Multidimensional Knowledge Graph Embeddings for International Trade Flow Analysis

Durgesh Nandini*[a], Simon Blöthner*[b], Mirco Schoenfeld*[c] and Mario Larch*[d]
University of Bayreuth, Bayreuth, Germany

Keywords: Knowledge Graph Embedding, Translational Embedding, KonecoTradeFlow Ontology, Multidimensional Data, International Economic Bilateral Trade Flow Data.

Abstract: Understanding the complex dynamics of high-dimensional, contingent, and strongly nonlinear economic data, often shaped by multiplicative processes, poses significant challenges for traditional regression methods as such methods offer limited capacity to capture the structural changes they feature. To address this, we propose leveraging the potential of knowledge graph embeddings for economic trade data, in particular, to predict international trade relationships. We implement KonecoKG, a knowledge graph representation of economic trade data with multidimensional relationships using SDM-RDFizer and transform the relationships into a knowledge graph embedding using AmpliGraph.

1 INTRODUCTION

Knowledge graphs (KG) are repositories for factual information in triple form and have been increasingly prevalent across various domains. Exploring knowledge graph embedding models has emerged as a novel approach for exploiting knowledge graphs. These graphs have been useful, promoting numerous downstream tasks (Kun et al., 2023; Abu-Salih, 2021). These embeddings represent nodes and, in some cases, edges as continuous vectors, providing several advantages over traditional graph structures (Cai et al., 2018; Goyal and Ferrara, 2018; Wang et al., 2017). Beyond this, graph-based methodologies offer a promising avenue for capturing and quantifying narratives, particularly through knowledge graphs (KGs) which map interactions between concepts or events relevant to the research subjects (Wang et al., 2017; Chen et al., 2020b). Numerous applications of these methods have demonstrated the efficacy of graph modelling and quantitative graph analysis in capturing complex economic relationships (Xia et al., 2021; Chen et al., 2020a).

Therefore, this study applies KG translational embedding techniques (Bordes et al., 2013) to solve inherent problems in empirical economic research. Economic research typically transforms the network of economic interactions into a format usable for (often even linear) inferential statistics or theoretical algebraic reasoning. However, this transformation can cause strong information and complexity compression, limiting the representativeness since the interaction and the underlying network structure have been almost completely ignored (Wolfram, 2002). Additionally, economic data has suffered from the problems of high-dimensionality, contingency and strong non-linearity, which originate from multiplicative dynamics (Donoho et al., 2000; Bolón-Canedo et al., 2016; Raudenbush and Bryk, 2002). This paper discusses these issues when further analysing economic data in Section 2.

To address these problems, we propose that every economic interaction can be represented within a network structure. In the latter, we establish the concept of an economic trade network as a system of interconnecting countries based on their trade relations. Our primary aim is to explore the predictive capabilities inherent within such a network, specifically focusing on forecasting flows between country pairs. To do this, we introduce KonecoKG, a downstream KG embedding model featuring multidimensional translational relationships for the international economic bilateral data. A multidimensional relationship in the context of KGs is one between entities encompass-

[a] https://orcid.org/0000-0002-9416-8554
[b] https://orcid.org/0009-0006-3462-4809
[c] https://orcid.org/0000-0002-2843-3137
[d] https://orcid.org/0000-0001-9355-2004
*All authors contributed equally.

Nandini, D., Blöthner, S., Schoenfeld, M. and Larch, M.
Multidimensional Knowledge Graph Embeddings for International Trade Flow Analysis.
DOI: 10.5220/0013028500003838
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024) - Volume 2: KEOD*, pages 63-73
ISBN: 978-989-758-716-0; ISSN: 2184-3228
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

ing multiple attributes or interactions simultaneously. Such a relationship offers various advantages because it facilitates capturing the combined effect of multiple attributes rather than one single entity-attribute relation. For example, a simple binary relationship might indicate only a single type of link, e.g., *"country A trades with country B"*, On the other hand, a multidimensional relationship captures a richer set of associations, such as fixed effects, and contextual information like trade volumes, geographical proximity and economic indicators. The latter include gross domestic product (GDP) and population size. By incorporating these diverse dimensions into the relationships, a KG can provide a more nuanced, comprehensive representation of the data, enabling more accurate and insightful analysis using the embedding model. By leveraging a trade network dataset, we anticipate future trade opportunities by integrating historical trade patterns with insights into the trading behaviours of neighbouring countries within the network. Additionally, multi-attributes such as trade agreements, geographic proximity and economic similarities act as network features to refine the accuracy of our predictions. The main contributions of this work are as follows:

- Establish a trade network as a graph representation of countries with relationships indicating trade flows, eliminating the problems of nonlinearity and non-hierarchical representations in international economic bilateral trade data.
- Introduce the KonecoTradeFlow ontology, which represents the concepts of the international economic bilateral trade data.
- Introduce KonecoKG, a downstream graph embedding model that applies translational techniques to forecast trade flows.

To the best of our knowledge, our study is one of the few pioneering efforts in utilising a large-scale economic trade network to predict trade flows between countries. The implications of accurately forecasting trade dynamics are significant, offering valuable insights for policymakers, businesses and investors to optimise international trade strategies. Additionally, the study identifies emerging market trends and encourages economic growth (Anand et al., 2021).

The rest of the paper is organised as follows: First, Section 2 gives an overview of the literature on conventional econometric approaches and graph-based methods and underlines the key challenges in economic research, establishing the significance of the current study's contribution. Having outlined the challenges, Section 3 focuses on the approach we are using and describes our process to construct the TradeFlow ontology, the embedding methods used, and the learning strategy. Section 4 focuses on the experimental setup and the evaluation metrics used. Section 5 provides insights into the results obtained and discusses their implications. Lastly, we highlight the findings of the research and conclude in Section 6, whereby we also layout ideas for future research in this field.

2 STATE OF THE ART

In this section, we discuss the challenges associated with the economic data comprising the foundation of this research. Additionally, this section reviews current methods used to address these challenges and identifies gaps in these methods, including those involving KGs, to underscore the necessity of the proposed approach.

Challenges of Economic Data. Many formal, data-driven efforts do not adequately address the unique characteristics of economic data (Schumpeter, 1933). Economic exchanges are shaped by subjectivity (Menger, 1871), creating context dependence and contingency, sometimes called localised knowledge (Hayek, 1945). Together, these characteristics hinder people from gathering reliable insights from economic data. Multi- or high-dimensionality requires incorporating many variables into models, which must be capable of untangling all the nonlinear interactions between these variables. Beyond this, many economic variables of interest exhibit strong power law behaviour, also called heavy- or fat-tailed behaviour (Gabaix, 2009; Di Giovanni et al., 2011; Axtell, 2001; Hinloopen and van Marrewijk, 2006). This process produces a slow convergence speed, leaving one in a world of pre-asymptotics with estimates which have not yet reached stable, reliable values. Even if such a value is reached, it is often unrepresentative of individual observations due to the large difference in magnitude (Taleb, 2020).

Figure 1 exemplifies this characteristic. Looking at all the bilateral trade flows grouped by year shows that the data has much heavier tails than a Gaussian distribution, also called the normal distribution. This can be seen by the mass of probability in the tails, as opposed to that in the body of the distribution. Notably, the distributions in Figure 1 are on a logarithmic scale, making the problem exponentially more pronounced. The distribution has a tail index $\alpha \approx 1$, leading to slow convergence and imprecise estimates.

All these phenomena are expressed to the highest degree when dealing with international trade flows, as they necessarily aggregate all the individual choices to the highest possible level (Blöthner and Larch, 2022).

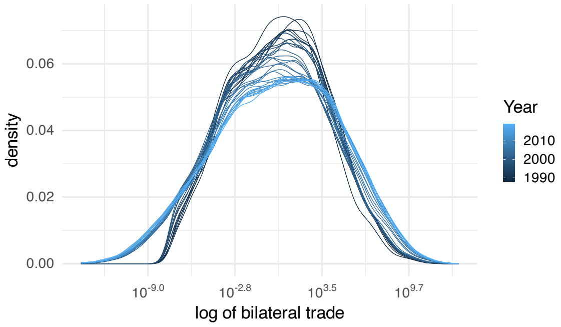


Figure 1: Log density of bilateral trade flows across time.

Methods for Economic Data Analysis. The standard empirical approach in economic data analysis, a field referred to as econometrics, is a regression model. To explain variations in bilateral international trade flows, the workhorse model is to estimate the theory-founded gravity equation using the Poisson pseudo-maximum likelihood (PPML) estimator (Santos Silva and Tenreyro, 2006; Head and Mayer, 2014; Yotov et al., 2016). Generally, these approaches rely on a large set of fixed effects to control for unobservable effects in various dimensions. This process includes dummy variables for every country, and sometimes for every country pair, as well as exporter-year and importer-year observations (Fally, 2015; Egger and Staub, 2016). We will also rely on this specification when comparing it to our KG model in section 5. Another approach is the descriptive analysis of networks such as in (De Benedictis and Tajoli, 2011; Basile et al., 2018). However, such work does not facilitate inference or the understanding of factors that drive certain characteristics within the network. Recent advances in informatics, especially the combination of machine learning models with graph structures, can provide new insights into the field of economics. However, due to their focus on causal explanation, traditional economic analysis methods have predominantly relied on linear models and supervised learning techniques.

Knowledge Graph for Economic Trade Flow Data. Relevant recent advances have been made in the field of neural networks and KG networks. (Sellami et al., 2024) used a Graph Convolution Network for predicting the trade relation between countries. Elsewhere, (Rincon-Yanez et al., 2023a) used a synthetic triple-generation algorithm for enhancing downstream tasks in KG embeddings based on the graph complement. (Rincon-Yanez et al., 2023b) leveraged KG embeddings for modelling international trade, focusing on link prediction using embeddings, and explored the integration of traditional machine learning methods with KG embeddings. (Meng, 2022) used an enterprise KG to predict China's Free Trade Zone. (Gastinger et al., 2023) used a KG to explain trade patterns among various countries. Other approaches to this process have been in the economic trade flow data analysis including economic planning (Shao et al., 2017), and industrial economic status (Quan, 2022).

3 METHODOLOGY

This section explains the creation of KonecoKG, applying embedding techniques, and predicting trade values. KonecoKG takes triples in the form of Subject (s), Predicate (p) and Object (o) as inputs for multiple relationships, and then forms embedding vectors for each relation. Next, the embedding vectors are combined as an average embedding vector to predict trade flows between countries as the final output. Figure 2 shows a diagrammatic representation of the methodology. The subsections here provide an extensive overview of the methodology followed.

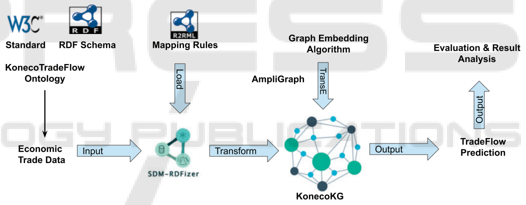


Figure 2: Trade flow prediction and analysis pipeline.

3.1 International Economic Trade Flow Data

The initial step entails identifying relevant aspects of the dataset. Using trade data from (Borchert et al., 2021), spanning 1986 to 2016, and encompassing 170 countries, we tackle the questions of economic drivers of trade flows. To determine this, we added explanatory data from (Gurevich and Herman, 2018) for GDP and population data, and (Mayer and Zignago, 2011) for information on geographic distances between countries. Lastly, we employed data about trade agreements from (Egger and Larch, 2008), a strong predictor of international trade flows. We aggregated this data into a tabular format, leaving us with over 2.5 million observations over the whole time frame.

3.2 Data Processing and Feature Selection

A detailed explanation of selected features is given in Table 1, comprising the key determinants of international trade. Economic theory predicts that larger countries, measured using population or economic size (GDP), are more able than smaller countries to trade with each other. Specifically, country size affects a country's division of labour and thus the 'roundaboutness' of production or how many intermediary capital goods for production are employed. As this number grows, countries develop greater potential to trade. In contrast, countries facing high trade costs will trade less. In contrast, countries facing high trade costs trade less. These costs can be either direct because they are far apart (distance, geographic position) or indirect due to other trade barriers which increase the transaction cost (triangulation, trust, transfer).

3.3 Data Modelling as KonecoTradeFlow Ontology

The subsequent step in the construction of the model involved creating a formal semantic representation of the dataset to serve as a structured framework for organising and categorising concepts, entities and relationships. The advantage to this method is that it captures the hierarchical structure and dependencies among these features, allowing for a nuanced understanding of their interplay in shaping trade dynamics (Chandrasekaran et al., 1999; Fensel and Fensel, 2001; Uschold and Gruninger, 1996). Figure 3 represents the hierarchical structure of our data as a class diagram. From the figure, we identify *'trade relation'* as our main class. A complete list of data properties (Uschold and Gruninger, 1996; Chandrasekaran et al., 1999) and object properties (Uschold and Gruninger, 1996; Chandrasekaran et al., 1999) is provided in Table 1.

3.4 Knowledge Graph Construction

In the next step, we use the KonecoTradeFlow ontology formulated in the above step to a structured representation in a KG, producing a set of triples.

To this end, we converted our dataset into KonecoKG using

SDM-RDFizer, an open-source tool and interpreter of the W3C Recommendations Standard R2RML[1] and its RDF Mapping Language (RML)[2] extension used for the semantification process and used in KG creation in prior research (Shahi, 2023). The RDF is a standardised data model used to describe resources on the web using subject-predicate-object statements, known as triples. Each triple comprises three components: subject, predicate, and object. The following are the detailed steps used to convert data into KonecoKG:

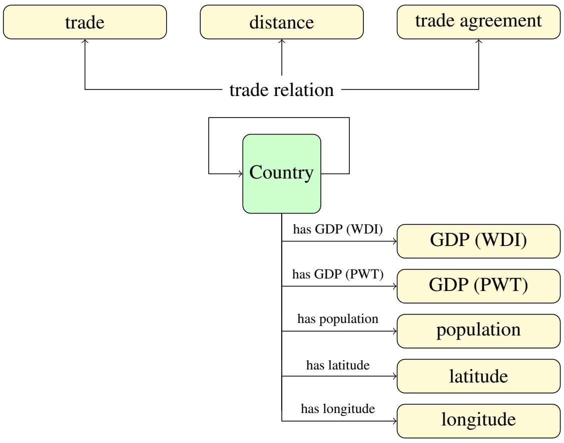


Figure 3: KonecoKG data model diagram.

- **Entity Identification:** we identified the entities or resources that we wanted to represent in RDF. For our use case, the entities were countries, specifically the exporters and the importers, their associated data properties, and their relationships among them.
- **Ontology:** We used the KonecoTradeFlow ontology as vocabulary to model trade data. For instance, the data property *trade* represents trade (in millions of US Dollars).
- **Mapping Rule:** Following the steps of SDM-RDFizer, mapping rules were created using R2RML. We assigned the base URL as *www.koneco.de*, and mapping rules assigned the data values to the corresponding subjects, predicates, and objects in the RDF triples and assign the appropriate Uniform Resource Identifier (URI) for the entities and properties and linked them to represent the relationships. For instance, the trade column of the dataset is mapped as *tradeValue*. A snapshot of the data properties from the KonecoTradeFlow ontology is given below.

rr:predicateObjectMap [

[1] https://www.w3.org/TR/r2rml/

[2] https://rml.io/specs/rml/

Table 1: Data & object properties and their description.

Data Property	**Description**
trade	volume of bilateral trade
distance	geodesic distance between the exporter and importer
trade agreement	whether a trade agreement exists between two countries
GDP (WDI)	GDP of a country as measured by the World Development Indicators
GDP (PWT)	GDP of a country as measured by the Penn World Tables
population	population of a country
latitude	geographical latitude of a country
longitude	geographical longitude of a country
Object Property	**Description**
tradesWith	indicates whether a trade relation exists between two countries

```
        rr:predicate kg:tradeValue;
        rr:objectMap [
            rml:reference "trade"
        ]
    ]
```

- **Serialising as RDF:** We serialised the RDF triples into a specific RDF serialisation format. We used the Turtle format to store and exchange RDF data while preserving the structure and semantics of the triples.

We represented *Facts* in a KG as relationships between entities — for instance, <*ARB_NZL hasTradeValue n*>, means Aruba exports, goods and services of value *n* to New Zealand. We build a series of such statements derived from the raw data collection to represent them as a graph.

3.5 Knowledge Graph Embeddings

After KonecoKG is created, we employed KG embeddings, generating embedding scores for each triple, thus encoding entities and relationships into numerical vectors. In this way, the model processes intricate patterns and semantic information as a continuous vector space, facilitating enhanced effective analysis and inference. Next, we trained the triples, derived from the KG, using three embedding models. Specifically, we employed TransE (Bordes et al., 2013), ComplEx (Trouillon et al., 2016), and DistMult (Dettmers et al., 2018).

The *TransE* is a deterministic approach which regards the relation as a translation operation from the head entity to the tail entity and utilises a distance-based scoring function to measure the plausibility of triples. Each of the latter offers unique advantages and facilitates different perspectives on capturing the semantics of the underlying data. On the other hand, the *ComplEx* and *DistMult* utilise tensor factorisation and model the interaction of entities and relations by vector-matrix product to obtain the expressive power of the data.

3.6 Prediction Model

This section explains this study's approach to finding trade relations using link prediction in KonecoKG. Link prediction is the process of exploiting the existing facts in a KG to infer missing ones. For triples <s,p,o> in KonecoKG, where <s> refers to a country pair, <p> represents countries' trade relation, and <o> represents the monetary value of the trade occurring between two countries. Then we used tail prediction to predict the values of o.

Subsequently, we adopted a corruption-based learning strategy (Bordes et al., 2013) to make predictions. This strategy entailed intentionally introducing corruptions or perturbations to the input data during the training process to enhance the model's ability to generalise and make accurate predictions. The rationale behind this approach is its ability to encourage the model to learn robust representations of the data resilient to noise and variations. Exposing the model to a diverse range of corrupted inputs during training caused it to become more adept at discerning meaningful patterns and relationships from the data, thus improving its predictive performance on unseen or noisy data.

Practically, the corruption strategy can be implemented by augmenting the training dataset with artificially corrupted samples or by introducing random perturbations to the input data during each training iteration. The degree and type of corruption introduced can be tailored based on the specific characteristics of the dataset and the desired robustness of the model. We have expanded on the use of the corruption model, adopted by us, in Section 4.

3.7 Evaluation

We evaluated the quality of the embedding model by measuring how well the model could complete facts. The prediction model predicted the tail of all the possible facts of KonecoKG.

We evaluated the embedding model using the Mean Reciprocal Rank (MRR) and Hits@N. Once the best embedding model was determined, we applied the Mean Squared Error (MSE) metric to calculate the error in the predicted values.

- **MRR** measures how well the model ranks the correct entity or relation among the candidates in the predicted list by measuring the average of the reciprocal ranks of the correct tail entities across all test triples. If the correct tail entity is ranked first, the reciprocal rank is 1; if it is ranked second, the reciprocal rank is 1/2, and so on. MRR is defined as:

$$\text{MRR} = \frac{1}{|\text{Test Triples}|} \sum_{i=1}^{|\text{Test Triples}|} \frac{1}{\text{Rank}_i}$$

- **Hits@N** measures the proportion of test triples where the correct answer appears within the top N predictions. Similar to MRR, we have a set of test triples and a ranked list of candidate tail entities for each test triple. This metric calculates the percentage of test triples for which the correct tail entity appears within the top N ranks in the predicted list. Hits@N is defined as follows:

$$\text{HITS@N} = \frac{\text{Number of Hits at Rank } \le N}{|\text{Test Triples}|}$$

- **MSE** is used to measure the error in the prediction model by computing the average squared difference between estimated trade values ($\hat{y}_i$) and actual trade values (y_i). MSE is defined as follows:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

4 EXPERIMENTAL SETUP

To begin with, we utilised Protégé[3], a widely used ontology editor and followed ontology design approach (Dutta et al., 2015b; Dutta et al., 2015a), to build and visualise the KonecoTradeFlow ontology. The importance of this initial step before any other experimental setup was to provide insights into formalising the concept for mapping the trade flow data to create KonecoKG. This initial step was crucial to provide a visual representation of the relationships between different entities, helping to clarify how various types are connected and ensuring consistent data structure. They also served to define the formal relationships between concepts and offered a shared understanding of the domain enabling reasoning over the data.

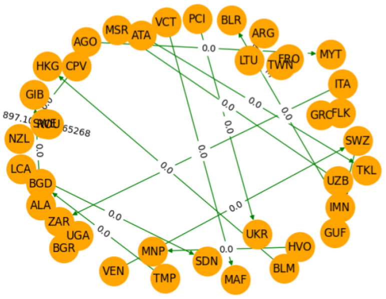


Figure 4: Sample trade network in KonecoKG.

In the second step, to simplify the start of the experimental process, we first filtered out data for each year from the entire dataset collection since the data comprises of trade information over a span of time. We did not deal with the temporal aspect of the data, rather we created a separate Kg for each year.

The third step was the conversion of trade flow data into a format suitable for KG embedding. To perform this, we employed the SDM-RDFizer. We started by formulating the required R2RML mapping rules in Turtle[4]. In the rules, we specified the classes, properties, and relationships we aim to necessitate in the graph. We used the mapping rules to generate <s,p,o> triples, also in the Turtle format. The result of all the triples (subdivisions of Classes, and Relationships) is the KonecoKG, which is also in the Turtle format.

In the fourth step, we used the generated Turtle output to parse the graph using the RDFLib[5] graph package in our model. Figure 4 shows a simplified glimpse of the trade network. In the figure, the nodes represent the countries and the edges represent a bilateral trade relationship between two countries. The labels of the edges represent the value of the monetary trade exchange in millions of US Dollar. A value of 0.0 indicates that there is no trade relation at all. The origin of the edge represents the export country, and the direction represents the import country.

In the fifth step, we employed the AmpliGraph (Costabello et al., 2019) Python library[6] to process

[3] https://protege.stanford.edu/

[4] https://www.w3.org/TR/turtle/

[5] https://rdflib.readthedocs.io/en/stable/apidocs/rdflib.html

[6] https://github.com/Accenture/AmpliGraph

the graph and to transform it into a vectorised multidimensional representation of the statements it contained. Several potential embedding model architectures were available through the AmpliGraph package with a variety of parameters. As mentioned in Section 3, to evaluate the performance of KG embedding models, we experimented with three algorithms: TransE, CompleX, and DistMult. To optimise the model parameters, we employed a grid search methodology, systematically exploring various combinations to identify the most effective settings. Table 4 presents the metric performance scores of the models obtained rounded off to the third decimal place.

Notably, our experiments revealed that the TransE model consistently outperformed the alternatives by 10%, for our data. Although ComplEx outperformed the other models for Hits@1 and Hits@10, however, upon further experiments, we found that when the N in Hits@N increased, the model's performance consistently decreased. On the other hand, with an increase in N, the Hits@N score for TransE consistently increased. Therefore, we decided to go ahead with the TransE model for further experimentation. We used the TransE to predict the trade values and evaluate the performance metrics.

In general, the model trains by comparing statements (s,p,o) known to be true against statements likely to be false based on local closed-world assumptions. This strategy measured the distances between different statements and aimed to minimise the said distance. An essential component of this experimental strategy is the corruption algorithm. The corruption algorithm creates negative triples by corrupting a true triple either by replacing the head or the tail entity with a random incorrect entity. This forces the model to distinguish between true and false facts, thereby enhancing model robustness.

Initially, we utilised the default corruption method provided by TransE. However, recognising the potential benefits of introducing controlled noise into the training process, we subsequently modified this strategy by corrupting trade values by a relative value of 20% of their true values. Through experimentation (20%, 50%, 70%, 100%, 120%), we determined that a corruption level of about 20% optimally enhanced the results. This adjustment appeared to simulate real-world variations and uncertainties in trade dynamics, thereby improving the model's ability to generalise to unseen data.

Subsequently, we trained our model for 1000 epochs, with an embedding size of 150 dimensions. These settings were chosen based on preliminary experiments and empirical observations to strike a balance between model performance and computational efficiency, ensuring timely convergence and effective learning. Tables 2 and 3 provides a full overview of the parameter values.

However, in our analysis, we noticed that we had to change the hyperparameters for a comparable prediction for the in-sample and out-of-sample methods, most notably in the epoch and batch size. The in-sample method required fewer epochs and lower negative sampling for predicting trade flows. Table 3 provides a full overview of the parameter values for in-sampling.

The trained model works by generalising relationships not yet seen by the neural network to predict the likelihood of a relationship being true with a given confidence.

Table 2: Out-of-sample embedding parameters

Parameter	Value
Epochs	1500
Embedding size	150
Corruptions	30
Batch size	30
Loss function	Pairwise
Initialiser	Xavier
Regulariser	LP, 'lambda': 0.01, 'p': 2
Optimiser	Adam
Learning rate	0.001

Table 3: In-sample embedding parameters.

Parameter	Value
Epochs	1000
Embedding size	150
Corruptions	10
Batch size	50
Loss function	Pairwise
Initialiser	Xavier
Regulariser	LP, 'lambda': 0.01,'p': 2
Optimiser	Adam
Learning rate	0.001

5 RESULTS AND DISCUSSION

To evaluate the effectiveness of our model on unseen data, we applied the leave-one-out cross-validation. We iterated over each country relation as the test set and used the rest of the chunk as the training set. Thus, we reported the average scores of the runs, each consisting of 1000 epochs. Then, we used the performance metrics MRR (Costabello et al., 2019), and Hits@N (Costabello et al., 2019) to evaluate the predictions generated by the model. As described in Section 3 we experimented with hyperparameters of three KG embedding models, namely, ComplEx, TransE

Table 4: Results of trade flow prediction.

(a) ComplEx

Metric	Score
MRR	0.483
Hits@1	**0.428**
Hits@10	**0.513**
Hits@100	0.512
Hits@1000	0.592

(b) TransE

Metric	Score
MRR	**0.587**
Hits@1	0.298
Hits@10	0.459
Hits@100	**0.576**
Hits@1000	**0.719**

(c) DistMult

Metric	Score
MRR	0.376
Hits@1	0.311
Hits@10	0.404
Hits@100	0.491
Hits@1000	0.504

and DistMult.

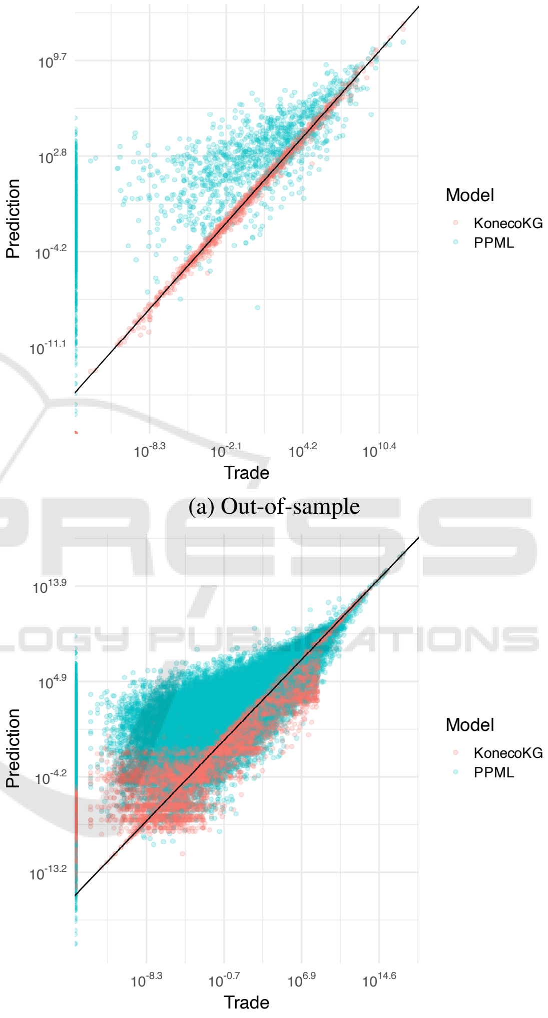
Furthermore, we also compared our results with a baseline regression model using the Mean Squared Error(MSE) metrics. We enlist the Mean Squared Error (MSE) (in millions) comparison in Table 5.

Table 5: Mean Squared Error by model.

Model	Mean Squared Error (in million)
PPML	2256.65
ComplEx	256.65
DistMult	196.26
TransE	**14.493564**

Lastly, we applied the traditional approach, for instance, PPML for predicting the trade value along with the proposed approach. This model vastly outperformed the conventional models in out-of-sample prediction tasks. Relying on the MSE, it is up to 155 times better than a comparable estimate using PPML with fixed effects, as seen from Table 5. In this vein, Figure 5a shows that the machine learning approach predicts values at every scale quite accurately. Notably the 45° line represents a perfect fit. Even in the in-sample case, KonecoKG outperforms PPML by a factor of 50. PPML is biased towards large values, which is a commonly observed result. Furthermore, our model predicts all the 0 trade flows correctly, a feature which is impossible for PPML.

Generally, conventional regression-based approaches aim to ascertain the average response of a variable of interest to a change, usually in policy. In our case, this process could involve signing a trade agreement between one or multiple countries. As motivated in Section 2 and reinforced by our results, the influence of certain factors is mediated by a plethora of contingent factors. Even if such an average response could be achieved, a complex interplay of dependencies could make the individual experience of an economic agent, such as a country, to differ wildly from the estimated average. This phenomenon has recently been addressed in economics (Peters, 2019). For this reason, we see great potential for graph-based learning algorithms to untangle the complexities at the heart of economic processes and to deepen our understanding of economic relationships.



(a) Out-of-sample

(b) In-sample

Figure 5: Comparison of predictions (on log-log scale).

6 CONCLUSION AND FUTURE WORK

In this work, we applied KG embedding techniques to predict trade flows in the international bilateral trade flow data by formulating KonecoKG, a downstream model. A significant advantage of introducing graph structure is that it alleviates the problems of non-linearity and hierarchical high-dimensional data. The proposed approach outperforms the state-of-the-art model in predicting trade values from 50, for in-sample tasks, to 155 times, for out-of-sample tasks. Currently, this approach has been applied for a limited number of properties.

Additionally, this approach can be extended by combining KGs built from other data sources which are nearly impossible to include in standard approaches, due to their unstructured nature. These sources could include text-based agreements, news, exchange and auction-based data, and market phenomena such as decentralised finance.

An alternate and immediate subsequent extension of the work would be explaining the embedding and the prediction model to identify the key determinants of the model. Additionally, post-hoc explainability models could be used to explore the results obtained.

Another possible step could be the use of the time dimension. As time is the medium through which any economic process is realised, this would offer a much more realistic picture. That is, much of recent econometric research has used this feature, facilitating the path dependence of multiplicative processes.

SUPPLEMENTAL MATERIAL

The raw data can be viewed and downloaded from *Mario Larch's* Regional Trade Agreements Database[7], Dynamic Gravity Dataset[8], International Trade and Production Database for Estimation (ITPD-E)[9]. In particular, we will release the ontology model, mapping rules for creating the KonecoTradeFlow ontology, code to tune hyperparameters for the ComplEx, TransE, and DistMult, code to train, and predict model using TransE.

The project, data, and the Python Code can also be found at the link Multidimensional Knowledge Graph Embeddings for International Trade Flow Analysis.

ACKNOWLEDGEMENT

The work has been done as the part of KONECO project, and it has received funding from the Bundesministerium für Bildung und Forschung (BMBF) under grant No 16DKWN095.

We also thank **Rebekka Koch**, our student assistant, for her valuable efforts in collecting related literature, and for efforts during various crucial parts of the paper development.

[7] https://www.ewf.uni-bayreuth.de/en/research/RTA-data/index.html

[8] https://www.usitc.gov/data/gravity/dgd.htm

[9] https://www.usitc.gov/data/gravity/itpde.htm

REFERENCES

Abu-Salih, B. (2021). Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications*, 185:103076.

Anand, J., McDermott, G., Mudambi, R., and Narula, R. (2021). Innovation in and from emerging economies: New insights and lessons for international business research.

Axtell, R. L. (2001). Zipf distribution of US firm sizes. *Science*, 293(5536):1818–1820.

Basile, R., Commendatore, P., De Benedictis, L., and Kubin, I. (2018). The impact of trade costs on the European regional trade network: An empirical and theoretical analysis. *Review of International Economics*, 26(3):578–609.

Blöthner, S. and Larch, M. (2022). Economic determinants of regional trade agreements revisited using machine learning. *Empirical Economics*, 63(4):1771–1807.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. *Progress in Artificial Intelligence*, 5:65–75.

Borchert, I., Larch, M., Shikher, S., and Yotov, Y. V. (2021). The international trade and production database for estimation (ITPD-E). *International Economics*, 166:140–166.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, 26.

Cai, H., Zheng, V. W., and Chang, K. C.-C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637.

Chandrasekaran, B., Josephson, J. R., and Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems and Their Applications*, 14(1):20–26.

Chen, X., Jia, S., and Xiang, Y. (2020a). A review: Knowledge reasoning over knowledge graph. *Expert Systems With Applications*, 141:112948.

Chen, Z., Wang, Y., Zhao, B., Cheng, J., Zhao, X., and Duan, Z. (2020b). Knowledge graph completion: A review. *IEEE Access*, 8:192435–192456.

Costabello, L., Pai, S., Le Van, C., McGrath, R., McCarthy, N., and Tabacof, P. (2019). Ampligraph: A library for representation learning on knowledge graphs. *Retrieved Oct*, 10:2019.

De Benedictis, L. and Tajoli, L. (2011). The world trade network. *The World Economy*, 34(8):1417–1454.

Dettmers, T., Minervini, P., Stenetorp, P., and Riedel, S. (2018). Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Di Giovanni, J., Levchenko, A. A., and Ranciere, R. (2011). Power laws in firm size and openness to trade: Measurement and implications. *Journal of International Economics*, 85(1):42–52.

Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1(2000):32.

Dutta, B., Chatterjee, U., and Madalli, D. P. (2015a). Yamo: yet another methodology for large-scale faceted ontology construction. *Journal of Knowledge Management*, 19(1):6–24.

Dutta, B., Nandini, D., and Shahi, G. K. (2015b). Mod: metadata for ontology description and publication. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative.

Egger, P. H. and Larch, M. (2008). Interdependent preferential trade agreement memberships: An empirical analysis. *Journal of International Economics*, 76(2):384–399.

Egger, P. H. and Staub, K. E. (2016). Glm estimation of trade gravity models with fixed effects. *Empirical Economics*, 50:137–175.

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

Fensel, D. and Fensel, D. (2001). *Ontologies*. Springer.

Gabaix, X. (2009). Power laws in economics and finance. *Annual Review of Economics*, 1(1):255–294.

Gastinger, J., Steinert, N., Gründer-Fahrer, S., and Martin, M. (2023). Dynamic representations of global crises: Creation and analysis of a temporal knowledge graph for conflicts, trade and value networks. In *D2R2*.

Goyal, P. and Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94.

Gurevich, T. and Herman, P. (2018). The dynamic gravity dataset: 1948-2016. *US International Trade Commission, Office of Economics Working Paper*.

Hayek, F. (1945). The use of knowledge in society. *American Economic Review*, 35(4).

Head, K. and Mayer, T. (2014). Gravity equations: Workhorse, toolkit, and cookbook. In *Handbook of International Economics*, volume 4, pages 131–195. Elsevier.

Hinloopen, J. and van Marrewijk, C. (2006). Comparative advantage, the rank-size rule, and Zipf's law. .

Kun, K. W., Liu, X., Racharak, T., Sun, G., Chen, J., Ma, Q., and Nguyen, L.-M. (2023). Weext: A framework of extending deterministic knowledge graph embedding models for embedding weighted knowledge graphs. *IEEE Access*.

Mayer, T. and Zignago, S. (2011). Notes on CEPII's distances measures: The geodist database. .

Meng, L. (2022). [retracted] information extraction and knowledge graph construction for enterprises in china's free trade zone. *Security and Communication Networks*, 2022(1):2962545.

Menger, C. (1871). *Grundsätze der Volkswirtschaftslehre*. Braumüller.

Peters, O. (2019). The ergodicity problem in economics. *Nature Physics*, 15(12):1216–1221.

Quan, J. (2022). Visualization and analysis model of industrial economy status and development based on knowledge graph and deep neural network. *Computational Intelligence and Neuroscience*, 2022(1):7008093.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage.

Rincon-Yanez, D., Mouakher, A., and Senatore, S. (2023a). Enhancing downstream tasks in knowledge graphs embeddings: A complement graph-based approach applied to bilateral trade. *Procedia Computer Science*, 225:3692–3700.

Rincon-Yanez, D., Ounoughi, C., Sellami, B., Kalvet, T., Tiits, M., Senatore, S., and Yahia, S. B. (2023b). Accurate prediction of international trade flows: Leveraging knowledge graphs and their embeddings. *Journal of King Saud University-Computer and Information Sciences*, 35(10):101789.

Santos Silva, J. and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, 88(4):641–658.

Schumpeter, J. (1933). The common sense of econometrics. *Econometrica*, pages 5–12.

Sellami, B., Ounoughi, C., Kalvet, T., Tiits, M., and Rincon-Yanez, D. (2024). Harnessing graph neural networks to predict international trade flows. *Big Data and Cognitive Computing*, 8(6):65.

Shahi, G. K. (2023). Fakekg: a knowledge graph of fake claims for improving automated fact-checking (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16320–16321.

Shao, L., Duan, Y., Sun, X., Zou, Q., Jing, R., and Lin, J. (2017). Bidirectional value driven design between economical planning and technical implementation based on data graph, information graph and knowledge graph. In *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 339–344. IEEE.

Taleb, N. N. (2020). Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications. *arXiv preprint arXiv:2001.10488*.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. (2016). Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080. PMLR.

Uschold, M. and Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*, 11(2):93–136.

Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Wolfram, S. (2002). *A new kind of science*, volume 5. Wolfram media Champaign, IL.

Xia, F., Sun, K., Yu, S., Aziz, A., Wan, L., Pan, S., and Liu, H. (2021). Graph learning: A survey. *IEEE Transactions on Artificial Intelligence*, 2(2):109–127.

Yotov, Y. V., Piermartini, R., Monteiro, J., and Larch, M. (2016). *An advanced guide to trade policy analysis: The structural gravity model*. available for download athttps://unctad.org/publication/advanced-guide-trade-policy-analysis-structural-gravity-model-volume-2.