Identification of Spillover Effects using Panel Data

Christiern Rose*

University of Queensland

christiern.rose@uq.edu.au

Latest version: March 2018. First version: June 2013

Abstract

In this paper I use panel data to identify and estimate spillover effects when the underlying network is sparse and unobserved. The outcome of each entity depends on its own covariates, the outcomes of others (endogenous effects), the covariates of others (contextual effects) and an entity-specific fixed effect. Sparsity restrictions are exclusion restrictions of unknown location, and lead to point identification if the network is suitably connected. Prior knowledge that a particular covariate does not generate contextual effects may also lead to point identification. I use the results of Gautier and Tsybakov (2014) for estimation, model selection and inference under sparsity. The results of a Monte Carlo experiment demonstrate applicability in realistic settings. I apply the approach to study R&D spillovers in an oligopoly model, finding that spillovers are sent predominantly by large firms and that R&D stocks are below the social optimum. My results do not depend on pre-specified competition and technology networks.

JEL codes: C31, L14, L24

Key words: Networks, spillovers, panel data, peer effects, R&D spillovers, high-dimensional econometrics

1. Introduction

Spillover effects occur in many applied settings, including technology adoption, treatment effects, risky behaviors, peer effects in education, economic growth, labor market performance and almost all strategic contexts. There are two types of spillover: *endogenous effects*, through which the outcomes of interest are simultaneously determined, and *contextual effects*, through which the outcomes depend on the covariates of others (Manski, 1993).

In order to empirically evaluate spillovers there must be some notion of which entities interact with one another and to what extent. That is to say, there must be an underlying

*I thank Stéphane Bonhomme, Aureo de Paula, Jean-Pierre Florens, Eric Gautier, Gregory Jolivet, Senay Sokullu and Frank Windmeijer for their constructive comments and advice. I am also grateful for comments and suggestions from seminar participants at the RES Annual Conference 2015, RES PhD Meetings 2015, the 2014 European Winter Meeting of the Econometric Society, Toulouse School of Economics, IAAE 2014, Econometric Study Group 2014 and the University of Bristol. I acknowledge financial support from the grant ERC POEMH, the ESRC, the IAAE and the RES. Any errors are my own.

network through which spillovers operate.[1] Most existing methodologies treat the network as observed and exogenous, and exploit its properties to identify some low dimensional parameters which capture the endogenous and contextual effects.[2] For example, in order to evaluate R&D spillovers in oligopoly, Bloom et al. (2013) and König et al. (2014) use data on sales and patents to construct competition and technology networks, through which the spillovers are assumed to operate.

In this paper, I suppose that the network is unobserved and sparse. This means that each entity has few direct neighbors compared to the feasible number. The main contribution is to propose a new identification strategy based around a sparsity restriction. In addition, I propose a novel means of conducting estimation, model selection and inference under sparsity through applying the results of Gautier and Tsybakov (2014).

From an applied perspective, the method is useful if network data are unavailable or subject to measurement error, which may arise if it is unclear which metric is appropriate to measure distance or due to censoring.[3] This is particularly important in economic applications, in which the notion of 'economic distance' is not well defined. Moreover, if the research objective is to study the identities or types of entities which send and receive spillovers, it is crucial to estimate the network rather than impose it. In the empirical application, I find that R&D spillovers are predominantly sent by large firms.

The baseline specification is a linear panel model with endogenous and contextual effects. Entities interact through an unobserved network, which determines the locations of nonzero effects. Spillovers may be heterogeneous and asymmetric. The parameters for the spillovers sent from entity j to $i \neq j$ are indexed by ij. It is for this reason that longitudinal data are required. I also study an extension of the model which incorporates entity-specific fixed effects.

I derive two identification results, one based on a sparsity restriction, and the other on a natural restriction on the contextual effects. Sparsity is equivalent to exclusion restrictions of unknown locations and leads to point identification if the network satisfies a condition related to the number and nature of its vertex-independent paths.[4] I derive rank and order conditions for point identification under sparsity, which reduce to the classical conditions for structural equations models if the network is observed.

The identification strategy under sparsity uses exogenous variation in the covariates of indirect neighbors to identify the endogenous effects exerted by the direct neighbors. This idea is studied in Bramoullé et al. (2009) under network observability. My innovation is to extend the approach to a setting where the researcher knows that the number of neighbors is bounded from above, but does not know neighbor identities. To account for the unknown network, up to twice as many restrictions are required to obtain point identification, which is attained if the network is sparse and suitably connected.

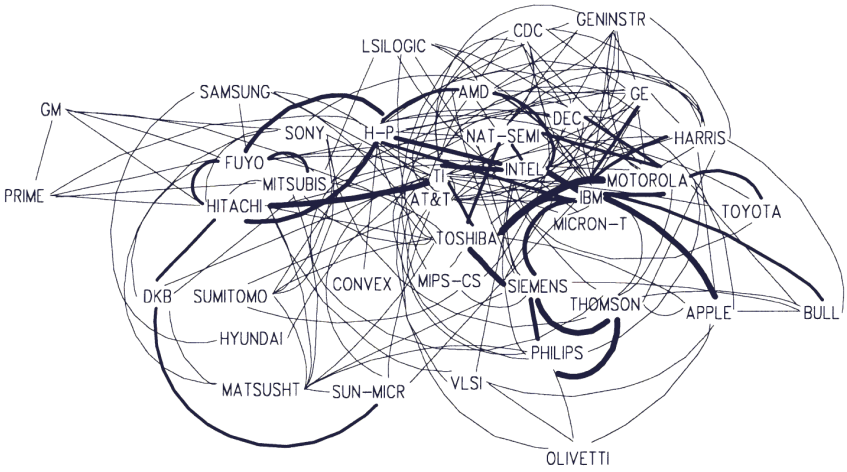
[1] There may be a different network for each source of spillovers, though most research in this area assumes a single common network.

[2] Exceptions include Manresa (2014); Lam and Souza (2013); de Paula et al. (2016); Gautier (2015) and Gautier and Rose (2016), which do not assume network observability. The precise relationship with these papers is discussed at the end of this section.

[3] In social network data, the number of friends than an individual may name is usually constrained from above. For example, in the Ad-Health data, students are asked to name up to 5 friends.

[4] A path is a sequence of neighboring vertices, beginning at one vertex and ending at another. Two paths are vertex independent if they have no vertex in common.

Figure 1: *R&D partnerships in the electronics industry (**Source**: Duysters et al. (1999))*



Many real world networks are sparse and connected. Social networks are a good example, as documented by Stanley Milgram's famous 'small world' experiment. For a concrete example, consider the R&D partnership network for major firms in the electronics industry, which is studied in Duysters et al. (1999) and depicted in figure 1.[5] The vertices represent firms, and the edges represent R&D collaborations as defined in the MERIT-CATI database in 1999.[6] Two things are immediately apparent. First, the network is sparse: each firm has relatively few partnerships relative to the number of firms. Second, the network is connected: any firm can be reached from any other firm through neighbors-of-neighbors.

The second identification result shows that prior knowledge that a particular covariate does not generate contextual effects can lead to point identification. If each entity has a covariate which determines its own outcome but not that of others, exogenous variation in this covariate may be used to instrument the outcome in the other equations. This approach does not rely on the properties of the network governing the endogenous effects, and is applicable in settings where there are entity-specific covariates which determine the costs and/or benefits of engaging in the outcome. For example, in the empirical application, each firm's output may be determined by exogenous, firm-specific variation in its costs or demand.

In typical applications, the number of periods (T) is small relative to the number of entities (N). This means that standard estimation and inference procedures are inapplicable because the number of parameters and instruments in each equation is linear in N, and may exceed the number of observations T. Moreover, in the absence of exclusion restrictions (i.e. under sparsity restrictions alone), the number of parameters exceeds the number of instruments. I show that, in addition to identifying the parameters, sparsity may also be exploited to address lack of data. To do this, I apply the Self Tuning Instrumental Variables (STIV) estimator developed in Gautier and Tsybakov (2014).

The STIV estimator is suited to high-dimensional linear settings with many endogenous covariates and many instruments. The objective function penalizes the ℓ_1 norm of the parameter

[5] Other examples of sparse and connected networks include lending networks, infrastructure networks, supply networks, trade networks and co-authorship networks.

[6] An edge between two firms exists if there is any type of R&D collaboration, including technology transfers, joint research, joint development, cross licensing, R&D contracts, joint ventures and research corporations.

vector, yielding sparse solutions. Gautier and Tsybakov (2014) show that the estimated set of nonzero parameters is a superset of the true set of nonzero parameters with high probability, provided that they are not too close to zero. Under a stronger assumption on the size of the nonzero parameters, the estimated and true sets coincide. Consequently, it is possible to exactly recover the true network if the spillovers are sufficiently large.

The convergence rate of the STIV estimator depends on the sparsity of the parameter vector and the ratio of the logarithm of the number of instruments and the sample size (Gautier and Tsybakov, 2014). This implies that even though the number of parameters in equation i grows linearly with N, the ℓ_2 norm of the estimation error converges to zero at the rate $\sqrt{|\mathcal{E}_i| \log(N)/T}$, where $|\mathcal{E}_i|$ is the number of direct neighbors of entity i.[7] Consistency requires $|\mathcal{E}_i| \log(N)/T \to 0$, which can be achieved even as $N/T \to \infty$ if the network is sufficiently sparse. I conduct simulations to evaluate the performance of the method, with results demonstrating that it is applicable in realistic settings.

I apply the approach to study R&D spillovers and product market rivalry in a structural oligopoly model. Results suggest that the ratio of the marginal social and private benefits of R&D was around 1.03 between 1981 and 2001. From a welfare perspective, this is indicative of underinvestment in R&D. Relative to existing work, the novelty lies in estimation rather than imposition of the competition and technology networks. This implies that my results are robust to misspecification of the networks, and permits the analysis of the identities and types of firms which send and receive spillovers. The remainder of this section reviews the existing literature and outlines the structure of the paper.

1.1. Related Literature

The majority of papers in the spillover effects literature assume that the network is observed and exogenous. A comprehensive review is provided by de Paula (2015). Identification of spillover effects if the network is not observed has only recently received attention. Blume et al. (2010) discuss identification prospects when the network is partially observed, such that the researcher knows neighbor identities but does not know the strength of the ties. The authors provide identification results for the case where the network is known to be circular and the weight of the ties decays geometrically in the distance between agents.

Blume et al. (2015) show that almost nothing can be learned about spillovers in the case where the researcher has no prior information on the underlying network, though prospects are improved under partial information.[8] If the researcher knows the network for the contextual effects, point identification can be attained if there are two entities which are known not to be neighbors in the endogenous effects network.[9] Blume et al. (2015) also show that point identification is feasible if there are sufficiently many disconnected entities in the endogenous and contextual effects networks, assuming that the researcher knows their identities. Souza (2014) supposes that the network is unobserved but depends exclusively on entities' observed

[7] To focus on the relative roles of N and T, the dimensions of the regressors $\boldsymbol{X}_{it}$ and instruments $\boldsymbol{Z}_{it}$ are fixed to be K and L respectively. For this reason, they do not appear in the rates.

[8] Blume et al. (2015) study the case where each source of spillovers may operate through a different adjacency matrix.

[9] There is an additional, technical condition which depends on the specifics of the model in Blume et al. (2015) and is omitted here for brevity.

characteristics, which are also assumed to be exogenous. In this event, the social effects parameters are partially identified.

The strand of the literature to which my work belongs has focussed on panel data models. Manresa (2014) considers estimation of contextual effects using panel data, treating the network as sparse and unobserved. The author proposes a pooled LASSO estimator, which is a variant of the LASSO estimator studied in Tibshirani (1996). Lam and Souza (2013) study estimation of contextual and endogenous effects using panel data, treating the network as unobserved and applying LASSO to estimate the parameters. Since the LASSO estimator does not allow for endogeneity, the authors assume that the variance of the structural error decays to zero asymptotically in order to obtain consistency results. The estimators in Manresa (2014) and Lam and Souza (2013) rely on an iterative procedure, the properties of which are not well understood.

de Paula et al. (2016) study identification and estimation of contextual and endogenous effects using panel data. The authors derive identification results similar to those of Bramoullé et al. (2009), which require linear independence of powers of the adjacency matrix. These results may be used to uniquely recover the spillovers and underlying network from the reduced form parameter matrix, which may be estimated directly by applying the adaptive LASSO of Zou (2006). This approach requires that the reduced form parameter matrix be sparse, which can be achieved if each entity is connected to few others, either directly or indirectly. This is stronger than the sparsity restriction I consider in this paper, which restricts only the direct connections. The authors also study estimation based on a sparse structural form by applying the adaptive GMM estimator of Caner and Zhang (2014). The properties of this estimator are known if $N/T \to 0$ as $N, T \to \infty$.

Relative to these papers, this paper makes two main innovations. First, I derive identification results under sparsity in a model with endogenous and contextual effects. Second, through applying the results of Gautier and Tsybakov (2014), I show that sparsity may be exploited to conduct estimation, model selection and inference even when N is large relative to T and the variance of the structural disturbance does not go to zero.

In statistics, the identification results in this paper are related to those of Candes and Tao (2007); Gautier and Tsybakov (2014) and Kang et al. (2016). Candes and Tao (2007) discuss conditions for exact recovery of β in the noiseless case in which $\boldsymbol{y} = \boldsymbol{X}\beta$, β is $p \times 1$ and $\boldsymbol{X}$ is $n \times p$ with $n < p$. Sparsity means that β has at most s nonzero elements. To perform exact recovery based on ℓ_1 penalization, the authors require a 'uniform uncertainty principle'. For this condition to hold, it is necessary that any sub-matrix of $\boldsymbol{X}$ formed from $2s$ columns has full column rank. This implies that there do not exist sparse $\beta_1 \neq \beta_2$ such that $\boldsymbol{X}\beta_1 = \boldsymbol{X}\beta_2$, and hence that β is identified.

Gautier and Tsybakov (2014) discuss identification in a single equation model with many endogenous regressors, many instruments and unknown exclusion restrictions. To see the argument, consider the population model $\boldsymbol{y} = \boldsymbol{X}'\beta + \epsilon, \mathbb{E}[\boldsymbol{Z}\epsilon] = \boldsymbol{0}$, where $\boldsymbol{y}$ and ϵ are scalars, $\boldsymbol{X}$ is $p \times 1$ and $\boldsymbol{Z}$ is $l \times 1$, and denote by $\boldsymbol{X}_{\boldsymbol{J}}$ the sub-vector of $\boldsymbol{X}$ formed from elements $\boldsymbol{J} \subseteq \{1, ..., p\}$. Under sparsity, the vector β has at most s nonzero entries, so the outcome equation can be written as $\boldsymbol{y} = \boldsymbol{X}'_{\boldsymbol{J}_\beta}\beta_{\boldsymbol{J}_\beta} + \epsilon$ where $\boldsymbol{J}_\beta$ is the support of β. Gautier and Tsybakov (2014) show that point identification is attained if $\mathbb{E}[\boldsymbol{Z}\boldsymbol{X}'_{\boldsymbol{J}_\beta}]$ has rank s and the vector $\mathbb{E}[\boldsymbol{Z}\boldsymbol{y}]$ lies in the range of

$\mathbb{E}[\boldsymbol{Z}\boldsymbol{X}'_{\boldsymbol{J}_\beta}]$ but not in the range of $\mathbb{E}[\boldsymbol{Z}\boldsymbol{X}'_{\tilde{\boldsymbol{J}}_\beta}]$ for some other $\tilde{\boldsymbol{J}}_\beta \subseteq \{1,...,p\}$ with cardinality at most s. These conditions respectively mean that there is a unique sparse parameter vector which could have generated the data, which also satisfies the usual rank condition.

Kang et al. (2016) study a single equation model with one endogenous regressor and many instruments. The authors study the case where all of the instruments are strong but it is not known which are excluded. The main result is to show that point identification is attained if at least half of the instruments are excluded.

In this paper, I provide identification results for linear systems of equations with endogeneity, and present rank and order conditions similar to the of the classical conditions for identification in structural equations models. To do this, I modify the approach of Candes and Tao (2007) to allow for endogeneity. Moreover, I relate the rank condition to the properties of the underlying network.

The remainder of the paper is structured as follows. Section 2 explains the notations and network theoretic terminology. Section 3 sets out the baseline model, and section 4 presents identification results. Section 5 presents the estimation strategy, which is applied to the simulations in section 6 and the empirical application in section 7. Section 8 concludes. All proofs are gathered in the appendix.

2. Notation

For $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, $\boldsymbol{X}_{\mathcal{Y},\mathcal{X}}$ denotes the sub-matrix formed from rows $\mathcal{Y} \subseteq \{1,...,m\}$ and columns $\mathcal{X} \subseteq \{1,...,n\}$. If $\mathcal{Y} = \{1,...,m\}$, the sub-matrix is $\boldsymbol{X}_{\cdot,\mathcal{X}}$ and if $\mathcal{X} = \{1,...,n\}$ it is $\boldsymbol{X}_{\mathcal{Y},\cdot}$. If $\boldsymbol{X}$ is square, $\text{diag}(\boldsymbol{X})$ is the column vector comprising its diagonal elements. The m dimensional identity matrix is $\boldsymbol{I}_m$ and vector of ones is $\boldsymbol{\iota}_m$. The cardinality of a set $\mathcal{S}$ is $|\mathcal{S}|$. The difference between two sets is $\mathcal{S}_1 \backslash \mathcal{S}_2$. The indicator function for condition c is $\mathbf{1}_c$.

A network is a tuple, $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ is a finite set of vertices and $\mathcal{E}$ is a finite set of edges, which are ordered pairs of vertices. I consider simple networks, in which there are no duplicate edges, and no edges from a vertex to itself. The network is undirected if $(j,i) \in \mathcal{E} \Leftrightarrow (i,j) \in \mathcal{E}$, and directed otherwise. My results apply equally to directed and undirected networks.

A walk $w = (v_1, e_1, v_2, e_2, ..., e_n, v_{n+1})$ is an alternating sequence of vertices and edges, such that for all $m = 1, ..., n : e_m = (v_m, v_{m+1}) \in \mathcal{E}$. There is a path from vertex j to vertex i if there is a walk beginning at j and ending at i. For two subsets of vertices, $\mathcal{V}_1, \mathcal{V}_2$ there is a path from $\mathcal{V}_2$ to $\mathcal{V}_1$ if there is a path from some $j \in \mathcal{V}_2$ to some $i \in \mathcal{V}_1$. Two paths from $\mathcal{V}_2$ to $\mathcal{V}_1$ are vertex-independent if they do not have a common vertex.

The connectivity of the network $\mathcal{C}$ is the set of ordered pairs of vertices such that $(j,i) \in \mathcal{C}$ if there is a path from j to i and $j \neq i$. This set satisfies $\mathcal{C} \supseteq \mathcal{E}$. The network $\mathcal{G}$ is connected if, for each pair of vertices, $(i,j) \in \mathcal{C}$ or $(j,i) \in \mathcal{C}$, and strongly connected if $(i,j) \in \mathcal{C}$ and $(j,i) \in \mathcal{C}$.

For vertex i, $\mathcal{E}_i = \{j \in \mathcal{V} : (j,i) \in \mathcal{E}\}$ are its neighbors and $\mathcal{C}_i = \{j \in \mathcal{V} : (j,i) \in \mathcal{C}\}$ is its neighborhood. These sets satisfy $\mathcal{C}_i \supseteq \mathcal{E}_i$. The complements of these sets are the non-neighbors $\mathcal{E}_i^c = \mathcal{V} \backslash \{\mathcal{E}_i \cup i\}$, and the non-neighborhood $\mathcal{C}_i^c = \mathcal{V} \backslash \{\mathcal{C}_i \cup i\}$. The in-degree is the number of neighbors $|\mathcal{E}_i|$ and the out-degree is the number of vertices for which i is a neighbor $|\{j \in \mathcal{V} : i \in \mathcal{E}_j\}|$. Vertex i is isolated if $\mathcal{E}_i = \emptyset$. The binary adjacency matrix $\boldsymbol{G} \in \{0,1\}^{N\times N}$

represents the network. Element $\boldsymbol{G}_{ij}$ is equal to one if $(j,i) \in \mathcal{E}$ and zero otherwise.

3. Model

There are N vertices $\mathcal{V} = \{1, 2, ..., N\}$, which interact for T periods through a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with connectivity $\mathcal{C}$. In period t, vertex i has a scalar outcome $\boldsymbol{Y}_{it}$, which is determined according to:

$$\boldsymbol{Y}_{it} = \left(\sum_{j \in \mathcal{E}_i} \boldsymbol{\psi}_{ij} \boldsymbol{Y}_{jt} + \boldsymbol{\gamma}_{ij}' \boldsymbol{X}_{jt} \right) + \boldsymbol{\gamma}_{ii}' \boldsymbol{X}_{it} + \boldsymbol{\epsilon}_{it} \tag{3.1}$$

where $\boldsymbol{X}_{it} \in \mathbb{R}^K$ is a vector of covariates and $\boldsymbol{\epsilon}_{it}$ is a scalar representing unobserved heterogeneity. The set of edges $\mathcal{E}$ is defined as the set of ordered pairs of vertices (j,i) such that j exerts at least one spillover on i:

$$\mathcal{E} = \{(j,i) \in \mathcal{V}^2 : j \neq i, \boldsymbol{\psi}_{ij} \neq 0 \text{ or } \boldsymbol{\gamma}_{ij} \neq \boldsymbol{0}\} \tag{3.2}$$

Associated with each vertex there is a vector of instrumental variables $\boldsymbol{Z}_{it} \in \mathbb{R}^L$, which may be constructed using current, past or future values of the covariates or outcomes of any of the vertices, or using some other external source of variation. In particular, if $\boldsymbol{X}_{it}$ is exogenous (see assumption 4.5), $\boldsymbol{Z}_{it} = \boldsymbol{X}_{it}$ is natural.

For clarity of exposition, the intercept is omitted from the outcome equation (3.1). This is rectified in section 4.3, which decomposes the disturbance to be the sum of vertex-specific fixed heterogeneity and a remaining vertex-period term. All of the identification results remain valid under a minor modification of the assumptions, which is presented in section 4.3. Aside from linearity, (3.1) is relatively general, and includes the specifications in Manresa (2014); Lam and Souza (2013) and de Paula et al. (2016) as special cases. In particular, it allows for endogenous and contextual spillovers which may be heterogeneous and directed.

Equation (3.1) is more compactly expressed as:

$$\boldsymbol{Y} = \boldsymbol{\Psi} \boldsymbol{Y} + \boldsymbol{\Gamma} \boldsymbol{X} + \boldsymbol{\epsilon} \tag{3.3}$$

where $\boldsymbol{Y}$ and $\boldsymbol{\epsilon}$ are $N \times T$, $\boldsymbol{\Psi}$ is $N \times N$ with diagonal elements equal to zero, $\boldsymbol{\Gamma}$ is $N \times NK$ and $\boldsymbol{X}$ is the $NK \times T$ matrix of stacked covariates. The $NL \times T$ matrix of stacked instruments is $\boldsymbol{Z}$. The $N \times N(1+K)$ matrix of structural parameters is denoted $\boldsymbol{\Theta} = (\boldsymbol{\Psi}\ \boldsymbol{\Gamma})$. The row vector $\boldsymbol{\Theta}_{i,\cdot}$ gives the spillovers received by vertex i. If the endogenous effects satisfy a stability condition (see assumption 4.4), the reduced form of (3.3) is:

$$\boldsymbol{Y} = \boldsymbol{\Pi} \boldsymbol{X} + \boldsymbol{\eta} \tag{3.4}$$

where $\boldsymbol{\Pi} = (\boldsymbol{I}_N - \boldsymbol{\Psi})^{-1} \boldsymbol{\Gamma}$ and $\boldsymbol{\eta} = (\boldsymbol{I}_N - \boldsymbol{\Psi})^{-1} \boldsymbol{\epsilon}$.

In the empirical application the vertices are firms, which are observed annually. Firms make R&D investments and engage in Cournot competition on the product market. Based on the production and demand specifications in section 7 equation (3.1) is a Cournot best response function in which $\boldsymbol{Y}_{it}$ is the log of real sales, $\boldsymbol{X}_{it}$ is the log R&D stock and $\boldsymbol{Z}_{it}$ are instruments

constructed from measures of firms' tax incentives for R&D investments and lags of the log R&D stock. The parameters $\mathbf{\Psi}$ and $\mathbf{\Gamma}$ respectively have structural interpretations in terms of demand elasticities and the parameters of the production function. The network $\mathcal{G}$ determines which firms exert spillovers on one another, either through competing on the product market, or through technology spillovers, which serve to reduce the beneficiaries costs.[10] Sections 4, 5 and 6 abstract away from the empirical application, to which we return in section 7.

4. Identification

A parametric model is point identified if the set of structural parameters which are consistent with the restrictions is a singleton. Otherwise the model may be partially identified. This section focuses on identification of $\mathbf{\Theta}$. Identification of the network $\mathcal{G}$ follows immediately using (3.2). The following assumptions are used in some or all of the propositions.

Assumption 4.1 (Weak stationarity)

$$(\boldsymbol{Y}_{\cdot,t}, \boldsymbol{X}_{\cdot,t}, \boldsymbol{Z}_{\cdot,t}) \text{ is weakly stationary.}$$

Assumption 4.2 (Linearly independent longitudinal variation)

$$rank\left(\mathbb{E}\left[\boldsymbol{X}_{\cdot,t}\boldsymbol{Z}'_{\cdot,t}\right]\right) = NK$$

Assumption 4.3 (Normalization)

$$diag(\mathbf{\Psi}) = \mathbf{0}$$

Assumption 4.4 (Stability)

$$\det(\boldsymbol{I}_N - \mathbf{\Psi}) \neq 0$$

Assumption 4.5 (Strict exogeneity)

$$\mathbb{E}\left[\boldsymbol{\epsilon}_{\cdot,t}\boldsymbol{Z}'_{\cdot,t}\right] = \mathbf{0}$$

Assumption 4.6 (s-sparsity)

$$\text{For a specified } \boldsymbol{s} \in \mathbb{N}^N, \quad |\mathcal{E}_i| \leq \boldsymbol{s}_i \quad \forall i \in \mathcal{V}$$

Assumption 4.1 is a weak stationarity assumption, which implies that the expectations are defined in the usual way. Assumption 4.2 is a full rank condition on the covariates and instruments. A necessary condition for assumption 4.2 is $L \geq K$. If $\boldsymbol{Z} = \boldsymbol{X}$, assumption 4.2 mandates that the time series of the NK covariates be linearly independent. Otherwise, the vector of instruments must generate linearly independent time series.

Assumption 4.3 is a normalization and assumption 4.4 is a stability condition for the endogenous effects. These assumptions are invoked in almost all structural equation models.

[10] I follow Bloom et al. (2013) and König et al. (2014) in assuming that R&D investments reduce costs. An alternative specification could consider R&D investments which increase demand.

Assumption 4.5 is a strict exogeneity assumption. Variants of this assumption, (usually with $\boldsymbol{Z} = \boldsymbol{X}$) are commonly invoked throughout the literature on identification of spillover effects (Blume et al., 2010; de Paula, 2015), and underpin the identification strategies of Lam and Souza (2013); Manresa (2014) and de Paula et al. (2016), among others. Many papers also make a stronger conditional exogeneity assumption such as $\mathbb{E}\left[\boldsymbol{\epsilon}_{\cdot,t}|\boldsymbol{Z}_{\cdot,t}\right] = \mathbf{0}$.

Assumption 4.6 is a sparsity restriction on the network $\mathcal{G}$. It restricts the in-degrees of the vertices, but leaves the out-degrees unrestricted.[11] This is important in applied settings where some vertices are more influential than others. For example, in the empirical application, it may be that particular firms are 'technological leaders'.

The sparsity restriction 4.6 is useful for both identification and estimation. From an identification perspective, it generates exclusion restrictions of unknown locations, which lead to point identification under a rank condition. From an estimation, model selection and inference perspective, sparsity is useful to address lack of data, and can lead to consistency and good finite sample properties even if N/T is large. This is studied in section 5.

The next part of this section studies the extent to which sparsity restrictions are useful for identification. To do this, first define the identified set, which is the the set of parameters which are compatible with assumptions 4.3-4.6 for some sparsity $\boldsymbol{s} \in \mathbb{N}^N$:

$$\mathcal{I}(\boldsymbol{s}) = \left\{ \begin{array}{ll} \boldsymbol{\Theta}: & \operatorname{diag}(\boldsymbol{\Psi}) = \mathbf{0}, \quad \det(\boldsymbol{I}_N - \boldsymbol{\Psi}) \neq 0, \\ & \mathbb{E}\left[\left(\boldsymbol{Y}_{\cdot,t} - \boldsymbol{\Psi}\boldsymbol{Y}_{\cdot,t} - \boldsymbol{\Gamma}\boldsymbol{X}_{\cdot,t}\right)\boldsymbol{Z}_{\cdot,t}'\right] = \mathbf{0}, \quad |\mathcal{E}_i| \leq \boldsymbol{s}_i \quad \forall i \in \mathcal{V} \end{array} \right\} \tag{4.1}$$

It is clear from the definition of $\mathcal{I}(\boldsymbol{s})$ that for $\boldsymbol{s}_1 \leq \boldsymbol{s}_2$ we have $\mathcal{I}(\boldsymbol{s}_1) \subseteq \mathcal{I}(\boldsymbol{s}_2)$. Consequently, strengthening the sparsity assumption weakly shrinks the identified set. If $\boldsymbol{s} \geq (N-1)\boldsymbol{\iota}_N$, the set $\mathcal{I}(\boldsymbol{s})$ is the identified set under unrestricted sparsity, which is denoted by $\mathcal{I}$ from this point. To fix ideas, the following lemma summarizes a well known non-identification result for structural equation models.

Lemma 4.7 (Non-identification of the unrestricted model)
Let $\boldsymbol{\Theta} \in \mathcal{I}$ and let assumptions 4.1 and 4.2 be satisfied. $\boldsymbol{\Theta}$ is not point identified.

Lemma 4.7 states that assumptions 4.1-4.5 are insufficient for point identification. In this case, the number of structural parameters exceeds the number of reduced form parameters by $N(N-1)$, implying that $\boldsymbol{\Pi}$ is not an injective mapping. The next sub-section studies identification under sparsity.

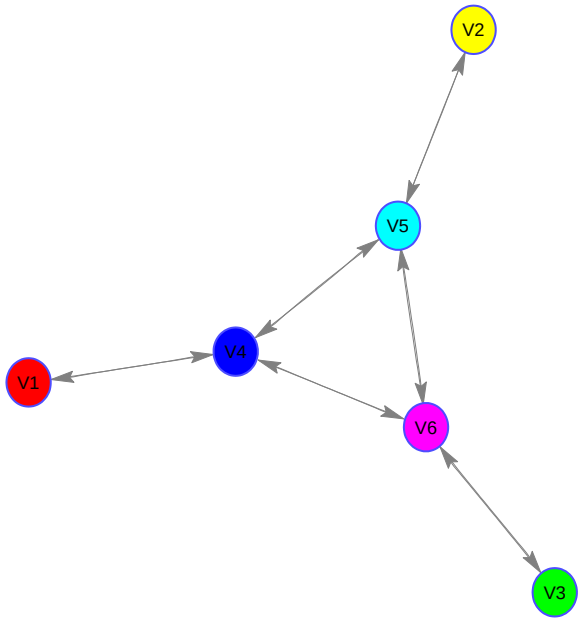
4.1. Sparsity Restrictions

For clarity of exposition, the results in this sub-section are presented for the case where there is a single covariate, such that $K = 1$. Identification results for general K require additional notation and complicate the argument. These results are provided in the appendix. To fix ideas, the next lemma considers identification under network observability.

Lemma 4.8 (Identification when the network is observed)
Let $\boldsymbol{\Theta} \in \mathcal{I}$ and let assumptions 4.1 and 4.2 be satisfied. Assume further that $\mathcal{G}$ is observed. $\boldsymbol{\Theta}_{i,\cdot}$ is point identified if the sub-matrix of $\boldsymbol{\Pi}$ with rows $\mathcal{E}_i$ and columns $\mathcal{V}\backslash\{\mathcal{E}_i \cup i\}$ has full row rank.

[11] Although clearly no vertex can have an out-degree exceeding $\boldsymbol{s}'\boldsymbol{\iota}_N$.

Figure 2: *A network in which* $\boldsymbol{\Theta}_{1,\cdot}$ *is point identifiable*



Lemma 4.8 presents the classical rank condition for point identification. The basic idea is to use exogenous variation in the covariates of vertices in the neighborhood of i as instruments for the outcomes of i's neighbors. This approach is studied in the spillover effects context by Bramoullé et al. (2009). As an example, consider the network in figure 2, and suppose that the neighbor identities are known. To point identify $\boldsymbol{\Theta}_{1,\cdot}$, one needs instruments for the outcome of vertex 4. Candidate instruments are the covariates of vertices 2, 3, 5 and 6, which are in the neighborhood of 1 but are not neighbors. This means that their covariates are excluded from equation 1. Point identification is attained if any of these have a reduced form effect on the outcome of vertex 4.

The next proposition extends the identification argument of Bramoullé et al. (2009) to the case where the researcher knows that the number of neighbors is bounded from above, but does not know their identities. To do this, I adapt the identification argument in Candes and Tao (2007) to incorporate endogeneity. Before stating the result, it is useful to define the observable connectivity:

$$\mathcal{C}^{\boldsymbol{\Pi}} = \{(j,i) : j \neq i, \quad \boldsymbol{\Pi}_{ij} \neq 0\} \tag{4.2}$$

This set describes the information on the the network $\mathcal{G}$ which may be inferred from the reduced form parameters. Similarly, I define the observable neighborhood of vertex i as $\mathcal{C}_i^{\boldsymbol{\Pi}} = \{j \in \mathcal{V} : (j,i) \in \mathcal{C}^{\boldsymbol{\Pi}}\}$.

The observable connectivity $\mathcal{C}^{\boldsymbol{\Pi}}$ need not equal the connectivity $\mathcal{C}$. This is because two vertices being connected is not sufficient for a nonzero reduced form effect. For example, if $\boldsymbol{\Gamma} = (\boldsymbol{I}_N - \boldsymbol{\Psi})$ then $\boldsymbol{\Pi} = \boldsymbol{I}_N$ and $\mathcal{C}^{\boldsymbol{\Pi}} = \emptyset$ regardless of $\mathcal{C}$. In this example, the endogenous and contextual effects exactly offset one another. From an identification perspective, this result is negative: one cannot learn anything about the set of edges $\mathcal{E}$ based on the support of $\boldsymbol{\Pi}$. All that we can say for certain is that if $j \neq i$ has a reduced form effect on i then j must be in i's neighborhood. Formally, this means $\mathcal{C} \supseteq \mathcal{C}^{\boldsymbol{\Pi}}$.

In order to learn more about the network based on the observable connectivity, one can also make the following assumption:

Assumption 4.9 (Reduced form effect of *i*'s neighbors)

$$\textit{For a given } i \in \mathcal{V}, \quad \mathbf{\Pi}_{ij} \neq 0 \quad \forall j \in \mathcal{E}_i$$

Assumption 4.9 requires that every neighbor of i has a reduced form effect on i's outcome. Under this assumption one can infer a superset of the neighbors of i from $\mathbf{\Pi}$, such that $\mathcal{C}_i \supseteq \mathcal{C}_i^{\mathbf{\Pi}} \supseteq \mathcal{E}_i$. The second part of proposition 4.10 uses assumption 4.9. Similar assumptions are made in Moffitt (2001); Lee (2007); Davezies et al. (2009); Bramoullé et al. (2009); Lam and Souza (2013) and de Paula et al. (2016).

Proposition 4.10 (Identification under sparsity)
Let $\mathbf{\Theta} \in \mathcal{I}(\boldsymbol{s})$ and let assumptions 4.1 and 4.2 be satisfied.

(i) $\mathbf{\Theta}_{i,\cdot}$ is point identified if for every

$$\tilde{\mathcal{E}}_i \in \{\tilde{\mathcal{E}}_i : \tilde{\mathcal{E}}_i \subseteq \mathcal{E}_i^c, |\tilde{\mathcal{E}}_i| = \min[\boldsymbol{s}_i, |\mathcal{E}_i^c|]\}, \tag{4.3}$$

the sub-matrix of $\mathbf{\Pi}$ with rows $\mathcal{E}_i \cup \tilde{\mathcal{E}}_i$ and columns $\mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$ has full row rank.

(ii) In addition, let assumption 4.9 be satisfied for vertex i. Then $\mathcal{C}_i \supseteq \mathcal{C}_i^{\mathbf{\Pi}} \supseteq \mathcal{E}_i$ and part (i) *applies with* (4.3) *replaced by*

$$\tilde{\mathcal{E}}_i \in \{\tilde{\mathcal{E}}_i : \tilde{\mathcal{E}}_i \subseteq \mathcal{E}_i^c, |\tilde{\mathcal{E}}_i| = \min[\boldsymbol{s}_i, |\mathcal{E}_i^c|]\} \cap \mathcal{C}_i^{\mathbf{\Pi}} \tag{4.4}$$

Proposition 4.10 extends the classical rank condition to the case where the network is unobserved and sparse. Sparsity is equivalent to placing exclusion restrictions of unknown locations. Point identification is more challenging compared to the case where the network is observed. This is because there may be more than one set of neighbors which could generate the same data. In proposition 4.10, the potentially observationally equivalent neighbors are denoted by $\tilde{\mathcal{E}}_i$.

The sparsity restriction $|\mathcal{E}_i| \leq \boldsymbol{s}_i$ places an upper bound on the number of neighbors of vertex i. This implies that in any observationally equivalent network $\tilde{\mathcal{G}}$, i can have at most $\boldsymbol{s}_i$ neighbors (i.e. $|\tilde{\mathcal{E}}_i| \leq \boldsymbol{s}_i$). The argument is as follows. If $\boldsymbol{s}_i$ is small enough, then at least some of the non-neighbors in the observationally equivalent network must also be non-neighbors in the true network.[12] These common non-neighbors are given by the set $\mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$, which is non-empty for sufficiently small $\boldsymbol{s}_i$. The covariates of the common non-neighbors are excluded in both networks, and may be used as instruments for the outcomes of $\mathcal{E}_i \cup \tilde{\mathcal{E}}_i$, which are neighbors of i in at least one of the networks. If the appropriate sub-matrix of the reduced form parameter matrix has full row rank, $\mathcal{E}_i = \tilde{\mathcal{E}}_i$ and the neighbors are uniquely determined.

[12] For $\boldsymbol{s}_i > (N-1)/2$ there need be no zeros in common for row i of the true and observationally equivalent adjacency matrices. However, for $\boldsymbol{s}_i = (N-1)/2 - a_i$ with $a_i \geq 0$, the number of zeros common to row i of the true and observationally equivalent adjacency matrices must be at least $1 + 2a_i$ if N is even and at least $2 + 2a_i$ if N is odd.

To attain point identification, the rank condition must hold over all of the possibly observationally equivalent neighbors. For part (i) of proposition 4.10, this set is:

$$\tilde{\mathcal{E}}_i \in \{\tilde{\mathcal{E}}_i : \tilde{\mathcal{E}}_i \subseteq \mathcal{V}, |\tilde{\mathcal{E}}_i| \leq \boldsymbol{s}_i\} \tag{4.5}$$

For this, it suffices that the rank condition holds over the smaller set in (4.3). This is because the smaller set in (4.3) considers only the largest possible sets of observationally equivalent neighbors.

In part (ii), it is additionally assumed that if j is a neighbor of i then at least one of j's covariates has a reduced form effect on the outcome of i. This identifies a superset of i's neighbors such that $\tilde{\mathcal{E}}_i \subseteq \mathcal{C}_i^{\mathbf{\Pi}}$ and the set of alternative neighbors over which the rank condition must hold is reduced.

The rank conditions in proposition 4.10 are not directly testable since they depend on the true neighbors of i, which are unknown. A sufficient condition is that every sub-matrix formed from each possible set of true and observationally-equivalent neighbors has full rank. This condition does not depend on the unknown neighbors of i, and is thus testable in principle. However, there are two practical barriers. First, the problem is combinatoric and therefore computationally intractable for large N. Second, to conduct such a test one must first estimate the reduced form parameter matrix $\mathbf{\Pi}$, which is not sparse.[13] This implies that $\mathbf{\Pi}$ is cannot be accurately estimated when N is large relative to T, as is typical in applications.

If the network is observed, the true and observationally equivalent networks are equal, such that $\mathcal{G} = \tilde{\mathcal{G}}$. This implies that we need only consider $\tilde{\mathcal{E}}_i = \mathcal{E}_i$ in which case the rank conditions in proposition 4.10 are identical to the classical rank condition in lemma 4.8. The following corollary gives the order conditions for point identification under sparsity.

Corollary 4.11 (Order condition)
The order conditions corresponding to the rank conditions in cases (i) and (ii) of proposition 4.10 are:

(i) $|\mathcal{E}_i| + \min[\boldsymbol{s}_i, |\mathcal{E}_i^c|] \leq \frac{(N-1)}{2}$

(ii) $|\mathcal{E}_i| + \min[\boldsymbol{s}_i, |\mathcal{E}_i^c \cap \mathcal{C}_i^{\mathbf{\Pi}}|] \leq \frac{(N-1)}{2}$

It follows immediately from lemma 4.8 that the order condition under network observability is $|\mathcal{E}_i| \leq (N-1)/2$. This implies that up to twice as many restrictions are required to account for the unobserved network. Exactly twice as many restrictions are required if $\boldsymbol{s}_i = |\mathcal{E}_i| \leq |\mathcal{E}_i^c|$, in which case the left hand side of part (i) of corollary 4.11 is $2|\mathcal{E}_i|$.

The rank conditions in proposition 4.10 entail restrictions on the network. The following lemma is useful to relate the rank conditions in proposition 4.10 to the properties of the network $\mathcal{G}$.

Lemma 4.12 (Rank and vertex-independent paths)
For two subsets of vertices, $\mathcal{V}_y \subseteq \mathcal{V}$, $\mathcal{V}_x \subseteq \mathcal{V}$, the rank of the sub-matrix of $\mathbf{\Pi}$ with rows $\mathcal{V}_y$ and columns $\mathcal{V}_x$ is less than or equal to the number of vertex-independent paths in $\mathcal{G}$ from $\mathcal{V}_x$ to $\mathcal{V}_y$.

[13] Assumption 4.6 implies that the rows of $\mathbf{\Psi}$ and $\mathbf{\Gamma}$ be sparse. In general, this does not imply that the rows of $\mathbf{\Pi}$ are sparse.

Corollary 4.13 uses lemma 4.12 to link the properties of the underlying network to the point identifiability of the spillovers received by vertex i.

Corollary 4.13 (Vertex-independent paths)
Let $\boldsymbol{\Theta} \in \mathcal{I}(\boldsymbol{s})$ and let assumptions 4.1 and 4.2 be satisfied.

(i) A necessary condition for the rank condition in part (i) of proposition 4.10 is that for every

$$\tilde{\mathcal{E}}_i \in \{\tilde{\mathcal{E}}_i : \tilde{\mathcal{E}}_i \subseteq \mathcal{E}_i^c, |\tilde{\mathcal{E}}_i| = \min[\boldsymbol{s}_i, |\mathcal{E}_i^c|]\}, \tag{4.6}$$

there are $|\mathcal{E}_i \cup \tilde{\mathcal{E}}_i|$ vertex-independent paths in $\mathcal{G}$ from $\mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$ to $\mathcal{E}_i \cup \tilde{\mathcal{E}}_i$.

(ii) In addition, let assumption 4.9 be satisfied for vertex i. Then $\mathcal{C}_i \supseteq \mathcal{C}_i^{\boldsymbol{\Pi}} \supseteq \mathcal{E}_i$ and a necessary condition for the rank condition in part (ii) of proposition 4.10 is that for every

$$\tilde{\mathcal{E}}_i \in \{\tilde{\mathcal{E}}_i : \tilde{\mathcal{E}}_i \subseteq \mathcal{E}_i^c, |\tilde{\mathcal{E}}_i| = \min[\boldsymbol{s}_i, |\mathcal{E}_i^c|]\} \cap \mathcal{C}_i^{\boldsymbol{\Pi}} \tag{4.7}$$

there are $|\mathcal{E}_i \cup \tilde{\mathcal{E}}_i|$ vertex-independent paths in $\mathcal{G}$ from $\mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$ to $\mathcal{E}_i \cup \tilde{\mathcal{E}}_i$.

Corollary 4.13 shows that point identification is attainable for networks which are sparse yet suitably connected. For a given network $\mathcal{G}$ and vertex $i \in \mathcal{V}$, the identification conditions in corollary 4.13 may be checked using standard graph theoretic software packages.[14] This is useful so as to gain an understanding of the types of networks for which identification is possible, even if the network is not observed in practice. Checking part (i) involves verifying the vertex-independent paths criteria for $|\mathcal{E}_i^c|$ choose $\min[\boldsymbol{s}_i, |\mathcal{E}_i^c|]$ sets of feasible neighbors. Checking part (ii) involves verifying the criteria over $|\mathcal{E}_i^c \cap \mathcal{C}_i^{\boldsymbol{\Pi}}|$ choose $\min[\boldsymbol{s}_i, |\mathcal{E}_i^c \cap \mathcal{C}_i^{\boldsymbol{\Pi}}|]$ sets.

As an example, consider identification of $\boldsymbol{\Theta}_{1,\cdot}$ if $\mathcal{G}$ is as depicted in figure 2. Suppose that we apply the sparsity restriction $|\mathcal{E}_1| \leq 1$. The neighbors are $\mathcal{E}_1 = \{4\}$, and the union of neighbors and observationally equivalent neighbors $\mathcal{E}_1 \cup \tilde{\mathcal{E}}_1$ can be $\{4,2\}$, $\{4,3\}$, $\{4,5\}$ or $\{4,6\}$. If $\mathcal{E}_1 \cup \tilde{\mathcal{E}}_1 = \{4,2\}$, there are two vertex-independent paths in $\mathcal{G}$ from $\{3,5,6\}$ to $\{2,4\}$: $(5,\{5,2\},2)$ and $(6,\{6,4\},4)$. By symmetry, the same is true for $\mathcal{E}_1 \cup \tilde{\mathcal{E}}_1 = \{4,3\}$. If $\mathcal{E}_1 \cup \tilde{\mathcal{E}}_1 = \{4,5\}$, there are also 2 vertex independent paths from $\{2,3,6\}$ to $\{4,5\}$: $(2,\{2,5\},5)$ and $(6,\{6,4\},4)$. By symmetry, the same is true for $\mathcal{E}_1 \cup \tilde{\mathcal{E}}_1 = \{4,6\}$. Hence, the network $\mathcal{G}$ satisfies identification condition in part (i) of corollary 4.13.

4.2. Exclusion Restrictions

This section studies a particular set of exclusion restrictions, which may be natural in some applications. The next proposition states the main result.

Proposition 4.14 (Identification using exclusion restrictions)
Let $\boldsymbol{\Theta} \in \mathcal{I}$ and let assumptions 4.1 and 4.2 be satisfied. $\boldsymbol{\Theta}$ is point identified if there is a covariate indexed by $k \in \{1,...,K\}$ such that the corresponding $N \times N$ sub-matrix of $\boldsymbol{\Gamma}$, denoted $\boldsymbol{\Gamma}_{(k)}$, is diagonal and has full rank.

Proposition 4.14 is useful if the researcher knows in advance that covariate k does not generate contextual effects. Restricting $\boldsymbol{\Gamma}_{(k)}$ to be diagonal places $N(N-1)$ exclusion restrictions.

[14] The Bioinformatics toolbox in MATLAB is useful for this.

This implies that exogenous variation in $\boldsymbol{X}_{(k)jt}$ may be used as an instrument for $\boldsymbol{Y}_{jt}$ in equation $i \neq j$. Proposition 4.14 is useful since the identification result holds for any network governing the endogenous effects, regardless of its properties.

This approach necessitates that the researcher apply economic theory to specify a covariate which determines individual outcomes but does not generate contextual effects. This is often the case in settings where there are vertex-specific determinants of the cost and/or benefit of the outcome.

4.3. Unobserved Heterogeneity

In this section I study unobserved heterogeneity which may cause the moment condition in assumption 4.5 to fail. For vertex i in period t, $\boldsymbol{\epsilon}_{it}$ may be decomposed as:

$$\boldsymbol{\epsilon}_{it} = \boldsymbol{\alpha}_i + \boldsymbol{\nu}_{it} \tag{4.8}$$

This specification models unobserved hetereogeneity as an additive function of fixed vertex-specific heterogeneity and the remaining vertex-period component. No restrictions are placed on $\boldsymbol{\alpha}_i$, which may be arbitrarily correlated with $\boldsymbol{X}, \boldsymbol{\nu}$ and $\boldsymbol{\Theta}$. Applying a transformation to the baseline model with the error structure in (4.8) yields:

$$\boldsymbol{Y}\boldsymbol{W} = \boldsymbol{\Psi}\boldsymbol{Y}\boldsymbol{W} + \boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{W} + \boldsymbol{\nu}\boldsymbol{W} \tag{4.9}$$

where $\boldsymbol{W}$ is any transformation matrix with T rows such that $\boldsymbol{\iota}_T'\boldsymbol{W} = \boldsymbol{0}$. In practice, one typically uses within-groups, first differences or forward orthogonal deviations.

All of the identification results stated thus far continue to apply if assumption 4.5 is replaced by the following assumption and the identified set $\mathcal{I}(\boldsymbol{s})$ is modified accordingly.

Assumption 4.15 (Strict exogeneity with unobserved heterogeneity)

$$\mathbb{E}\left[\boldsymbol{\nu}_{\cdot,t} | \boldsymbol{Z}\right] = \boldsymbol{0}$$

Beyond the conditional moment restriction in assumption 4.15, no additional assumptions on $\boldsymbol{\nu}$ are necessary for the purposes of identification. Nevertheless, for the purposes of estimation, model selection and inference, the next section assumes that $\boldsymbol{\nu}_{it}$ is i.i.d. over t. This is to apply the results of Gautier and Tsybakov (2014), and can also be relaxed to an i.n.i.d. assumption, though for brevity, I do not discuss this further.[15]

Whilst the i.i.d. assumption may appear quite strong, it places no restrictions on the cross-sectional correlation in the unobservables. Indeed, it allows for unrestricted dependence between $\boldsymbol{\nu}_{it}$ and $\boldsymbol{\nu}_{js}$ for $j \neq i$. This allows for general forms of cross-sectional dependence, which is more pertinent than temporal dependence in the spillovers setting.

The model of unobserved heterogeneity in (4.8) does not allow for common period level heterogeneity, in which case (4.8) would include an additional $\boldsymbol{\lambda}_t$ term. In this case, the transformed model is:

$$\boldsymbol{V}\boldsymbol{Y}\boldsymbol{W} = \boldsymbol{V}\boldsymbol{\Psi}\boldsymbol{Y}\boldsymbol{W} + \boldsymbol{V}\boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{W} + \boldsymbol{V}\boldsymbol{\nu}\boldsymbol{W} \tag{4.10}$$

[15] For details, see section 5 of Gautier and Tsybakov (2014).

where $\boldsymbol{V}$ is any transformation matrix with N columns such that $\boldsymbol{V}\boldsymbol{\iota}_N = \boldsymbol{0}$. I do not consider this type of unobserved heterogeneity, since the transformation $\boldsymbol{V}$ is applied to the parameters rather than the data, which leads to significant complications in the identification and estimation approach. This type of heterogeneity is permitted in Gautier (2015) and Gautier and Rose (2016), which are discussed in detail in section 8.

5. Estimation, Model Selection & Inference

The results thus far have studied the role of sparsity in obtaining point identification, through showing that it may induce an injective mapping from the structural parameters to the reduced form parameters. In this section, the focus is on exploiting sparsity to address lack of data. To do this, I apply the STIV estimator of Gautier and Tsybakov (2014). The remainder of this section demonstrates the approach for the baseline model augmented to include vertex-fixed effects and transformed according to (4.9).

If T is small relative to N, standard instrumental variables approaches are inapplicable. In each equation, in the absence of exclusion restrictions there are $N-1+NK$ parameters, NL instruments and T observations. If $N-1+NK > \min(T, NL)$ or $NL > T$, the linear systems on which IV and GMM are based are rank deficient. This is the case if $\boldsymbol{Z} = \boldsymbol{X}$ or if T is small relative to N. These cases are typical for many applications, including the empirical application in section 7.

The STIV estimator is applicable in high-dimensional linear settings with many endogenous covariates and many instruments. The number of regressors and instruments may be large relative to the sample size. Moreover, it is not necessary to specify which instruments are excluded from the right hand side in advance, such that $\boldsymbol{X} = \boldsymbol{Z}$ is permitted and no exclusion restrictions need be made.

The estimator searches for a sparse parameter vector among a 'small' set of parameters which satisfy a relaxation of the sample analogue of the moment condition. The moment condition is relaxed since attempting to find an exact solution can induce a large error if the system is rank deficient. For equation i the 'small' set is:

$$\widehat{\mathcal{I}}(r\sigma_i) = \left\{ \boldsymbol{\Theta}_{i,\cdot} : T^{-1} \| (\boldsymbol{Y}_{i,\cdot} - \boldsymbol{\Psi}_{i,\cdot}\boldsymbol{Y} - \boldsymbol{\Gamma}_{i,\cdot}\boldsymbol{X})\boldsymbol{W}\boldsymbol{Z}' \|_\infty \leq r\sigma_i \right\} \tag{5.1}$$

where $r > 0$ is computed from the data as discussed below and σ_i is the unknown level of the noise $T^{-1/2}\|(\boldsymbol{Y}_{i,\cdot} - \boldsymbol{\Psi}_{i,\cdot}\boldsymbol{Y} - \boldsymbol{\Gamma}_{i,\cdot}\boldsymbol{X})\boldsymbol{W}\|_2$. For clarity, I omit the the rescaling of the data in Gautier and Tsybakov (2014) from the exposition, though it is applied in the simulations and empirical application.[16]

The STIV estimator searches for sparse parameter vectors in the set $\mathcal{I}(r\sigma_i)$ by minimizing a sparsity inducing criterion. The estimator for $(\boldsymbol{\Theta}_{i,\cdot}, \sigma_i)$ is defined as a solution to the following

[16] To ensure that the method is invariant to the scale of the data, Gautier and Tsybakov (2014) rescale the data such that each row of $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}$ has ℓ_2 norm equal to $\sqrt{T}$.

conic program:

$$(\widehat{\boldsymbol{\Theta}}_{i\cdot}, \widehat{\sigma}_i) \in \underset{\substack{\boldsymbol{\Theta}_{i,\cdot} \in \widehat{\mathcal{I}}(r\sigma_i) \\ \widehat{Q}(\boldsymbol{\Theta}_{i,\cdot}) \leq \sigma_i^2}}{\arg\min} \|\boldsymbol{\Theta}_{i,\cdot}\|_1 + c\sigma_i \tag{5.2}$$

where $c > 0$ is a tuning parameter which controls the sparsity and

$$\widehat{Q}(\boldsymbol{\Theta}_{i,\cdot}) = T^{-1}\|(\boldsymbol{Y}_{i,\cdot} - \boldsymbol{\Psi}_{i,\cdot}\boldsymbol{Y} - \boldsymbol{\Gamma}_{i,\cdot}\boldsymbol{X})\boldsymbol{W}\|_2^2 \tag{5.3}$$

is the noise level.

The estimator penalizes the ℓ_1 norm of the parameter vector and the standard deviation of the disturbances whilst ensuring that any violation of the moment conditions is small. This leads to a sparse solution which approximately satisfies the moment conditions. To impose assumptions 4.3 and 4.4, I restrict $\boldsymbol{\Psi}_{ii} = 0$ and $\sum_{j=1}^{N} \boldsymbol{\Psi}_{ij} \in (-1, 1)$ when estimating equation i.[17] It is straightforward to impose exclusion restrictions in a similar manner. I exempt $\boldsymbol{\Gamma}_{ii}$ from the ℓ_1 penalty, as it is only the network which is assumed to be sparse.

The value of r is chosen such that

$$\boldsymbol{\Theta}_{i,\cdot} \in \widehat{\mathcal{I}}\left(r\sqrt{\widehat{Q}(\boldsymbol{\Theta}_{i,\cdot})}\right) \quad \forall i \in \mathcal{V} \tag{5.4}$$

with probability at least $1 - \alpha$ for some pre-prescribed α. A reference choice is

$$r \sim \sqrt{\frac{\log(NL)}{T}} \tag{5.5}$$

To compute r exactly, Gautier and Tsybakov (2014) suggest five different sets of distributional assumptions covering the i.i.d. case and various i.n.i.d. cases. Each set of distributional assumptions leads to a different choice of r. Here, I take r equal to the $1 - \alpha/N$ quantile of $T^{-1}|\boldsymbol{e}'\boldsymbol{W}\boldsymbol{Z}|_\infty$, where $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_T)$.[18] This statistic is straightforward to compute by simulation. This choice is based on corollary 2.1 in Chernozhukov et al. (2013) and delivers

$$\mathbb{P}\left(\boldsymbol{\Theta}_{i,\cdot} \in \widehat{\mathcal{I}}\left(r\sqrt{\widehat{Q}(\boldsymbol{\Theta}_{i,\cdot})}\right) \quad \forall i \in \mathcal{V}\right) \geq 1 - \alpha \tag{5.6}$$

asymptotically as $N, T \to \infty$ under the following assumption:

Assumption 5.1 (Data generating process)
The disturbances $\boldsymbol{\nu}_{it}$ are i.i.d. over t with bounded fourth moments, the transformed instruments $(\boldsymbol{ZW})_{\cdot,t}$ are bounded and independent of $\boldsymbol{\nu}_{it}$ and there exist B_T and constants $\bar{C} > 0$ and $\bar{c} > 0$ such that $B_T^4 \log(NLT)^7/T \leq \bar{C}T^{-\bar{c}}$.

Choosing r in this way allows for unrestricted correlation of $\boldsymbol{\nu}_{it}$ with $\boldsymbol{\nu}_{js}$ for $j \neq i$. In addition, it allows for heteroskedasticity of the form $\mathbb{E}\left[\boldsymbol{\nu}_{it}^2\right] = \sigma_i^2$. It does, however, rule out

[17] The linear restriction $\sum_{j=1}^{N} \boldsymbol{\Psi}_{ij} \in (-1, 1) \quad \forall i \in \mathcal{V}$ is sufficient for $\det(\boldsymbol{I}_N - \boldsymbol{\Psi}) \neq 0$ due to diagonal dominance. It is convenient since it preserves the convexity of the optimization problem and does not involve cross-equation restrictions.

[18] I use the $1 - \alpha/N$ quantile to guarantee a joint confidence level of $1 - \alpha$ using a union bound. This is because we wish to control the probability of the intersection of N dependent events.

serial correlation in ν_{it} and time varying heteroskedasticity. In principle it is feasible to relax independence over t using notions of weak dependence. This would result in a larger value of r.

Under assumption 5.1 and an identifiability assumption,[19] for $c \in (0, 1/r)$ and fixed K and L, Gautier and Tsybakov (2014) show that:

$$\|\mathbf{\Theta}_{i,\cdot} - \widehat{\mathbf{\Theta}}_{i,\cdot}\|_2 \leq \mathcal{O}\left(\sqrt{|\mathcal{E}_i| \log(N)/T}\right) \tag{5.7}$$

with probability at least $1 - \alpha$. This means that even though the dimension of $\mathbf{\Theta}_{i,\cdot}$ is linear in N, under sparsity, the rate depends only on $|\mathcal{E}_i| \log(N)$, which can be much smaller than N. In particular, consistency requires $|\mathcal{E}_i| \log(N)/T \to 0$, which can be achieved even as $N/T \to \infty$.

Gautier and Tsybakov (2014) also derive model selection results. Under the same assumptions as for the rate, and an additional assumption that the absolute values of the nonzero components of $\mathbf{\Theta}_{i,\cdot}$ are sufficiently large (see theorem 7.1 (iii)), it can also be shown that the estimated set of neighbors

$$\widehat{\mathcal{E}}_i = \left\{ j \in \mathcal{V} : j \neq i, \widehat{\boldsymbol{\psi}}_{ij} \neq 0 \text{ or } \widehat{\boldsymbol{\gamma}}_{ij} \neq \mathbf{0} \right\} \tag{5.8}$$

is a superset of $\mathcal{E}_i$ with high probability. Under a stronger assumption on the size of the spillovers (see assumption 8.1 and theorem 8.1), one obtains $\check{\widehat{\mathcal{E}}}_i = \mathcal{E}_i$ with high probability, where

$$\check{\widehat{\mathcal{E}}}_i = \left\{ j \in \mathcal{V} : j \neq i, |\widehat{\boldsymbol{\psi}}_{ij}| > \boldsymbol{\omega}_{ij}^{\psi} \text{ or } |\widehat{\boldsymbol{\gamma}}_{ij}| > \boldsymbol{\omega}_{ij}^{\gamma} \right\} \tag{5.9}$$

is a thresholded estimator and $\boldsymbol{\omega}_{ij}^{\psi}$, $\boldsymbol{\omega}_{ij}^{\gamma}$ are thresholds computed from the data. Consequently, by applying the STIV estimator one can exactly recover the true network provided that the spillovers are sufficiently large.

The inference procedure of Gautier and Tsybakov (2014) is based on (5.6). It applies equally under point identification, partial identification and non-identification and is uniform over the class of data generating processes compatible with assumption 5.1. Robustness to identification implies that the confidence sets may have infinite volume if there is insufficient sparsity, and/or the instruments are weak.

6. Simulations

This section studies the performance of the method for realistic data generating processes. The exposition focuses on estimation and model selection. This is because the confidence sets are too large to be informative, an issue which is discussed in detail in section 8.

To generate the data, I set $N = 100$, $K = 1$ and generate a random directed network with adjacency matrix $\boldsymbol{G}$, where all edges are i.i.d. with link probability p and there are no edges from a vertex to itself. I then take every row of $\boldsymbol{G}$ with at least one nonzero entry, and rescale so that the elements sum to one. This implies that the rows of $\boldsymbol{G}$ sum to zero or one, which is

[19] A precise statement of the identifiability assumption requires a significant amount of additional notation. It is stated in assumption 7.1 of Gautier and Tsybakov (2014).

a common assumption in the social effects literature (see, for example, Blume et al. (2015) or de Paula et al. (2016)). I study two cases: one in which the network is fixed over the data sets, and another in which the network is redrawn in each dataset.

I set $\boldsymbol{\Psi} = 0.5\boldsymbol{G}$ and $\boldsymbol{\Gamma} = -2(\boldsymbol{G} + \boldsymbol{I}_N)$. This specification implies that for those vertices which are not isolated, the outcomes depend on the mean outcome of their neighbors, the mean covariate of their neighbors and their own covariate with respective weights 0.5, -2 and -2.

I let $\boldsymbol{X}_{it}$ follow a multivariate normal distribution, with $\mathbb{E}[\boldsymbol{X}_{it}] = 0$ for every i and t and

$$\mathbb{COV}[\boldsymbol{X}_{it}, \boldsymbol{X}_{js}] = \begin{cases} 1 & \text{if } i = j \text{ and } s = t \\ \boldsymbol{\rho}/N & \text{if } i \neq j \text{ and } s = t \\ \boldsymbol{\rho} & \text{if } i = j \text{ and } |s - t| = 1 \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

I let $\boldsymbol{\epsilon}_{it} = \boldsymbol{\alpha}_i + \boldsymbol{\nu}_{it}$, where $\boldsymbol{\alpha}_i = 1 + T^{-1}\sum_{t=1}^{T} \boldsymbol{X}_{it}$ and $\boldsymbol{\nu}_{it}$ follows a multivariate normal distribution (independent of $\boldsymbol{X}$) with $\mathbb{E}[\boldsymbol{\nu}_{it}] = 0$ for every i and t and

$$\mathbb{COV}[\boldsymbol{\nu}_{it}, \boldsymbol{\nu}_{js}] = \begin{cases} 1 & \text{if } i = j \text{ and } s = t \\ \boldsymbol{\rho}/N & \text{if } i \neq j \text{ and } s = t \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

I set $\boldsymbol{\rho} = 0.5$ to allow for moderate dependence in the data and apply the within-groups transformation $\boldsymbol{W} = \boldsymbol{I}_T - \boldsymbol{\iota}_T\boldsymbol{\iota}_T'/T$, leading to the outcome equation in (4.9). The instruments are $\boldsymbol{Z} = \boldsymbol{X}$. The design implies that the number of parameters in each equation is $2N - 1 = 199$, and the number of instruments is $N = 100$. Since there are more parameters than instruments, conventional instrumental variables estimation is inapplicable.

To vary the sample size and sparsity, I study each configuration over $(T, p) \in \{50, 100, 200, 500\} \times \{1/N, \log(N)/N\}$. The values for T span the cases where the number of observations is much less than, approximately equal to, and greater than the number of parameters. For $T = 50$, the number of instruments is also much greater than the sample size. The choice of $p = \log(N)/N$ coincides with the asymptotic threshold for strong connectedness of directed random graphs.

For each configuration, I apply the STIV estimator with $c = 0.99/r$. For comparison, I also compute the infeasible 2SLS estimator based on the true network for configurations with $T \geq 100$.[20] I do this for 1000 data sets, over which I compute percentiles of the statistics reported below.

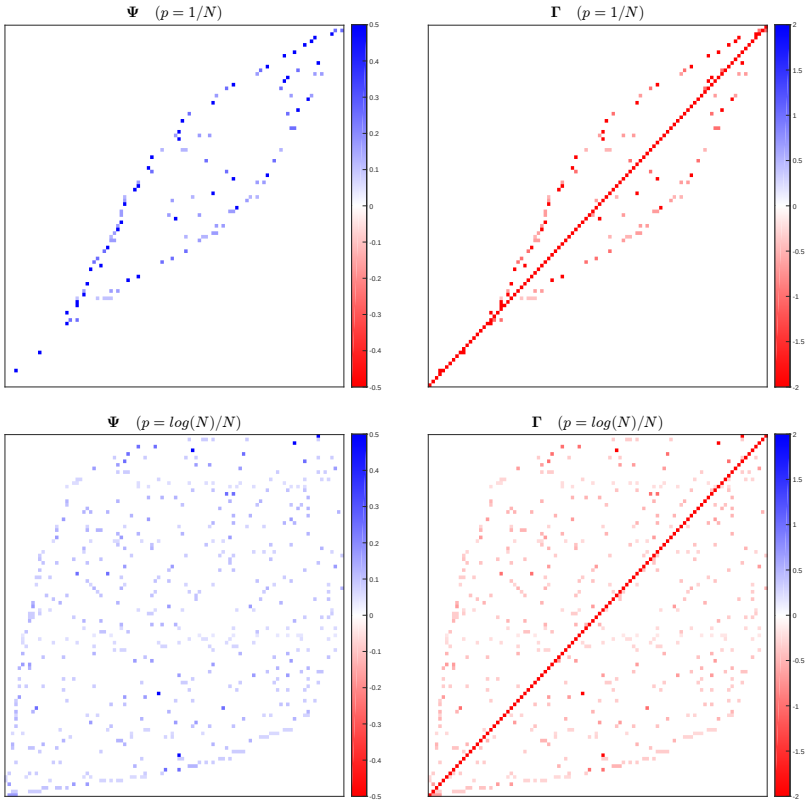
6.1. Results for a Fixed Network

This sub-section presents estimation results for a specific network, which is held fixed over the data sets. The true parameters are depicted in figure 3. For $p = 1/N$, each vertex has around 1 neighbor on average, with a maximum of 5 and a minimum of 0. For $p = \log(N)/N$,

[20] Applicability of the 2SLS estimator requires that the number of instruments $N = 100$ does not exceed the sample size T.

each vertex has around 5 neighbors on average, with a maximum of 12 and a minimum of 0. These choices imply that the number of nonzero parameters in each equation is at most 11 for $p = 1/N$ and 25 for $p = \log(N)/N$. Both networks satisfy the identification condition in part (i) of corollary 4.13.

Figure 3: $\mathbf{\Psi}$ *and* $\mathbf{\Gamma}$ *with edge probabilities* $1/N$ *and* $\log(N)/N$



Notes: $\mathbf{\Psi}$ and $\mathbf{\Gamma}$ are generated as follows. First, generate the $N \times N$ adjacency matrix $\boldsymbol{G}$ of a random graph with i.i.d. edge probability p. Next, rescale the nonzero rows of $\boldsymbol{G}$ to sum to 1 and specify $\mathbf{\Psi} = 0.5\boldsymbol{G}$ and $\mathbf{\Gamma} = -2(\boldsymbol{G} + \boldsymbol{I}_N)$. For clarity of presentation the vertices are reordered to concentrate the mass around the 45 degree line.

Figures 8-15 in the appendix depict percentiles of the point estimates and estimation errors over the configurations. The performance of the estimator depends on the sample size and the sparsity in the expected way. When the sample size is small or the network is dense, the ℓ_1 penalty in the estimator leads to greater shrinkage, and hence median point estimates for the nonzero parameters are closer to zero. As the sample size or sparsity increase, median point estimates for the nonzero parameters move closer to their true values. For $p = 1/N$, aside from the shrinkage, the estimator performs reasonably well in terms of the signs and locations of edges even when $T = 100$ or $T = 50$. This is not the case for $p = \log(N)/N$.

6.2. Results for Random Networks

In this sub-section, the network is redrawn for every data set. This ensures that the results are not driven by any particular choice of network. Table 1 summarizes the model selection results over the configurations. For each data set, I compute the proportion of parameters which are correctly and incorrectly selected, and take percentiles of these statistics over all of the data

sets. I exclude the diagonal elements of $\mathbf{\Gamma}$ as these are exempt from the ℓ_1 penalty.

Table 1: *Percentiles of selection frequencies*

		$p = 1/N$		$p = \log(N)/N$	
	Percentile	$\widehat{\mathbf{\Theta}}_{ij} \neq 0 \vert \mathbf{\Theta}_{ij} \neq 0$	$\widehat{\mathbf{\Theta}}_{ij} \neq 0 \vert \mathbf{\Theta}_{ij} = 0$	$\widehat{\mathbf{\Theta}}_{ij} \neq 0 \vert \mathbf{\Theta}_{ij} \neq 0$	$\widehat{\mathbf{\Theta}}_{ij} \neq 0 \vert \mathbf{\Theta}_{ij} = 0$
$T = 500$	05	0.98	0.00	0.93	0.00
	50	**0.99**	**0.00**	**0.95**	**0.01**
	95	0.99	0.00	0.97	0.01
$T = 200$	05	0.97	0.00	0.77	0.00
	50	**0.99**	**0.00**	**0.81**	**0.01**
	95	0.99	0.00	0.84	0.01
$T = 100$	05	0.92	0.00	0.53	0.00
	50	**0.96**	**0.00**	**0.58**	**0.01**
	95	0.98	0.00	0.63	0.01
$T = 50$	05	0.77	0.00	0.28	0.00
	50	**0.85**	**0.00**	**0.32**	**0.00**
	95	0.90	0.00	0.37	0.01

Notes: The STIV estimator is applied with $c = 0.99/r$. The $\widehat{\mathbf{\Theta}}_{ij} \neq 0 \vert \mathbf{\Theta}_{ij} \neq 0$ column reports the selection frequency of the nonzero spillover parameters $\left(\sum_{i,j \neq i+N} \mathbf{1}_{\widehat{\mathbf{\Theta}}_{ij} \neq 0, \mathbf{\Theta}_{ij} \neq 0}\right) / \left(\sum_{i,j \neq i+N} \mathbf{1}_{\mathbf{\Theta}_{ij} \neq 0}\right)$. The $\widehat{\mathbf{\Theta}}_{ij} \neq 0 \vert \mathbf{\Theta}_{ij} = 0$ column reports the selection frequency of the nonzero parameters $\left(\sum_{i,j \neq i+N} \mathbf{1}_{\widehat{\mathbf{\Theta}}_{ij} \neq 0, \mathbf{\Theta}_{ij} = 0}\right) / \left(\sum_{i,j \neq i+N} \mathbf{1}_{\mathbf{\Theta}_{ij} = 0}\right)$. The condition $j \neq i + N$ is included in the sum to ignore the diagonal elements of $\mathbf{\Gamma}$, which are exempt from the ℓ_1 penalty. Percentiles are taken over 1000 simulations.

Model selection performance depends on the sample size and the sparsity in the expected way. For $T = 500$ and $p = 1/N$, the nonzero parameters are selected with very high frequency, the median of which is 0.99. As the sample size decreases, the frequency of correct selection falls, though it remains at 0.96 for $T = 100$ and 0.85 for $T = 50$. For $p = \log(N)/N$, the frequency of correct selection is lower than for $p = 1/N$ for every sample size. Its median falls from 0.95 for $T = 500$ to 0.32 for T=50. Across all configurations, the frequency of incorrect selection is close to zero.

Table 2 reports percentiles of the estimation errors for the low dimensional parameters

$$\bar{\mathbf{\Psi}} = N^{-1} \sum_{i,j \neq i} \mathbf{\Psi}_{ij} \tag{6.3}$$

$$\bar{\mathbf{\Gamma}} = N^{-1} \sum_{i,j \neq i} \mathbf{\Gamma}_{ij} \tag{6.4}$$

$$\bar{\mathbf{\Delta}} = N^{-1} \sum_{i} \mathbf{\Gamma}_{ii} \tag{6.5}$$

which are the main quantities of interest in Manski (1993); Moffitt (2001); Lee (2007); Davezies et al. (2009); Bramoullé et al. (2009); Blume et al. (2015); Lam and Souza (2013); Manresa (2014) and de Paula et al. (2016). The true values vary in each data set due to redrawing of the network. For this reason, table 2 reports estimation errors standardized by the true value. Shrinkage leads point estimates of $\bar{\mathbf{\Psi}}$ and $\bar{\mathbf{\Gamma}}$ to be too small in absolute value in all specifications. The shrinkage declines as the number of observations and sparsity increases, and is larger for $\bar{\mathbf{\Gamma}}$ than for $\bar{\mathbf{\Psi}}$. The parameter $\bar{\mathbf{\Delta}}$ is precisely estimated across all configurations. This is most likely because there is no model selection uncertainty around the diagonal elements of $\mathbf{\Gamma}$, which are

exempt from the penalty.

Table 2: *Percentiles of the estimation error of low dimensional parameters*

		$p = 1/N$			$p = \log(N)/N$		
	Percentile	$\frac{\widehat{\bar{\Psi}}-\bar{\Psi}}{\bar{\Psi}}$	$\frac{\widehat{\bar{\Gamma}}-\bar{\Gamma}}{\bar{\Gamma}}$	$\frac{\widehat{\bar{\Delta}}-\bar{\Delta}}{\bar{\Delta}}$	$\frac{\widehat{\bar{\Psi}}-\bar{\Psi}}{\bar{\Psi}}$	$\frac{\widehat{\bar{\Gamma}}-\bar{\Gamma}}{\bar{\Gamma}}$	$\frac{\widehat{\bar{\Delta}}-\bar{\Delta}}{\bar{\Delta}}$
$T = 500$	05	-0.08	-0.12	-0.02	-0.07	-0.40	-0.01
	50	**-0.06**	**-0.09**	**-0.01**	**-0.05**	**-0.36**	**0.00**
	95	-0.04	-0.08	-0.01	-0.02	-0.32	0.00
$T = 200$	05	-0.17	-0.18	-0.01	-0.22	-0.59	-0.01
	50	**-0.13**	**-0.14**	**0.00**	**-0.15**	**-0.55**	**0.00**
	95	-0.09	-0.12	0.01	-0.12	-0.50	0.01
$T = 100$	05	-0.27	-0.27	-0.01	-0.47	-0.76	0.00
	50	**-0.22**	**-0.22**	**0.00**	**-0.37**	**-0.71**	**0.01**
	95	-0.17	-0.18	0.01	-0.28	-0.67	0.03
$T = 50$	05	-0.46	-0.43	-0.01	-0.76	-0.89	0.00
	50	**-0.36**	**-0.37**	**0.00**	**-0.66**	**-0.85**	**0.03**
	95	-0.29	-0.31	0.03	-0.53	-0.82	0.05

Notes: The STIV estimator is applied with $c = 0.99/r$. The $\widehat{\bar{\Psi}}$ and $\widehat{\bar{\Gamma}}$ columns report $N^{-1}\sum_{i,j\neq i}\widehat{\Psi}_{ij}$ and $N^{-1}\sum_{i,j\neq i}\widehat{\Gamma}_{ij}$. The $\widehat{\bar{\Delta}}$ column reports $N^{-1}\sum_{i}\widehat{\Gamma}_{ii}$. Percentiles are taken over 1000 simulations.

Table 3 reports percentiles of the standardized ℓ_2 estimation error for the STIV estimator and the infeasible 2SLS estimator of the nonzero spillovers based on the true network. This comparison is useful to quantify the loss in performance attributable to non-observability of the network. The loss is relatively mild for moderate to large sample sizes and high sparsity, but can be large if there is insufficient sparsity or the sample is too small. For $T \geq 200$ and $p = 1/N$, the STIV estimator performs almost as well as the infeasible 2SLS estimator. The performance gap widens as T decreases and as the sparsity decreases.

Table 3: *Percentiles of the ℓ_2 estimation error for the STIV and infeasible 2SLS estimators*

		$p = 1/N$		$p = \log(N)/N$	
	Percentile	$\frac{\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\|_2}{\|\mathbf{\Theta}\|_2}$	$\frac{\|\widehat{\mathbf{\Theta}}_{\text{IV}}-\mathbf{\Theta}\|_2}{\|\mathbf{\Theta}\|_2}$	$\frac{\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\|_2}{\|\mathbf{\Theta}\|_2}$	$\frac{\|\widehat{\mathbf{\Theta}}_{\text{IV}}-\mathbf{\Theta}\|_2}{\|\mathbf{\Theta}\|_2}$
$T = 500$	05	0.06	0.05	0.16	0.08
	50	**0.07**	**0.06**	**0.17**	**0.09**
	95	0.10	0.13	0.18	0.12
$T = 200$	05	0.09	0.06	0.24	0.12
	50	**0.10**	**0.08**	**0.25**	**0.13**
	95	0.12	0.13	0.27	0.16
$T = 100$	05	0.14	0.08	0.32	0.17
	50	**0.15**	**0.09**	**0.33**	**0.18**
	95	0.18	0.13	0.34	0.21
$T = 50$	05	0.22	-	0.39	-
	50	**0.25**	**-**	**0.41**	**-**
	95	0.28	-	0.42	-

Notes: The STIV estimator is applied with $c = 0.99/r$. The $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2/\|\mathbf{\Theta}\|_2$ column reports the normalized ℓ_2 estimation error for the STIV estimator. The $\|\widehat{\mathbf{\Theta}}_{\text{IV}} - \mathbf{\Theta}\|_2/\|\mathbf{\Theta}\|_2$ column reports the normalized ℓ_2 estimation error for the infeasible 2SLS estimator based on the true network. Percentiles are taken over 1000 simulations.

7. R&D Spillovers & Product Market Rivalry

R&D spillovers have attracted much attention in the industrial organisation, growth and productivity literatures over a sustained period of time. R&D investments have two competing effects on other firms (Bloom et al., 2013). The first is the knowledge spillover effect, through which R&D investments increase other firms' productivity.[21] The second is the product market rivalry effect: R&D investments allow firms to steal business from their rivals.

Bloom et al. (2013) argue that separate identification of the two channels is crucial to conduct welfare analysis and inform policy. The authors' empirical analysis uses measures of firms' positions in technology and product market space, which are constructed using information on the distribution of patenting over technology classes and sales over four digit industry codes. This approach permits estimation of panel regression models with spillover effects. The central result is that both channels are non-negligible and knowledge spillovers dominate product market rivalry. This implies that the social returns to R&D are larger than the private returns, and hence that R&D stocks are below the social optimum.

König et al. (2014) apply a structural framework to study R&D spillovers in a Cournot oligopoly. The authors specify a game in which firms simultaneously choose R&D investments and output given the product market competition and R&D networks, characterise an equilibrium and conduct welfare analysis. The networks are constructed using data on R&D partnerships and industry codes.

The remainder of this section applies the methods developed in this paper to study R&D spillovers in a Cournot oligopoly similar to that of König et al. (2014). The novelty lies in estimating the structure of firm interactions, rather than imposing it. This implies that the results are robust to misspecification of the networks, and permits the analysis of the identities and types of firms which send and receive spillovers.

7.1. Model

There are N firms which make R&D investment decisions and compete on the product market. Firms maximize the stream of expected, discounted future profits. In period t, firm i has a Cobb-Douglas production function of the form:

$$q_{it} = \sum_{j \in \mathcal{E}_i} \mu_{ij} rd_{jt} + \mu_{ii} rd_{it} + \alpha_i l_{it} + \beta_i k_{it} + \eta^q_{it} \tag{7.1}$$

where q_{it} is log output, l_{it} is log labor, k_{it} is log capital and rd_{it} is the log R&D stock, which is determined by firms' R&D investments.[22] Technology spillovers are incorporated through allowing the R&D stocks of other firms to enter the production function through the network $\mathcal{G}$. For log wage w_{it} and log rental rate of capital r_{it}, the log total cost is:

$$tc_{it} = \frac{q_{it} - \sum_{j \in \mathcal{E}} \mu_{ij} rd_{jt} - \mu_{ii} rd_{it} + \alpha_i w_{it} + \beta_i r_{it} - \eta^q_{it}}{\alpha_i + \beta_i} + \ln\left(\left(\frac{\alpha_i}{\beta_i}\right)^{\frac{\beta_i}{\alpha_i+\beta_i}} + \left(\frac{\beta_i}{\alpha_i}\right)^{\frac{\alpha_i}{\alpha_i+\beta_i}} \right) \tag{7.2}$$

[21] As in Bloom et al. (2013); König et al. (2014), I consider a setting in which R&D increases productivity. In some settings it may be more appropriate to suppose that R&D investments increase demand.

[22] For a description of the R&D accumulation process, see equation (7.8) and the discussion thereunder.

In each period firms engage in Cournot competition on the product market. The inverse demand function takes the form:

$$p_{it} = \sum_{j \in \mathcal{E}_i} \kappa_{ij} q_{jt} + \kappa_{ii} q_{it} + \eta^p_{it} \tag{7.3}$$

where p_{it} is the log price. For each firm to have a unique best response function, it is necessary to assume that there are constant returns to labor and capital ($\beta_i + \alpha_i = 1 \quad \forall i \in \mathcal{V}$) and that the own-price elasticity of demand is nonzero ($\kappa_{ii} < 0 \quad \forall i \in \mathcal{V}$). Under these assumptions, it is straightforward to show that the Cournot best response function is:

$$q_{it} = \left(\sum_{j \in \mathcal{E}_i} \Psi_{ij} q_{jt} + \Gamma_{ij} rd_{jt} \right) + \Gamma_{ii} rd_{it} + \tau_{it} \tag{7.4}$$

where $\Psi_{ij} = -\kappa_{ij}/\kappa_{ii}$, $\Gamma_{ij} = -\mu_{ij}/\kappa_{ii}$ and:

$$\tau_{it} = -\frac{1 + \kappa_{ii} + \ln\left(\left(\frac{\alpha_i}{1-\alpha_i} \right)^{1-\alpha_i} + \left(\frac{1-\alpha_i}{\alpha_i} \right)^{\alpha_i} \right)}{\kappa_{ii}} + \frac{\alpha_i w_{it} + (1-\alpha_i) r_{it} - \eta^p_{it} - \eta^q_{it}}{\kappa_{ii}} \tag{7.5}$$

One can then decompose the log wage $w_{it} = \phi^w_i + \nu^w_{it}$ and similarly for the log rental rate r_{it} and the productivity and demand shocks η^q_{it} and η^p_{it}, yielding:

$$q_{it} = \left(\sum_{j \in \mathcal{E}_i} \Psi_{ij} q_{jt} + \Gamma_{ij} rd_{jt} \right) + \Gamma_{ii} rd_{it} + \phi_i + \nu_{it} \tag{7.6}$$

To transform out the fixed effects, I use forward orthogonal deviations for W, and estimate:

$$qW = \Psi qW + \Gamma rdW + \nu W \tag{7.7}$$

7.2. Instrumental Variables

Following Bloom et al. (2013), the R&D stocks are assumed to be determined by the capital accumulation process:

$$RD_{it} = (1-\delta) RD_{it-1} + I_{it} \tag{7.8}$$

where RD_{it} is R&D stock, I_{it} is R&D investment and δ is the depreciation rate. Since firms endogenously allocate R&D investments to maximise expected future profits, the R&D stocks are endogenous. In order to identify the parameters of the best response function, it is necessary to specify instrumental variables.

Due to the capital formation process (7.8), candidate instruments are current and past values of any covariate which determines the cost of R&D investment, as well as lags of the R&D stock. Following Bloom et al. (2013), I use tax induced changes to the cost of R&D.[23] The tax price component of the cost of R&D for firm i in state s in period t is given by $\rho_{it} = (1 - D_{it}) \,/\, (1 - \tau_{st})$ where D_{it} is the discounted value of R&D tax credits and τ_{st} is the

[23] Full details of this approach can be found in Appendix B.3 of Bloom et al. (2013).

rate of corporation tax. If $\rho_{it} = 1$, R&D is tax neutral, whilst $\rho_{it} < 1$ implies that there are tax incentives for R&D.

The tax component varies at the firm-year level for two reasons, each of which can be exploited to specify instruments. First, firms typically conduct R&D across many states, each with different R&D tax credits and rates of corporation tax. Since firms have differential distributions of R&D activity over states, changes to tax credits or corporation taxes have heterogeneous impacts across firms. Using the distribution of R&D activity over states, Bloom et al. (2013) construct a firm-specific measure of the 'state tax credit component', which is adopted here. Second, the federal rules pertaining to what type of activity is permissible for R&D tax credits depend on a firm-specific base level, which is determined by the firm's past R&D investments. Bloom et al. (2013) construct a firm-specific measure of the 'federal tax credit component' based on these rules, which is also adopted here.

Due to the capital formation process (7.8), one may also use lags of the log R&D stock as instruments for its present value. This is possible owing to the use of forward orthogonal deviations in transforming the model. One then requires that the log R&D stock be uncorrelated with the future disturbances. In addition to the tax based instruments, I also use the first lag of the log R&D stock as an instrument, and construct the 5×1 vector of instrumental variables $\boldsymbol{Z}_{it}$ from the natural logarithms of the period t and $t-1$ measures of state and federal tax credits for firm i and the period $t-1$ R&D stock for firm i.

A potential concern related to the instruments is that tax policy may be endogenously determined based on macroeconomic conditions. Bloom et al. (2013) argue that there is a substantial degree of randomness in R&D tax credits and find that past changes in R&D expenditures and GDP do not have a statistically significant association with policy. In addition, under the forward orthogonal deviations transformation, we require only that current and past tax credits be uncorrelated with the current and future disturbances.

7.3. Data

The data are identical to Bloom et al. (2013). I obtain firm level accounting data from the U.S. COMPUSTAT 1980-2001 and match it to the U.S. Patent and Trademark Office data available through the NBER. All variables are deflated to 1996 values using the CPI.

R&D investments are observed directly, whilst the R&D stock is calculated using the perpetual inventory method described in Bloom et al. (2013). The perpetual inventory method uses equation (7.8) and assumes that in the first period each firms' stock of knowledge is at the steady-state level $\boldsymbol{K}_1 = \left(\frac{1}{\delta+g}\right)\boldsymbol{I}_1$ where g is the steady-state growth rate in the R&D stock. The values $\delta = 0.15$ and $g = 0.05$ in Bloom et al. (2013) are adopted here. The R&D stock in subsequent periods evolves according to (7.8).

Output is measured by deflating sales using the industry price index,[24] and the measures of state and federal tax credits for R&D are constructed in an identical manner to Bloom et al. (2013). To prevent the number of firms becoming too large relative to the number of years, I use data for a single industry defined by the two digit SIC code for electronics (36). I focus on

[24] A lack of firm-specific prices leads to measurement error in output. Bloom et al. (2013) argue that this is unlikely to be severe.

electronics because there is significant scope for R&D spillovers and the number of firms is not too large relative to the number of observations.

The sample for the electronics industry comprises all firms which have a positive R&D stock over the sample period, yielding a balanced panel of 26 firms observed over 21 years. These firms are inherently large and conduct a significant amount of research. Table 4 summarizes the data.

Table 4: *Summary statistics for the electronics industry from 1980-2001*

Variable	Symbol	Units	Mean	St.Dev.	Min.	Max.
Real sales	Q_{it}	**Million** 1996 $	2254.51	5235.58	20.79	39725.64
Real R&D investments	I_{it}	**Million** 1996 $	186.86	548.19	1.00	4769.00
Real R&D stock	RD_{it}	**Million** 1996 $	869.70	2123.73	3.21	17770.88
State tax credit	$TX1_{it}$	-	1.28	0.12	1.10	1.51
Federal tax credit	$TX2_{it}$	-	0.95	0.07	0.58	1.23
N	26					
T	21					

Notes: The sample comprises 26 firms in the electronics industry observed annually between 1981 and 2001. For the state and federal tax credits, a value of 1 implies that R&D investments are tax neutral, and values less than 1 imply that there are tax incentives for R&D investment.

Due to the high-dimensional setting, the usual first stage statistics for weak identification cannot be computed. Instead, to verify that the instruments are strong predictors of the log R&D stock, I regress the log R&D stock of each firm on its own log tax credit measures and their first lag, and the first lag of its own log R&D stock. That is, I regress rd_{it} on $Z_{it} = (tx1_{it}, tx2_{it}, tx1_{it-1}, tx2_{it-1}, rd_{it-1})'$ for each $i \in \{1, ..., N\}$. The F-statistics associated with these regressions are large. The median F-statistic over the N regressions is 233.44, and the minimum is 5.42. The associated p-values for the null hypothesis that the instruments do not predict the log R&D stock are small, with the largest equal to 0.01. The tax credits instruments are not redundant: even if the $t-1$ log R&D stock is omitted from the right hand side, the instruments remain strong predictors of the log R&D stock, with a median F-statistic of 8.78.

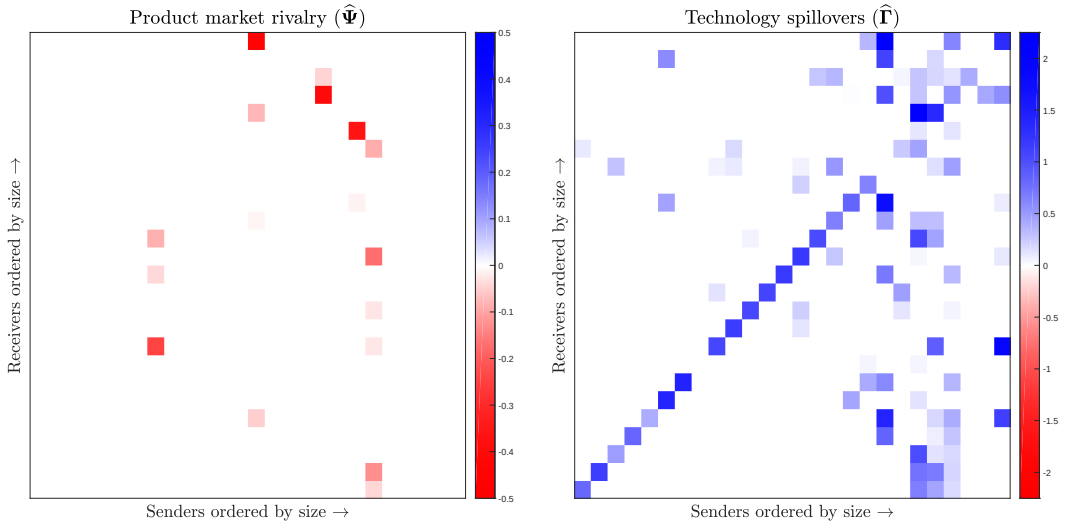
7.4. Results

The transformed Cournot best response function (7.7) is estimated using the STIV estimator with $c = 0.99/r$ under the restrictions $\mathbf{\Psi} \leq \mathbf{0}$ and $\mathbf{\Gamma} \geq \mathbf{0}$.[25] The exposition focuses on estimation and model selection. This is because confidence sets are too large to be informative, which is discussed in detail in section 8. The lack of informative inference implies that point estimates ought to be interpreted with caution, and ought not to be used for policy purposes. Nevertheless, the simulation results in section 6 demonstrate that the point estimates may be relatively reliable provided that there is sufficient sparsity, though they are also likely to be too small in magnitude. Moreover, the estimator may fail to select spillovers which are not large

[25] In the absence of sign restrictions, the majority of the estimated spillovers are correctly signed. However, one large negative estimated R&D spillover has a large impact on the welfare analysis. For this reason, I present results under sign restrictions.

enough. For this reason, the parameter estimates in this section should be interpreted as lower bounds on the extensive and intensive margins of the effects.

Figure 4: *Estimated product market competition and R&D elasticities in the electronics industry*



Notes: Point estimates are for the STIV estimator of (7.7) with $c = 0.99/r$. The sample comprises 26 firms in the electronics industry observed annually between 1981 and 2001. Instruments are the logs of the current tax credits and the first lag of the log R&D stock. The forward orthogonal deviations transformation is applied to remove the firm fixed-effects. Firms are ordered by size, which is measured by total real sales over the sample period.

Figure 4 shows point estimates of the parameters for (7.7). All variables are in natural logarithms, hence parameters may be interpreted as elasticities. On each axis, firms are ordered by size, which I measure by total real sales over the sample period. The mass of R&D spillovers is concentrated on the right hand side of the plot. This means that R&D spillovers are sent predominantly by large firms and received by both large and small firms. Bloom et al. (2013) also find that large firms send more R&D spillovers than small firms. This is because large firms conduct R&D over a relatively broad range of technology classes, and are hence centrally located in the technology network which the authors construct.

There are few nonzero point estimates for the product market competition parameters. This could be due to the relatively small sample size, and/or because few firms exert large competitive pressure on one another. In any case, the estimates of R&D spillovers are larger and more frequently nonzero than those of the product market competition effect, which was also concluded by Bloom et al. (2013).

Table 5 summarizes the magnitudes and selection frequencies of the estimated output elasticities with respect to own R&D, other firms' R&D and other firms' output. The 'Own R&D' row summarizes the elasticity of real output with respect to a firm's own R&D stock. A 10% increase in a firms' R&D stock typically leads to a 7.8% increase in real output.

The 'Extensive margin of R&D spillovers' and 'Intensive margin of R&D spillovers' rows summarize the frequency and magnitudes of the elasticities with respect to other firms' R&D stocks. Around 11% of possible R&D spillovers are estimated to be nonzero. On average, a 10% increase in a neighbor's R&D stock leads to a 4.5% increase in real output. These numbers suggest that R&D spillovers are large, though a firm's own R&D stock has a greater impact on its productivity than the R&D stock of one of its neighbors.

Table 5: *Estimated elasticities of real output and spillovers in the electronics industry*

Description	Definition	(7.7)
Own R&D	$\frac{\sum_i -\mu_{ii}/\kappa_{ii}}{N}$	0.7828
Extensive margin of R&D spillovers	$\frac{\sum_{i,j\neq i} \mathbf{1}_{\lvert\mu_{ij}/\kappa_{ii}\rvert\neq 0}}{N(N-1)}$	0.1169
Intensive margin of R&D spillovers	$\frac{\sum_{i,j\neq i} -\mu_{ij}/\kappa_{ii}}{\sum_{i,j\neq i} \mathbf{1}_{\lvert\mu_{ij}/\kappa_{ii}\rvert\neq 0}}$	0.4496
Extensive margin of product market competition	$\frac{\sum_{i,j\neq i} \mathbf{1}_{\lvert\kappa_{ij}/\kappa_{ii}\rvert\neq 0}}{N(N-1)}$	0.0262
Intensive margin of product market competition	$\frac{\sum_{i,j\neq i} -\kappa_{ij}/\kappa_{ii}}{\sum_{i,j\neq i} \mathbf{1}_{\lvert\kappa_{ij}/\kappa_{ii}\rvert\neq 0}}$	-0.1320

Notes: Point estimates are for the STIV estimator of (7.7) with $c = 0.99/r$. The sample comprises 26 firms in the electronics industry observed annually between 1981 and 2001. Instruments are the logs of the current tax credits and the first lag of the log R&D stock.

The 'Extensive margin of product market competition' and 'Intensive margin of product market competition' rows summarize the frequency and magnitudes of the elasticities with respect to other firms' real output. Around 3% of possible competitive relationships are estimated to be nonzero. On average, a 10% increase in the real output of a neighbor leads to a 1.3% decrease in real output. These numbers appear to suggest that there is weak product market competition in the electronics industry. However, due to the small sample size and the shrinkage applied by the estimator, it is likely that the estimated elasticities are biased towards zero.

7.5. Welfare

The remainder of this section uses parameter estimates to conduct welfare analyses. Following Bloom et al. (2013), the marginal private return and marginal social return for firm i in period t are defined as the increase in private and aggregate output attributed to a marginal increase in firm i's period t stock of R&D:

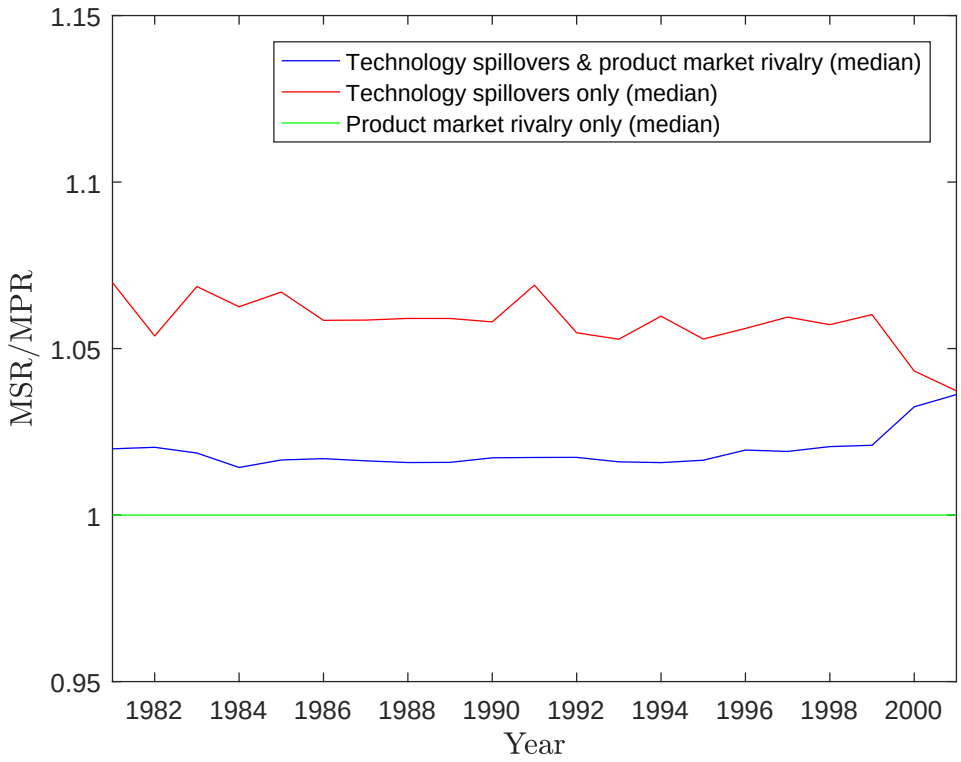
$$\boldsymbol{MPR}_{it} = \frac{\boldsymbol{\Pi}_{ii}\boldsymbol{Q}_{it}}{\boldsymbol{RD}_{it}} \tag{7.9}$$

$$\boldsymbol{MSR}_{it} = \frac{\sum_{j=1}^{N} \boldsymbol{\Pi}_{ij}\boldsymbol{Q}_{jt}}{\boldsymbol{RD}_{it}} \tag{7.10}$$

These can be estimated for any firm-year pair using the data and parameter estimates. From a welfare perspective, there is underinvestment in R&D if the marginal social return exceeds the marginal private return. The signs of the elements of $\boldsymbol{\Pi}$ depend on which spillover is dominant. This implies that the social return may be larger than, equal to or smaller than the private return. It is worth pointing out that the high-dimensional setting implies that the estimated

welfare effects are likely to be imprecise, particularly since the confidence sets for the estimated parameters are too wide to be informative. Nevertheless, the analysis may be useful to give some indication of the optimality of R&D stocks from a welfare perspective.

Figure 5: *Ratio of marginal social and private returns to R&D for the median firm in the electronics industry*



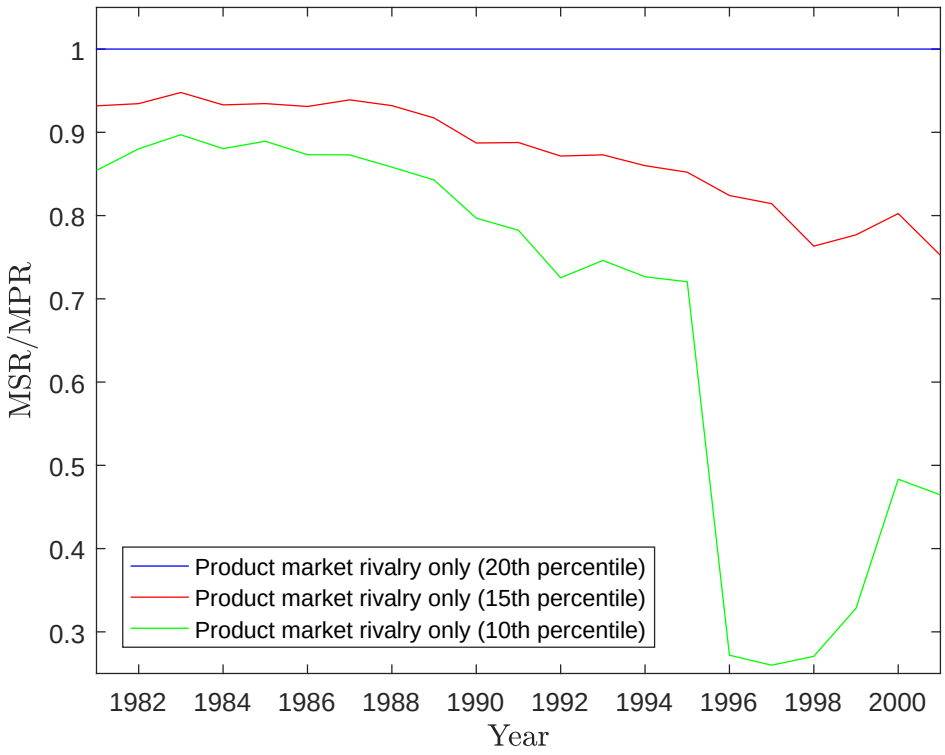
Notes: The figure depicts the estimated ratio of the private and social returns to R&D for the median firm from 1981 to 2001. For the blue line, the ratio for any firm-year pair is estimated based on (7.9) and (7.10), replacing $\mathbf{\Pi}$ with $\widehat{\mathbf{\Pi}}$. For the red line, the ratio is based on replacing $\mathbf{\Pi}$ with $\widehat{\mathbf{\Gamma}}$. For the green line, the ratio is based on replacing $\mathbf{\Pi}$ with $(\boldsymbol{I}_N - \widehat{\mathbf{\Psi}})^{-1}\text{diag}(\widehat{\mathbf{\Gamma}})$.

Figure 5 depicts the estimated the ratio of marginal social returns and marginal private returns for the median firm over the sample period. The blue line shows this statistic when both channels are active, whereas for the red line product market rivalry is deactivated, and for the green line the R&D spillover effect is deactivated.

Looking first at the case where both channels are active, the estimated ratio is larger than 1 in every year. This suggests that the technology spillover effect dominates the product market rivalry effect and is indicative of underinvestment in R&D. The ratio is relatively stable at around 1.02 from 1981 to 1999 before rising in 2000 and 2001. Bloom et al. (2013) also find that there is underinvestment in R&D, and estimate the median ratio of the private and social returns over all firm year pairs as 2.76, which is much larger than the ratios depicted in figure 5. This difference may indicate that levels of R&D in the electronics industry are closer to the social optimum than in the other industries included in the analysis of Bloom et al. (2013). Additionally, the shrinkage applied by the estimator may have lead the point estimates to be too small in magnitude, which could equally account for the difference.

Looking next at the case where there are only R&D spillovers, the ratio is around 1.06. Finally, looking at the case where there is only product market rivalry, the ratio is 1 in every year for the median firm. In order to better undestand the variation over the distribution, figure 6 depicts the 10^{th}, and 15^{th} and 20^{th} percentiles. The 10^{th} and 15^{th} percentiles are less than 1 in every year, and have both declined over the sample period. The 20^{th} percentile is 1 in every year.

Figure 6: *Ratio of marginal social and private returns to R&D attributable to product market rivalry*



Notes: The figure depicts percentiles of the estimated ratio of the private and social returns to R&D from 1981 to 2001 attributable to the product market rivalry channel. The ratios are computed using (7.9) and (7.10), replacing $\mathbf{\Pi}$ with $(\boldsymbol{I}_N - \widehat{\mathbf{\Psi}})^{-1}\text{diag}(\widehat{\mathbf{\Gamma}})$.

8. Conclusion

In this paper, I use panel data to identify and estimate spillover effects when the underlying network is sparse and unobserved. I show that sparsity restrictions can lead to point identification if the network is suitably connected and provide identification results for the case where the researcher has ex-ante knowledge that a given covariate does not generate contextual effects. I apply the STIV estimator of Gautier and Tsybakov (2014) to conduct estimation and model selection using simulated data and data for the electronics industry.

In the simulations and application, I find that the confidence sets are too large to be informative, and sometimes have infinite volume. There are several possible explanations. First, it may be that the parameters are weakly identified. Second, it may be that the sample sizes are too small relative to the sparsity. Finally, it may be because the confidence sets are conservative.

This results from the fact that the confidence sets constructed by Gautier and Tsybakov (2014) do not use (5.6) directly. This is because

$$\widehat{\mathcal{I}}\left(r\sqrt{\widehat{Q}(\boldsymbol{\Theta}_{i,\cdot})}\right) \tag{8.1}$$

is a non-convex set, which introduces computational issues. Instead, Gautier and Tsybakov (2014) work with a convex superset, which leads to conservative inference.

These issues are addressed in Gautier (2015) and Gautier and Rose (2016). As a possible remedy to weak identification, these papers allow for additional structure through considering simultaneous estimation of the system of equations. This permits cross-equation restrictions in addition to the within-equation restrictions considered here, and also allows for a more general sparsity assumption, since, rather than placing an upper bound on the in-degree of each vertex we can instead place an upper bound on the sum of the in-degrees of every vertex.

In addition, Gautier (2015) and Gautier and Rose (2016) allow for a richer specification of unobserved heterogeneity, in which the disturbance is decomposed as $\boldsymbol{\epsilon}_{it} = \boldsymbol{\alpha}_i + \boldsymbol{\lambda}_t + \boldsymbol{\nu}_{it}$ and no assumptions are made on $\boldsymbol{\alpha}_i$ nor on $\boldsymbol{\lambda}_t$. This is important, since the $\boldsymbol{\lambda}_t$ term is a common means of representing *correlated effects*, through which heterogeneity common to all vertices may be correlated with both their outcomes and characteristics. This type of endogeneity is particularly pertinent in the spillovers setting (Manski, 1993; Bramoullé et al., 2009).

Gautier (2015) and Gautier and Rose (2016) also allow for arbitrary linear restrictions on the parameters and restrictions on the sparsity pattern, and develop a new inference procedure which provides confidence sets for linear functionals of the parameters. This is important, since although each element of the parameter matrices may be weakly identified, a linear functional of the parameters, such as the mean entry of $\boldsymbol{\Psi}$, may not be. This can lead to informative inference even if T is small relative to N.

References

Bloom, N., M. Schankerman and J. Van Reenen, "Identifying technology spillovers and product market rivalry," *Econometrica* 81 (2013), 1347–1393.

Blume, L. E., W. A. Brock, S. N. Durlauf and Y. M. Ioannides, "Identification of social interactions," *Available at SSRN 1660002* (2010).

Blume, L. E., W. A. Brock, S. N. Durlauf and R. Jayaraman, "Linear social interactions models," *Journal of Political Economy* 123 (2015), 444–496.

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

Candes, E. and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *The Annals of Statistics* (2007), 2313–2351.

Caner, M. and H. H. Zhang, "Adaptive elastic net for generalized methods of moments," *Journal of Business & Economic Statistics* 32 (2014), 30–47.

Chernozhukov, V., D. Chetverikov, K. Kato et al., "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," *The Annals of Statistics* 41 (2013), 2786–2819.

Davezies, L., X. d'Haultfoeuille and D. Fougère, "Identification of peer effects using group size variation," *The Econometrics Journal* 12 (2009), 397–413.

de Paula, À., "Econometrics of Network Models," *CEMMAP Working Papers* (2015).

de Paula, A., I. Rasul and P. C. Souza, "Identifying and Estimating Social Connections from Outcome Data," *UCL and PUC-Rio working paper* (2016).

Duysters, G., A.-P. De Man and L. Wildeman, "A network approach to alliance management," *European Management Journal* 17 (1999), 182–187.

Gautier, E., "Inference on social effects when the network is unknown by convex (and linear) programming," *Manuscript* (2015).

Gautier, E. and C. Rose, "Inference on social effects when the network is sparse and unobserved," *Working paper* (2016).

Gautier, E. and A. Tsybakov, "High-dimensional instrumental variables regression and confidence sets," *Working Paper* (2014).

Kang, H., A. Zhang, T. T. Cai and D. S. Small, "Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization," *Journal of the American Statistical Association* 111 (2016), 132–144.

König, M., X. Liu and Y. Zenou, "R&D networks: Theory, empirics and policy implications," *CEPR Discussion Paper No. DP9872* (2014).

Lam, C. and P. Souza, "Regularization for High-Dimensional Spatial Models Using the Adaptive LASSO," *LSE working paper* (2013).

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

Manresa, E., "Recovery of Networks using Panel Data," *Working Paper* (2014).

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

Moffitt, R. A., "Policy interventions, low-level equilibria, and social interactions," *Social dynamics* (2001), 45–82.

Souza, P., "Estimating Network Effects without Network Data," *LSE working paper* (2014).

Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

Zou, H., "The adaptive lasso and its oracle properties," *Journal of the American statistical association* 101 (2006), 1418–1429.

9. Appendix

9.1. Proofs

Proof of lemma 4.7 *Fix $\boldsymbol{\Theta} \in \mathcal{I}$. Under assumption (4.2) the baseline model yields the reduced form parameter matrix:*

$$\boldsymbol{\Pi} = \mathbb{E}[\boldsymbol{Y}_{\cdot,t}\boldsymbol{Z}'_{\cdot,t}]\mathbb{E}[\boldsymbol{X}_{\cdot,t}\boldsymbol{Z}'_{\cdot,t}]' \left(\mathbb{E}[\boldsymbol{X}_{\cdot,t}\boldsymbol{Z}'_{\cdot,t}]\mathbb{E}[\boldsymbol{X}_{\cdot,t}\boldsymbol{Z}'_{\cdot,t}]'\right)^{-1} = (\boldsymbol{I}_N - \boldsymbol{\Psi})^{-1}\boldsymbol{\Gamma} \tag{9.1}$$

Now consider $\tilde{\boldsymbol{\Theta}}$ and suppose that $\tilde{\boldsymbol{\Gamma}} = (\boldsymbol{I}_N - \tilde{\boldsymbol{\Psi}})(\boldsymbol{I}_N - \boldsymbol{\Psi})^{-1}\boldsymbol{\Gamma}$. Then we have:

$$(\boldsymbol{I}_N - \boldsymbol{\Psi})^{-1}\boldsymbol{\Gamma} = (\boldsymbol{I}_N - \tilde{\boldsymbol{\Psi}})^{-1}\tilde{\boldsymbol{\Gamma}} \tag{9.2}$$

and $\boldsymbol{\Theta} \neq \tilde{\boldsymbol{\Theta}}$. □

Proof of lemma 4.8 *Fix $\boldsymbol{\Theta} \in \mathcal{I}$. Under assumption (4.2) the baseline model yields the reduced form parameter matrix in (9.1). Pre-multiplying (9.1) by $(\boldsymbol{I}_N - \boldsymbol{\Psi})$ yields $\boldsymbol{\Pi} = \boldsymbol{\Psi}\boldsymbol{\Pi} + \boldsymbol{\Gamma}$. Now, suppose that there is $\tilde{\boldsymbol{\Theta}} \in \mathcal{I}$ with corresponding network $\tilde{\mathcal{G}}$. Then by the same arguments we have $\boldsymbol{\Pi} = \tilde{\boldsymbol{\Psi}}\boldsymbol{\Pi} + \tilde{\boldsymbol{\Gamma}}$. This implies:*

$$\boldsymbol{\Psi}\boldsymbol{\Pi} + \boldsymbol{\Gamma} = \tilde{\boldsymbol{\Psi}}\boldsymbol{\Pi} + \tilde{\boldsymbol{\Gamma}} \tag{9.3}$$

Looking at the rows:

$$(\boldsymbol{\Psi}_{i,\cdot} - \tilde{\boldsymbol{\Psi}}_{i,\cdot})\boldsymbol{\Pi} + (\boldsymbol{\Gamma}_{i,\cdot} - \tilde{\boldsymbol{\Gamma}}_{i,\cdot}) = \boldsymbol{0} \quad \forall i \in \mathcal{V} \tag{9.4}$$

Next consider the sub-vector with elements $\mathcal{X}_i = \mathcal{V}\backslash\{\mathcal{E}_i \cup i\}$:

$$(\boldsymbol{\Psi}_{i,\cdot} - \tilde{\boldsymbol{\Psi}}_{i,\cdot})\boldsymbol{\Pi}_{\cdot,\mathcal{X}_i} + (\boldsymbol{\Gamma}_{i,\mathcal{X}_i} - \tilde{\boldsymbol{\Gamma}}_{i,\mathcal{X}_i}) = \boldsymbol{0} \quad \forall i \in \mathcal{V} \tag{9.5}$$

Now, since $\mathcal{V}\backslash\{\mathcal{E}_i \cup i\}$ are non-neighbors of vertex i in $\mathcal{G}$ and $\tilde{\mathcal{G}} = \mathcal{G}$ by assumption, we have $\boldsymbol{\Psi}_{i,\mathcal{X}_i} = \tilde{\boldsymbol{\Psi}}_{i,\mathcal{X}_i} = \boldsymbol{0}$, $\boldsymbol{\Gamma}_{i,\mathcal{X}_i} = \tilde{\boldsymbol{\Gamma}}_{i,\mathcal{X}_i} = \boldsymbol{0}$ and $\boldsymbol{\Psi}_{ii} = \tilde{\boldsymbol{\Psi}}_{ii} = 0$ (by assumption 4.3). Hence:

$$(\boldsymbol{\Psi}_{i,\mathcal{E}_i} - \tilde{\boldsymbol{\Psi}}_{i,\mathcal{E}_i})\boldsymbol{\Pi}_{\mathcal{E}_i,\mathcal{X}_i} = \boldsymbol{0} \quad \forall i \in \mathcal{V} \tag{9.6}$$

So $\boldsymbol{\Psi}_{i,\mathcal{E}_i} = \tilde{\boldsymbol{\Psi}}_{i,\mathcal{E}_i}$ if $\boldsymbol{\Pi}_{\mathcal{E}_i,\mathcal{X}_i}$ has full row rank, in which case $\boldsymbol{\Theta}_{i,\cdot} = \widetilde{\boldsymbol{\Theta}}_{i,\cdot}$. □

Proof of proposition 4.10 *Fix $\boldsymbol{\Theta} \in \mathcal{I}(\boldsymbol{s})$. Under assumption (4.2) the baseline model yields the reduced form parameter matrix in (9.1). Pre-multiplying (9.1) by $(\boldsymbol{I}_N - \boldsymbol{\Psi})$ yields $\boldsymbol{\Pi} = \boldsymbol{\Psi}\boldsymbol{\Pi} + \boldsymbol{\Gamma}$. Now, suppose that there is $\tilde{\boldsymbol{\Theta}} \in \mathcal{I}(\boldsymbol{s})$ with corresponding network $\tilde{\mathcal{G}}$. Then by the same arguments we have $\boldsymbol{\Pi} = \tilde{\boldsymbol{\Psi}}\boldsymbol{\Pi} + \tilde{\boldsymbol{\Gamma}}$. Hence:*

$$\boldsymbol{\Psi}\boldsymbol{\Pi} + \boldsymbol{\Gamma} = \tilde{\boldsymbol{\Psi}}\boldsymbol{\Pi} + \tilde{\boldsymbol{\Gamma}} \tag{9.7}$$

Looking at the rows:

$$(\mathbf{\Psi}_{i,\cdot} - \tilde{\mathbf{\Psi}}_{i,\cdot})\mathbf{\Pi} + (\mathbf{\Gamma}_{i,\cdot} - \tilde{\mathbf{\Gamma}}_{i,\cdot}) = \mathbf{0} \quad \forall i \in \mathcal{V} \tag{9.8}$$

Next consider the sub-vector with elements $\mathcal{X}_i = \mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$*:*

$$(\mathbf{\Psi}_{i,\cdot} - \tilde{\mathbf{\Psi}}_{i,\cdot})\mathbf{\Pi}_{\cdot,\mathcal{X}_i} + (\mathbf{\Gamma}_{i,\mathcal{X}_i} - \tilde{\mathbf{\Gamma}}_{i,\mathcal{X}_i}) = \mathbf{0} \quad \forall i \in \mathcal{V} \tag{9.9}$$

Now, since $\mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$ *are non-neighbors of vertex* i *in* $\mathcal{G}$ *and* $\tilde{\mathcal{G}}$*, we have* $\mathbf{\Psi}_{i,\mathcal{X}_i} = \tilde{\mathbf{\Psi}}_{i,\mathcal{X}_i} = \mathbf{0}$*,* $\mathbf{\Gamma}_{i,\mathcal{X}_i} = \tilde{\mathbf{\Gamma}}_{i,\mathcal{X}_i} = \mathbf{0}$ *and* $\mathbf{\Psi}_{ii} = \tilde{\mathbf{\Psi}}_{ii} = 0$ *(by assumption 4.3). Hence, for* $\mathcal{Y}_i = \mathcal{E}_i \cup \tilde{\mathcal{E}}_i$*:*

$$(\mathbf{\Psi}_{i,\mathcal{Y}_i} - \tilde{\mathbf{\Psi}}_{i,\mathcal{Y}_i})\mathbf{\Pi}_{\mathcal{Y}_i,\mathcal{X}_i} = \mathbf{0} \quad \forall i \in \mathcal{V} \tag{9.10}$$

Suppose that $\mathbf{\Pi}_{\mathcal{Y}_i,\mathcal{X}_i}$ *has full row rank. This implies* $\mathbf{\Psi}_{i,\mathcal{Y}_i} = \tilde{\mathbf{\Psi}}_{i,\mathcal{Y}_i}$*, in which case* $\mathbf{\Theta}_{i,\cdot} = \widetilde{\mathbf{\Theta}}_{i,\cdot}$*. Finally, note that it is only necessary to consider:*

$$\tilde{\mathcal{E}}_i \in \{\tilde{\mathcal{E}}_i : \tilde{\mathcal{E}}_i \subseteq \mathcal{E}_i^c, |\tilde{\mathcal{E}}_i| = \min[\boldsymbol{s}_i, |\mathcal{E}_i^c|]\} \tag{9.11}$$

which lead to sub-matrices with the maximal number of rows and minimal number of columns. This proves part (i).

To prove part (ii), it is sufficient to note that assumption 4.9 implies $\mathcal{C} \supseteq \mathcal{C}^{\mathbf{\Pi}} \supseteq \mathcal{E}$ *and hence* $\tilde{\mathcal{C}} \supseteq \mathcal{C}^{\mathbf{\Pi}} \supseteq \tilde{\mathcal{E}}$*.* □

Proof of lemma 4.12 Let $\mathbf{\Pi}_{\mathcal{V}_y,\mathcal{V}_x}$ sub-matrix of $\mathbf{\Pi}$ with rows $\mathcal{V}_y$ and columns $\mathcal{V}_x$. Suppose that there are $v \leq \min\left[|\mathcal{V}_y|, |\mathcal{V}_x|\right]$ vertex-independent paths in $\mathcal{G}$ from $\mathcal{V}_x$ to $\mathcal{V}_y$. Then one of the following is true:

1. $\mathbf{\Pi}_{\mathcal{V}_y,\mathcal{V}_x}$ has at least $|\mathcal{V}_y| - v$ rows of zeros.
2. $\mathbf{\Pi}_{\mathcal{V}_y,\mathcal{V}_x}$ has at least $|\mathcal{V}_y| - v$ linearly dependent rows.

In either case, $\text{rank}\left(\mathbf{\Pi}_{\mathcal{V}_y,\mathcal{V}_x}\right) \leq \min\left[|\mathcal{V}_y|, |\mathcal{V}_x|\right] - \left(|\mathcal{V}_y| - v\right) \leq v$. □

Proof of proposition 4.14 *Fix* $\mathbf{\Theta} \in \mathcal{I}$*. Under assumption (4.2) the baseline model yields the reduced form parameter matrix in (9.1). The* $N \times N$ *sub-matrix of* $\mathbf{\Pi}$ *corresponding to covariate* k *is* $\mathbf{\Pi}_{(k)} = (\boldsymbol{I}_N - \mathbf{\Psi})^{-1}\mathbf{\Gamma}_{(k)}$*. If* $\mathbf{\Gamma}_{(k)}$ *is diagonal with rank* N*,* $\mathbf{\Pi}_{(k)}$ *is invertible and:*

$$(\mathbf{\Pi}^{-1})_{(k)ii} = 1/\mathbf{\Gamma}_{(k)ii} \qquad \forall i \in \mathcal{V} \tag{9.12}$$

$$(\mathbf{\Pi}^{-1})_{(k)ij} = -\mathbf{\Psi}_{(k)ij}/\mathbf{\Gamma}_{(k)ii} \qquad \forall i, j \neq i \in \mathcal{V}^2 \tag{9.13}$$

Now suppose that there is $\tilde{\mathbf{\Theta}} \in \mathcal{I}$ *and* $\tilde{\mathbf{\Gamma}}_{(k)}$ *is diagonal with rank* N*. Then* $\mathbf{\Pi} = (\boldsymbol{I}_N - \tilde{\mathbf{\Psi}})^{-1}\tilde{\mathbf{\Gamma}}$ *and:*

$$1/\mathbf{\Gamma}_{(k)ii} = 1/\tilde{\mathbf{\Gamma}}_{(k)ii} \qquad \forall i \in \mathcal{V} \tag{9.14}$$

$$-\mathbf{\Psi}_{ij}/\mathbf{\Gamma}_{(k)ii} = -\tilde{\mathbf{\Psi}}_{ij}/\tilde{\mathbf{\Gamma}}_{(k)ii} \qquad \forall i, j \neq i \in \mathcal{V}^2 \tag{9.15}$$

Solving yields $\tilde{\mathbf{\Psi}} = \mathbf{\Psi}, \tilde{\mathbf{\Gamma}}_{(k)} = \mathbf{\Gamma}_{(k)}$, *which implies* $\tilde{\mathbf{\Gamma}} = \mathbf{\Gamma}$. □

9.2. Identification under sparsity for general K

The following additional notation is required to extend the results of sub-section 4.1 to $K > 1$. First, the observable connectivity is generalized as follows:

$$\mathcal{C}^{\mathbf{\Pi}} = \{(j,i) : \exists k \in \{1,...,K\}, \quad \mathbf{\Pi}_{(k)ij} \neq 0\} \tag{9.16}$$

where $\mathbf{\Pi}_{(k)}$ is the $N \times N$ reduced form parameter matrix for covariate k. Proposition 4.10 generalizes as follows:

Proposition 9.1 (Identification under sparsity)
Let $\mathbf{\Theta} \in \mathcal{I}(\boldsymbol{s})$ *and let assumptions 4.1 and 4.2 be satisfied.*

(i) $\mathbf{\Theta}_{i,\cdot}$ *is point identified if for every*

$$\tilde{\mathcal{E}}_i \in \{\tilde{\mathcal{E}}_i : \tilde{\mathcal{E}}_i \subseteq \mathcal{E}_i^c, |\tilde{\mathcal{E}}_i| = \min[\boldsymbol{s}_i, |\mathcal{E}_i^c|]\}, \tag{9.17}$$

the sub-matrix of $\mathbf{\Pi}$ *with rows* $\mathcal{E}_i \cup \tilde{\mathcal{E}}_i$ *and the columns corresponding to the covariates of* $\mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$ *has full row rank.*

(ii) In addition, let assumption 4.9 be satisfied for vertex i*. Then* $\mathcal{C}_i \supseteq \mathcal{C}_i^{\mathbf{\Pi}} \supseteq \mathcal{E}_i$ *and part* (i) *applies with* (9.17) *replaced by*

$$\tilde{\mathcal{E}}_i \in \{\tilde{\mathcal{E}}_i : \tilde{\mathcal{E}}_i \subseteq \mathcal{E}_i^c, |\tilde{\mathcal{E}}_i| = \min[\boldsymbol{s}_i, |\mathcal{E}_i^c|]\} \cap \mathcal{C}_i^{\mathbf{\Pi}} \tag{9.18}$$

The only difference from proposition 4.10 is that in part (i) the columns correspond to all of the covariates of $\mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$. For $K = 1$, these are columns $\mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$, whereas for general K we require columns $\mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}, N + \mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}, ..., N(K-1) + \mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$, where the addition is applied to every element in the set. The order condition in corollary 4.11 generalizes as follows:

Corollary 9.2 (Order condition)
The order conditions corresponding to the rank conditions in cases (i) and (ii) of proposition 4.10 are:

(i) $|\mathcal{E}_i| + \min[\boldsymbol{s}_i, |\mathcal{E}_i^c|] \leq \frac{K(N-1)}{1+K}$

(ii) $|\mathcal{E}_i| + \min[\boldsymbol{s}_i, |\mathcal{E}_i^c \cap \mathcal{C}_i^{\mathbf{\Pi}}|] \leq \frac{K(N-1)}{1+K}$

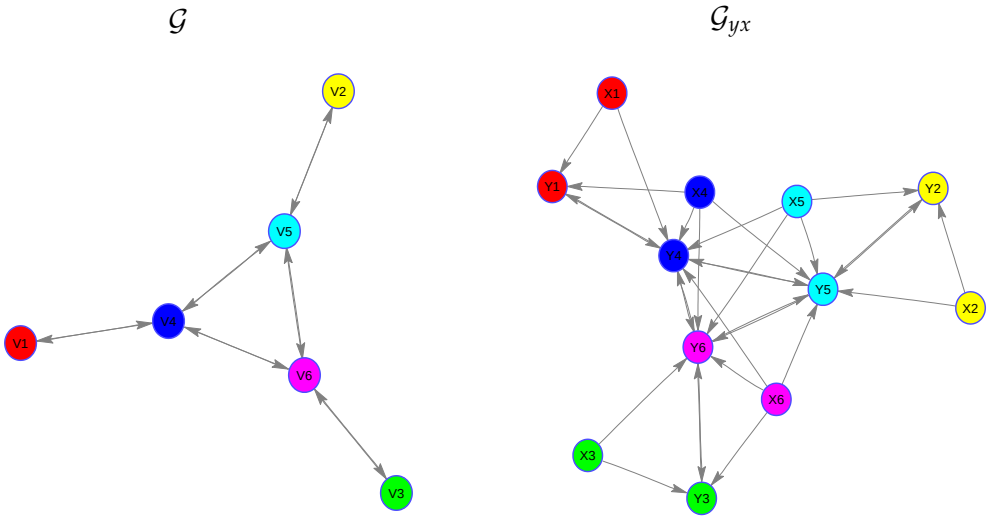
To extend the result on vertex-independent paths in corollary 4.13, we must first define a new network which relates the covariates to the outcomes. First, define the network $\mathcal{G}_{yx} = (\mathcal{V}_{yx}, \mathcal{E}_{yx})$, which represents the system of equations in (3.1). The set of vertices is:

$$\mathcal{V}_{yx} = \left\{y_1, ..., y_N, x_{(1)1}, ..., x_{(1)N}, ..., x_{(K)N}\right\} \tag{9.19}$$

and the set of edges $\mathcal{E}_{yx}$ is uniquely determined by K and $\mathcal{G}$, such that $(j,i) \in \mathcal{E} \iff (y_j, y_i) \in \mathcal{E}_{yx}$ and $(j,i) \in \mathcal{E} \iff (x_{(k)j}, y_i) \in \mathcal{E}_{yx} \quad \forall k \in \{1,...,K\}$. An example of $\mathcal{G}$ and $\mathcal{G}_{yx}$ is depicted in figure 7.

Using the network $\mathcal{G}_{yx}$, the result is:

Figure 7: *An example of $\mathcal{G}$ and $\mathcal{G}_{yx}$*



Corollary 9.3 (Vertex-independent paths)
Let $\boldsymbol{\Theta} \in \mathcal{I}(\boldsymbol{s})$ and let assumptions 4.1 and 4.2 be satisfied.

(i) A necessary condition for the rank condition in part (i) of proposition 9.1 is that for every

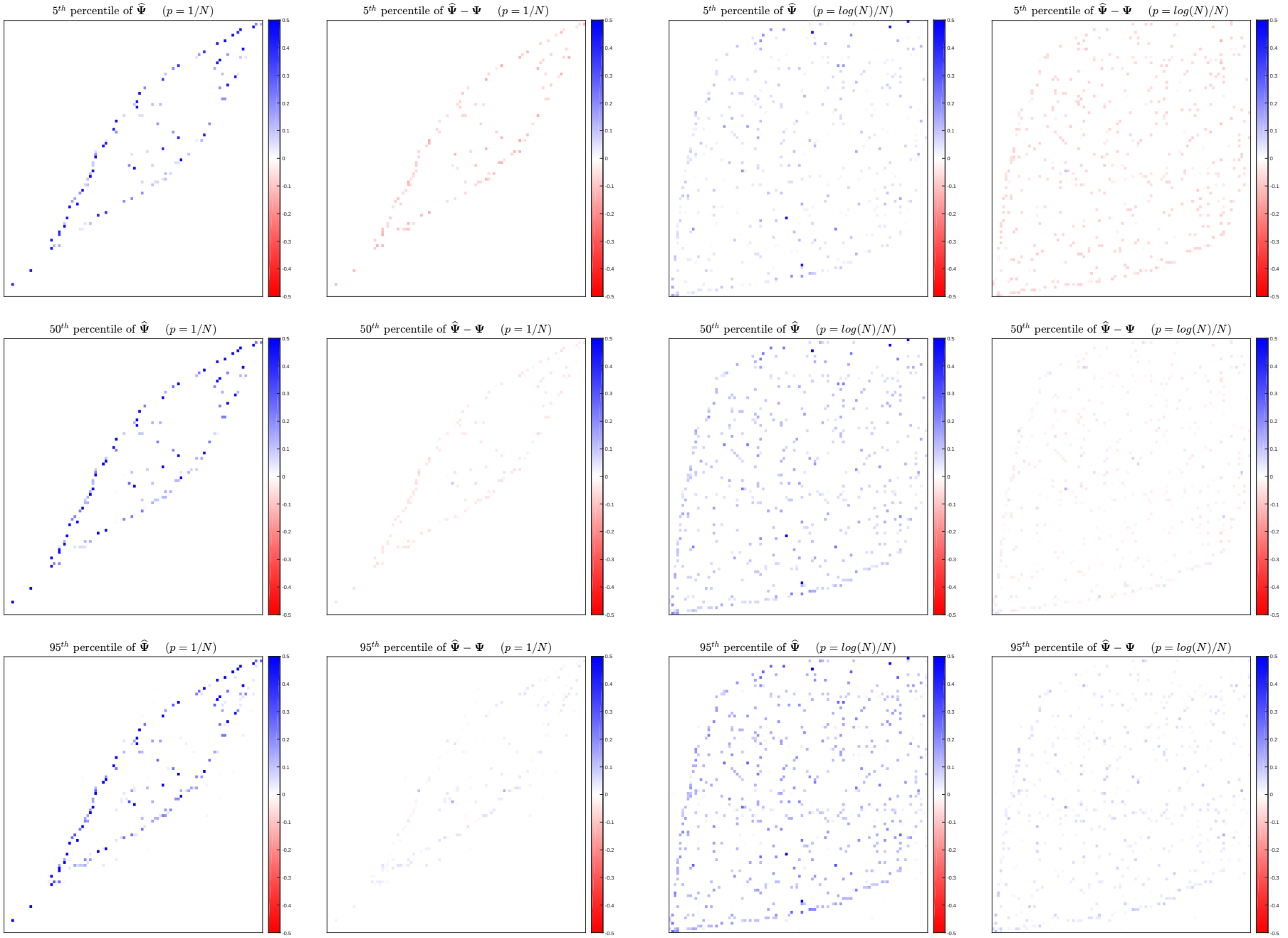
$$\tilde{\mathcal{E}}_i \in \{\tilde{\mathcal{E}}_i : \tilde{\mathcal{E}}_i \subseteq \mathcal{E}_i^c, |\tilde{\mathcal{E}}_i| = \min[\boldsymbol{s}_i, |\mathcal{E}_i^c|]\}, \tag{9.20}$$

there are $|\mathcal{E}_i \cup \tilde{\mathcal{E}}_i|$ vertex-independent paths in $\mathcal{G}_{yx}$ from the covariates of $\mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$ to the outcomes of $\mathcal{E}_i \cup \tilde{\mathcal{E}}_i$.

(ii) In addition, let assumption 4.9 be satisfied for vertex i. Then $\mathcal{C}_i \supseteq \mathcal{C}_i^{\mathbf{\Pi}} \supseteq \mathcal{E}_i$ and a necessary condition for the rank condition in part (ii) of proposition 9.1 is that for every

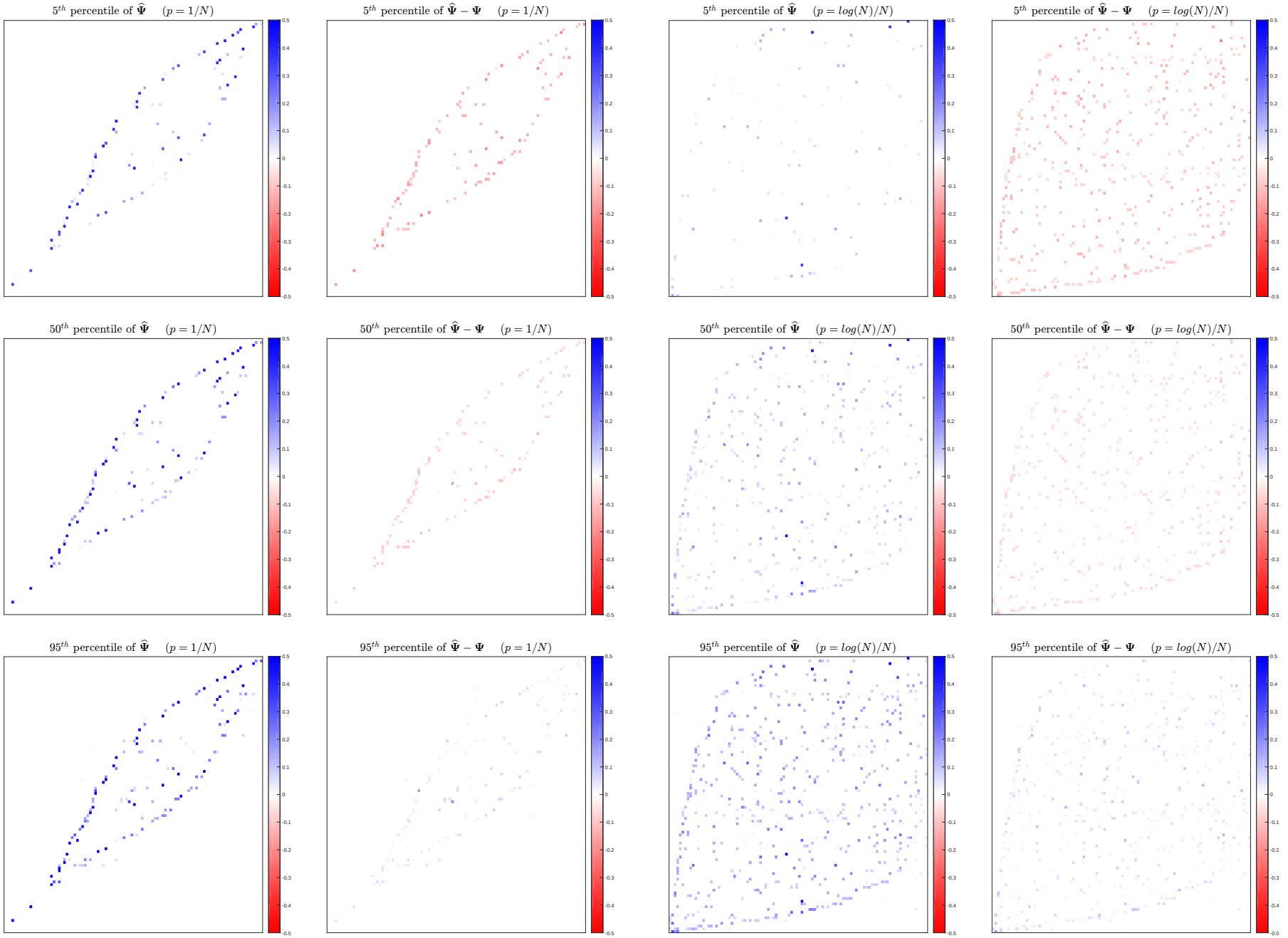
$$\tilde{\mathcal{E}}_i \in \{\tilde{\mathcal{E}}_i : \tilde{\mathcal{E}}_i \subseteq \mathcal{E}_i^c, |\tilde{\mathcal{E}}_i| = \min[\boldsymbol{s}_i, |\mathcal{E}_i^c|]\} \cap \mathcal{C}_i^{\mathbf{\Pi}} \tag{9.21}$$

there are $|\mathcal{E}_i \cup \tilde{\mathcal{E}}_i|$ vertex-independent paths in $\mathcal{G}_{yx}$ from the covariates of $\mathcal{V}\backslash\{\mathcal{E}_i \cup \tilde{\mathcal{E}}_i \cup i\}$ to the outcomes of $\mathcal{E}_i \cup \tilde{\mathcal{E}}_i$.



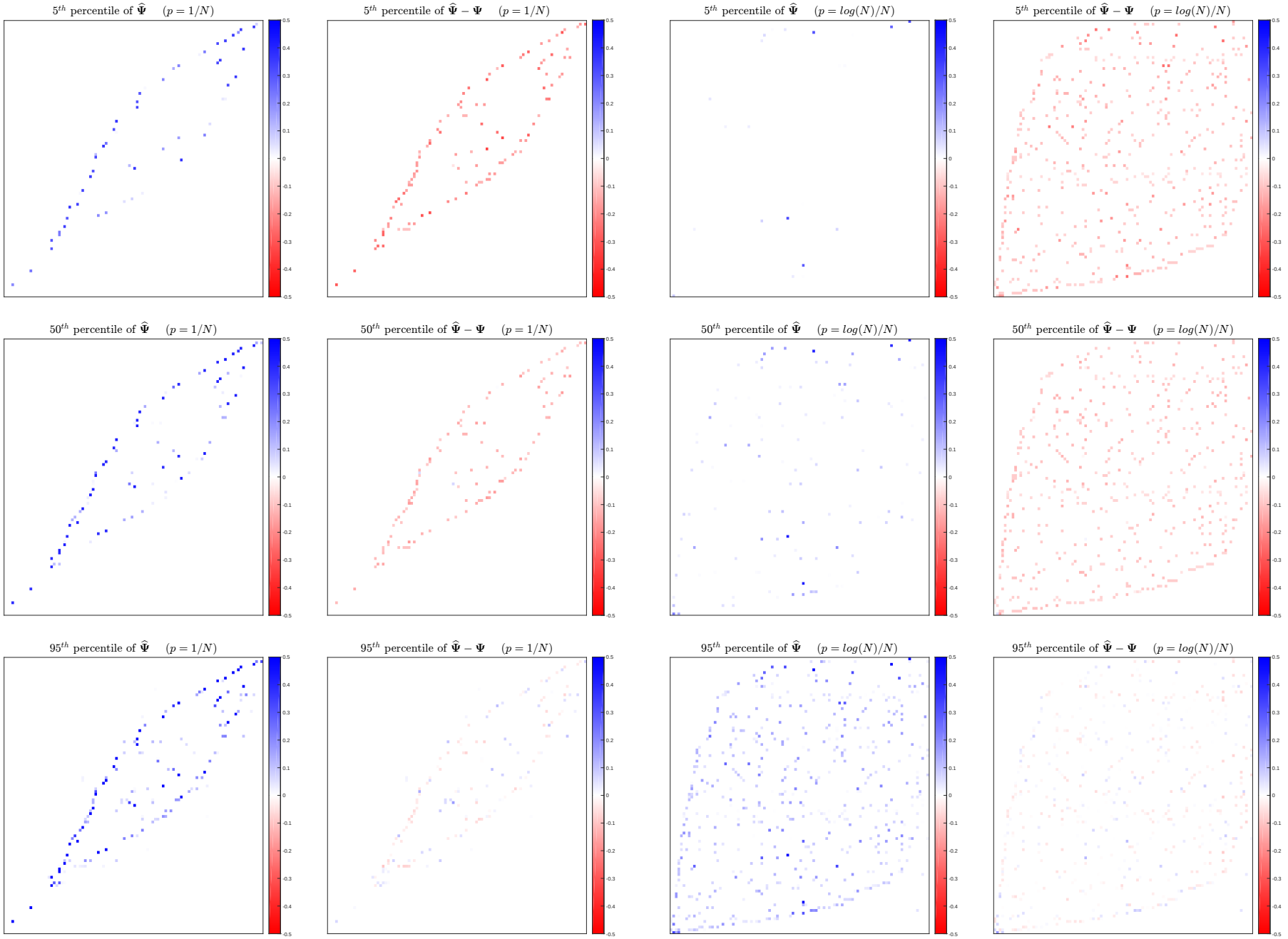
Notes: The true parameters are in figure 3. The STIV estimator is applied with $c = 0.99/r$. Percentiles are taken over 1000 simulations. Vertices are reordered to concentrate the mass around the 45 degree line.

Figure 8: *Percentiles of point estimates and estimation errors of* $\mathbf{\Psi}$ *for* $T = 500$



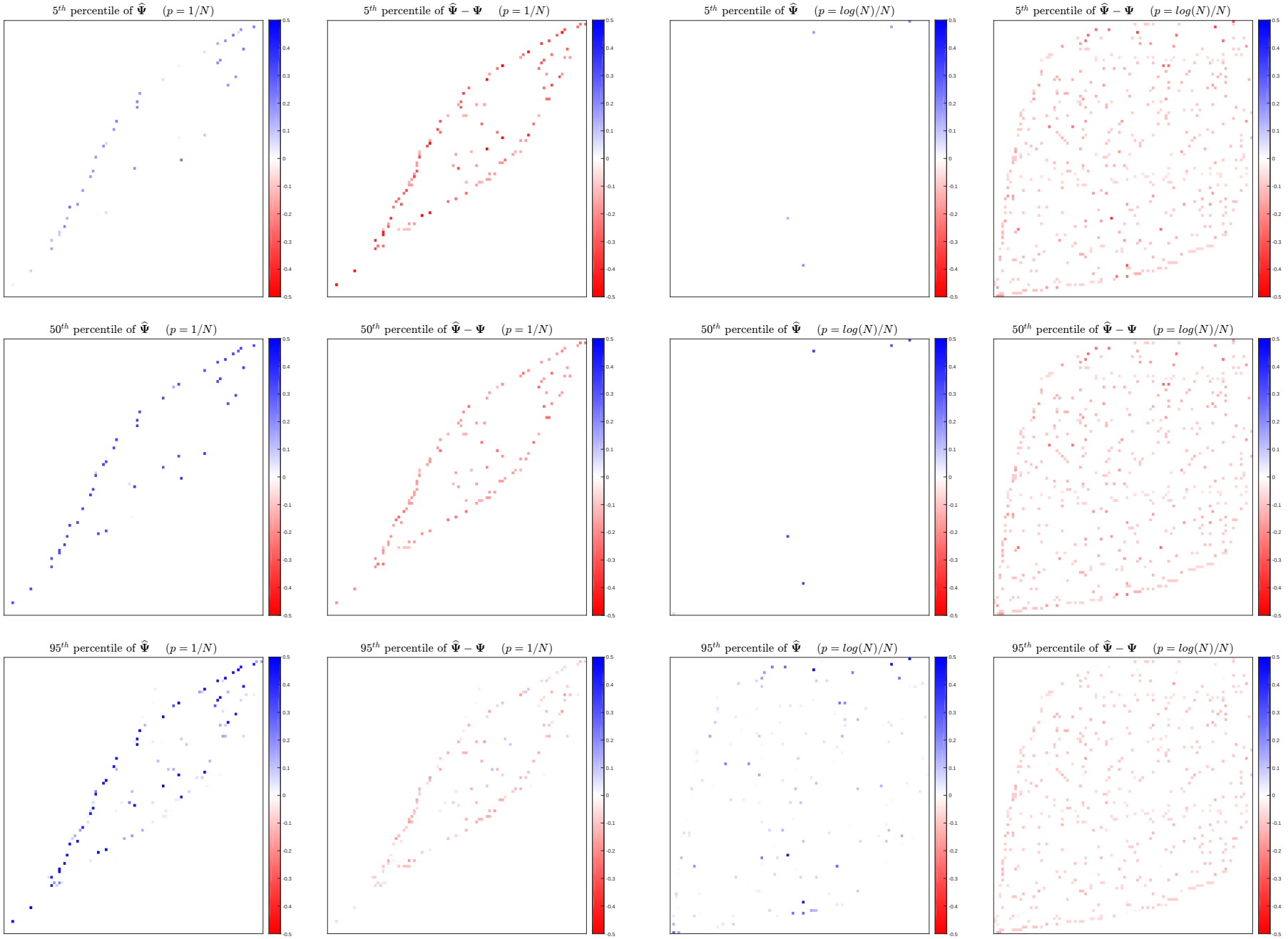
Notes: The true parameters are in figure 3. The STIV estimator is applied with $c = 0.99/r$. Percentiles are taken over 1000 simulations. Vertices are reordered to concentrate the mass around the 45 degree line.

Figure 9: *Percentiles of point estimates and estimation errors of* $\mathbf{\Psi}$ *for* $T = 200$



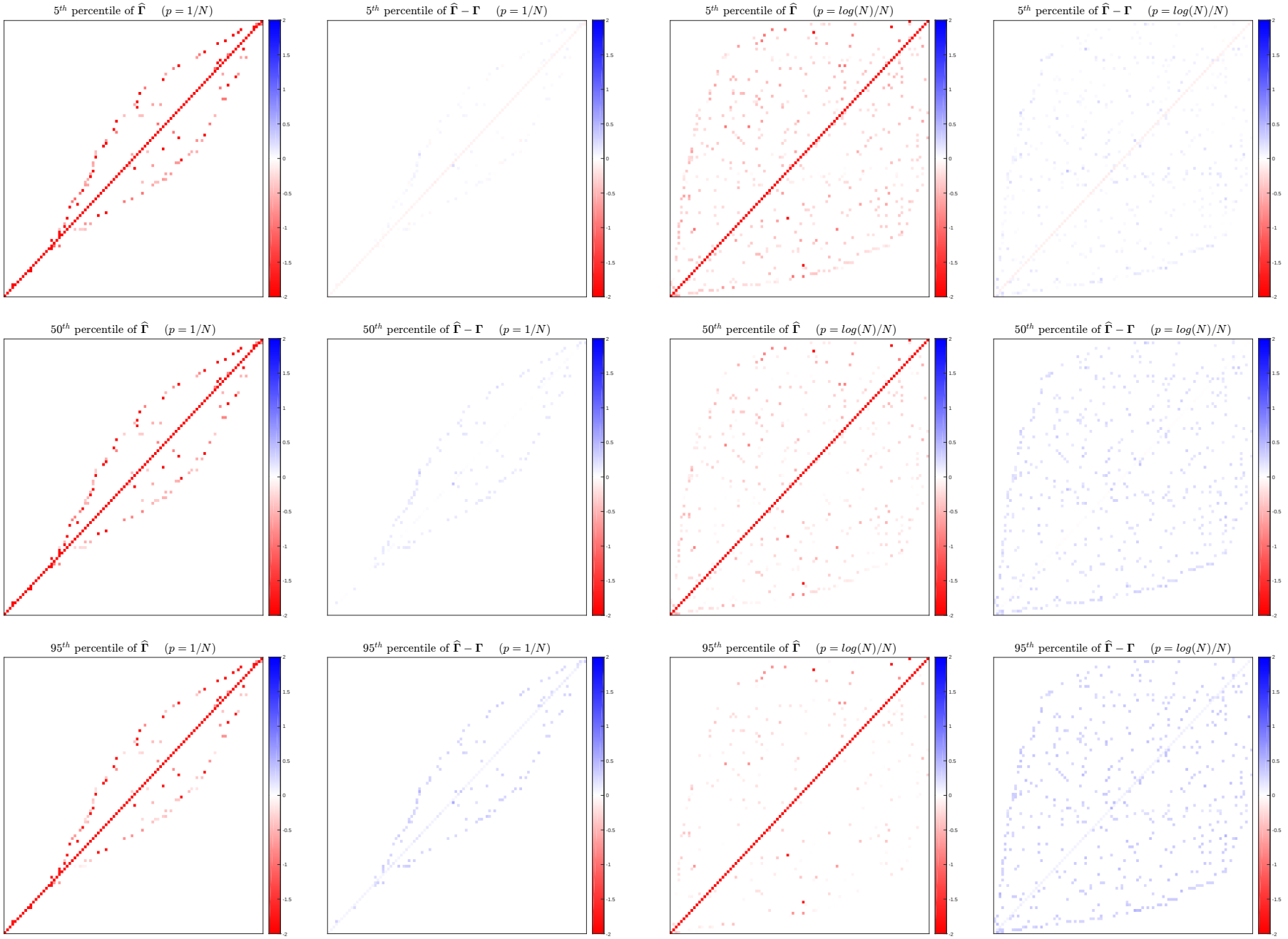
Notes: The true parameters are in figure 3. The STIV estimator is applied with $c = 0.99/r$. Percentiles are taken over 1000 simulations. Vertices are reordered to concentrate the mass around the 45 degree line.

Figure 10: *Percentiles of point estimates and estimation errors of* $\mathbf{\Psi}$ *for* $T = 100$



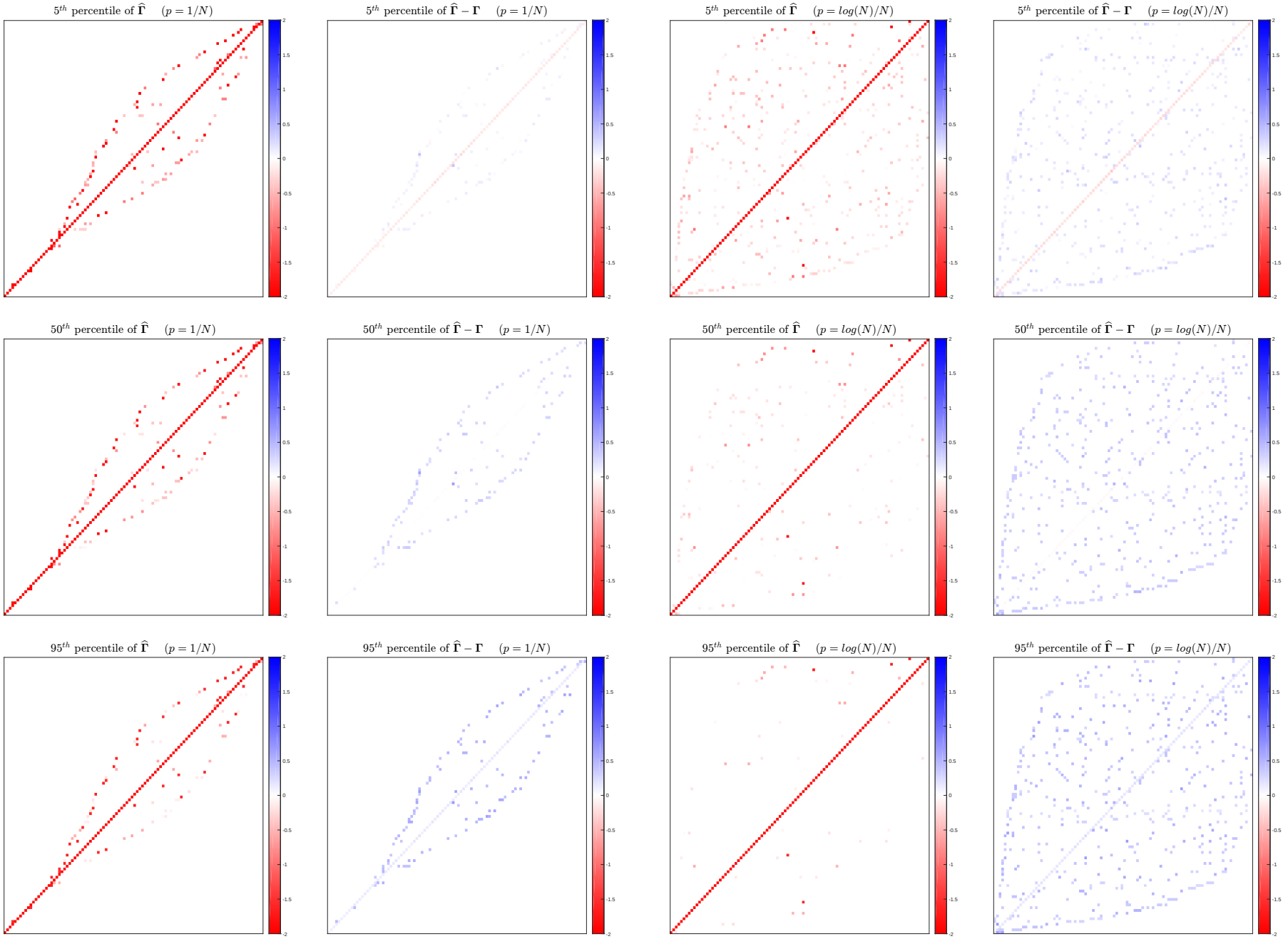
Notes: The true parameters are in figure 3. The STIV estimator is applied with $c = 0.99/r$. Percentiles are taken over 1000 simulations. Vertices are reordered to concentrate the mass around the 45 degree line.

Figure 11: *Percentiles of point estimates and estimation errors of* $\mathbf{\Psi}$ *for* $T = 50$



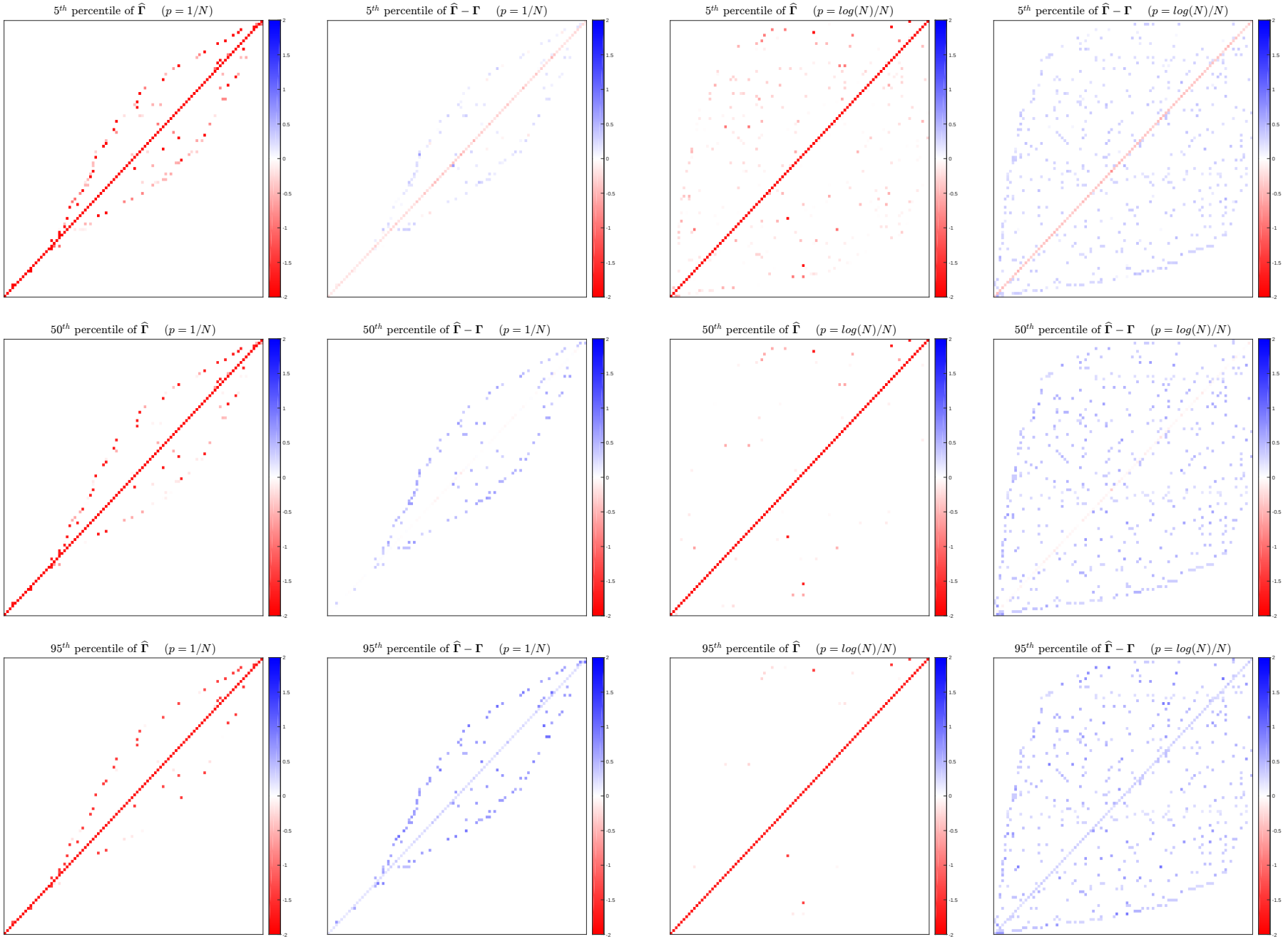
Notes: The true parameters are in figure 3. The STIV estimator is applied with $c = 0.99/r$. Percentiles are taken over 1000 simulations. Vertices are reordered to concentrate the mass around the 45 degree line.

Figure 12: *Percentiles of point estimates and estimation errors of* $\mathbf{\Gamma}$ *for* $T = 500$



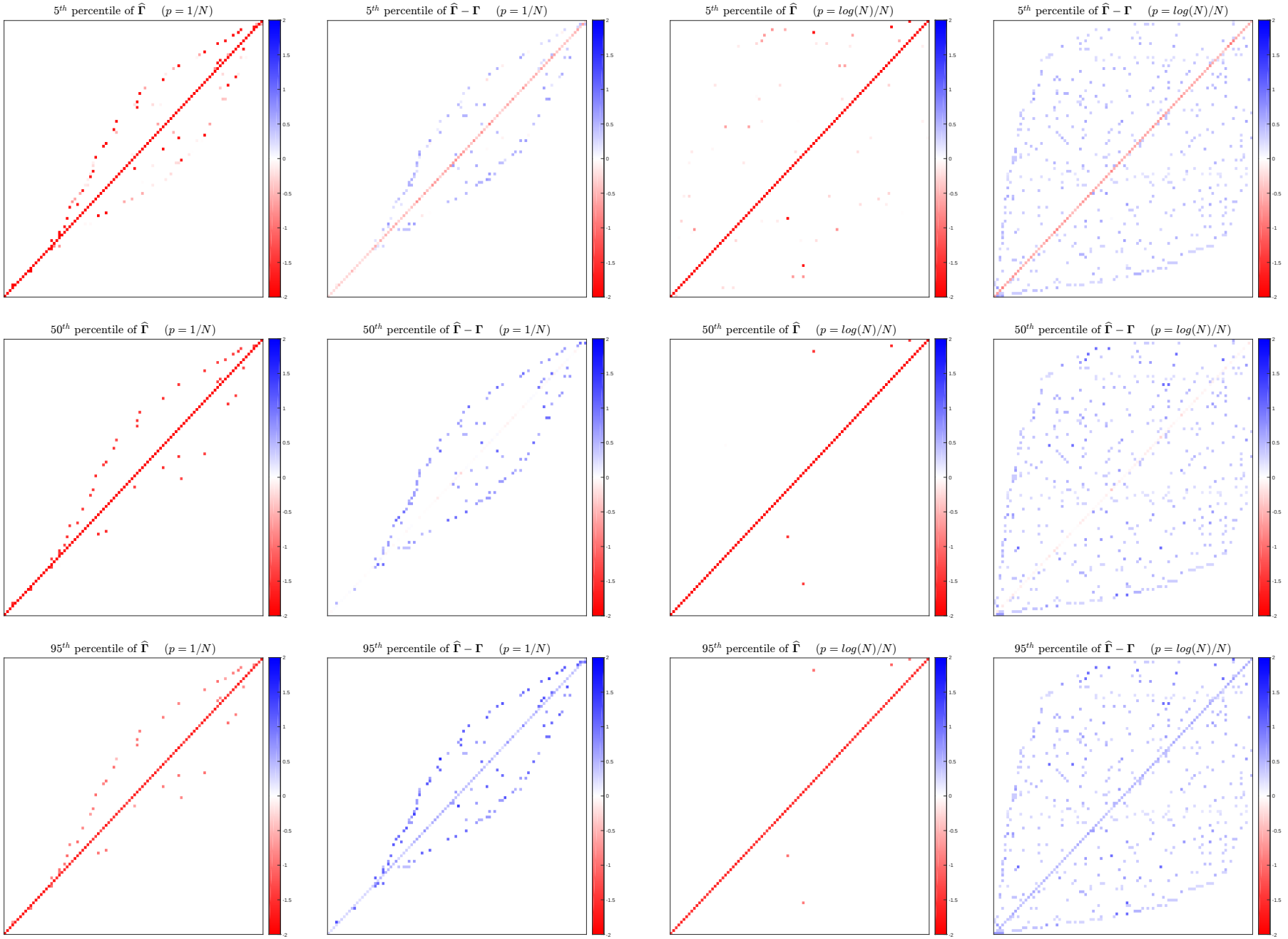
Notes: The true parameters are in figure 3. The STIV estimator is applied with $c = 0.99/r$. Percentiles are taken over 1000 simulations. Vertices are reordered to concentrate the mass around the 45 degree line.

Figure 13: *Percentiles of point estimates and estimation errors of* $\mathbf{\Gamma}$ *for* $T = 200$



Notes: The true parameters are in figure 3. The STIV estimator is applied with $c = 0.99/r$. Percentiles are taken over 1000 simulations. Vertices are reordered to concentrate the mass around the 45 degree line.

Figure 14: *Percentiles of point estimates and estimation errors of* $\mathbf{\Gamma}$ *for* $T = 100$



Notes: The true parameters are in figure 3. The STIV estimator is applied with $c = 0.99/r$. Percentiles are taken over 1000 simulations. Vertices are reordered to concentrate the mass around the 45 degree line.

Figure 15: *Percentiles of point estimates and estimation errors of* $\mathbf{\Gamma}$ *for* $T = 50$