Machine Learning the Gravity Equation for International Trade *

Sergiy Verstyuk† Michael R. Douglas‡

First version: 8 March 2022
This version: 29 March 2022

Abstract

Machine learning (ML) is becoming more and more important throughout the mathematical and theoretical sciences. In this work we apply modern ML methods to gravity models of pairwise interactions in international economics. We explain the formulation of graphical neural networks (GNNs), models for graph-structured data that respect the properties of exchangeability and locality. GNNs are a natural and theoretically appealing class of models for international trade, which we demonstrate empirically by fitting them to a large panel of annual-frequency country-level data. We then use a symbolic regression algorithm to turn our fits into interpretable models with performance comparable to state of the art hand-crafted models motivated by economic theory. The resulting symbolic models contain objects resembling market access functions, which were developed in modern structural literature, but in our analysis arise *ab initio* without being explicitly postulated. Along the way, we also produce several model-consistent and model-agnostic ML-based measures of bilateral trade accessibility.

Keywords: gravity, bilateral accessibility, market access function, graph theory, graph neural network, black-box model interpretation

*Acknowledgements: William Hamilton, Gordon Hanson, Scott Kominers, Marinka Zitnik. SV was supported by the Center of Mathematical Sciences and Applications at Harvard University.

†Contact address: verstyuk@cmsa.fas.harvard.edu.

‡Contact address: mdouglas@cmsa.fas.harvard.edu.

Contents

1 Introduction

This study deals with international economics, focusing on gravity equation for international trade. Taking a large panel of annual country-level data (68 years for 181 countries), we are interested in leveraging recent developments in Machine Learning (ML) to assess what insights it may offer to existing empirical and theoretical knowledge. Specifically, we use Neural Network (NN) models to construct a flexible formulation of the gravity equation (and, separately, a measure of bilateral accessibility of an importer to an exporter), and then utilize algorithmic tools to help us interpret the internal mechanics of the optimally fitted gravity NN.

The observation that trade flows between pairs of countries can be well captured by an expression with a striking resemblance to Isaac Newton's universal law of gravitation,[1]

$$\text{Trade}(i \leftrightarrow j) \approx \gamma \frac{\text{GDP}_i \times \text{GDP}_j}{\text{Distance}(i,j)}, \tag{1}$$

goes back to Jan Tinbergen (1962). Since then international economics has expanded our understanding tremendously, both theoretically and empirically, but we believe that new tools can bring further progress on this front.

Modern ML has achieved impressive results over the last several years, with applications of NNs to image recognition (Krizhevsky et al., 2012; Szegedy et al., 2014), natural language processing (Vaswani et al., 2017) or chemistry (Jumper et al., 2021), as well as reinforcement learning to game playing (Silver et al., 2018), theorem proving (Kaliszyk et al., 2018; Zombori et al., 2020) and knowledge discovery (Gauthier, 2020; Davies et al., 2021; Wagner, 2021).[2] Recognizing our focus area, international trade, as a canonical example of the mathematical structure known as a graph, we make use of Graph Neural Network (GNN) models, which have demonstrated strong results in the corresponding application domains (Kipf, 2016).

Two key ingredients in many theoretical gravity models are so-called multilateral resistance terms, or market access functions, which scale each of the exporter (importer) country's market size by the size of its respective import (export) partners, thus defining them in relative terms and reflecting the degree of competition in each market.[3] In graph

[1] Below, γ is a constant parameter. We also replaced the GNP used in the original formulation with the GDP, and set power coefficients on three right hand side (r.h.s.) variables to "1", "1" and "-1", respectively (thus making the equation symmetric). Also note that in this formulation the left hand side (l.h.s.) is understood as a 2-dimensional vector since trade is flowing in both directions, or as a scalar if their sum is considered.

[2] Similarly to computer science and other disciplines, economics undergoes a renewed interest in NN methods. Theoretical studies establishing rigorous foundations behind NNs include Chen (2007), Hartford et al. (2017), Farrell et al. (2021). Empirical works demonstrate applications to auctions design (Dütting et al., 2021), finance (Gu et al., 2020), and macroeconometrics (Verstyuk, 2020). Computer scientists directly tried their hands in economics as well (Yang et al., 2020).

[3] The resulting more complicated objects replace countries' GDPs in the original formulation of Jan Tinbergen.

theory and network theory, it is a central observation that the interaction between two vertices depends not only on the edge connecting them but also on their larger neighborhood within the graph. This type of dependence is innate to GNN models, which learn by aggregating information over the neighborhood of each vertex. Instead of postulating specific non-pairwise interactions, a GNN model can represent a very general class of such interactions, constrained by locality and interchangeability, and learn the specific forms which are best suited to a particular application case.

Generally speaking, NNs can be understood as a collection of highly flexible functions with many degrees of freedom (i.e., free parameters) that take in some explanatory variables and produce predicted values for the variables to be explained. Their rich parameterization does not prevent them from learning useful features of the data in the training data sample (In-Sample, henceforth IS) and predicting for the test sample (Out-Of-Sample, OOS). Our unrestricted NN-based model demonstrates a much stronger fit both IS and OOS than the alternative that explicitly includes restrictions stemming from economic theory; it also fares strongly in comparison to the "gold-standard" model with time-varying country fixed effects.[4] It is also reassuring that the terms embodying export and import market characteristics that are generated by our NN model are closely correlated with their counterparts from the fixed effects benchmark model.

However, fit accuracy in such models is not useful per se, hence our work also focuses on producing model-consistent measures of bilateral accessibility for a broad set of country pairs and years that can be used for a wide range of empirical applications. We verify and confirm that they behave in a reasonable manner (i) on aggregate over the whole length of our dataset, and also in particular periods such as (ii) financial crisis of 2008-2009 and ensuing Great recession as well as (iii) enlargement of the European Union over 2004-2013.

Our measures of bilateral accessibility rely on 17 different cultural, geographic, economic, legal and political trade cost-related variables that are expressed by an auxiliary NN, which can be incorporated into any econometric model (including simple linear regressions). On that account, we additionally construct a relatively agnostic measure of bilateral accessibility based on our auxiliary NN with controls for time-varying exporter and importer country effects, which may be of independent interest to empirical trade economists. A notable property that turns out to hold for all our measures of bilateral accessibility is their asymmetry, that is a given pair of countries exhibits different accessibility depending on which trade direction we are looking at.[5]

The rich parameterization and overall complexity of NNs means that they generally do not admit an *a priori* interpretation. For many applications (for example supporting crucial human decisions such as medical diagnosis or public policy), this problem is very serious, and so many techniques have been developed to aid in interpreting them. In our

[4]The latter two classes of gravity models are well explained in, e.g., Head and Mayer (2014); section §2 provides additional literature sources.

[5]In addition to the above, it is also worth noting the relatively small direct contributions of these measures to corresponding models' overall explanatory performance.

work, we use an interpretation method called Symbolic Regression (SR) to find standard mathematical functions that are able to accurately replicate our NN's predictions and performance. Many SR algorithms (including the one we used) provide a list of candidate interpretations and their use is often guided by common sense and intuition. But as we will see, SR combined with some economic intuition produced fairly well determined interpretations for our NN model. This approach allowed us to uncover its internal mechanics, including the values of its reduced-form parameters.

The interpretation we obtain in this way seems to "rediscover" and refine some of the functional relationships encountered in existing gravity models founded on economic theory, while outperforming its theory-motivated counterparts in IS and OOS predictions. For instance, we detect objects similar to multilateral resistance terms. Furthermore, our interpretation exercise appears to "discover" some functional forms we could not find analogues of in the literature.[6] Some of them are rather counter-intuitive, hence we also consider their alternative specification (of the Constant Elasticity of Substitution form), which allows to enforce more intuitive economic relationships between the input variables.

The paper is organized as follows. In the next section we provide references to related work. Section §3 describes our methods, with a brief review of graph theory as well as of neural network models and their interpretation. In the following section we formally define the gravity models used. Data and estimation details can be found in sections §5 and §6. Section §7 contains the main estimation results as well as their discussion and informal interpretation, including some empirical illustrations. Section §8 provides a more formal approach to their interpretation. The final section concludes.

2 Related literature

Standard models for gravity-type relationships in international economics in modern paradigm start from Eaton and Kortum (2002) and Anderson and van Wincoop (2003). Many gravity models rely on Constant Elasticity of Substitution preferences, but there are alternatives built on, e.g., translog preferences (Novy, 2013) or Almost Ideal Demand System (Fajgelbaum and Khandelwal, 2016). Higher flexibility is also possible with semi/non-parametric approaches, as in Adão et al. (2017), Lind and Ramondo (2018) or Adão et al. (2020). This paper can be viewed as a continuation of this effort.

More recent developments in the area focus on rationalizing the role of distance in gravity models (Chaney, 2018), adding temporal dimension to them in order to capture path dependence in trade (Morales et al., 2019), further progress in their generalization (Allen et al., 2020), which resonates with our motivations as well.

An important concept in modern trade literature is a market access function, developed

[6]It also provides additional evidence (specifically, properly aggregated measures of production have a larger quantitative impact than those of expenditures) for the dominance of the supply side of the market in our data, which agrees with the variance decomposition results obtained using original NN's outputs.

by Hanson (2005), Redding and Venables (2004), Head and Mayer (2011), Donaldson and Hornbeck (2016), also see Costinot et al. (2019). It is intended to explicitly model the impact of third parties on bilateral trade, the effect that is at the heart of our own GNN-based approach.

Research on gravity models in international economics is vast, and we only scratched the surface. More references can be found in excellent reviews offered by Head and Mayer (2014), Costinot and Rodríguez-Clare (2014), Donaldson (2015), Redding and Rossi-Hansberg (2017).

3 Methodological approach

3.1 Graphs

We will start with introducing the basics of graph (equivalently, network) theory.[7,8] In economics, these ideas have already been used in various network models, such as Long and Plosser (1983); Kakade et al. (2004); Hausmann and Hidalgo (2011); Acemoglu et al. (2012), but their roots can be traced as far back as Leontief (1941) (also see von Neumann's (1945) take).

Mathematically, a graph is defined by a set V of vertices (nodes), and a set E of edges which is a subset of the set of pairs of vertices, denoted $V \times V$.[9] The case in which every pair of distinct vertices is connected by an edge, so $E = \{(a, b) \in V \times V | a \neq b\}$, is referred to as the "complete" graph.[10] Given this structure, one can label (or "decorate") the vertices and edges with numerical or categorical data.

For example, in the case of international trade each country (or firm) corresponds to a vertex, while the trading relationship between a pair of countries corresponds to an edge. The variables in equation (1) is then naturally associated to vertices (for example, GDP or other country-specific data) or edges (trade volumes, geographical distances and relationships).[11]

We next review some basic concepts of graph theory. The two vertices a and b incident to an edge (a, b) are referred to as its head and tail. A path is a sequence of edges

[7]Standard textbooks are West (2000) and Diestel (2010), with Even and Even (2012), Newman (2010), and Jackson (2008) providing a somewhat more applied perspective.

[8]To avoid confusion, in the text we try to adhere to using the term "graph" for correspondingly-structured data, while reserving the term "network" for the formal structure of NN models.

[9]By "graph" we will mean an undirected (or unoriented) graph, one in which edges are symmetric in the sense that $(a, b) \in E \leftrightarrow (b, a) \in E$. The more general (directed) graphs not satisfying this condition also arise in economics, for example in input-output networks.

[10]We will use the (essentially) complete graph in our model, but just to have another example let us define the "geography" graph G as follows: the vertices are countries, and a pair of countries are connected by an edge if and only if they share a land border.

[11]In some cases, e.g., when trade volume is zero, edges may be left-out or their associated variables may be set to 0.

$(e_1, e_2, \ldots, e_n)$ such that for each i, the tail of e_i is equal to the head of e_{i+1}. The number of edges n in a path is its length, while the head of e_1 and tail of e_n are its starting point and ending point. The k-neighborhood of a vertex a is the set of all vertices which are the ending points of paths of length less than or equal to k which start at a, while the k-neighborhood of an edge (a, b) is the union of the k-neighborhoods of a and b. We will also use a second (slightly different) definition of neighborhood which contains edges as well; the 0-neighborhood of a vertex a includes all edges incident to a.

We next note two important properties of equation (1), *locality* and *exchangeability*, which we will maintain in generalizing it. In fact the two properties are stronger together than separately. To explain this, let us refer to the individual equations obtained from (1) by specializing to a specific pair of countries (fixing the indices a and b) as the "component equations."

Locality states that each component equation only relates data contained in a single k-neighborhood of the graph. If (as is the case for equation 1) the l.h.s. refers to a single edge (or vertex), we mean the k-neighborhood of that edge (or vertex), though the definition can be generalized. Since the r.h.s. of equation (1) depends only on data in the 0-neighborhood of the edge (i, j), it is local.

A next generalization of equation (1) would be to allow the r.h.s. to depend on the 1-neighborhood of the edge (i, j). The strength of this restriction depends on the structure of the graph; for example it is quite restrictive for the geography graph G, but for the complete graph it imposes no restriction at all.

Exchangeability states that a general relabeling of the vertices preserves the set of component equations. Thus, specific choices of vertex cannot appear; only general indices such as the i and j of equation (1). An example of an allowed generalization would be for the r.h.s. to additionally depend on quantities such as $\sum_{k \in V, k \neq i,j} \mathrm{GDP}_i \mathrm{GDP}_k / \mathrm{Distance}(i, k)$, in which every possible choice of the third vertex k appears with equal weight. This example can be modified to satisfy locality as well by restricting the sum over k to include only those vertices connected by an edge to i.

It may not be obvious how to write the most general set of equations which satisfy both locality and exchangeability. In §3.2 we will state a prescription, known as message passing, which produces a wide variety of such equations. The basic idea is to introduce latent variables associated to the vertices and edges, each determined by equations which can depend on the 0-neighborhood of the specific vertex or edge. By composing these equations, or equivalently allowing each equation to depend on the previously defined latent variables in its 0-neighborhood, one systematically incorporates longer range dependencies. And as we explain in §3.2, one can get a rich family of parameterized equations by using the "layer" of (feedforward) neural networks as the basic equation in this construction. This setup is known as a graph neural network or GNN.

3.2 Neural networks

Feed-Forward Neural Networks: To give a brief introduction,[12] NNs are formed by a hierarchy (composition) of simple nonlinear functions that transform input data into outputs, thus producing predictions. The most basic Feed-Forward NN (FFNN) is defined as follows:

$$\boldsymbol{f}(\boldsymbol{z}_0, \boldsymbol{\theta}) := \boldsymbol{h}_{L+1}(\boldsymbol{h}_L(\cdots \boldsymbol{h}_2(\boldsymbol{h}_1(\boldsymbol{z}_0))\cdots)), \tag{2}$$

where each hidden layer $\ell = 1, \ldots, L$ and output layer $\ell = L+1$ is formed by N^ℓ neurons,

$$\boldsymbol{h}_\ell(\boldsymbol{z}_{\ell-1}) := \boldsymbol{g}_\ell(\boldsymbol{b}_\ell + \boldsymbol{W}_\ell \boldsymbol{z}_{\ell-1}), \tag{3}$$

with an activation function $\boldsymbol{g}_\ell : \mathbb{R}^{N_{\ell-1}} \mapsto \mathbb{R}^{N_\ell}$ (which is, usually, a nonlinear transformation, e.g., sigmoid or Rectified Linear Unit; except in the output layer, in which it is application-specific and is, oftentimes, a linear or logistic transformation);
$\boldsymbol{\theta}_\ell := [\boldsymbol{b}_\ell, \boldsymbol{W}_\ell]$ are parameters of conformable shapes, $\boldsymbol{b}_\ell \in \mathbb{R}^{N_\ell}$ and $\boldsymbol{W}_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$;
$\boldsymbol{z}_{\ell-1}$ is a corresponding function's input (with sample data $\boldsymbol{z}_0$ sometimes called the input layer).

FFNN's output seeks to approximate some unknown function:

$$\mathring{\boldsymbol{f}}(\boldsymbol{z}_0, \boldsymbol{\theta}) = \boldsymbol{f}(\boldsymbol{z}_0, \boldsymbol{\theta}) + \boldsymbol{\xi} \approx \boldsymbol{f}(\boldsymbol{z}_0, \boldsymbol{\theta}),$$

where $\mathring{\boldsymbol{f}}(\cdot)$ is the unknown data-generating function;
$\boldsymbol{\xi}$ is an approximation error term (residual).

Thus, effectively NNs are flexible functionals used for mapping explanatory input variables into the explained output variables.

The theoretical appeal of NNs rests on their universal approximation property that allows them to capture functional representations of complex interrelationships present in the data with arbitrary precision (Cybenko, 1989; Hornik et al., 1989). Achieving high approximation accuracy requires high expressive capacity, determined by the number of layers and neurons (i.e., architecture and parameterization). However, the dependence of accuracy on these choices is intricate. NNs fitting flexible functions have so many parameters that they should be thought of as infinitely-dimensional and effectively non-parametric objects. In general, the expressive capacity, complexity and performance of NNs are not related to the number of parameters in a simple way.

The current understanding of this relation postulates two statistical learning regimes. In the "classical" regime, the number of parameters is low relatively to the number of observations ($|\boldsymbol{\theta}|$ smaller than $|\boldsymbol{D}|$), and the OOS error is convex in $|\boldsymbol{\theta}|$ with a sharp minimum. That is, IS error is decreasing in $|\boldsymbol{\theta}|$, but for $|\boldsymbol{\theta}|$ beyond some optimal value the OOS error starts rising. This phenomenon is called "overfitting" and is well understood

[12] Good textbooks on the topic are Bishop (2006) or Goodfellow et al. (2016)

in terms of concepts such as the bias-variance tradeoff, or more intuitively as the idea that simpler models will generalize better, a version of "Occam's razor."[13]

As one increases the number of parameters $|\boldsymbol{\theta}|$ at fixed training set size, eventually one reaches a number (much greater than $|\boldsymbol{D}|$) where the model can fit the training observations perfectly, and the IS error is zero. In the "interpolation" or "overparameterized" regime the model can perfectly fit **any** set of training observations of size $|\boldsymbol{D}|$. Surprisingly, by appropriate use of ML optimization techniques, and if the dataset has structure which makes prediction possible, as one continues to increase $|\boldsymbol{\theta}|$ the OOS error again decreases, undergoing a "double descent." Indeed the OOS error can be smaller in this regime than its optimum in the classical regime. This phenomenon has been called "benign overfitting" and appears to contradict the traditional dogmas of statistics. Evidence for and analysis of this phenomenon can be found in Zhang et al. (2017); Belkin et al. (2019); Bartlett et al. (2020); Nakkiran et al. (2019).

In this paper, we focus on the more established classical regime. But the above points justify why, by ML standards, even our largest model complexity is fairly modest.

Graph Neural Networks: A relatively recent addition to NN frameworks is GNN (see Kipf (2016), Morris et al. (2019), You et al. (2020); textbook treatments are available in Hamilton (2020) and, from a broader theoretical perspective, Bronstein et al. (2021)). Intuitively, Convolutional NNs are well-defined only over grid-structured data (e.g., images), while Recurrent NNs only for sequential data (e.g., text). GNN formulates a generalization of convolutions to the non-Euclidean domain (or, to put it differently, a generalization of recurrencies to the non-uniquely ordered graph domain).

From our perspective, the modeling challenge graphs pose is that they have a complex geometric structure without an absolute order in space or time. More concretely, graph data satisfy exchangeability and locality. Therefore, the model should respect them too: (i) it must be defined over permutation-equivariant inputs; (ii) it must be applied to neighborhoods rather than a full set of nodes.

In what follows, we will use objects called "node embeddings" that compactly represent the nodes in the graph: these are latent variables $\boldsymbol{\varepsilon}(u) \in \mathbb{R}^{d_\varepsilon}$ that for each node $u \in V$ summarize the available local information around it (i.e., both node- and edge-related inputs) in a task-dependent way. The idea is to place nodes into a low-dimensional space by compressing the relevant information in a way that allows accurate reconstruction of the original high-dimensional graph (this is achieved by maintaining distances between points in two spaces).

Now, mathematically any node-level function $\boldsymbol{\varphi} : \mathbb{R}^{|V|\times|V|} \mapsto \mathbb{R}^{|V|\times d}$, when applied to graph data should satisfy permutation equivariance property (or, abusing slightly the terminology, exchangeability). That is, $\boldsymbol{\varphi}(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^\intercal) = \boldsymbol{P}\boldsymbol{\varphi}(\boldsymbol{A})$, where $\boldsymbol{A}$ is an adjacency

[13] A related but slightly different method of achieving simplicity is regularization. This could be a term in the objective function which favors sparseness, or a step which shrinks parameter values toward 0.

matrix and $\boldsymbol{P}$ is a permutation matrix.[14]

Next, aggregation of information from nodes should satisfy locality. Define neighborhood (0-neighborhood, to be precise) of some node u as $V(u) := \{v \in V : (u, v) \in E\}$. An admissible local function $\boldsymbol{\mu}(\{\boldsymbol{\varepsilon}(v), \forall v \in V(\cdot)\}) : \mathbb{R}^{|V(\cdot)| \times d_\varepsilon} \mapsto \mathbb{R}^{d_\mu}$ operates on a neighborhood $V(u)$, $\forall u \in V$, collects local nodes' embeddings and generates "messages" of dimension d_μ. The function's input is a set, hence its first operation must be some kind of uniform aggregation over all set elements (e.g., $\sum(\cdot)$, $\max(\cdot)$, $\min(\cdot)$, $\text{mean}(\cdot)$, etc.). This still satisfies exhangeability.

Lastly, a function $\boldsymbol{\upsilon}(\boldsymbol{\varepsilon}(\cdot), \boldsymbol{\mu}(\{\boldsymbol{\varepsilon}(v), \forall v \in V(\cdot)\})) : \mathbb{R}^{d_\varepsilon + d_\mu} \mapsto \mathbb{R}^{d_\varepsilon}$ iteratively updates the node's embedding by combining its current embedding with the latest message aggregating information about its neighborhood. Formally, for each $u \in V$,

$$\boldsymbol{\varepsilon}(u) = \boldsymbol{\upsilon}(\boldsymbol{\varepsilon}^{(-1)}(u), \boldsymbol{\mu}(\{\boldsymbol{\varepsilon}^{(-1)}(v), \forall v \in V(u)\})),$$

where "(-1)" superscript denotes value from the previous iteration (which is initialized at, say, node-specific variables $\boldsymbol{z} \in \mathbb{R}^{d_z}$ compressed to conformable dimensions). After k iterations (sometimes also called "layers"), node's embedding reaches out to and captures information contained in its k-neighborhood.[15]

To sum up, function $\boldsymbol{\mu}(\cdot)$ is called aggregation function, it produces a message aggregating information from the node's local neighborhood. Function $\boldsymbol{\upsilon}(\cdot)$ is called update function, it collects messages with aggregated local information and updates the node's embedding. Both functions can be formulated as NNs. This learning algorithm is called "message passing".

Perhaps the simplest example takes the following form:

$$\boldsymbol{\mu}(\{\boldsymbol{\varepsilon}(v), \forall v \in V(u)\}) := \sum_{v \in V(u)} \boldsymbol{\varepsilon}(v),$$

$$\boldsymbol{\upsilon}(\boldsymbol{\varepsilon}(u), \boldsymbol{\mu}(\{\boldsymbol{\varepsilon}(v), \forall v \in V(u)\})) := \boldsymbol{g}(\boldsymbol{b} + \boldsymbol{W}_\varepsilon \, \boldsymbol{\varepsilon}(u) + \boldsymbol{W}_\mu \, \boldsymbol{\mu}(\{\boldsymbol{\varepsilon}(v), \forall v \in V(u)\})),$$

where notation is the same as in equation (3).[16]

[14]Strictly speaking, permutation equivariance is a weaker property than permutation invariance, which is equivalent to the notion of exchangeability in statistics. A stronger notion is needed for dealing with whole graphs. Then, a graph-level function $\boldsymbol{\psi} : \mathbb{R}^{|V| \times |V|} \mapsto \mathbb{R}^d$ should satisfy permutation invariance, which requires $\boldsymbol{\psi}(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^\intercal) = \boldsymbol{\psi}(\boldsymbol{A})$.

[15]As long as $\boldsymbol{\upsilon}(\boldsymbol{\varepsilon}(\cdot), \boldsymbol{\mu}(\cdot))$ is a contraction map, this iterative process converges to its fixed point from any initial value exponentially fast, according to Banach's fixed point theorem. However, as more iterations are performed, the learned node embeddings may become very similar to one another and relatively uninformative ("over-smoothed"): the information being aggregated from the node neighbors during message passing begins to dominate the updated node representation at the expense of more local information preserved in the node representation from the previous iterations. (The problem of over-smoothing is also present in more familiar CNNs.)

[16]The node-level equation above can be restated as a graph-level system. Using matrix notation, it becomes: $\boldsymbol{\mathcal{E}} = \boldsymbol{G}(\boldsymbol{B} + \boldsymbol{A}\boldsymbol{\mathcal{E}}^{(-1)}\boldsymbol{W}_\mu + \boldsymbol{\mathcal{E}}^{(-1)}\boldsymbol{W}_\varepsilon)$, where $\boldsymbol{\mathcal{E}} \in \mathbb{R}^{|V| \times d_\varepsilon}$ is a nodes-by-embedding dimension matrix of $\boldsymbol{\varepsilon}(\cdot)$-s, $\boldsymbol{B}$ is a matrix of transposed and vertically stacked $\boldsymbol{b}$-s, and $\boldsymbol{A}$ is the graph adjacency matrix.

Now we can see why GNN is a natural framework for models in international economics.

(a) Exchangeability of labels makes it possible to formulate an economic model that is applicable to any generic country or their subset.

(b) Node embeddings summarize multidimensional information collected in disparate country-specific variables with approximate reduced-dimensional representations. In our case, node embeddings embody supply- and demand-market characteristics, and are generated by corresponding NNs.

(c) Standard graph theory formalism built on directed edge connections with continuous, potentially multi-dimensional edge weights allows for heterogeneous (and possibly asymmetric) economic links between any two origin and destination trading partners. In our model, edge-specific variables are compressed by NNs into scalar weights (that is, bilateral accessibility measures).

(d) Message passing allows to implement the idea of supply and demand market access functions: that each country's trade with another country depends not only on the variables specific to origin and destination countries themselves, but also indirectly on the third countries that are important for each of the two direct trading partners. Moreover, iterative message passing in principle allows for higher-order influences, fourth countries affecting third countries and so on.[17]

3.3 Interpretability

Given the complexity of most ML models, especially in the case of NNs, their interpretation is a challenging task. For more details, see recent reviews and references there in (Molnar et al., 2020; Xu et al., 2020). We will now describe two approaches particularly relevant for NN models in economics.

The first approach is to develop attribution methods which measure the impact and importance of each explanatory variable. A variable's marginal impact is measured as a change in a model's predicted output given a small change in the input variable, a quantity which can be calculated using standard automatic differentiation libraries. For NN models one can reuse the gradients computed for optimization when training the model; e.g., see Sundararajan et al. (2017). A variable's causal impact is measured as an effect on the modeled phenomenon (i.e., dependent, explained variable) of a particular input factor (explanatory, independent variable), while isolating it from the interference of confounding variables. Given the prominence of this question to economists, there already

[17]Such effects are less important for trade between countries in modern, "globalized" period, but were more important at earlier times. They are still important for modern trade between firms and, even more so, between individuals. Movements of entities other than goods and services, such as international human migration, is another possible application domain.

are techniques for implementing this calculation, as proposed in Hartford et al. (2017) or Farrell et al. (2021). Uncertainty quantification for the estimates described above can be done using standard bootstrap procedures (Efron and Tibshirani, 1993) or relying on the outputs of widely-used dropout techniques (Gal and Ghahramani, 2016; Hartford et al., 2017).

In turn, variable importance quantifies the contribution of each input variable to model's output and allows us to rank inputs accordingly. This can be done using the gradients discussed above, or their squares or absolute values, e.g. see Dimopoulos et al. (1995). A simpler method is to measure the deterioration in the model's performance when the variable is dropped (in practice one usually sets its value to the sample mean).

A second approach uses surrogate models. This amounts to constructing a secondary "cartoon" model that approximates the behavior of the original "black box" model but has a clearer interpretation. One popular class of surrogate model is the decision tree, *e.g.* see Augasta and Kathirvalavakumar (2012); Guidotti et al. (2018); Frosst and Hinton (2017); Bastani et al. (2017). A variation of this approach tailored specifically to GNNs is proposed in Ying et al. (2019).

One might advocate the automatic construction of a surrogate model by an algorithm which uses few if any *a priori* choices. But in practice, researchers usually propose and fit a variety of candidate surrogate models, guided by attribution methods to keep the most important variables, and choosing among model classes and parameter counts. The final choice between these models is based on fidelity to the original NN model but also on human judgement and intuition. A sign that this approach is appropriate given the current state of the art is that it is followed even for applications to pure mathematics, as in Davies et al. (2021). These works do not claim to perform "automated model discovery" but rather that they work with more general models which make fewer *a priori* assumptions than a hand-crafted model.

In this work we focus on the surrogate approach. The formal method we adopt is *symbolic regression* (SR) (Schmidt and Lipson, 2009). This procedure searches the equation space using a genetic algorithm to find algebraic (as well as some transcendental) functional relationships that closely replicate the performance of a given model, and at the same time are transparent to and interpretable by humans. Symbolic regression was used for GNNs in Cranmer et al. (2020) to discover modified physical laws of gravity in cosmological simulations, and this was one inspiration for the present work. We followed a similar approach, now combining SR with common sense and economic intuition to maximize the interpretability of our results.

4 Models

4.1 Theoretical framework

The *general* form of the gravity equation for trade is, following Head and Mayer (2014),

$$X(i \leftarrow j) =: X_{ij} = GM_i S_j \phi_{ij}, \tag{4}$$

where, with countries $i, j = 1, \ldots, K$ and time periods $t = 1, \ldots, T$ (the latter subscripts are left out for readability), the following notation is used:
$X_{ij} \in \mathbb{R}^+$ — trade volume directed from exporter j to importer i;
$M_i \in \mathbb{R}^+$ — characteristics of import market i that promote demand from all origins;
$S_j \in \mathbb{R}^+$ — capabilities of exporter j as a supplier to all destinations;
$\phi_{ij} \in [0, 1]$ — bilateral accessibility of importer i to exporter j;
$G \in \mathbb{R}^+$ — parameter that is constant $\forall i, j$ (and possibly $\forall t$).[18]

Furthermore, $\phi_{ij} := \tau_{ij}^{\varepsilon}$, with τ_{ij} denoting trade costs and ε denoting elasticity; while S_k and M_k are functions of ϕ_{ij} and country-specific variables.[19]

A large class of *structural* economic models specializes the gravity equation (4) with the following restrictions on S_k and M_k:

$$M_i := \frac{X_i}{\Phi_i}, \tag{5}$$

$$S_j := \frac{Y_j}{\Omega_j}, \tag{6}$$

$$\Phi_i = \sum_{k=1}^{K} \frac{\phi_{ik} Y_k}{\Omega_k}, \tag{7}$$

$$\Omega_j = \sum_{k=1}^{K} \frac{\phi_{kj} X_k}{\Phi_k}, \tag{8}$$

where $X_i := \sum_{j=1}^{K} X_{ij}$ is the value of the importer's expenditure on all origin countries (including itself);
$Y_j := \sum_{i=1}^{K} X_{ij}$ is the value of the exporter's production purchased by all destination countries (including itself);
Φ_i, Ω_j are multilateral resistance terms called market access, or market potential, functions. (They also subsume the constant G.)

The last two objects deserve some elaboration. Φ_i is an index of consumer market access, it measures the set of opportunities available to consumers in destination country

[18] In our data, some observations on X_{ij} (as well as X_i, Y_j) and ϕ_{ij} are missing, hence the variables that are available for our analysis are strictly positive quantities. The missing observations are implicitly equalized to 0.

[19] For example, an intuitive, naïve gravity equation implies $S_k = M_k = Y_k, \forall k = 1, \ldots, K$. An influential study by Anderson and van Wincoop (2003) assumes balanced trade ($X_i = Y_i$) and symmetric trade costs ($\phi_{ij} = \phi_{ji}$), which implies symmetric market access functions ($\Phi_i = \Omega_i$) and thus symmetric gravity equation ($S_i = M_i$).

i, and is understood as accessibility-weighted sum of the exporter capabilities S_j:

$$\Phi_i = \sum_k S_k \phi_{ik},$$

which reduces to a production-weighted average of relative potential (equation 7).

Ω_j is an index of supplier market access, it measures the set of opportunities available to producers in origin country j, and is understood as accessibility-weighted sum of the importer characteristics M_i:

$$\Omega_j = \sum_k M_k \phi_{kj},$$

which reduces to an expenditure-weighted average of relative access (equation 8).

Relationship between general and structural formulations: If the general model 4 holds exactly, it can always be recast in terms of a structural model (which involves economically intuitive observables, multivariate resistance terms Ω_j and Φ_i). Of course, empirically this is quite far from the truth, hence the two formulations are not equivalent. In the interest of completeness, let us give a brief argument for when these formulations coincide.

Let us write the general equation (4) in matrix format,

$$\boldsymbol{X} = G\boldsymbol{M}\boldsymbol{\phi}\boldsymbol{S}, \tag{9}$$

where $\boldsymbol{X}$ and $\boldsymbol{\phi}$ are rectangular matrices with elements (X_{ij}) and (ϕ_{ij}), respectively, while $\boldsymbol{M}$ and $\boldsymbol{S}$ are diagonal matrices with diagonal elements M_i and S_j.[20]

Consider the case when matrices $\boldsymbol{M}$ and $\boldsymbol{S}$ are full rank, as are $\boldsymbol{\Omega}$ and $\boldsymbol{\Phi}$. Then, the structural equations (5–8) can be written as

$$\begin{aligned} \mathbf{1}^\intercal \boldsymbol{\Omega} \boldsymbol{S} &:= \mathbf{1}^\intercal \boldsymbol{X}, & (10)\\ \boldsymbol{M}\boldsymbol{\Phi}\mathbf{1} &:= \boldsymbol{X}\mathbf{1}, & (11)\\ \mathbf{1}^\intercal \boldsymbol{\Omega} &= \mathbf{1}^\intercal \boldsymbol{M}\boldsymbol{\phi}, & (12)\\ \boldsymbol{\Phi}\mathbf{1} &= \boldsymbol{\phi}\boldsymbol{S}\mathbf{1}, & (13) \end{aligned}$$

where $\mathbf{1}$ is a column vector of ones (and where, say, $\boldsymbol{\Phi}$ subsumes the $1/G$ term).

Now, without loss of generality, rewrite the general equation (9) with a matrix of multiplicative mean-one residual terms $\boldsymbol{\varsigma}$ accounting for potential modeling errors:

$$\boldsymbol{X} = G\boldsymbol{M}\boldsymbol{\phi}\boldsymbol{S} \odot \boldsymbol{\varsigma}, \tag{14}$$

where $\odot$ stands for an element-wise (Hadamard) product.

[20] If $\boldsymbol{\phi}$ were a general matrix (even with ϕ_{ij} in $[0, 1]$) this would be an overparameterized model without content. It gains content by postulating a model which determines ϕ_{ij} in terms of country-specific data, as we discuss in later sections. However the argument here does not depend on this.

If we additionally postulate the presence of multilateral resistance terms,

$$\mathbf{1}^\intercal \check{\boldsymbol{\Omega}} \boldsymbol{S} := \mathbf{1}^\intercal \boldsymbol{X}, \tag{15}$$

$$\boldsymbol{M} \check{\boldsymbol{\Phi}} \mathbf{1} := \boldsymbol{X} \mathbf{1}, \tag{16}$$

then equations (14) and (15–16) produce

$$\mathbf{1}^\intercal \check{\boldsymbol{\Omega}} := \mathbf{1}^\intercal \boldsymbol{M} \boldsymbol{\phi} \boldsymbol{S} \odot \boldsymbol{\varsigma} \boldsymbol{S}^{-1}, \tag{17}$$

$$\check{\boldsymbol{\Phi}} \mathbf{1} := \boldsymbol{\phi} \boldsymbol{S} \odot \boldsymbol{\varsigma} \mathbf{1}. \tag{18}$$

Clearly, the last two equations coincide with the structural equations (12–13) if $\boldsymbol{\varsigma}$ is a matrix of ones (that is, if there are no modeling errors in the general gravity equation 14). More generally, the residuals $\boldsymbol{\varsigma}$ will break the equivalence between two model classes, differentiating them and modifying the corresponding estimation problems (as well as their optimal solutions).

4.2 Structural model

To start with, we consider the standard parametric formulation that boils down to equation (4) with structural restrictions (5)–(6):

$$\ln X_{ij,t} = \ln G + \ln \frac{X_{i,t}}{\Phi_{i,t}} + \ln \frac{Y_{j,t}}{\Omega_{j,t}} + \ln \phi_{ij,t} + \xi^X_{ij,t}, \tag{19}$$

where $\Phi_{i,t}$ and $\Omega_{j,t}$ are defined analogously to above, and $\xi^X_{ij,t}$ is a residual term.

Of course, there is no specific structural model directly matching such an abstract formulation, but this parametric implementation provides the upper bound on potential fit for the structural class of gravity models.

Crucially, the unknown and unobservable $\phi_{ij,t}$ is modeled as a latent variable represented by a NN, which is discussed next in §4.3.

4.3 Auxiliary neural network for bilateral accessibility

The unobservable $\phi_{ij,t}$ is assumed to be some unknown function of relevant variables, whose form is approximated by a NN, that is an unconstrained and flexible function of many theoretically motivated variables:

$$f^\phi(\boldsymbol{D}^\phi_{ij,t}) + \xi^\phi_{ij,t} \approx f^\phi(\boldsymbol{D}^\phi_{ij,t}), \tag{20}$$

where $f^\phi(\cdot)$ is the NN that approximates the unknown function $\mathring{f}^\phi$, $\boldsymbol{D}^\phi_{ij,t}$ is a vector of exogenous trade cost proxy variables and $\xi^\phi_{ij,t}$ is a residual term (with the whole expression ultimately mapped onto $[0, 1]$ interval).[21]

[21] Note that we compress several different trade cost proxies into a single-dimensional bilateral accessibility measure, but producing an intermediate multidimensional trade cost measure can also be easily done.

4.4 Flexible model

Now, we propose a more flexible (and effectively) nonparametric formulation that models equation (4) as a Graph Neural Network. Its form is captured by equation

$$\ln X_{ij,t} = \ln G + \ln M_{i,t} + \ln S_{j,t} + \ln \phi_{ij,t} + \xi^X_{ij,t}, \tag{21}$$

where $\ln \phi_{ij,t}$ is obtained as above in §4.3, $\xi^X_{ij,t}$ is a residual term, while $\ln M_{i,t}$ and $\ln S_{j,t}$ are modeled as functions of $X_{i,t}$, $Y_{j,t}$ as well as collections $\{\phi_{kj,t}, X_{k,t}\}_{k=1}^K$ and $\{\phi_{ik,t}, Y_{k,t}\}_{k=1}^K$, with the latter two multi-objects allowing us to account for some role of market access, i.e.,

$$\ln M_{i,t} \approx f^M \left(X_{i,t}, \{\phi_{ik,t}, Y_{k,t}\}_k, \{\phi_{kj.t}, X_{k,t}\}_k\right)$$

and

$$\ln S_{j,t} \approx f^S \left(Y_{j,t}, \{\phi_{kj,t}, X_{k,t}\}_k, \{\phi_{ik,t}, Y_{k,t}\}_k\right),$$

defining functions $f^M(\cdot)$ and $f^S(\cdot)$ above as NN approximations.

The aggregation function we choose to use is $\sum(\cdot)$, which is very natural for the application at hand, but satisfies the exchangeability requirement.[22] Apart from this, the outer and inner nonlinear functions are unconstrained and completely flexible. Thus, we specialize the above formulation to two alternatives.

First, along with the unconstrained form of the outer functions $f^M(\cdot)$ and $f^S(\cdot)$ that are NN approximations, we impose a multiplicative form on the inner nonlinear functions of ϕ_{ik} and Y_k as well as ϕ_{kj} and X_k:

$$\ln M_{i,t} \approx f^M \left(X_{i,t}, \sum_k \phi_{ik,t} Y_{k,t}, \sum_k \phi_{kj,t} X_{k,t}\right)$$

and

$$\ln S_{j,t} \approx f^S \left(Y_{j,t}, \sum_k \phi_{kj,t} X_{k,t}, \sum_k \phi_{ik,t} Y_{k,t}\right).$$

Second, in contrast to the above, we allow the inner functions $f^M_{\phi Y}(\cdot)$, $f^M_{\phi X}(\cdot)$, $f^S_{\phi X}(\cdot)$ and $f^S_{\phi Y}(\cdot)$ to be NN approximations as well:

$$\ln M_{i,t} \approx f^M \left(X_{i,t}, \sum_k f^M_{\phi Y}(\phi_{ik,t}, Y_{k,t}), \sum_k f^M_{\phi X}(\phi_{kj,t}, X_{k,t})\right)$$

and

$$\ln S_{j,t} \approx f^S \left(Y_{j,t}, \sum_k f^S_{\phi X}(\phi_{kj,t}, X_{k,t}), \sum_k f^S_{\phi Y}(\phi_{ik,t}, Y_{k,t})\right).$$

In the second alternative, we further allow $f^M_{\phi Y}(\cdot)$, $f^M_{\phi X}(\cdot)$, $f^S_{\phi X}(\cdot)$ and $f^S_{\phi Y}(\cdot)$ to range either on $\mathbb{R}$ or on $\mathbb{R}_+$.[23]

[22] Hence our model may also be viewed as a convolutional GNN.

[23] Although multilateral resistance terms $\Phi_{i,t}$ and $\Omega_{j,t}$ are not explicitly included in the Flexible model's formulation, implicitly they can still make an effect via the second and third arguments of the outer functions.

Lastly, note that, in parallel with the Structural model, inner functions $f^M_{\phi Y}(\cdot)$ and $f^S_{\phi Y}(\cdot)$ (as well as $f^M_{\phi X}(\cdot)$ and $f^S_{\phi X}(\cdot)$) can be conveniently thought of as measures of bilateral accessibility-mediated aggregate production (respectively, expenditures).

4.5 Fixed Effects model

In agreement with existing literature and as an additional check, we also take a formulation with fixed effects accounting for year and/or country identities.[24] It represents equation (4) by either

$$\ln X_{ij,t} = \beta^X_0 + \beta^X_t + \beta^X_i + \beta^X_j + \zeta^X_{ij,t}, \tag{22}$$

where β^X_t, β^X_i and β^X_j are year, country j and country i effects, respectively, and $\zeta^X_{ij,t}$ is a residual term subsuming both the bilateral accessibility variable $\ln \phi_{ij,t}$ and fitting errors (with appropriate restrictions on β^X_t, β^X_i and β^X_j); or by a richer alternative such as

$$\ln X_{ij,t} = \beta^X_0 + \beta^X_{i,t} + \beta^X_{j,t} + \zeta^X_{ij,t}, \tag{23}$$

where $\beta^X_{i,t}$ and $\beta^X_{j,t}$ are corresponding country-by-year effects, and $\zeta^X_{ij,t}$ is a residual term.

Lastly, we consider a hybrid model combining country-by-year fixed effects with an explicit bilateral accessibility variable $\ln \phi_{ij,t}$ (obtained similarly to §4.3):

$$\ln X_{ij,t} = \beta^X_0 + \beta^X_{i,t} + \beta^X_{j,t} + \ln \phi_{ij,t} + \xi^X_{ij,t}, \tag{24}$$

where the notation is unchanged from above except for the residual term $\xi^X_{ij,t}$, which no longer subsumes the bilateral accessibility term. Effectively, here is the general model (4), and this concrete implementation provides the upper bound on potential fit[25] for the general class of gravity models.

4.6 Remarks

Structural model is interesting because its formulation stems from economic theory, so (i) the restrictions it imposes, as long as they correctly reflect true causal structure, help to (efficiently) fit data, and (ii) it can be used for empirical tests of respective theories. The formulation we adopt here is fairly abstract and is satisfied by a large class of economic theories, but it can be further specialized to match concrete micro-founded theoretical models.

Flexible models presented above put a very adaptable and elastic framework on interactions between countries and interdependencies between economic variables characterizing each country. This lets economic data speak for itself, allowing to identify novel empirical regularities (thus helping "discover" new theoretical models). However, while not fully general, they are relatively unconstrained, having many "degrees of freedom"

[24] Essentially, this is the current workhorse approach in the literature.

[25] This concept is parallel to what is known as "oracle" in computer science.

to fit the data. Of course, given the vast expressive capacity of NNs that are employed there, predictive performance beyond the training data set and prevention of overfitting are important factors to keep in mind when implementing these models.

Fixed Effects models can be viewed here as "gold-standard" benchmarks. Existing studies recognize them as being fairly efficient at controlling for supply and demand market characteristics, providing an approximate measure of bilateral accessibility (albeit contaminated with model residuals in the standard approaches). As long as underlying theoretical causes can be abstracted away from and bilateral accessibility is the only object of interest, these models can do the job. This economic theory-agnostic setup is also their major weakness, prohibiting generalization beyond the training data set and requiring additional analysis for testing general equilibrium theoretical models.

The auxiliary NN model for bilateral accessibility relies on NN's universal approximation property to capture the functional representation of a broad collection of disparate proxies for trade costs (and their elasticity). It can be viewed as a "black box", ignoring complex arrangement and interrelationships inside it; but eliciting $\phi_{ij,t}$'s sensitivity to ingredient variables is also straightforward using the corresponding gradients, which are computed automatically by NN optimization algorithm.

5 Data

We have annual data on 181 countries over 68 years (1949–2016). Full data sample is split into training (1949–2013) and testing (2014–2016) subsamples.

From the modeling perspective, our data variables are of two types:

(a) country-specific node variables: contig, gdp, gatt;

(b) relationship-specific edge variables: comlang_off, comlang_ethno, comcol, col45, distw, tdiff, heg_o, heg_d, colony, curcol, comcur, comrelig, fta_wto, txg_fob_usd, tmg_cif_usd.

Basically, the former include each country's economic, geographic and political attributes (most importantly, GDP); the latter include economic, geographic, political and religious attributes (such as volume of imports, common currency, etc.) Variable definitions and data sources are available in Appendix §A.

Thus, X_{ij} stands for the value of imports of country i from country j that is represented by tmg_cif_usd; X_i stands for country i's expenditures and is represented by gdp plus its total imports tmg_cif_usd minus total exports txg_fob_usd; while Y_j stands for country j's production and is represented by gdp. Lastly, $\boldsymbol{D}_{ij}^{\phi}$ includes a broad collection of variables potentially related to bilateral accessibility ϕ_{ij}, specifically comlang_off, comlang_ethno, comcol, col45, distw, tdiff, heg_o, heg_d, colony, curcol, comcur, comrelig, fta_wto; as well as contig and gatt both for origin country j and destination country i.

6 Estimation details

Conceptually, we would like to estimate linear or non-linear (NN-based) approximations to some unknown functions of interest. Throughout the paper, $f^{\phi}(\cdot)$, $f^{M}(\cdot)$ and so on each designate a chosen NN specification (as a function of and thus indexed by hyperparameters defining its architecture, which is not made explicit for readability); they approximate unknown functions $\mathring{f}^{\phi}$, $\mathring{f}^{M}$ and so on (i.e., ϕ, $\ln M$, etc.); and their estimates are denoted by $\hat{f}^{\phi}(\cdot)$, $\hat{f}^{M}(\cdot)$ and so on.

The structural model in §4.2 is estimated by Nonlinear Least Squares: the Mean Squared Error criterion is minimized by Stochastic Gradient Descent. Regressor inputs are computed in the following way: the function $f^{\phi}(\cdot)$ for latent variable $\phi_{ij,t}$ from §4.3 is approximated by a NN with L hidden layers each containing N^{ϕ} neurons,[26] while objects $\Phi_{i,t}$ and $\Omega_{j,t}$ are recomputed at every iteration of the SGD algorithm. We consider NN architectures with hyperparameters $L \in \{1, 2, 3, 4\}$ and $N^{\phi} \in \{20, 50, 100\}$.[27]

Flexible models in §4.4 are also estimated by Nonlinear Least Squares, minimizing the MSE criterion by SGD. Regressor functions $f^{\phi}(\cdot)$ for $\phi_{ij,t}$, as well as $f^{M}(\cdot)$ and $f^{S}(\cdot)$—and, where necessary, $f^{M}_{\phi Y}(\cdot)$, $f^{M}_{\phi X}(\cdot)$, $f^{S}_{\phi X}(\cdot)$ and $f^{S}_{\phi Y}(\cdot)$—for $\ln M_{j,t}$ and $\ln S_{j,t}$ are each approximated by an L-layered NN with N^{b}_{a} neurons[28] in every layer.[29] We consider architectures with $L \in \{1, 2, 3, 4\}$, $N^{\phi} \in \{20, 50, 100\}$, and $N^{M}, N^{S}, N^{M}_{\phi X}, N^{M}_{\phi Y}, N^{S}_{\phi X}, N^{S}_{\phi Y} \in \{10, 20, 50\}$.

Given that nodes in our graph are very densely connected (i.e., every country trades with almost every other country), economic intuition suggests that direct trade flows inevitably dominate the outcomes. Hence, our GNNs are formulated with effectively a single iteration of message passing (this is relevant to the Flexible model only). [30]

Fixed Effects models in §4.5 are estimated by a variation of Ordinary Least Squares, minimizing the MSE criterion by SGD, with regressor inputs given by appropriately defined categorical dummy variables. The exception is the hybrid model, which combines the approach of other Fixed Effects models and an L-layered N^{ϕ}-neuron NN approxi-

[26] The return of NN $f^{\phi}(\cdot)$ belongs to $\mathbb{R}$, and is mapped onto $[0, 1]$ by applying to it the standard logistic function, which is defined as $S(x) := 1/1 + e^{-x}$. Formally, we model the relationship $S^{-1}(\phi_{ij,t}) = f^{\phi}(\boldsymbol{D}^{\phi}_{ij,t}) + \xi^{\phi}_{ij,t} \approx f^{\phi}(\boldsymbol{D}^{\phi}_{ij,t})$; naïvely recovering $\phi_{ij,t}$ with the non-linear transformation $S\left(f^{\phi}(\boldsymbol{D}^{\phi}_{ij,t})\right)$ poses obvious problems, but for "small" $\xi^{\phi}_{ij,t}$ they can be ignored.

[27] In theory, 1 hidden layer allows to achieve the universal approximation property, but in practice with finite number of neurons and bounded learning time deeper networks with more than 1 hidden layers often seem to perform better.

[28] Specifically, $a \in \{\varnothing, \phi X, \phi Y\}$, $b \in \{\phi, M, S\}$.

[29] The return of NN $f^{\phi}(\cdot)$ is on $\mathbb{R}$ and then goes through $S(\cdot)$; the returns of NNs $f^{M}(\cdot)$ and $f^{S}(\cdot)$ are on $\mathbb{R}$; while the returns of NNs $f^{M}_{\phi Y}(\cdot)$, $f^{M}_{\phi X}(\cdot)$, $f^{S}_{\phi X}(\cdot)$ and $f^{S}_{\phi Y}(\cdot)$ are either on $\mathbb{R}$ or $\mathbb{R}_{+}$, in the latter case this is achieved by adding an additional overlay and applying an $\exp(\cdot)$ function (e.g., formally we model the object $f^{M}_{\phi Y}(\cdot)$, but as an input to downstream functions provide its transformed counterpart $\exp\left(f^{M}_{\phi Y}(\cdot)\right)$).

[30] Such an abridgement also relieves us of the over-smoothing problem. For an example of using a single-iteration message passing, but in a more challenging setup, see Kampffmeyer et al. (2019).

Table 1: Estimation results, best models from each class

Model for $\ln X_{ij,t}$:	Structural	Flexible			Fixed Effects		
		$\times$	NN to $\mathbb{R}$	NN to $\mathbb{R}_+$	t, i, j	$t \times i$, $t \times j$	$t \times i$, $t \times j$, BA
L	1	4	3	4	-	-	1
N^ϕ	20	50	50	100	-	-	20
N^M, N^S	-	20	20	50	-	-	-
$N^M_{\phi X}$, $N^M_{\phi Y}$, $N^S_{\phi X}$, $N^S_{\phi Y}$	-	-	20	50	-	-	-
Parameters	382	11324	9774	63704	431	24617	24998
R^2 IS	0.53	0.61	0.65	0.68	0.54	0.62	0.74
R^2 OOS	0.51	0.60	0.65	0.68	-	-	-
Var. decomposition							
$\ln M_{i,t}$	0.21	0.18	0.19	0.21	0.18	0.22	0.24
$\ln S_{j,t}$	0.32	0.30	0.37	0.34	0.32	0.39	0.41
$\ln \phi_{ij,t}$	0.04	0.10	0.06	0.06	-	-	0.08
β^X_t	-	-	-	-	0.01	-	-
$\xi_{ij,t}$ or $\zeta_{ij,t}$	0.44	0.42	0.36	0.33	0.49	0.38	0.27

Notes: Number of observations for each model is 515345 in the training sample and 54355 in the testing sample. Number of epochs for each model is 1000. "-" stands for "not applicable". Models are listed in columns from left to right in the order of their introduction in text. Model architectures are captured by rows "L"; "N^ϕ"; "N^M, N^S"; "$N^M_{\phi X}$, $N^M_{\phi Y}$, $N^S_{\phi X}$, $N^S_{\phi Y}$"; and "Parameters". Goodness-of-fit measure R^2 is calculated for the training (IS) and testing (OOS) samples. Variance decomposition is performed for the training sample. See text for additional estimation details.

mation of the function $f^\phi(\cdot)$ for bilateral accessibility $\phi_{ij,t}$. For the latter, we consider architectures with hyperparameters $L \in \{1, 2, 3, 4\}$ and $N^\phi \in \{20, 50, 100\}$.

7 Results

7.1 Aggregate aspects

Results for the best model in each class are available in Table 1. Results for all the considered models are collected in Tables 7 and 8 of Appendix B.[31]

Architectures, fits, decompositions: Table 1 suggests that the Structural model requires just 1 hidden layer in its NN for bilateral accessibility $\phi_{ij,t}$, which is different from

[31]Note that an explicit time trend term is not included in our models. The results seem not to object: we do not see a common time trend across country pairs' residuals.

other model classes and is probably due to restrictions this model imposes on the way $\phi_{ij,t}$ can interact with other variables. Optimal architectures in Flexible model classes require 3-4 layers. Fixed Effects model with a bilateral accessibility term, similarly to Flexible models, also favors deep architectures; but given its already rich parameterizaton with time-varying country effects, we focus on the most minimalist specification (which performs very strongly nevertheless). The number of parameters used in top-performing Flexible and Fixed Effects models is relatively high; however, (i) as discussed earlier in §3.2, for NNs such overparameterization is not a problem but may actually be beneficial, and (ii) this is fairly common for Fixed Effects models applied to large panel data sets. It is worth emphasizing that the expressive capacity of NNs allows Flexible models to avoid using a plethora of country- and time-specific categorical dummy variables utilized in Fixed Effects models, which is an advantage since overabundance of categorical dummies may absorb some important effects and obstruct the detailed picture.

The goodness of fit, that is the amount of variance that can be explained in-sample (IS) by either of the three Flexible models is considerably higher than what is achieved by the Structural model (with our preferred Flexible architecture reaching 0.68, versus 0.53 in the case of Structural model); Flexible models also perform better out-of-sample (OOS). Fixed Effects models plays the role of benchmark for other model classes here, with time-varying country effects it exhibits the IS fit of 0.62, and even as high as 0.74 when we include a bilateral accessibility term (as mentioned earlier, they do not generalize to OOS by construction). It is encouraging to see that in terms of IS fit, our preferred Flexible model is not far behind the strongest Fixed Effects model.

The bottom panel of Table 1 presents calculations analogous to variance decomposition.[32] It is approximately applicable in our case since r.h.s. variables in Fixed Effects models are by construction mutually orthogonal, and they are close to being orthogonal in the other two model classes. One interesting observation is that $\ln M_{i,t}$ and $\ln S_{j,t}$ objects contribute less to explaining variance in the Structural and Flexible models than they do in two larger Fixed Effects models. Another observation is that for all models, contribution to explanatory performance of $\ln S_{j,t}$ is almost twice as large as that of $\ln M_{i,t}$. Lastly, the standalone bilateral accessibility term $\ln \phi_{ij,t}$ has a very modest contribution in all gravity models.

Supply and demand market characteristics: Table 2 compares $\ln S_{j,t}$ and $\ln M_{i,t}$ objects from our preferred Structural and Flexible models, and relates them to their counterparts from Fixed Effects models (that is, to $\beta^X_{j,t}$ and $\beta^X_{i,t}$). Both Flexible and Structural model outputs are highly correlated with the their Fixed Effects benchmarks,

[32] It is well know that for a linear regression equation $y_t = \sum_i z_{i,t} + \varepsilon_t$, where orthogonal variables $z_{i,t}$ are, for example, $\beta_i x_{i,t}$, the following variance decomposition relationship holds. Defining variance composition share $\text{Share}[y; z_i] := \text{Cov}[y, z_i]/\text{Var}[y]$, we have $\sum_i \text{Share}[y; z_i] + \text{Share}[y; \varepsilon] \equiv 1$, with all shares $\sum_i \text{Share}[y; z_i]$, by construction, summing to regression R^2. For a reference, see Klenow and Rodríguez-Clare (1997).

Table 2: Correlations, supply and demand market characteristics

	$\ln M_{i,t}$				$\ln S_{j,t}$			
Model for $\ln X_{ij,t}$:	Structural	Flex	FEs	FEs+BA	Structural	Flex	FEs	FEs+BA
Structural	1.00	0.91	0.85	0.83	1.00	0.94	0.89	0.86
Flex (NN to $\mathbb{R}_+$)	0.89	1.00	0.89	0.88	0.94	1.00	0.94	0.91
FEs ($t \times i$, $t \times j$)	0.83	0.88	1.00	0.94	0.89	0.93	1.00	0.95
FEs+BA ($t \times i$, $t \times j$)	0.81	0.87	0.93	1.00	0.85	0.91	0.94	1.00
Structural	1.00	0.95	-	-	1.00	0.97	-	-
Flex (NN to $\mathbb{R}_+$)	0.96	1.00	-	-	0.98	1.00	-	-
FEs ($t \times i$, $t \times j$)	-	-	-	-	-	-	-	-
FEs+BA ($t \times i$, $t \times j$)	-	-	-	-	-	-	-	-

Notes: Pearson's correlation coefficients are presented in the upper-right corner, Spearman's correlation coefficients are in the lower-left corner. Results for the training sample are presented in the top panel, for the testing sample are in the bottom panel. "-" stands for "not applicable".

but this interrelationship is especially strong for the Flexible model.

Bilateral accessibility: One of the main practical uses of gravity models for international trade is measuring bilateral accessibility, its driving factors, impact on other variables, time evolution. We can get a sense of its aggregate behavior over time from Figure 1.[33] An interesting finding of our analysis is that bilateral accessibility for imports and exports between a given pair of countries is asymmetric[34], which is evident from plots such as those in Figures 3 and 4 (and is broadly confirmed by statistical comparisons).

Table 3 compares bilateral accessibility across different models. First, $\ln \phi_{ij,t}$ correlations for Flexible and Structural models are high, but far from perfect (both Pearson's and Spearman's correlation coefficients are close to 0.9 in either training or testing sample), which confirms the differences between models and is reflected in, say, Figure 1. Second, relating bilateral accessibility objects from the Structural and Flexible models to their counterpart from the Fixed Effects model (that is, comparing $\ln \phi_{ij,t}$ or, more conservatively, $\ln \phi_{ij,t} + \xi^X_{ij,t}$, to $\zeta^X_{ij,t}$) shows that for both Structural and Flexible models these objects are highly correlated with their Fixed Effects benchmarks, but this is especially true for the Flexible model (i.e., both Pearson's and Spearman's correlations are 5-10 percentage points higher in the case of Flexible model).

[33]Figures mentioned in this paragraph are discussed in more detail in the next subsection.

[34]That is, $\phi_{ij,t}$ and $\phi_{ji,t}$ are not equal to each other for $i \neq j$.

Table 3: Correlations, bilateral accessibility

	$\ln \phi_{ij,t}$ or $\zeta^X_{ij,t}$				$(\ln \phi_{ij,t} + \xi^X_{ij,t})$ or $\zeta^X_{ij,t}$			
Model for $\ln X_{ij,t}$:	Structural	Flex	FEs	FEs+BA	Structural	Flex	FEs	FEs+BA
Structural	1.00	0.85	0.43	0.69	1.00	0.91	0.83	0.79
Flex (NN to $\mathbb{R}_+$)	0.87	1.00	0.51	0.79	0.88	1.00	0.89	0.86
FEs ($t \times i$, $t \times j$)	0.40	0.45	1.00	0.42	0.78	0.85	1.00	0.91
FEs+BA ($t \times i$, $t \times j$)	0.63	0.76	0.38	1.00	0.73	0.83	0.88	1.00
Structural	1.00	0.88	-	-	1.00	0.92	-	-
Flex (NN to $\mathbb{R}_+$)	0.91	1.00	-	-	0.89	1.00	-	-
FEs ($t \times i$, $t \times j$)	-	-	-	-	-	-	-	-
FEs+BA ($t \times i$, $t \times j$)	-	-	-	-	-	-	-	-

Notes: Pearson's correlation coefficients are presented in the upper-right corner, Spearman's correlation coefficients are in the lower-left corner. Results for the training sample are presented in the top panel, for the testing sample are in the bottom panel. "-" stands for "not applicable".

7.2 Particularities

Now we zoom into particular economic situations in the past to inspect how different models behaved there.

Full sample: To provide a reference point, Figure 1 depicts the aggregate time series of (i) bilateral accessibility $\phi_{ij,t}$ for the Structural, our preferred Flexible and the Fixed Effects model that explicitly includes it as well as (ii) its closest counterparts $\zeta^X_{ij,t}$ for the plain Fixed Effects model and $(\ln \phi_{ij,t} + \xi_{ij,t})$ for the other three models, taking the same subset of 448 trading directions between 40 countries and tracking them for the whole training sample. We can see that bilateral accessibility improves over time for the Flexible model as well as for the Fixed Effects model including it, but it deteriorates for the Structural model and starts improving only in early 1990s with the start of "Globalization". We are not able to separately identify this object for the plain Fixed Effects model, but from the right panel it seems to improve over time until 1990s, when it starts deteriorating. The behavior exhibited by the Flexible and Fixed Effects model with bilateral accessibility is in line with our understanding of the actual events.

Financial crisis and Great recession: Figure 2 depicts the aggregate dynamics of bilateral accessibility at times of the 2008-2009 financial crisis and ensuing "Great recession". In spite of narratives about protectionism as political responses to economic difficulties,[35] which would have been manifested in measures of bilateral accessibility (specifically, in effective bilateral accessibility that comprises the residual term as captured by the Fig-

[35]For some details, see Evenett (2009).

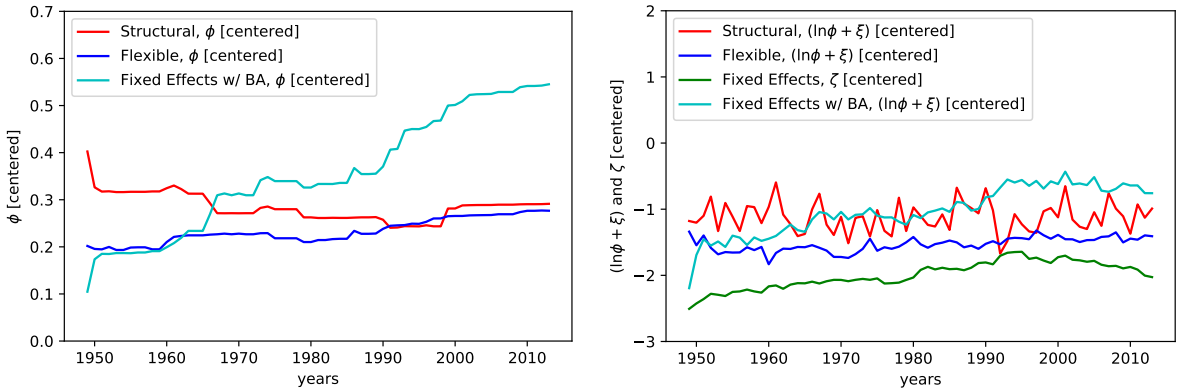

Figure 1: Aggregate bilateral accessibility: Full training sample.
[Bilateral accessibility as a separate variable, $\phi_{ij,t}$, as well as bilateral accessibility amalgamated with gravity equation's residuals, $(\ln\phi_{ij,t} + \xi_{ij,t})$ or $\zeta^X_{ij,t}$, implied by different models (Flexible, Structural, Fixed Effects, Fixed Effects with bilateral accessibility); centering done by, respectively, multiplying/adding G, $M_{i,t}$, $S_{j,t}$ averaged over the whole training sample, thus producing $G\bar{M}\bar{S}\phi$ as well as $(\ln G\bar{M}\bar{S} + \ln\phi + \xi)$ or $(\ln G\bar{M}\bar{S} + \zeta)$; values are arithmetic averages over 488 trade directions (edges) between 40 countries; time period used is 1949–2013.]

ure's right panel, since in our application "institutional" variables entering function $f^{\phi}(\cdot)$ are too slow-moving), we do not observe such effects in any of our models' output.

EU enlargement: While the previous paragraph considers virtually all available countries over time, we consider next a small subset of countries, differentiating them into two sub-classes. Figures 3 and 4 depicts the aggregate dynamics of bilateral accessibility during the process of European Union [EU] enlargement over 2004–2013 based on imports by the "old" EU members from the "new" EU members as well as imports by the "new" members from the "old" ones, respectively. Economic integration is supposed to improve bilateral accessibility, which indeed happens to be the case: it is clearly seen for the Structural and Fixed Effects model with bilateral accessibility, although it seems to have much more modest effect for the Flexible model. Looking into less clean measures of accessibility on the right panels, we can see that all four models exhibit an improvement.[36]

8 Interpreting the "black box"

Interpretation of neural networks is a challenging task. To aid with it, we are using a procedure called *symbolic regression* (Schmidt and Lipson, 2009; Cranmer et al., 2020; Cranmer, 2020), which searches the equation space using a genetic algorithm to find algebraic (and some transcendental) functional relations that approximate our model's

[36] A much broader recent study on this theme is Head and Mayer (2021).

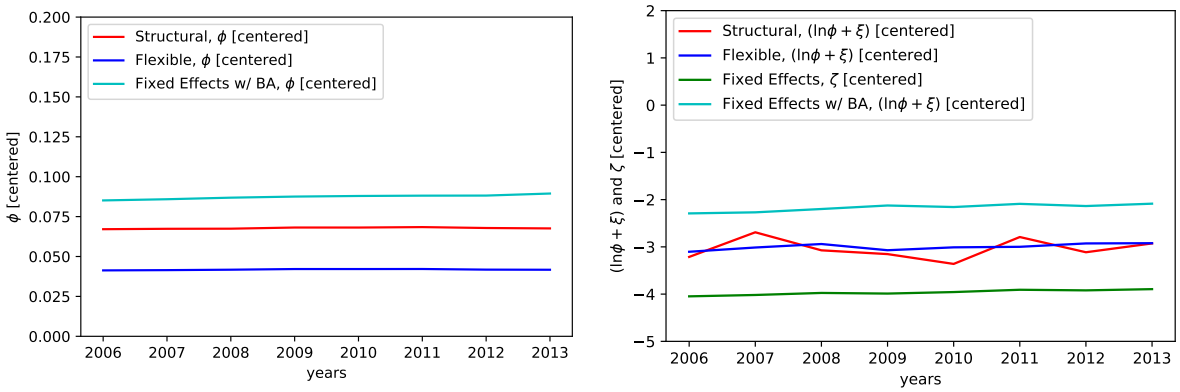

Figure 2: Aggregate bilateral accessibility: Financial crisis and Great recession.
[Bilateral accessibility as a separate variable, $\phi_{ij,t}$, as well as bilateral accessibility amalgamated with gravity equation's residuals, $(\ln\phi_{ij,t} + \xi_{ij,t})$ or $\zeta^X_{ij,t}$, implied by different models (Flexible, Structural, Fixed Effects, Fixed Effects with bilateral accessibility); centering done by, respectively, multiplying/adding G, $M_{i,t}$, $S_{j,t}$ averaged over the whole training sample, thus producing $G\bar{M}\bar{S}\phi$ as well as $(\ln G\bar{M}\bar{S} + \ln\phi + \xi)$ or $(\ln G\bar{M}\bar{S} + \zeta)$; values are arithmetic averages over 13542 trade directions (edges) between 180 countries; time period used is 2006–2013.]

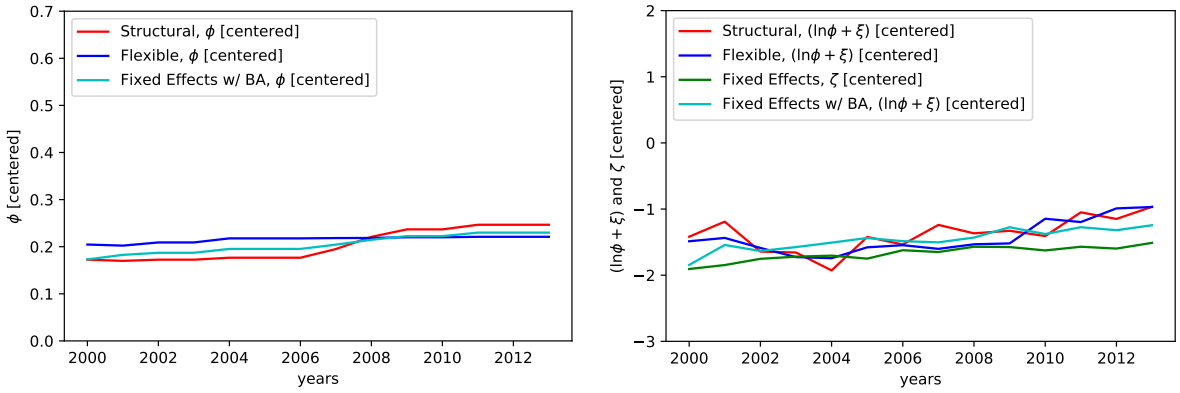

Figure 3: Aggregate bilateral accessibility: EU enlargement, imports by old members from new members.
[Bilateral accessibility as a separate variable, $\phi_{ij,t}$, as well as bilateral accessibility amalgamated with gravity equation's residuals, $(\ln\phi_{ij,t} + \xi_{ij,t})$ or $\zeta^X_{ij,t}$, implied by different models (Flexible, Structural, Fixed Effects, Fixed Effects with bilateral accessibility); centering done by, respectively, multiplying/adding G, $M_{i,t}$, $S_{j,t}$ averaged over the whole training sample, thus producing $G\bar{M}\bar{S}\phi$ as well as $(\ln G\bar{M}\bar{S} + \ln\phi + \xi)$ or $(\ln G\bar{M}\bar{S} + \zeta)$; values are arithmetic averages over 195 trade directions (edges) between European Union countries (15 old and 13 new members); time period used is 2000–2013.]

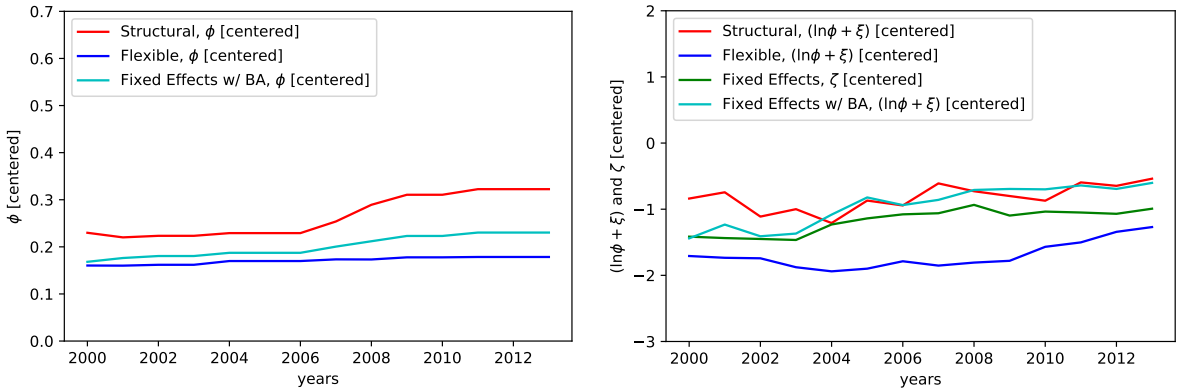

Figure 4: Aggregate bilateral accessibility: EU enlargement, imports by new members from old members.

[Bilateral accessibility as a separate variable, $\phi_{ij,t}$, as well as bilateral accessibility amalgamated with gravity equation's residuals, $(\ln\phi_{ij,t} + \xi_{ij,t})$ or $\zeta^X_{ij,t}$, implied by different models (Flexible, Structural, Fixed Effects, Fixed Effects with bilateral accessibility); centering done by, respectively, multiplying/adding G, $M_{i,t}$, $S_{j,t}$ averaged over the whole training sample, thus producing $G\bar{M}\bar{S}\phi$ as well as $(\ln G\bar{M}\bar{S} + \ln\phi + \xi)$ or $(\ln G\bar{M}\bar{S} + \zeta)$; values are arithmetic averages over 194 trade directions (edges) between European Union countries (15 old and 13 new members); time period used is 2000–2013.]

performance, and at the same time can be inspected and (potentially) understood by humans.

This procedure is applied to our preferred Flexible specification, and our implementation includes the following steps.

First, denoting by $\hat{f}(\boldsymbol{z}, \boldsymbol{\theta})$ with a constant parameter vector $\boldsymbol{\theta}$ and explanatory input vector $\boldsymbol{z}$ an estimate of NN approximation $f(\cdot)$ (i.e., $f^M(\cdot)$ and $f^S(\cdot)$ as well as $f^M_{\phi Y}(\cdot)$, $f^M_{\phi X}(\cdot)$, $f^S_{\phi X}(\cdot)$ and $f^S_{\phi Y}(\cdot)$ with their respective arguments), we take such estimates and apply the SR algorithm to each of them separately. We view the bilateral accessibility ϕ_{ij} as an unobserved latent variable and approximate its functional relationship to explanatory variables $\mathring{f}^{\phi}$ by a NN $f^{\phi}(\cdot)$, but we do not aim to identify internal mechanics of the latter and do not apply the SR algorithm to its estimate $\hat{f}^{\phi}(\cdot)$, treating $\hat{\phi}_{ij}$ in this exercise as a given input.

Second, we take the raw output of this machine learning algorithm and use our human common sense and intuition to choose function specifications that balance accuracy against complexity (but also maintaining plausible symmetries).

Third, the raw functional specifications usually include some constant parameters. We re-estimate their values when fitting the final symbolic interpretation of our NN model. Before that, we insert additional free parameters in places where functional forms specify coefficients of "1" or where the additive intercept is set to "0" (our additions are weakly redundant and the final optimization algorithm is allowed to set them back to "1" or "0"). This gives us symbolic interpretations of estimated NN approximations $\hat{f}(\boldsymbol{z}, \boldsymbol{\theta})$

(and, presumably, of NN approximations $f(\cdot)$ as well as, hopefully, of unknown functions $\mathring{f}$ themselves), generally denoted as $\tilde{f}(\boldsymbol{z}, \boldsymbol{\vartheta})$ for a constant parameter vector $\boldsymbol{\vartheta}$ and explanatory input vector $\boldsymbol{z}$.

An illustrative output of symbolic regression with top candidate equations ranked by complexity is available in Appendix C. Reassuringly, we ran the algorithm with many restarts, and a lot of functional forms appeared again and again in independent runs. The final results following the steps described above are presented in Table 4.

In terms of model performance, symbolic model fits are quite accurate, with the overall encompassing model's symbolic version approaching the numeric original in terms of accuracy. This is achieved with a vastly smaller number of parameters.

Qualitatively speaking, the first aspect that catches our eyes is that SR independently learned that the logarithmic transformation is warranted for outer functions. At the same time, it notably chose not to "undo" with a logarithm the exponential transformation imposed manually on inner functions' outputs (in order to ensure their mapping to $\mathbb{R}_+$). Quantitatively, the salient feature here is that elasticities of trade flows with respect to supply and demand capacities (i.e., the leading pre-multiplying coefficients in the outer functions) are close to 1, as is commonly found in the literature.

Even though our Flexible model is, in general, different from the Structural model, it is instructive to compare the two. Overall, the "discovered" functions seem to depart substantially from the Structural specifications.[37] Nevertheless, in the outer functions we also find a kind of multilateral resistance terms; albeit without the need for the $\sum_k f^b_{\phi X}(\cdot)$, $b \in \{M, S\}$, object determining the supplier market access in the Structural model, which is particularly notable in the case of supplier capabilities $f^S(\cdot)$ formulation. The effect of $\sum_k f^b_{\phi X}(\cdot)$, $b \in \{M, S\}$, term in outer functions is quantitatively small, however (also, it is hard to justify the exponent of it that shows up in a multiplicative gravity equation); while the effect of $\sum_k f^b_{\phi Y}(\cdot)$, $b \in \{M, S\}$, is much larger.

As to the inner functions, they look very different from what we have seen in the Structural models. Here, only the effect of aggregated exporter's production Y_k variable is quantitatively important, with an ultimately positive marginal effect on the gravity equation's l.h.s. In particular, the role of bilateral accessibility ϕ_{ij} is quantitatively small in the inner functions, effectively restricting its impact to be constant across country pairs and time periods. This may be due to the fact that ϕ_{ij} is not too volatile overall, across both country pairs and years: on a fixed set of 488 bilateral country edges across 65 years, its coefficient of variation[38] is a modest 0.17. Also, it is highly correlated across country pairs: its first principal component explains 74% of variance (and the second one only 4%). (Of course, ϕ_{ij} still plays a role as a standalone term; its contribution to explained variance is fairly small though, see table 1).

[37] Some of the additive constant parameters, as those seen in the denominators of $f^M(\cdot)$ and $f^S(\cdot)$, perform a relatively innocuous centering of input variables, and we leave aside their discussion.

[38] A coefficient of variation of a random variable is defined as its standard deviation divided by the mean.

Table 4: Interpretation by unrestricted Symbolic Regression, best Flexible model

Model specification	R^2 IS	R^2 OOS
$\ln X_{ij} = \ln G + f_i^M + f_j^S + \ln \phi_{ij} + \xi_{ij}^X$	0.61	0.58

Composition

Estimated approximations		Symbolic form		
Function, $\hat{f}(\boldsymbol{z}, \boldsymbol{\theta})$	Arguments, $\boldsymbol{z}$	Function, $\tilde{f}(\boldsymbol{z}, \boldsymbol{\vartheta})$	R^2 IS	MSE IS
$\hat{f}^M(\cdot)$	X_i, $\sum_k \hat{f}_{\phi Y}^M(\cdot)$, $\sum_k \hat{f}_{\phi X}^M(\cdot)$	$\vartheta_1 \ln(\frac{\vartheta_{10}+\vartheta_{11} z_1}{\vartheta_{20}+\vartheta_{21} z_2}) + \vartheta_3 z_3$	0.96	0.15
$\hat{f}^S(\cdot)$	Y_j, $\sum_k \hat{f}_{\phi X}^S(\cdot)$, $\sum_k \hat{f}_{\phi Y}^S(\cdot)$	$\vartheta_1 \ln(\frac{\vartheta_{10}+\vartheta_{11} z_1}{\vartheta_{30}+\vartheta_{31} z_3}) + \vartheta_2 z_2$	0.95	0.21
$\ln \hat{f}_{\phi X}^M(\cdot)$, $\ln \hat{f}_{\phi X}^S(\cdot)$	$\hat{\phi}_{kj}$, X_k	$\vartheta_0 + \vartheta_1 z_1 + \vartheta_2 z_2$	0.99	70.28
$\ln \hat{f}_{\phi Y}^M(\cdot)$, $\ln \hat{f}_{\phi Y}^S(\cdot)$	$\hat{\phi}_{ik}$, Y_k	$\vartheta_0 + \frac{\vartheta_{10}+\vartheta_{11} z_1}{\vartheta_{20}+\vartheta_{21} z_2}$	0.98	62.51

Descriptive statistics		Numeric form
Mean [St.Dev.]	Mean [St.Dev.]	Function, $\tilde{f}(\boldsymbol{z}, \hat{\boldsymbol{\vartheta}})$
-1.39 $[1.97]$	$150.47, 16.95, 7.19$ $[790.89, 13.40, 7.22]$	$0.9853 \ln(\frac{0.1818+1.2808 z_1}{5.0208+0.5757 z_2}) + 0.0570 z_3$
-2.05 $[2.16]$	$149.71, 3.87, 19.61$ $[771.65, 4.75, 14.17]$	$1.1857 \ln(\frac{0.8767+0.9773 z_1}{6.7095+0.6026 z_3}) - 0.0755 z_2$
$-30.53, -30.53$ $[80.24, 80.24]$	$0.06, 355.45$ $[0.18, 1259.07]$	$-0.0042 - 0.0041 z_1 + 0.0002 z_2$
$-37.17, -37.17$ $[57.90, 57.90]$	$0.06, 345.01$ $[0.18, 1222.06]$	$-0.9710 + \frac{0.0014+0.0006 z_1}{1.2024+0.2622 z_2}$

Notes: Number of observations for each component is 515345 in the training sample and 54355 in the testing sample (the latter sample is not used separately for outer and inner functions). Number of epochs for each component is 1000. Bilateral accessibility term $\ln \phi_{ij}$ is taken as an exogenous input. Parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ as well as input variable vectors $\mathbf{z}$ are component-specific. Time subscripts are omitted throughout. See text for additional details on symbolic regression.

To sum up, some of the "discovered" functional forms and relationships resemble those present in the existing theoretical and empirical literature.

As a result of imposing an exponential transformation on the inner functions' outputs, the additive terms $\sum_k f^b_{\phi X}(\cdot)$, $b \in \{M, S\}$, in the outer functions enter the final (multiplicative) gravity equation in an economically dubious form. The exponential transformation, which is defined by a continuous and differentiable everywhere function, was used for streamlining the optimization algorithm rather than for reasons stemming from economic theory. At the same time, the inner functions' formulations favored by the SR algorithm are also not entirely satisfactory from an economic standpoint. Therefore, we consider the alternative, arguably more natural for our application, that is offered by a general Constant Elasticity of Substitution (CES) function.[39] Effectively, we take the estimated NN approximations and fit them with the CES specification in place of the specification based on a linear/fractional inner function formulation followed by an exponential transformation, testing the hypothesis that the latter is in fact the true ingredient of the unknown data-generating function. Table 5 summarizes this exercise.

With the CES restriction, inner functions fit slightly worse than in the original symbolic formulation from Table 4. However, the overall performance of the encompassing symbolic model is now even stronger and exceeds that of the unrestricted symbolic alternative both IS and OOS.[40]

In contrast to the original formulation in Table 4, inner functions $f^M_{\phi Y}(\cdot)$ and $f^S_{\phi Y}(\cdot)$ (as well as $f^M_{\phi X}(\cdot)$ and $f^S_{\phi X}(\cdot)$, which are still negligible in quantitative sense) now combine bilateral accessibility ϕ_{ik} (ϕ_{kj}) and production Y_k (expenditures X_k) with positive signs and in a complementary manner, although not in a simple multiplicative form. This agrees with economic intuition, and is also found in the Structural model.

[39] CES utility functions are ubiquitous in the gravity literature, while CES production functions (as well as their linear, Cobb-Douglas or Leontief special cases) are equally common in international and macroeconomics. We use a generalized CES function which further relaxes standard constraints on its coefficients.

[40] In principle, the CES-based formulation is not an explicit interpretation of our Flexible NN (c.f. imperfect inner functions' fits) but, strictly speaking, a different model. Hence, the overall performances of this new formulation and the Flexible model or its original symbolic interpretation are not cleanly comparable.

Table 5: Interpretation by Symbolic Regression with CES restriction, best Flexible model

Model specification			
		R^2 IS	R^2 OOS
$\ln X_{ij} = \ln G + f_i^M + f_j^S + \ln \phi_{ij} + \xi_{ij}^X$		0.66	0.67
Composition			
Estimated approximations		Symbolic form	
Function, $\hat{f}(\boldsymbol{z}, \boldsymbol{\theta})$	Arguments, $\boldsymbol{z}$	Function, $\tilde{f}(\boldsymbol{z}, \boldsymbol{\vartheta})$	R^2 IS / MSE IS
$\hat{f}^M(\cdot)$	X_i, $\sum_k \hat{f}_{\phi Y}^M(\cdot)$, $\sum_k \hat{f}_{\phi X}^M(\cdot)$	$\vartheta_1 \ln(\frac{\vartheta_{10}+\vartheta_{11} z_1}{\vartheta_{20}+\vartheta_{21} z_2}) + \vartheta_3 z_3$	0.96 / 0.15
$\hat{f}^S(\cdot)$	Y_j, $\sum_k \hat{f}_{\phi X}^S(\cdot)$, $\sum_k \hat{f}_{\phi Y}^S(\cdot)$	$\vartheta_1 \ln(\frac{\vartheta_{10}+\vartheta_{11} z_1}{\vartheta_{30}+\vartheta_{31} z_3}) + \vartheta_2 z_2$	0.95 / 0.21
$\ln \hat{f}_{\phi X}^M(\cdot)$, $\ln \hat{f}_{\phi X}^S(\cdot)$	$\hat{\phi}_{kj}$, X_k	$\ln(\vartheta_0(\vartheta_{10} z_1^{\vartheta_{11}} + \vartheta_{20} z_2^{\vartheta_{21}})^{\vartheta_1})$	0.90 / 664.83
$\ln \hat{f}_{\phi Y}^M(\cdot)$, $\ln \hat{f}_{\phi Y}^S(\cdot)$	$\hat{\phi}_{ik}$, Y_k	$\ln(\vartheta_0(\vartheta_{10} z_1^{\vartheta_{11}} + \vartheta_{20} z_2^{\vartheta_{21}})^{\vartheta_1})$	0.96 / 118.91
Descriptive statistics		Numeric form	
Mean [St.Dev.]	Mean [St.Dev.]	Function, $\tilde{f}(\boldsymbol{z}, \hat{\boldsymbol{\vartheta}})$	
-1.39 [1.97]	$150.47, 16.95, 7.19$ [$790.89, 13.40, 7.22$]	$1.0055 \ln(\frac{0.0618+1.9559 z_1}{11.7018+0.9271 z_2}) + 0.4388 z_3$	
-2.05 [2.16]	$149.71, 3.87, 19.61$ [$771.65, 4.75, 14.17$]	$0.9763 \ln(\frac{0.7828+1.5192 z_1}{7.5665+0.1926 z_3}) - 0.0435 z_2$	
$-30.53, -30.53$ [$80.24, 80.24$]	$0.06, 355.45$ [$0.18, 1259.07$]	$\ln(0.1986(6.4708 z_1^{15.0746} + 0.1988 z_2^{0.0162})^{1.1231})$	
$-37.17, -37.17$ [$57.90, 57.90$]	$0.06, 345.01$ [$0.18, 1222.06$]	$\ln(0.2979(0.0050 z_1^{1.4748} + 0.2335 z_2^{0.8234})^{0.9847})$	

Notes: Number of observations for each component is 515345 in the training sample and 54355 in the testing sample (the latter sample is not used separately for outer and inner functions). Number of epochs for each component is 1000. Bilateral accessibility term $\ln \phi_{ij}$ is taken as an exogenous input. Parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ as well as input variable vectors $\mathbf{z}$ are component-specific. Time subscripts are omitted throughout. See text for additional details on symbolic regression and CES specification.

9 Conclusions

We constructed several gravity models for international trade based on a GNN framework, producing results of both theoretical and empirical interest. Our Flexible NN models avoid imposing the strong assumptions stemming from economic theory which are made for Structural models, and even so they allow us to identify important empirical economic relationships. In terms of IS fit, our preferred Flexible model (i) outperforms the best Structural alternative, as well as (ii) outperforms the standard Fixed Effects model, while not lagging far behind the NN-enhanced Fixed Effects model. In terms of generalization to OOS, it maintains this strong performance (which turns out not to be the case for the Structural model, and is by construction precluded for the Fixed Effects models).

In the process of fitting our Flexible model (as well as its Structural alternative), as a by-product we obtain a model-consistent measure of bilateral accessibility for a broad set of country pairs and years, which may be used in stand-alone studies. Additionally, we produce bilateral accessibility measure utilizing a hybrid Fixed Effects model that includes the corresponding term (also utilizing a NN approximation, similarly to other models considered). The latter, relatively agnostic alternative may be useful to practical economists interested exclusively in an accurate bilateral accessibility measure that controls for time-varying exporter and importer country effects without explicit modeling of supply and demand market characteristics.

Focusing on the models' properties, we observe that the gravity equation's components such as supply and demand market characteristics as well as bilateral accessibility for Structural and, especially, Flexible models are closely correlated with their Fixed Effects benchmarks. Moreover, across all models considered the supply market capabilities component contributes to explaining the disproportionate amount of variance, almost double that of its demand market counterpart. A quantitatively dominant role of the bilateral accessibility-mediated aggregate production term versus that of aggregate expenditures is another asymmetry that suggests the leading role of the supply side of the market in our results. An additional interesting observation is the asymmetry of bilateral accessibility measures implied by our models (as well as their relatively small direct contribution to explanatory performance).

Zooming deeper into our preferred Flexible model's properties, we tried to interpret its NN modules using the Symbolic Regression algorithm. Taking the outputs of SR and to some extent relying on our intuition, we select a low-complexity collection of standard mathematical functions (and their compositions) that successfully replicate the bulk of our Flexible model's predictive performance (and, presumably, its internal mechanics). Interestingly, a symbolic interpretation of our estimated Flexible model obtained in this way seems to reinvent and refine some of Structural model's internal components such as multilateral resistance terms, while outperforming the latter in IS and OOS predictions. In addition to that, it also exhibits features that do not seem to have analogues in the Structural model. For instance, the obtained functions representing bilateral accessibility-

mediated aggregate production and expenditures are somewhat counter-intuitive. Thus, we also consider their alternative specifications provided by a well-known CES function, which produces economically intuitive relationships between the input variables and demonstrates an even higher explanatory performance.

As an illustration and a real-life check, we apply the assortment of our gravity models to several specific economic events. The following findings are worth highlighting: (i) only our preferred Flexible model as well as the Fixed Effects model with bilateral accessibility term demonstrate a more or less steady improvement in accessibility over time, which agrees with common economic sense; (ii) during the 2008–2009 financial crisis and the ensuing Great Recession, none of the models exhibit a discernible aggregate shock to bilateral accessibility (even if gravity equations' residuals are used as a measure of effective accessibility); (iii) during the European Union enlargement period of 2004–2013 all of our models show, in line with economic sense, an improvement in bilateral accessibility, which is especially pronounced for the Structural and Fixed-Effects-with-bilateral-accessibility models.

The flexibility and ease of use of NNs will, in our opinion, lead to more and more demand for interpretation of their inner workings. On the theoretical and methodological side, it seems promising to push further the automatic approach to NN interpretation taken in this paper. With a more applied agenda in mind, it can be fruitful to further study the results produced by our interpretation of the NN-based gravity model, and to consider their potential contributions to new theoretical developments in this area. On the empirical side, the new bilateral accessibility measures constructed here may prove to be advantageous for the practice of international economics and trade.

References

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

Adão, R., Arkolakis, C., and Ganapati, S. (2020). Aggregate implications of firm heterogeneity: A nonparametric analysis of monopolistic competition trade models. Manuscript.

Adão, R., Costinot, A., and Donaldson, D. (2017). Nonparametric counterfactual predictions in neoclassical models of international trade. *American Economic Review*, 107(3):633–89.

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

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

Augasta, M. and Kathirvalavakumar, T. (2012). Rule extraction from neural networks - a comparative study. In *Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering*, pages 404–408. IEEE.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Bastani, O., Kim, C., and Bastani, H. (2017). Interpreting blackbox models via model extraction. *CoRR*, abs/1705.08504.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Bronstein, M. M., Bruna, J., Cohen, T., and Velickovic, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478.

Chaney, T. (2018). The gravity equation in international trade: An explanation. *Journal of Political Economy*, 126(1):150–177.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models, chapter 76. volume 6 of *Handbook of Econometrics*, pages 5549–5632. Elsevier.

Costinot, A., Donaldson, D., Kyle, M., and Williams, H. (2019). The More We Die, The More We Sell? A Simple Test of the Home-Market Effect*. *The Quarterly Journal of Economics*, 134(2):843–894.

Costinot, A. and Rodríguez-Clare, A. (2014). Trade theory with numbers: Quantifying the consequences of globalization. In Gopinath, G., Helpman, E., and Rogoff, K., editors, *Handbook of International Economics*, volume 4 of *Handbook of International Economics*, chapter 4, pages 197–261. Elsevier.

Cranmer, M. (2020). Pysr: Fast & parallelized symbolic regression in python/julia.

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S. (2020). Discovering symbolic models from deep learning with inductive biases.

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

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., Lackenby, M., Williamson, G., Hassabis, D., and Kohli, P. (2021). Advancing mathematics by guiding human intuition with ai. *Nature*, 600(600):70–74.

Diestel, R. (2010). *Graph Theory*. Graduate Texts in Mathematics, Volume 173. Springer-Verlag, Heidelberg, 4th edition.

Dimopoulos, Y., Bourret, P., and Lek, S. (1995). Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Processing Letters*, 2(6):1–4.

Donaldson, D. (2015). The gains from market integration. *Annual Review of Economics*, 7(1):619–647.

Donaldson, D. and Hornbeck, R. (2016). Railroads and American Economic Growth: A "Market Access" Approach. *The Quarterly Journal of Economics*, 131(2):799–858.

Dütting, P., Feng, Z., Narasimhan, H., Parkes, D. C., and Ravindranath, S. S. (2021). Optimal auctions through deep learning. *Communications of the ACM*, 64(8):109–116.

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

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL.

Even, S. and Even, G. (2012). *Graph Algorithms, Second Edition*. Cambridge University Press.

Evenett, S. J. (2009). Crisis-era protectionism one year after the washington g20 meeting: A gta update, some new analysis, and a few words of caution. In Baldwin, R. E., editor, *The Great Trade Collapse: Causes, Consequences and Prospects*, pages 37–45.

Fajgelbaum, P. D. and Khandelwal, A. K. (2016). Measuring the Unequal Gains from Trade. *The Quarterly Journal of Economics*, 131(3):1113–1180.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Frosst, N. and Hinton, G. E. (2017). Distilling a neural network into a soft decision tree. *CoRR*, abs/1711.09784.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning.

Gauthier, T. (2020). Deep reinforcement learning for synthesizing functions in higher-order logic. In Albert, E. and Kovacs, L., editors, *LPAR23. LPAR-23: 23rd International Conference on Logic for Programming, Artificial Intelligence and Reasoning*, volume 73 of *EPiC Series in Computing*, pages 230–248. EasyChair.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., and Giannotti, F. (2018). Local rule-based explanations of black box decision systems. *CoRR*, abs/1805.10820.

Hamilton, W. L. (2020). *Graph Representation Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, Vol. 14, No. 3, Pages 1-159.*

Hanson, G. H. (2005). Market potential, increasing returns and geographic concentration. *Journal of International Economics*, 67(1):1–24.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1414–1423. PMLR.

Hausmann, R. and Hidalgo, C. A. (2011). The network structure of economic output. *Journal of Economic Growth*, 16:309–342.

Head, K. and Mayer, T. (2011). Gravity, market potential and economic development. *Journal of Economic Geography*, 11(2):281–294.

Head, K. and Mayer, T. (2014). 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*, chapter 3, pages 131–195. Elsevier.

Head, K. and Mayer, T. (2021). The united states of europe: A gravity model evaluation of the four freedoms. *Journal of Economic Perspectives*, 35(2):23–48.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Jackson, M. O. (2008). *Social and economic networks*. Princeton University Press, Princeton, NJ.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(596):583–589.

Kakade, S. M., Kearns, M., and Ortiz, L. E. (2004). Graphical economics. In *International Conference on Computational Learning Theory*, pages 17–32. Springer.

Kaliszyk, C., Urban, J., Michalewski, H., and Olšák, M. (2018). Reinforcement learning of theorem proving. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kampffmeyer, M., Chen, Y., Liang, X., Wang, H., Zhang, Y., and Xing, E. P. (2019). Rethinking knowledge graph propagation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kipf, Thomas N; Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*, 5(1):61–80.

Klenow, P. J. and Rodríguez-Clare, A. (1997). The neoclassical revival in growth economics: Has it gone too far? *NBER Macroeconomics Annual*, 12:73–103.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Leontief, W. W. (1941). *The Structure of American Economy, 1919–1929: An Empirical Application of Equilibrium Analysis*. Harvard University Press.

Lind, N. and Ramondo, N. (2018). Trade with correlation. NBER Working Paper 24380, National Bureau of Economic Research.

Long, J. B. and Plosser, C. I. (1983). Real business cycles. *Journal of Political Economy*, 91(1):39–69.

Molnar, C., Casalicchio, G., and Bischl, B. (2020). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *arXiv:2010.09337 [cs, stat]*. arXiv: 2010.09337.

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

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. (2019). Weisfeiler and leman go neural: Higher-order graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4602–4609.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2019). Deep double descent: Where bigger models and more data hurt.

Neumann, J. v. (1945). A Model of General Economic Equilibrium. *The Review of Economic Studies*, 13(1):1–9.

Newman, M. E. J. (2010). *Networks : an introduction*. Oxford University Press, Oxford; New York.

Novy, D. (2013). International trade without ces: Estimating translog gravity. *Journal of International Economics*, 89(2):271–282.

Redding, S. and Venables, A. J. (2004). Economic geography and international inequality. *Journal of International Economics*, 62(1):53–82.

Redding, S. J. and Rossi-Hansberg, E. (2017). Quantitative spatial economics. *Annual Review of Economics*, 9(1):21–58.

Schmidt, M. and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions. *CoRR*, abs/1409.4842.

Tinbergen, J. (1962). An analysis of world trade flows. In Tinbergen, J., editor, *Shaping the World Economy*, pages 262–293. New York: Twentieth Century Fund.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Verstyuk, S. (2020). Modeling multivariate time series in economics: From auto-regressions to recurrent neural networks. SSRN Working Paper 3118395, SSRN.

Wagner, A. Z. (2021). Constructions in combinatorics via neural networks.

West, D. B. (2000). *Introduction to Graph Theory*. Prentice Hall, 2 edition.

Xu, G., Duong, T. D., Li, Q., Liu, S., and Wang, X. (2020). Causality learning: A new perspective for interpretable machine learning. *ArXiv*, abs/2006.16789.

Yang, Y., Zheng, Z., and E, W. (2020). Interpretable Neural Networks for Panel Data Analysis in Economics. *arXiv:2010.05311 [cs, econ, q-fin, stat]*. arXiv: 2010.05311.

Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). GNN explainer: A tool for post-hoc explanation of graph neural networks. *CoRR*, abs/1903.03894.

You, J., Ying, Z., and Leskovec, J. (2020). Design space for graph neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17009–17021. Curran Associates, Inc.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization.

Zombori, Z., Urban, J., and Brown, C. E. (2020). Prolog technology reinforcement learning prover. In Peltier, N. and Sofronie-Stokkermans, V., editors, *Automated Reasoning*, pages 489–507, Cham. Springer International Publishing.

A Data details

A.1 Variables

Table 6: Variable definitions

Variable	Definition
contig	1 for contiguity
gdp	GDP (current US$)
gatt	1 if GATT/WTO member
comlang_off	1 for common official or primary language
comlang_ethno	1 if a language is spoken by at least 9% of the population in both countries
comcol	1 for common colonizer post 1945
col45	1 for pairs in colonial relationship post 1945
distw	weighted distance (pop-wt, km)
tdiff	nb of hours difference between ex and im
heg_o	1 if origin is current or former hegemon of destination
heg_d	1 if destination is current or former hegemon of origin
colony	1 for pair ever in colonial relationship
curcol	1 for pair currently in colonial relationship
comcur	1 for common currency
comrelig	common religion
fta_wto	1=RTA
txg_fob_usd	Goods, Value of Exports, Free on board (FOB), US$
tmg_cif_usd	Goods, Value of Imports, Cost, Insurance, Freight (CIF), US$

Notes: sources are Head and Mayer (2014) and IMF.

A.2 Sources

CEPII Gravity Dataset: gravity estimation dataset including trade, GDP, population, trade agreements (224 economies, 1948 to 2016 period, bilateral).

Source: Head and Mayer (2014), `https://sites.google.com/site/hiegravity/`, `http://www.cepii.fr/CEPII/en/bdd_modele/bdd.asp`

Directions of Trade Statistics: country and area distribution of countries' exports and imports by their partners (139 economies, 1948 to 2020 period, bilateral).

Source: International Monetary Fund, `https://www.imf.org/en/Data`

B Full estimation results

Table 7: Estimation results, all models from Structural and Fixed Effects classes

Model	L	N^{ϕ}	N^{M}, N^{S}	$N^{M}_{\phi X}$, $N^{M}_{\phi Y}$, $N^{S}_{\phi X}$, $N^{S}_{\phi Y}$	Params	R^2 train	R^2 test
Structural	1	20	-	-	382	0.53	0.51
Structural	1	50	-	-	952	0.50	0.55
Structural	1	100	-	-	1902	0.44	0.53
Structural	2	20	-	-	802	0.46	0.53
Structural	2	50	-	-	3502	0.45	0.49
Structural	2	100	-	-	12002	0.46	0.52
Structural	3	20	-	-	1222	0.49	0.54
Structural	3	50	-	-	6052	0.47	0.53
Structural	3	100	-	-	22102	0.46	0.53
Structural	4	20	-	-	1642	0.48	0.53
Structural	4	50	-	-	8602	0.43	0.46
Structural	4	100	-	-	32202	0.46	0.52
FEs (t, i, j)	-	-	-	-	431	0.54	-
FEs ($t \times i$, $t \times j$)	-	-	-	-	24617	0.62	-
FEs+BA ($t \times i$, $t \times j$)	1	20	-	-	24998	0.74	-
FEs+BA ($t \times i$, $t \times j$)	1	50	-	-	25568	0.75	-
FEs+BA ($t \times i$, $t \times j$)	1	100	-	-	26518	0.75	-
FEs+BA ($t \times i$, $t \times j$)	2	20	-	-	25418	0.75	-
FEs+BA ($t \times i$, $t \times j$)	2	50	-	-	28118	0.76	-
FEs+BA ($t \times i$, $t \times j$)	2	100	-	-	36618	0.77	-
FEs+BA ($t \times i$, $t \times j$)	3	20	-	-	25838	0.74	-
FEs+BA ($t \times i$, $t \times j$)	3	50	-	-	30668	0.76	-
FEs+BA ($t \times i$, $t \times j$)	3	100	-	-	46718	0.78	-
FEs+BA ($t \times i$, $t \times j$)	4	20	-	-	26258	0.75	-
FEs+BA ($t \times i$, $t \times j$)	4	50	-	-	33218	0.77	-
FEs+BA ($t \times i$, $t \times j$)	4	100	-	-	56818	0.78	-

Notes: Number of observations for each model is 515345 in the training sample and 54355 in the testing sample. Number of epochs for each model is 1000. "-" stands for "not applicable". See text for additional estimation details.

Table 8: Estimation results, all models from Flexible class

Model	L	N^{ϕ}	N^M, N^S	$N^M_{\phi X}$, $N^M_{\phi Y}$,$N^S_{\phi X}$, $N^S_{\phi Y}$	Params	R^2 train	R^2 test
Flex ($\times$)	1	20	10	-	484	0.23	0.14
Flex ($\times$)	1	50	20	-	1154	0.29	0.21
Flex ($\times$)	1	100	50	-	2404	0.28	0.13
Flex ($\times$)	2	20	10	-	1124	0.59	0.59
Flex ($\times$)	2	50	20	-	4544	0.34	0.36
Flex ($\times$)	2	100	50	-	17604	0.32	0.16
Flex ($\times$)	3	20	10	-	1764	0.56	0.50
Flex ($\times$)	3	50	20	-	7934	0.46	0.39
Flex ($\times$)	3	100	50	-	32804	0.06	0.04
Flex ($\times$)	4	20	10	-	2404	0.59	0.57
Flex ($\times$)	4	50	20	-	11324	0.61	0.60
Flex ($\times$)	4	100	50	-	48004	0.57	0.58
Flex (NN to $\mathbb{R}$)	1	20	10	10	564	0.54	0.57
Flex (NN to $\mathbb{R}$)	1	50	20	20	1314	0.54	0.58
Flex (NN to $\mathbb{R}$)	1	100	50	50	2804	0.55	0.45
Flex (NN to $\mathbb{R}$)	2	20	10	10	1424	0.62	0.63
Flex (NN to $\mathbb{R}$)	2	50	20	20	5544	0.33	0.37
Flex (NN to $\mathbb{R}$)	2	100	50	50	23104	0.45	0.42
Flex (NN to $\mathbb{R}$)	3	20	10	10	2284	0.64	0.64
Flex (NN to $\mathbb{R}$)	3	50	20	20	9774	0.65	0.65
Flex (NN to $\mathbb{R}$)	3	100	50	50	43404	0.64	0.64
Flex (NN to $\mathbb{R}$)	4	20	10	10	3144	0.64	0.65
Flex (NN to $\mathbb{R}$)	4	50	20	20	14004	0.61	0.62
Flex (NN to $\mathbb{R}$)	4	100	50	50	63704	0.62	0.63
Flex (NN to $\mathbb{R}_+$)	1	20	10	10	564	0.63	0.63
Flex (NN to $\mathbb{R}_+$)	1	50	20	20	1314	0.57	0.57
Flex (NN to $\mathbb{R}_+$)	1	100	50	50	2804	0.58	0.59
Flex (NN to $\mathbb{R}_+$)	2	20	10	10	1424	0.63	0.64
Flex (NN to $\mathbb{R}_+$)	2	50	20	20	5544	0.61	0.61
Flex (NN to $\mathbb{R}_+$)	2	100	50	50	23104	0.55	0.56
Flex (NN to $\mathbb{R}_+$)	3	20	10	10	2284	0.57	0.57
Flex (NN to $\mathbb{R}_+$)	3	50	20	20	9774	0.64	0.64
Flex (NN to $\mathbb{R}_+$)	3	100	50	50	43404	0.54	0.53
Flex (NN to $\mathbb{R}_+$)	4	20	10	10	3144	0.53	0.44
Flex (NN to $\mathbb{R}_+$)	4	50	20	20	14004	0.66	0.65
Flex (NN to $\mathbb{R}_+$)	4	100	50	50	63704	0.68	0.68

Notes: Number of observations for each model is 515345 in the training sample and 54355 in the testing sample. Number of epochs for each model is 1000. "-" stands for "not applicable". See text for additional estimation details.

C Symbolic Regression illustration

Table 9: Symbolic Regression results, illustration

Complexity	MSE	Score	Functional form
1	4.69454	0.00E+00	-1.90094
2	4.675339	4.10E-03	inv(-0.5007871)
4	2.436301	3.26E-01	log(Abs(x1)) - 2.6198666
6	1.227358	3.43E-01	log(Abs(x0/x2)) - 1.1418222
7	0.987196	2.18E-01	inv(-0.48061988) + log(Abs(x0/x2))
8	0.870141	1.26E-01	log(Abs(inv(inv(0.012780067*x0 + 0.012780067*x1))))
9	0.866631	4.04E-03	inv(-0.48061988) + log(Abs((x0 + 0.27159384)/x2))
10	0.359099	8.81E-01	log(1.0233755*Abs(x0/(x2 + 2.4021518))) - 1.5363793
12	0.254029	1.73E-01	log(Abs((x0 + 0.20765327)/(0.34836403*x2 + 0.69086415))) - 2.4887078
13	0.243078	4.41E-02	log(Abs((x0 + 0.27159384)/(0.39762166*x2 + 0.8343353))) - 2.4887078
14	0.239832	1.34E-02	log(Abs((x0 + 0.16599981)/(1.0676638**x2 + 0.17757954*x2))) - 2.3656318
15	0.22983	4.26E-02	log(1.0233755**re(x1)*Abs((x0 + 0.23992212)/(x2 + 2.4021518))) - 1.5363793
16	0.221448	3.72E-02	log(0.9678089*1.0335835**re(x1)*Abs((x0 + 0.3486238)/(x2 + 2.4686007))) - 1.651166
17	0.221448	9.03E-08	log(0.9702488*1.0335946**re(x1)*Abs(x0 + 0.34837505)/(Abs(x2) + 2.4677477)) - 1.6536945
19	0.192199	7.08E-02	log(Abs(Abs(x0 + 0.1795876)/(0.3598067**x1*x0 + 0.31959492*x2 + 0.60487235))) - 2.4887078
20	0.169029	1.28E-01	log(Abs((x0 + 0.16601108)/(0.35681313**x1*x0 + 1.0649966**x2 + 0.17763309*x2))) - 2.3653913

Notes: The explained variable is $\ln(S_{j,t})$, the optimization criterion is MSE loss. In the output, "Complexity" column contains total complexity of operators, constants and explanatory variables used in the function specification; "MSE" column contains Mean-Square Errors; "Score" column contains an improvement in MSE relatively to additional complexity costs, and is measured as $-\ln\left(\frac{\text{MSE}_2/\text{MSE}_1}{\text{Complexity}_2-\text{Complexity}_1}\right)$; "Functional form" column contains functions selected by the algorithm in a symbolic format.