Multilateral Resistance-Guided Graph Message Passing for Trade Flow Prediction

Quansi Li
Xiangtan University
Xiangtan, Hunan 411105, China
342419420@qq.com

Haijing Yu
Xiangtan University
Xiangtan, Hunan 411105, China
yuhajing1513875@163.com

Abstract

Trade flows are a central variable for studying international trade networks and the propagation of policy shocks. Traditional economic models provide theory-grounded representations of bilateral trade flows, but often rely on fixed functional forms, have limited ability to incorporate high-dimensional dyadic features, and do not directly capture trade inertia. Graph neural networks are naturally compatible with networked trade data, yet their aggregation rules are usually weakly interpretable from an economic perspective. We propose MRTGNN, a theory-guided graph neural network whose message-passing layer is motivated by the multilateral-resistance (MRT) fixed-point equations of structural gravity. Theoretically, we give a conditional finite-depth error bound for decoded neural approximations to MRT updates on a compact domain. Empirically, on a panel of 176 countries from 2002 to 2019, ten matched runs show modest, consistent RMSE gains over strong temporal and mechanism-matched baselines, while the Spearman difference from the temporal MLP baseline is not statistically significant. Under China-related tariff shocks, MRTGNN produces a Network Spillover Ratio one order of magnitude higher than a generic graph-attention baseline. Additional evaluations cover forecasts through 2024, sector networks, and gravity-benchmark alignment. These results support theory-guided prediction and interpretable network-sensitivity analysis; the shock responses are predictive sensitivities rather than causal or general-equilibrium effects.

1 Introduction

One central challenge in trade-flow prediction is to build a model that captures the key factors shaping country-level trade flows and describes how exogenous shocks propagate through the global trade network. The classical gravity model, one of the most widely used empirical tools in international trade, explains trade flows through bilateral factors such as exporter size, importer demand, and bilateral trade costs [1, 2, 3]. The structural gravity model of Anderson and van Wincoop extends this bilateral view to a multilateral one by introducing multilateral resistance terms, which are typically controlled by fixed effects or solved through iterative general-equilibrium procedures [4, 5, 6, 7, 8]. However, accurate and mechanism-grounded trade-flow prediction in large-scale, high-dimensional, and dynamic global trade networks remains an open problem.

Graph neural networks have recently been used for trade-flow prediction, supply-chain modeling, and other economic network tasks [9, 10, 11]. Generic architectures such as GCN, GraphSAGE, GAT, GATv2, and temporal GNNs provide flexible mechanisms for neighborhood aggregation and dynamic prediction [12, 13, 14, 15, 16]. However, their message-passing rules are usually selected for expressive power rather than derived from economic theory; consequently, the learned aggregation has limited ex ante interpretability. Recent work has begun to introduce gravity priors into

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

neural trade models. Minakawa et al. [17] incorporate a gravity prior into a graph auto-encoder and formulate bilateral trade-flow prediction as graph edge-weight prediction. *Although these models combine graph structure with economic priors, how to formally connect neural aggregation to the core mechanism of structural gravity remains unclear.*

We propose MRTGNN to address this gap. The model has two synergistic components. First, we start from the multilateral-resistance fixed-point equations in structural gravity and embed their structure into graph message passing through three mechanisms: a trade-cost gate, edge-augmented messages, and attention weights constructed from asymmetric message-source and message-destination projections. These mechanisms organize learnable partner aggregation according to economic roles specified before training. Second, we introduce temporal edge features and a bilateral residual pathway. Temporal edge features capture trade inertia and sunk-cost effects, including lagged trade, trade growth, tariff changes, and quantity changes. The bilateral residual pathway preserves direct dyadic gravity factors from raw country features, allowing the final prediction to use both the MRT network path and the direct bilateral path.

We evaluate MRTGNN on a global bilateral trade panel covering 176 countries from 2002 to 2019 against representative baselines spanning gravity models, tree models, MLPs, static GNNs, temporal GNNs, and graph-attention models. MRTGNN+T+R achieves competitive predictive performance, with modest but consistent RMSE gains in ten matched runs; T and R denote temporal edge features and the bilateral residual pathway. The component comparisons support combining temporal features, MRT-guided aggregation, and the bilateral residual pathway, while the mechanism-matched comparison evaluates the joint propagation design. Counterfactual shock experiments further show that MRTGNN produces substantially stronger spillovers under continuous trade-cost perturbations, especially in the China tariff-shock scenario, with additional analyses examining responses across countries and periods.

2 Related Work

Structural gravity and quantitative trade. The classical gravity equation [1, 2, 3] has been microfounded by structural gravity, which derives bilateral exports from constant elasticity of substitution (CES) preferences and equilibrium multilateral-resistance terms [4, 5, 6, 7, 8]. Quantitative trade models build on this foundation to perform counterfactual general-equilibrium analysis of tariffs and welfare [18, 19, 20, 21, 22]. Estimation typically relies on fixed effects or iterative general-equilibrium (GE) solvers; the resulting models are interpretable but use fixed functional forms and do not naturally exploit high-dimensional dyadic features. Our work keeps the equilibrium-motivated role of multilateral resistance, but recasts its fixed-point computation as a learnable graph message-passing operator for trade-flow prediction and counterfactual spillover analysis.

Machine learning for trade prediction. A growing literature applies machine learning, including knowledge graphs and graph neural networks, to bilateral trade prediction and supply-chain modeling [9, 10, 11]. Most generic message-passing architectures, including GCN, GraphSAGE, GAT, GATv2, GIN, and temporal graph models such as DySAT, are designed for general graph tasks rather than trade-specific structure [12, 13, 14, 15, 23, 16]. Minakawa et al. [17] introduce a gravity prior into a graph auto-encoder for bilateral trade-flow prediction, but combine the prior with graph aggregation rather than deriving the aggregation rule from the structural-gravity equations. MRTGNN uses the multilateral-resistance updates to organize edge gating, asymmetric message-source/destination projections, and aggregation, assigning economic roles to the propagation components before training.

Theory-guided and equilibrium-based neural networks. Beyond the trade domain, several lines of work embed mechanistic structure into neural networks. Deep equilibrium models compute representations as fixed points of a parametric operator [24], providing context for our analysis of finite-depth MRT-style updates. Generic graph message-passing frameworks [25, 26, 27] provide the architectural language we instantiate. More broadly, explicit geometric structure and comparison-based supervision have been used to shape representations for new intent discovery [28] and transferable personality modeling [29], respectively. We complement these approaches by deriving the message-passing form from a domain-specific economic fixed-point system and by providing a conditional finite-depth analysis of decoded neural approximations to multilateral-resistance updates.

3 Preliminaries

We first formalize the directed bilateral trade network and the prediction task, then summarize the structural-gravity multilateral-resistance (MRT) fixed-point system that motivates our model. Extended derivations and economic intuition are deferred to Appendix A.

3.1 Directed Bilateral Trade Network and Prediction Task

We consider N countries indexed by $\mathcal{V} = \{1, \dots, N\}$. For each year t, the global trade system is represented as a directed graph $\mathcal{G}^t = (\mathcal{V}, \mathcal{E}^t)$, where each directed edge $(i, j) \in \mathcal{E}^t$ represents exports from country i to country j. Each country i has a node feature vector $\mathbf{x}_i^t \in \mathbb{R}^{d_x}$ encoding macro variables (GDP, population, imports, exports), and each directed edge has a feature vector $\mathbf{e}_{ij}^t \in \mathbb{R}^{d_e}$ encoding trade-cost and dynamic variables (tariffs, FTAs, distance, lagged trade, trade growth). Let X_{ij}^{t+1} denote exports from i to j in year $t+1$. The prediction target is the log-transformed flow $y_{ij}^{t+1} = \log(1 + X_{ij}^{t+1})$, and the model takes the form $\hat{y}_{ij}^{t+1} = f_\theta(\mathcal{G}^t, \mathbf{x}_i^t, \mathbf{x}_j^t, \mathbf{e}_{ij}^t)$, trained by minimizing squared error on $\mathcal{D}_{\text{train}}$.

3.2 Structural Gravity and the Multilateral-Resistance Fixed Point

The structural gravity model writes bilateral exports as $X_{ij} = (Y_i E_j / Y_W)(t_{ij}/(P_i \Pi_j))^{1-\sigma}$, where Y_i is exporter output, E_j is importer expenditure, Y_W is world output, $t_{ij} > 0$ is bilateral trade cost, $\sigma > 1$ is the elasticity of substitution, and P_i, Π_j are outward and inward multilateral-resistance terms. MRTs are not exogenous: they solve the fixed-point system

$$P_i^{1-\sigma} = \sum_j \left(\frac{t_{ij}}{\Pi_j}\right)^{1-\sigma} \frac{E_j}{Y_W}, \qquad \Pi_j^{1-\sigma} = \sum_i \left(\frac{t_{ij}}{P_i}\right)^{1-\sigma} \frac{Y_i}{Y_W}. \tag{1}$$

These equations show that each country's MRT depends on all partners' resistance states and bilateral trade costs. Let $p_i = \log P_i$, $\pi_j = \log \Pi_j$, and $\mathbf{z} = (p_1, \dots, p_N, \pi_1, \dots, \pi_N)^\top$. Section 5 defines the full-support log-MRT update and a fixed normalization. Edge-dependent trade costs, directional partner aggregation, and repeated updating provide the structural motivation for MRTGNN.

3.3 Graph Message Passing as MRT Iteration

Graph neural networks update node representations by aggregating neighbor messages [25, 26, 27]: $\mathbf{h}_i^{(\ell+1)} = \mathrm{AGG}(\{\mathrm{MSG}(\mathbf{h}_j^{(\ell)}, \mathbf{e}_{ij}) : j \in \mathcal{N}(i)\})$; MRT iteration repeatedly aggregates information from all trade partners. This correspondence motivates our method: we translate MRT fixed-point aggregation into a learnable message-passing layer that updates country representations through economically structured partner aggregation.

4 Method

We formulate trade-flow prediction as a theory-guided graph learning problem, then introduce the MRTGNN layer (Figure 1) and the prediction head with bilateral residual pathway. Detailed component-by-component motivation, the explicit Log-Sum-Exp connection, and additional design discussions are given in Appendix B.

4.1 MRTGNN: A Theory-Guided Message-Passing Layer

Given the layer-ℓ country embedding $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^d$ and directed edge feature $\mathbf{e}_{ij}$, MRTGNN updates node i as

$$\mathbf{h}_i^{(\ell+1)} = \varphi\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(\ell)}\, g(\mathbf{e}_{ij}) \left(\mathbf{W}_V \mathbf{h}_j^{(\ell)} + \mathbf{W}_E \mathbf{e}_{ij}\right)\right), \tag{2}$$

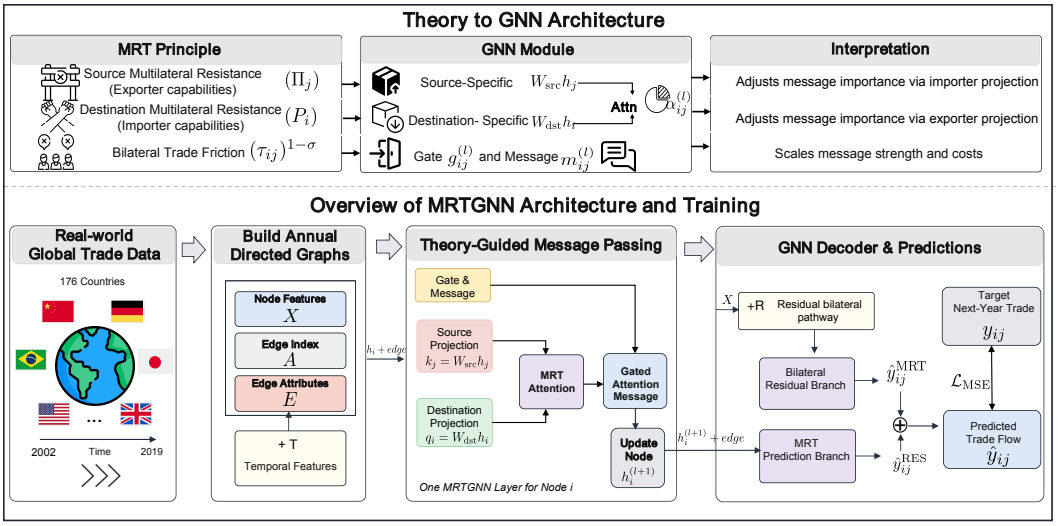

Figure 1: Architecture of MRTGNN. The top panel shows how structural-gravity components are mapped to GNN operations, where multilateral resistance terms are modeled by asymmetric attention-based node projections and bilateral trade friction is modeled by edge-conditioned gates and messages. The bottom panel shows the full prediction pipeline: yearly directed trade graphs over 176 countries are processed by stacked MRTGNN layers, and an MRT branch is combined with a bilateral residual branch to predict trade flows.

where $\varphi(\cdot)$ is a nonlinearity and $\alpha_{ij}^{(\ell)}$ is an attention weight. The vector $\mathbf{h}_i^{(\ell)}$ is a learned high-dimensional trade-network representation motivated by outward and inward MRT roles. MRTGNN organizes three familiar message-passing mechanisms around this economic structure.

Trade-cost gate. The structural factor $t_{ij}^{1-\sigma}$ motivates an edge-dependent gate $g(\mathbf{e}_{ij}) = \text{sigmoid}(\mathbf{W}_g\mathbf{e}_{ij})$. The gate modulates message magnitude using bilateral features, while attention normalizes relative influence across partners.

Asymmetric message-source/destination projections. International trade is directional, while MRT aggregation updates a focal country from partner information. Following attention-based message passing [14, 15], we use distinct projections to compute

$$\eta_{ij}^{(\ell)} = \mathbf{a}^\top \text{LeakyReLU}\big(\mathbf{W}_{\text{src}}\mathbf{h}_j^{(\ell)} + \mathbf{W}_{\text{dst}}\mathbf{h}_i^{(\ell)} + \mathbf{W}_{\text{edge}}\mathbf{e}_{ij}\big), \qquad \alpha_{ij}^{(\ell)} = \text{softmax}_j(\eta_{ij}^{(\ell)}). \quad (3)$$

Here j is the message source, or trade partner, and i is the message destination, or focal country. These roles refer to message direction rather than exporter/importer direction. The theoretical outward and inward updates below use distinct row and column sums.

Edge-augmented messages. The message content $\mathbf{m}_{ij}^{(\ell)} = \mathbf{W}_V\mathbf{h}_j^{(\ell)} + \mathbf{W}_E\mathbf{e}_{ij}$ lets bilateral trade-cost and dynamic features influence both *which* partners are weighted and *what* they transmit, paralleling the dual role of t_{ij} in Eq. (1).

Connection to the Log-Sum-Exp MRT operator. Before gauge normalization, the outward update is $[\mathcal{U}(\mathbf{z})]_i = \frac{1}{1-\sigma}\log\sum_j \tau_{ij}q_j$, where $\tau_{ij} = t_{ij}^{1-\sigma}$ and $q_j = \exp((\sigma-1)\pi_j)s_j$. The relative partner allocation is $q_j/\sum_k q_k$. This motivates normalized attention and a separate bilateral gate, but recovering the full Log-Sum-Exp update also requires its unnormalized scale and outer transformation. Table 1 summarizes these design roles. Representability of the complete update is an explicit assumption of the conditional analysis below.

Table 1: Design correspondences between structural gravity and MRTGNN.

Structural gravity component	**MRTGNN component**
Trade-cost weight $t_{ij}^{1-\sigma}$	Trade-cost gate $g(\mathbf{e}_{ij})$
Normalized partner-share allocation $q_j / \sum_k q_k$	Attention weight $\alpha_{ij}^{(\ell)}$
Directional MRT dependencies	Asymmetric message projections $(\mathbf{W}_{\mathrm{src}}, \mathbf{W}_{\mathrm{dst}})$
Bilateral trade-cost features	Edge-augmented message $\mathbf{W}_E \mathbf{e}_{ij}$
Repeated MRT-style updating	Stacked message-passing layers
Outer continuous transformation	Idealized decoded readout in the analysis

4.2 Inference Framework

Dynamic full-graph aggregation. We use $\mathcal{N}(i) = \mathcal{V} \setminus \{i\}$ for international message passing. The full structural-gravity operator analyzed in Section 5 also includes domestic terms. We further include temporal edge features, including lagged trade, trade growth, tariff changes, lagged quantity, and quantity growth, so that aggregation depends on static bilateral conditions, trade persistence, and recent trade dynamics. In the full model MRTGNN+T+R, T denotes these temporal edge features.

MRT-informed prediction head. After L MRTGNN layers, we form the MRT-path prediction as $\hat{y}_{ij}^{\mathrm{MRT}} = \mathrm{MLP}_{\mathrm{edge}}([\mathbf{h}_i^{(L)} \| \mathbf{h}_j^{(L)} \| \mathbf{e}_{ij}])$. We then add a bilateral residual pathway $\hat{y}_{ij}^{\mathrm{res}} = \mathrm{MLP}_{\mathrm{res}}([\mathbf{x}_i^t \| \mathbf{x}_j^t])$ to preserve a direct gravity channel from raw country features. This pathway is useful when network effects are weak or when graph aggregation smooths out local dyadic signals. The final prediction is $\hat{y}_{ij} = \hat{y}_{ij}^{\mathrm{MRT}} + \hat{y}_{ij}^{\mathrm{res}}$. The residual is applied in the prediction head.

Training and counterfactual inference. The model is trained by minimizing the mean squared prediction error. Once trained, MRTGNN supports next-year trade-flow prediction and predictive counterfactual analysis by perturbing node or edge features at inference time, including tariffs, embargo-style edge-feature masking, and country-level GDP features. A single forward pass yields direct and higher-order prediction changes with the remaining macroeconomic state held fixed. Full counterfactual protocols are given in Appendix G.

5 Theoretical Analysis

We formalize a conditional connection between neural updates and MRT iteration, complementing the neural fixed-point perspective [24]. The one-step statement assumes an idealized expressive decoded operator class; the finite-depth bound concerns compositions of those decoded operators. Full conditions and proofs are in Appendix C.

5.1 MRT Fixed-Point Operator

Let $\tau_{ij} = t_{ij}^{1-\sigma}$, $s_j = E_j / Y_W$, $r_i = Y_i / Y_W$, and $\mathbf{z} = (p_1, \ldots, p_N, \pi_1, \ldots, \pi_N)^\top$. The full-support raw log-MRT update is

$$[\mathcal{U}(\mathbf{z})]_i = \tfrac{1}{1-\sigma} \log \sum_{j=1}^{N} \tau_{ij} \exp((\sigma-1) z_{N+j}) s_j, \quad [\mathcal{U}(\mathbf{z})]_{N+j} = \tfrac{1}{1-\sigma} \log \sum_{i=1}^{N} \tau_{ij} \exp((\sigma-1) z_i) r_i. \tag{4}$$

These sums include domestic terms, as in Eq. (1). For the analysis, set $\mathcal{T} = \mathcal{R} \circ \mathcal{U}$, where $\mathcal{R}$ fixes $p_1 = 0$ (Appendix A). This gauge fixes the scale convention.

5.2 Assumptions

We use six assumptions, detailed in Appendix C: (A1)–(A2) positive, uniformly bounded trade costs and economic shares; (A3) a chosen compact normalized domain $\mathcal{K}$; (A4) a Lipschitz constant K for $\mathcal{T}$ on $\mathcal{K}$; (A5) exact and approximate trajectories remain in $\mathcal{K}$ for the finite horizon considered; and (A6) uniform density of an idealized decoded operator class in the continuous maps on $\mathcal{K}$. A6 specifies the idealized decoded expressivity used in Proposition 1.

5.3 Approximation Results

Proposition 1 (Conditional one-step representation)**.** *Under (A1), (A2), (A3), and (A6), for every* $\epsilon > 0$ *there exists a decoded map* $\widetilde{\mathcal{T}} = \psi \circ F_\theta \circ \phi$ *in the idealized class such that* $\sup_{\mathbf{z}\in\mathcal{K}} \|\widetilde{\mathcal{T}}(\mathbf{z}) - \mathcal{T}(\mathbf{z})\| \le \epsilon$.

Sketch. Positive bounded weights make the log-sum updates continuous on the compact domain. The gauge map is continuous, and A6 supplies the stated approximation.

Theorem 1 (Finite-depth decoded-update error)**.** *Under (A3)–(A5), suppose decoded updates* $\widetilde{\mathcal{T}}_\ell$ *satisfy* $\sup_{\mathbf{u}\in\mathcal{K}} \|\widetilde{\mathcal{T}}_\ell(\mathbf{u}) - \mathcal{T}(\mathbf{u})\| \le \epsilon_\ell$. *For an initial state* $\mathbf{z} \in \mathcal{K}$ *whose trajectories satisfy A5,*

$$\left\|\widetilde{\mathcal{T}}_L \circ \cdots \circ \widetilde{\mathcal{T}}_1(\mathbf{z}) - \mathcal{T}^L(\mathbf{z})\right\| \le \sum_{\ell=1}^{L} K^{L-\ell}\epsilon_\ell. \tag{5}$$

Sketch. On the stated trajectories, the triangle inequality and A4 give $E_\ell \le \epsilon_\ell + KE_{\ell-1}$, with $E_0 = 0$.

The bound concerns decoded compositions. An uninterrupted latent-layer stack inherits it under the encoding/readout compatibility condition in Appendix C. If $K < 1$ and $\epsilon_\ell \le \epsilon$ uniformly, its right-hand side is at most $\epsilon(1 - K^L)/(1 - K)$.

6 Experiments

We evaluate MRTGNN on next-year bilateral trade-flow prediction and on counterfactual shock propagation, organized around three questions: **Q1** (Predictive performance) does MRTGNN improve over strong econometric, tabular, generic GNN, and temporal GNN baselines? **Q2** (Component contribution) which components drive the improvement, beyond temporal features or model capacity? **Q3** (Counterfactual spillover) does the model propagate trade shocks through the network in a way pairwise models and generic GNNs do not? Predictive accuracy alone is not sufficient: a model can fit historical trade well from lagged dyadic features yet fail to propagate policy shocks.

6.1 Evaluation Setup

Data and split. We construct a directed country-level trade panel covering 176 countries from 2002 to 2019. Yearly snapshots are directed graphs with country-year macroeconomic node features, dyad-year trade-cost edge features, time-invariant country attributes, and next-year trade-flow targets. The raw 2002–2019 panel contains 554,400 directed dyad-year records. The supervised dataset uses feature years 2002–2018 and contains 523,600 observations with 23 features per observation; the target is $\log(1 + \text{trade}_{t+1})$. We split by feature year: train 2002–2013, validation 2014–2015, test 2016–2018. Continental coverage appears in Figure 2; sources [30, 31, 32] and details are in Appendix D.

Baselines and metrics. We compare MRTGNN+T+R against econometric, tabular, static GNN, attention GNN, and temporal GNN families [33, 34, 12, 13, 23, 14, 15, 16]. Tables 2 and 4 retain the original three-seed cross-family and component evaluations. The reliability assessment uses a separate set of ten matched runs, excluding the original three seeds (Table 3); these estimates supersede the corresponding three-seed estimates for inference. We report RMSE, MAE, R^2, and Spearman correlation, with configurations in Appendix E.

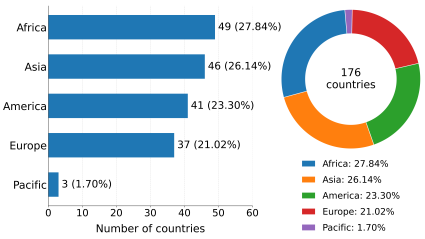

Figure 2: Continental coverage of the 176-country panel.

6.2 Main Predictive Results

Table 2 reports competitive performance across the original baseline suite on the 2016–2018 test set. Absolute gains are modest because trade is highly persistent (Appendix D). The separate ten-run evaluation provides the basis for our reliability claims.

Table 2: Original cross-family comparison on feature-year test set 2016–2018. Mean ± std over three seeds; the focal ten-run comparison is in Table 3. GNNs share the data and evaluation protocol; residual pathways and temporal wrappers differ as indicated by the model variants. Gray row: our model. Bold: best per column.

Type	**Model**	RMSE ↓	R^2 ↑	Spearman ↑
Econometric	Gravity	1.4639 ± 0.0000	0.9092 ± 0.0000	0.8913 ± 0.0000
Tree	RF+T	1.4071 ± 0.0002	0.9161 ± 0.0000	0.9025 ± 0.0000
Tree	GBDT+T	1.3978 ± 0.0006	0.9172 ± 0.0001	0.9038 ± 0.0001
Neural	MLP	1.4469 ± 0.0008	0.9113 ± 0.0001	0.8985 ± 0.0003
Neural	MLP+T	1.3913 ± 0.0014	0.9180 ± 0.0002	0.9036 ± 0.0005
Static GNN	GCN+T	1.4104 ± 0.0057	0.9157 ± 0.0007	0.8998 ± 0.0007
Static GNN	SAGE+T	1.3868 ± 0.0006	0.9185 ± 0.0001	0.9030 ± 0.0002
Static GNN	GIN+T	1.5539 ± 0.0116	0.8977 ± 0.0015	0.8959 ± 0.0013
Static GNN	GAT+T (v1)	1.4148 ± 0.0182	0.9151 ± 0.0022	0.9009 ± 0.0011
Temporal GNN	GCN-GRU	1.9383 ± 0.4176	0.8334 ± 0.0728	0.8769 ± 0.0161
Temporal GNN	GAT-GRU	1.4607 ± 0.0053	0.9096 ± 0.0007	0.8951 ± 0.0006
Temporal GNN	DySAT	1.4480 ± 0.0140	0.9111 ± 0.0017	0.8981 ± 0.0009
Ours	MRTGNN+T	1.3840 ± 0.0062	0.9188 ± 0.0007	0.9029 ± 0.0003
Ours	MRTGNN+T+R	$\mathbf{1.3773 \pm 0.0024}$	$\mathbf{0.9196 \pm 0.0003}$	$\mathbf{0.9049 \pm 0.0002}$

Table 3: Focal comparison over ten new matched seeds, excluding the original three, on feature-year test set 2016–2018. RMSE is mean ± standard deviation. Δ is baseline minus MRTGNN RMSE; positive values favor MRTGNN. CI: paired 95% confidence interval. GATv2+T is descriptive; GATv2+T+R is the mechanism-matched control.

Model	RMSE	Δ	95% CI	Wins
MLP+T	1.3921 ± 0.0023	0.00890	$[0.00625, 0.01156]$	10/10
GATv2+T+R	1.3926 ± 0.0065	0.00942	$[0.00349, 0.01535]$	9/10
GATv2+T	1.3874 ± 0.0059	0.00420	—	—
MRTGNN+T+R	1.3832 ± 0.0034	—	—	—

Statistical significance. Ten matched runs give RMSE 1.3832 ± 0.0034 for MRTGNN+T+R, versus 1.3921 ± 0.0023 for MLP+T and 1.3926 ± 0.0065 for the mechanism-matched GATv2+T+R. The paired improvements have positive 95% confidence intervals, with wins in 10/10 and 9/10 runs, respectively (Table 3). Against MLP+T, R^2 improves in all ten runs, while its Spearman difference is not statistically significant. Appendix I.1 reports additional uncertainty estimates. Frozen 2020–2024 forecasts, sector and country-group analyses, and allocation diagnostics are summarized in Appendix I; the latter show closer gravity-benchmark alignment for MRTGNN than matched GATv2 (Spearman 0.500 versus 0.300).

6.3 Ablation: Component Contribution

Table 4 retains the original component comparison. MLP improves over Gravity, and temporal features substantially improve MLP, supporting the use of nonlinear interactions and trade persistence. MLP+T has the lowest MAE in this comparison. MRTGNN+T+R improves on MRTGNN+T in this evaluation, supporting a direct bilateral residual pathway alongside network aggregation. The ten-seed comparison against GATv2+T+R (Table 3) is the focal operator comparison: both models include temporal features and the residual pathway. Their total parameter counts are close (55,114 versus 53,873), while MRTGNN exhibits a modest, consistent RMSE gain. Further discussion is in Appendix F.

6.4 Counterfactual Spillover Analysis

We evaluate predictive network sensitivities by modifying inputs at inference time while holding the remaining macroeconomic state fixed. We consider three shocks: (i) Russia trade embargo, by masking edge features on Russia–Western trade links with a 24-country sanctions coalition; (ii) China-related tariff shock, by tripling tariffs on China-related edges involving 12 major economies (Ap-

Table 4: Original component-comparison ablation on feature-year test set 2016–2018. Mean ± std over three seeds. Lower RMSE/MAE is better; higher R^2/Spearman is better. Gray rows: MRTGNN variants. Bold: best per column.

Model	RMSE ↓	MAE ↓	R^2 ↑	Spearman ↑
Gravity	1.4639 ± 0.0000	0.9534 ± 0.0000	0.9092 ± 0.0000	0.8913 ± 0.0000
MLP	1.4469 ± 0.0008	0.9100 ± 0.0062	0.9113 ± 0.0001	0.8985 ± 0.0003
MLP+T	1.3913 ± 0.0014	$\mathbf{0.8624 \pm 0.0013}$	0.9180 ± 0.0002	0.9036 ± 0.0005
GATv2	1.4450 ± 0.0140	0.9170 ± 0.0171	0.9115 ± 0.0017	0.8972 ± 0.0012
GATv2+T+R	1.3958 ± 0.0070	0.8955 ± 0.0080	0.9174 ± 0.0008	0.9023 ± 0.0013
MRTGNN+T	1.3840 ± 0.0062	0.8746 ± 0.0130	0.9188 ± 0.0007	0.9029 ± 0.0003
MRTGNN+T+R	$\mathbf{1.3773 \pm 0.0024}$	0.8676 ± 0.0070	$\mathbf{0.9196 \pm 0.0003}$	$\mathbf{0.9049 \pm 0.0002}$

Table 5: Network Spillover Ratio (NSR) under three counterfactual shocks. Higher NSR indicates greater relative net propagation. Pairwise MLP+T has zero spillover. Bold marks the largest NSR per scenario.

Model	Russia embargo (feature mask)	China-related tariff (edge cost)	China node features -30%
MLP+T	0.0000	0.0000	0.0000
GCN+T	≈ 0	≈ 0	0.0000
GATv2+T+R	**0.0025**	0.0215	0.0000
MRTGNN+T+R	0.0008	**0.2426**	0.0000

pendix G lists both country sets); and (iii) China node contraction, by contracting China's GDP, export, and import feature values by 30%. They cover edge-feature masking, continuous edge-cost perturbation, and node-level macro perturbation. The main metric is the Network Spillover Ratio, $\mathrm{NSR} = |\Delta_{\mathrm{spillover}}|/|\Delta_{\mathrm{direct}}|$, where each Δ sums signed log-prediction changes over direct or spillover dyads. The ratio is defined for nonzero aggregate direct change; pairwise models then have $\mathrm{NSR} = 0$ by construction. It measures net propagation and permits cancellation.

Table 5 reports NSR for the three shocks. The largest qualitative difference appears under the China-related tariff shock, where MRTGNN+T+R reaches $\mathrm{NSR} = 0.2426$ versus 0.0215 for GATv2+T+R (an $\approx 11.3\times$ gap) and ≈ 0 for GCN+T. This tests sensitivity of the joint MRT-guided propagation design to continuous bilateral trade-cost perturbations. Under the Russia embargo, GATv2+T+R yields a slightly higher NSR (0.0025) than MRTGNN+T+R (0.0008); this feature-masking intervention provides a separate boundary stress test. Under the China node contraction, all evaluated models give zero NSR under the current edge-level decomposition, reflecting that NSR with edge-level direct sets is less informative for node-level perturbations. Appendix G gives the original protocols. The expanded seven-economy analysis separates focal-country and third-country node responses and reports selected country-period tariff comparisons (Appendix I.6).

7 Conclusion

We proposed MRTGNN, a theory-guided graph neural network for bilateral trade-flow prediction. Its message-passing layer combines trade-cost gating, asymmetric message-source/destination projections, and edge-augmented messages; the full model adds temporal features and a bilateral residual pathway. On a panel of 176 countries, MRTGNN+T+R achieves modest, consistent RMSE gains over strong temporal and mechanism-matched controls in ten new matched runs. Extended evaluations show competitive forecasts through 2024 and closer alignment of learned allocations with a structural-gravity benchmark, while a generic transformer attains slightly lower point RMSE. The MRT correspondence provides a design principle and a conditional approximation analysis. Shock experiments support interpretable predictive network-sensitivity analysis under prescribed perturbations with the remaining macroeconomic state held fixed.

Acknowledgments and Disclosure of Funding

This research was supported by the China Postdoctoral Science Foundation (General Program, Grant No. 2024M752688). The authors declare no competing interests.

References

[1] Jan Tinbergen. *Shaping the World Economy: Suggestions for an International Economic Policy*. The Twentieth Century Fund, New York, 1962.

[2] James E. Anderson. A theoretical foundation for the gravity equation. *American Economic Review*, 69(1):106–116, 1979.

[3] Jeffrey H. Bergstrand. The gravity equation in international trade: Some microeconomic foundations and empirical evidence. *Review of Economics and Statistics*, 67(3):474–481, 1985.

[4] James E. Anderson and Eric van Wincoop. Gravity with gravitas: A solution to the border puzzle. *American Economic Review*, 93(1):170–192, 2003.

[5] Keith Head and Thierry Mayer. Gravity equations: Workhorse, toolkit, and cookbook. In *Handbook of International Economics*, volume 4, pages 131–195. Elsevier, 2014.

[6] Yoto V. Yotov, Roberta Piermartini, José-Antonio Monteiro, and Mario Larch. *An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model*. World Trade Organization and United Nations, 2016.

[7] Thibault Fally. Structural gravity and fixed effects. *Journal of International Economics*, 97(1): 76–85, 2015.

[8] Robert Dekle, Jonathan Eaton, and Samuel Kortum. Global rebalancing with gravity: Measuring the burden of adjustment. *IMF Staff Papers*, 55(3):511–540, 2008.

[9] Diego Rincon-Yanez, Chahinez Ounoughi, Bassem Sellami, Tarmo Kalvet, Marek Tiits, Sabrina Senatore, and Sadok Ben Yahia. Accurate prediction of international trade flows: Leveraging knowledge graphs and their embeddings. *Journal of King Saud University - Computer and Information Sciences*, 35(10):101789, 2023.

[10] Bassem Sellami, Chahinez Ounoughi, Tarmo Kalvet, Marek Tiits, and Diego Rincon-Yanez. Harnessing graph neural networks to predict international trade flows. *Big Data and Cognitive Computing*, 8(6):65, 2024.

[11] Serina Chang, Zhiyin Lin, Benjamin Yan, Swapnil Bembde, Qi Xiu, Chi Heem Wong, Yu Qin, Frank Kloster, Xi Luo, Raj Palleti, and Jure Leskovec. Learning production functions for supply chains with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27878–27886, 2025.

[12] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[13] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017.

[14] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[15] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022.

[16] Aravind Sankar, Yanhong Wu, Liang Gou, Wei Zhang, and Hao Yang. DySAT: Deep neural representation learning on dynamic graphs via self-attention networks. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 519–527, 2020.

[17] Naoto Minakawa, Kiyoshi Izumi, and Hiroki Sakaji. Bilateral trade flow prediction by gravity-informed graph auto-encoder. In *2022 IEEE International Conference on Big Data*, pages 2327–2332, 2022.

[18] Costas Arkolakis, Arnaud Costinot, and Andrés Rodríguez-Clare. New trade models, same old gains? *American Economic Review*, 102(1):94–130, 2012.

[19] Arnaud Costinot and Andrés Rodríguez-Clare. Trade theory with numbers: Quantifying the consequences of globalization. In *Handbook of International Economics*, volume 4, pages 197–261. Elsevier, 2014.

[20] Lorenzo Caliendo and Fernando Parro. Estimates of the trade and welfare effects of NAFTA. *Review of Economic Studies*, 82(1):1–44, 2015. doi: 10.1093/restud/rdu035.

[21] Mary Amiti, Stephen J. Redding, and David E. Weinstein. The impact of the 2018 tariffs on prices and welfare. *Journal of Economic Perspectives*, 33(4):187–210, 2019. doi: 10.1257/jep.33.4.187.

[22] Pablo D. Fajgelbaum, Pinelopi K. Goldberg, Patrick J. Kennedy, and Amit K. Khandelwal. The return to protectionism. *Quarterly Journal of Economics*, 135(1):1–55, 2020. doi: 10.1093/qje/qjz036.

[23] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

[24] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2019.

[25] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[26] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272, 2017.

[27] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[28] Kai Tang, Junbo Zhao, Xiao Ding, Runze Wu, Lei Feng, Gang Chen, and Haobo Wang. Learning geometry-aware representations for new intent discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5641–5654. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.306. URL `https://aclanthology.org/2024.acl-long.306/`.

[29] Kai Tang, Rui Wang, Renyu Zhu, Minmin Lin, Xiao Ding, Tangjie Lv, Changjie Fan, Runze Wu, and Haobo Wang. Towards transferable personality representation learning based on triplet comparisons and its applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10050–10066. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.509. URL `https://aclanthology.org/2025.emnlp-main.509/`.

[30] Guillaume Gaulier and Soledad Zignago. BACI: International trade database at the product-level. the 1994–2007 version. Working Paper 2010-23, CEPII, 2010.

[31] Maddalena Conte, Pierre Cotterlaz, and Thierry Mayer. The CEPII gravity database. Working Paper 2022-05, CEPII, 2022.

[32] Andreas Dür, Leonardo Baccini, and Manfred Elsig. The design of international trade agreements: Introducing a new dataset. *Review of International Organizations*, 9(3):353–375, 2014.

[33] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[34] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[35] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[36] Zijie Pan, Stepan Gordeev, Jiahui Zhao, Ziyi Meng, Caiwen Ding, Sandro Steinbach, and Dongjin Song. International trade flow prediction with bilateral trade provisions. *arXiv preprint arXiv:2407.13698*, 2024. doi: 10.48550/arXiv.2407.13698. URL https://arxiv.org/abs/2407.13698.

[37] Costanza Bosone and Alessandro Spelta. A Boltzmann generator framework for modelling and forecasting international trade networks. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnag083, 2026. doi: 10.1093/jrsssa/qnag083. URL https://doi.org/10.1093/jrsssa/qnag083.

[38] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjing Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1548–1554. International Joint Conferences on Artificial Intelligence Organization, 2021. doi: 10.24963/ijcai.2021/214. URL https://www.ijcai.org/proceedings/2021/214.

A Extended Preliminaries

Economic intuition for multilateral resistance. Multilateral resistance terms show that trade between countries i and j depends not only on their bilateral cost t_{ij}, but also on how difficult it is for both countries to trade with all other potential partners. Bilateral trade is therefore not an isolated country-pair relation but a systemic outcome embedded in the global trade network. Intuitively, MRT is a relative-accessibility or normalization factor: t_{ij} alone describes the direct difficulty of trading between i and j, but does not say whether this relationship remains attractive relative to other partners. If only two countries exist, they may trade heavily even at high bilateral cost; introducing a third low-cost partner dilutes the original relation. MRT captures exactly this effect: P_i and Π_j are equilibrium adjustment terms that place bilateral costs within the system of alternative trade opportunities.

Scale indeterminacy and normalization. The transformation $(\mathbf{p}, \boldsymbol{\pi}) \mapsto (\mathbf{p} + c\mathbf{1}, \boldsymbol{\pi} - c\mathbf{1})$ leaves every $p_i + \pi_j$ unchanged. For the theoretical analysis we choose the linear gauge map $\mathcal{R}(\mathbf{p}, \boldsymbol{\pi}) = (\mathbf{p} - p_1\mathbf{1}, \boldsymbol{\pi} + p_1\mathbf{1})$, with normalized state space $\mathcal{H} = \{(\mathbf{p}, \boldsymbol{\pi}) : p_1 = 0\}$, and define $\mathcal{T} = \mathcal{R} \circ \mathcal{U}$. The compact domain is chosen independently of this gauge.

Structural motivation. The raw full-support update $\mathcal{U}$ has edge-dependent trade-cost weights, distinct outward and inward dependencies, and repeated partner aggregation. These features motivate the gate, asymmetric message projections, and layer stacking.

B Method Details

High-dimensional MRT state. The vector $\mathbf{h}_i^{(\ell)}$ is a learned trade-network representation. Following the message-direction convention, $\mathbf{W}_{\mathrm{src}}\mathbf{h}_j^{(\ell)}$ represents the partner, or message source, and $\mathbf{W}_{\mathrm{dst}}\mathbf{h}_i^{(\ell)}$ the focal country, or message destination.

Trade-cost gate vs. attention. The gate $g(\mathbf{e}_{ij}) = \mathrm{sigmoid}(\mathbf{W}_g\mathbf{e}_{ij})$ modulates message magnitude, while softmax attention redistributes relative influence across partners. For $q_j = \exp((\sigma - 1)\pi_j)s_j$, the structural sum $\sum_j \tau_{ij}q_j$ combines relative partner weights $q_j/\sum_k q_k$ with the scale $\sum_k q_k$. A6 covers representation of the complete update, including its scale and nonlinear readout.

Flexible approximation of MRT aggregation. MRTGNN learns feature transformations and partner aggregation from trade data, accommodating uncertainty in the elasticity, effective trade-cost map, and relation between observed features and trade frictions. The theoretical analysis studies conditional representability of MRT updates, while the allocation diagnostics measure alignment of trained propagation with a structural-gravity benchmark.

Bilateral residual interpretation. $\mathrm{MLP}_{\mathrm{res}}([\mathbf{x}_i^t \| \mathbf{x}_j^t])$ provides a direct source-target gravity channel based on raw country features such as country scale, demand capacity, and macroeconomic conditions. Added at the prediction layer, it retains raw country-level signals when network effects are weak or graph aggregation smooths out local source-target information.

C Full Theoretical Proofs

C.1 Restated Assumptions

Assumption 1 (Bounded trade costs)**.** *For the full-support theoretical operator,* $0 < \underline{t} \le t_{ij} \le \overline{t} < \infty$ *for all* i, j*, including domestic terms, and* $\sigma > 1$ *is fixed.*

Assumption 2 (Bounded economic shares)**.** *The theoretical shares are fixed and satisfy* $0 < \underline{s} \le s_j, r_i \le \overline{s} < \infty$.

Assumption 3 (Compact normalized domain)**.** *Choose a compact set* $\mathcal{K} \subset \mathcal{H}$*, where* $\mathcal{H}$ *is the gauge-fixed state space defined in Appendix A.*

Assumption 4 (Lipschitz continuity)**.** *On* $\mathcal{K}$*,* $\|\mathcal{T}(\mathbf{z}) - \mathcal{T}(\mathbf{z}')\| \le K\|\mathbf{z} - \mathbf{z}'\|$ *for a finite* K*, using the same norm throughout the analysis.*

Assumption 5 (Finite-horizon domain condition)**.** *For the initial state and decoded updates under consideration, both* $\mathbf{x}_\ell = \mathcal{T}^\ell(\mathbf{z})$ *and* $\mathbf{u}_\ell = \widetilde{\mathcal{T}}_\ell \circ \cdots \circ \widetilde{\mathcal{T}}_1(\mathbf{z})$ *belong to* $\mathcal{K}$ *for* $0 \le \ell \le L$.

Assumption 6 (Idealized decoded representability)**.** *An idealized class* $\mathcal{D} = \{\psi \circ F_\theta \circ \phi\}$ *of continuous maps from* $\mathcal{K}$ *to* $\mathcal{H}$ *is uniformly dense in* $C(\mathcal{K}, \mathcal{H})$*. The encoder and readout are part of this specified class. Decoded-class density is posited as an analytical assumption; classical universal-approximation results provide background on expressive neural function classes [35].*

C.2 Proof of Proposition 1

Under A1–A2, every sum in Eq. (4) is positive. On the compact domain in A3, its arguments are bounded, and the sum is bounded away from zero. Exponentiation, summation, and the logarithm therefore give a continuous raw update $\mathcal{U}$ on $\mathcal{K}$. The linear gauge map $\mathcal{R}$ is continuous, so $\mathcal{T} = \mathcal{R} \circ \mathcal{U} \in C(\mathcal{K}, \mathcal{H})$. A6 then yields a decoded map within any prescribed uniform error ϵ. □

C.3 Proof of Theorem 1

Let $\mathbf{x}_0 = \mathbf{u}_0 = \mathbf{z}$, $\mathbf{x}_\ell = \mathcal{T}(\mathbf{x}_{\ell-1})$, $\mathbf{u}_\ell = \widetilde{\mathcal{T}}_\ell(\mathbf{u}_{\ell-1})$, and $E_\ell = \|\mathbf{u}_\ell - \mathbf{x}_\ell\|$. A5 ensures that the one-step error bound and A4 apply at both preceding states. Thus

$$E_\ell \le \|\widetilde{\mathcal{T}}_\ell(\mathbf{u}_{\ell-1}) - \mathcal{T}(\mathbf{u}_{\ell-1})\| + \|\mathcal{T}(\mathbf{u}_{\ell-1}) - \mathcal{T}(\mathbf{x}_{\ell-1})\| \le \epsilon_\ell + KE_{\ell-1}.$$

Starting from $E_0 = 0$ and expanding this recursion proves Eq. (5). □

C.4 Compatibility with a latent-layer stack

Decoded composition generally differs from $\psi \circ F_L \circ \cdots \circ F_1 \circ \phi$, because the latter has no intermediate decoding and re-encoding. A sufficient compatibility condition is that shared maps ϕ, ψ satisfy $\psi \circ \phi = \mathrm{id}$ on the visited states and

$$F_\ell(\phi(\mathbf{u})) = \phi(\widetilde{\mathcal{T}}_\ell(\mathbf{u}))$$

for every visited state at layer ℓ. Induction then gives $F_L \circ \cdots \circ F_1 \circ \phi(\mathbf{z}) = \phi(\mathbf{u}_L)$; applying ψ recovers the decoded composition and its bound.

C.5 Scope of the analysis

The analysis concerns finite-horizon decoded approximation under the stated assumptions. Applying it to the fixed-width experimental architecture requires establishing decoded representability, the trajectory-domain condition, and encoding/readout compatibility. Learned weights are evaluated through predictive performance and allocation diagnostics; structural identification and equilibrium convergence require separate analysis.

D Dataset and Implementation Details

Feature groups. Node features describe country-level economic scale (GDP, population, imports, exports, real-GDP-related indicators). Static edge features describe bilateral costs and dyadic conditions (tariffs, FTA indicators, product diversity, geographic attributes, historical trade presence) following standard international-trade resources [30, 31, 32]. Temporal edge features include lagged trade flows, trade growth, tariff changes, lagged tariffs, lagged quantities, and quantity growth.

Trade persistence. Bilateral trade prediction is highly persistent. We report lag-one correlation between consecutive log trade flows

$$\rho_1 = \mathrm{Corr}(y_{ij,t}, y_{ij,t-1}), \qquad y_{ij,t} = \log(1 + X_{ij,t}), \tag{6}$$

over exporter–importer pairs with consecutive observations. On our 176-country panel, the reported lag-one correlation is $\rho_1 \approx 0.939$ ($n \approx 4.93 \times 10^5$ consecutive dyad-year pairs), indicating strong persistence. This helps explain why temporal baselines such as MLP+T are strong competitors and why aggregate graph-based gains are numerically modest.

Implementation. In the original GNN comparisons, models share the data split, loss, optimizer, and evaluation protocol. Aggregation operators, temporal wrappers, and residual pathways differ according to the specified variants; the MRTGNN+T+R versus GATv2+T+R comparison matches the temporal and residual components. Hyperparameters were tuned on 2014–2015 validation. The original broad and component evaluations report mean $\pm$ standard deviation over three seeds (42, 123, 456); the focal reliability comparison uses ten new matched runs with no overlap with those three seeds (Table 3), and the sector evaluation uses five seeds (Appendix I.4). The original experiments use Adam, weight decay 10^{-5}, learning rate 10^{-3} with cosine decay, batch size 4096 dyads sampled per training step, and early stopping on validation RMSE with patience 10 epochs. Continuous features are standardized; the trade-flow target is $\log(1 + X_{ij,t+1})$. Hyperparameters of MRTGNN and the strongest GNN baselines are summarized in Table 6.

Table 6: Hyperparameters used for the main MRTGNN model and the strongest GNN baselines. Listed hyperparameters are shared across seeds; the indicated aggregation and residual components distinguish the rows.

Model	Hidden dim	Layers L	Heads	Dropout	LR	Edge-feat. dim
GCN+T	128	2	—	0.10	10^{-3}	18
SAGE+T	128	2	—	0.10	10^{-3}	18
GATv2+T	128	2	4	0.10	10^{-3}	18
GATv2+T+R	128	2	4	0.10	10^{-3}	18
MRTGNN+T	128	2	4	0.10	10^{-3}	18
MRTGNN+T+R	128	2	4	0.10	10^{-3}	18

Computational complexity. For a dense directed graph with N countries and a fixed number of layers, heads, and feature/hidden dimensions, MRTGNN inference scales quadratically in N. With variable widths, the cost also depends on dense edge projections, shared node projections, and the edge decoder. A trained network evaluates each predictive perturbation in one forward pass.

E Baseline Configurations

We list the baselines used in the main comparison (Table 2) and the controlled component-wise ablation (Table 4). *Econometric:* ridge-regularized gravity model. *Tabular:* random forest and gradient-boosted decision trees with temporal features (+T); multilayer perceptrons with and without temporal features [33, 34]. *Static GNNs:* GCN, GraphSAGE, GIN [12, 13, 23]. *Attention GNNs:* GAT (v1) and GATv2 [14, 15]. *Temporal GNNs:* DySAT [16], alongside GCN-GRU and GAT-GRU sequence wrappers. The baseline suite uses a shared data and evaluation protocol. Comparisons of aggregation operators are distinguished from ablations of temporal features or the residual pathway; only the mechanism-matched +T+R comparison holds both components fixed. Hyperparameters are tuned on the 2014–2015 validation split.

F Additional Ablation and Sensitivity Discussion

The component-wise discussion in Section 6.3 compares nonlinear prediction, temporal features, graph aggregation, and the residual pathway. The ten-run MRTGNN+T+R versus GATv2+T+R comparison isolates the aggregation operator under closely matched parameter counts. The improvement from adding the bilateral residual (R) to MRTGNN+T is consistent with retaining direct dyadic signals alongside partner aggregation.

Effect of depth L. The decoded-update analysis motivates viewing depth as a finite number of aggregation steps. We observed that $L \in \{2, 3\}$ suffices in practice: $L = 1$ underperformed deeper settings, while $L \geq 4$ shows diminishing returns and increased risk of oversmoothing on the directed dense trade graph. These qualitative observations motivate shallow unrolling.

Effect of hidden dimension and heads. Increasing the hidden dimension from 64 to 128 improved RMSE and Spearman; further increases yielded no consistent improvement. Multi-head attention with 4 heads was sufficient; larger head counts marginally increased variance across seeds without consistent gains. These observations are reported as qualitative trends from the validation sweep on 2014–2015; the configurations of the principal GNN comparisons are listed in Table 6.

Roles of the gate and edge-augmented messages. The gate controls message magnitude, while attention normalizes relative partner influence. The term $\mathbf{W}_E \mathbf{e}_{ij}$ incorporates bilateral features directly into message content. Table 3 evaluates these components jointly in the MRT-guided propagation design.

Effect of the bilateral residual (R). The residual pathway provides a direct exporter–importer prediction channel that bypasses graph smoothing; Table 4 reports its contribution to MRTGNN.

G Counterfactual Protocols and Discussion

Scenarios. (i) Russia trade embargo: at inference time we mask the directed edges between Russia and a 24-country Western sanctions coalition (USA, GBR, DEU, FRA, ITA, JPN, KOR, CAN, AUS, ESP, POL, SWE, FIN, NLD, BEL, AUT, CHE, CZE, PRT, GRC, DNK, NOR, IRL, ISL) by zeroing trade-presence and lagged-trade edge features. (ii) China-related tariff shock: we triple bilateral tariffs and force tariff-change features upward on China-related directed edges with a 12-economy coalition (USA, GBR, DEU, FRA, ITA, IND, JPN, KOR, CAN, AUS, BRA, MEX). (iii) China node contraction: we multiply China's GDP, export, and import feature values by 0.7 for both the exporter and importer roles. These cover edge-feature masking, continuous edge-cost perturbation, and node-level macro perturbation respectively.

Metric. Let $\mathcal{S}_{\text{dir}}$ and $\mathcal{S}_{\text{spil}}$ denote directly perturbed and remaining dyads, and write $\delta_{ij} = \hat{y}_{ij}^{\text{shock}} - \hat{y}_{ij}^{\text{base}}$. The original edge-level statistic is

$$\text{NSR} = \frac{|\sum_{(i,j)\in\mathcal{S}_{\text{spil}}} \delta_{ij}|}{|\sum_{(i,j)\in\mathcal{S}_{\text{dir}}} \delta_{ij}|}, \tag{7}$$

for a nonzero denominator. NSR measures relative net changes in log predictions: signed responses can cancel, and the ratio is sensitive to small direct changes. Pairwise models have zero spillover numerator. The added node-shock diagnostic separately reports shares of absolute responses on focal-country and third-country edges (Appendix I.6).

Results. Under the Russia embargo, GATv2+T+R achieves $\text{NSR} = 0.0025$ versus 0.0008 for MRTGNN+T+R; under the China-related tariff shock, MRTGNN+T+R achieves $\text{NSR} = 0.2426$ versus 0.0215 for GATv2+T+R and ≈ 0 for GCN+T; under the China node contraction, all evaluated models give zero NSR under the current direct/spillover decomposition. Pairwise MLP+T has zero spillover numerator by construction across all scenarios.

Interpretation and limitations. The China-related tariff experiment measures the joint design's predictive sensitivity to continuous trade-cost perturbations. The Russia embargo is a feature-masking stress test; explicit edge-formation or edge-deletion dynamics could extend the model to discrete network restructuring. For the China node contraction, the expanded analysis uses focal-country and third-country edges to match the node-level intervention (Appendix I.6). These experiments hold the remaining macroeconomic state fixed. General-equilibrium policy effects would additionally require endogenous income/expenditure updates and market clearing.

Comparison with classical structural-gravity counterfactuals. Classical structural-gravity counterfactuals re-solve Eq. (1) after a shock and update predicted trade flows from the new equilibrium MRTs. The MRT subproblem costs $\mathcal{O}(K_{\text{iter}}N^2)$ per scenario, using a chosen trade-cost function and calibrated σ. MRTGNN instead evaluates a fitted predictive response in one forward pass. The two approaches address complementary tasks: structural gravity models equilibrium adjustment, while MRTGNN provides scalable analysis of learned network sensitivities.

H Broader Impact and Reproducibility

Broader impact. Bilateral trade-flow prediction and counterfactual analysis are inputs into public policy discussions on tariffs, sanctions, free-trade agreements, and supply-chain resilience. We see two main risks in deploying MRTGNN-style models in this setting. First, predictive counterfactuals approximate the network propagation channel of structural gravity but do not endogenously update market clearing, so users should not interpret MRTGNN outputs as full general-equilibrium welfare estimates. Second, model errors could be unevenly distributed across small or under-represented economies, where data quality is lower; we therefore recommend using the model in conjunction with domain expertise and conventional structural-gravity diagnostics. Potential benefits include greater transparency in trade-network modeling and lower computational cost for exploring predictive responses across policy scenarios.

Reproducibility. Hyperparameters of the principal GNN comparisons are reported in Table 6. Tables 2, 4, and 5 retain the original three-seed results (42, 123, 456). Table 3 reports a disjoint set of ten matched runs; the sector analysis uses five seeds. Additional protocols and uncertainty estimates are provided in Appendix I. Data sources are referenced in Appendix D. In the original comparisons, the feature-year split (train 2002–2013, validation 2014–2015, test 2016–2018) and loss are kept identical across baselines and MRTGNN variants.

I Extended Empirical Evaluation

I.1 Matched-run evaluation and uncertainty

The ten-seed comparisons use a new seed set that excludes the original three seeds (42, 123, 456). Within this new set, models are matched on observations, targets, information cutoffs, splits, seeds, validation-only selection, and training budgets. GATv2+T+R is the mechanism-matched control; GATv2+T is descriptive. The lag-one log-trade correlation is approximately 0.939 over 4.93×10^5 dyad-years.

Across the ten runs, the paired RMSE improvements (baseline minus MRTGNN) are 0.00890 ± 0.00371 against MLP+T and 0.00942 ± 0.00829 against GATv2+T+R (mean $\pm$ standard deviation). Ensemble RMSE differences (MRTGNN minus baseline) have dyad-clustered 95% intervals: -0.00922 $[-0.01218, -0.00636]$ against MLP+T and -0.00796 $[-0.00915, -0.00676]$ against GATv2+T+R. These complement the seed-level intervals for training variability. Against MLP+T, R^2 improves in ten of ten runs (difference 0.00105, 95% interval $[0.00074, 0.00136]$), while the Spearman difference is not statistically significant (0.00016, $[-0.00051, 0.00082]$).

I.2 Frozen evaluation through 2024

The extension uses BACI HS02 flows and H2-aligned TRAINS tariffs[1] for 176 countries and 30,800 directed dyads, retaining the original features and feature-year splits: 2002–2013 training and 2014–2015 validation. Models are selected on validation data and frozen thereafter; each forecast uses information only through the preceding year. The adaptations preserve GGAE edge prediction [17], use second-order factorized interactions inspired by PTA-based trade prediction [36], and retain conditional Boltzmann whole-network generation with separate link and flow heads [37]. Unlike the original Boltzmann study's expanding-window re-estimation, all models here follow the frozen protocol. The generic graph-transformer baseline uses TransformerConv [38].

Table 7: RMSE by target period and pooled over 2020–2024 under the frozen protocol. Point estimates are shown; — denotes an unreported period estimate.

Model	2020–2021	2022–2023	2024	Pooled
MRTGNN+T+R	1.377	1.468	1.826	1.512
PTA-FM-inspired adaptation	1.379	1.477	1.889	1.532
Protocol-matched GGAE	1.477	1.567	1.933	1.613
Protocol-matched Boltzmann	2.278	2.472	2.702	2.446
TransformerConv+T	—	—	—	1.496

MRTGNN's R^2 is $0.921/0.913/0.870$ for 2020–2021/2022–2023/2024; TransformerConv achieves slightly lower pooled RMSE (1.496 versus 1.512).

I.3 Structural-gravity allocation diagnostics

On unseen 2019–2023 networks, we compare learned allocations with a tariff-based gravity benchmark using observed BACI shares. Table 8 distinguishes the complete normalized allocation from its relative-partner component.

Table 8: Alignment with two structural-gravity benchmark components. Higher is better.

Learned quantity	Spearman	Top-5 overlap	NDCG@5
Complete normalized gravity weight			
MRTGNN attention × gate	0.500	0.225	0.336
Near-parameter-matched GATv2 attention	0.300	0.082	0.134
TransformerConv attention allocation	-0.284	0.021	0.036
Relative-partner component			
MRTGNN attention	0.621	0.284	0.428
GATv2 attention	0.297	0.081	0.134

MRTGNN's complete-allocation advantages over TransformerConv have country-clustered 95% intervals of $[0.749, 0.821]$, $[0.189, 0.218]$, and $[0.283, 0.316]$ for Spearman, Top-5 overlap, and NDCG@5. Advantages persist across seeds, years, elasticities $\sigma \in \{3, 5, 7\}$, and fixed-2018 economic shares. GATv2+T+R and MRTGNN+T+R contain 53,873 and 55,114 parameters. TransformerConv's second layer is larger than GATv2's (68,352 versus 35,584 parameters), yet its allocation alignment is weaker.

I.4 Sector networks and economically meaningful prediction metrics

The BACI HS02 sectors are agriculture/food (HS01–24), textiles (HS50–63), and machinery/electrical (HS84–85). Across five seeds, MRTGNN improves on persistence and matched GATv2 in all three; TransformerConv has slightly lower point RMSE.

Over 2020–2024, MRTGNN and matched GATv2 both obtain export-destination NDCG@10 of 0.969 and import-source NDCG@10 of 0.985 at the reported precision.

[1] UNCTAD TRAINS tariff-data and nomenclature documentation: `https://wits.worldbank.org/witsapiintro.aspx?lang=en`.

Table 9: Sector RMSE and MRTGNN's reduction relative to persistence.

Sector	Reduction	MRTGNN	Matched GATv2	TransformerConv
Agriculture/food	8.0%	1.318	1.323	1.314
Textiles	9.7%	1.032	1.036	1.027
Machinery/electrical	12.2%	1.255	1.267	1.247

I.5 Heterogeneity across exporter groups

Country quartiles use training-period information only. MRTGNN's gains over matched GATv2 are larger for lower-income, smaller, and less-connected exporters (Table 10); the income contrast is positive in all ten seeds.

Table 10: MRTGNN's RMSE advantage over matched GATv2, with paired intervals for the lower-minus-higher group contrast.

Exporter characteristic	Lower group	Higher group	Contrast: 95% CI
Income	0.01393	0.00637	$[0.00502, 0.01009]$
Economic size	0.01116	0.00333	$[0.00122, 0.01445]$
Export-network strength	0.01180	0.00324	$[0.00148, 0.01565]$

I.6 Country and period variation in predictive shock responses

The shock analysis covers the USA, China, Germany, Japan, India, Brazil, and Russia. Each uses its 12 largest partners measured during the 2002–2013 training feature years. Feature years 2002–2023 and targets 2003–2024 include training and later periods; responses average five frozen extension models. Table 11 presents four selected economies.

Table 11: Tariff NSR for four selected economies. Partner selection and shock calibration differ from the original China scenario; compare levels within this table.

Focal economy	Pre-GFC	COVID	Recovery	2024
Brazil	0.668	0.671	0.742	0.666
Germany	0.458	0.488	0.530	0.501
Russia	0.535	0.416	0.363	0.203
United States	0.360	0.347	0.364	0.337

Continuous tariff changes, embargo-style edge-feature masking, and node contractions probe distinct input sensitivities; masking provides a boundary stress test. Across seven recovery-period economies, 84%–99% of absolute node-shock response lies on focal-country edges and 1%–16% on third-country edges.

These experiments measure predictive network sensitivity with macroeconomic conditions held fixed apart from the specified perturbations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction distinguish the conditional MRT approximation analysis from empirical predictive performance and allocation alignment. Counterfactual outputs are described as predictive sensitivities rather than causal elasticities or general-equilibrium effects.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The theoretical scope and the limits of the fixed-width implementation are discussed in the theoretical analysis and Appendix C. Appendix G discusses the lack of endogenous general-equilibrium feedback and the distinct interpretation of edge-feature masking and node shocks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical analysis and Appendix C state the domain, normalization, regularity, and approximation assumptions used in each result and provide the corresponding proofs. The finite-depth error bound concerns decoded updates under the stated approximation, trajectory, and compatibility conditions.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: The paper specifies the data construction, temporal split, features, training objective, metrics, and counterfactual scenarios in Appendix D and Appendix G. Complete reproduction of the additional evaluations is not yet specified, including all seed identifiers and implementation details of the allocation diagnostic and extended baselines.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper cites data sources and describes the experimental protocol, but provides no public code repository or processed-data download link. Appendix H summarizes the reported reproducibility information.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: Appendix D gives the original model settings, temporal split, and training protocol, while Appendix I describes the additional evaluations. Complete seed identifiers and per-baseline hyperparameter settings are not provided for every result.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 2 and 4 report variation across three training seeds, and Table 3 reports paired 95% confidence intervals and win counts across ten matched runs. Appendix I.1 distinguishes training-run variability from dyad-clustered uncertainty in ensemble predictions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The training and evaluation protocol is described in Appendix D, but the paper does not report the exact compute hardware, memory, execution time, or total compute consumption.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses economic and trade datasets and does not involve human-subject experiments or private personal data. Appendix H discusses the intended use and potential risks of trade-flow prediction and network-sensitivity analysis.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix H discusses potential benefits for transparent trade-network analysis and risks from confusing predictive sensitivities with welfare effects or from uneven errors across economies. It recommends using the model with domain expertise and structural-gravity diagnostics.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with the high-risk characteristics described in this question. It studies trade-flow prediction on economic network data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper cites the principal data sources, but does not explicitly document the license, terms of use, and exact version of every asset used. No processed-data or code release is provided.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release a new dataset, pretrained checkpoint, or software package. It presents a method and empirical evaluations.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects, participant recruitment, or private participant data, so IRB review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not an original or non-standard component of the research methods, experiments, or data construction. Writing and editing assistance does not form part of the proposed model or evaluation protocol.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.