}\|_2^2\right)^{1/2}.\end{aligned}$$

It follows from (E.12) and (E.17) that

$$\|\mathcal{R}_{n1}\|_2 = O_p(w_{2n}+w_{2n}^2). \tag{E.18}$$

For $\mathcal{R}_{n2}$ and $\mathcal{R}_{n3}$, define

$$D_l := (W^l)^\top W - (W_0^l)^\top W_0.$$

Because

$$D_l = A_l^\top W + (W_0^l)^\top E,$$

we have

$$\|D_l\|_2 \le \|A_l\|_2\|W\|_2 + \|W_0\|_2^l\|E\|_2.$$

Thus,

$$\max_{0\le l\le 2}\|D_l\|_2 = O_p(w_{2n}+w_{2n}^2). \tag{E.19}$$

Each (j,l) entry of $\mathcal{R}_{n2}$ satisfies

$$\begin{aligned}\left|\frac{1}{nT}\sum_{t=1}^{T}X_{jt}^\top D_l Y_t\right| \le \|D_l\|_2&\left(\frac{1}{nT}\sum_{t=1}^{T}\|X_{jt}\|_2^2\right)^{1/2} \\ &\times\left(\frac{1}{nT}\sum_{t=1}^{T}\|Y_t\|_2^2\right)^{1/2}.\end{aligned}$$

Therefore, by (E.12), (E.15), and (E.19),

$$\|\mathcal{R}_{n2}\|_2 = O_p(w_{2n}+w_{2n}^2). \tag{E.20}$$

Similarly, every (j,k,l) entry of $\mathcal{R}_{n3}$ satisfies

$$\left|\frac{1}{nT}\sum_{t=1}^{T} X_{jt}^{\top} D_l X_{kt}\right| \le \|D_l\|_2 \left(\frac{1}{nT}\sum_{t=1}^{T}\|X_{jt}\|_2^2\right)^{1/2} \times \left(\frac{1}{nT}\sum_{t=1}^{T}\|X_{kt}\|_2^2\right)^{1/2}.$$

Hence

$$\|\mathcal{R}_{n3}\|_2 = O_p(w_{2n} + w_{2n}^2). \tag{E.21}$$

Combining (E.18)–(E.21), we obtain

$$\|\widehat{G}(W) - \widehat{G}(W_0)\|_2 = O_p(w_{2n} + w_{2n}^2).$$

We next consider the sample moment. Define

$$\varepsilon_t(\theta_0, M) := Y_t - \lambda_0 M Y_t - X_t\beta_0 - M X_t \gamma_0.$$

Then

$$\varepsilon_t(\theta_0, W_0) = \varepsilon_t$$

and

$$\delta_t := \varepsilon_t(\theta_0, W) - \varepsilon_t = -E(\lambda_0 Y_t + X_t\gamma_0) = -EU_t.$$

Consequently,

$$\begin{aligned}
&\bar{g}_{nT}(\theta_0, W) - \bar{g}_{nT}(\theta_0, W_0) \\
&\quad = \frac{1}{nT}\sum_{t=1}^{T} Z_t^{\top}(W_0)\delta_t + \frac{1}{nT}\sum_{t=1}^{T}\left[Z_t^{\top}(W) - Z_t^{\top}(W_0)\right]\varepsilon_t(\theta_0, W) \\
&\quad =: \mathcal{R}_1 + \mathcal{R}_2.
\end{aligned}$$

For each instrument component of $\mathcal{R}_1$,

$$\left|\frac{1}{nT}\sum_{t=1}^{T} X_{jt}^{\top}(W_0^l)^{\top} E U_t\right| \le \|E\|_2 \left(\frac{1}{nT}\sum_{t=1}^{T}\|W_0^l X_{jt}\|_2^2\right)^{1/2} \times \left(\frac{1}{nT}\sum_{t=1}^{T}\|U_t\|_2^2\right)^{1/2}.$$

Because $\|W_0\|_2 = O(1)$, (E.12) and (E.16) imply

$$\|\mathcal{R}_1\|_2 = O_p(w_{2n}).$$

For $\mathcal{R}_2$, use

$$\varepsilon_t(\theta_0, W) = \varepsilon_t + \delta_t$$

to write

$$\mathcal{R}_2 = \underbrace{\frac{1}{nT}\sum_{t=1}^{T}\left[Z_t^\top(W) - Z_t^\top(W_0)\right]\varepsilon_t}_{\mathcal{R}_{21}} + \underbrace{\frac{1}{nT}\sum_{t=1}^{T}\left[Z_t^\top(W) - Z_t^\top(W_0)\right]\delta_t}_{\mathcal{R}_{22}}.$$

For every (j,l) component of $\mathcal{R}_{21}$,

$$\left|\frac{1}{nT}\sum_{t=1}^{T} X_{jt}^\top A_l^\top \varepsilon_t\right| \le \|A_l\|_2 \left(\frac{1}{nT}\sum_{t=1}^{T}\|X_{jt}\|_2^2\right)^{1/2} \times \left(\frac{1}{nT}\sum_{t=1}^{T}\|\varepsilon_t\|_2^2\right)^{1/2}.$$

Thus,

$$\|\mathcal{R}_{21}\|_2 = O_p(w_{2n} + w_{2n}^2).$$

Finally, since $\delta_t = -EU_t$, every component of $\mathcal{R}_{22}$ satisfies

$$\left|\frac{1}{nT}\sum_{t=1}^{T} X_{jt}^\top A_l^\top E U_t\right| \le \|A_l\|_2\|E\|_2 \left(\frac{1}{nT}\sum_{t=1}^{T}\|X_{jt}\|_2^2\right)^{1/2} \times \left(\frac{1}{nT}\sum_{t=1}^{T}\|U_t\|_2^2\right)^{1/2}.$$

It follows that

$$\|\mathcal{R}_{22}\|_2 = O_p(w_{2n}^2 + w_{2n}^3) = O_p(w_{2n}^2),$$

where the final equality uses $w_{2n} = O(1)$.

Combining the preceding bounds gives

$$\|\bar{g}_{nT}(\theta_0, W) - \bar{g}_{nT}(\theta_0, W_0)\|_2 = O_p(w_{2n} + w_{2n}^2),$$

which completes the proof of part (a).

Part (b): $|\cdot|_{1,1}$-based bound.

Throughout this part, $d = 2$. We use the inequalities

$$|AB|_{1,1} \le \|A\|_1 |B|_{1,1}, \qquad |AB|_{1,1} \le \|B\|_\infty |A|_{1,1}. \tag{E.22}$$

For vectors $x, y \in \mathbb{R}^n$, we also use

$$|x^\top A y| \le \|x\|_{\max} |A|_{1,1} \|y\|_{\max}. \tag{E.23}$$

Recall the definitions

$$A_l := W^l - W_0^l, \qquad D_l := (W^l)^\top W - (W_0^l)^\top W_0, \qquad l = 0, 1, 2.$$

Since $W = W_0 + E$,

$$A_0 = 0, \qquad A_1 = E,$$

and

$$A_2 = W_0 E + E W_0 + E^2.$$

Using (E.22),

$$\begin{aligned} |A_2|_{1,1} &\le |W_0 E|_{1,1} + |E W_0|_{1,1} + |E^2|_{1,1} \\ &\le (\|W_0\|_1 + \|W_0\|_\infty + \|E\|_\infty)\, |E|_{1,1}. \end{aligned}$$

By the additional hypothesis of part (b), $\|W_0\|_1 = O(1)$. Moreover, Assumption 4(iii) gives $\|W_0\|_\infty = O(1)$, while $\|E\|_\infty = O_p(1)$ by assumption. Therefore,

$$\max_{0 \le l \le 2} |A_l|_{1,1} = O_p(|E|_{1,1}). \tag{E.24}$$

Moreover,

$$D_l = A_l^\top W + (W_0^l)^\top E.$$

Therefore,

$$\begin{aligned} |D_l|_{1,1} &\le |A_l^\top W|_{1,1} + |(W_0^l)^\top E|_{1,1} \\ &\le \|W\|_\infty |A_l^\top|_{1,1} + \|(W_0^l)^\top\|_1 |E|_{1,1} \\ &= \|W\|_\infty |A_l|_{1,1} + \|W_0^l\|_\infty |E|_{1,1}. \end{aligned}$$

Since

$$\|W\|_\infty \le \|W_0\|_\infty + \|E\|_\infty = O_p(1)$$

and

$$\|W_0^l\|_\infty \le \|W_0\|_\infty^l = O(1),$$

it follows that

$$\max_{0 \le l \le 2} |D_l|_{1,1} = O_p(|E|_{1,1}). \tag{E.25}$$

We next record bounds for the data factors. Since

$$Y_t = B_0 \left(X_t \beta_0 + W_0 X_t \gamma_0 + \varepsilon_t \right),$$

Assumption 4(iii) gives

$$\|B_0\|_\infty = O(1), \qquad \|W_0\|_\infty = O(1).$$

Together with the boundedness of β_0 and γ_0 and the fact that K is fixed, this gives

$$\|Y_t\|_{\max} \lesssim \sum_{k=1}^{K} \|X_{kt}\|_{\max} + \|\varepsilon_t\|_{\max}.$$

By the additional maximum-norm assumptions in part (b),

$$\frac{1}{T}\sum_{t=1}^{T}\|Y_t\|_{\max}^2 = O_p(1). \tag{E.26}$$

Similarly, defining

$$U_t := \lambda_0 Y_t + X_t\gamma_0,$$

we have

$$\frac{1}{T}\sum_{t=1}^{T}\|U_t\|_{\max}^2 = O_p(1). \tag{E.27}$$

With $\mathcal{R}_{n1}$, $\mathcal{R}_{n2}$, and $\mathcal{R}_{n3}$ defined as in part (a), we first bound the gradient difference. For $\mathcal{R}_{n1}$, every (j,k,l) entry satisfies

$$\begin{aligned}&\left|\frac{1}{nT}\sum_{t=1}^{T}X_{jt}^\top A_l^\top X_{kt}\right| \\ &\quad\le \frac{|A_l|_{1,1}}{n}\left(\frac{1}{T}\sum_{t=1}^{T}\|X_{jt}\|_{\max}^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\|X_{kt}\|_{\max}^2\right)^{1/2}.\end{aligned}$$

Thus, by (E.24) and the additional maximum-norm assumptions,

$$\|\mathcal{R}_{n1}\|_2 = O_p\left(n^{-1}|E|_{1,1}\right).$$

For $\mathcal{R}_{n2}$, every (j,l) entry satisfies

$$\begin{aligned}\left|\frac{1}{nT}\sum_{t=1}^{T}X_{jt}^\top D_l Y_t\right| &\le \frac{|D_l|_{1,1}}{n}\left(\frac{1}{T}\sum_{t=1}^{T}\|X_{jt}\|_{\max}^2\right)^{1/2} \\ &\quad\times\left(\frac{1}{T}\sum_{t=1}^{T}\|Y_t\|_{\max}^2\right)^{1/2}.\end{aligned}$$

Therefore, by (E.25), (E.26), and the additional maximum-norm assumptions,

$$\|\mathcal{R}_{n2}\|_2 = O_p\left(n^{-1}|E|_{1,1}\right).$$

Similarly, every (j,k,l) entry of $\mathcal{R}_{n3}$ satisfies

$$\begin{aligned}\left|\frac{1}{nT}\sum_{t=1}^{T}X_{jt}^\top D_l X_{kt}\right| &\le \frac{|D_l|_{1,1}}{n}\left(\frac{1}{T}\sum_{t=1}^{T}\|X_{jt}\|_{\max}^2\right)^{1/2} \\ &\quad\times\left(\frac{1}{T}\sum_{t=1}^{T}\|X_{kt}\|_{\max}^2\right)^{1/2}.\end{aligned}$$

Hence,

$$\|\mathcal{R}_{n3}\|_2 = O_p\left(n^{-1}|E|_{1,1}\right).$$

Since K, d, and the number of instruments are fixed, the preceding entrywise bounds also hold for the corresponding matrix 2-norms. Combining the three terms gives

$$\|\widehat{G}(W) - \widehat{G}(W_0)\|_2 = O_p\left(n^{-1}|E|_{1,1}\right).$$

We next consider the sample-moment difference. Recall that

$$\delta_t := \varepsilon_t(\theta_0, W) - \varepsilon_t = -EU_t.$$

As in part (a), decompose

$$\begin{aligned}
&\bar{g}_{nT}(\theta_0, W) - \bar{g}_{nT}(\theta_0, W_0) \\
&\quad = \underbrace{\frac{1}{nT}\sum_{t=1}^{T} Z_t^\top(W_0)\delta_t}_{\mathcal{R}_1} + \underbrace{\frac{1}{nT}\sum_{t=1}^{T}\left[Z_t^\top(W) - Z_t^\top(W_0)\right]\varepsilon_t(\theta_0, W)}_{\mathcal{R}_2}.
\end{aligned}$$

For each instrument component of $\mathcal{R}_1$,

$$\begin{aligned}
&\left|\frac{1}{nT}\sum_{t=1}^{T} X_{jt}^\top (W_0^l)^\top E U_t\right| \\
&\qquad \le \frac{|(W_0^l)^\top E|_{1,1}}{n}\left(\frac{1}{T}\sum_{t=1}^{T}\|X_{jt}\|_{\max}^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\|U_t\|_{\max}^2\right)^{1/2}.
\end{aligned}$$

Furthermore,

$$|(W_0^l)^\top E|_{1,1} \le \|(W_0^l)^\top\|_1 |E|_{1,1} = \|W_0^l\|_\infty |E|_{1,1} = O(|E|_{1,1}).$$

Hence,

$$\|\mathcal{R}_1\|_2 = O_p\left(n^{-1}|E|_{1,1}\right).$$

For $\mathcal{R}_2$, write

$$\mathcal{R}_2 = \mathcal{R}_{21} + \mathcal{R}_{22},$$

where

$$\mathcal{R}_{21} := \frac{1}{nT}\sum_{t=1}^{T}\left[Z_t^\top(W) - Z_t^\top(W_0)\right]\varepsilon_t,$$

$$\mathcal{R}_{22} := \frac{1}{nT}\sum_{t=1}^{T}\left[Z_t^\top(W) - Z_t^\top(W_0)\right]\delta_t.$$

For each component of $\mathcal{R}_{21}$,

$$\left|\frac{1}{nT}\sum_{t=1}^{T} X_{jt}^{\top} A_l^{\top} \varepsilon_t\right| \le \frac{|A_l|_{1,1}}{n}\left(\frac{1}{T}\sum_{t=1}^{T}\|X_{jt}\|_{\max}^2\right)^{1/2} \times \left(\frac{1}{T}\sum_{t=1}^{T}\|\varepsilon_t\|_{\max}^2\right)^{1/2}.$$

Thus,

$$\|\mathcal{R}_{21}\|_2 = O_p\left(n^{-1}|E|_{1,1}\right).$$

Finally, because $\delta_t = -EU_t$, every component of $\mathcal{R}_{22}$ satisfies

$$\left|\frac{1}{nT}\sum_{t=1}^{T} X_{jt}^{\top} A_l^{\top} E U_t\right| \le \frac{|A_l^{\top}E|_{1,1}}{n}\left(\frac{1}{T}\sum_{t=1}^{T}\|X_{jt}\|_{\max}^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\|U_t\|_{\max}^2\right)^{1/2}.$$

Using (E.22) and (E.24),

$$|A_l^{\top}E|_{1,1} \le \|E\|_{\infty}|A_l^{\top}|_{1,1} = \|E\|_{\infty}|A_l|_{1,1} = O_p(|E|_{1,1}).$$

Therefore,

$$\|\mathcal{R}_{22}\|_2 = O_p\left(n^{-1}|E|_{1,1}\right).$$

Combining the preceding bounds yields

$$\|\bar{g}_{nT}(\theta_0, W) - \bar{g}_{nT}(\theta_0, W_0)\|_2 = O_p\left(n^{-1}|E|_{1,1}\right),$$

which completes the proof of part (b).

Proof of part (c). Define

$$\widehat{W} := \mathrm{off}(\widehat{L} + \widehat{S}), \qquad \Delta_W := \widehat{W} - W_0,$$

where

$$\mathrm{off}(A) = A - \mathrm{diag}(A).$$

The off-diagonal operator is redundant under the zero-diagonal constraint, but we retain it to make the diagonal bookkeeping explicit. Since $\mathrm{diag}(W_0) = 0$,

$$\Delta_W = \mathrm{off}(\Delta L) + \mathrm{off}(\Delta S).$$

Recall that

$$\rho_n := w_{2n} + \mathcal{R}_s, \qquad s_n := \mathcal{R}_s\sqrt{m_s(S_0)}, \qquad \kappa_n := 1 + \|W_0\|_2 + \rho_n.$$

Theorem 1 and its cone condition imply

$$\|\Delta L\|_* = O_p\big((r \vee 1)\rho_n\big), \tag{E.28}$$

$$\|\Delta L\|_F + \|\Delta S\|_F = O_p\big(\sqrt{r \vee 1}\,\rho_n\big), \tag{E.29}$$

$$|\operatorname{off}(\Delta S)|_{1,1} = O_p(s_n), \tag{E.30}$$

$$\|\Delta_W\|_2 = O_p(\rho_n). \tag{E.31}$$

Consequently,

$$\|\widehat{W}\|_2 \le \|W_0\|_2 + \|\Delta_W\|_2 = O_p(\kappa_n).$$

Let $\mathcal{E} = \sigma(E)$. Since $\widehat{W}$ is constructed from $W = W_0 + E$, the matrices $\widehat{W}$, Δ_W, ΔL, and ΔS are $\mathcal{E}$-measurable.

We first record the bilinear-form bound used below. Consider

$$\mathcal{B}_{nT}(A) = \frac{1}{nT}\sum_{t=1}^{T} a_t^\top A b_t,$$

where (a_t, b_t) is one of

$$(X_{jt}, X_{kt}), \qquad (X_{jt}, Y_t), \qquad (X_{jt}, U_t), \qquad (X_{jt}, \varepsilon_t),$$

and A is $\mathcal{E}$-measurable.

Suppose first that

$$A = P_E \operatorname{off}(\Delta L) Q_E,$$

where

$$\|P_E\|_2 \|Q_E\|_2 = O_p\left((1 + \|B_0\|_2)\kappa_n^{d+1}\right).$$

Define

$$M_E = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[a_t b_t^\top \mid \mathcal{E}].$$

The conditional-moment condition gives

$$\|M_E\|_2 = O_p(1).$$

Since the off-diagonal operator is self-adjoint under the Frobenius inner product,

$$\begin{aligned} |\mathbb{E}[\mathcal{B}_{nT}(A) \mid \mathcal{E}]| &= \frac{1}{n}\left|\left\langle \operatorname{off}(\Delta L), P_E^\top M_E Q_E^\top \right\rangle_F\right| \\ &= \frac{1}{n}\left|\left\langle \Delta L, \operatorname{off}\left(P_E^\top M_E Q_E^\top\right)\right\rangle_F\right|. \end{aligned}$$

For any matrix N,

$$\|\operatorname{off}(N)\|_2 \le \|N\|_2 + \|\operatorname{diag}(N)\|_2 \le 2\|N\|_2.$$

Therefore, nuclear–operator norm duality and (E.28) give

$$|\mathbb{E}[\mathcal{B}_{nT}(A) \mid \mathcal{E}]| \le \frac{2}{n}\|\Delta L\|_*\|P_E\|_2\|M_E\|_2\|Q_E\|_2 = O_p\left((1+\|B_0\|_2)\kappa_n^{d+1}\frac{(r\vee 1)\rho_n}{n}\right). \tag{E.32}$$

Moreover,

$$\begin{aligned}\|A\|_F &\le \|P_E\|_2\|\operatorname{off}(\Delta L)\|_F\|Q_E\|_2\\ &\le \|P_E\|_2\|\Delta L\|_F\|Q_E\|_2\\ &= O_p\left((1+\|B_0\|_2)\kappa_n^{d+1}\sqrt{r\vee 1}\,\rho_n\right).\end{aligned}$$

The conditional variance condition and conditional Chebyshev's inequality yield

$$\mathcal{B}_{nT}(A) - \mathbb{E}[\mathcal{B}_{nT}(A) \mid \mathcal{E}] = O_p\left(\frac{\|A\|_F}{n\sqrt{T}}\right) = O_p\left((1+\|B_0\|_2)\kappa_n^{d+1}\frac{\sqrt{r\vee 1}\,\rho_n}{n\sqrt{T}}\right). \tag{E.33}$$

If instead

$$A = P_E \operatorname{off}(\Delta S) Q_E,$$

then entrywise ℓ_1–max norm duality gives

$$\begin{aligned}|\mathbb{E}[\mathcal{B}_{nT}(A) \mid \mathcal{E}]| &= \frac{1}{n}\left|\left\langle \operatorname{off}(\Delta S), P_E^\top M_E Q_E^\top\right\rangle_F\right|\\ &\le \frac{1}{n}|\operatorname{off}(\Delta S)|_{1,1}\|P_E^\top M_E Q_E^\top\|_{\max}.\end{aligned}$$

Since

$$\|P_E^\top M_E Q_E^\top\|_{\max} \le \|P_E^\top M_E Q_E^\top\|_2 \le \|P_E\|_2\|M_E\|_2\|Q_E\|_2,$$

we obtain

$$|\mathbb{E}[\mathcal{B}_{nT}(A) \mid \mathcal{E}]| = O_p\left((1+\|B_0\|_2)\kappa_n^{d+1}\frac{s_n}{n}\right). \tag{E.34}$$

Similarly,

$$\begin{aligned}\|A\|_F &\le \|P_E\|_2\|\operatorname{off}(\Delta S)\|_F\|Q_E\|_2\\ &\le \|P_E\|_2\|\Delta S\|_F\|Q_E\|_2,\end{aligned}$$

and hence

$$\mathcal{B}_{nT}(A) - \mathbb{E}[\mathcal{B}_{nT}(A) \mid \mathcal{E}] = O_p\left((1+\|B_0\|_2)\kappa_n^{d+1}\frac{\sqrt{r\vee 1}\,\rho_n}{n\sqrt{T}}\right). \tag{E.35}$$

Because $r \vee 1 \ge 1$ and $T \ge 1$,

$$\frac{\sqrt{r\vee 1}\,\rho_n}{n\sqrt{T}} \le \frac{(r\vee 1)\rho_n}{n}.$$

Thus every bilinear form considered above is

$$O_p\left((1+\|B_0\|_2)\kappa_n^{d+1}\frac{(r\vee 1)\rho_n+s_n}{n}\right)=O_p(\mathcal{R}^*_{nT}). \tag{E.36}$$

Left factors such as $W_0^l X_{jt}$ or $\widehat{W}^a X_{jt}$ are handled by absorbing the corresponding $\mathcal{E}$-measurable network matrices into P_E or Q_E. This explains why the conditional-moment conditions are stated using the primitive vectors X_{jt}.

Moment difference. As in part (a), write

$$\begin{aligned}
&\bar{g}_{nT}(\theta_0,\widehat{W})-\bar{g}_{nT}(\theta_0,W_0)\\
&\quad=\frac{1}{nT}\sum_{t=1}^{T}Z_t(W_0)^\top\left[\varepsilon_t(\theta_0,\widehat{W})-\varepsilon_t(\theta_0,W_0)\right]\\
&\qquad+\frac{1}{nT}\sum_{t=1}^{T}\left[Z_t(\widehat{W})-Z_t(W_0)\right]^\top\varepsilon_t(\theta_0,\widehat{W})\\
&\quad=:\mathcal{R}_1+\mathcal{R}_2.
\end{aligned}$$

Let

$$F_t:=B_0(X_t\beta_0+W_0X_t\gamma_0).$$

Since

$$Y_t=F_t+B_0\varepsilon_t$$

and

$$U_t=\lambda_0Y_t+X_t\gamma_0,$$

we have

$$\begin{aligned}
\mathcal{R}_1&=-\lambda_0\frac{1}{nT}\sum_{t=1}^{T}Z_t(W_0)^\top\Delta_W B_0\varepsilon_t\\
&\quad-\lambda_0\frac{1}{nT}\sum_{t=1}^{T}Z_t(W_0)^\top\Delta_W F_t\\
&\quad-\frac{1}{nT}\sum_{t=1}^{T}Z_t(W_0)^\top\Delta_W X_t\gamma_0\\
&=:A_{1nT}+A_{2nT}+A_{3nT}.
\end{aligned}$$

Every coordinate of A_{1nT} is a finite sum of terms involving the pair (X_{jt},ε_t), with B_0 and the powers of W_0 absorbed into P_E and Q_E. Hence,

$$\|A_{1nT}\|_2=O_p(\mathcal{R}^*_{nT}).$$

After writing

$$F_t=B_0X_t\beta_0+B_0W_0X_t\gamma_0,$$

every coordinate of A_{2nT} involves a pair (X_{jt}, X_{kt}). Therefore,

$$\|A_{2nT}\|_2 = O_p(\mathcal{R}^*_{nT}).$$

The same argument using (X_{jt}, X_{kt}) gives

$$\|A_{3nT}\|_2 = O_p(\mathcal{R}^*_{nT}).$$

Consequently,

$$\|\mathcal{R}_1\|_2 = O_p(\mathcal{R}^*_{nT}). \tag{E.37}$$

For $\mathcal{R}_2$, write

$$\begin{aligned}\mathcal{R}_2 &= \frac{1}{nT}\sum_{t=1}^{T}\left[Z_t(\widehat{W}) - Z_t(W_0)\right]^\top \varepsilon_t \\ &\quad + \frac{1}{nT}\sum_{t=1}^{T}\left[Z_t(\widehat{W}) - Z_t(W_0)\right]^\top \left[\varepsilon_t(\theta_0, \widehat{W}) - \varepsilon_t\right] \\ &=: \mathcal{R}_{21} + \mathcal{R}_{22}.\end{aligned}$$

For every $l \le d$,

$$\widehat{W}^l - W_0^l = \sum_{a=0}^{l-1} \widehat{W}^a \Delta_W W_0^{l-1-a}. \tag{E.38}$$

Thus each coordinate of $\mathcal{R}_{21}$ is a finite sum of bilinear forms involving (X_{jt}, ε_t) and at least one factor Δ_W. It follows that

$$\|\mathcal{R}_{21}\|_2 = O_p(\mathcal{R}^*_{nT}).$$

Moreover,

$$\varepsilon_t(\theta_0, \widehat{W}) - \varepsilon_t = -\Delta_W U_t.$$

Thus each coordinate of $\mathcal{R}_{22}$ involves the pair (X_{jt}, U_t) and at least two recovery-error factors. One factor is paired with the data bilinear form, while every remaining factor is bounded using

$$\|\Delta_W\|_2 = O_p(\rho_n).$$

These additional factors are absorbed into κ_n^{d+1}. Therefore,

$$\|\mathcal{R}_{22}\|_2 = O_p(\mathcal{R}^*_{nT}).$$

Combining these bounds yields

$$\|\mathcal{R}_2\|_2 = O_p(\mathcal{R}^*_{nT}). \tag{E.39}$$

Equations (E.37) and (E.39) give

$$\|\bar{g}_{nT}(\theta_0, \widehat{W}) - \bar{g}_{nT}(\theta_0, W_0)\|_2 = O_p(\mathcal{R}^*_{nT}).$$

Gradient difference. As in part (a), decompose

$$\begin{aligned}\|\widehat{G}(\widehat{W}) - \widehat{G}(W_0)\|_2 \lesssim & \left\| \frac{1}{nT} \sum_{t=1}^{T} \left[Z_t(\widehat{W}) - Z_t(W_0) \right]^\top X_t \right\|_2 \\ & + \left\| \frac{1}{nT} \sum_{t=1}^{T} \left[Z_t(\widehat{W})^\top \widehat{W} - Z_t(W_0)^\top W_0 \right] Y_t \right\|_2 \\ & + \left\| \frac{1}{nT} \sum_{t=1}^{T} \left[Z_t(\widehat{W})^\top \widehat{W} - Z_t(W_0)^\top W_0 \right] X_t \right\|_2 \\ =: & \|\mathcal{R}_{n1}\|_2 + \|\mathcal{R}_{n2}\|_2 + \|\mathcal{R}_{n3}\|_2.\end{aligned}$$

For an instrument column of the form $W^l X_{jt}$, define

$$A_l := \widehat{W}^l - W_0^l, \qquad D_l := (\widehat{W}^l)^\top \widehat{W} - (W_0^l)^\top W_0.$$

By (E.38), A_l is a finite sum of terms containing at least one factor Δ_W. Moreover,

$$D_l = A_l^\top \widehat{W} + (W_0^l)^\top \Delta_W.$$

Therefore, every term in D_l also contains at least one factor Δ_W.

Each coordinate of $\mathcal{R}_{n1}$ has the form

$$\frac{1}{nT} \sum_{t=1}^{T} X_{jt}^\top A_l^\top X_{kt},$$

and hence involves the pair (X_{jt}, X_{kt}). Therefore,

$$\|\mathcal{R}_{n1}\|_2 = O_p(\mathcal{R}_{nT}^*).$$

Each coordinate of $\mathcal{R}_{n2}$ has the form

$$\frac{1}{nT} \sum_{t=1}^{T} X_{jt}^\top D_l Y_t,$$

and hence involves the pair (X_{jt}, Y_t). Thus,

$$\|\mathcal{R}_{n2}\|_2 = O_p(\mathcal{R}_{nT}^*).$$

Finally, each coordinate of $\mathcal{R}_{n3}$ has the form

$$\frac{1}{nT} \sum_{t=1}^{T} X_{jt}^\top D_l X_{kt},$$

which again involves the pair (X_{jt}, X_{kt}). Hence,

$$\|\mathcal{R}_{n3}\|_2 = O_p(\mathcal{R}_{nT}^*).$$

Combining the three bounds gives

$$\|\widehat{G}(\widehat{W}) - \widehat{G}(W_0)\|_2 = O_p(\mathcal{R}^*_{nT}).$$

This completes the proof of part (c). □

Proof of Theorem 2. **Step 1** The proof consists of two steps. First, we show that when W_0 is correctly specified, we establish the large sample performance of $\widehat{\theta}_p(W_0)$. In the first step, when W_0 is correctly specified, $\varepsilon_t(\theta_0, W_0) = \varepsilon_t$ and hence

$$\bar{g}_{nT}(\theta_0, W_0) = \sum_{t=1}^{T} Z_t^\top(W_0)\varepsilon_t/(nT).$$

Note that the gradient

$$\frac{\partial \bar{g}_{nT}(\theta, W)}{\partial \theta^\top} = \frac{\partial \frac{1}{nT}\sum_{t=1}^{T} Z_t(W)^\top(Y_t - \lambda W Y_t - X_t\beta - WX_t\gamma)}{\partial \theta^\top} = -\frac{1}{nT}\sum_{t=1}^{T} Z_t(W)^\top(WY_t, X_t, WX_t),$$

with $Y_t = (I_n - \lambda_0 W_0)^{-1}(X_t\beta_0 + W_0X_t\gamma_0 + \varepsilon_t)$. Define $\mathbb{E}[\widehat{G}(W)] = G(W)$. Recall that

$$\widehat{G}(W) = \frac{\partial \bar{g}_{nT}(\theta, W)}{\partial \theta^\top}.$$

Recall $Q_{0t} = (W_0(I_n - \lambda_0 W_0)^{-1}(X_t\beta_0 + W_0X_t\gamma_0), X_t, W_0X_t)$, and we have

$$\begin{aligned}\frac{\partial \bar{g}_{nT}(\theta, W_0)}{\partial \theta^\top} &= -\frac{\sum_{t=1}^{T} Z_t^\top(W_0)Q_{0t}}{nT} - \frac{\sum_{t=1}^{T} Z_t^\top(W_0)(W_0(I_n - \lambda_0 W_0)^{-1}\varepsilon_t, \mathbf{0}_{n\times(2K)})}{nT} \\ &:= \mathcal{R}_1 + \mathcal{R}_2\end{aligned}$$

To prove $\mathcal{R}_1 + \mathcal{R}_2 \to -\Sigma_{Z_0Q_0}$ holds, Assumption 6 yields the first term $\mathcal{R}_1$ converges to $-\Sigma_{Z_0Q_0}$. For the second term $\mathcal{R}_2$ we have $\frac{1}{nT}\sum_{t=1}^{T} Z_t^\top(W_0)(W_0(I_n - \lambda_0 W_0)^{-1}\varepsilon_t)$, let us compute the variance,

$$A_z = \frac{\sigma_0^2 \sum_{t=1}^{T} Z_t^\top(W_0)(W_0(I_n - \lambda_0 W_0)^{-1}(I_n - \lambda_0 W_0)^{-1\top}W_0^\top Z_t(W_0)}{n^2T^2}.$$

Assume that the column sums $\sum_{i=1}^{n}|W_{0,im}| \le c_n$. From Assumption 1 and its remark we have $\|S_0\|_1 = O(\sqrt{n})$, and moreover, from Assumption 1, $\|L_0\|_2 \le \sqrt{m_r(L_0)m_c(L_0)}\|L_0\|_{\max}$ is bounded and we have $\|L_0\|_1 = O(\sqrt{n}\|L_0\|_2)$. So we can deduce that $\|W_0\|_1 = O(\sqrt{n})$. Since we have

$$\begin{aligned}\|A_z\|_\infty &\le \frac{\sigma_0^2 \sum_{t=1}^{T} \|Z_t^\top(W_0)W_0(I_n - \lambda_0 W_0)^{-1}\|_\infty \|(I_n - \lambda_0 W_0)^{-1\top}W_0^\top Z_t(W_0)\|_\infty}{n^2T^2} \\ &\le \frac{\sigma_0^2 \sum_{t=1}^{T} \|Z_t^\top(W_0)\|_\infty \|W_0(I_n - \lambda_0 W_0)^{-1}\|_\infty \|(I_n - \lambda_0 W_0)^{-1\top}W_0^\top Z_t(W_0)\|_\infty}{n^2T^2},\end{aligned}$$

for $d = 0, 1, 2$.

To fill in details since there are finite number of columns of W_0 to have unbounded column sums of order bounded by c_n, we have by Lemma 7 with m and M being constants,

$\|X_t^\top W_0^{l\top}\|_\infty \lesssim \|X_t^\top\|_{\max} c_n + \|X_t^\top\|_\infty \lesssim n$, and

$$\|(I_n - \lambda_0 W_0)^{-1\top} W_0^\top Z_t(W_0)\|_\infty \lesssim \|(I_n - \lambda_0 W_0)^{-1\top} W_0^\top\|_\infty \|Z_t(W_0)\|_\infty \lesssim c_n,$$

$$\|A_z\|_\infty \leq \frac{\sigma_0^2 \sum_{t=1}^T \|Z_t^\top(W_0)\|_\infty \|(I_n - \lambda_0 W_0)^{-1\top} W_0^\top Z_t(W_0)\|_\infty}{n^2 T^2} \lesssim \frac{nTc_n}{n^2T^2}.$$

Thus to ensure that $\|A_z\|_\infty \to 0$ we need to have $\frac{c_n}{nT} \to 0$ to ensure $\mathcal{R}_2 \to_p 0$. It shall be noted that this is already ensured by our Assumptions 1. It is negligible since its 2-norm is of order $O_p(\sqrt{c_n/(nT)}) = O_p(1/(n^{1/4}\sqrt{T})) = o_p(1)$.
By the linearity of the moment condition, we have

$$\bar{g}_{nT}(\widehat{\theta}_p(W_0), W_0) = \bar{g}_{nT}(\theta_0, W_0) + \widehat{G}(W_0)(\widehat{\theta}_p(W_0) - \theta_0).$$

Hence the first order condition yields that

$$\mathbf{0} = \widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \bar{g}_{nT}(\widehat{\theta}_p(W_0), W_0) = \widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \bar{g}_{nT}(\theta_0, W_0) + \widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \widehat{G}(W_0)(\widehat{\theta}_p(W_0) - \theta_0),$$

or equivalently,

$$\widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \widehat{G}(W_0)(\widehat{\theta}_p(W_0) - \theta_0) = -\widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \bar{g}_{nT}(\theta_0, W_0).$$

By Assumption 6 and 7, we conclude that $\widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \to_p -\Sigma_{Z_0Q_0}^\top \Lambda^{-1}$, and

$$\widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \widehat{G}(W_0) \to_p \Sigma_{Z_0Q_0}^\top \Lambda^{-1} \Sigma_{Z_0Q_0},$$

which is nonsingular as Assumption 6 assumes $\Sigma_{Z_0Q_0}$ is of full column rank and Assumption 7 assumes Λ is a positive definite matrix.

We shall first prove the consistency. For convenience we introduce the vector $O_p(.)$ notation. Let a_n be a positive sequence. A finite dimension vector $A_n = O_p(a_n)$ means that $\|A_n/a_n\| = O_p(1)$ for any norm $\|.\|$. Thus

$$\widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \widehat{G}(W_0)(\widehat{\theta}_p(W_0) - \theta_0) = -\widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \bar{g}_{nT}(\theta_0, W_0),$$

$$[\Sigma_{Z_0Q_0}^\top \Lambda^{-1} \Sigma_{Z_0Q_0}](\widehat{\theta}_p(W_0) - \theta_0) + o_p(\|(\widehat{\theta}_p(W_0) - \theta_0)\|_2) = -\widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \bar{g}_{nT}(\theta_0, W_0).$$

By Assumptions 6 and 7, we have that

$$-\widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \bar{g}_{nT}(\theta_0, W_0) = -G(W_0)^\top \Lambda^{-1} \bar{g}_{nT}(\theta_0, W_0) + O_p(\|\bar{g}_{nT}(\theta_0, W_0)\|_2).$$

By Lemma 10 and the above derivation, we have that ,

$$\| - \widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \bar{g}_{nT}(\theta_0, W_0)\|_2 = O_p(\frac{1}{\sqrt{nT}}).$$

Thus the consistency follows, $\|\widehat{\theta}_p(W_0) - \theta_0\|_2 = o_p(1)$.

Furthermore, the variance matrix $\mathrm{Var}(\sqrt{nT}\bar{g}_{nT}(\theta_0, W_0)) = \mathrm{Var}(\sum_{t=1}^T Z_t^\top(W_0)\varepsilon_t/\sqrt{nT})$ asymptotically converges to $\sigma_0^2 \Sigma_{Z_0Z_0}$, i.e.

$$\sqrt{nT}\sigma_0^{-1} \Sigma_{Z_0Z_0}^{-1/2} \bar{g}_{nT}(\theta_0, W_0) \xrightarrow{d} \mathbb{N}(\mathbf{0}, I_m). \tag{E.40}$$

We shall try to understand the asymptotic of the above term by apply Lemma 6.

Compute the covariance to ensure finite variances:

$$\text{Cov}(Z_t^\top(W_0)\varepsilon_t/\sqrt{nT}) = \mathbb{E}\left[\frac{1}{nT}Z_t^\top(W_0)\varepsilon_t\varepsilon_t^\top Z_t(W_0)\right].$$

Since by Assumption 5 the elements of ε_t are i.i.d. with $\mathbb{E}[\varepsilon_t\varepsilon_t^\top] = \sigma_0^2 I_n$:

$$\text{Cov}(Z_t^\top(W_0)\varepsilon_t/\sqrt{nT}) = \frac{\sigma_0^2}{nT}\mathbb{E}[Z_t^\top(W_0)Z_t(W_0)].$$

Without loss of generality, we prove the result assuming no collinearity (i.e., the vectors are linearly independent). $m = 3K$. The proof can be extended to cases with collinearity. Compute $Z_t^\top(W_0)Z_t(W_0)$:

$$Z_t(W_0) = (X_t, W_0X_t, W_0^2X_t),$$

$$Z_t^\top(W_0)Z_t(W_0) = \begin{pmatrix} X_t^\top X_t & X_t^\top W_0X_t & X_t^\top W_0^2X_t \\ (W_0X_t)^\top X_t & (W_0X_t)^\top W_0X_t & (W_0X_t)^\top W_0^2X_t \\ (W_0^2X_t)^\top X_t & (W_0^2X_t)^\top W_0X_t & (W_0^2X_t)^\top W_0^2X_t \end{pmatrix}.$$

Since by our Assumption 6, $\frac{1}{nT}\sum_{t=1}^T \mathbb{E}[Z_t^\top(W_0)Z_t(W_0)]$ converge to $\Sigma_{Z_0Z_0}$. Thus condition i) in Lemma 6 is verified.

To verify the Lindeberg condition (condition ii)) for Lemma 6 applied to $\bar{g}_{nT}(\theta_0, W_0) = \frac{1}{nT}\sum_{t=1}^T Z_t^\top(W_0)\varepsilon_t$. Without loss of generality, we consider $2+\delta$ moment with $\delta = 1$. We thus compute the third moment of $Y_{n,it} = \frac{Z_{t,i}(W_0)\varepsilon_{it}}{\sqrt{nT}}$, a scalar component of $\frac{Z_{t,ij}(W_0)\varepsilon_{it}}{\sqrt{nT}}$, with $Z_{t,i}(W_0)$ the i-th row of $Z_t(W_0) = (X_t, W_0X_t, W_0^2X_t)$, X_t an $n \times K$ matrix, and $\varepsilon_{it} \sim (0, \sigma_0^2)$ with $\mathbb{E}[|\varepsilon_{it}|^3] = \mu_3 < \infty$. The matrix W_0 has row sums $\sum_{m=1}^n |W_{0,im}| \leq 1$ and column sums $\sum_{i=1}^n |W_{0,im}| \leq c_n$. However since the maximum absolute elements of W_0 is 1. For component j, we calculate $\mathbb{E}[|Y_{n,i,t}|^3] = \frac{\mathbb{E}[|Z_{t,ij}(W_0)|^3]\mu_3}{(nT)^{3/2}}$.

For the first block ($j = 1, \ldots, K$), $Z_{t,ij}(W_0) = X_{t,ij}$, so $\mathbb{E}[|(Z_{t,ij}(W_0))|^3] = O(1)$.

For the second block ($j = K + p = K + 1, \ldots, 2K$), $(Z_t(W_0))_{i,K+p} = \sum_{m=1}^n W_{0,im}X_{t,mp}$, and row sums ≤ 1 give $\mathbb{E}[|(Z_t(W_0))_{i,K+p}|^3] \leq \max_m[|X_{t,mp}|^3] = O(1)$.

For the third block ($j = 2K + p = 2K + 1, \ldots, 3K$), $(Z_t(W_0))_{i,2K+p} = \sum_{m=1}^n W_{0,im}^2X_{t,mp}$, with only for mth column of W_0 with infinite sums we need to enforce the following bound.

$$|W_{0,im}^2| \leq \sum_{r=1}^n |W_{0,ir}||W_{0,rm}| \leq \sum_{r=1}^n |W_{0,ir}| \leq 1.$$

Then use the fact that

$$|\sum_{m=1}^n (W_{0,im}^2)X_{t,mp}| \leq \sum_{r=1}^n |W_{0,ir}| \max_{1\leq r\leq n} \sum_{m=1}^n |W_{0,rm}| \max_{1\leq m\leq n, 1\leq p\leq K} |X_{t,mp}|.$$

So $\mathbb{E}[|(Z_t(W_0))_{i,2K+p}|^3] \leq \mathbb{E}[|(X_t)_{mp}|^3] = O(1)$. Thus, $\mathbb{E}[|Y_{n,i,t}|^3] = O\left(\frac{1}{n^{3/2}T^{3/2}}\right)$, and the Lindeberg term is $\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}[|Y_{n,i,t}|^3] = O\left(\frac{1}{(nT)^{1/2}}\right) = o(1)$. Hence we do not need any further restrictions on c_n for verifying this condition.

Define the variance covariance matrix

$$\Sigma_{GMM} = (\Sigma_{Z_0Q_0}^\top \Lambda^{-1}\Sigma_{Z_0Q_0})^{-1}\Sigma_{Z_0Q_0}^\top \Lambda^{-1}(\sigma_0^2\Sigma_{Z_0Z_0})\Lambda^{-1}\Sigma_{Z_0Q_0}(\Sigma_{Z_0Q_0}^\top \Lambda^{-1}\Sigma_{Z_0Q_0})^{-1}$$

Together with equation (E.40), we have

$$\begin{aligned}\sqrt{nT}\,\Sigma_{GMM}^{-1/2}(\widehat{\theta}_p(W_0) - \theta_0) &= -\Sigma_{GMM}^{-1/2}\big(G(W_0)^\top\Lambda^{-1}G(W_0)\big)^{-1}G(W_0)^\top\Lambda^{-1}\sqrt{nT}\,\bar{g}_{nT}(\theta_0, W_0) + o_p(1)\\ &\xrightarrow{d} \mathbb{N}(\mathbf{0}, I_{2K+1}).\end{aligned}$$

Step 2

Second, we study the deviation between $\widehat{\theta}_p(\widehat{W})$ and $\widehat{\theta}_p(W_0)$ to establish the large-sample performance of $\widehat{\theta}_p(\widehat{W})$. By the linearity of the sample moments,

$$\widehat{\theta}_p(\widehat{W}) - \theta_0 = -\left(\widehat{G}(\widehat{W})^\top\widehat{\Lambda}_{nT}^{-1}\widehat{G}(\widehat{W})\right)^{-1}\widehat{G}(\widehat{W})^\top\widehat{\Lambda}_{nT}^{-1}\bar{g}_{nT}(\theta_0, \widehat{W}).$$

Decompose

$$\big(\widehat{\theta}_p(\widehat{W}) - \theta_0\big) - \big(\widehat{\theta}_p(W_0) - \theta_0\big) = \mathcal{R}_{1,nT} + \mathcal{R}_{2,nT} + \mathcal{R}_{3,nT},$$

where

$$\begin{aligned}\mathcal{R}_{1,nT} &= \Big[-\left(\widehat{G}(\widehat{W})^\top\widehat{\Lambda}_{nT}^{-1}\widehat{G}(\widehat{W})\right)^{-1} + \left(\widehat{G}(W_0)^\top\widehat{\Lambda}_{nT}^{-1}\widehat{G}(W_0)\right)^{-1}\Big]\\ &\quad\times\widehat{G}(\widehat{W})^\top\widehat{\Lambda}_{nT}^{-1}\bar{g}_{nT}(\theta_0, \widehat{W}),\\ \mathcal{R}_{2,nT} &= \left(\widehat{G}(W_0)^\top\widehat{\Lambda}_{nT}^{-1}\widehat{G}(W_0)\right)^{-1}\\ &\quad\times\Big[-\widehat{G}(\widehat{W})^\top + \widehat{G}(W_0)^\top\Big]\widehat{\Lambda}_{nT}^{-1}\bar{g}_{nT}(\theta_0, \widehat{W}),\\ \mathcal{R}_{3,nT} &= \left(\widehat{G}(W_0)^\top\widehat{\Lambda}_{nT}^{-1}\widehat{G}(W_0)\right)^{-1}\widehat{G}(W_0)^\top\widehat{\Lambda}_{nT}^{-1}\\ &\quad\times\Big[\bar{g}_{nT}(\theta_0, W_0) - \bar{g}_{nT}(\theta_0, \widehat{W})\Big].\end{aligned}$$

We apply Lemma 10(c) to bound each remainder. Recall $\mathcal{R}_{nT}^*$ from (E.11). Lemma 10(c) gives

$$\|\widehat{G}(\widehat{W}) - \widehat{G}(W_0)\|_2 = O_p(\mathcal{R}_{nT}^*), \tag{E.41}$$

$$\|\bar{g}_{nT}(\theta_0, \widehat{W}) - \bar{g}_{nT}(\theta_0, W_0)\|_2 = O_p(\mathcal{R}_{nT}^*). \tag{E.42}$$

For $\mathcal{R}_{3,nT}$, since

$$\left\| \left(\widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \widehat{G}(W_0) \right)^{-1} \right\|_2 = O_p(1)$$

and

$$\left\| \widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \right\|_2 = O_p(1),$$

Equation (E.45) gives

$$\|\mathcal{R}_{3,nT}\|_2 = O_p(\mathcal{R}_{nT}^*).$$

For $\mathcal{R}_{2,nT}$, Step 1 and Equation (E.45) give

$$\|\bar{g}_{nT}(\theta_0, \widehat{W})\|_2 \leq \|\bar{g}_{nT}(\theta_0, W_0)\|_2 + \|\bar{g}_{nT}(\theta_0, \widehat{W}) - \bar{g}_{nT}(\theta_0, W_0)\|_2 = O_p\left(\frac{1}{\sqrt{nT}} + \mathcal{R}_{nT}^* \right).$$

Together with Equation (E.44), this yields

$$\|\mathcal{R}_{2,nT}\|_2 = O_p\left[\mathcal{R}_{nT}^* \left(\frac{1}{\sqrt{nT}} + \mathcal{R}_{nT}^* \right) \right] = o_p(\mathcal{R}_{nT}^*),$$

provided $\mathcal{R}_{nT}^* = o(1)$.

For $\mathcal{R}_{1,nT}$, define

$$\widehat{A}_{nT} = \widehat{G}(\widehat{W})^\top \widehat{\Lambda}_{nT}^{-1} \widehat{G}(\widehat{W}), \qquad A_{0,nT} = \widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \widehat{G}(W_0).$$

Equation (E.44) implies

$$\|\widehat{A}_{nT} - A_{0,nT}\|_2 = O_p(\mathcal{R}_{nT}^*).$$

Using the inverse identity

$$\widehat{A}_{nT}^{-1} - A_{0,nT}^{-1} = \widehat{A}_{nT}^{-1}(A_{0,nT} - \widehat{A}_{nT})A_{0,nT}^{-1},$$

together with

$$\|\widehat{A}_{nT}^{-1}\|_2 + \|A_{0,nT}^{-1}\|_2 = O_p(1),$$

we obtain

$$\|\widehat{A}_{nT}^{-1} - A_{0,nT}^{-1}\|_2 = O_p(\mathcal{R}_{nT}^*).$$

Therefore,

$$\|\mathcal{R}_{1,nT}\|_2 = O_p\left[\mathcal{R}_{nT}^* \left(\frac{1}{\sqrt{nT}} + \mathcal{R}_{nT}^* \right) \right] = o_p(\mathcal{R}_{nT}^*),$$

provided $\mathcal{R}_{nT}^* = o(1)$.

Combining the three terms and recalling from Step 1 that

$$\|\widehat{\theta}_p(W_0) - \theta_0\|_2 = O_p\left(\frac{1}{\sqrt{nT}} \right),$$

we obtain

$$\begin{aligned} \|\widehat{\theta}_p(\widehat{W}) - \theta_0\|_2 &\leq \|\widehat{\theta}_p(W_0) - \theta_0\|_2 + \|\mathcal{R}_{1,nT}\|_2 + \|\mathcal{R}_{2,nT}\|_2 + \|\mathcal{R}_{3,nT}\|_2 \\ &= O_p\left(\frac{1}{\sqrt{nT}} + \mathcal{R}_{nT}^* \right). \end{aligned}$$

By the definition in (E.11),

$$\mathcal{R}^*_{nT} = (1 + \|B_0\|_2)\kappa_n^{d+1} \frac{(r \vee 1)\rho_n + s_n}{n},$$

where

$$\rho_n = w_{2n} + \mathcal{R}_s, \qquad s_n = \mathcal{R}_s \sqrt{m_s(S_0)}, \qquad \kappa_n = 1 + \|W_0\|_2 + \rho_n.$$

Thus the preceding display retains all spectral factors without requiring them to be uniformly bounded.

Under the boundedness conditions in Theorem 2, and when $\rho_n = O(1)$, we have $\kappa_n = O(1)$ and hence

$$\mathcal{R}^*_{nT} = O\left(\frac{(r \vee 1)(w_{2n} + \mathcal{R}_s) + \mathcal{R}_s \sqrt{m_s(S_0)}}{n} \right).$$

Rate comparison. In contrast, using the noisy matrix $W = W_0 + E$ directly without denoising, Lemma 10(a) gives

$$\|\widehat{\theta}_p(W) - \theta_0\|_2 = O_p\left(\frac{1}{\sqrt{nT}} + w_{2n} + w_{2n}^2 \right).$$

Therefore, the denoised estimator has a strictly smaller plug-in error whenever

$$\mathcal{R}^*_{nT} = o(w_{2n} + w_{2n}^2).$$

Equivalently,

$$(1 + \|B_0\|_2)\kappa_n^{d+1}\big\{(r \vee 1)\rho_n + s_n\big\} = o\left(n(w_{2n} + w_{2n}^2)\right).$$

Finally, if

$$\mathcal{R}^*_{nT} = o\left(\frac{1}{\sqrt{nT}} \right),$$

then

$$\sqrt{nT}\left(\widehat{\theta}_p(\widehat{W}) - \widehat{\theta}_p(W_0) \right) = o_p(1).$$

Combining this result with Step 1 and Slutsky's theorem gives

$$\sqrt{nT}\, \Sigma_{GMM}^{-1/2}\left(\widehat{\theta}_p(\widehat{W}) - \theta_0 \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I_{2K+1}).$$

□

E.6. **Proof of Proposition 1.**

Proof. Recall that

$$\bar{g}_{nT}(\theta, W, M) = \frac{1}{nT} \sum_{t=1}^{T} Z_t(M)^\top \left(Y_t - \lambda W Y_t - X_t\beta - W X_t \gamma \right).$$

and $Z_t(M)$ is of the form $M^\ell X_t$ for $\ell = 0, 1, 2$. Moreover, $\mathbb{E}\{\widehat{G}(M)\} = G(M)$. Define $\widehat{\theta}_p(\widehat{W}, M)$ as the plug-in estimator using $Z_t(M)$:

$$\widehat{\theta}_p(\widehat{W}, M) = \left[\widetilde{X}(\widehat{W})^\top Z(M)\widehat{\Lambda}_{nT}^{-1} Z(M)^\top \widetilde{X}(\widehat{W})\right]^{-1} \widetilde{X}(\widehat{W})^\top Z(M)\widehat{\Lambda}_{nT}^{-1} Z(M)^\top Y. \tag{E.43}$$

We compare with the estimator using W:

$$\widehat{\theta}_p(W, M) = \widehat{\theta}_{GMM} = \left[\widetilde{X}(W)^\top Z(M)\widehat{\Lambda}_{nT}^{-1} Z(M)^\top \widetilde{X}(W)\right]^{-1} \widetilde{X}(W)^\top Z(M)\widehat{\Lambda}_{nT}^{-1} Z(M)^\top Y.$$

By the first-order condition for $\widehat{\theta}_p(\widehat{W}, M)$ and linearity of the moment function in θ, we have the exact expansion

$$\widehat{\theta}_p(\widehat{W}, M) - \theta_0 = -\left[\widehat{G}(\widehat{W}, M)^\top \widehat{\Lambda}_{nT}^{-1} \widehat{G}(\widehat{W}, M)\right]^{-1} \widehat{G}(\widehat{W}, M)^\top \widehat{\Lambda}_{nT}^{-1}\, \bar{g}_{nT}(\theta_0, \widehat{W}, M).$$

Under Assumptions 1–7, $\widehat{G}(\widehat{W}, M) \to_p -\Sigma_{Z_0 Q_0}(M)$ and $\widehat{\Lambda}_{nT} \to_p \Lambda$, so that

$$\widehat{\theta}_p(\widehat{W}, M) - \theta_0 = [\Sigma_{Z_0 Q_0}(M)^\top \Lambda^{-1} \Sigma_{Z_0 Q_0}(M)]^{-1} \Sigma_{Z_0 Q_0}(M)^\top \Lambda^{-1}\, \bar{g}_{nT}(\theta_0, \widehat{W}, M)\,(1 + o_p(1)).$$

The deviation from the limiting form is o_p relative to $\|\bar{g}_{nT}(\theta_0, \widehat{W}, M)\|_2$, so the rate of $\|\widehat{\theta}_p(\widehat{W}, M) - \theta_0\|_2$ matches that of $\|\bar{g}_{nT}(\theta_0, \widehat{W}, M)\|_2$ up to multiplicative constants. Let $B_0 := (I_n - \lambda_0 W_0)^{-1}$ and $\mathcal{R}_s := (w_{\max} n + 1/\sqrt{m_r(L_0) m_c(L_0)})\sqrt{m_s(S_0)}$.

Rate of the GMM estimator (using W). Let Z_{jt} be the jth column of the $n \times m$ matrix $Z_t(M)$. The residual at θ_0 when using W is $\varepsilon_t(\theta_0, W) = \varepsilon_t - \lambda_0 E Y_t - E X_t \gamma_0$, where $E = W - W_0$. Substituting $Y_t = B_0(X_t\beta_0 + W_0 X_t \gamma_0 + \varepsilon_t)$, the jth element of $\bar{g}_{nT}(\theta_0, W, M)$ is

$$\begin{aligned}
&\frac{1}{nT}\sum_{t=1}^{T} Z_{jt}^\top(M)[\varepsilon_t - \lambda_0 E Y_t - E X_t \gamma_0] \\
&\quad = \frac{1}{nT}\sum_{t=1}^{T} Z_{jt}^\top(M)\varepsilon_t - \frac{\lambda_0}{nT}\sum_{t=1}^{T} Z_{jt}^\top(M) E B_0(X_t\beta_0 + W_0 X_t\gamma_0) \\
&\qquad - \frac{1}{nT}\sum_{t=1}^{T} Z_{jt}^\top(M) E X_t \gamma_0 - \frac{\lambda_0}{nT}\sum_{t=1}^{T} Z_{jt}^\top(M) E B_0 \varepsilon_t \\
&\quad =: O_p(1/\sqrt{nT}) + \mathcal{R}_{1,j} + \mathcal{R}_{3,j} + \mathcal{R}_{2,j}.
\end{aligned}$$

For each i, let

$$E_{i,-i} = (E_{i1}, \ldots, E_{i,i-1}, E_{i,i+1}, \ldots, E_{in})^\top \in \mathbb{R}^{n-1}$$

denote the vector of off-diagonal entries in row i of E. Let

$$\widetilde{\Sigma}_E = \sigma_E^2 \Sigma_E$$

and define

$$\delta_n = \left(n^{-2s}\widetilde{\Sigma}_E\right)^{-1} n^{-s}\sigma_{E\varepsilon},$$

where

$$\sigma_{E\varepsilon} = \sigma_0 \sigma_E \rho_{\varepsilon E}^\top \in \mathbb{R}^{n-1}.$$

By Assumption 8 and joint normality,

$$\mathbb{E}[\varepsilon_{it} \mid E] = E_{i,-i}^{\top}\delta_n.$$

Define

$$\mu_E = \left(E_{1,-1}^{\top}\delta_n, \ldots, E_{n,-n}^{\top}\delta_n\right)^{\top}.$$

Then

$$\varepsilon_t = \mu_E + \xi_t,$$

where ξ_t is independent of E and has conditional mean zero. Under the uniform eigenvalue bounds on Σ_E,

$$\|\mu_E\|_2 = O_p\big(\sqrt{n}\,\|\rho_{\varepsilon E}\|_2\big).$$

Decompose

$$\begin{aligned}\mathcal{R}_{2,j} = &-\frac{\lambda_0}{nT}\sum_{t=1}^{T} Z_{jt}(M)^{\top} E B_0 \mu_E \\ &-\frac{\lambda_0}{nT}\sum_{t=1}^{T} Z_{jt}(M)^{\top} E B_0 \xi_t.\end{aligned}$$

The first term is the endogeneity-bias component, whereas the second is conditionally centered. Using the bounds in Assumption 8, we obtain

$$\max_j |\mathcal{R}_{2,j}| \lesssim_p n^{1/2-s}\|\rho_{\varepsilon E}\|_2 + \|B_0\|_2 w_{2n}.$$

Then

$$\begin{aligned}\max_j |\mathcal{R}_{1,j}| &\le |\lambda_0|\frac{1}{nT}\sum_t \|Z_{jt}(M)\|_2 \|B_0\|_2 \|X_t\beta_0 + W_0 X_t\gamma_0\|_2 \|E\|_2 \\ &\lesssim_p \|B_0\|_2(1+\|W_0\|_2)\, w_{2n}, \\ \max_j |\mathcal{R}_{3,j}| &\le \frac{1}{nT}\sum_t \|Z_{jt}(M)\|_2 \|E\|_2 \|X_t\gamma_0\|_2 \lesssim_p w_{2n},\end{aligned}$$

Hence

$$\|\bar{g}_{nT}(\theta_0, W, M)\|_2 = O_p\big(n^{1/2-s}\|\rho_{\varepsilon E}\|_2 + w_{2n}\|B_0\|_2(1+\|W_0\|_2) + 1/\sqrt{nT}\big),$$

yielding

$$\|\widehat{\theta}_{GMM}(W, M) - \theta_0\|_2 \lesssim_p n^{1/2-s}\|\rho_{\varepsilon E}\|_2 + \|B_0\|_2(1+\|W_0\|_2)\, w_{2n} + 1/\sqrt{nT},$$

which under the boundedness assumption $\|B_0\|_2(1+\|W_0\|_2) = O(1)$ gives the rate (13).

Step 2: Effect of replacing W_0 by $\widehat{W}$.

We now study the deviation between $\widehat{\theta}_p(\widehat{W})$ and $\widehat{\theta}_p(W_0)$. Because the sample moment is linear in θ, the first-order condition gives the exact representation

$$\widehat{\theta}_p(\widehat{W}) - \theta_0 = -\left[\widehat{G}(\widehat{W})^{\top}\widehat{\Lambda}_{nT}^{-1}\widehat{G}(\widehat{W})\right]^{-1}\widehat{G}(\widehat{W})^{\top}\widehat{\Lambda}_{nT}^{-1}\bar{g}_{nT}(\theta_0, \widehat{W}).$$

Similarly,

$$\widehat{\theta}_p(W_0) - \theta_0 = -\left[\widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \widehat{G}(W_0)\right]^{-1} \widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \bar{g}_{nT}(\theta_0, W_0).$$

Define

$$\widehat{A}_{nT} := \widehat{G}(\widehat{W})^\top \widehat{\Lambda}_{nT}^{-1} \widehat{G}(\widehat{W}), \qquad A_{0,nT} := \widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \widehat{G}(W_0).$$

Then

$$\left\{\widehat{\theta}_p(\widehat{W}) - \theta_0\right\} - \left\{\widehat{\theta}_p(W_0) - \theta_0\right\} = \mathcal{R}_{1,nT} + \mathcal{R}_{2,nT} + \mathcal{R}_{3,nT},$$

where

$$\mathcal{R}_{1,nT} = \left(-\widehat{A}_{nT}^{-1} + A_{0,nT}^{-1}\right) \widehat{G}(\widehat{W})^\top \widehat{\Lambda}_{nT}^{-1} \bar{g}_{nT}(\theta_0, \widehat{W}),$$

$$\mathcal{R}_{2,nT} = A_{0,nT}^{-1} \left[-\widehat{G}(\widehat{W})^\top + \widehat{G}(W_0)^\top\right] \widehat{\Lambda}_{nT}^{-1} \bar{g}_{nT}(\theta_0, \widehat{W}),$$

$$\mathcal{R}_{3,nT} = A_{0,nT}^{-1} \widehat{G}(W_0)^\top \widehat{\Lambda}_{nT}^{-1} \left[\bar{g}_{nT}(\theta_0, W_0) - \bar{g}_{nT}(\theta_0, \widehat{W})\right].$$

Recall the rate defined in (E.11):

$$\mathcal{R}_{nT}^* = (1 + \|B_0\|_2)\kappa_n^{d+1} \frac{(r \vee 1)\omega_n + s_n}{n},$$

where

$$\omega_n = w_{2n} + \mathcal{R}_s, \qquad s_n = \mathcal{R}_s \sqrt{m_s(S_0)}, \qquad \kappa_n = 1 + \|W_0\|_2 + \omega_n.$$

Lemma 10(c) gives

$$\|\widehat{G}(\widehat{W}) - \widehat{G}(W_0)\|_2 = O_p(\mathcal{R}_{nT}^*), \tag{E.44}$$

$$\|\bar{g}_{nT}(\theta_0, \widehat{W}) - \bar{g}_{nT}(\theta_0, W_0)\|_2 = O_p(\mathcal{R}_{nT}^*). \tag{E.45}$$

By Step 1,

$$\|\bar{g}_{nT}(\theta_0, W_0)\|_2 = O_p\left(\frac{1}{\sqrt{nT}}\right), \qquad \|\widehat{G}(W_0)\|_2 = O_p(1),$$

and

$$A_{0,nT} \to_p G_0^\top \Lambda^{-1} G_0.$$

Since G_0 has full column rank and Λ^{-1} is positive definite,

$$\|A_{0,nT}^{-1}\|_2 = O_p(1).$$

Moreover, Equation (E.44) gives

$$\|\widehat{A}_{nT} - A_{0,nT}\|_2 = O_p(\mathcal{R}_{nT}^*).$$

If $\mathcal{R}_{nT}^* = o(1)$, Weyl's inequality implies

$$\lambda_{\min}(\widehat{A}_{nT}) \geq \lambda_{\min}(A_{0,nT}) - \|\widehat{A}_{nT} - A_{0,nT}\|_2,$$

and hence

$$\|\widehat{A}_{nT}^{-1}\|_2 + \|A_{0,nT}^{-1}\|_2 = O_p(1).$$

Using the inverse identity

$$\widehat{A}_{nT}^{-1} - A_{0,nT}^{-1} = \widehat{A}_{nT}^{-1}\left(A_{0,nT} - \widehat{A}_{nT}\right)A_{0,nT}^{-1},$$

we obtain

$$\|\widehat{A}_{nT}^{-1} - A_{0,nT}^{-1}\|_2 = O_p(\mathcal{R}_{nT}^*). \tag{E.46}$$

Next, Equation (E.45) and Step 1 imply

$$\begin{aligned}\|\bar{g}_{nT}(\theta_0, \widehat{W})\|_2 &\leq \|\bar{g}_{nT}(\theta_0, W_0)\|_2 + \|\bar{g}_{nT}(\theta_0, \widehat{W}) - \bar{g}_{nT}(\theta_0, W_0)\|_2 \\ &= O_p\left(\frac{1}{\sqrt{nT}} + \mathcal{R}_{nT}^*\right).\end{aligned} \tag{E.47}$$

For $\mathcal{R}_{3,nT}$, Equations (E.45) and the boundedness of the multiplying matrices give

$$\|\mathcal{R}_{3,nT}\|_2 = O_p(\mathcal{R}_{nT}^*).$$

For $\mathcal{R}_{2,nT}$, Equations (E.44) and (E.47) give

$$\begin{aligned}\|\mathcal{R}_{2,nT}\|_2 &= O_p\left[\mathcal{R}_{nT}^*\left(\frac{1}{\sqrt{nT}} + \mathcal{R}_{nT}^*\right)\right] \\ &= o_p(\mathcal{R}_{nT}^*),\end{aligned}$$

provided $\mathcal{R}_{nT}^* = o(1)$.

Similarly, Equations (E.46) and (E.47) yield

$$\begin{aligned}\|\mathcal{R}_{1,nT}\|_2 &= O_p\left[\mathcal{R}_{nT}^*\left(\frac{1}{\sqrt{nT}} + \mathcal{R}_{nT}^*\right)\right] \\ &= o_p(\mathcal{R}_{nT}^*).\end{aligned}$$

Combining the three remainder bounds gives

$$\left\|\widehat{\theta}_p(\widehat{W}) - \widehat{\theta}_p(W_0)\right\|_2 = O_p(\mathcal{R}_{nT}^*).$$

Since Step 1 establishes

$$\|\widehat{\theta}_p(W_0) - \theta_0\|_2 = O_p\left(\frac{1}{\sqrt{nT}}\right),$$

we conclude that

$$\|\widehat{\theta}_p(\widehat{W}) - \theta_0\|_2 = O_p\left(\frac{1}{\sqrt{nT}} + \mathcal{R}_{nT}^*\right). \tag{E.48}$$

Rate comparison. Using the noisy matrix $W = W_0 + E$ directly, Lemma 10(a) and the same first-order-condition argument give

$$\|\widehat{\theta}_p(W) - \theta_0\|_2 = O_p\left(\frac{1}{\sqrt{nT}} + w_{2n} + w_{2n}^2\right),$$

provided the corresponding sample Jacobian remains nonsingular. Therefore, the denoised estimator has a strictly smaller network-error contribution whenever

$$\mathcal{R}^*_{nT} = o(w_{2n} + w_{2n}^2).$$

Asymptotic normality. Suppose, in addition, that

$$\sqrt{nT}\,\mathcal{R}^*_{nT} = o(1),$$

or equivalently,

$$\mathcal{R}^*_{nT} = o\left(\frac{1}{\sqrt{nT}}\right).$$

Then

$$\sqrt{nT}\left\{\widehat{\theta}_p(\widehat{W}) - \widehat{\theta}_p(W_0)\right\} = o_p(1).$$

Combining this result with the oracle central limit theorem from Step 1 and applying Slutsky's theorem yields

$$\sqrt{nT}\left(\widehat{\theta}_p(\widehat{W}) - \theta_0\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma_{GMM}).$$

Equivalently, the asymptotic-normality condition can be written as

$$(1 + \|B_0\|_2)\kappa_n^{d+1}\big\{(r \vee 1)\omega_n + s_n\big\} = o\left(\sqrt{\frac{n}{T}}\right).$$

When T is fixed, this reduces to

$$(1 + \|B_0\|_2)\kappa_n^{d+1}\big\{(r \vee 1)\omega_n + s_n\big\} = o(\sqrt{n}).$$

□

E.7. **Proof of Theorem 3.**

Proof. This step estimates the weight matrix $W_0 = L_0 + S_0$ jointly with the regression coefficient θ. Specifically, we minimize the objective in (8) using a block-coordinate procedure. When L and S are given, θ is estimated by standard GMM using the plug-in weight matrix $L + S$. When θ is given, minimization over (L, S) becomes a penalized regression problem involving a nuclear-norm penalty on L and an elementwise ℓ_1 penalty on the off-diagonal entries of S.

To streamline the exposition, we establish the result using the instrumental variable $Z_t(M)$ for an arbitrary fixed matrix M. The same proof applies when M is exogenous or $\sigma(E)$-measurable and satisfies the corresponding norm and moment conditions. Additional arguments are required when M is estimated using the same outcome data as the supervised estimator.

Throughout the proof, we impose

$$\operatorname{diag}(L + S) = 0,$$

so that all estimated adjacency matrices have zero diagonal.

Recall that

$$\bar{g}_{nT}(\theta, W, M) = \frac{1}{nT} \sum_{t=1}^{T} Z_t^\top(M) \left(Y_t - \lambda W Y_t - X_t\beta - W X_t \gamma\right).$$

For

$$\theta = (\lambda, \beta^\top, \gamma^\top)^\top,$$

define the $n^2 \times m$ matrix

$$\widetilde{X}_t(\theta, M) := \frac{1}{n}(\lambda Y_t + X_t\gamma) \otimes Z_t(M),$$

and let

$$\overline{\widetilde{X}}_{nT}(\theta, M) := \frac{1}{T} \sum_{t=1}^{T} \widetilde{X}_t(\theta, M).$$

At the true parameter, write

$$\widetilde{X}_t(M) := \widetilde{X}_t(\theta_0, M) = \frac{1}{n}(\lambda_0 Y_t + X_t\gamma_0) \otimes Z_t(M).$$

The vectorization identity

$$\operatorname{vec}\left(Z_t^\top(M) W (\lambda Y_t + X_t\gamma)\right) = \left[(\lambda Y_t + X_t\gamma) \otimes Z_t(M)\right]^\top \operatorname{vec}(W)$$

implies that, for fixed θ and M,

$$\bar{g}_{nT}(\theta, W_1, M) - \bar{g}_{nT}(\theta, W_2, M) = -\overline{\widetilde{X}}_{nT}(\theta, M)^\top \operatorname{vec}(W_1 - W_2). \tag{E.49}$$

Define

$$q_{nT}(\theta, W, M) := \bar{g}_{nT}(\theta, W, M)^\top \widehat{\Lambda}_{nT}^{-1} \bar{g}_{nT}(\theta, W, M)$$

and the adjacency-only objective

$$Q_n^0(L, S) := \frac{1}{2}\|W - L - S\|_F^2 + \nu_{nT}\|L\|_* + \tau_{nT} \sum_{i \neq j} |S_{ij}|.$$

The complete supervised objective can therefore be written as

$$Q_{nT}(\theta, L, S; M) = \xi_{nT} q_{nT}(\theta, L + S, M) + Q_n^0(L, S).$$

Let $(\widehat{L}^{[0]}, \widehat{S}^{[0]})$ be the adjacency-only estimator obtained by minimizing $Q_n^0(L, S)$, with the same tuning parameters ν_{nT} and τ_{nT} as those appearing in the supervised objective. Set

$$\widehat{W}^{[0]} := \widehat{L}^{[0]} + \widehat{S}^{[0]}.$$

Because $(\widehat{L}^{[0]}, \widehat{S}^{[0]})$ is constructed using only the observed adjacency matrix $W = W_0 + E$, it is measurable with respect to $\sigma(E)$.

Let

$$\omega_n := w_{2n} + \mathcal{R}_s, \qquad s_n := \mathcal{R}_s \sqrt{m_s(S_0)}.$$

By Theorem 1,

$$\begin{aligned} \|\widehat{L}^{[0]} - L_0\|_2 &= O_p(\omega_n), \\ \|\widehat{S}^{[0]} - S_0\|_2 &= O_p(\mathcal{R}_s), \\ \left|\mathrm{off}(\widehat{S}^{[0]} - S_0)\right|_{1,1} &= O_p(s_n). \end{aligned} \tag{E.50}$$

Thus, $\widehat{W}^{[0]}$ satisfies the conditions of Lemma 10(c).

The supervised estimator is obtained by the following block-coordinate procedure, initialized at $(\widehat{L}^{[0]}, \widehat{S}^{[0]})$.

a) Given $(\widehat{L}^{[0]}, \widehat{S}^{[0]})$, define

$$\widehat{\theta}^{[1]} = \arg\min_{\theta \in \Theta} q_{nT}(\theta, \widehat{W}^{[0]}, M).$$

b) Given $\widehat{\theta}^{[1]}$ and $\widehat{L}^{[0]}$, define

$$\widehat{S}^{[1]} = \arg \min_{\substack{S \in \mathbb{R}^{n \times n}: \\ \mathrm{diag}(\widehat{L}^{[0]} + S) = 0}} \Bigg\{ \xi_{nT} q_{nT}\big(\widehat{\theta}^{[1]}, \widehat{L}^{[0]} + S, M\big) + \frac{1}{2}\|W - \widehat{L}^{[0]} - S\|_F^2 + \tau_{nT} \sum_{i \neq j} |S_{ij}| \Bigg\}.$$

c) Given $\widehat{\theta}^{[1]}$ and $\widehat{S}^{[1]}$, define

$$\widehat{L}^{[1]} = \arg \min_{\substack{L \in \mathbb{R}^{n \times n}: \\ \mathrm{diag}(L + \widehat{S}^{[1]}) = 0}} \Bigg\{ \xi_{nT} q_{nT}\big(\widehat{\theta}^{[1]}, L + \widehat{S}^{[1]}, M\big) + \frac{1}{2}\|W - L - \widehat{S}^{[1]}\|_F^2 + \nu_{nT}\|L\|_* \Bigg\}.$$

d) Set

$$\widehat{W}^{[1]} := \widehat{L}^{[1]} + \widehat{S}^{[1]}$$

and define

$$\widehat{\theta}_s = \arg\min_{\theta \in \Theta} q_{nT}(\theta, \widehat{W}^{[1]}, M).$$

We now establish the rate of the resulting estimator.

Step a). The estimator $\widehat{\theta}^{[1]}$ is the plug-in GMM estimator based on the adjacency-only estimator $\widehat{W}^{[0]}$. Because $\widehat{W}^{[0]}$ is $\sigma(E)$-measurable and satisfies (E.50), Theorem 2, using Lemma 10(c), gives

$$\|\widehat{\theta}^{[1]} - \theta_0\|_2 = O_p\left(\frac{1}{\sqrt{nT}} + \mathcal{R}^*_{nT}\right). \tag{E.51}$$

For conciseness, define

$$a_{nT} := \frac{1}{\sqrt{nT}} + \mathcal{R}^*_{nT}.$$

Then

$$\|\widehat{\theta}^{[1]} - \theta_0\|_2 = O_p(a_{nT}).$$

Because $\widehat{\theta}^{[1]}$ minimizes $q_{nT}(\theta, \widehat{W}^{[0]}, M)$,

$$q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M) \leq q_{nT}(\theta_0, \widehat{W}^{[0]}, M). \tag{E.52}$$

The oracle sampling bound and Lemma 10(c) imply

$$\|\bar{g}_{nT}(\theta_0, \widehat{W}^{[0]}, M)\|_2 = O_p(a_{nT}).$$

Under Assumption 7, there exist constants $0 < c < C < \infty$ such that, with probability approaching one,

$$c\|v\|_2^2 \leq v^\top \widehat{\Lambda}_{nT}^{-1} v \leq C\|v\|_2^2$$

for every conformable vector v. Consequently,

$$\begin{aligned} c\|\bar{g}_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M)\|_2^2 &\leq q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M) \\ &\leq q_{nT}(\theta_0, \widehat{W}^{[0]}, M) \\ &\leq C\|\bar{g}_{nT}(\theta_0, \widehat{W}^{[0]}, M)\|_2^2. \end{aligned}$$

It follows that

$$\|\bar{g}_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M)\|_2 = O_p(a_{nT}). \tag{E.53}$$

We also verify the operator-norm bound used below. Let

$$U_t(\theta) := \lambda Y_t + X_t\gamma.$$

For each of the finitely many columns $Z_{t,\cdot j}(M)$ of $Z_t(M)$,

$$\left\| \frac{1}{n} U_t(\theta) \otimes Z_{t,\cdot j}(M) \right\|_2 = \frac{1}{n} \|U_t(\theta)\|_2 \|Z_{t,\cdot j}(M)\|_2.$$

By Lemma 9 and the moment conditions on the regressors and instruments, both norms on the right-hand side are $O_p(\sqrt{n})$, uniformly for θ in a neighborhood of θ_0. Since the number of instruments m is fixed,

$$\|\overline{\widetilde{X}}_{nT}(\theta_0, M)\|_2 = O_p(1).$$

Moreover, (E.51) gives $\widehat{\theta}^{[1]} - \theta_0 = o_p(1)$ under $a_{nT} = o(1)$. Therefore,

$$\|\overline{\widetilde{X}}_{nT}(\widehat{\theta}^{[1]}, M)\|_2 = O_p(1). \tag{E.54}$$

Steps b) and c). These two updates are performed with $\widehat{\theta}^{[1]}$ fixed. By the definition of $\widehat{S}^{[1]}$,

$$Q_{nT}\big(\widehat{\theta}^{[1]}, \widehat{L}^{[0]}, \widehat{S}^{[1]}; M\big) \leq Q_{nT}\big(\widehat{\theta}^{[1]}, \widehat{L}^{[0]}, \widehat{S}^{[0]}; M\big).$$

Similarly, by the definition of $\widehat{L}^{[1]}$,

$$Q_{nT}\big(\widehat{\theta}^{[1]}, \widehat{L}^{[1]}, \widehat{S}^{[1]}; M\big) \le Q_{nT}\big(\widehat{\theta}^{[1]}, \widehat{L}^{[0]}, \widehat{S}^{[1]}; M\big).$$

Combining these inequalities,

$$Q_{nT}\big(\widehat{\theta}^{[1]}, \widehat{L}^{[1]}, \widehat{S}^{[1]}; M\big) \le Q_{nT}\big(\widehat{\theta}^{[1]}, \widehat{L}^{[0]}, \widehat{S}^{[0]}; M\big). \tag{E.55}$$

Let

$$\Delta_L := \widehat{L}^{[1]} - \widehat{L}^{[0]}, \qquad \Delta_S := \widehat{S}^{[1]} - \widehat{S}^{[0]},$$

and define

$$H_{nT} := \widehat{W}^{[1]} - \widehat{W}^{[0]} = \Delta_L + \Delta_S.$$

We first establish a lower bound for the change in the adjacency-only objective. The function

$$Q_n^0(L, S) = \frac{1}{2}\|W - L - S\|_F^2 + \nu_{nT}\|L\|_* + \tau_{nT}\sum_{i\neq j}|S_{ij}|$$

is convex in (L, S) and is 1-strongly convex in the direction $L + S$. More precisely, for any two feasible pairs (L_1, S_1) and (L_2, S_2), the strong-convexity remainder in the quadratic part is

$$\frac{1}{2}\|(L_1 + S_1) - (L_2 + S_2)\|_F^2.$$

Because $(\widehat{L}^{[0]}, \widehat{S}^{[0]})$ minimizes Q_n^0 over the convex zero-diagonal feasible set, its first-order optimality condition and the strong convexity of the quadratic part imply

$$\begin{aligned} &Q_n^0(\widehat{L}^{[1]}, \widehat{S}^{[1]}) - Q_n^0(\widehat{L}^{[0]}, \widehat{S}^{[0]}) \\ &\quad\ge \frac{1}{2}\left\|(\widehat{L}^{[1]} + \widehat{S}^{[1]}) - (\widehat{L}^{[0]} + \widehat{S}^{[0]})\right\|_F^2 \\ &\quad= \frac{1}{2}\|H_{nT}\|_F^2. \end{aligned} \tag{E.56}$$

For fixed θ and M, equation (E.49) shows that $q_{nT}(\theta, W, M)$ is a convex quadratic function of $\mathrm{vec}(W)$. Its gradient is

$$\nabla_{\mathrm{vec}(W)} q_{nT}(\theta, W, M) = -2\overline{\widetilde{X}}_{nT}(\theta, M)\widehat{\Lambda}_{nT}^{-1}\bar{g}_{nT}(\theta, W, M). \tag{E.57}$$

The exact quadratic expansion around $\widehat{W}^{[0]}$ gives

$$\begin{aligned} &q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M) - q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[1]}, M) \\ &\quad= -\left\langle \nabla_{\mathrm{vec}(W)} q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M), \mathrm{vec}(H_{nT})\right\rangle \\ &\qquad - \mathrm{vec}(H_{nT})^\top \overline{\widetilde{X}}_{nT}(\widehat{\theta}^{[1]}, M)\widehat{\Lambda}_{nT}^{-1}\overline{\widetilde{X}}_{nT}(\widehat{\theta}^{[1]}, M)^\top \mathrm{vec}(H_{nT}). \end{aligned}$$

The final term is nonpositive. Therefore,

$$\begin{aligned} & q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M) - q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[1]}, M) \\ & \quad \leq \left| \left\langle \nabla_{\mathrm{vec}(W)} q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M), \mathrm{vec}(H_{nT}) \right\rangle \right| \\ & \quad \leq \left\| \nabla_{\mathrm{vec}(W)} q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M) \right\|_2 \|H_{nT}\|_F. \end{aligned} \tag{E.58}$$

On the other hand, the descent inequality (E.55) implies

$$\begin{aligned} & Q_n^0(\widehat{L}^{[1]}, \widehat{S}^{[1]}) - Q_n^0(\widehat{L}^{[0]}, \widehat{S}^{[0]}) \\ & \quad \leq \xi_{nT} \left[q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M) - q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[1]}, M) \right]. \end{aligned} \tag{E.59}$$

Combining (E.56), (E.58), and (E.59), we obtain

$$\frac{1}{2}\|H_{nT}\|_F^2 \leq \xi_{nT} \left\| \nabla_{\mathrm{vec}(W)} q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M) \right\|_2 \|H_{nT}\|_F.$$

Hence,

$$\|H_{nT}\|_F \leq 2\xi_{nT} \left\| \nabla_{\mathrm{vec}(W)} q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M) \right\|_2. \tag{E.60}$$

By (E.57), (E.53), (E.54), and $\|\widehat{\Lambda}_{nT}^{-1}\|_2 = O_p(1)$,

$$\begin{aligned} & \left\| \nabla_{\mathrm{vec}(W)} q_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M) \right\|_2 \\ & \quad \leq 2\|\overline{\widetilde{X}}_{nT}(\widehat{\theta}^{[1]}, M)\|_2 \|\widehat{\Lambda}_{nT}^{-1}\|_2 \|\bar{g}_{nT}(\widehat{\theta}^{[1]}, \widehat{W}^{[0]}, M)\|_2 \\ & \quad = O_p(a_{nT}). \end{aligned}$$

It follows from (E.60) that

$$\|\widehat{W}^{[1]} - \widehat{W}^{[0]}\|_F = O_p(\xi_{nT} a_{nT}). \tag{E.61}$$

This stability bound is important because $\widehat{W}^{[1]}$ depends on the outcome data and therefore is not $\sigma(E)$-measurable. We do not apply Lemma 10(c) directly to $\widehat{W}^{[1]}$. Instead, we apply that lemma only to the initial estimator $\widehat{W}^{[0]}$ and control the additional supervised update through (E.61).

The bound also implies the aggregate recovery inequality

$$\|\widehat{W}^{[1]} - W_0\|_F \leq \|\widehat{W}^{[1]} - \widehat{W}^{[0]}\|_F + \|\widehat{W}^{[0]} - W_0\|_F.$$

This inequality controls the aggregate adjacency estimator, but does not by itself give separate recovery rates for $\widehat{L}^{[1]}$ and $\widehat{S}^{[1]}$.

Step d). Define

$$\widehat{G}_s := \frac{\partial \bar{g}_{nT}(\theta, \widehat{W}^{[1]}, M)}{\partial \theta^\top}$$

and

$$\widehat{G}^{[0]} := \frac{\partial \bar{g}_{nT}(\theta, \widehat{W}^{[0]}, M)}{\partial \theta^\top}.$$

We first verify that the GMM Jacobian evaluated at $\widehat{W}^{[1]}$ remains well conditioned. Because the Jacobian is affine in W, the same design bounds used above give

$$\|\widehat{G}_s - \widehat{G}^{[0]}\|_2 \lesssim_p \|\widehat{W}^{[1]} - \widehat{W}^{[0]}\|_F = O_p(\xi_{nT} a_{nT}). \tag{E.62}$$

When $\xi_{nT} = O(1)$ and $a_{nT} = o(1)$, the right-hand side is $o_p(1)$.

Moreover, the fixed-M version of Lemma 10(c), applied only to the $\sigma(E)$-measurable estimator $\widehat{W}^{[0]}$, and the oracle law of large numbers give

$$\begin{aligned}\|\widehat{G}^{[0]} - G_0(M)\|_2 &\le \|\widehat{G}^{[0]} - \widehat{G}(W_0, M)\|_2 + \|\widehat{G}(W_0, M) - G_0(M)\|_2 \\ &= O_p(\mathcal{R}^*_{nT}) + o_p(1) = o_p(1).\end{aligned} \tag{E.63}$$

Combining (E.62) and (E.63), we obtain

$$\widehat{G}_s = G_0(M) + o_p(1).$$

By the identification condition,

$$\lambda_{\min}\left(G_0(M)^\top \Lambda^{-1} G_0(M)\right) > 0.$$

Since $\widehat{\Lambda}^{-1}_{nT} \to_p \Lambda^{-1}$, Weyl's inequality implies that, for some constant $c > 0$,

$$\lambda_{\min}\left(\widehat{G}_s^\top \widehat{\Lambda}^{-1}_{nT} \widehat{G}_s\right) \ge c + o_p(1).$$

Consequently,

$$\left\| \left(\widehat{G}_s^\top \widehat{\Lambda}^{-1}_{nT} \widehat{G}_s\right)^{-1} \widehat{G}_s^\top \widehat{\Lambda}^{-1}_{nT} \right\|_2 = O_p(1). \tag{E.64}$$

Because the moment function is affine in θ,

$$\bar{g}_{nT}(\widehat{\theta}_s, \widehat{W}^{[1]}, M) = \bar{g}_{nT}(\theta_0, \widehat{W}^{[1]}, M) + \widehat{G}_s(\widehat{\theta}_s - \theta_0).$$

The first-order condition for the final GMM step is

$$\widehat{G}_s^\top \widehat{\Lambda}^{-1}_{nT} \bar{g}_{nT}(\widehat{\theta}_s, \widehat{W}^{[1]}, M) = 0.$$

Therefore,

$$\widehat{\theta}_s - \theta_0 = -\left(\widehat{G}_s^\top \widehat{\Lambda}^{-1}_{nT} \widehat{G}_s\right)^{-1} \widehat{G}_s^\top \widehat{\Lambda}^{-1}_{nT} \bar{g}_{nT}(\theta_0, \widehat{W}^{[1]}, M). \tag{E.65}$$

Using (E.49),

$$\begin{aligned}&\bar{g}_{nT}(\theta_0, \widehat{W}^{[1]}, M) - \bar{g}_{nT}(\theta_0, \widehat{W}^{[0]}, M) \\ &\quad = -\overline{\widetilde{X}}_{nT}(\theta_0, M)^\top \operatorname{vec}(\widehat{W}^{[1]} - \widehat{W}^{[0]}).\end{aligned} \tag{E.66}$$

It follows that

$$\begin{aligned}\|\bar{g}_{nT}(\theta_0, \widehat{W}^{[1]}, M)\|_2 \leq \|\bar{g}_{nT}(\theta_0, \widehat{W}^{[0]}, M)\|_2 \\ + \|\overline{\widehat{\dot{X}}}_{nT}(\theta_0, M)\|_2 \|\widehat{W}^{[1]} - \widehat{W}^{[0]}\|_F \\ = O_p(a_{nT}) + O_p(\xi_{nT} a_{nT}).\end{aligned}$$

Combining this bound with (E.65) and (E.64), we obtain

$$\|\widehat{\theta}_s - \theta_0\|_2 = O_p(a_{nT}) + O_p(\xi_{nT} a_{nT}).$$

If $\xi_{nT} = O(1)$, it follows that

$$\|\widehat{\theta}_s - \theta_0\|_2 = O_p\left(\frac{1}{\sqrt{nT}} + \mathcal{R}^*_{nT}\right). \tag{E.67}$$

This proves the rate in (14).

Finally, we establish the asymptotic distribution. Define

$$\widehat{\Psi}(W) := \left[\widehat{G}(W)^\top \widehat{\Lambda}^{-1}_{nT} \widehat{G}(W)\right]^{-1} \widehat{G}(W)^\top \widehat{\Lambda}^{-1}_{nT}.$$

The GMM expansions for $\widehat{\theta}_s$ and $\widehat{\theta}^{[1]}$ give

$$\begin{aligned}\widehat{\theta}_s - \widehat{\theta}^{[1]} = & -\left[\widehat{\Psi}(\widehat{W}^{[1]}) - \widehat{\Psi}(\widehat{W}^{[0]})\right] \bar{g}_{nT}(\theta_0, \widehat{W}^{[0]}, M) \\ & - \widehat{\Psi}(\widehat{W}^{[1]}) \left[\bar{g}_{nT}(\theta_0, \widehat{W}^{[1]}, M) - \bar{g}_{nT}(\theta_0, \widehat{W}^{[0]}, M)\right].\end{aligned}$$

The Jacobian perturbation bound and the inverse identity imply

$$\|\widehat{\Psi}(\widehat{W}^{[1]}) - \widehat{\Psi}(\widehat{W}^{[0]})\|_2 = O_p\left(\|\widehat{W}^{[1]} - \widehat{W}^{[0]}\|_F\right).$$

Together with (E.53), (E.66), and (E.61), this yields

$$\begin{aligned}\|\widehat{\theta}_s - \widehat{\theta}^{[1]}\|_2 &= O_p\left(a_{nT} \|\widehat{W}^{[1]} - \widehat{W}^{[0]}\|_F\right) + O_p\left(\|\widehat{W}^{[1]} - \widehat{W}^{[0]}\|_F\right) \\ &= O_p(\xi_{nT} a_{nT}).\end{aligned}$$

Suppose, in addition, that

$$\sqrt{nT}\, \mathcal{R}^*_{nT} = o(1) \qquad \text{and} \qquad \xi_{nT} = o(1).$$

Then

$$\sqrt{nT}\, \xi_{nT} a_{nT} = \xi_{nT}\left(1 + \sqrt{nT}\, \mathcal{R}^*_{nT}\right) = o(1).$$

Consequently,

$$\sqrt{nT}\left(\widehat{\theta}_s - \widehat{\theta}^{[1]}\right) = o_p(1).$$

The asymptotic distribution of $\widehat{\theta}_s$ is therefore the same as that of the initial plug-in estimator $\widehat{\theta}^{[1]}$ established in Theorem 2. Thus, supervision may affect finite-sample performance, but under $\xi_{nT} = o(1)$ it is asymptotically negligible at the $\sqrt{nT}$ scale and does not alter the first-order limiting distribution. This completes the proof.

□

E.8. **Proof of Corollary 2.** The proof follows the same steps as in Theorem 2, with the only difference being the substitution of the supervised estimator's rate.