Connected Trade Flows via Trade Costs: A Spatial Autoregressive Framework

Hanbat Jeong* Jieun Lee†

January 30, 2026

Abstract

This paper proposes a new econometric specification for gravity equations grounded in the countries' network-leveraging features. Building on spatial autoregressive specifications, we endogenize trade costs by characterizing them as a function of a network of countries' proximities. Under the resulting model specification, the network-leveraging feature induces interdependence among trade flows and pair-specific heterogeneity. The conventional iceberg cost is a special case when countries do not leverage their network connections. For estimation, we use the PPML estimator and develop a robust method that accommodates heteroskedasticity and unknown error correlations for inference. Moreover, we develop a novel algorithm for a linear transformation from a large-dimensional network multiplier matrix, enabling much faster computation. In the empirical application, we find evidence of endogenous trade costs. Our model significantly improved fit relative to the conventional gravity equation, with gains of up to 30%. In detail, it implies the existence of pair-specific heterogeneities not captured by the existing models. By counterfactual analysis, we show that a bilateral increase in trade costs leads to substantial reallocations of import shares across a wide range of third countries, once network interactions are accounted for.

Keywords: Origin-destination flows, International trade, Gravity equation, Network, Iceberg cost, Spatial autoregressive model, Poisson pseudo-maximum likelihood estimation, Eigenvalues

JEL codes: C13, C31, F1.

*Department of Economics, Macquarie Business School, Macquarie University. E-mail: hanbat.jeong@mq.edu.au.

†Department of Economics, Emory University. E-mail: jieunlee.sophia@gmail.com.

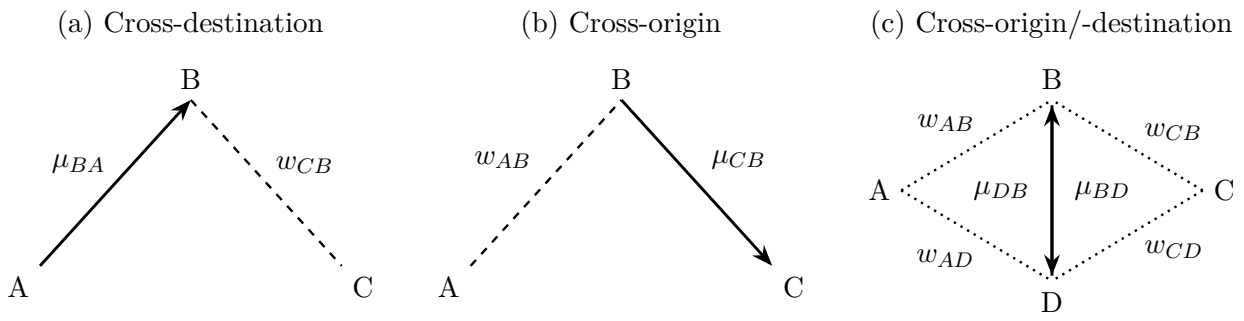
1 Introduction

How do bilateral trade policy shocks—such as sanctions, or supply-chain disruptions—propagate through the trade network to affect third-country trade patterns? The recent China-US trade war may significantly affect their nearby countries like Korea and Japan, and beyond, the far remote countries. Similarly, trade policy shocks are rarely confined to the countries they directly target. In an interconnected trade network, changes in bilateral trade costs can propagate through origin-, destination-, and hub-based linkages, reshaping trade patterns far beyond the affected pair. This paper asks whether existing gravity-based policy evaluations systematically understate the scope and consequences of trade policy interventions. Throughout our new model, we show that once trade costs respond to network-wide reallocations, bilateral shocks generate large third-country reallocation and concentration effects that standard gravity largely misses.

Gravity equations are the workhorse for explaining bilateral trade flows. In the structural gravity framework, how trade costs are specified is crucial because it shapes both the composition of trading partners and the quantities exchanged. The iceberg-cost specification—under which a fraction of the shipped good "melts" in transit—has become the canonical choice: it is microfounded across standard models, enters trade shares multiplicatively, and maps cleanly to identification and counterfactual analyses. In empirical work, distance and tariffs are used as proxies for this ad-valorem–type cost. This conventional formulation treats trade costs as exogenous and unavoidable. Moreover, these costs are assumed to form independently across country pairs, even when some pairs (e.g., A–C or B–D) are geographically or logistically connected. This is distinct from multilateral resistance, which captures general-equilibrium expenditure reallocation; our focus is on how trade costs themselves become network-dependent across origins, destinations, and third-country hubs.

This paper revisits that conventional view by asking to what extent countries respond to bilateral trade costs by exploiting information about expected third-party flows—for instance, by routing or collaborating shipments. That is, higher expected flows can proxy route reliability/scale economies; but may raise costs under capacity constraints. Figure 1 sketches the core mechanisms. Consider three countries, A, B, and C. If exports from A to B are expected to be large and B is strongly linked to C (panel a), routing via B allows sharing of schedules and fixed logistics, reducing the effective A→C cost. Suppose instead that A and B are closely connected and a large B→C flow is expected (panel b); then A can bundle A→B shipments with B→C services, again lowering the effective cost to C. Third, expected flows among third parties (e.g., B↔D), combined with strong proximities (A–D and B–C), can make multi-leg routes (A→D→B→C or A→B→D→C) cost-effective (panel

Figure 1: Network linkages in the formation of bilateral trade costs



Note: μ denotes an expected trade flow and w a connectivity measure. Panels (a)–(c) depict three channels—cross-destination, cross-origin, and hub (third-country)—through which bilateral trade costs can become network-dependent. These linkages may reduce effective A→C trade costs, but may also increase them when congestion or capacity constraints bind.

c). Conversely, if shipping capacity is constrained, large expected flows on adjacent routes (e.g., A→B, B→C, B↔D) may raise the effective cost of A→C. Hence, trade costs depend on expectations about third-party flows, generating cross-origin and cross-destination linkages in trade costs that a purely pairwise iceberg view misses. These considerations motivate the development of a theoretical and econometric framework that endogenizes trade costs through network-based linkages.[1]

We develop a gravity model with microfoundations in which trade costs are endogenized. To do so, we adopt a spatial autoregressive (SAR) operator structure to capture network interactions. The core hypothesis is that expected trade flows, filtered through countries' connectivity, shape bilateral costs, rather than costs being driven solely by geographic distance. Let y_{ij} denote the observed flow from j to i and μ_{ij} its systematic component. The model proceeds in three stages. Stage 1 shapes the country (long-run) connectivity matrix $W = (w_{ij})$: each w_{ij} indicates how strongly i is linked to j in trade-relevant proximity. Hence, W captures the relevance of nearby routes/hubs for each country-pair's shipping options. Stage 2 specifies ad-valorem–type trade costs as the product of a network-driven component (a function of expected flows and W) and standard bilateral covariates. Stage 3 characterizes equilibrium trade flows via a CES demand system à la Anderson and van

[1]Hub ports (e.g., Singapore or Dubai) illustrate how third-party flows create network economies: by collaborating shipments across origins and destinations, they reduce effective costs even for countries that are not directly linked. Similarly, in container shipping between Asia and North America, eastbound flows (Asia→US) are typically much larger than westbound flows, leading to costly empty-container repositioning. When westbound exports expand, carriers exploit backhaul opportunities, thereby lowering average eastbound costs. Analogous mechanisms appear in air cargo between Asia and Europe and in long-haul trucking, where return-leg demand reduces unit costs by sharing fixed logistics and scheduling across directions.

Wincoop (2003).

The endogenous part of the cost function is multiplicative (ad-valorem–type) and depends on geometric averages of expected flows weighted by connectivity $\{w_{ij}\}$. Three sets of terms capture distinct network channels: (i) outflows from a common origin (A$\rightarrow \cdot$) induce cross-destination linkages (routing, backhaul); (ii) inflows to a common destination ($\cdot \rightarrow$C) induce cross-origin linkages (collaboration toward the same market); and (iii) third-party flows capture hub-and-spoke and multi-leg routing effects.

We take the equilibrium mapping from the theoretical framework to the data and estimate a gravity equation in which bilateral trade costs depend on (i) network spillover elasticities—capturing how connected flows across origins, destinations, and third-country hubs shape the effective cost of trading—and (ii) the usual elasticities on bilateral cost shifters such as distance and borders. Network spillovers can lower effective trade costs through coordination, routing, and bundling, but they can also raise costs when capacity constraints bind. Accordingly, the sign of the spillover elasticities distinguishes efficiency amplification from congestion transmission. Crucially, these spillovers generate systematic pair-specific heterogeneity that conventional gravity specifications would leave in the residual. Empirically, the model uncovers systematic pair-specific variation that conventional gravity absorbs into residuals—variation that is correlated with proximity to common routes and hubs. Setting the network spillover elasticities to zero nests the standard iceberg-cost gravity model.

To ensure a well-defined econometric model, we characterize equilibrium trade flows and their uniqueness. Our model's equilibrium is characterized by a combination of two contraction mappings: (i) conventional multilateral resistance terms, and (ii) the network SAR operator. Then, the derived equilibrium follows the SAR structure embedded in an exponential functional form. This representation is a semi-reduced form obtained by only inverting the network SAR operator, as the fixed-effect components are implicit functions of the systematic part of trade flows—they include multilateral resistance terms, which are aggregations of countries' expenditure shares weighted by endogenized trade costs. This semi-reduced form highlights a simple propagation logic: a shock to one bilateral pair changes expected flows elsewhere, which feeds back into trade costs and reallocates trade across the entire network. The resulting econometric model represents y_{ij} as a combination of the expected trade flow μ_{ij} and an error term.

We consider the Poisson pseudo-maximum likelihood (PPML) estimator. This method only requires a correctly specified conditional mean of y_{ij}, so we allow for arbitrary heteroskedasticity and correlation in the error terms and construct heteroskedasticity- and

spatial-correlation-robust (HAC) standard errors. Robust inference is crucial because the network dependence embedded in the conditional mean may also induce spatial dependence in the errors, so we remain agnostic about the correlation structure. We specify identification conditions and derive the asymptotic distributions of the PPML estimators. Relative to PPML estimation in conventional gravity, our model further restricts the admissible parameter space due to network feedback. We assess finite-sample performance via Monte Carlo simulations and find small bias and accurate coverage in realistic sample sizes.

Methodologically, we further develop a spectral algorithm for computing transformations by the network multiplier matrix (inverse of the network SAR operator). This procedure is essential since inverting the network SAR operator is sufficient for estimation. However, if there are n countries in a sample, the network multiplier matrix has dimension $N \times N$ with $N = n^2$. Rather than inverting this $N \times N$ matrix, the algorithm exploits the eigendecomposition of W to simultaneously diagonalize the three Kronecker-type network operators for cross-destination, cross-origin, and joint origin–destination linkages. As a result, transforming a vector by the network multiplier matrix can be implemented via simple elementwise operations in the eigenbasis. This approach dramatically reduces the computational burden relative to the naive inverse-based approach. In the case of our empirical application ($n \simeq 150$), the proposed spectral algorithm is more than four orders of magnitude faster (relative to direct matrix inversion), rendering the estimation procedure computationally feasible even for moderately large n.

Using our model, we study how network effects in countries' trade flows shape and reflect effective trade costs. Four benchmark phases of global trade that correspond to major shifts in the international environment are considered: Phase 1 (Trade liberalization), Phase 2 (Active NAFTA implementation), Phase 3 (Emergence of the China trade shock), and Phase 4 (Expansion of global supply chains).[2] We find significant spillovers in trade flows and pronounced sign changes in both cross-destination and cross-origin linkages, while the third-country (hub-related) component is significantly positive across all phases. Importantly, incorporating these spillovers via our model substantially improves the model fit relative to a conventional gravity equation with an iceberg cost specification (with gains of up to 30% in terms of a likelihood measure), providing quantitative evidence that third–party proximities and network-mediated trade costs play a first–order role in international trade flows.

We show that allowing bilateral trade costs to respond endogenously to network-wide reallocations materially changes the effects of trade policy. In conventional structural gravity, a bilateral cost shock can influence third-country trade shares through general-equilibrium

[2]Note that our model is for cross-sectional data. Hence, we separately estimate the model across the designated phases.

adjustments summarized by multilateral resistance, but trade costs themselves remain pair-specific and do not directly inherit network linkages. In our framework, by contrast, changes in flows along connected origins, destinations, and third-country hubs feed back into effective bilateral costs, creating an additional propagation mechanism beyond multilateral resistance.

In counterfactuals motivated by the recent U.S.–China trade war, incorporating these network interactions implies that a bilateral increase in trade costs triggers sizable reallocations of import shares across a broad set of third countries, reshaping import concentration and diversification even in economies not directly targeted by the policy. This structural distinction explains why our model predicts larger and more widely dispersed third-country adjustments than conventional gravity benchmarks.

These results underscore that the global trading system is better viewed as an interconnected network than as a set of independent bilateral relationships. By allowing bilateral trade costs to be network-dependent, our framework links policy shocks to both direct bilateral responses and broader third-country reallocations that operate through connected origins, destinations, and hubs. The policy implication is straightforward: trade interventions should be evaluated and designed in a network-aware manner—anticipating rerouting and spillovers—and their incidence should be assessed not only at the targeted pair but across the wider set of connected economies. For policy evaluation, this means that assessments based solely on bilateral wedges and multilateral resistance can understate both the magnitude and the incidence of trade shocks, especially in economies tied to major hubs.

1.1 Our Contribution and Related Literature

The overall contribution of this work is to establish a new spatial econometric framework that illustrates gravity models with network interactions via endogenized trade costs. For this, we draw on two literatures—spatial econometrics and trade and spatial economics—and offer a distinct framework that links them through endogenous, network-dependent trade costs.

First, we contribute to the endogenization of trade costs by introducing countries' network-leveraging features in gravity models. This approach goes beyond traditional gravity-based trade models based on the canonical iceberg-cost specification (Samuelson (1952, 1954)), which treat bilateral trade costs as exogenous and pair-specific. Since the early gravity models of Tinbergen (1962), Anderson (1979), and Helpman and Krugman (1985), this framework has also been foundational in modern structural gravity, spanning (i) Armington-based models (McCallum, 1995; Anderson and van Wincoop, 2003; Arkolakis et al., 2012; Allen and Arkolakis, 2022; Wong, 2022), (ii) Ricardian-based models (Eaton and Kortum, 2002; Caliendo and Parro, 2015; Lind and Ramondo, 2023), and (iii) models with heteroge-

neous firms, fixed costs, and extensive margins (Melitz, 2003; Chaney, 2008; Helpman et al., 2008; Morales et al., 2019). Considering spillovers from/to third countries has also been regarded as the essential object even in the conventional framework based on the iceberg-cost specification. However, the spillover channel operates through multilateral resistance terms; consequently, these spillovers are summarized as country-specific fixed effects in the econometric model and limit the ability to illustrate heterogeneities arising from the relationships themselves.

A close parallel to our starting point for endogenizing trade costs is Brancaccio et al. (2020), who likewise recognize that trade costs reflect transportation-sector conditions and are therefore interrelated across links in the global network. Whereas Brancaccio et al. (2020) discipline trade costs using micro-level shipping data and explicit transport-sector foundations to establish some stylized facts, we provide a gravity-based econometric specification and implementation. That is, we identify and quantify network dependence in trade costs that is not fully captured by the multilateral resistance terms, and use these estimates to inform counterfactual policy analysis using standard gravity data and a country connectivity matrix.

To keep the focus on the formation of trade costs, we retain the workhorse Armington-based structural gravity framework (Anderson and van Wincoop, 2003) to illustrate trade-flow formation. This demand-side foundation provides a transparent mapping from multiplicative trade-cost wedges to aggregate bilateral flows, making our network channel directly estimable with gravity data using PPML. In this sense, our departure from the canonical framework lies in how trade costs are formed and become interdependent across pairs, not in the demand system that allocates expenditure across origins. The resulting model introduces a third set of parameters governing network-induced trade costs, in addition to standard demand and supply elasticities (Allen et al., 2020). Identification comes from the restriction that each bilateral cost loads on origin-, destination-, and hub-based network aggregates constructed from connectivity, which generates systematic cross-pair variation distinct from bilateral cost shifters.

Second, our modeling framework builds on the spatial econometrics literature, which models interdependence across space or networks through spatial lag structures. The specification is motivated by the SAR model (Cliff and Ord (1995); Ord (1975); Lee (2004, 2007); LeSage and Pace (2008)), in which outcomes for one unit depend on those of neighboring units. Recent extensions of SAR models to origin–destination (OD) flows (Jeong et al. (2023); Jeong and Lee (2024)) and international trade applications (Behrens et al. (2012); Jin et al. (2023)) have advanced the modeling of bilateral interdependence. As a computational contribution, we exploit the network structure to compute the required linear transformations

of the LeSage–Pace (2008) network multiplier efficiently.

However, existing SAR models for OD flows cannot directly address zero trade flows because they rely on log-linearized specifications. As noted by Santos Silva and Tenreyro (2006), log transformations for gravity equations can lead to biased inference and lead to an ad hoc manner in treating zeros. Building on our microfounded structure, we formulate the model at the original level, thereby avoiding the log-transformation issue associated with zero trade flows. For gravity equation estimation, we then extend the PPML framework (Gourieroux et al. (1984); Santos Silva and Tenreyro (2006); Head and Mayer (2014); Santos Silva and Tenreyro (2022)) and follow the same path of recent developments (Nagengast and Yotov (2025); Kwon et al. (2025)). Our derived asymptotic distribution of the PPML estimator is asymptotically bias-free, which is consistent with recent econometric advances in the estimation of linear/nonlinear models with fixed effects (Kapetanios et al. (2021); Weidner and Zylkin (2021); Fernandez-Val and Weidner (2016, 2018); Chen et al. (2021)). For inference, we adopt heteroskedasticity- and autocorrelation-consistent (HAC) methods for spatial data (Kelejian and Prucha (2007); Kim and Sun (2011); Conley et al. (2023)).

Third, our empirical results speak directly to the recent trade(-cost) literature. Using gravity data and a connectivity-based structure, we find strong evidence that bilateral trade costs are network-dependent, complementing the transportation-sector microfoundations in Brancaccio et al. (2020). Our estimates also align with the view that cross-country linkages are heterogeneous rather than uniform (Lind and Ramondo, 2023). Importantly, allowing for network dependence converts what conventional gravity would treat as residual variation into systematic pair-specific heterogeneity in trade flows. Across all phases, we find economically meaningful elasticities on linkages to nearby routes and hubs, consistent with evidence on the role of hubs in shaping trade and reallocation (Wong, 2022; Ganapati et al., 2024). Moreover, in phases associated with trade liberalization and the emergence of China-related trade shocks, we estimate negative network interactions, suggesting congestion or capacity constraints that offset network efficiencies. These patterns imply that effective policy evaluation and resilience analysis should be network-aware, with particular attention to capacity management at critical routes and hubs, echoing lessons from Allen and Arkolakis (2022) and Fuchs and Wong (2024).

The rest of the paper is organized as follows. In Section 2, we present the microfoundations of our econometric model. We then introduce an econometric specification and estimation/inference strategy. Section 3 provides the statistical analysis, including the asymptotic properties of our estimator and simulation evidence. In Section 4, we apply our framework to empirical data on world trade flows from the Center for International Data at UC Davis. Section 5 concludes.

2 Model

We will characterize a spatial model for an origin-destination (OD) flow y_{ij} $(i, j = 1, \cdots, n)$, denoting the directed flow from origin j to destination i. Suppose we have $N = n^2$ flow observations ($N = n(n-1)$ observations if intra flows are excluded), where we represent this data structure as an $n \times n$ matrix Y or $\mathbf{y} = \text{vec}(Y)$.[3] For its indexes, we define a pair of cross-sectional units $ij \equiv (i, j)$ to indicate a case originating from j and destined for i. The following introduces the location setting outlined by Jenish and Prucha (2009, 2012).

Assumption 2.1. Each $i \in \{1, \cdots, n\}$ is located in $\mathcal{D}_n \subset \mathcal{D}$, where $\mathcal{D}$ denotes a set of all potential locations in $\mathbb{R}^d$. We assume $\lim_{n\to\infty} \#(\mathcal{D}_n) = \infty$ and $\min_{i\neq j} d(l(i), l(j)) \geq 1$, where $\#(\mathcal{D}_n)$ is the cardinality of $\mathcal{D}_n$, $l : i \mapsto l(i) \in \mathcal{D}$ stands for an injective location function, and $d(l(i), l(j))$ is a distance between i and j.

Assumption 2.1 is widely employed in the spatial econometric literature (Qu and Lee, 2015; Xu and Lee, 2015a,b, 2018; Jeong and Lee, 2024). Beyond the geographic space, the concepts of $\mathcal{D}_n$ and $\mathcal{D}$ can be extended to a characteristic space that captures the economic and political locations of regions.

Traditional spatial autoregressive (SAR) models (Cliff and Ord, 1995; Ord, 1975; Lee, 2004, 2007) treat observations (y_i, x_i) collected based on Assumption 2.1, where y_i denotes an i's outcome and x_i is a regressor vector. The SAR model specifies how y_i's are interrelated:

$$y_i = \lambda \sum_{j=1}^{n} w_{ij} y_j + x_i'\beta + v_i, \text{ for } i = 1, \cdots, n, \tag{2.1}$$

where λ and β are the model's parameters, w_{ij} (element of an $n \times n$ spatial weighting matrix W) represents a strength of i being influenced by j, and v_i is an error. When $S \equiv I_n - \lambda W$ is invertible, (2.1) can be represented by

$$y_i = \sum_{j=1}^{n} (S^{-1})_{ij}(x_j\beta + v_j) = \sum_{j=1}^{n} (I_n + \lambda W + \lambda^2 W^2 + \cdots)_{ij}(x_j\beta + v_j). \tag{2.2}$$

Hence, $\mathbb{E}(y_i|x) = \sum_{j=1}^{n} (S^{-1})_{ij} x_j \beta$ and $\mathsf{Var}(y_i|x) = \sigma^2 (S^{-1}S^{-1\prime})_{ii}$ if $\mathbb{E}(v_i|x) = 0$ and $\mathsf{Var}(v_i|x) = \sigma^2 > 0$ for all $i = 1, \cdots, n$. Since equation (2.2) is a unique solution to (2.1), the SAR model represents how outcomes y_is are formed by consensus. In consequence, this model captures

[3] This notation scheme is called the *destination centric ordering* (see LeSage and Pace (2008)). We use this scheme since this is consistent with (i) a matrix-based notation (i.e., y_{ij} is the (i, j)-element of Y) and (ii) spatial/network econometric interpretations. According to spatial/network econometric literature, Y can be interpreted as a directed weighted network, and each y_{ij} denotes a signal from j to i. In the trade literature, Eaton and Kortum (2002) and Head and Mayer (2014) utilize this scheme.

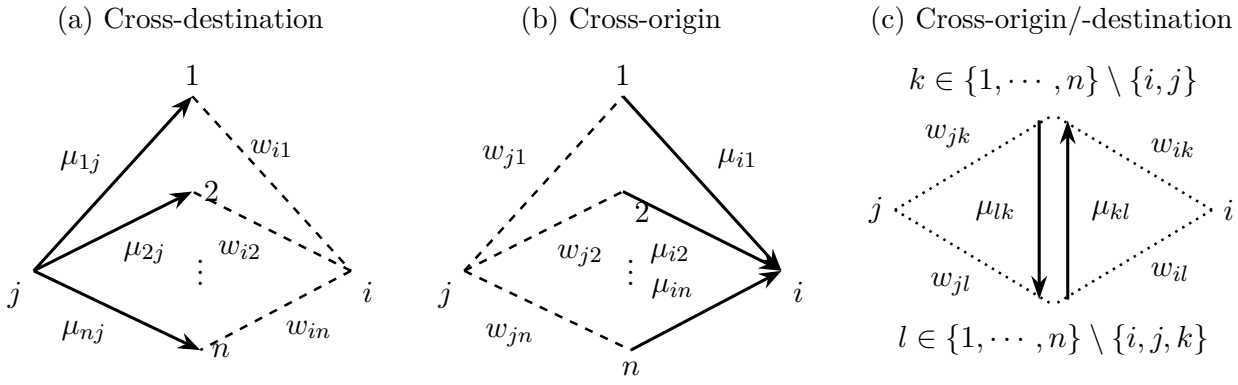
regional heterogeneities and spillovers stemming from their interconnectivities.

2.1 Theoretical foundation

This subsection provides microfoundations of our new model. Our goal is to specify μ_{ij}—the systematic component of y_{ij}—by endogenizing trade costs.

We now formalize the model in three stages. Building on ideas of Brancaccio et al. (2020), we model bilateral trade costs as functions of origin and destination conditions and their network-linked neighborhoods, summarized by a connectivity matrix and expected (equilibrium) flows.

Figure 2: Formation of key trade partners



Stage 1 (partner (hub) selection). Consider a destination i and an origin j with $i \neq j$. For each shipping opportunity ν, i forms a probability distribution over potential partners $k \in \{1, \cdots, n\} \setminus \{i\}$, and j forms a probability distribution over $l \in \{1, \cdots, n\} \setminus \{j\}$:

$$\Pr(k \text{ is } i\text{'s partner at } \nu) = w^d_{ik} \text{ and } \Pr(l \text{ is } j\text{'s partner at } \nu) = w^o_{jl}.$$

We collect these probabilities into row–stochastic matrices $W^d = (w^d_{ik})$ and $W^o = (w^o_{jl})$ with zero diagonals, i.e., $\sum_{k=1}^n w^d_{ik} = 1$, $w^d_{ii} = 0$ and $\sum_{l=1}^n w^o_{jl} = 1$, $w^o_{jj} = 0$. We assume draws are stationary and ergodic across ν; hence, by the law of large numbers, we have

$$\frac{1}{M}\sum_{\nu=1}^{M} \mathbb{I}\{k \text{ chosen by } i \text{ at } \nu\} \xrightarrow{a.s.} w^d_{ik}, \text{ and } \frac{1}{M}\sum_{\nu=1}^{M} \mathbb{I}\{l \text{ chosen by } j \text{ at } \nu\} \xrightarrow{a.s.} w^o_{jl}$$

as $M \to \infty$. Intuitively, w_{ik}^d measures the destination–side relevance of origin k for i, while w_{jl}^o captures the origin–side relevance of destination l for j. For empirical application, we adopt a single set of proximity weights and impose $W^d = W^o = W$; this aligns with the construction of W from historical flows in Section 4.

Assumption 2.2. We posit that W is constructed by row–normalizing a symmetric base matrix $\widetilde{W}$ (e.g., geographic/logistical affinity), $W = \mathrm{Diag}^{\mathrm{sum}}(\widetilde{W})^{-1}\widetilde{W}$, allowing W itself to be asymmetric after normalization.

This construction provides a parsimonious representation of network-mediated proximity in the spirit of Brancaccio et al. (2020).

Stage 2 (endogenous cost formation). Let τ_{ij} be a measure of a trade cost from j to i. Consistent with Brancaccio et al. (2020), we allow τ_{ij} to depend on network-linked trading activity around the origin and destination in addition to bilateral characteristics. We posit the following cost function that endogenizes trade costs.

Assumption 2.3. (i) For each ij, we assume

$$\tau_{ij}^{+} = D_{ij,1}^{\tilde{\beta}_1} \cdots D_{ij,K}^{\tilde{\beta}_K},$$

where $D_{ij,k}$ $(k = 1, \cdots, K)$ represents a bilateral characteristic affecting τ_{ij}. $\tilde{\beta}_1, \cdots, \tilde{\beta}_K$ are parameters. We assume that the baseline cost τ_{ij}^{+} satisfies the triangle inequality: for arbitrary three countries i, j, and k, $\tau_{ij}^{+} \leq \tau_{ik}^{+} \cdot \tau_{kj}^{+}$.

(ii) If i chooses $k \in \{1, \cdots, n\} \setminus \{i\}$ with probability w_{ik} and j selects $l \in \{1, \cdots, n\} \setminus \{j\}$ with probability w_{jl} as partners (hubs), the trade cost from j to i through k and l is

$$\widetilde{\tau}_{ij}\left(\boldsymbol{\mu}; k, l\right) = \mu_{kj}^{-\tilde{\lambda}_d} \mu_{il}^{-\tilde{\lambda}_o} \mu_{kl}^{-\tilde{\lambda}_w} \cdot \tau_{ij}^{+},$$

where $\tilde{\lambda}_d$, $\tilde{\lambda}_o$ and $\tilde{\lambda}_w$ are coefficients and $\boldsymbol{\mu} = (\mu_{11}, \mu_{21}, \cdots, \mu_{n1}, \cdots, \mu_{1n}, \mu_{2n}, \cdots, \mu_{nn})'$.

(iii) i's and j's partner choices are independent, so the probability of using the route k and l is $w_{ik} w_{jl}$.

(iv) Then, the overall trade cost from j to i is defined as

$$\tau_{ij}(\boldsymbol{\mu}) = \exp\left(\mathbb{E}_W\left[\ln(\widetilde{\tau}_{ij}\left(\boldsymbol{\mu}; k, l\right))\right]\right), \text{ where } \mathbb{E}_W(\cdot) = \sum_{k,l=1}^{n} w_{ik} w_{jl}(\cdot).$$

As a result, we have

$$\tau_{ij}(\boldsymbol{\mu}) = \tau^e_{ij}(\boldsymbol{\mu}) \cdot \tau^+_{ij}, \text{ where } \tau^e_{ij}(\boldsymbol{\mu}) = \left(\overline{\mu}^i_{.j}\right)^{-\tilde{\lambda}_d} \left(\overline{\mu}^j_{i.}\right)^{-\tilde{\lambda}_o} \left(\overline{\mu}^{ij}_{..}\right)^{-\tilde{\lambda}_w} \tag{2.3}$$

with $\overline{\mu}^i_{.j} = \prod_{k=1}^n \mu_{kj}^{w_{ik}}$; $\overline{\mu}^j_{i.} = \prod_{l=1}^n \mu_{il}^{w_{jl}}$; and $\overline{\mu}^{ij}_{..} = \prod_{k,l=1}^n \mu_{kl}^{w_{ik}w_{jl}}$.

(v) All stages—**Stage 1**, **Stage 2**, and **Stage 3**—operate under no informational frictions, meaning that the expected flows in **Stage 2** are the same as the equilibrium flows in **Stage 3**, $\boldsymbol{\mu}^* = (\mu^*_{11}, \mu^*_{21}, \cdots, \mu^*_{n1}, \cdots, \mu^*_{1n}, \mu^*_{2n}, \cdots, \mu^*_{nn})'$.

Under these settings, **Stage 2** forms $\tau_{ij}(\boldsymbol{\mu}^*)$ for $i, j = 1, \cdots, n$.

Assumption 2.2 illustrates properties of a resulting country connectivity matrix (W) from **Stage 1**. The symmetry of the source matrix ($\widetilde{W}$) is conventional and adopted for computational convenience (see Sec. 2.3). Assumption 2.3 describes how trade costs are endogenously shaped by information in trade networks. The three panels in Figure 2 correspond directly to the three network-based terms in equation (2.3). Unlike conventional models with exogenous τ_{ij}, here trade costs are themselves functions of the expected flows weighted by countries' connectivities. This provides a new channel of spillovers not captured in conventional frameworks such as Anderson and van Wincoop (2003).

The component $\tau^e_{ij}(\boldsymbol{\mu})$ within $\tau_{ij}(\boldsymbol{\mu})$ captures the network-based part of trade costs:

1. Outflows from j ($\overline{\mu}^i_{.j}$) summarize the common-origin with cross-destination linkages;
2. Inflows to i ($\overline{\mu}^j_{i.}$) summarize the common-destination with cross-origin linkages; and
3. Third-party flows ($\overline{\mu}^{ij}_{..}$), which do not share the same origin or destination, influence costs through cross-origin and cross-destination linkages.

Note that $w_{i1}, \ldots, w_{in}$ from **Stage 1** in $\overline{\mu}^i_{.j}$ capture cross-destination linkages; $w_{j1}, \ldots, w_{jn}$ in $\overline{\mu}^j_{i.}$ capture cross-origin linkages; and both sets of weights appear in $\overline{\mu}^{ij}_{..}$ to capture combined cross-destination and cross-origin linkages. $\tilde{\lambda}_d > 0$ implies network-based efficiency gains, reducing effective trade costs by the cross-destination linkages. On the other hand, $\tilde{\lambda}_d < 0$ means congestion effects from the cross-destination linkages if nearby routes or hubs operate at capacities. Analogously, $\tilde{\lambda}_o$ governs cross-origin linkages (e.g., cooperation vs. destination-side bottlenecks), and $\tilde{\lambda}_w$ captures third-party routing (hub-and-spoke economies vs. congestion on multi-leg paths).

Our specification nests the conventional iceberg cost specification since $\tau_{ij}(\boldsymbol{\mu}) = \tau^+_{ij}$ when $\tilde{\lambda}_d = \tilde{\lambda}_o = \tilde{\lambda}_w = 0$. In this case, cross-border arbitrage satisfies the triangle inequality, highlighting the effectiveness of geographic barriers (e.g., Eaton and Kortum (2002)). By

contrast, in our general case, the triangle inequality need not hold, implying that countries can achieve effective cost reductions based on the expected trade network.

Stage 3 (trade flow formation). Given τ_{ij} from **Stage 2**, a country i's consumer chooses $\{c_{i1}, \cdots, c_{in}\}$ by the following problem:

$$\max_{\{c_{ij}\}_{j=1}^n} U_i = \left(\sum_{j=1}^n \chi_j^{\frac{1}{\varrho}} \cdot c_{ij}^{\frac{\varrho-1}{\varrho}} \right)^{\frac{\varrho}{\varrho-1}} \text{ subject to } \sum_{j=1}^n \tau_{ij} c_{ij} = G_i, \tag{2.4}$$

where χ_j denotes a preference parameter, and G_i stands for the exogenously given country i's budget. Here, (2.4) is a standard CES aggregator with elasticity $\varrho > 1$.[4]

The following conditions characterize the uniqueness of the nominal value of the optimal trade flow $\mu_{ij}^* = \tau_{ij}(\boldsymbol{\mu}^*)c_{ij}^*$, where c_{ij}^* $(j = 1, \cdots, n)$ denotes the solution of (2.4). For this, we define $\mathbf{S} = I_N - \mathbf{A}$ (network SAR operator) where $\mathbf{A} = \lambda_d(I_n \otimes W) + \lambda_o(W \otimes I_n) + \lambda_w(W \otimes W)$ with $\lambda_d = (\varrho - 1)\tilde{\lambda}_d$, $\lambda_o = (\varrho - 1)\tilde{\lambda}_o$, $\lambda_w = (\varrho - 1)\tilde{\lambda}_w$. Here, $I_n \otimes W$, $W \otimes I_n$, and $W \otimes W$ characterize respectively the cross-destination, cross-origin, and cross-origin and destination linkages.

Assumption 2.4. (i) We assume

$$\max\{\lambda_d + \lambda_o + \lambda_w, \lambda_d\varphi_{\min} + \lambda_o + \lambda_w\varphi_{\min}, \lambda_d + \lambda_o\varphi_{\min} + \lambda_w\varphi_{\min}, \lambda_d\varphi_{\min} + \lambda_o\varphi_{\min} + \lambda_w\varphi_{\min}^2\} < 1, \tag{2.5}$$

where $\varphi_{\min}$ is the minimum eigenvalue of W. Then, $\mathbf{S}$ is invertible.

(ii) $\boldsymbol{\mu}^*$ satisfies the following condition:

$$\sup_{i,j} \sum_{k,l=1}^n \left| \sum_{p,q=1}^n s_{ij,pq} \left(\frac{\partial \left(\alpha_q(\boldsymbol{\mu}) + \eta_p(\boldsymbol{\mu}) \right)}{\partial \ln(\mu_{kl})} \right) \right| < 1,$$

where $s_{ij,kl}$ denotes the $((j-1)n + i, (l-1)n + k)$-element of $\mathbf{S}^{-1}$. Further,

$$\begin{aligned} \alpha_j(\boldsymbol{\mu}) &= -\frac{1}{2}\ln(G^W) + \ln(G_j) + \ln\left(\Pi_j^{\varrho-1}(\boldsymbol{\mu})\right) \text{ for } j = 1, \cdots, n \text{ and} \\ \eta_i(\boldsymbol{\mu}) &= -\frac{1}{2}\ln(G^W) + \ln(G_i) + \ln\left(P_i^{\varrho-1}(\boldsymbol{\mu})\right), \text{ for } i = 1, \cdots, n, \end{aligned} \tag{2.6}$$

[4]We adopt a demand-side focus to highlight how network-leveraged trade costs shape flows, taking the production side as exogenous. Microfounding network dependence on the production side (e.g., extending Eaton and Kortum (2002)) would require an alternative specification (e.g., Lind and Ramondo (2023)); we leave this for future research. See Appendix 1.3.3 for related discussion.

where $\Pi_j(\boldsymbol{\mu}) = \left(\sum_{i=1}^n \frac{G_r}{G^W}\left(\frac{\tau_{ij}(\boldsymbol{\mu})}{P_i(\boldsymbol{\mu})}\right)^{1-\varrho}\right)^{\frac{1}{1-\varrho}}$, $P_i(\boldsymbol{\mu}) = \left(\sum_{j=1}^n \frac{G_j}{G^W}\left(\frac{\tau_{ij}(\boldsymbol{\mu})}{\Pi_j(\boldsymbol{\mu})}\right)^{1-\varrho}\right)^{\frac{1}{1-\varrho}}$, and $G^W = \sum_{i=1}^n G_i$.

Our model's equilibrium can be characterized by a combination of two contraction mappings: $\mathsf{T}_1(\cdot)$ is the conventional mapping for the multilateral resistance terms, and $\mathsf{T}_2(\cdot)$ characterizes the network SAR operator $\mathbf{S}$. Under conditions (i) and (ii) in Assumption 2.4, there is a unique $\boldsymbol{\mu}^*$ satisfying

$$\mu_{ij}^* = \frac{G_i G_j}{G^W}\left(\frac{\tau_{ij}(\boldsymbol{\mu}^*)}{P_i(\boldsymbol{\mu}^*)\Pi_j(\boldsymbol{\mu}^*)}\right)^{1-\varrho} = \exp\left(\sum_{k,l=1}^n s_{ij,kl}\left(x_{kl}'\beta + \alpha_l(\boldsymbol{\mu}^*) + \eta_k(\boldsymbol{\mu}^*)\right)\right), \tag{2.7}$$

where $x_{ij} = (\ln(D_{ij,1}), \cdots, \ln(D_{ij,K}))'$ and $\beta = (\beta_1, \cdots, \beta_K)'$ with $\beta_k = (1-\varrho)\tilde{\beta}_k$. Under the no information friction setting, the expected flows coincide with the optimized flows, i.e., $\boldsymbol{\mu}^*$. Equation (2.7) is a semi-reduced form and establishes the main econometric equation by regarding $\alpha_j(\boldsymbol{\mu}^*)$ and $\eta_i(\boldsymbol{\mu}^*)$ as fixed-effect components. Note that equation (2.7) is not a full reduced form since the fixed-effect components ($\alpha_j(\boldsymbol{\mu})$ and $\eta_i(\boldsymbol{\mu})$) still rely on $\boldsymbol{\mu}$. See Appendix 1.3.2 for technical details.

Assumption 2.4 (i) ensures that $\mathsf{T}_2(\cdot)$ is a contraction mapping. It implies the well-definedness of $\{s_{ij,kl}\}$, and then $\ln(\mu_{ij}^*)$ is an aggregation of $x_{kl}'\beta + \alpha_l(\boldsymbol{\mu}^*) + \eta_k(\boldsymbol{\mu}^*)$ weighted by $s_{ij,kl}$. This condition is equivalent that $\rho_{\text{spec}}(\mathbf{A}) < 1$, where $\rho_{\text{spec}}(\mathbf{A})$ denotes the spectral radius of $\mathbf{A}$.[5] Note that the minimum eigenvalue of W ($\varphi_{\min}$) is a key network statistic showing bipartiteness (if $\varphi_{\min} \to -1$) and averaging rate (if $\varphi_{\min} \to 0$) (see Chung (1997); Bramoullé et al. (2014) for more details). This structure originates from the SAR framework (see LeSage and Pace (2008); Sec. C.3.2 of LeSage and Fischer (2010); Jeong et al. (2023); and Jeong and Lee (2024)).[6] However, while our formulation borrows from SAR models, its interpretation differs fundamentally: the expected trade flows here influence costs, rather than directly determining flows. From $\mathbf{S} = I_N - \mathbf{A}$, $\mathbf{S}^{-1}$ (network multiplier matrix) characterizes how network spillovers propagate across regions. Each element of $\mathbf{S}^{-1}$, $s_{ij,kl}$, represents the influence from pair kl to ij.[7] When $\lambda_d = \lambda_o = \lambda_w = 0$, (2.7) becomes the conventional gravity equation.

By Assumption 2.4 (ii), $\mathsf{T}_1(\cdot)$ is a contraction mapping leading to the unique forms of $\eta_i(\boldsymbol{\mu}^*)$ $(i = 1, \cdots, n)$ and $\alpha_j(\boldsymbol{\mu}^*)$ $(j = 1, \cdots, n)$. The main implication of Assumption 2.4 (ii) is that the cumulative network influence should not excessively perturb the multilateral

[5] $\rho_{\text{spec}}(\mathbf{A}) < 1$ is more general condition for invertibility of $\mathbf{S}$ when W has complex eigenvalues. Since we consider a row-normalized W from a symmetric relationship, all eigenvalues are real.

[6] A linear SAR model specifies the log-transformed equation based on (2.7) by replacing μ_{ij}^* by y_{ij}.

[7] The detailed structures of $s_{ij,kl}$ can be found in Section 1.3.4 of the supplement file.

resistance terms. Consequently, the resulting μ_{ij}^* is a unique function of the exogenous factors $\{x_{kl}\}_{k,l=1}^n$.

2.2 Econometric model specification

This subsection introduces our econometric model based on the established theoretical foundation. The true data-generating process (DGP) can be specified by

$$y_{ij} = \mu_{ij}^0 \times \xi_{ij}, \text{ where } \mu_{ij}^0 = \exp\left(\sum_{k,l=1}^{n} s_{ij,kl}\left(x_{kl}'\beta^0 + \alpha_l^0 + \eta_k^0\right)\right). \tag{2.8}$$

Here, $\mu_{ij}^0 = \mathbb{E}(y_{ij}|\mathbf{x})$; $\mathbf{x} = (x_{11}', \cdots, x_{n1}', \cdots, x_{1n}', \cdots, x_{nn}')'$ stands for a vector of exogenous characteristics; ξ_{ij} is a multiplicative CEF disturbance satisfying $\mathbb{E}(\xi_{ij}|\mathbf{x}) = 1$, $\lambda^0 = (\lambda_d^0, \lambda_o^0, \lambda_w^0)'$ denotes a vector of the true spatial influence parameters, $\beta^0 = (\beta_1^0, \cdots, \beta_K^0)'$ is the true parameter for x_{kl}, α_j^0 and η_i^0 represent the true origin- and destination- fixed effects, respectively. Importantly, the advantage of this specification is that a practitioner only needs $\mathbb{E}(\xi_{ij}|\mathbf{x}) = 1$ for estimation. For analytic simplicity, we can define the additive error $u_{ij} = \mu_{ij}^0(\xi_{ij} - 1)$ for each ij to have $y_{ij} = \mu_{ij}^0 + u_{ij}$ and $\mathbb{E}(u_{ij}|\mathbf{x}) = 0$.

Equation (2.8) addresses three essential issues. First, equation (2.8) correctly specifies $\mathbb{E}(y_{ij}|\mathbf{x})$. The logarithmic transformation approach, as LeSage and Pace (2008), encounters several significant issues. As Santos Silva and Tenreyro (2006) emphasize, applying a logarithmic transformation in constant elasticity models can lead to inconsistent estimates. This inconsistency arises because the transformation specifies $\mathbb{E}(\ln(y_{ij})|\mathbf{x})$ instead of $\mathbb{E}(y_{ij}|\mathbf{x})$, and due to Jensen's inequality, $\mathbb{E}(\ln(y_{ij})|\mathbf{x}) \neq \ln(\mathbb{E}(y_{ij}|\mathbf{x}))$. The gap, $\mathbb{E}(\ln(y_{ij})|\mathbf{x}) - \ln(\mathbb{E}(y_{ij}|\mathbf{x})) = \mathbb{E}(\ln(\xi_{ij})|\mathbf{x})$, characterizing the bias from the log-transformed model increases when (i) some y_{ij}s take huge positive values or (ii) many zero OD flows are contained in a sample. Detailed analysis can be found in Sec. 1 of the supplement file.[8]

The second issue arises from the specific feature of spatial econometric models. Let v_{ij} be the additive error term in the logarithmic transformed model (LeSage and Pace (2008)). To illustrate how a distributional assumption on $\{v_{ij}\}$ affects $\mathbb{E}(y_{ij}|\mathbf{x})$, consider two scenarios based on the distribution of v_{ij}. By assuming $y_{ij} > 0$ for all ij, LeSage and Pace's (2008) model can be rewritten as $y_{ij} = \exp\left(\sum_{k,l=1}^n s_{ij,kl}x_{kl}'\beta\right)\prod_{k,l=1}^n \exp(v_{kl})^{s_{ij,kl}}$.

[8] In addition to the issue $\mathbb{E}(\ln(\xi_{ij})|\mathbf{x}) \neq 0$, various dependence structures between ξ_{ij} and a component in $\mathbf{x}$ can also affect the magnitude of the bias from the log-transformed model. Since $\ln(\xi_{ij})$ can be expressed by Maclaurin series expansion, $\mathbb{E}(\ln(\xi_{ij})|\mathbf{x})$ is a function of the infinite-order of the moments, $h_p(\mathbf{x}) \equiv \mathbb{E}\left((\xi_{ij}^-)^p|\mathbf{x}\right)$ for $p = 2, 3, \cdots$ where $\xi_{ij}^- = \xi_{ij} - 1$. Then, a possible moment for estimating the log-transformed model $\mathbb{E}(x_{ij}\ln(\xi_{ij}))$ depends on the correlations between x_{ij} and $h_p(\mathbf{x})$ for $p = 2, 3, \cdots$.

1. If $v_{ij}|\mathbf{x} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ as LeSage and Pace (2008),

$$\mathbb{E}(y_{ij}|\mathbf{x}) = \exp\left(\sum_{k,l=1}^{n} s_{ij,kl} x'_{kl}\beta\right) \exp\left(\frac{\sigma^2}{2} \sum_{k,l=1}^{n} s^2_{ij,kl}\right)$$

since $\mathbb{E}(\exp(v_{kl})^{s_{ij,kl}}|\mathbf{x}) = \exp(\frac{\sigma^2 s^2_{ij,kl}}{2})$,

2. If $v_{ij}|\mathbf{x} \overset{i.i.d.}{\sim} \text{logGamma}(\theta_{\text{shape}}, \theta_{\text{rate}})$ with $\frac{\theta_{\text{shape}}}{\theta_{\text{rate}}} = 1$,

$$\mathbb{E}(y_{ij}|\mathbf{x}) = \exp\left(\sum_{k,l=1}^{n} s_{ij,kl} x'_{kl}\beta\right) \frac{\theta_{\text{shape}}^{\sum_{k,l=1}^{n} s_{ij,kl}} \prod_{k,l=1}^{n} \Gamma(\theta_{\text{shape}} + s_{ij,kl})}{\Gamma(\theta_{\text{shape}})^{n^2}}$$

since $\mathbb{E}(\exp(v_{kl})^{s_{ij,kl}}|\mathbf{x}) = \frac{\Gamma(\theta_{\text{shape}}+s_{ij,kl})}{\Gamma(\theta_{\text{shape}})\theta_{\text{rate}}^{s_{ij,kl}}} = \frac{\Gamma(\theta_{\text{shape}}+s_{ij,kl})\theta_{\text{shape}}^{s_{ij,kl}}}{\Gamma(\theta_{\text{shape}})}$.

Hence, $\mathbb{E}(y_{ij}|\mathbf{x})$ of the LeSage and Pace's (2008) model has different forms across distributional specifications on $\{v_{kl}\}$.

Third, ad hoc transformations of y_{ij} become necessary because a logarithmic transformed equation only accommodates strictly positive outcomes. Considering the undefined nature of $\ln(0)$, practitioners often use $\ln(y_{ij} + c)$ where $c > 0$, commonly setting $c = 1$. However, as Chen and Roth (2024) and Mullahy and Norton (2024) emphasize, this choice of c lacks a theoretical basis and leads to variations in the specification of $\mathbb{E}(\ln(y_{ij} + c)|\mathbf{x})$, which can result in biases and misleading interpretations of a covariate's marginal effect. We verify that the magnitude of the bias from employing $\ln(y_{ij} + c)$ becomes larger when the frequency of values close to zero grows.

2.3 Estimation

This subsection describes an estimation method for the parameters in (2.8), $\boldsymbol{\theta}^0 = \left(\theta^{0\prime}, \boldsymbol{\phi}^{0\prime}\right)'$, $\theta^0 = (\lambda^{0\prime}, \beta^{0\prime})'$, $\boldsymbol{\phi}^0 = (\boldsymbol{\alpha}^{0\prime}, \boldsymbol{\eta}^{0\prime})'$, where $\boldsymbol{\alpha}^0 = (\alpha_1^0, \cdots, \alpha_n^0)'$ and $\boldsymbol{\eta}^0 = (\eta_1^0, \cdots, \eta_n^0)'$. For possible values of the parameters, we denote $\boldsymbol{\theta} = (\theta', \boldsymbol{\phi}')'$, $\theta = (\lambda', \beta')'$ and $\boldsymbol{\phi} = (\boldsymbol{\alpha}', \boldsymbol{\eta}')'$, where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)'$ and $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)'$.

Assumption 2.5. Let Λ be the parameter space of λ. For each $\lambda \in \Lambda$, we define

$$\mathbf{A}(\lambda) = \lambda_d(I_n \otimes W) + \lambda_o(W \otimes I_n) + \lambda_w(W \otimes W) \text{ and } \mathbf{A} = \mathbf{A}(\lambda^0).$$

We assume $\sup_n \sup_{\lambda\in\Lambda} \|\mathbf{A}(\lambda)\|_\infty < 1$.

Assumption 2.5 guarantees the existence of the semi-reduced form at each $\lambda \in \Lambda$.[9]

The Poisson pseudo maximum likelihood (PPML) estimation method. We utilize the Poisson pseudo maximum likelihood (PPML) estimator developed by (Gourieroux et al., 1984; Santos Silva and Tenreyro, 2006). The log-likelihood function for $\boldsymbol{\theta}$ is:

$$\ell_N(\boldsymbol{\theta}) = \sum_{i,j=1}^{n} \ell_{ij}(\boldsymbol{\theta}) - \frac{1}{2}\left(\sum_{j=1}^{n}\alpha_j - \sum_{i=1}^{n}\eta_i\right)^2, \tag{2.9}$$

where $\ell_{ij}(\boldsymbol{\theta}) = -\mu_{ij}(\boldsymbol{\theta}) + y_{ij}\ln(\mu_{ij}(\boldsymbol{\theta})) - \ln(y_{ij}!)$ and the second term in (2.9) denotes a penalty term for the normalization (see Sec.2 in Fernandez-Val and Weidner (2016)).[10] This method focuses on correctly specifying $\mathbb{E}(y_{ij}|\mathbf{x})$ without transforming the model.[11] As the most notable advantage, hence, no additional assumptions are required on ξ_{ij} in estimation except $\mathbb{E}(\xi_{ij}|\mathbf{x}) = 1$. In conventional spatial econometric literature, the normal-based likelihood is employed. Since this conventional method uses the y's conditional mean as well as its variance structure, it is vulnerable to unknown heteroskedasticity (Lin and Lee, 2010). On the other hand, since this method does not specify a specific variance structure, we can not only achieve consistency of the PPMLE but it can also accommodate a robust standard error for heteroskedasticity and autocorrelation (see Sec. 3.2).

Then, the PPML estimator (PPMLE) can be obtained by

$$\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}, \widehat{\boldsymbol{\phi}}\right) = \underset{\theta\in\Theta_\theta, \boldsymbol{\phi}\in\mathbb{R}^{2n}}{\arg\max}\ \ell_N(\theta, \boldsymbol{\phi}),$$

where Θ_θ is a parameter space of θ. For further analysis, let $\widehat{\boldsymbol{\phi}}(\theta) = \arg\max_{\boldsymbol{\phi}\in\mathbb{R}^{2n}} \ell_N(\theta, \boldsymbol{\phi})$ for each $\theta \in \Theta_\theta$, and $\ell_N^c(\theta) = \ell_N(\theta, \widehat{\boldsymbol{\phi}}(\theta))$ denote the concentrated penalized log-likelihood function. For each element in $\widehat{\boldsymbol{\phi}}(\theta)$, we define $\widehat{\boldsymbol{\alpha}}(\theta) = (\widehat{\alpha}_1(\theta), \cdots, \widehat{\alpha}_n(\theta))'$ and $\widehat{\boldsymbol{\eta}}(\theta) = (\widehat{\eta}_1(\theta), \cdots, \widehat{\eta}_n(\theta))'$ for each $\theta \in \Theta_\theta$.

[9] This condition is slightly stronger than condition 2.5 in Assumption 2.4 (i). The purpose of this assumption is for asymptotic analysis.

[10] The key insight of (2.9) is that we can work with an unconstrained optimization problem. Indeed, (2.9) imposes a single linear constraint, $v'\boldsymbol{\phi} = 0$ where $v = (l_n', -l_n')'$, to eliminate the identification issue originated from the fixed effects' additive feature: $\alpha_j + \eta_i = \alpha_j^* + \eta_i^*$ where $\alpha_j^* = \alpha_j + c$ and $\eta_i^* = \eta_i - c$ for any c. Note that a normalization restriction for this issue is not unique: Fernandez-Val and Weidner (2016) additionally mention the possibility of $\alpha_1 = 0$, while Lee and Yu (2010) employs $\sum_{i=1}^n \eta_i = 0$.

[11] Another approach for estimating $\mathbb{E}(y_{ij}|\mathbf{x})$ without transforming the model is employing non-linear least squares (NLS). However, this method tends to give disproportionate weight to noisy observations, leading to inefficient estimations. This inefficiency arises because the method heavily depends on a relatively small number of observations (Silva and Tenreyro (2006, Sec. III A). For details, refer to Section 2.1 in the supplement.

Efficient computation of the network multiplier matrix $\mathbf{S}^{-1}$. Computation of our model is more complex than that of conventional gravity models, as it involves two contraction mappings. In estimation, however, inverting $\mathbf{S}$ is sufficient since it leads to the semi-reduced-form (2.7). In detail, we encounter the evaluation of $\mathbf{S}^{-1}(\lambda)\mathbf{Z}(\boldsymbol{\theta})$ for each $\boldsymbol{\theta}$, where $\mathbf{Z}(\boldsymbol{\theta}) = \mathbf{X}\beta + \boldsymbol{\alpha} \otimes l_n + l_n \otimes \boldsymbol{\eta}$ with $\mathbf{Z} = \mathbf{Z}(\boldsymbol{\theta}^0)$. This term must be carried out repeatedly during likelihood maximization. As n increases, the computation cost grows very rapidly with n. By exploiting the structure of the three networks $I_n \otimes W$, $W \otimes I_n$, and $W \otimes W$ generated by a single row-normalized connectivity matrix W, we can design a much more efficient computation procedure.

By Assumption 2.2 and the spectral decomposition theorem, $W = QDQ^{-1}$, where $D = \text{diag}(\varphi_1, \cdots, \varphi_n)$ and Q denotes the eigenvector matrix. Because W is constructed from symmetric relationships, all φ_i are real. Since $I_n \otimes W$, $W \otimes I_n$, and $W \otimes W$ share the same eigenvector basis, an eigenvalue of $\mathbf{A}(\lambda)$ can be represented by

$$\mathbf{A}(\lambda)(q_i \otimes q_j) = (\lambda_d \varphi_j + \lambda_o \varphi_i + \lambda_w \varphi_i \varphi_j)(q_i \otimes q_j) \text{ for } i, j = 1, \cdots, n, \tag{2.10}$$

where q_i denotes the ith column vector of Q.[12]

Define $Z^{\text{mat}}(\boldsymbol{\theta})$ satisfying $\mathbf{Z}(\boldsymbol{\theta}) = \text{vec}\,(Z^{\text{mat}}(\boldsymbol{\theta}))$. Then, we want to obtain the matrix fixed-point $T(\boldsymbol{\theta})$ satisfying $Z^{\text{mat}}(\boldsymbol{\theta}) = T(\boldsymbol{\theta}) - \lambda_d W T(\boldsymbol{\theta}) - \lambda_o T(\boldsymbol{\theta}) W' - \lambda_w W T(\boldsymbol{\theta}) W'$ ($\Leftrightarrow$ $\text{vec}\,(T(\boldsymbol{\theta})) = \mathbf{S}^{-1}(\lambda)\mathbf{Z}(\boldsymbol{\theta})$). Note that Q and D are invariant in the estimation procedure. Hence, we have

$$\widetilde{Z}^{\text{mat}}(\boldsymbol{\theta}) = \widetilde{T}(\boldsymbol{\theta}) - \lambda_d D \widetilde{T}(\boldsymbol{\theta}) - \lambda_o \widetilde{T}(\boldsymbol{\theta}) D - \lambda_w D \widetilde{T}(\boldsymbol{\theta}) D,$$

where $\widetilde{Z}^{\text{mat}}(\boldsymbol{\theta}) = Q^{-1} Z^{\text{mat}}(\boldsymbol{\theta}) Q^{-1\prime}$ and $\widetilde{T}(\boldsymbol{\theta}) = Q^{-1} T(\boldsymbol{\theta}) Q^{-1\prime}$. It implies

$$(\widetilde{T}(\boldsymbol{\theta}))_{ij} = \frac{(\widetilde{Z}^{\text{mat}}(\boldsymbol{\theta}))_{ij}}{1 - \lambda_d \varphi_i - \lambda_o \varphi_j - \lambda_w \varphi_i \varphi_j} \text{ for } i, j = 1, \cdots, n, \tag{2.11}$$

by (2.10). Then, we can easily recover $T(\boldsymbol{\theta}) = Q \widetilde{T}(\boldsymbol{\theta}) Q'$.

Identification condition. Next, we study the identification of $\boldsymbol{\theta}^0$. We define the following notations for this purpose:

[12] In terms of model structure, we can allow different proximity matrices for cross-origin and cross-destination linkages (W and M) such as Jeong and Lee (2024). In this case, however, it might not be possible to dramatically reduce computation costs since two connectivity matrices do not share the eigenvector basis in general. We leave this issue for future research.

Let $\mathbf{H}^{\theta\theta}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{H}^{\lambda\lambda}(\boldsymbol{\theta}) & \mathbf{H}^{\beta\lambda\prime}(\boldsymbol{\theta}) \\ \mathbf{H}^{\beta\lambda}(\boldsymbol{\theta}) & \mathbf{H}^{\beta\beta}(\boldsymbol{\theta}) \end{bmatrix}$ with $\mathbf{H}^{\lambda\lambda}(\boldsymbol{\theta}) = \left(h_{ab}^{\lambda\lambda}(\boldsymbol{\theta})\right)$ $(a, b \in \{d, o, w\})$, $\mathbf{H}^{\beta\lambda}(\boldsymbol{\theta}) = \left[h_d^{\beta\lambda}(\boldsymbol{\theta}) \quad h_o^{\beta\lambda}(\boldsymbol{\theta}) \quad h_w^{\beta\lambda}(\boldsymbol{\theta})\right]$, $\mathbf{H}^{\beta\beta}(\boldsymbol{\theta}) = \mathbf{0}_{K\times K}$, $\mathbf{H}^{\phi\theta}(\boldsymbol{\theta}) = \left[h_d^{\phi\theta}(\boldsymbol{\theta}) \quad h_o^{\phi\theta}(\boldsymbol{\theta}) \quad h_w^{\phi\theta}(\boldsymbol{\theta}) \quad \mathbf{0}_{2n\times K}\right]$, and $\mathbf{H}^{\phi\phi} = -\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes l_n l_n'$. Here,

- $h_{ab}^{\lambda\lambda}(\boldsymbol{\theta}) = \left(2\mathbf{W}_a\mathbf{W}_b\mathbf{S}^{-3}(\lambda)\mathbf{Z}(\boldsymbol{\theta})\right)'\mathbf{u}(\boldsymbol{\theta})$ for $a, b \in \{d, o, w\}$, $h_a^{\beta\lambda}(\boldsymbol{\theta}) = \left(\mathbf{W}_a\mathbf{S}^{-2}(\lambda)\mathbf{X}\right)'\mathbf{u}(\boldsymbol{\theta})$ and $h_a^{\phi\theta}(\boldsymbol{\theta}) = \left(\mathbf{W}_a\mathbf{S}^{-2}(\lambda)\mathbf{D}\right)'\mathbf{u}(\boldsymbol{\theta})$ for $a \in \{d, o, w\}$, where $\mathbf{W}_d = I_n \otimes W$, $\mathbf{W}_o = W \otimes I_n$, and $\mathbf{W}_w = W \otimes W$;
- $\mathbf{X}$ denotes an $N \times K$ matrix which has $x_{ij,k}$ as the $((j-1)n+i, k)$- element of $\mathbf{X}$; $\mathbf{D} = [\mathbf{I}_n \otimes l_n, l_n \otimes \mathbf{I}_n]$ is an $N \times 2n$ matrix for dummy variables;
- $\boldsymbol{\mu}(\boldsymbol{\theta}) = \left(\exp\left(\tilde{\mu}_{11}(\boldsymbol{\theta})\right), \cdots, \exp\left(\tilde{\mu}_{n1}(\boldsymbol{\theta})\right), \cdots, \exp\left(\tilde{\mu}_{1n}(\boldsymbol{\theta})\right), \cdots, \exp\left(\tilde{\mu}_{nn}(\boldsymbol{\theta})\right)\right)$ with $\boldsymbol{\mu}^0 = \boldsymbol{\mu}(\boldsymbol{\theta}^0)$ and $\tilde{\mu}_{ij}(\boldsymbol{\theta}) = \sum_{k,l=1}^n s_{ij,kl}(\lambda)\left(x_{kl}'\beta + \alpha_l + \eta_k\right)$;
- $\mathbf{u}(\boldsymbol{\theta}) = \mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta})$ with $\mathbf{u} = \mathbf{y} - \boldsymbol{\mu}^0$.

Assumption 2.6 (Identification)**.** Let $\boldsymbol{\Theta} = \Theta_\theta \times \boldsymbol{\Phi}$ be the parameter space of $\boldsymbol{\theta}$, where Θ_θ denotes a compact parameter space of θ and $\boldsymbol{\Phi}$ represents a parameter space of $\boldsymbol{\phi}$. Here, $\boldsymbol{\Phi} \subset [-C, C]^{2n}$ for some finite constant $C > 0$.

(i) For each $(\theta, \boldsymbol{\phi}) \in \boldsymbol{\Theta}$, define $\mathbf{J}_N^{\phi\phi}(\boldsymbol{\theta}) = \frac{1}{N}\Big(\mathbf{D}'\mathbf{S}^{-1\prime}(\lambda)\,\mathrm{Diag}\left(\boldsymbol{\mu}(\boldsymbol{\theta})\right)\mathbf{S}^{-1}(\lambda)\mathbf{D} - \mathbf{H}^{\phi\phi}\Big)$. Assume $\liminf_{n\to\infty}\ \inf_{\theta\in\Theta_\theta}\inf_{\phi\in\boldsymbol{\Phi}}\varphi_{\min}\Big(\mathbf{J}_N^{\phi\phi}(\theta, \boldsymbol{\phi})\Big) > 0$. Then, for each $\theta \in \Theta_\theta$ and for n sufficiently large, $\widehat{\boldsymbol{\phi}}(\theta) = \arg\max_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \ell_N(\theta, \boldsymbol{\phi})$ is unique.

(ii) For each $(\theta, \boldsymbol{\phi}) \in \boldsymbol{\Theta}$, define

$$\mathbf{J}_N^{\theta\theta}(\boldsymbol{\theta}) = \frac{1}{N}\Big(\mathbf{G}(\boldsymbol{\theta})'\mathbf{S}^{-1\prime}(\lambda)\,\mathrm{Diag}\left(\boldsymbol{\mu}(\boldsymbol{\theta})\right)\mathbf{S}^{-1}(\lambda)\mathbf{G}(\boldsymbol{\theta}) - \mathbf{H}^{\theta\theta}(\boldsymbol{\theta})\Big),$$
$$\mathbf{J}_N^{\theta\phi}(\boldsymbol{\theta}) = \frac{1}{N}\Big(\mathbf{G}(\boldsymbol{\theta})'\mathbf{S}^{-1\prime}(\lambda)\,\mathrm{Diag}\left(\boldsymbol{\mu}(\boldsymbol{\theta})\right)\mathbf{S}^{-1}(\lambda)\mathbf{D} - \mathbf{H}^{\phi\theta}(\boldsymbol{\theta})'\Big), \text{ and } \mathbf{J}_N^{\phi\theta}(\boldsymbol{\theta}) = \left(\mathbf{J}_N^{\theta\phi}(\boldsymbol{\theta})\right)'.$$

Here, $\mathbf{G}(\boldsymbol{\theta}) = \left[\mathbf{W}_d\mathbf{S}^{-1}(\lambda)\mathbf{Z}(\boldsymbol{\theta}), \mathbf{W}_o\mathbf{S}^{-1}(\lambda)\mathbf{Z}(\boldsymbol{\theta}), \mathbf{W}_w\mathbf{S}^{-1}(\lambda)\mathbf{Z}(\boldsymbol{\theta}), \mathbf{X}\right]$. For each $\theta \in \Theta_\theta$, let

$$\widehat{\mathbf{J}}_N^{\theta\theta}(\theta) = \mathbf{J}_N^{\theta\theta}(\theta, \widehat{\boldsymbol{\phi}}(\theta)), \quad \widehat{\mathbf{J}}_N^{\theta\phi}(\theta) = \mathbf{J}_N^{\theta\phi}(\theta, \widehat{\boldsymbol{\phi}}(\theta)), \quad \widehat{\mathbf{J}}_N^{\phi\phi}(\theta) = \mathbf{J}_N^{\phi\phi}(\theta, \widehat{\boldsymbol{\phi}}(\theta)).$$

Assume $\liminf_{n\to\infty}\ \inf_{\theta\in\Theta_\theta}\varphi_{\min}\Big(\widehat{\mathbf{H}}(\theta)\Big) > 0$, where $\widehat{\mathbf{H}}(\theta) = \widehat{\mathbf{J}}_N^{\theta\theta}(\theta) - \widehat{\mathbf{J}}_N^{\theta\phi}(\theta)\Big[\widehat{\mathbf{J}}_N^{\phi\phi}(\theta)\Big]^{-1}\widehat{\mathbf{J}}_N^{\phi\theta}(\theta)$.

Then, for n sufficiently large, $\widehat{\theta} = \arg\max_{\theta\in\Theta_\theta}\ell_N^c(\theta)$ is unique.

Assumption 2.6 (i) guarantees that, for each $\theta \in \Theta_\theta$, the population criterion $\ell_\infty(\theta, \boldsymbol{\phi})$ is strictly concave in $\boldsymbol{\phi}$, so that the maximizer $\boldsymbol{\phi}(\theta)$ is unique. Assumption 2.6 (ii) ensures

that the profiled criterion $\ell_\infty^c(\theta) = \ell_\infty(\theta, \boldsymbol{\phi}(\theta))$ is strictly concave in θ, which implies that the θ^0 is the unique maximizer of $\ell_\infty^c(\theta)$. Consequently, $(\theta^0, \boldsymbol{\phi}^0)$ is point-identified in large samples, where $\boldsymbol{\phi}^0 = \boldsymbol{\phi}(\theta^0)$. The matrices $\mathbf{H}^{\theta\theta}(\boldsymbol{\theta})$, $\mathbf{H}^{\phi\theta}(\boldsymbol{\theta})$, and $\mathbf{H}^{\phi\phi}(\boldsymbol{\theta})$ are additional components in the Hessian that arise from the model's network interactions. Unlike the traditional gravity equation, the presence of spatial spillovers makes the curvature of the objective function more complex. Intuitively, when $\mathbf{S}^{-1}(\lambda)$ is well defined, the leading part of $\widehat{\mathbf{H}}(\theta)$ is likely positive definite; if the perturbation induced by the $\mathbf{H}$-terms is not too large, the conditions in Assumption 2.6 are satisfied. Motivated by this identification argument, we recommend first estimating the conventional gravity equation to obtain preliminary estimates $\widetilde{\beta}$, $\widetilde{\boldsymbol{\alpha}}$, and $\widetilde{\boldsymbol{\eta}}$. One can then set the initial parameter vector for maximizing (2.9) as $\widehat{\boldsymbol{\theta}}^{(0)} = \left(0, 0, 0, \widetilde{\beta}', \widetilde{\boldsymbol{\alpha}}', \widetilde{\boldsymbol{\eta}}'\right)'$, since $\mathbf{S}^{-1}(\lambda) = I_N$ when $\lambda = \mathbf{0}$. Details are provided in Lemmas 2.4 and 2.5 of Sec. 2.3 of the supplement file.

3 Statistical Analysis

This section derives the asymptotic distribution of the PPMLE, presents relevant statistical inference, and presents simulation results for finite samples.

3.1 Asymptotic distribution of the PPMLE

To derive the asymptotic distribution of $\widehat{\theta}$ and $\widehat{\boldsymbol{\phi}}$, we impose the following regularity assumption for Theorems 3.1 and 3.2. Details can be found in Sec. 2 of the supplement file.

Assumption 3.1. (i) $\{x_{ij}\}$, $\{\eta_i^0\}$, and $\{\alpha_j^0\}$ are random fields satisfying $\max_k \sup_{i,j,n} |x_{ij,k}| < C$, $\sup_{i,n} |\eta_i^0| < C$, and $\sup_{j,n} |\alpha_j^0| < C$, where $C > 0$ denotes a generic finite constant.

(ii) $\{\xi_{ij}\}$ is a random field satisfying $\sup_{i,j,n} \mathbb{E}|\xi_{ij}|^{2+c} < C$ for some $c > 0$.

(iii) $\mathbb{E}(\xi_{ij}|\mathbf{x}) = 1$ for all $i, j = 1, \cdots, n$.

Then, the same type of properties follow for the additive error u_{ij}: $\mathbb{E}(u_{ij}|\mathbf{x}) = 0$ and $\sup_{i,j,n} \mathbb{E}|u_{ij}|^{2+c} < C$. The theorems below state the asymptotic properties of the PPMLE. The asymptotic properties of the PPMLE for the fixed-effect parameters are used to examine the structure of the multilateral resistance terms. Details are in Sec. 2.3 of the supplement.

Theorem 3.1. Suppose that Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 3.1, and 3.2 hold. Let

$$\boldsymbol{\Sigma}_{\theta,N} = \frac{1}{N}\mathbf{G}'\mathbf{S}^{-1\prime}\mathrm{Diag}(\boldsymbol{\mu})\mathbf{M}_{\mathbf{D}}\mathbf{S}^{-1}\mathbf{G}, \text{ and } \boldsymbol{\Omega}_{\theta,N} = \frac{1}{N}\mathbf{G}'\mathbf{S}^{-1\prime}\mathbf{M}_{\mathbf{D}}'\mathbb{E}\left(\mathbf{u}\mathbf{u}'\right)\mathbf{M}_{\mathbf{D}}\mathbf{S}^{-1}\mathbf{G},$$

where $\mathbf{M_D} = I_N - \mathbf{P_D}\text{Diag}(\boldsymbol{\mu})$ with $\mathbf{P_D} = \mathbf{S}^{-1}\mathbf{D}(\widetilde{\mathbf{D}'\mathbf{D}})^{-1}\mathbf{D}'\mathbf{S}^{-1\prime}$ and $\widetilde{\mathbf{D}'\mathbf{D}} = \mathbf{D}'\mathbf{S}^{-1\prime}\,\text{Diag}(\boldsymbol{\mu})\mathbf{S}^{-1}\mathbf{D} - \mathbf{H}^{\phi\phi} = N\mathbf{J}_N^{\phi\phi}(\boldsymbol{\theta}^0)$.

Suppose (i) $\liminf_{n\to\infty}\varphi_{\min}(\boldsymbol{\Omega}_{\theta,N}) > 0$ and (ii) $\boldsymbol{\Sigma}_\theta = \text{plim}_{n\to\infty}\boldsymbol{\Sigma}_{\theta,N}$ exists and is positive definite. Then, we have

$$\sqrt{N}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\Omega}_\theta\boldsymbol{\Sigma}_\theta^{-1}\right) \text{ as } n \to \infty, \tag{3.1}$$

where $\boldsymbol{\Omega}_\theta = \text{plim}_{n\to\infty}\boldsymbol{\Omega}_{\theta,N}$.

Theorem 3.2. Suppose that Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 3.1 and 3.2 hold. Let

$$\mathbf{V}_{\phi,N} = n\left(\widetilde{\mathbf{D}'\mathbf{D}}\right)^{-1}\mathbf{D}'\mathbf{S}^{-1\prime}\mathbf{M}_{\boldsymbol{\phi}}'\mathbb{E}(\mathbf{u}\mathbf{u}')\mathbf{M}_{\boldsymbol{\phi}}\mathbf{S}^{-1}\mathbf{D}\left(\widetilde{\mathbf{D}'\mathbf{D}}\right)^{-1},$$

where $\mathbf{M}_{\boldsymbol{\phi}} = I_N - \mathbf{M_D}\mathbf{S}^{-1}\mathbf{G}\left(\mathbf{G}'\mathbf{S}^{-1\prime}\text{Diag}(\boldsymbol{\mu})\mathbf{M_D}\mathbf{S}^{-1}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{S}^{-1\prime}\text{Diag}(\boldsymbol{\mu})$. Then,

$$\begin{aligned}\sqrt{n}\left(\widehat{\alpha}_j - \alpha_j^0\right) &\xrightarrow{d} \mathcal{N}(0, \lim_{n\to\infty} e_{2n,j}'\mathbf{V}_{\boldsymbol{\phi},N}e_{2n,j}), \text{ and}\\ \sqrt{n}\left(\widehat{\eta}_i - \eta_i^0\right) &\xrightarrow{d} \mathcal{N}(0, \lim_{n\to\infty} e_{2n,n+i}'\mathbf{V}_{\boldsymbol{\phi},N}e_{2n,n+i})\end{aligned}$$

as $n \to \infty$, where $e_{2n,j}$ denotes the $2n$-dimensional unit vector with its j-th element equal to 1 and all other elements equal to 0.

By Assumption 2.6(i), we obtain both a well-defined $\widehat{\boldsymbol{\phi}}(\theta)$ and good conditioning of $\widetilde{\mathbf{D}'\mathbf{D}}$, which in turn ensures that $\mathbf{P_D}$ and $\mathbf{M_D}$ are well-defined. Assumption 2.6 as a whole guarantees identification of θ^0 and, in particular, implies that $\text{plim}_{n\to\infty}\widehat{\mathbf{H}}(\theta^0) = \boldsymbol{\Sigma}_\theta$, i.e., $\widehat{\mathbf{H}}(\theta^0)$ is asymptotically equivalent to $\boldsymbol{\Sigma}_{\theta,N}$. Details are provided in Lemmas 2.4 and 2.5 of Sec. 2.3 of the supplement file.

3.2 Variance estimation

This subsection provides a method for spatial HAC estimation of the covariance matrices in Theorems 3.1 and 3.2 of the PPMLE. Because errors in bilateral flows may exhibit network dependence across origin–destination pairs like the network dependence embedded in the conditional mean, we conduct inference using spatial HAC covariance estimation without imposing a parametric dependence structure. Our suggested method extends the existing literature on cross-sectional data to OD flows (Kelejian and Prucha (2007) and Kim and Sun (2011)), spatial extensions of the time-series literature (Newey and West (1987); Andrews (1991); de Jong and Davidson (2000)). Since an estimate of $\boldsymbol{\Sigma}_\theta$ is obtained by plugging $\widehat{\boldsymbol{\theta}}$ in, we focus on estimating $\boldsymbol{\Omega}_\theta$ here. Also, note that estimating $\lim_{n\to\infty} e_{2n,j}'\mathbf{V}_{\boldsymbol{\phi},N}e_{2n,j}$ and

$\lim_{n\to\infty} e'_{2n,n+i}\mathbf{V}_{\phi,N}e_{2n,n+i}$ for $i,j=1,\cdots,n$ is applicable by the same way of estimating $\mathbf{\Omega}_\theta$.

First, we provide the regularity assumptions.

Assumption 3.2. (i) For the structure of $\mathbf{u}=(u_{11},\cdots,u_{n1},\cdots,u_{1n},\cdots,u_{nn})'$, we assume

$$\mathbf{u}=\mathbf{BH}\boldsymbol{\epsilon}, \tag{3.2}$$

where $\mathbf{B}$ denotes some $N\times N$ matrix, $\mathbf{H}=\text{diag}\,(\sigma^*_{11},\cdots,\sigma^*_{n1},\cdots,\sigma^*_{1n},\cdots,\sigma^*_{nn})$, and $\boldsymbol{\epsilon}=(\epsilon_{11},\cdots,\epsilon_{n1},\cdots,\epsilon_{1n},\cdots,\epsilon_{nn})'$ is an $N\times 1$ vector of innovations.

(ii) $\epsilon_{ij}\overset{i.i.d.}{\sim}(0,1)$ across ij with $\sup_{n,i,j}\mathbb{E}|\epsilon_{ij}|^4<\infty$.

(iii) $0<\inf_{i,j,n}\sigma^*_{ij}\leq\sup_{i,j,n}\sigma^*_{ij}<\infty$.

(iv) $\mathbf{B}$ is nonsingular and $\sup_n\max\{\|\mathbf{B}\|_\infty,\|\mathbf{B}\|_1\}<\infty$.

Assumption 3.2 (i) describes the basic covariance structure. Hence, we have

$$\mathbf{\Omega}_{\theta,N}=\frac{1}{N}\mathbf{R}'\mathbf{R}=\frac{1}{N}\sum_{i,j,k,l=1}^{n}R_{ij}R'_{kl}=\frac{1}{N}\sum_{i,j,k,l=1}^{n}\mathbb{E}\left(\left(\mathbf{G}'\mathbf{S}^{-1\prime}\mathbf{M}'_{\mathbf{D}}\mathbf{u}\right)_{.,ij}\left(\mathbf{u}'\mathbf{M}_{\mathbf{D}}\mathbf{S}^{-1}\mathbf{G}\right)_{kl,.}\right),$$

where $\mathbf{R}=\mathbf{HB}'\mathbf{M}_{\mathbf{D}}\mathbf{S}^{-1}\mathbf{G}$, R_{ij} denotes the $((j-1)n+i)$-th column of $\mathbf{R}$, $\left(\mathbf{G}'\mathbf{S}^{-1\prime}\mathbf{M}'_{\mathbf{D}}\mathbf{u}\right)_{.,ij}$ denotes the $(j-1)n+i$-th column of $\mathbf{G}'\mathbf{S}^{-1\prime}\mathbf{M}'_{\mathbf{D}}\mathbf{u}$ and $\left(\mathbf{u}'\mathbf{M}_{\mathbf{D}}\mathbf{S}^{-1}\mathbf{G}\right)_{kl,.}$ is the $(l-1)n+k$-th row of $\mathbf{u}'\mathbf{M}_{\mathbf{D}}\mathbf{S}^{-1}\mathbf{G}$. Sec. 3.3 provides an example of (3.2). Conditions (ii) and (iii) in Assumption 3.2 are conventional. Condition (iv) comes from Kelejian and Prucha (2007).[13]

Based on this setting, we can characterize the spatial HAC estimator

$$\widehat{\mathbf{\Omega}}_{\theta,N}=\frac{1}{N}\sum_{i,j,k,l=1}^{n}\left(\widehat{\mathbf{G}}'\widehat{\mathbf{S}}^{-1\prime}\widehat{\mathbf{M}}'_{\mathbf{D}}\hat{\mathbf{u}}\right)_{.,ij}\left(\hat{\mathbf{u}}'\widehat{\mathbf{M}}_{\mathbf{D}}\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{G}}\right)_{kl,.}\mathsf{K}\left(\frac{d^*_{ij,kl}}{d_N}\right),$$

where $\widehat{\cdot}$ is the plug-in estimator from $\widehat{\boldsymbol{\theta}}$, $\mathsf{K}(\cdot)$ denotes a real-valued kernel function (e.g., Bartlett, Parzen, and Tukey-Hanning), $d^*_{ij,kl}$ is a distance measure between two pairs ij and kl, and d_N is a bandwidth. For details, refer to Sec. 2.4 in the supplement file.

3.3 Monte Carlo Simulations

Performance of the spectral algorithm for $\mathbf{S}^{-1}\mathbf{Z}(\boldsymbol{\theta})$. First, we study the performance of our suggested algorithm for evaluating $\mathbf{S}^{-1}\mathbf{Z}(\boldsymbol{\theta})$ introduced in Section 2.3. Table 1 reports how the computation time of the two methods varies with n (in seconds) for completing the

[13]Still, it can be relaxed by the ideas of Pesaran and Yang (2020, 2021). For example, if the error structure follows (3.3) in the simulation section, we can allow a finite number of moderate dominant units. We will leave this issue for future research.

maximization of the log-likelihood function (2.9). Method A maximizes the log-likelihood based on the direct computation of $\mathbf{S}^{-1}(\lambda)$, while Method B utilizes our algorithm. As n increases from 9 to 64, the running time of Method A rises dramatically from 0.57 seconds to about 9,466 seconds (approximately 2.6 hours), whereas Method B increases only modestly from 0.35 seconds to 9.28 seconds. Consequently, the ratio A/B grows sharply from 1.61 to about 1,020, indicating that Method B is only slightly faster for very small problems but becomes more than three orders of magnitude faster for $n = 64$.

Table 1: Computation time comparison

n	$N = n^2$	Method A	Method B	A/B
9	81	0.5668	0.3518	1.6112
25	625	12.3862	0.3967	31.2265
49	2401	1252.0077	3.9725	315.1687
64	4096	9465.7145	9.2816	1019.8380

Note: Each number in this table represents seconds for finishing maximization of the log-likelihood (2.9). For the maximization, we use `fminunc`. Method A is maximizing the log-likelihood function based on the calculation of $\mathbf{S}^{-1}$ and multiplication by $\mathbf{Z}(\boldsymbol{\theta})$. Method B is maximizing the log-likelihood function based on (2.11). In the last column, we provide the ratios of the computation times for the two methods (in seconds). All computations were performed in MATLAB R2023b on a Windows 11 machine equipped with a 13th Gen Intel Core i5-1340P CPU (12 cores, 1.9 GHz) and 16 GB of RAM.

Data generating process. We consider $n = 49$ regions. Following Pesaran and Yang (2020, 2021), the regional proximity matrix W is constructed to feature two "dominant" regions (units 1 and 2). For regions $i \in \{1, 2\}$, the code draws relatively large link weights $\sim \mathcal{U}[0.8, 1]$ to n^{δ_i} neighbors (with $\delta_1 = 0.25$, $\delta_2 = 0.1$), then spreads the remaining mass across two randomly chosen non-hub neighbors so that each row approximately sums to one. Self-links are zero. Let the total world GDP be $G^W = 1400$. Country i's GDP G_i is proportional to its out-degree in W: $G_i = G^W \cdot s_i$ and $s_i = \frac{\sum_j (w_{ij}+0.01)}{\sum_{i,k=1}^{n}(w_{ik}+0.01)}$. We impose balanced trade targets so that exports by origin and imports by destination both equal G. Preferences follow an Armington structure with elasticity of substitution $\varrho = 5$ (Anderson and van Wincoop, 2003). First-stage parameters $\tilde{\lambda} = \left(\tilde{\lambda}_d, \tilde{\lambda}_o, \tilde{\lambda}_w\right)' = (0.05, 0.05, 0.025)'$ and $\tilde{\beta} = (-0.15, -0.05)'$ are mapped to reduced-form coefficients $\lambda^0 = (\varrho - 1)\tilde{\lambda} = (0.2, 0.2, 0.1)'$ and $\beta^0 = (1-\varrho)\tilde{\beta} = (0.6, 0.2)'$. Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2] \in \mathbb{R}^{N\times 2}$ be bilateral covariates drawn from $\mathcal{U}[0, 0.75]$, and $X^U = (x_1^u, \cdots, x_n^u)' \in \mathbb{R}^{n\times 1}$ a standardized country characteristic. Second-stage loadings are $\tilde{\gamma}_o = \tilde{\gamma}_d = 0.01$ with the same scaling $\gamma = (1 - \varrho)\tilde{\gamma} = -0.04$.

Starting with this setting, the initial destination and origin fixed effects are specified by

$$\boldsymbol{\alpha}^{(0)} = X^U\gamma_o, \text{ and } \boldsymbol{\eta}^{(0)} = X^U\gamma_d.$$

Conditional on $(\boldsymbol{\alpha}^{(0)}, \boldsymbol{\eta}^{(0)})$, $\boldsymbol{\mu}^{(0)}$. Given $\boldsymbol{\mu}^{(0)}$ is determined. Given $\boldsymbol{\mu}^{(0)}$ and initial $\left(P^{(0)}, \Pi^{(0)}\right) = (\mathbf{1}_n, \mathbf{1}_n)$, we compute multilateral resistances via contraction mappings by (2.6):

$$\boldsymbol{\alpha}^{(\ell)} = c_0\mathbf{1}_n + X^U\gamma_o + \log G + (\varrho-1)\log\Pi^{(\ell-1)} \text{ and } \boldsymbol{\eta}^{(\ell)} = c_0\mathbf{1}_n + X^U\gamma_d + \log G + (\varrho-1)\log P^{(\ell-1)},$$

for $\ell = 1, 2, \cdots$, where $c_0 = -\frac{1}{2}\log G^W$ and $G = (G_1, \cdots, G_n)'$. Then, we have $\boldsymbol{\mu}^{(\ell)}$ for $\ell = 1, 2, \cdots$. Note that we apply the normalization $\sum_i \alpha_i^{(\ell)} - \sum_i \eta_i^{(\ell)} = 0$ for each iternation ℓ. We iterate until $\max\{\|\boldsymbol{\alpha}^{(\ell)} - \boldsymbol{\alpha}^{(\ell-1)}\|_\infty, \|\boldsymbol{\eta}^{(\ell)} - \boldsymbol{\eta}^{(\ell-1)}\|_\infty\} < 10^{-12}$. Condition in Assumption 2.4 is utilized for guaranteeing this convergence. For each simulation replication, we allow random variations in the fixed effects:

$$\boldsymbol{\alpha}^0 = \boldsymbol{\alpha}^{(\infty)} + \boldsymbol{\varepsilon}^\alpha, \text{ and } \boldsymbol{\eta}^0 = \boldsymbol{\eta}^{(\infty)} + \boldsymbol{\varepsilon}^\eta,$$

where $\boldsymbol{\varepsilon}^\alpha = (\varepsilon_1^\alpha, \cdots, \varepsilon_n^\alpha) \sim \mathcal{N}(\mathbf{0}, 0.08^2 I_n)$ and $\boldsymbol{\varepsilon}^\eta = (\varepsilon_1^\eta, \cdots, \varepsilon_n^\eta) \sim \mathcal{N}(\mathbf{0}, 0.08^2 I_n)$. Consequently, we have $\boldsymbol{\mu}^0 = \exp\left(\mathbf{S}^{-1}\left(\mathbf{X}\beta^0 + \boldsymbol{\alpha}^0 \otimes \mathbf{1}_n + \mathbf{1}_n \otimes \boldsymbol{\eta}^0\right)\right)$.

Next, we generate the error components for each simulation replication:

1. First, we generate $\xi_{ij}^* \overset{i.i.d.}{\sim} \text{Lognormal}\left(-\frac{1}{2}\sigma^2, \sigma^2\right)$ across ij with $\sigma^2 = 0.125^2$. Then, $\mathbb{E}\left(\xi_{ij}^*|\mathbf{x}\right) = \mathbb{E}\left(\xi_{ij}^*\right) = 1$.
2. Let $\epsilon_{ij}^* = \mu_{ij}^0$ for all ij. Then, $\mathbb{E}\left(\epsilon_{ij}^*|\mathbf{x}\right) = \mu_{ij}^0 \cdot \left(\mathbb{E}\left(\xi_{ij}^*|\mathbf{x}\right) - 1\right) = 0$ and $\mathsf{Var}\left(\epsilon_{ij}^*|\mathbf{x}\right) = \left(\mu_{ij}^0\right)^2 \cdot \mathsf{Var}\left(\xi_{ij}^*\right) = \left(\mu_{ij}^0\right)^2 \cdot (\exp(\sigma^2) - 1)$.
3. Last, we generate

$$u_{ij} = 0.008\sum_{k=1}^n w_{ik}^*\epsilon_{kj}^* + 0.008\sum_{l=1}^n w_{jl}^*\epsilon_{li}^* + 0.002\sum_{k,l=1}^n w_{ik}^* w_{jl}^*\epsilon_{kl}^*, \tag{3.3}$$

where $W^* = (w_{ij}^*)$ is a row-normalized one characterized by the adjacency based on W (i..e., $w_{ij}^* = \frac{\widetilde{w}_{ij}^*}{\sum_{k=1}^n \widetilde{w}_{ik}^*}$ where $\widetilde{w}_{ij}^* = \mathbb{I}\{w_{ij} + w_{ji} > 0\}$). This error structure follows (3.2) since $\sigma_{ij}^* = \mu_{ij}^0\sqrt{\exp(\sigma^2) - 1}$ and $\mathbf{B} = 0.008(I_n \otimes W^*) + 0.008(W^* \otimes I_n) + 0.002(W^* \otimes W^*)$.

Basic information. Four criteria are used to evaluate the finite sample performance of the PPMLE: (i) empirical bias, (ii) empirical standard deviation (STD), (iii) standard error (s.e.), and (iv) the coverage probability of a nominal 95% confidence interval (CP). To

evaluate the standard errors, we consider three kernel functions: (i) Bartlett, (ii) Parzen, and (iii) Tukey-Hanning. For the distance measurem we first conduct the adjacency matrix $A = (a_{ij})$, $a_{ij} = \max\{\mathbb{I}(w_{ij} > 0), \mathbb{I}(w_{ji} > 0)\}$. Then, we evaluate the geodesic distance d^*_{ij}. To gauge the distance between two pairs, we consider the three types of measures: (i) L^1 distance: $d^{*,1}_{ij,kl} = d^*_{ik} + d^*_{jl}$, (ii) L^2 (Euclidean) distance: $d^{*,2}_{ij,kl} = \sqrt{(d^*_{ik})^2 + (d^*_{jl})^2}$, and (iii) L^∞ distance: $d^{*,\infty}_{ij,kl} = \max\{d^*_{ik}, d^*_{jl}\}$. We set a bandwidth d_N to be the 25th-percentile of $\{d^*_{ij,kl}\}$. For the fixed-effect parameters, we report $\widehat{\alpha}_{49}$ and $\widehat{\eta}_{49}$ as representatives. We consider 1,000 sample repetitions. Table 2 summarizes the results.

Interpretations. Across designs, the PPMLE performs well in finite samples. We see small upward biases in $\widehat{\lambda}_d$ and the fixed effects $(\widehat{\alpha}_{49}, \widehat{\eta}_{49})$, and mild downward biases in $\widehat{\lambda}_o$, $\widehat{\lambda}_w$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$. For instance, the empirical biases are $+0.0123$ for $\widehat{\lambda}_d$ and -0.0163 for $\widehat{\lambda}_w$. These finite-sample biases attenuate as n increases; see Section 3.

For the main parameters $(\widehat{\lambda}, \widehat{\beta})$, our standard errors track the empirical standard deviations reasonably closely, albeit somewhat below them. As a result, coverage probabilities (CPs) are slightly below the 95% nominal rate. For example, under Parzen with an L^2 pair distance, the empirical STD and reported s.e. are $(0.0266, 0.0229)$ for $\widehat{\lambda}_d$ and $(0.0137, 0.0120)$ for $\widehat{\beta}_1$, with CPs of 0.933 and 0.897, respectively. Among kernel–distance combinations, we recommend the Parzen kernel with the L^2 pair distance for the main parameters, which delivers the most stable CPs across coefficients.

For the fixed-effect estimates $(\widehat{\alpha}_{49}, \widehat{\eta}_{49})$, the reported s.e. tend to be understated relative to the empirical STDs (e.g., $\widehat{\alpha}_{49}$: STD 0.0845 vs s.e. 0.0299 under Parzen–L^∞), yielding CPs around 0.86–0.90. The gap narrows with larger n (Section 3). Within our design, Parzen with an L^∞ pair distance performs best for fixed effects. Intuitively, the L^∞ metric better captures the effective dependence radius in the FE direction by guarding against long-range pairwise links, which helps reduce s.e. underestimation.

As a rule of thumb, we recommend Parzen–L^2 for $(\widehat{\lambda}, \widehat{\beta})$ and Parzen–L^∞ for $(\widehat{\alpha}, \widehat{\eta})$. Monte Carlo uncertainty for a 95% CP with 1,000 replications is about 0.7 percentage points, so differences below $\approx$1–2 points are not statistically meaningful, whereas the 4–8 point gaps we observe are. In applications, a slightly larger bandwidth (e.g., a +5–10 percentile increase) can further mitigate undercoverage for the fixed effects.

Table 2: Simulation results

	λ_d	λ_o	λ_w	β_1	β_2	α_{49}	η_{49}
Empirical bias	0.0123	-0.0004	-0.0163	-0.0011	-0.0001	0.0131	0.0044
Empirical STD	0.0266	0.0260	0.0355	0.0137	0.0130	0.0845	0.0850
s.e. (Bartlett, L^1)	0.0228	0.0226	0.0326	0.0117	0.0115	0.0283	0.0196
CP (Bartlett, L^1)	0.9180	0.9250	0.9050	0.9030	0.9110	0.8430	0.8290
s.e. (Bartlett, L^2)	0.0227	0.0227	0.0326	0.0117	0.0115	0.0241	0.0215
CP (Bartlett, L^2)	0.9190	0.9270	0.9120	0.9000	0.9160	0.8560	0.8630
s.e. (Bartlett, L^∞)	0.0227	0.0228	0.0326	0.0117	0.0116	0.0294	0.0224
CP (Bartlett, L^∞)	0.9180	0.9280	0.9120	0.8960	0.9070	0.8580	0.8830
s.e. (Parzen, L^1)	0.0228	0.0232	0.0330	0.0120	0.0118	0.0292	0.0216
CP (Parzen, L^1)	0.9310	0.9260	0.9200	0.9000	0.9180	0.8580	0.8710
s.e. (Parzen, L^2)	0.0229	0.0232	0.0331	0.0120	0.0118	0.0252	0.0228
CP (Parzen, L^2)	0.9330	0.9340	0.9280	0.8970	0.9180	0.8630	0.8910
s.e. (Parzen, L^∞)	0.0229	0.0233	0.0332	0.0120	0.0118	0.0299	0.0232
CP (Parzen, L^∞)	0.9320	0.9310	0.9290	0.8960	0.9160	0.8660	0.9030
s.e. (Tukey–Hanning, L^1)	0.0227	0.0228	0.0326	0.0118	0.0115	0.0284	0.0197
CP (Tukey–Hanning, L^1)	0.9170	0.9270	0.9080	0.9000	0.9140	0.8410	0.8220
s.e. (Tukey–Hanning, L^2)	0.0227	0.0229	0.0327	0.0118	0.0116	0.0242	0.0217
CP (Tukey–Hanning, L^2)	0.9140	0.9260	0.9080	0.9000	0.9170	0.8560	0.8680
s.e. (Tukey–Hanning, L^∞)	0.0227	0.0229	0.0327	0.0118	0.0116	0.0295	0.0224
CP (Tukey–Hanning, L^∞)	0.9220	0.9250	0.9130	0.8970	0.9150	0.8580	0.8840

4 Empirical Application

4.1 Basic setting

In this application, we quantify how network effects in bilateral trade flows shape effective trade costs through connections to third countries. Countries may reduce trade costs by routing or integrating shipments via third-party hubs, but congestion or capacity constraints at those hubs can also raise effective costs and dampen trade among neighboring countries.

We collect the international trade flows y_{ij} from the Center for International Data at UC Davis (`https://cid.ucdavis.edu/worldtradeflows`).[14] Then, y_{ij} denotes gross bilateral merchandise trade flows, including only *cross-border* transactions and excluding services, domestic intermediate transactions, and non-traded sectors. In the analysis, we therefore exclude the domestic trade (intra-trade flows y_{ii}) that are not directly observed in the data.[15]

[14]Relevant explanations can be found in Feenstra et al. (2005).

[15]Although domestic trade is a well-defined theoretical concept, it is not directly observable in the data. When we define $y_{ii} = G_i - \sum_{j=1, j\neq i}^{n} y_{ij}$ using the country i's GDP G_i, y_{ii} overwhelms the share since the GDP includes non-traded goods, services, and domestic intermediate transactions. Because our objective is to study how bilateral trade-cost shocks reshape the distribution of international trade relationships, we do not explicitly consider the intra flows y_{ii}. This approach is standard in the empirical trade and trade-network literature (Anderson and van Wincoop, 2003; Head and Mayer, 2014). While recent structural gravity literature emphasizes the importance of intra-national trade flows for identifying the border effect

For the covariates, we use the same set as in Helpman et al. (2008). FTA data are collected from WTO (`https://www.wto.org/english/res_e/statis_e/statis_e.htm`).

Understanding the evolution of global trade requires recognizing that international linkages are not static but shaped by major economic shifts. Hence, we suspect that the structure and sources of network effects vary across four key phases of global trade: Phase 1 (1986, trade liberalization), Phase 2 (1997, active NAFTA implementation), Phase 3 (2007, emergence of the China trade shock), and Phase 4 (2016, expansion of global supply chains). That is, we utilize the four cases of y_{ij}: $y_{ij}^{\text{Phase}} = y_{ij,t_{\text{Phase}}}$ denotes the trade flow from j to i in year t_{Phase}, and $t_{\text{Phase=1}} = 1986$, $t_{\text{Phase=2}} = 1997$, $t_{\text{Phase=3}} = 2007$, and $t_{\text{Phase=4}} = 2016$. The number of countries included in each phase is 136, 142, 146, and 147, respectively.[16]

Descriptive statistics. Table 3 presents the descriptive statistics for the main variables used in the analysis across the four phases of global trade. The mean bilateral trade flow (y_{ij}), measured in 2015 constant U.S. dollars, increases markedly from Phase 1 to Phase 3 before stabilizing at a similar level in Phase 4. This pattern reflects the steady expansion of global trade and the progressive deepening of international production networks. Meanwhile, the share of zero trade flows declines over time—from roughly 53% in Phase 1 to about 28% in Phase 4—indicating that more country pairs established trading relationships as global markets became increasingly interconnected. In addition, y_{ij}^{+} denotes the mean conditional on strictly positive flows, which rises from about USD 0.43 million in Phase 1 to about USD 1.0 million in Phase 3–4.

Countries' connectivity matrix To operationalize these phases, we constructed the countries' connectivity matrix (W) from trade flows observed prior to each referential phase, so as to avoid simultaneity with the outcome period. Let $\mathcal{T}^{\text{Phase}}$ denote the set of years used to construct the connectivity matrix for a given phase.[17] For any $i \neq j$ and for $t \in \mathcal{T}^{\text{Phase}}$,

$$w_{ij}^{\text{Phase}} = \frac{\widetilde{w}_{ij}^{\text{Phase}}}{\sum_{k=1}^{n} \widetilde{w}_{ik}^{\text{Phase}}}, \text{ where } \widetilde{w}_{ij}^{\text{Phase}} = \frac{1}{\#(\mathcal{T}^{\text{Phase}})} \sum_{t \in \mathcal{T}^{\text{Phase}}} (y_{ij,t} + y_{ji,t})$$

with $w_{ii}^{\text{Phase}} = 0$. This produces a row–stochastic W with a symmetric base $\widetilde{W}$, consistent with Assumption 2.2. Each w_{ij}^{Phase} admits a probability interpretation as the phase–specific likelihood of choosing j as i's partner.

(Yotov, 2022), our study focuses specifically on the reallocation effects within the international trade network.

[16] The full list of country names is provided in Table A.1.

[17] In detail, $\mathcal{T}^{\text{Phase=1}} = \{1984, 1985\}$ (as data collection begins in 1984), $\mathcal{T}^{\text{Phase=2}} = \{1993, 1994, 1995, 1996\}$, $\mathcal{T}^{\text{Phase=3}} = \{2000, 2001, \cdots, 2006\}$, and $\mathcal{T}^{\text{Phase=4}} = \{2010, 2011, \cdots, 2015\}$.

Table 3: Descriptive statistics

Phase	1			2			3			4		
	Mean	Median	STD	Mean	Median	STD	Mean	Median	STD	Mean	Median	STD
y_{ij}	201,347	0	2,604,211	380,573	382	4,151,519	753,486	1,127	7,055,185	718,068	865	7,174,976
(Zero freq.)	(0.5307)			(0.3007)			(0.2488)			(0.2835)		
y_{ij}^{+}	429,005	8,105	3,788,568	544,197	4,726	4,955,465	1,003,100	7,249	8,125,016	1,002,158	8,909	8,459,553
Distance	0.2898	0.2635	0.1821	0.2864	0.2598	0.1801	0.2853	0.2589	0.1796	0.2829	0.2543	0.1787
Border	0.0184	0.0000	0.1344	0.0188	0.0000	0.1357	0.0190	0.0000	0.1365	0.0186	0.0000	0.1352
Legal	0.3760	0.0000	0.4844	0.3629	0.0000	0.4808	0.3623	0.0000	0.4807	0.3673	0.0000	0.4821
Language	0.3254	0.0000	0.4685	0.3100	0.0000	0.4625	0.3027	0.0000	0.4594	0.3035	0.0000	0.4598
Colony	0.0117	0.0000	0.1073	0.0110	0.0000	0.1042	0.0105	0.0000	0.1019	0.0104	0.0000	0.1016
Currency	0.0109	0.0000	0.1038	0.0100	0.0000	0.0994	0.0094	0.0000	0.0967	0.0093	0.0000	0.0961
Islands	0.3824	0.0000	0.5541	0.3662	0.0000	0.5450	0.3699	0.0000	0.5472	0.3673	0.0000	0.5457
Landlock	0.3088	0.0000	0.5091	0.3099	0.0000	0.5099	0.3151	0.0000	0.5134	0.2993	0.0000	0.5028
FTA	0.0004	0.0000	0.0209	0.0010	0.0000	0.0316	0.0060	0.0000	0.0775	0.0107	0.0000	0.1030

Note: Phase 1 (1986, trade liberalization), Phase 2 (1997, active NAFTA implementation), Phase 3 (2007, emergence of the China trade shock), and Phase 4 (2016, expansion of global supply chains). The definitions of explanatory variables are adopted from Helpman et al. (2008) as follows: *Distance:* the distance between importer i's and exporter j's capitals; *Border:* a binary variable that equals one if country j and country i are neighbors that meet a common physical boundary, and zero otherwise; *Legal:* a binary variable that equals one if country j and country i share the same legal system, and zero otherwise; *Language:* a binary variable that equals one if country i and j share the same language system, and zero otherwise; *Colony:* a binary variable that equals one if country j ever colonized country i or vice versa, and zero otherwise; *Currency*: a binary variable that equals one if the country j and country i use the same currency or if within the country pair money was interchangeable at a 1:1 exchange rate for an extended period of time; *Islands:* a binary variable that equals one if both importer i and exporter j are islands, and zero otherwise; *Landlock:* a binary variable that equals one if both exporting country j and importing country i have no coastline or direct access to sea, and zero otherwise; *FTA:* a binary variable that equals one if country j and country i belong to a common regional trade agreement, and zero otherwise.

To better understand the role of W, we describe its main properties and their implications via their network statistics. Since W is row-normalized, its maximum eigenvalue ($\varphi_{\max} = \varphi_{(1)}$) is equal to 1. Because the matrix also has zero diagonal elements (i.e., $0 = \text{tr}(W) = \sum_{i=1}^{n} \varphi_i$), its minimum eigenvalue must be negative, though still greater than or equal to -1. The minimum eigenvalue of W ($\varphi_{\min}$) represents the network's bipartiteness, and this tendency becomes stronger if $\varphi_{\min} \to -1$. On the other hand, if the second-largest ($\varphi_{(2)}$) and minimum eigenvalues ($\varphi_{\min}$) are both close to zero, it indicates that W has a high averaging rate (Chung, 1997; Bramoullé et al., 2014). By contrast, if both $\varphi_{(2)}$ and $\varphi_{\min}$ diverge from zero, the weighted influence of neighboring units does not collapse into something close to a constant (such as a uniform average across units), but instead reflects meaningful variation driven by the structure of the network.

Table 4 documents how the countries' connectivity matrix evolves across phases. Panel A reports Frobenius distances between the phase-specific networks, restricted to the 135 countries that appear in all periods. The distances grow monotonically with the time gap between phases, indicating that the structure of trade linkages has not only shifted over time but has done so in a cumulative way. In particular, the connectivity matrix in 2016 (Phase 4) is substantially further from the 1986 benchmark (Phase 1) than from the intermediate phases, consistent with the idea that trade liberalization, regional integration (NAFTA), the rise of China, and the expansion of global supply chains have jointly transformed the geography of trade relationships.

Panel B summarizes key network statistics. Two main patterns stand out. First, the connectivity networks become progressively denser and more diversified. Average degree rises from 81 to 138 between Phase 1 and Phase 4, while network density increases from about 0.49 to above 0.90. At the same time, the dispersion in degree shrinks, suggesting that the distinction between highly connected and poorly connected countries has weakened over time. Both the HHI-based and entropy-based effective numbers of partners, n^{HHI} and n^{E}, also increase across phases. The former implies that the number of economically "dominant" partners per country rises from about 8 to about 12, whereas the latter suggests that the overall diversification of partner portfolios—including small "long-tail" partners—grows from roughly 14 to 21 effective partners.

Second, these changes are not purely mechanical consequences of higher density. The normalized Shannon partner diversification entropy (PDE) increases across phases, while its cross-country dispersion declines. This indicates that countries have not only added more links, but have also allocated trade more evenly across partners. At the same time, the high-intensity degree exhibits a large cross-country dispersion in all phases, which is consistent with the presence of a small set of hub countries that account for a disproportionate share of strong connections. The eigenvalue statistics reinforce this picture: the second-largest eigenvalue increases modestly, and the most negative eigenvalue moves toward zero, suggesting that the network becomes more integrated and less polarized over time, departing from a bipartite-type structure. Overall, the evidence is consistent with a transition from a relatively sparse and unequal trade network toward a nearly fully connected and more balanced system, while retaining a prominent role for a limited number of global hubs.

Country-specific statistics reveal a persistent core–periphery structure across all four phases (see Appendix Tables 1–4). A small set of high-income economies—most notably the United States, Germany, France, the United Kingdom, China, Japan, India, Singapore, Korea and Australia—consistently emerge as hubs, with degrees close to the maximum, large hub indices, low concentration (low HHI and high n^{HHI}) and high entropy-based effective numbers of partners n^{E}, indicating that their links are spread relatively evenly over a wide range of partners. By contrast, many small or low-income countries, particularly in Africa, the Caribbean, and Oceania, display medium to low degrees but high concentration (high n^{HHI}, low n^{HHI}), implying dependence on a few core partners and highly skewed partner distributions. Over time, degrees increase and the connectivity network becomes denser—with China in particular rapidly converging to the hub group—but the concentration and divergence patterns for peripheral economies change little, so that the core–periphery gap in partner diversification remains pronounced despite the overall thickening of the network. More details can be found in Appendix 4.

Table 4: Network statistics for the countries' connectivity matrix

Panel A. Relationships among the connectivity networks across phases

	$W^{\text{Phase=1}}$	$W^{\text{Phase=2}}$	$W^{\text{Phase=3}}$	$W^{\text{Phase=4}}$
$W^{\text{Phase=2}}$	2.3728 (0.0159)	0	*	*
$W^{\text{Phase=3}}$	2.8813 (0.0193)	1.9212 (0.0129)	0	*
$W^{\text{Phase=4}}$	3.6703 (0.0246)	2.8833 (0.0193)	1.9474 (0.0130)	0

Panel B. Summary network statistics

	linear-in-means	Bipartite	Phase 1	Phase 2	Phase 3	Phase 4
degree (deg)	149.0000	75.0000	81.3088	118.9859	136.3973	138.4354
Std(deg)	0.0000	0.0000	34.9280	21.9047	12.5476	12.1543
high-intensity degree ($\deg^+$)	0.0000	0.0000	6.7500	7.0493	7.2466	7.3469
Std($\deg^+$)	0.0000	0.0000	20.0458	19.1505	18.1331	18.0075
Herfindahl–Hirschman Index (HHI)	0.0067	0.0133	0.1673	0.1419	0.1249	0.1182
Std(HHI)	0.0000	0.0000	0.1206	0.1040	0.0932	0.0881
n^{HHI}	149.0000	75.0000	8.3230	9.7423	11.0296	11.5994
Partner diversification entropy (PDE)	1.0000	0.8628	0.5182	0.5623	0.5853	0.5921
Std(PDE)	0.0000	0.0000	0.1080	0.0968	0.0905	0.0924
n^{E}	149.0000	75.0000	14.3640	17.8278	20.0889	20.9272
$\varphi_{(2)}$	-0.0067	$\simeq$0	0.5413	0.5653	0.5752	0.5553
$\varphi_{\min}$	-0.0067	-1	-0.5296	-0.5151	-0.5088	-0.4419
Density	1	0.5034	0.4948	0.7560	0.8910	0.9105

Note: Phase 1 (1986, trade liberalization), Phase 2 (1997, active NAFTA implementation), Phase 3 (2007, emergence of the China trade shock), and Phase 4 (2016, expansion of global supply chains). The first panel reports the Frobenius-norm-based distances between the connectivity matrices of the two phases, along with their normalized values in parentheses, for the 135 common countries. That is, we report $\|W^{\text{Phase}} - W^{\text{Phase}\prime}\|_F$ (and $\frac{1}{\sqrt{n(n-1)}}\|W^{\text{Phase}} - W^{\text{Phase}\prime}\|_F$ in the parentheses). In the second panel, key network statistics are illustrated. For comparison, we also report network statistics for the *linear-in-means* network and the *bipartite* network (a perfectly polarized structure in which nodes are divided into two mutually connected groups with no intra-group links). $\deg_i = \sum_{j=1}^n \mathbb{I}(w_{ij} > 0)$, and $\deg_i^+ = \sum_{j=1}^n \mathbb{I}(w_{ij} > w_{0.95})$, where $w_{0.95}$ denotes the 95% percentile of $\{w_{ij}\}_{j\neq i}$. $\text{HHI}_i = \sum_{j=1}^n w_{ij}^2$, and $n_i^{\text{HHI}} = \frac{1}{\text{HHI}_i}$. For the partner diversification entropy, we report the normalized Shannon entropy $\widetilde{H}_i = \frac{H_i}{\ln(n-1)} \in [0,1]$, where $H_i = -\sum_{j=1}^n w_{ij}\ln(w_{ij})$ and the effective number of partner $n_i^{\text{E}} = \exp(H_i)$. $\varphi_{(2)}$ denotes the second-largest eigenvalue, $\varphi_{\min}$ denotes the smallest eigenvalue, and *Density* represents the proportion of nonzero elements (edges) in the network.

4.2 Estimation results

Table 5 presents how network effects evolve across phases. We focus on the three network parameters $(\lambda_d, \lambda_o, \lambda_w)$, which govern how bilateral trade flows respond to changes in neighboring flows along the three network dimensions $I_n \otimes W$, $W \otimes I_n$, and $W \otimes W$.

As a key quantitative finding, our model structure, which allows for these network spillovers, substantially improves model fit relative to the conventional gravity equation with a purely iceberg-cost specification. The McFadden R^2 reported in Table 5 ranges from about 0.10 in Phase 3 to about 0.28 in Phase 2, with values of 0.13 and 0.17 in Phases 1 and 4, respectively. Since the McFadden R^2 here is defined relative to a conventional gravity model, these figures indicate that incorporating third–party proximities and network interactions improves the log-likelihood by roughly 10–30 percent, depending on the phase.

Because a well-measured cost specification is key to gravity equations that reduce residuals, this provides evidence of the significant roles of third-party proximities in international trade flows that are not captured by multilateral resistance terms (fixed-effect components). In other words, the pair-specific heterogeneities generated by our model may not be captured by the conventional gravity model and thus constitute residuals in it. This means that specific channels in our model do not operate within the conventional gravity equation, as they do not induce changes in economic behavior (see Sec. 4.3 for details).

In Phase 1 (1986, trade liberalization), both λ_d and λ_o are negative and statistically significant, indicating that trade flows sharing the same exporter or the same importer behaved as *substitutes*. In our framework, $\lambda_d < 0$ implies that, holding other factors constant, an increase in exports from a given origin j to some destinations tends to reduce the weighted average exports from j to its other partners, while $\lambda_o < 0$ implies an analogous substitution pattern across alternative exporters serving the same destination i. These results are consistent with a situation in which, at the onset of liberalization, exporters and importers faced capacity or market–access constraints, so that expanding one bilateral relationship came at the expense of others. In network terms, the indirect effects propagated through the $I_n \otimes W$ and $W \otimes I_n$ components primarily *dampen* neighboring flows, so that network competition dominates network complementarity along the exporter and importer dimensions.

The parameter on the third–country dimension, λ_w, is also negative and significant in Phase 1. Since λ_w multiplies the $W \otimes W$ component, this finding suggests that, in the early post-liberalization period, increased trade among countries that are jointly close to i and j (common or two–step neighbors in W) tended to *crowd out* flows on the ij link. One interpretation is that hub countries and transit routes were still subject to strong congestion or capacity constraints, so that routing more trade through these hubs increased effective

trade costs for their neighbors.

In Phase 2 (1997, active NAFTA implementation), both λ_d and λ_o switch sign and become positive and strongly significant, reflecting a structural shift toward *complementarity* among trade flows. Within our model, $\lambda_d > 0$ means that when an exporter deepens trade with some destinations, the weighted average flows from that exporter to other markets also tend to rise; $\lambda_o > 0$ analogously captures complementarity among alternative suppliers into a given destination. This pattern is consistent with reductions in trade frictions and institutional integration in the late 1990s—for example, those associated with NAFTA and other regional arrangements—that made it easier for firms to serve multiple destinations and to diversify their sourcing. Bilateral flows thus become more mutually reinforcing within the trade network, and the indirect effects along $I_n \otimes W$ and $W \otimes I_n$ operate more like increasing returns to network connectivity rather than crowding–out.

At the same time, λ_w turns positive and is precisely estimated, indicating the emergence of robust *third–party spillover effects* once cross–country trade networks are more firmly in place. A positive λ_w implies that the trade flow from j to i increases when countries that are simultaneously well connected to both i and j expand their trade, so that bilateral trade is amplified when partners share access to the same hubs.

In Phase 3 (2007, emergence of the China trade shock), λ_d and λ_o again become negative and statistically significant. Within our framework, this means that, conditional on fundamentals, an increase in the averaged flows that share the same exporter or the same importer now *reduces* the equilibrium flow on a given link ij. When $\lambda_d, \lambda_o < 0$, the off-diagonal entries of $\mathbf{S}^{-1}$ associated with the $I_n \otimes W$ (destination-side) and $W \otimes I_n$ (origin-side) directions are effectively negative, so that a positive shock to one trade link crowds out flows on neighboring links that involve the same exporter or the same importer. In other words, trade flows behave as *substitutes* along these two network dimensions. This pattern is consistent with the emergence of China as a dominant global exporter: the sharp rise of China-centered flows intensified competition for market share, inducing exporters to reallocate sales across destinations and importers to reshuffle sourcing across competing suppliers.

The magnitude of λ_w rises sharply in Phase 3, becoming large and precisely estimated. This indicates that hub–mediated, two–step propagation of trade volumes and trade costs became particularly important in that period: trade among common neighbors of i and j substantially boosts the bilateral flow from j to i, even as direct competition along the exporter and importer dimensions remains strong.

By Phase 4 (2016, expansion of global supply chains), λ_d becomes relatively small in magnitude and only weakly significant, while λ_o remains positive and highly significant but

with a reduced magnitude relative to its Phase 2 level. This pattern suggests a partial stabilization of trade interdependencies within an increasingly dense and geographically diversified trade network. The modest value of λ_d points to a weaker role for substitution or complementarity along the "common–destination" dimension, whereas the still positive λ_o implies that some residual complementarity continues to operate primarily through common export origins. In other words, exporters that are important suppliers to a given destination tend to be essential suppliers to its other partners as well, but this reinforcement is more modest than in earlier phases.

The third–country parameter λ_w remains positive, sizeable, and precisely estimated in Phase 4, though its magnitude is lower than the peak observed in 2007. Overall, the evolution of λ_w suggests a transition from early hub congestion and competition (negative third–party effects in Phase 1) to strong and persistent hub–driven complementarities once global supply chains are fully in place (positive and large third–party effects from Phase 2 onward).

4.3 Counterfactual analysis

Motivated by the significant estimated spillover parameters, we investigate how changes in the economic environment alter the distribution of trade flows in our model and in the conventional gravity model. Accordingly, our framework provides a tool for ex-ante stress testing of trade networks. Because counterfactual trade shares are generated from the same estimated structure, policymakers can use the model to simulate hypothetical shocks—such as tariff escalations, decoupling strategies, or supply-chain disruptions—and trace how these shocks propagate through the network, reshaping import concentration and diversification.

Focusing on Phase 4, we consider a threefold increase in $\tau^{+}_{\text{US,CN}}$ and $\tau^{+}_{\text{CN,US}}$, thereby illustrating the recent US-China trade war as a counterfactual scenario. Specifically, we compare predicted trade flows using our estimates $(\widehat{\cdot})$ with those obtained from the counterfactual scenario $(\widetilde{\cdot})$. We set the elasticity of substitution to $\varrho = 5$, following Anderson and van Wincoop (2003).[18] Based on the counterfactual fixed-effect components $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\eta}}$ obtained

[18]Head and Mayer (2014) also set the value of ϱ around 5 (i.e., $\varrho = 5.03$) as their preferred estimate from their meta analysis.

Table 5: Estimation Results by Phase

Phase	1	2	3	4
λ_d	-0.1261	0.3002	-0.6440	0.1370
	(0.0480)	(0.0362)	(0.1106)	(0.0712)
λ_o	-0.1900	0.3510	-0.6246	0.2830
	(0.0968)	(0.0373)	(0.0913)	(0.0649)
λ_w	-0.1533	0.3354	1.3110	0.5734
	(0.0561)	(0.0336)	(0.0742)	(0.0418)
Distance	-1.5011	-1.5184	-1.5808	-1.7511
	(0.7203)	(0.2449)	(0.3730)	(0.2849)
Border	1.1213	0.9973	-0.1417	0.9402
	(0.3638)	(0.2134)	(0.2256)	(0.1575)
Legal	0.2610	0.2029	0.3701	0.1337
	(0.0757)	(0.0682)	(0.0926)	(0.0698)
Language	0.0706	-0.0159	0.1536	0.0225
	(0.1017)	(0.0547)	(0.0484)	(0.0697)
Colony	0.0662	0.0000	-0.2150	0.1386
	(0.1397)	(0.0985)	(0.1567)	(0.0871)
Currency	0.4327	-0.0682	0.2835	-0.0622
	(0.3968)	(0.4260)	(0.2833)	(0.3185)
Islands	18.7876	-8.0810	5.5282	-6.3638
	(10.2236)	(0.0784)	(0.5768)	(0.0639)
Landlock	18.4614	-7.2785	8.9211	-10.5669
	(14.0051)	(0.0649)	(5.9597)	(0.0482)
FTA	0.8220	0.4123	0.2586	0.4968
	(0.2338)	(0.1717)	(0.1303)	(0.0851)
# of observations	18,360	20,022	21,170	21,462
Log-likelihood	-774,526,152	-1,323,341,030	-5,561,319,422	-3,009,628,447
McFadden's R^2	0.1295	0.2848	0.1037	0.1739

Note: Standard errors are evaluated by the Parzen kernel with the L^2-based distance, and are reported in parentheses. Phase 1 (1986, trade liberalization), Phase 2 (1997, active NAFTA implementation), Phase 3 (2007, emergence of the China trade shock), and Phase 4 (2016, expansion of global supply chains). The definitions of explanatory variables are adopted from Helpman et al. (2008) as follows: *Distance:* the distance between importer i's and exporter j's capitals; *Border:* a binary variable that equals one if country j and country i are neighbors that meet a common physical boundary, and zero otherwise; *Legal:* a binary variable that equals one if country j and country i share the same legal system, and zero otherwise; *Language:* a binary variable that equals one if country i and j share the same language system, and zero otherwise; *Colony:* a binary variable that equals one if country j ever colonized country i or vice versa, and zero otherwise; *Currency*: a binary variable that equals one if the country j and country i use the same currency or if within the country pair money was interchangeable at a 1:1 exchange rate for an extended period of time; *Islands:* a binary variable that equals one if both importer i and exporter j are islands, and zero otherwise; *Landlock:* a binary variable that equals one if both exporting country j and importing country i have no coastline or direct access to sea, and zero otherwise; *FTA:* a binary variable that equals one if country j and country i belong to a common regional trade agreement, and zero otherwise. When we evaluate the log-likelihood values, we consider $\ln(\Gamma(y_{ij}+1))$ for $\ln y_{ij}!$, where $\Gamma(\cdot)$ denotes the gamma function. The McFadden's R^2 (McFadden, 1972) here is defined by $1-\frac{\widehat{\ell}_N}{\widehat{\ell}_N^{trad.}}$, where $\widehat{\ell}_N$ and $\widehat{\ell}_N^{trad.}$ denote respectively the log-likelihood values at the PPMLEs from our model and the conventional gravity model.

by (2.6), we generate the counterfactual trade flows $\widetilde{\boldsymbol{\mu}}$.[19]

Next, we compute the estimated and counterfactual shares to understand changes in trade flows from different economic environments. In both models, importer budget G_i is taken as given/absorbed by fixed effects, so the counterfactual is not designed to predict aggregate trade volumes. The economically meaningful adjustment margin is the reallocation of a fixed import budget across partners, which is naturally summarized by import shares. Hence, the counterfactual is best interpreted as a redistribution exercise: the shock primarily changes the composition of each importer's sourcing across partners, rather than providing a forecast of aggregate trade volume. For this reason, we summarize equilibrium adjustments using changes in import shares, which differ across importers and isolate reallocation across partners.[20] For each $i = 1, \cdots, n$,

$$\widehat{s}_{ij} = \frac{\widehat{\mu}_{ij}}{\sum_{k=1,k\neq i} \widehat{\mu}_{ik}} \text{ and } \widetilde{s}_{ij} = \frac{\widetilde{\mu}_{ij}}{\sum_{k=1,k\neq i}^{n} \widetilde{\mu}_{ik}} \text{ for } j \neq i.$$

For comparison, we also compute the estimated and counterfactual shares $\widehat{s}_{ij}^{\text{con}}$ and $\widetilde{s}_{ij}^{\text{con}}$ from the conventional gravity model. We report the results by selecting 10 countries with four categories: (i) Category 1 (Hub countries having high degree with low concentration): U.S., Germany, and Japan, (ii) Category 2 (Countries with structural changes): China and South Korea, (iii) Category 3 (Countries with high concentration): Canada and Mexico, and (iv) Category 4 (Countries with low concentration): Sudan, Egypt, and Kenya.

Model's mechanism. By increasing $\tau^{+}_{\text{US,CN}}$ and $\tau^{+}_{\text{CN,US}}$, we study how a bilateral cost shock propagates through the equilibrium mapping and ultimately reshapes the import shares $\{s_{ij}\}_{j\neq i}$.

- Conventional gravity model ($\lambda = 0$):

[19] In detail, we generate $\widetilde{\boldsymbol{\mu}}$ by the following steps.

- **Step 1**. Evaluate $\widehat{\tau}^{+}_{ij} = \exp\left(\frac{1}{1-\varrho} \cdot x'_{ij}\widehat{\beta}\right)$ using $\widehat{\beta}$ and $\varrho = 5$.
- **Step 2**. Construct $\widetilde{\tau}^{+}_{ij} = \widehat{\tau}^{+}_{ij} + \Delta^{+}_{ij}$, where $\Delta^{+}_{ij} = \begin{cases} 2\widehat{\tau}^{+}_{ij} & \text{if } ij \in \{(\text{US},\text{CN}),(\text{CN},\text{US})\}, \\ 0 & \text{otherwise.} \end{cases}$.
- **Step 3**. Generate $\widetilde{\boldsymbol{\mu}}$ with $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\eta}}$ as the data generating process in Sec 3.3.

[20] Both our model and the conventional gravity framework (Anderson and van Wincoop, 2003) embed a CES demand system with multilateral resistance. Our counterfactual perturbs only the targeted bilateral trade-cost components and recomputes equilibrium objects within the same estimated structure, so the exercise is naturally interpreted as tracing *reallocation* across partners. Reporting changes in levels μ_{ij} may conflate two margins: changes in the overall scale of importer i's imports and changes in the *composition* of imports across partners. Import shares provide a scale-free, directly comparable summary of the equilibrium adjustment across models and countries.

- Direct effect: These changes directly lead to decreasing $\mu_{\text{US,CN}}$ and $\mu_{\text{CN,US}}$.
- Via the multilateral resistance terms: These shocks change the multilateral resistance terms $\{P_i, \Pi_j\}_{i,j=1}^n$ through price-index/expenditure-share re-optimization, implying adjustments in third-country flows μ_{kl} (for $k \neq i$ or $l \neq j$) through the multilateral resistance terms (the magnitude depends on substitution patterns and baseline trade shares).[21]

- Our model's additional channels (if $\widehat{\lambda}_d > 0$, $\widehat{\lambda}_o > 0$, and $\widehat{\lambda}_w > 0$):
  - Endogenous network trade-cost channel: The increase in $\tau^+_{\text{US,CN}}$ and $\tau^+_{\text{CN,US}}$ reduces the targeted flows, reallocates $\boldsymbol{\mu}$ in equilibrium, and thereby changes $\tau^e_{kl}(\boldsymbol{\mu})$ for many non-targeted pairs kl. Hence, trade flows μ_{kl} can change even when τ^+_{kl} is held fixed.
  - Multilateral resistance amplification: Because network costs $\tau^e_{ij}(\boldsymbol{\mu})$ respond endogenously to the reallocation of $\boldsymbol{\mu}$, the adjustment in $\{P_i(\boldsymbol{\mu}), \Pi_j(\boldsymbol{\mu})\}$ is typically larger and more system-wide than under $\lambda = 0$, generating additional propagation of the bilateral shock to third-country trade flows. This is because $\tau^e(\mu)$ shifts the entire vector of bilateral trade costs that enters the CES price indices.

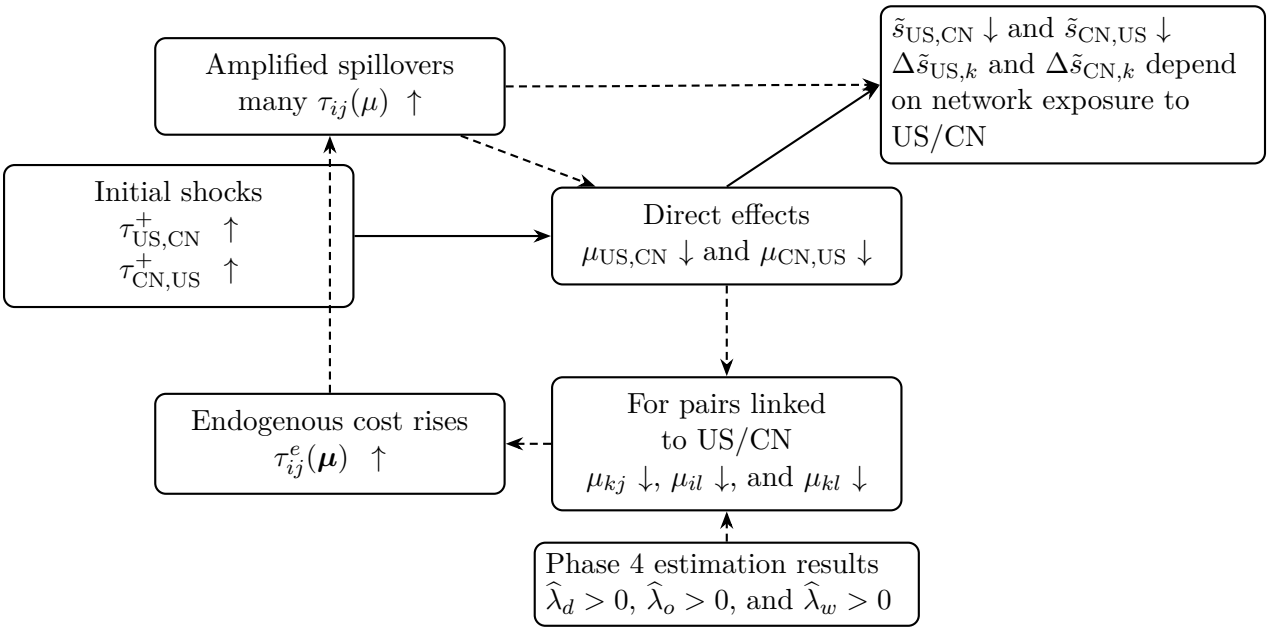
Figure 3 illustrates the model's main mechanism as a diagram.

Summary of the key findings. While higher U.S.–China trade costs naturally reshape the import composition of the two countries involved, our network model implies substantial adjustments in the import shares of third countries. In contrast, under the conventional gravity framework, these third-country effects are largely absent. This difference arises because, in our model, trade flows are jointly determined through the network multiplier matrix $\mathbf{S}^{-1}$, so that a bilateral shock propagates through origin-, destination-, and hub-based linkages.

Category 1: Hub economies. For the U.S., the increase in U.S.–China trade costs results in an almost complete collapse in China's import share (from $\widehat{s}_{\text{US,CN}} = 0.1790$ to $\widetilde{s}_{\text{US,CN}} = 0.0024$). In our model, this loss is absorbed by a broad set of alternative suppliers (e.g., Germany, Japan, the UK, France, Canada, etc.), resulting in pronounced declines in the Gini coefficient. By contrast, under conventional gravity, China's share also falls (from

[21] The signs of ΔP_i, $\Delta \Pi_j$, and $\Delta \mu_{kl}$ are generally ambiguous and depend on substitution patterns and initial trade shares.

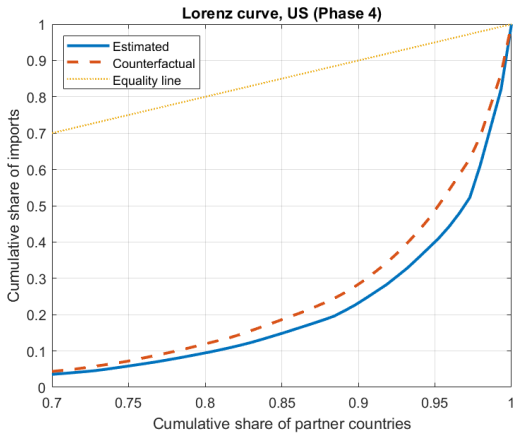
Figure 3: Model's mechanism



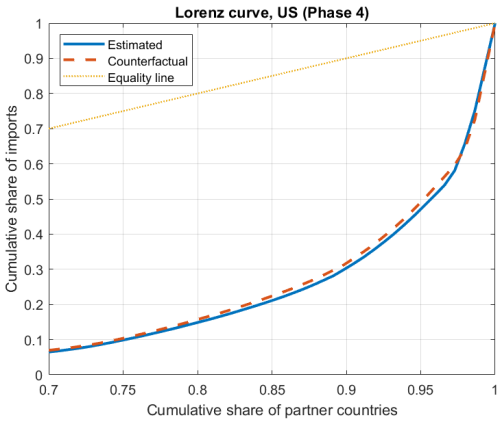
Note: Solid arrows indicate channels that are present in the conventional model (i.e., $\lambda = 0$). In contrast, dashed arrows represent additional propagation mechanisms implied by our estimated network spillovers in Phase 4 ($\widehat{\lambda}_d > 0$, $\widehat{\lambda}_o > 0$, $\widehat{\lambda}_w > 0$). The initial bilateral shock raises the exogenous trade-cost components for the targeted pair (US–CN and CN–US), reducing the corresponding predicted flows. In our model, this reallocation alters the endogenous network trade costs $\tau^{e}_{ij}(\boldsymbol{\mu})$, thereby amplifying adjustments in multilateral resistance terms and transmitting the shock to non-targeted country pairs. The "Share view" box summarizes the counterfactual reporting object: changes in import shares $\widetilde{s}_{ij}$ for $j \neq i$.

Figure 4: Lorenz curve (Category 1: Hub economies)

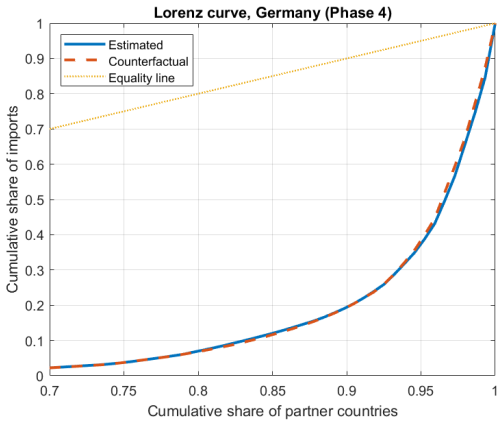
(a) U.S. (Our model)



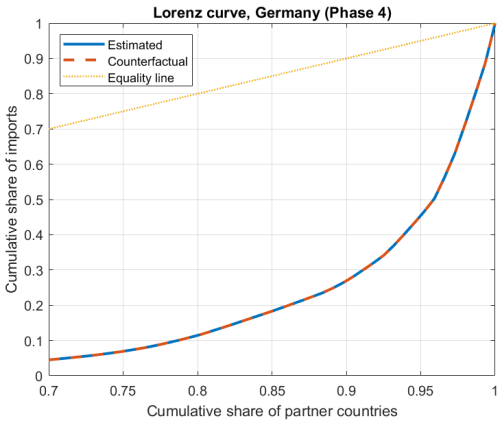
(b) U.S. (Conventional model)



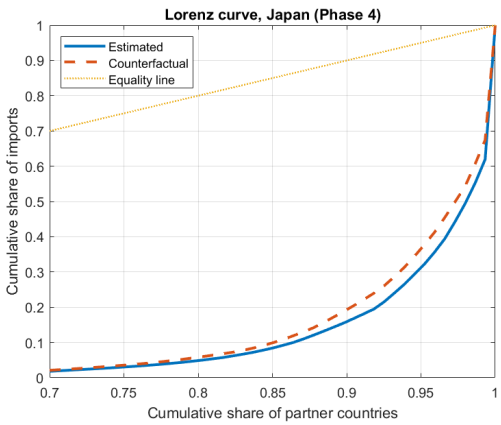
(c) Germany (our model)



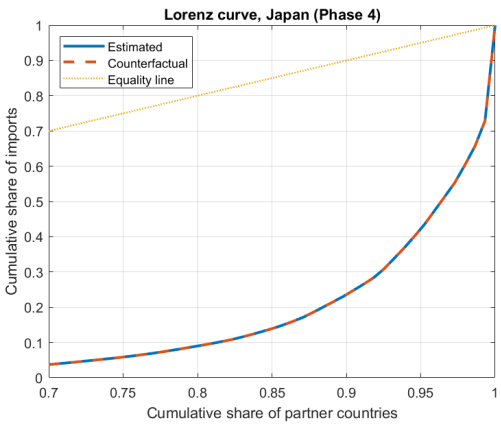
(d) Germany (Conventional model)



(e) Japan (our model)



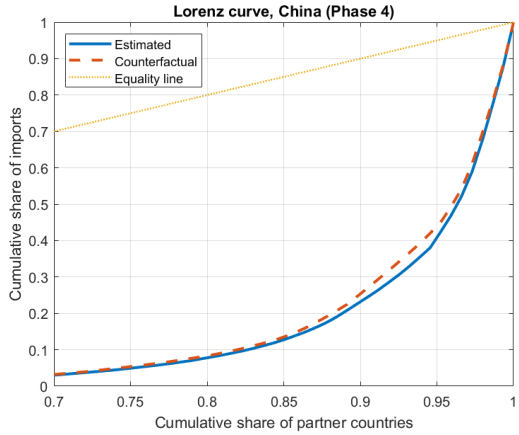
(f) Japan (Conventional model)



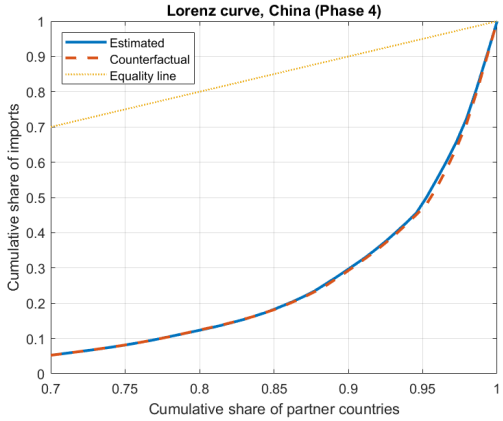
Note: The Gini coefficients from the Lorenz curves are as follows: (i) For Panels (a) and (b), $\text{Gini}(\widehat{s}_{\text{US},\cdot}) = 0.8574$, $\text{Gini}(\widetilde{s}_{\text{US},\cdot}) = 0.8304$, $\text{Gini}(\widehat{s}^{\text{con}}_{\text{US},\cdot}) = 0.8146$, $\text{Gini}(\widetilde{s}^{\text{con}}_{\text{US},\cdot}) = 0.8074$; (ii) For Panels (c) and (d), $\text{Gini}(\widehat{s}_{\text{DE},\cdot}) = 0.8726$, $\text{Gini}(\widetilde{s}_{\text{DE},\cdot}) = 0.8712$, $\text{Gini}(\widehat{s}^{\text{con}}_{\text{DE},\cdot}) = 0.8331$, $\text{Gini}(\widetilde{s}^{\text{con}}_{\text{DE},\cdot}) = 0.8331$; (iii) For Panels (e) and (f), $\text{Gini}(\widehat{s}_{\text{JP},\cdot}) = 0.8996$, $\text{Gini}(\widetilde{s}_{\text{JP},\cdot}) = 0.8849$, $\text{Gini}(\widehat{s}^{\text{con}}_{\text{JP},\cdot}) = 0.8571$, $\text{Gini}(\widetilde{s}^{\text{con}}_{\text{JP},\cdot}) = 0.8571$.

Figure 5: Lorenz curve (Category 2: Structural-changed economies)

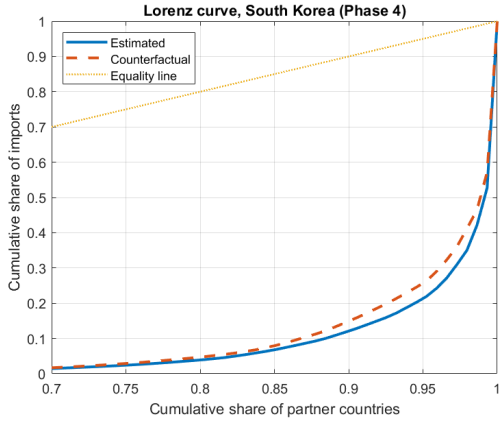
(a) China (Our model)



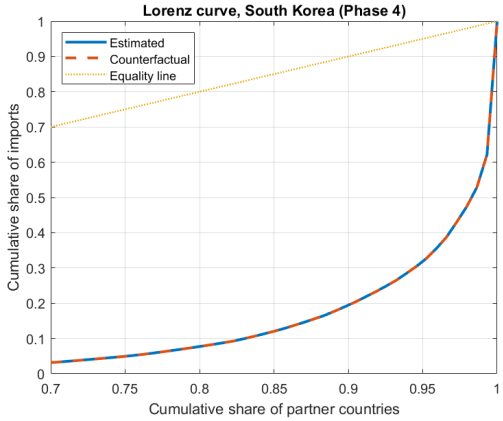
(b) China (Conventional model)



(c) South Korea (our model)



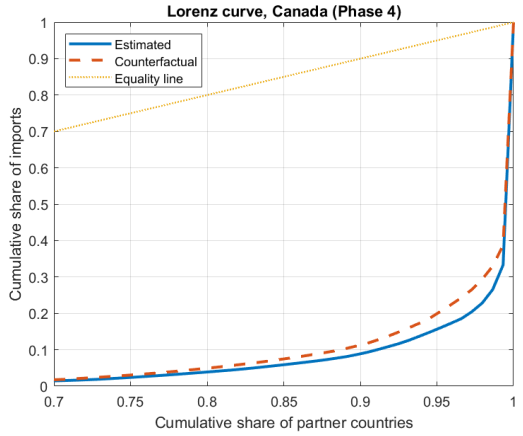
(d) South Korea (Conventional model)



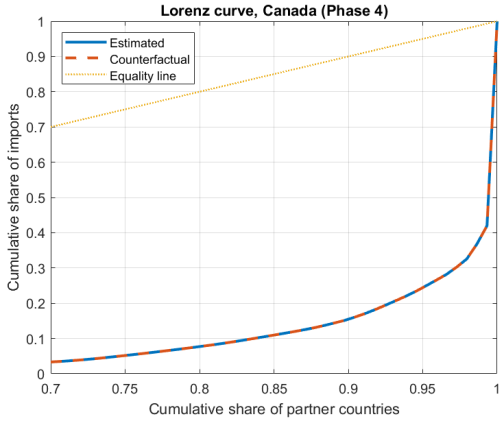
Note: The Gini coefficients from the Lorenz curves are as follows: (i) For Panels (a) and (b), $\text{Gini}(\widehat{s}_{\text{CN},\cdot}) = 0.8601$, $\text{Gini}(\widetilde{s}_{\text{CN},\cdot}) = 0.8516$, $\text{Gini}(\widehat{s}^{\text{con}}_{\text{CN},\cdot}) = 0.8232$, $\text{Gini}(\widetilde{s}^{\text{con}}_{\text{CN},\cdot}) = 0.8250$; (ii) For Panels (c) and (d), $\text{Gini}(\widehat{s}_{\text{KR},\cdot}) = 0.9229$, $\text{Gini}(\widetilde{s}_{\text{KR},\cdot}) = 0.9102$, $\text{Gini}(\widehat{s}^{\text{con}}_{\text{KR},\cdot}) = 0.8827$, $\text{Gini}(\widetilde{s}^{\text{con}}_{\text{KR},\cdot}) = 0.8827$.

Figure 6: Lorenz curve (Category 3: Highly concentrated economies)

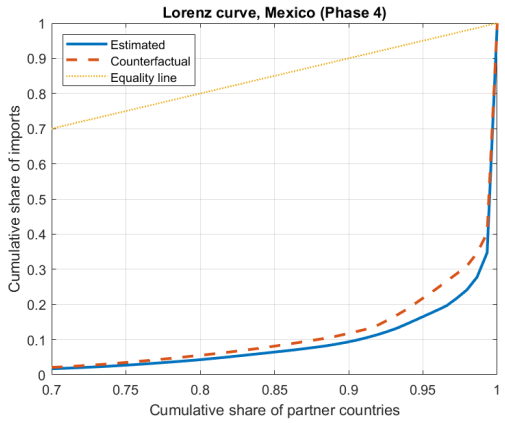
(a) Canada (Our model)



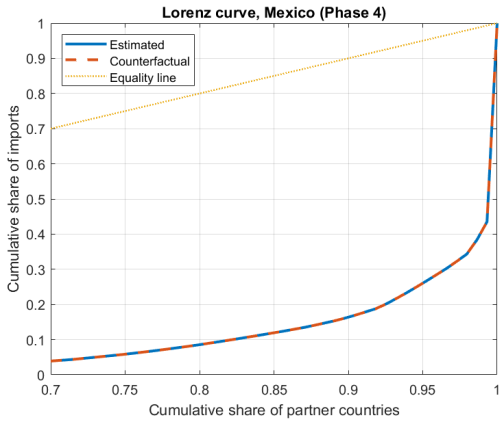
(b) Canada (Conventional model)



(c) Mexico (our model)



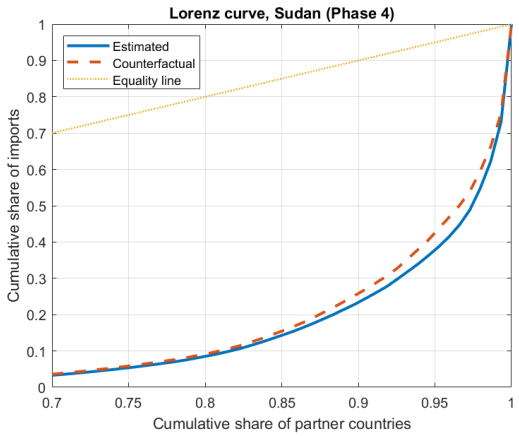
(d) Mexico (Conventional model)



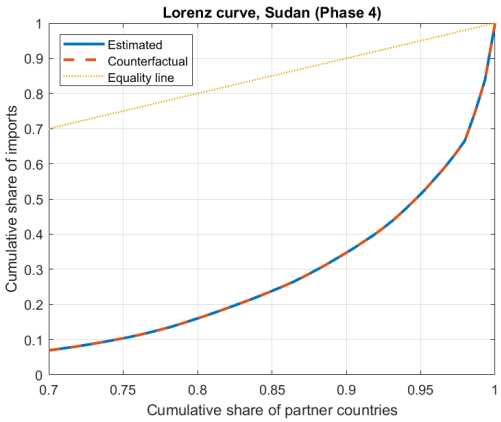
Note: The Gini coefficients from the Lorenz curves are as follows: (i) For Panels (a) and (b), $\text{Gini}(\widehat{s}_{\text{CA},\cdot}) = 0.9399$, $\text{Gini}(\widetilde{s}_{\text{CA},\cdot}) = 0.9260$, $\text{Gini}(\widehat{s}^{\text{con}}_{\text{CA},\cdot}) = 0.9019$, $\text{Gini}(\widetilde{s}^{\text{con}}_{\text{CA},\cdot}) = 0.9019$; (ii) For Panels (c) and (d), $\text{Gini}(\widehat{s}_{\text{MX},\cdot}) = 0.9359$, $\text{Gini}(\widetilde{s}_{\text{MX},\cdot}) = 0.9206$, $\text{Gini}(\widehat{s}^{\text{con}}_{\text{MX},\cdot}) = 0.8945$, $\text{Gini}(\widetilde{s}^{\text{con}}_{\text{MX},\cdot}) = 0.8945$.

Figure 7: Lorenz curve (Category 4: Low-concentration economies)

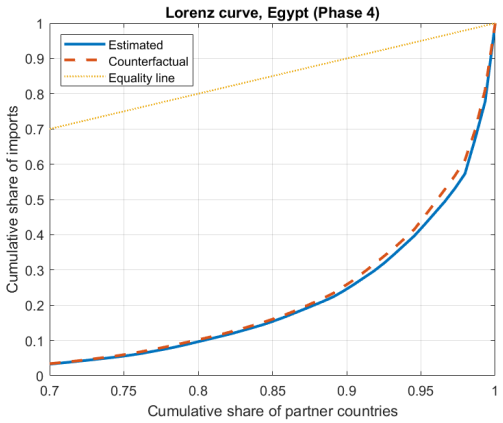
(a) Sudan (Our model)



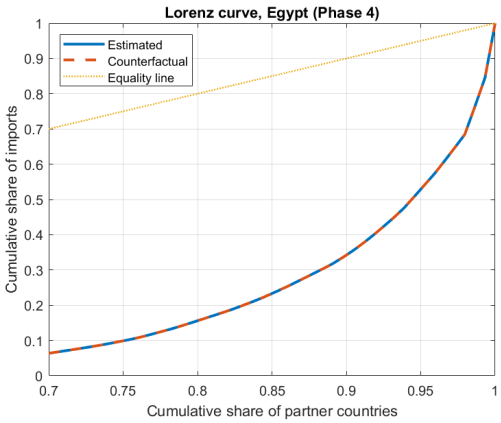
(b) Sudan (Conventional model)



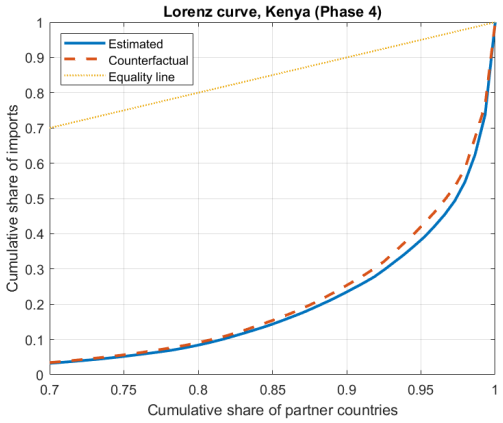
(c) Egypt (our model)



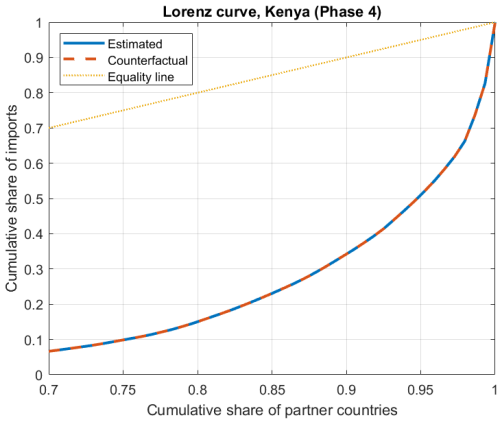
(d) Egypt (Conventional model)



(e) Kenya (our model)



(f) Kenya (Conventional model)



Note: The Gini coefficients from the Lorenz curves are as follows: (i) For Panels (a) and (b), $\text{Gini}(\widehat{s}_{\text{SS},.}) = 0.8660$, $\text{Gini}(\widetilde{s}_{\text{SS},.}) = 0.8544$, $\text{Gini}(\widehat{s}^{\text{con}}_{\text{SS},.}) = 0.8010$, $\text{Gini}(\widetilde{s}^{\text{con}}_{\text{SS},.}) = 0.8010$; (ii) For Panels (c) and (d), $\text{Gini}(\widehat{s}_{\text{EG},.}) = 0.8568$, $\text{Gini}(\widetilde{s}_{\text{EG},.}) = 0.8497$, $\text{Gini}(\widehat{s}^{\text{con}}_{\text{EG},.}) = 0.8029$, $\text{Gini}(\widetilde{s}^{\text{con}}_{\text{EG},.}) = 0.8029$; (iii) For Panels (e) and (f), $\text{Gini}(\widehat{s}_{\text{KE},.}) = 0.8657$, $\text{Gini}(\widetilde{s}_{\text{KE},.}) = 0.8559$, $\text{Gini}(\widehat{s}^{\text{con}}_{\text{KE},.}) = 0.8054$, $\text{Gini}(\widetilde{s}^{\text{con}}_{\text{KE},.}) = 0.8054$.

$\widehat{s}^{\text{con}}_{\text{US,CN}} = 0.0951$ to $\widetilde{s}^{\text{con}}_{\text{US,CN}} = 0.0013$), but the adjustment is concentrated among a small number of large partners, leaving overall concentration nearly unchanged.

For Germany and Japan—countries not directly targeted by the bilateral shock—the conventional gravity counterfactual yields virtually no change in import shares. In sharp contrast, our model predicts sizable reallocations, reflecting the propagation of the US–China shock through the countries' network connectivity. These effects on Japan are greater than those on Germany due to Japan's connectivity to the U.S. and China.

Category 2: Structural-changed economies. China and Korea exhibit particularly strong responses. Both economies initially exhibit a high degree of concentration among a small set of partners. Following the shock, our model predicts substantial diversification, as indicated by declining Gini coefficients. These adjustments are largely absent under conventional gravity.

Category 3: Highly concentrated economies. Canada and Mexico initially rely heavily on the U.S. market. In the counterfactual, the reorganization of U.S. import demand partially unwinds this dependence, leading to lower concentration and greater diversification of import shares.

Category 4: Low-concentration economies. Even for relatively peripheral economies, whose exogenous bilateral trade costs are left unchanged, our model predicts noticeable redistribution of import shares.

4.4 Discussion

Our counterfactual analysis of the U.S.–China trade war highlights that trade policy shocks operate not only through direct bilateral channels but also through the countries' connectivity network. In our framework, a change in bilateral trade costs alters equilibrium trade flows through origin-based, destination-based, and hub-mediated linkages, generating reallocations that extend well beyond the directly targeted country pair. This mechanism produces rich distributional responses in import shares that are largely absent under a conventional gravity model.

First, trade policy evaluation must be network-aware. While the increase in U.S.–China trade costs naturally reshapes the import composition of the two countries involved, our results show that third countries experience substantial reallocations of import shares even when their (exogenous) bilateral trade costs remain unchanged. In particular, hub economies

such as Japan—neither of which is directly targeted by the shock—exhibit sizable changes in import shares under our model (In contrast, conventional gravity predicts little response in import shares.). This contrast implies that ignoring network propagation risks severely understates the scope of policy effects.

Second, hub countries play a systemic role in transmitting trade shocks. Countries with strong network connectivity act as conduits through which bilateral shocks are redistributed across the global system. In the U.S., the decline in China's import share is offset by a broad range of alternative suppliers, leading to a marked decrease in import concentration. This diversification arises precisely because trade flows are jointly determined through network interactions. From a policy perspective, this implies that shocks or interventions affecting highly connected hubs—whether through tariffs, sanctions, or industrial policy—can generate disproportionately large spillovers, both stabilizing and destabilizing, for the rest of the network.

Third, network effects shape the distributional consequences of trade shocks. Our comparison of Lorenz curves shows that the same policy shock can generate distinct concentration responses across countries, depending on their initial positions in their connectivity networks. Economies with initially concentrated import structures, such as Canada and Mexico, experience substantial diversification as dependence on the U.S. market is partially unwound. By contrast, peripheral economies with low initial concentration—such as Sudan, Egypt, and Kenya—also experience noticeable redistribution of import shares, despite facing no direct change in bilateral trade costs. Without an explicit network-based framework, trade policy can reallocate gains and losses among partners in highly difficult-to-anticipate ways.

5 Conclusion

The gravity equation has long served as a foundation for analyzing origin–destination flows, tracing back to Isard (1954) and Tinbergen (1962). Although these early formulations were groundbreaking in highlighting the role of geographic distance in trade, they remained within an exogenous framework. This bilateral focus isolated pairwise effects—such as shared borders, common languages, or trade agreements—but failed to capture how indirect and higher-order connections among countries jointly shaped trade outcomes. Recognizing the growing importance of trade networks, we developed a microfoundation-based specification and econometric framework that endogenized trade costs as a function of network structure, thereby making them both endogenous and interdependent, thereby amplifying heterogeneity across trade pairs.

From a theoretical standpoint, we derived the specification from microfoundations relying on a spatial autoregressive structure that naturally captured multilateral interdependence. Our framework departed from the traditional iceberg-cost assumption by allowing trade costs to reflect network linkages, in which countries leveraged their trade connections as a resource.

Methodologically, we extended the Poisson pseudo maximum likelihood estimator (PPMLE) with accommodating heteroskedasticity- and autocorrelation-robust standard errors to ensure valid inference in the presence of arbitrary correlation in the error structure. To implement PPMLE, we develop a spectral algorithm for computing transformations by the network multiplier matrix. Rather than forming and inverting a large-dimensional matrix for the network multiplier matrix, our algorithm utilizes the eigendecomposition of the row-normalized countries' connectivity matrix to simultaneously diagonalize the three Kronecker-type network operators for cross-destination, cross-origin, and joint origin–destination linkages. As a result, linearly transforming a vector by the network multiplier matrix can be implemented via simple elementwise operations in the eigenbasis. This approach substantially decreases the computational burden compared to the naive inverse-based approach.

Empirically, we identified significant and distinct patterns of network effects across four key phases of global trade. In light of the recent U.S.–China trade war, our counterfactual analysis further demonstrates that a bilateral increase in trade costs induces substantial reallocation of import shares across a wide range of third countries when network interactions are accounted for. These results underscore the importance of viewing the global trading system as an interconnected network rather than a collection of independent bilateral relationships.

As final remarks, we note two possible extensions. First, our framework might be extended to interrelated sectors and countries. As an initial point of view, we focused on aggregate goods rather than sector-specific goods. As Caliendo and Parro (2015) point out, however, sectors are also interrelated. In this case, a possible data structure might be $y_{ip,jq}$ representing flows from country j's sector q to country i's sector p. Second, although our primary focus was on international trade flows, our proposed framework might be broadly applicable to other types of origin-destination flows (commuting or migration flows). In particular, our model would be useful for explaining systemized flows when cross-sectional units are interconnected, and the formation of flows is heavily influenced by their interconnectedness.

Table A.1: Countries List

Countries	Phase 1	2	3	4	Countries	Phase 1	2	3	4	Countries	Phase 1	2	3	4	Countries	Phase 1	2	3	4
Afghanistan	NA	NA	*	*	Dominican Republic	*	*	*	*	Lebanon	NA	*	*	*	Saint Kitts and Nevis	*	*	*	*
Albania	*	*	*	*	Ecuador	*	*	*	*	Liberia	*	*	*	*	Saudi Arabia	*	*	*	*
Algeria	*	*	*	*	Egypt	*	*	*	*	Libya	*	*	*	*	Senegal	*	*	*	*
Angola	*	*	*	*	El Salvador	*	*	*	*	Madagascar	*	*	*	*	Serbia	NA	NA	*	*
Argentina	*	*	*	*	Equatorial Guinea	*	*	*	*	Malawi	*	*	*	*	Seychelles	*	*	*	*
Australia	*	*	*	*	Ethiopia	*	*	*	*	Malaysia	*	*	*	*	Sierra Leone	*	*	*	*
Austria	*	*	*	*	Fiji	*	*	*	*	Maldives	*	*	*	*	Singapore	*	*	*	*
Bahamas	*	*	*	*	Finland	*	*	*	*	Mali	*	*	*	*	Solomon Islands	*	*	*	*
Bahrain	*	*	*	*	France	*	*	*	*	Malta	*	*	*	*	Somalia	*	*	*	*
Bangladesh	*	*	*	*	French Guiana	NA	NA	NA	NA	Mauritania	*	*	*	*	South Africa	*	*	*	*
Barbados	*	*	*	*	Gabon	*	*	*	*	Mauritius	*	*	*	*	South Korea	*	*	*	*
Belgium	NA	NA	*	*	Gambia	*	*	*	*	Mexico	*	*	*	*	Spain	*	*	*	*
Belize	*	*	*	*	Germany	*	*	*	*	Mongolia	*	*	*	NA	Sri Lanka	*	*	*	*
Benin	*	*	*	*	Ghana	*	*	*	*	Morocco	*	*	*	*	Sudan	*	*	*	*
Bermuda	*	*	*	*	Greece	*	*	*	*	Mozambique	*	*	*	*	Suriname	*	*	*	*
Bhutan	*	*	*	*	Greenland	*	*	*	*	Myanmar	*	*	*	*	Sweden	*	*	*	*
Bolivia	*	*	*	*	Guadeloupe	NA	NA	NA	NA	Nepal	*	*	*	*	Switzerland	*	*	*	*
Brazil	*	*	*	*	Guatemala	*	*	*	*	Netherlands	*	*	*	*	Syria	*	*	*	*
Brunei	*	*	*	*	Guinea	*	*	*	*	Netherlands Antilles	NA	NA	NA	NA	Taiwan	NA	NA	NA	NA
Bulgaria	*	*	*	*	Guinea-Bissau	*	*	*	*	New Caledonia	NA	NA	NA	NA	Tanzania	*	*	*	*
Burkina Faso	*	*	*	*	Guyana	*	*	*	*	New Zealand	*	*	*	*	Thailand	*	*	*	*
Burundi	*	*	*	*	Haiti	*	*	*	*	Nicaragua	*	*	*	*	Togo	*	*	*	*
Cote D'Ivoire	*	*	*	*	Honduras	*	*	*	*	Niger	*	*	*	*	Trinidad and Tobago	*	*	*	*
Cambodia	*	*	*	*	Hong Kong	*	*	*	*	Nigeria	*	*	*	*	Tunisia	*	*	*	*
Cameroon	*	*	*	*	Hungary	*	*	*	*	North Korea	NA	NA	NA	NA	Turkey	*	*	*	*
Canada	*	*	*	*	Iceland	*	*	*	*	Norway	*	*	*	*	Turks and Caicos Islands	NA	NA	NA	*
Cayman Islands	NA	NA	*	*	India	*	*	*	*	Oman	*	*	*	*	Uganda	*	*	*	*
Central African Republic	*	*	*	*	Indonesia	*	*	*	*	Pakistan	*	*	*	*	United Arab Emirates	*	*	*	*
Chad	*	*	*	*	Iran	*	*	*	*	Panama	*	*	*	*	United Kingdom	*	*	*	*
Chile	*	*	*	*	Iraq	*	*	*	*	Papua New Guinea	*	*	*	*	United States	*	*	*	*
China	*	*	*	*	Ireland	*	*	*	*	Paraguay	*	*	*	*	Uruguay	*	*	*	*
Colombia	*	*	*	*	Israel	*	*	*	*	Peru	*	*	*	*	Venezuela	NA	NA	NA	NA
Comoros	*	*	*	*	Italy	*	*	*	*	Philippines	*	*	*	*	Vietnam	*	*	*	*
Costa Rica	*	*	*	*	Jamaica	*	*	*	*	Poland	NA	*	*	*	Western Sahara	NA	NA	NA	NA
Cuba	*	*	*	*	Japan	*	*	*	*	Portugal	*	*	*	*	Yemen	NA	*	*	*
Cyprus	*	*	*	*	Jordan	*	*	*	*	Qatar	*	*	*	*	Zambia	*	*	*	*
Czechia	NA	*	*	*	Kenya	*	*	*	*	Reunion	NA	NA	NA	NA	Zimbabwe	*	*	*	*
Democratic Republic of the Congo	*	*	*	*	Kiribati	*	*	*	*	Romania	NA	*	*	*					
Denmark	*	*	*	*	Kuwait	*	*	*	*	Russia	NA	*	*	*					
Djibouti	NA	NA	NA	*	Laos	*	*	*	*	Rwanda	*	*	*	*					

References

Allen, T. and Arkolakis, C. (2022). The welfare effects of transportation infrastructure improvements. *The Review of Economic Studies*, 89(6):2911–2957.

Allen, T., Arkolakis, C., and Takahashi, Y. (2020). Universal gravity. *Journal of Political Economy*, 128(2):393–433.

Anderson, J. (1979). A theoretical foundation for the gravity equation. *The American economic review*, 69(1):106–116.

Anderson, J. and van Wincoop, E. (2003). Gravity with gravitas: a solution to the border puzzle. *American Economic Review*, 93:170–192.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858.

Arkolakis, C., Costinot, A., and Rodríguez-Clare, A. (2012). New trade models, same old gains? *American Economic Review*, 102(1):94–130.

Behrens, K., Ertur, C., and W., K. (2012). 'dual' gravity: using spatial econometrics to control for multilateral resistance. *Journal of Applied Econometrics*, 27:773–794.

Bramoullé, Y., Kranton, R., and D'Amours, M. (2014). Strategic interaction and networks. *American Economic Review*, 104(3):898–930.

Brancaccio, G., Kalouptsidi, M., and Papageorgiou, T. (2020). Geography, transportation, and endogenous trade costs. *Econometrica*, 88(2):657–691.

Caliendo, L. and Parro, F. (2015). Estimates of the trade and welfare effects of nafta. *The Review of Economic Studies*, 82(1 (290)):1–44.

Chaney, T. (2008). Distorted gravity: the intensive and extensive margins of international trade. *American Economic Review*, 98(4):1707–21.

Chen, J. and Roth, J. (2024). Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics*, 139(2):891–936.

Chen, M., Fernandez-Val, I., and Weidner, M. (2021). Nonlinear factor models for network and panel data. *Journal of Econometrics*, 220:296–324.

Chung, F. R. (1997). *Spectral graph theory*, volume 92. American Mathematical Soc.

Cliff, A. D. and Ord, J. (1995). *Spatial Autocorrelation*. Pion Ltd, London.

Conley, T., Gonçalves, S., Kim, M., and Peron, B. (2023). Bootstrap inference under cross-sectional dependence. *Quantitative Economics*, 14:511–569.

de Jong, R. M. and Davidson, J. (2000). Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica*, 68(2):407–423.

Eaton, J. and Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, 70(5):1741–1779.

Feenstra, R. C., Lipsey, R. E., Deng, H., Ma, A. C., and Mo, H. (2005). World trade flows: 1962-2000. Working Paper 11040, National Bureau of Economic Research.

Fernandez-Val, I. and Weidner, M. (2016). Individual and time effects in nonlinear panel models with large n, t. *Journal of Econometrics*, 192:291–312.

Fernandez-Val, I. and Weidner, M. (2018). Fixed effects estimation of large-t panel data models. *Annual Review of Economics*, 10:109–138.

Fuchs, S. and Wong, W. F. (2024). Multimodal transport networks. *Available at SSRN*.

Ganapati, S., Wong, W. F., and Ziv, O. (2024). Entrepôt: Hubs, scale, and trade costs. *American Economic Journal: Macroeconomics*, 16(4):239–78.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: applications to poisson models. *Econometrica*, 52:701–720.

Head, K. and Mayer, T. (2014). Chapter 3 - gravity equations: Workhorse,toolkit, and cookbook. In Gopinath, G., Helpman, E., and Rogoff, K., editors, *Handbook of International Economics*, volume 4 of *Handbook of International Economics*, pages 131–195. Elsevier.

Helpman, E. and Krugman, P. (1985). *Market Structure and Foreign Trade*. Cambridge, MA: The MIT Press.

Helpman, E., Melitz, M., and Rubinstein, Y. (2008). Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics*, 123(2):441–487.

Isard, W. (1954). Location theory and trade theory: Short-run analysis. *Quarterly Journal of Economics*, 68(2):305–320.

Jenish, N. and Prucha, I. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics*, 150:86–98.

Jenish, N. and Prucha, I. (2012). On spatial processes and asymptotic inference under near-epoch dependence. *Journal of Econometrics*, 170:178–190.

Jeong, H. and Lee, L. (2024). Maximum likelihood estimation of a spatial autoregressive model for origin-destination flow variables. *Journal of Econometrics*, 242:105790.

Jeong, H., Lin, Y., and Lee, L. (2023). Estimation of spatial autoregressive models for origin-destination flows: A partial likelihood approach. *Economics Letters*, 229:111202.

Jin, F., Lee, L., and Yu, J. (2023). Estimating flow data models of international trade: Dual gravity and spatial interactions. *Econometric Reviews*, 42:157–194.

Kapetanios, G., Serlenga, L., and Shin, Y. (2021). Estimation and inference for multi-dimensional heterogeneous panel datasets with hierarchical multi-factor error structure. *Journal of Econometrics*, 220(2):504–531. Annals Issue: Celebrating 40 Years of Panel Data Analysis: Past, Present and Future.

Kelejian, H. and Prucha, I. (2007). HAC estimation in a spatial framework. *Journal of Econometrics*, 140:131–154.

Kim, M. and Sun, Y. (2011). Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix. *Journal of Econometrics*, 160:349–371.

Kwon, O., Yoon, J., and Yotov, Y. V. (2025). A generalized poisson-pseudo maximum likelihood estimator*. *Journal of Business & Economic Statistics*, 0(ja):1–27.

Lee, L. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models. *Econometrica*, 72:1899–1925.

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

Lee, L. and Yu, J. (2010). A spatial dynamic panel data model with both time and individual fixed effects. *Econometric Theory*, 26:564–597.

LeSage, J. and Fischer, M. (2010). *Spatial econometric methods for modeling origin-destination flows*. In: Fischer M., Getis A. (eds) Handbook of Applied Spatial Analysis. Springer, Berlin, Heidelberg.

LeSage, J. and Pace, R. (2008). Spatial econometric modeling of origin-destination flows. *Journal of Regional Science*, 48:941–967.

Lin, X. and Lee, L.-f. (2010). GMM estimation of spatial autoregressive models with unknown heteroskedasticity. *Journal of Econometrics*, 157(1):34–52.

Lind, N. and Ramondo, N. (2023). Trade with correlation. *American Economic Review*, 113(2):317–53.

McCallum, J. (1995). National borders matter: Canada-U.S. regional trade patterns. *American Economic Review*, 85:615–623.

McFadden, D. (1972). Conditional logit analysis of qualitative choice behavior.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6):1695–1725.

Morales, E., Sheu, G., and Zahler, A. (2019). Extended gravity. *The Review of Economic Studies*, 86(6):2668–2712.

Mullahy, J. and Norton, E. (2024). Why transform Y? A the pitfalls of transformed regressions with a mass at zero. *Oxford Bulletin of Economics and Statistics*, 86:417 – 447.

Nagengast, A. J. and Yotov, Y. V. (2025). Staggered difference-in-differences in gravity settings: Revisiting the effects of trade agreements. *American Economic Journal: Applied Economics*, 17(1):271–96.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Ord, J. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70:120–126.

Pesaran, M. H. and Yang, C. F. (2020). Econometric analysis of production networks with dominant units. *Journal of Econometrics*, 219(2):507–541. Annals Issue: Econometric Estimation and Testing: Essays in Honour of Maxwell King.

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

Qu, X. and Lee, L. (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *Journal of Econometrics*, 184:209–232.

Samuelson, P. (1952). The transfer problem and transport costs: The terms of trade when impediments are absent. *Economic Journal*, pages 265–289.

Samuelson, P. (1954). The transfer problem and trasponrt costs ii: Analysis of effects of trade impediments. *Economic Journal*, pages 265–289.

Santos Silva, J. and Tenreyro, S. (2006). The log of gravity. *Review of Economics and Statistics*, 88:641–658.

Santos Silva, J. and Tenreyro, S. (2022). The log of gravity at 15. *Portuguese Economic Journal*, 21:423–437.

Tinbergen, J. (1962). *Shaping the World Economy*. Net York: The Twentieth Century Fund.

Weidner, M. and Zylkin, T. (2021). Bias and consistency in three-way gravity models. *Journal of International Economics*, 132:103513.

Wong, W. F. (2022). The round trip effect: Endogenous transport costs and international trade. *American Economic Journal: Applied Economics*, 14(4):127–66.

Xu, X. and Lee, L. (2015a). Maximum likelihood estimation of a spatial autoregressive Tobit model. *Journal of Econometrics*, 188:264–280.

Xu, X. and Lee, L. (2015b). A spatial autoregressive model with a nonlinear transformation of the dependent variable. *Journal of Econometrics*, 186:1–18.

Xu, X. and Lee, L. (2018). Sieve maximum likelihood estimation of the spatial autoregressive Tobit model. *Journal of Econometrics*, 203:96–112.

Yotov, Y. V. (2022). On the role of domestic trade flows for estimating the gravity model of trade. *Contemporary Economic Policy*, 40(3):526–540.