Dyadic Regression with Block-Specific Fixed Effects and Application to Input-Output Matrix

Seongjoo Min*

Abstract

Viewing the Input-Output matrix as a network of industries, this article studies how exogenous shocks, such as changes in tax rates and import duty rates, can affect relationships between industries within the U.S. economy. It also proposes a method for making counterfactual predictions on the Input-Output Matrix. A key innovation is utilizing a stochastic blockmodel to group industries into blocks, so that pairs of industries that belong to the same pair of blocks share common pairwise block-specific fixed effects. It uses dyadic regression model to estimate model parameters, together with pairwise block-specific fixed effects. An empirical analysis using the 2017 U.S. Input-Output Matrix shows that the supplier-buyer relationship between commodities (or industries) are sensitive to changes in import duty rates.

Keywords: networks, input-output, stochastic blockmodel

JEL codes: L14, R15, C21

*Department of Economics, University of Iowa, S324 Pappajohn Business Building, Iowa City, IA 52242.

1 Introduction

No industry produces output that is entirely consumed by the final user. Instead, a portion of an industry's output is used by itself or by other industries as intermediate goods and services to produce their output. Similarly, no industry produces output from the ground up, as they use intermediate goods and services produced by other industries as inputs. In fact, all industries are interconnected via such supplier-buyer relationships, forming a network of industries.

The Input-Output Matrix introduced by Leontief (1951) describes how interactions of industries as suppliers and buyers produce the total output of an economy. Each industry produces goods and services or "commodities".[1] The $(i, j)^{\text{th}}$ component of the Input-Output Matrix is the amount of commodity i that is used to produce \$1 of commodity j. Thus the j^{th}column lays out the amount of all commodities used to produce \$1 of commodity j. The j^{th} column essentially describes how the production of commodity j depends itself and other commodities. By putting together all columns, the Input-Output Matrix describes the dependence structure in the production of all commidities within an economy. By utilizing the Input-Output Matrix, one can relate the demand for final use to the total demand for various goods and services.[2] It is widely used by researchers to quantify how an economic shock to an industry can propagate to the whole economy.

This article studies the formation of the Input-Output Matrix as the outcome of observable characteristics, including the demand for final use, net tax rate, and import duty rate. Viewing the dependence structure across commodities as a network, it utilizes dyadic regression model to study how observable characteristics influence the dependence between pairs of commodities. Similar to the Jochmans (2017), it uses two-way pair-specific fixed effects to account for the unobserved heterogeneity specific to each pair of commodities. It

[1] Leontief's model classifies goods and services either as industries that produce them or commodity groups they belong to. There are two versions of the Input-Output Matrix: one for industries and the other for commodities. Obtaining the industry version from the commodity version (or vice versa) only requires simple matrix operations.

[2] Total demand is equal to sum of the demand for intermediate use and the demand for final use.

is two-way because how the production of commodity j depends on commodity i could be different from how the production of i depends on j. For the former, (i, j) denotes the pair, and the underlying industry of commodity i serves the role of supplier and j the buyer. For the latter, (j, i) denotes the pair, and the commodity j serves the role of supplier and i the buyer.

A novel contribution is utilizing a stochastic blockmodel to generate two partitions over the commodities. For one partition, each commodity is assigned membership into a block based on its role as a supplier, so that commodities that belong to the same block share common block-specific fixed effect as a supplier. For the other partition, each commodity is assigned membership into a block based on its role as a buyer, so that commodities that belong to the same block share common block-specific fixed effect as a buyer. If commodities i, k belong to the same supplier block and j, l belong to the same buyer block, then the pairs (i, j) and (k, l) are endowed with the same pairwise block-specific fixed effects. Because pairs (i, j) and (j, i) are not the same, they are not necessarily endowed with the same pairwise block-specific fixed effects. The number of pairwise block-specific fixed effects is significantly smaller than the number of commodities, so the model is not subject to the incidental parameter bias.[3] Because the numbers of blocks are unknown, generating each partition requires estimating the number of blocks and estimating the block memberships. Under certain conditions, pairwise block-specific fixed effects and model parameters can be consistently estimated, which enables making counterfactual predictions on the Input-Output Matrix.

Estimation involves two steps. First, it employs the stochastic blockmodel proposed by Rohe et al. (2016) to generate two partitions over the commodities. In the second step, it estimates parameters of the dyadic regression model, together with two-way block-specific fixed effects. Similar to Santos Silva and Tenreyro (2006), it uses Poisson pseudo-likelihood specification. Covariance matrix estimation utilizes the method proposed by Fafchamps

[3]It is a difficulty in estimating agent-specific fixed effects first highlighted by Neyman and Scott (1948).

and Gubert (2007). It also shows that the two-step estimator is consistent under mild assumptions.

An empirical analysis using the 2017 Commodity-by-Commodity Input-Output Matrix of the U.S. economy and the domestic and foreign trade accounts published by the BEA (Bureau of Economic Analysis, 2024) suggests

- If import duty rate on a commodity increases, less of it is used in the domestic production of other commodities, but there is no significant change in the domestic production of it.

The rest of the article is organized as the following. Section 2 discusses a selection of related research on dyadic regression models and panel data analysis with latent group structure. Section 3 presents the model and provides a justification for using stochastic blockmodels for better model specification. Section 4 describes the two-step estimation procedure. Section 5 presents the empirical analysis of the U.S. Input-Output Matrix. Section 6 concludes.

2 Related Research

This article is founded upon preceding research on networks and dyadic regression models. Fafchamps and Gubert (2007) uses dyadic regression model to study the formation of risk-sharing networks. Using private money lending and borrowing practice in the rural Philippines, it shows that geographical distance is an important determinant for whether or not to lend and borrow money. It also proposes a method for computing robust standard error, which Graham (2020) shows is asymptotically valid. This article uses a modified version of it.

Graham (2017) introduces agent-specific fixed effects to dyadic regression model with binary outcome variable. Extending the method of Charbonneau (2017), it uses pairs of pairs, or tetrads, to difference away the fixed effects under a maximum likelihood estimator with logistic error term. It shows that the resulting estimator, called Tetrad Logit, is consistent and asymptotically normally distributed. Empirically, it shows that homophily, which is the tendency that people with similar characteristics are more likely to be connected, is crucial for explaining the formation of social networks.

Santos Silva and Tenreyro (2006) models international trade networks using a Poisson pseudo-likelihood specification, which is dyadic regression model with nonnegative outcome variable. Jochmans (2017) introduces pairwise agent-specific fixed effects to the model of Santos Silva and Tenreyro (2006) and differences away the fixed effects by extending the method of Charbonneau (2017). It also shows consistency and asymptotic normality of the estimator.

This article is related to the research on the analysis of panel data with latent group structure. Su, Shi, et al. 2016 proposes a method for identification of latent group structure and parameter estimation by utilizing a variant of lasso. Su and Wang 2021 proposes a method analyzing panel data while utilizing a break detection method or a stochastic blockmodel to identify latent group structure. Ma et al. (2021) proposes a method for consistently estimating the number of groups (blocks) while utilizing a stochastic blockmodel to identify latent group structure.

This article contributes to these areas of research by (1) utilizing a stochastic blockmodel to reduce the dimension of pair-specific fixed effects in dyadic regression model, and (2) empirically analyzing the U.S. Input-Output Matrix and making counterfactual predictions.

3 The Model

Let N denote the number of agents (commodities) that appear on the Input-Output Matrix. For $i, j \in \{1, 2, \ldots, N\}$, $[TR_{ij}] \in \mathbb{R}_+^{N \times N}$ denotes the Input-Output Matrix. Inspection of 2007, 2012, and 2017 U.S. Input-Output Matrices reveals TR_{ii} are close to 1. For this reason, the i^{th} column of $[TR_{ij}]$ is divided by TR_{ii}, so that all elements on the main diagonal are identically 1. With Y_i the i^{th} column after the normalization, let $\mathbf{Y} \overset{def}{=} [Y_1, Y_2, \ldots, Y_N] \in \mathbb{R}_+^{N\times N}$ and $Y_{ij} \overset{def}{=} [\mathbf{Y}]_{ij}$. The interpretation of Y_{ij} is the amount of commodity i used to produce \$1 of commodity j, relative to the amount j used in the production. In other words, Y_{ij} is the dependence between (i, j) with i as the supplier and j as the buyer, relative to the dependence of j on itself. Naturally, Y_{ij} and Y_{ji} are not necessarily equal. In this case, $\mathbf{Y}$ is called a directed network because the direction of the association matters. Let $X_i, X_j \in \mathbb{R}^K$ denote observable characteristics of i and j, respectably. Let $\mathbf{X} \overset{def}{=} [X_1, X_2, \ldots, X_N] \in \mathbb{R}^{K\times N}$, the matrix containing the characteristics of all agents.

3.1 Background: Exchangeable Random Arrays

$\mathbf{Y}$ is essentially an array of $N(N-1)$ random variables, and its components are not necessarily independent. The non-independence is clear because the Input-Output Matrix is obtained by matrix inversion. Intuitively, Y_{ki} and Y_{kj} are not independent because both agents i and j buy from agent k. Similarly, Y_{ik} and Y_{jk} are not independent because both i and j supply to k. Such dependence structure makes specifying a probability distribution for $\mathbf{Y}|\mathbf{X}$ nontrivial.

The theory of exchangeable random arrays, which was studied separately by Aldous (1981) and Hoover (1979) and then refined by Crane and Towsner (2018), provides a foundation for specifying the probability distribution (also see Graham (2020)). $\mathbf{Y}$ is said to be relatively exchangeable with respect to $\mathbf{X}$ if $\left[Y_{\sigma(i)\sigma(j)}\right] \overset{D}{=} [Y_{ij}]$ for all permutations σ such that $X_{\sigma(i)} = X_i,\ i \in \{1, 2, \ldots, N\}$. In other words, $\mathbf{Y}$ is relatively exchangeable with respect to $\mathbf{X}$ if the distribution of $[Y_{ij}]$ is invariant under any re-branding (or exchange) of agents that are homogeneous in X. If the relative exchangeability condition holds, one may proceed as if $\mathbf{Y}$ is generated according to

$$Y_{ij} = h\left(X_i, X_j, U_i, U_j, V_{ij}\right) \tag{1}$$

for some graph function h. U_i and V_{ij} are such that $\{(X_i, U_i)\}_{i=1}^N$ and $\{V_{ij}\}_{\substack{i,j=1\\ i\neq j}}^N$ are IID, and they are independent from each other.[4] This result justifies using dyadic regression model to describe the formation of $\mathbf{Y}$ using $\mathbf{X}$. That is, one may specify $\Pr(\mathbf{Y}|\mathbf{X})$ as a function of $\Pr(Y_{ij}|X_i, X_j, U_i, U_j)$ after assigning a certain (parametric) distribution to V_{ij}. Here, U_i represents unobserved heterogeneity specific to i and V_{ij} unobserved heterogeneity specific to the pair (i, j). The data generating process implies that Y_{ij} and Y_{kl} are dependent if (i, j) and (k, l) share at least one common agent and independent if otherwise. Also note that one can construct any pair-level characteristics for (i, j) as a function of the agent-level characteristics X_i and X_j.

If $\mathbf{Y}$ satisfies the relative exchangeability condition, then we can describe its distribution as a collection of pair-level outcomes Y_{ij} for $i, j = 1, 2, \ldots, N$. That is, relative exchangeability is a required *a priori* assumption for dyadic regression models. Whether or not $\mathbf{Y}$ actually satisfies the condition depends on the quality of information contained in $\mathbf{X}$. For example, consider a model of worldwide trade flows. Let $\mathbf{Y}$ represent the trade flows between countries,[5] and $\mathbf{X}$ the country-level characteristics such as GDP, import duty rate, latitude, longitude, cultural aspects, and so forth. Consider country pairs (i_1, j_1) and (i_2, j_2). $\mathbf{Y}$ being exchangeable with respect to $\mathbf{X}$ implies that if $X_{i_1} = X_{i_2}$ and $X_{j_1} = X_{j_2}$, then $Y_{i_1 j_1} \overset{d}{=} Y_{i_2 j_2}$. Now consider the trade flow between South Korea and North Korea. The two countries are geographically close,[6] culturally similar, and the South has a much larger GDP than the North. The fact that they have a very small trade flow is because of political reasons.[7] If $\mathbf{X}$ does not contain information on political aspects of countries, then the relative exchangeability condition fails. In this case, it may not be sensible to model worldwide trade flows $(\mathbf{Y})$ as a collection of pairwise trade flows $(Y_{ij},\ i, j = 1, 2, \ldots, N)$.

[4] One may use a more flexible specification by replacing $\{V_{ij}\}_{\substack{i,j=1\\ i\neq j}}^{N}$ with $\{(V_{ij}, V_{ji})\}_{\substack{i,j=1\\ i\neq j}}^{N}$. Here, V_{ij} and V_{ji} are not necessarily independent.

[5] One can set $Y_{ii} = 0$ for all i to specify that a country does not trade with itself.

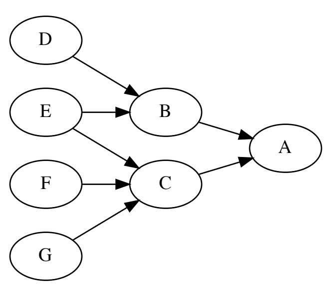
[6] One can use the latitude, longitude, and the Euclidean norm to compute geographical distance.

[7] As a comparison, North Korea has a large volume of trade (relative to its GDP) with China. China and North Korea are geographically close, and China has a much larger GDP than North Korea.

3.2 Pair-Specific Fixed Effects

Under certain cases, such as the settings studied by Fafchamps and Gubert (2007) and Santos Silva and Tenreyro (2006), it may be plausible to assume relative exchangeability and proceed with a standard dyadic regression model. For the Input-Ouput matrix, there is no guarantee that $\mathbf{Y}$ is relatively exchangeable with respect to $\mathbf{X}$. For example, the relationship between the steel industry and the automobile industry could be different from the relationship between the agriculture industry and the semiconductor industry even if all observed industry-level characteristics were identical for the two pair of industries. Such difference is a *structural restriction* of the network. Figure 1 illustrates a hierarchical supply chain, which is a network with extreme structural restrictions. There are three agent tiers: (1) {D, E, F, G}, (2) {B, C}, and (3) {A}. Agents within the same tier cannot supply to or buy from each other. Agents in tier (1) can only supply to agents in tier (2). Finally, agents in tier (2) can only supply to agent(s) in tier (3). All other associations between agents are structurally impossible. Relative exchangeability does not hold if agent-level characteristics fail to capture the structural restrictions.

Figure 1: Hierarchical Supply Chain



The hierarchical supply chain is a network with extreme structural restriction. It is structurally impossible for agents D, E, F, G to be associated with each other or with agent A. Moreover, they can only supply to, not buy from, agents B or C. Similarly, it is structurally impossible for agents B and C to be associated with each other. They can only supply to agent A or buy from agents D, E, F, G. Finally, agent A can only buy from agents B and C. Relative exchangeability fails unless agent-level characteristics capture such structural impossibilities.

In general cases in which structural restrictions are present (and directly modeling the structural restrictions is impossible), a plausible workaround is to include pair-specific fixed effects in dyadic regression model. Here, the *a priori* modeling assumption is that $\mathbf{Y}$ is relatively exchangeable with respect to $\mathbf{X}$ and pair-specific fixed effects. In other words, we expect the pair-specific fixed effects to capture structural restrictions of the network not explained by $\mathbf{X}$. Graham (2017) and Jochmans (2017) are examples of recent research in this topic. In particular, Graham (2017) studies the estimation of dyadic regression models with pair-specific fixed effects for undirected networks (i.e., $Y_{ij} = Y_{ji}$ for all i, j), and Jochmans (2017) studies the estimation for directed networks (i.e., not necessarily $Y_{ij} = Y_{ji}$). Naturally, pair-specific fixed effects are undirected for an undirected network, and they are directed for a directed network. The Input-Ouput matrix is a directed network; thus, similar to Jochmans (2017) the fixed effects associated with (i, j) is not the same as the fixed effects associated with (j, i). In particular, the former captures the fixed effects associated with industry i as a supplier and industry j as a buyer, and the latter with the the roles of i and j switched.

A difficulty in estimation is that the number of pair-specific effects is $O(N^2)$, where N is the number of agents in the network, and there are insufficient number of "observations" to consistently estimate the pair-specific fixed effects. Using the expression of Neyman and Scott (1948), the pair-specific fixed effects are *incidental*, and the estimation of model parameter is inconsistent because of them. To avoid the incidental parameter problem, Graham (2017)[8] and Jochmans (2017) both utilize pairs of pairs, or tetrads, to difference away the pair-specific fixed effects. Such method allows one to consistently estimate model parameters and make inferences on average marginal effects. However, the number of tetrads is $O(N^4)$, and computing differences of functions across 370^4 number of tetrads is a demanding task even for a modern workstation. Moreover, one cannot make counterfactual predictions of $\mathbf{Y}$ because the pair-specific fixed effects are not estimated. Lastly, differencing away pair-specific fixed effects works only if the model's distribution is a member of the exponential

[8]Here, I am referring specifically to the "tetrad logit", which is one of the two estimators proposed by Graham (2017).

family. This article alleviates these problems by using group-level fixed effects, where the groups are determined by a stochastic blockmodel.

3.3 Stochastic co-Blockmodel

A stochastic blockmodel, first introduced by Holland et al. (1983), partitions a network into isomorphism classes or *blocks*, where two agents are isomorphic if their relationships to the rest of the network are identical. That is, for any pair of blocks (B, B^*), one expects that $Y_{i_1 j_1} \overset{d}{=} Y_{i_2 j_2}$ if $i_1, i_2 \in B$ and $j_1, j_2 \in B^*$. I utilize the stochastic co-blockmodel proposed by Rohe et al. (2016), called "DI-SIM". DI-SIM is a stochastic blockmodel for directed networks, and it generates two partitions to account for different roles of agents in a pair. The next paragraph briefly explains the idea of DI-SIM.

In the network perspective, $Y_{ij} > 0$ if seller i *sends* a "link" to buyer j or, equivalently, buyer j *receives* a "link" from seller i. With S (number of supplier blocks) and R (number of buyer blocks) determined *a priori,* DI–SIM performs singular value decomposition on $\mathbf{Y}$ and then K-means clustering on the singular vectors. In particular, performing K-means clustering on the left (with S clusters) and right singular vectors of $\mathbf{Y}$ (with R clusters) assign each agent the membership to a supplier block and a buyer block, respectively. This process partitions the network into supplier blocks $\{B_s\}_{s=1}^S$ and buyer blocks $\{B_r^*\}_{r=1}^R$, so that each agent belongs to exactly one of the supplier blocks based on its role as a supplier and belongs to exactly one of the buyer blocks based on its role as a buyer in the network. For any pair of (supplier, buyer) blocks (B_s, B_r^*), one expects that

$$Y_{i_1 j_1} \overset{d}{=} Y_{i_2 j_2} \text{ if } i_{1,} i_2 \in B_s \text{ and } j_1, j_2 \in B_r^*. \tag{2}$$

A stochastic blockmodel is an estimation, and there could be errors in the assignment of block memberships. The output of DI-SIM is *weakly consistent* for $\{B_s\}_{s=1}^S$ and $\{B_r^*\}_{r=1}^R$ that satisfies expression (2), meaning that the fraction of incorrectly specified block memberships converges in probability to zero as $N \to \infty$, provided that the average degree of the network grows faster than $\log(N)$ rate. To be precise, if the average degree $\lambda_N \overset{def}{=} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} Y_{ij}$ satisfies

$$\lim_{N\to\infty} \frac{\lambda_N}{\log(N)} = \infty, \tag{3}$$

then $\frac{\delta_N}{N} = o_p(1)$, where δ_N is the number of incorrectly assigned block memberships. The requirement $\lim_{N\to\infty} \frac{\lambda_N}{\log(N)} = \infty$ can be satisfied if (1) every agent in the network has a "link" with infinitely many other agents as $N \to \infty$ or (2) there are several agents that have infinitely many "links" while others only have finite number of links as $N \to \infty$. Case (1) is called a dense network. For example, an Erdös Rényi network that assigns a strictly positive $\Pr(Y_{ij} > 0)$ is dense. Case (2) is an example of a sparse network with several "important" agents. The Input-Output Matrix belongs to case (1).

3.4 Dyadic Regression Model with Block-Specific Fixed Effects

Expression (2) states that $\mathbf{Y}$ is relatively exchangeable with respect to the block memberships. This statement is trivially true because exchanging the block memberships between two agents is synonymous to exchanging their relationships to the rest of the network. Moreover, because subsets within each block must conform to block-level relationships, it is also true that

$$Y_{i_1 j_1} \overset{d}{=} Y_{i_2 j_2} \text{ if } i_1, i_2 \in B_s,\ j_1, j_2 \in B_r^*,\ \text{and}\ (X_{i_1}, X_{j_1}) = (X_{i_2}, X_{j_2}) \tag{4}$$

for all $B_s,\ s = 1, 2, \ldots, S$ and $B_r^*,\ r = 1, 2, \ldots, R$, provided that the support of X_i is the same for all i. Let $M_i \in \{B_s\}_{s=1}^S \times \{B_r^*\}_{r=1}^R$ denote the (supplier block, buyer block) i belongs to. Define $\mathbf{M}$ as the collection of M_i for all $i = 1, 2, \ldots, N$. Expression (4) suggests that $\mathbf{Y}$ is relatively exchangeable with respect to $(\mathbf{M}, \mathbf{X})$, and thus one may describe the distribution of $\mathbf{Y}$ as a collection of

$$Y_{ij} = h\left(M_i, M_j, X_i, X_j, U_i, U_j, V_{ij}\right),\ i, j = 1, 2, \ldots, N. \tag{5}$$

I assume $\{M_i\}_{i=1}^N$ are IID. Similar to expression 1 discussed in section 4.1, $\{(M_i, X_i, U_i)\}_{i=1}^N$ and $\{(V_{ij})\}_{\substack{i,j=1\\ i\neq j}}^N$ are IID and independent from each other.

Because $\{X_i\}_{i=1}^N$ are agent-level IID covariates, $\{M_i\}_{i=1}^N$ being IID is synonymous to $\left\{M_i \middle| X_i\right\}_{i=1}^N$ being IID. As described in section 4.2, M_i conditional on X_i is precisely the structural restriction associated with agent i. That is, what remains in M_i after controlling for X_i represents the relationship of agent i to the rest of the network which cannot be explained by X_i. For the Input-Output Matrix, one could imagine that nature randomly endows industries with "popularity" as a seller and as a buyer. Then the characteristics of industry i, X_i, determines the remaining part of M_i.

With $\mathbf{Y}$ represented as a collection of expression (5), one may construct a (composite) likelihood function for $\mathbf{Y}$ by specifying a distribution for V_{ij}. Conditional on (M_i, X_i, M_j, X_j) and (U_i, U_j), the randomness in Y_{ij} is driven entirely by V_{ij}, and constructing the likelihood function is straightforward. However, (U_i, U_j) is latent and unconditional on (U_i, U_j), Y_{ij} and Y_{kl} are not independent if $k \in \{i, j\}$ or $l \in \{i, j\}$. Without making assumptions on the joint distribution of $(U_1, U_2, \ldots, U_N)$, one can only construct a composite likelihood function.

Let $\alpha_S \in \mathbb{R}^S$ and $\alpha_R \in \mathbb{R}^R$ denote the vector of fixed effect coefficients as a supplier and as a buyer. Observe that the dimension of α_S and α_R are, respectively, equal to the number of supplier and buyer blocks. Let $M_{S,i} \in \{0,1\}^S$ be a unit vector whose s^{th} component is 1 if i belongs to supplier block B_s. Similarly, let $M_{R,i} \in \{0,1\}^R$ be a unit vector whose r^{th} component is 1 if i belongs to buyer block B_r^*. Thus, $\alpha_S' M_{S,i} + \alpha_R' M_{R,j}$ denotes the two-way block-fixed effects associated with the pair (i, j) if i belongs to supplier block B_s and j belongs to buyer block B_r^*. A general form of the log-likelihood function for the pair

(i, j) is

$$l_{ij}\left(\theta; Y_{ij}, M_i, M_j, X_i, X_j\right) = f\left(Y_{ij}, \alpha'_S M_{S,i} + \alpha'_R M_{R,j} + \beta' X_{ij}\right) \tag{6}$$

where $\theta = (\alpha_S, \alpha_R, \beta)$ denotes the estimand. $X_{ij} = g\left(X_i, X_j\right)$ is a vector of pair-level covariates. The expression $g\left(X_i, X_j\right)$ indicates that any pair-level covariates can be constructed as a function of individual-level characteristics. Note that the density function $f(\cdot, \cdot)$ needs not be a member of the exponential family. The associated composite likelihood is

$$L_N\left(\theta; \mathbf{Y}, \mathbf{M}, \mathbf{X}\right) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} l_{ij}\left(\theta; Y_{ij}, M_i, M_j, X_i, X_j\right). \tag{7}$$

4 Estimation

4.1 Two-Stage Estimation

Assume (1) the parameter space Θ is compact, (2) l_{ij} is twice continuously differentiable in θ a.s., and (3) the true parameter $\theta_0 \in \Theta$ is identified. In a hypothetical world, suppose one knew the true block memberships $\mathbf{M}$ that satisfies condition (2) for all pairs of blocks. The maximizer of the *oracle* composite likelihood $L_N\left(\theta; \mathbf{Y}, \mathbf{M}, \mathbf{X}\right)$ is consistent for θ_0 and asymptotically normally distributed after scaling by $\sqrt{N}$, provided that the marginal densities l_{ij} are properly specified (Graham 2020, p. 144). One may estimate the associated asymptotic covariance using the method proposed by Fafchamps and Gubert (2007) or another method that accounts for the dependence between l_{ij} (see Graham 2020 pp. 150-154).

However, the block memberships generated by DI-SIM is an estimator of $\mathbf{M}$. Let $\hat{\mathbf{M}}$ denote the output of DI-SIM. The augmented-data composite likelihood is

$$L_N\left(\theta; \mathbf{Y}, \hat{\mathbf{M}}, \mathbf{X}\right), \tag{8}$$

and, it contains uncertainty in $\hat{\mathbf{M}}$. The next proposition suggests that working with the augmented-data composite likelihood is asymptotically equivalent (in probability) to working with the oracle composite likelihood if $\hat{\mathbf{M}}$ is weakly consistent for $\mathbf{M}$.

Theorem 1. *Maintain the three assumptions above. If $\hat{\mathbf{M}}$ is weakly consistent for $\mathbf{M}$, then*

$$\sup_{\theta\in\Theta}\left|L_N\left(\theta;\mathbf{Y},\hat{\mathbf{M}},\mathbf{X}\right)-L_N\left(\theta;\mathbf{Y},\mathbf{M},\mathbf{X}\right)\right|=o_p\left(1\right). \tag{9}$$

Proof of Theorem 1 is given in the appendix. A natural corollary is that

$$\hat{\theta}=\arg\max_{\theta\in\Theta}L_N\left(\theta;\mathbf{Y},\hat{\mathbf{M}},\mathbf{X}\right) \tag{10}$$

is asymptotically equivalent to the maximizer of $L_N\left(\theta;\mathbf{Y},\mathbf{M},\mathbf{X}\right)$. Thus, $\hat{\theta}$ is consistent for θ_0 and $\sqrt{N}$-asymptotically normal.

4.2 Choosing the Number of Blocks

To construct $\hat{\mathbf{M}}$, one must choose the number of supplier blocks S and the number of buyer blocks R. In this section, I provide a broad guideline on the optimal choice of (S,R).

First, S and R must be chosen so that the model is not subject to the incidental parameter bias. Define T_s as the number of outgoing "links" associated with the supplier block B_s. In other words, T_s is the total number of "links" that agents in B_s send to agents outside of B_s. Similarly, define T_r as the number of incoming "links" associated with the buyer block B_r. As before, T_r is the number of "links" that agents in B_r receives from agents outside of B_r. That is

$$\begin{aligned}T_s&=\sum_{i\in B_s}\sum_{j\notin B_s}\mathbf{1}\left(Y_{ij}>0\right)\\T_r&=\sum_{j\in B_r}\sum_{i\notin B_r}\mathbf{1}\left(Y_{ij}>0\right).\end{aligned} \tag{11}$$

It is natural to assume that all agents in the network sends and receives at least one link. Thus $T_s,T_r\to\infty$ as $\left|B_s\right|,\left|B_r\right|\to\infty$.

Observe that T_s and T_r are numbers of effective observations for estimating the fixed effect associated with the block B_s and B_r, respectively. To avoid the incidental parameter problem, we must have (c.f. Fernandez-Val and Weidner, 2018)

$$\frac{S}{\min\{T_s\}_{s=1}^{S}}, \frac{R}{\min\{T_r\}_{r=1}^{R}} \to 0, \text{ as } N \to \infty. \tag{12}$$

Expression (12) suggests upper bounds for S and R. For example, one may choose S and R that satisfy $S < \sqrt{\min\{T_s\}_{s=1}^{S}}$ and $R < \sqrt{\min\{T_r\}_{r=1}^{R}}$.

5 Empirical Analysis of Input-Output Matrix

The empirical analysis uses the 2017 Commodity-by-Commodity Input-Output Matrix of the U.S. economy and the domestic and foreign trade accounts published by the BEA (Bureau of Economic Analysis, 2024). The Input-Output Matrix contains information on supplier-buyer relationships across 370 commodities.[9] Inspection of the Input-Output Matrix for all years reveals that elements on the main diagonal are close to 1. Due to the lack of variation, elements of each column has been normalized so that all elements on the main diagonal are exactly 1. Due to the normalization, the interpretation of Y_{ij} becomes the amount of commodity i, relative to commodity j, that is necessary to produce \$1 of commodity j. In other words, Y_{ij} is the dependence of j on the input from i relative to its own. The empirical analysis does not use the main diagonal.

Similar to Santos Silva and Tenreyro (2006), the composite likelihood function is specified as

$$l_{ij} = Y_{ij} \log(\mu_{ij}) - \mu_{ij}, \tag{13}$$

where

$$\mu_{ij} = \exp(\alpha_S' M_{S,i} + \alpha_R' M_{R,j} + \beta' X_{ij}) \tag{14}$$

[9] The original Input-Output Matrix contains 405 commodities. They are consolidated to 370.

is the mean of the pseudo-Poisson distribution. Note that for each i, I divided the entries on the i^{th} column by Y_{ii}. X_{ij} contains 17 pair-level covariates,[10] excluding block-specific fixed effects.

Table 1: Empirical Analysis of 2017 Commodity-by-Commodity Input-Output Matrix

	w/ Constant Term		**w/ Fixed Effects $S = 5, R = 7$**	
Covariate	**Coef**	**SE**	**Coef**	**SE**
Log Final Demand (supplier)	0.0986**	0.0465	0.0947**	0.0501
Log Final Demand (buyer)	0.0004	0.0091	0.0024	0.0093
Net Tax Rate (supplier)	-4.1495	3.1003	-4.1044	3.291
Net Tax Rate (buyer)	-1.4002**	0.5079	-1.3022**	0.5312
Import Duty Rate (supplier)	-9.9164	6.4110	-11.2378*	6.4702
Import Duty Rate (buyer)	4.1163**	1.4798	3.6851**	1.4982

** indicates the coefficient is different from 0 with 95 percent confidence level.

* indicates the coefficient is different from 0 with 90 percent confidence level.

Table 1 reports the estimation results. The second column reports estimation result with a constant term, and the third column reports it with block-specific fixed effects, respectively. **Describe the table.**

6 Concluding Remarks

If observed covariates fail to explain structural restrictions over pairs of agents, then pairs of agents are not exchangeable. Then even if two pairs of agents are observationally indistin-

[10]The table below shows 6 covariates. The remaining 11 covariates are amount of import, export, import used for production (rather than end-use consumption), trade margin (profit from reselling), transportation cost for i and j, and an indicator for i and j having the same first 2-digit NAICS code.

guishable, the distribution over the connection between agents in one pair can be different from the distribution over the other pair. In the presence of such structural restrictions, using dyadic regression model to describe a network can suffer from model mis-specification. Including pair-specific fixed effects can alleviate this issue. However, we lose the ability to make counterfactual predictions, as fixed effects are differenced away to avoid the incidental parameter problem.

This article uses a stochastic blockmodel to group agents into blocks, so that agents who belong to the same pair of blocks share common pairwise block-specific fixed effects. Doing so allows making counterfactual predictions on the network and also reduces computation burden. It shows that block-specific fixed effects and other model parameters can be consistently estimated and are asymptotically normally distributed provided that the network is sufficiently dense. Moreover, an empirical analysis using the 2017 U.S. Input-Output Matrix shows that the supplier-buyer relationship between commodities (or industries) are sensitive to changes in import duty rates.

A potential weakness in making counterfactual predictions is that blocks are fixed. If we assign counterfactual characteristics to certain agents, their and possibly others' memberships to blocks could change. However, addressing it requires modeling block formations within a network, which is a burgeoning area of research. I leave for future research to implement endogenous block formation.

References

Aldous, D. J. (1981). "Representations for Partially Exchangeable Arrays of Random Variables". *Journal of Multivariate Analysis.*

Bureau of Economic Analysis (2024). "Input Output Accounts Data". Retrieved from BEA website. https://www.bea.gov/industry/input-output-accounts-data (accessed March 22, 2024).

Charbonneau, K. (2017). "Multiple Fixed Effects in Binary Response Panel Data Models". *The Econometrics Journal* 20, S1–S13.

Crane, H. and H. Towsner (2018). "Relatively Exchangeable Structures". *Journal of Symbolic Logic.*

Fafchamps, M. and F. Gubert (2007). "Risk Sharing and Network Formation". *American Economic Review.*

Fernandez-Val, I. and M. Weidner (2018). "Individual and Time Effects in Nonlinear Panel Models witrh Large N, T". *Journal of Econometrics.*

Graham, B. S. (2017). "An Econometric Model of Network Formation with Degree Heterogeneity". *Econometrica.*

— (2020). "Network Data". *Handbook of Econometrics* 7A, 111–218.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). "Stochastic Blockmodels: First Steps". *Social Networks.*

Hoover, D. N. (1979). "Relations on Probability Spaces and Arrays of Random Variables".

Jochmans, K. (2017). "Two Way Models for Gravity". *Review of Economics and Statistics.*

Leontief, W. W. (1951). "Input-Output Economics". *Scientific American.*

Ma, S., L. Su, and Y. Zhang (2021). "Determining the Number of Communities in Degree-corrected Stochastic Block Models". *Journal of Machine Learning Research* 22, 1–63.

Neyman, J. and E. E. Scott (1948). "Consistent Estimates Based on Partially Consistent Observations". *Econometrica.*

Rohe, K., T. Qin, and B. Yu (2016). “Co-clustering Directed Graphs to Discover Asymmetries and Directional Communities”. *Proceedings of the National Academy of Sciences.*

Santos Silva, J. M. and S. Tenreyro (2006). “The Log of Gravity”. *Review of Economics and Statistics.*

Su, L., Z. Shi, and P. C. B. Phillips (2016). “Identifying Latent Structures in Panel Data”. *Econometrica.*

Su, L. and W. Wang (2021). “Identifying Latent Group Structures in Nonlinear Panels”. *Journal of Econometrics* 220.2, 272–295.

Appendix

A. Proofs

Lemma 2. *Let W denote the number agents with an incorrectly assigned block membership. For $W \geq 2$, the number of incorrectly pair-level objective functions is at most*

$$2NW - W^2 - W. \tag{15}$$

Proof. Suppose agent i has incorrectly assigned block memberships (supplier block, buyer block, or both). There are $N-1$ pair-level objective functions with i as the supplier and $N-1$ pair-level objective functions with i as the buyer, thus leading to at most $2(N-1)$ incorrect pair-level objective functions. If W number of agents have incorrectly assigned block memberships, ignoring double counting, there are at most $2(N-1)W$ incorrect pair-level objective functions. Observe that double-counting occurs only if the block memberships of both (i,j) are incorrectly assigned, and there are $W(W-1)$ the number of ordered pairs among W number of agents. □

Proof. **(Proof Theorem 1)** Let $\theta = (\alpha_S, \alpha_R, \beta)^{\mathrm{T}} \in \Theta$. We have

$$L_N\left(\theta; \mathbf{Y}, \hat{\mathbf{M}}, \mathbf{X}\right) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} l_{ij}\left(\theta; \hat{M}_i, \hat{M}_j, X_i, X_j\right) \tag{16}$$

$$L_N\left(\theta; \mathbf{Y}, \mathbf{M}, \mathbf{X}\right) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} l_{ij}\left(\theta; M_i, M_j, X_i, X_j\right) \tag{17}$$

where

$$l_{ij}\left(\theta; \hat{M}_i, \hat{M}_j, X_i, X_j\right) = f\left(Y_{ij}, \alpha_S' \hat{M}_{S,i} + \alpha_R' \hat{M}_{R,j} + \beta' X_{ij}\right) \tag{18}$$

$$l_{ij}\left(\theta; M_i, M_j, X_i, X_j\right) = f\left(Y_{ij}, \alpha_S' M_{S,i} + \alpha_R' M_{R,j} + \beta' X_{ij}\right). \tag{19}$$

Fix $\theta \in \Theta$. It is clear that

$$\left|\left(\alpha_S' \hat{M}_{S,i} + \alpha_R' \hat{M}_{R,j}\right) - \left(\alpha_S' M_{S,i} + \alpha_R' M_{R,j}\right)\right| < \infty \ a.s. \tag{20}$$

and thus

$$\left| l_{ij}\left(\theta; \hat{M}_i, \hat{M}_j, X_i, X_j\right) - l_{ij}\left(\theta; M_i, M_j, X_i, X_j\right) \right| < \infty \ a.s. \tag{21}$$

Moreover, because Θ is compact and l_{ij} is differentiable in θ,

$$\sup_{\theta \in \Theta} \left| l_{ij}\left(\theta; \hat{M}_i, \hat{M}_j, X_i, X_j\right) - l_{ij}\left(\theta; M_i, M_j, X_i, X_j\right) \right| \stackrel{def}{=} \lambda < \infty \ a.s. \tag{22}$$

Thus for sufficiently large N, with probability 1,

$$\sup_{\theta \in \Theta} \left| L_N\left(\theta; \mathbf{Y}, \hat{\mathbf{M}}, \mathbf{X}\right) - L_N\left(\theta; \mathbf{Y}, \mathbf{M}, \mathbf{X}\right) \right| \leq \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \lambda \tag{23}$$

$$\leq \frac{1}{N(N-1)} \left[2NW - W^2 - W\right] \lambda \tag{24}$$

$$\leq \frac{2NW}{N(N-1)}\lambda - \frac{W^2}{N(N-1)}\lambda - \frac{W}{N(N-1)}\lambda, \tag{25}$$

where the second inequality follows from Lemma 2. Weak consistency of $\hat{\mathbf{M}}$ gives

$$\frac{W}{N} = o_p(1). \tag{26}$$

□

B. Figures and Tables

Table 2: Empirical Analysis of 2007 Commodity-by-Commodity Input-Output Matrix

	w/ Constant Term		**w/ Fixed Effects $S, R = 5$**	
Covariate	**Coef**	**SE**	**Coef**	**SE**
Log Final Demand (supplier)	0.1552**	0.0430	0.1586**	0.0456
Log Final Demand (buyer)	-0.0041**	0.0087	-0.0029	0.0085
Net Tax Rate (supplier)	-1.1057	4.3546	-0.4768	4.1456
Net Tax Rate (buyer)	-1.9540**	0.4877	-1.7541**	0.4254
Import Duty Rate (supplier)	-11.2995*	6.8424	-13.4337**	6.6916
Import Duty Rate (buyer)	3.3973**	1.4799	3.2288**	1.5172

** indicates the coefficient is different from 0 with 95 percent confidence level.

* indicates the coefficient is different from 0 with 90 percent confidence level.