Chaos, Solitons and Fractals 176 (2023) 114071

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

journal homepage: www.elsevier.com/locate/chaos

Relatedness in the era of machine learning

Andrea Tacchella [a], Andrea Zaccaria [b,a,*], Marco Miccheli [b,c,d], Luciano Pietronero [a,b]

[a] Centro Ricerche Enrico Fermi, Rome, Italy
[b] Istituto dei Sistemi Complessi - CNR, UOS Sapienza, Rome, Italy
[c] Translated Srl, Rome, Italy
[d] Dipartimento di Fisica, Università Sapienza, Rome, Italy

ARTICLE INFO

Keywords:
Economic complexity
Machine learning
Relatedness
Industry upgrading

ABSTRACT

Relatedness is a quantification of how much two human activities are similar in terms of the inputs and contexts needed for their development. Under the idea that it is easier to move between related activities than towards unrelated ones, empirical approaches to quantify relatedness are currently used as predictive tools to inform policies and development strategies in governments, international organizations, and firms. Here we show that the standard, widespread approach of estimating Relatedness through the co-location of activities (e.g. Product Space) generates a measure of relatedness that performs worse than trivial auto-correlation prediction strategies. In this paper, working on data about countries' trade, technologies, and scientific production, we show two main findings. First, we find that a shift from two-product correlations (network-density based) to many-product correlations (decision trees) can dramatically improve the quality of forecasts, allowing the possibility to assist policymakers in optimizing decisions to promote growth. Then, we propose a new methodology to empirically estimate Relatedness that we call Continuous Projection Space (CPS). CPS, which represents a general network embedding technique, vastly outperforms all the co-location, network-based approaches, while retaining similar interpretability in terms of pairwise distances. Depending on the dataset the best approach is always either CPS or machine learning algorithms based on decision trees.

1. Introduction

The concept of Relatedness [1] is a key element for both economic and social sciences, with applications ranging from smart specialization strategies [2] to the study of countries development [3,4], to recommender systems [5]. Two human activities are considered to be *related* if they share a common set of capabilities that are needed for their development [6]. The larger the intersection of the needed capabilities, the stronger the Relatedness. For example, the industrial production of radars is closely related to the production of radio broadcasting apparatus, but not so much to crude oil refining. In recent years, driven by the increasing popularity and adoption of the Economic Complexity framework [3,4,7–14], Relatedness has been gaining importance in informing diversification or specialization strategies across a wide range of policy-making institutions such as the World Bank [15] and the European Commission [16,17]. Under the idea that it is easier to develop activities that are similar to those already developed in a region, decision makers can rely on a quantitative tool to design policies that can be adapted to several strategic approaches (e.g. vertical or horizontal policies [18]. See [19] for a study about the relationship between relatedness and economic complexity metrics). So a country, or a region, that is currently competitive in the production of radars can have development opportunities in radio broadcasting apparatus, as a consistent set of the needed capabilities is likely already present. By forecasting future activities we quantify how close countries are to a given industrial or technological sector; this is not a recommendation to walk the easiest path. Instead, with this paper we address the problem of providing a reliable quantification of the feasibility of these transitions; such assessment can be a valuable input for strategic decisions.

Clearly, a poor or inconsistent quantification of Relatedness, and therefore a wrong estimation of the feasibility of transitions to new industries, represents a huge risk for policymakers basing their decisions on it. In fact, despite the potentially great impact that such ideas can have in shaping policies, an important point that needs to be addressed is the fact that there is no direct way to estimate the Relatedness of real-world activities from first principles, i.e. through the quantification of common inputs. A notable exception is the skill relatedness [20], which provides a similarity measure between sectors by using the information on cross-industry labor flows. While some theoretical work has been done on the combinatorics of very specific or synthetic

* Corresponding author at: Istituto dei Sistemi Complessi - CNR, UOS Sapienza, Rome, Italy.
E-mail address: andrea.zaccaria@cnr.it (A. Zaccaria).

https://doi.org/10.1016/j.chaos.2023.114071
Received 19 August 2023; Accepted 14 September 2023
Available online 19 October 2023

networks where the inputs layer is observed (such as letters-words or ingredient-recipes networks [4,21]), in any real scenario involving human activities (e.g. industries, technologies, scientific research, etc.) we do not have access to any 'book of recipes', not even to the 'list of possible ingredients', that would allow for a principled computation of Relatedness in terms of shared input capabilities.

For this reason, research has been focusing on how to recover an effective measure of Relatedness from location-activity data [3,4,6,22,23]. The core idea is that if two activities require similar inputs they tend to co-occur within the same locations more than randomly [6,24,25]. Therefore, suitably normalized counts of co-occurrences can be used as a proxy for Relatedness. The problem that we address with this paper is that this estimation is inherently difficult for two reasons:

- The number of activities is very often much larger than the number of locations in which to count co-occurrences. This means that the correlation structure that emerges is mostly random. E.g. in the countries-products case, one would estimate a 5000*5000 (in the typical Harmonized System 6-digits classification setting) co-occurrence matrix of products out of approximately 170 observations (countries). It is possible to reduce the number of activities by aggregation, for instance at 2 digits, for a total of about 100 economic sectors [26–28], but this typically leads to a very coarse-grained Relatedness structure that is often trivial and of no practical interest. In the literature, a compromise is usually made and scholars work at 4 digits. Also, the approach to increase the number of locations by reducing granularity (i.e. going at the subnational level), is of little use due to the fact that harmonized regional-level data is often unavailable, and the number of observations needed to produce a good estimate of a 5000*5000 correlation matrix is easily in the tens or hundreds of thousands [29].
- Very often, location-activity bipartite networks have a very strong nested structure [30]. As opposed to a block-diagonal structure, that would immediately lead to a definition of sectors-communities, in these networks the Relatedness signal is of second order with respect to the drive towards diversification that generates the nested structure.

The basic idea of counting co-occurrences to infer Relatedness has been refined and generalized in a wide variety of approaches [4,6,24,25,31–38]. All such approaches give rise to a network of Relatedness relations between couples of activities, i.e., in the language of statistical physics, a two-bodies correlation structure.

In order to give an objective assessment of the quality of these proxies for Relatedness, we test the ability of these networks to perform an out-of-sample link prediction task in three location-activity bipartite networks (countries vs. products, technologies, and scientific research, see Methods section for details). We perform the tests in a cross-validation setting that is detailed in the Methods section. The finding is that the link prediction ability of the co-occurrence methods is generally poor, in some cases only marginally better than the one produced by a random network, and sometimes even inferior to trivial prediction strategies. From this evidence one could draw two kinds of conclusions: either (i) Relatedness is an unimportant concept in predicting the development paths of countries, or (ii) the co-occurrence proxy is able to provide a poor quantification of Relatedness. Our findings are strongly in favor of (ii). While we find poor link-prediction performances from co-occurrence-based topologies, at the same time we are able to build much better Relatedness proxies through more advanced machine learning-based embedding techniques. These algorithms are based on high dimensional representations of the single activities; such representations are the output of supervised machine learning models trained to forecast which country will engage in a target activity in the near future. However, we also find that describing countries' development paths as the sum of binary Relatedness relationships is an oversimplification, and that, in some cases, much better results can be obtained considering higher-order interactions (i.e. patterns of absence/presence of many activities) through more complex but less interpretable tree-based models (Fig. 1 panels A and B). This is true, in particular, when the number of activities is large (as in the export dataset in this paper). An example is Boosted Trees [39], a supervised machine learning algorithm based on decision trees. In this case, we train one model for each target activity, and the present country basket is used as an input for the model to decide which activities are more related to the target one. These models can be learned with very effective strategies of *data augmentation* (bagging), where many models are learned on randomized subsamples of data and then averaged, and *boosting*, where the models are learned in sequence, with each new model trained to minimize the residuals of the previous ones (see Fig. 1 panel C). These strategies allow us to better cope with the relative scarcity of data, which makes it problematic to properly learn a correlation matrix with standard methods, and to learn more complex and effective models. Here we also mention early attempts to adopt machine learning approaches in economic complexity [40–43], which however either lack the systematic comparison in prediction tasks we show here, use different data, discuss only very specific test cases, or propose methodologies which are not suitable for the type and amount of data relevant here. In [44] we propose a scientifically testable prediction framework to compare various machine learning algorithms; however, that approach does not tackle the interpretability issues discussed here. In particular, in [44] we showed that tree-based algorithms outperform logistic regressions in assessing relatedness.

2. Results

2.1. Definition of the problem

We consider the temporal location-activity network defined by the bipartite adjacency matrix

$$M_{cp}^{t} = \begin{cases} 1, & \text{if } RCA_{cp}^{t} \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

where RCA_{cp}^{t} is the Revealed Comparative Advantage [45] of country c in activity p in year t (see Methods section for the precise definition for the different datasets). Roughly speaking, $M_{cp} = 1$ means that country c is competitive with respect to other countries in activity p. We want to test if co-occurrence-based methods are a good proxy for the Relatedness of countries' activities. In order to provide a quantitative and objective evaluation of how good these proxies are, we make use of the standard assumption that countries are more likely to develop new products that are related to the ones they already produce [3,4,23]. Therefore, our validation criteria are all related to the ability of any inferred Relatedness topology to predict the basket of activities of a country in year $t + \delta$ given its basket in year t. Here we discuss the case $\delta = 5$. More precisely, we implement a *leave-k-countries-out* cross-validation strategy (see Methods section), so that we learn both Relatedness topologies and predictive models from a set of countries and then we test them on a different, non-overlapping set of countries. We consider various classes of predictive models, whose exact specifications are given in the Methods section:

- Baselines: benchmark models that completely disregard or randomize the co-occurrence signal. The RCA method is based on the auto-correlation of the data.
- Bipartite Projections: inference of a Relatedness graph based on the monopartite projection of the bipartite network connecting countries with activities. Co-occurrence-based techniques such as the Product Space [3] and the Taxonomy Network [4] belong to this class.
- Description-based: Relatedness topology here is based on the textual similarity [46] between activities descriptions.

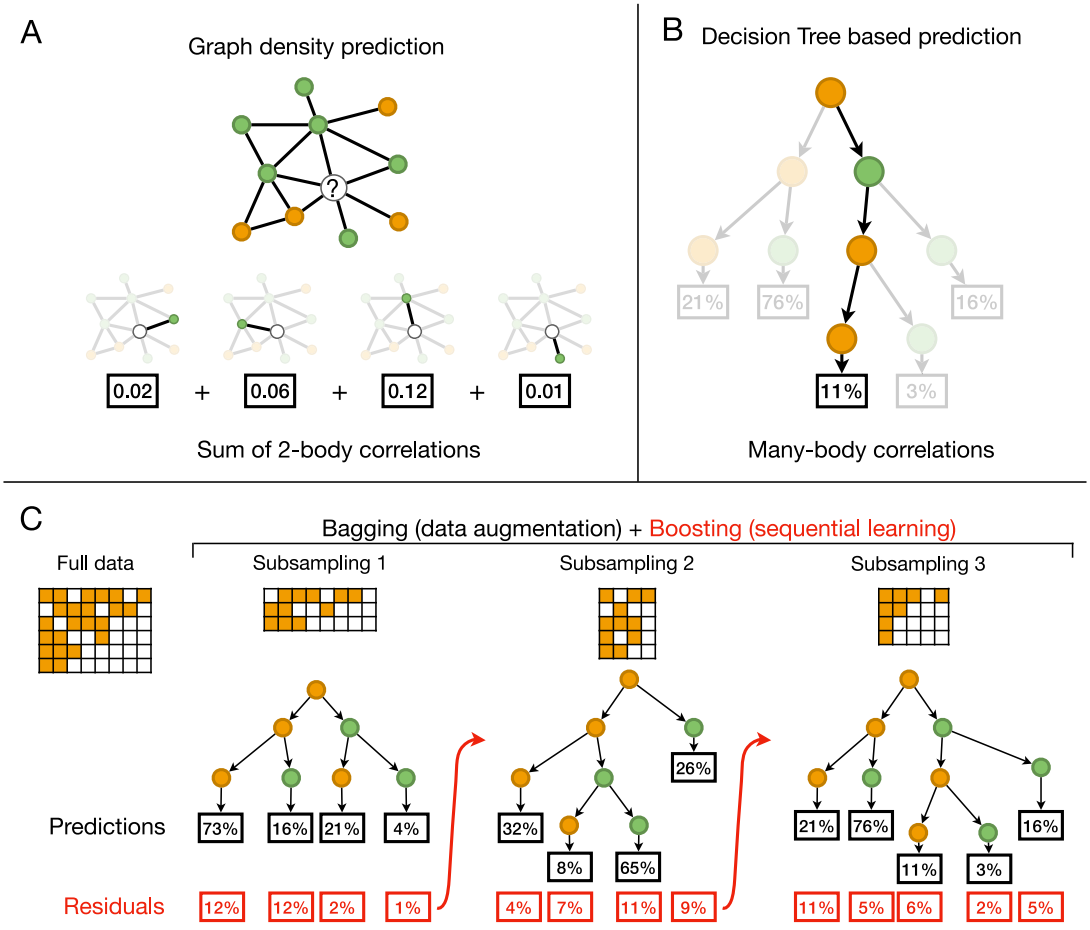

Fig. 1. **A** Density-based predictions on a graph are a linear sum of two-products relations. **B.** Decision tree-based predictions are based on many-products relations, i.e. the full path on the decision tree. A change in the value of one of the nodes has a non-linear effect on the prediction. **C.** Visual explanation of the Bagging and Boosting paradigms implemented in the XGBoost tree-based models. Bagging: The full training data is sub-sampled (for the sake of visualization here we represent one subsampling per tree, but can be done at the node level) and different *weak* models are learned on each subsample. Boosting: each model is trained to optimize the residuals of the averaged previous models.

- Tree Based: tree-based machine learning algorithms that make a link prediction based on complex patterns of presence–absence of many links (many-body correlations). In particular, here we use XGBoost [39,47] (XGB).
- CPS: low-dimensional representations of suitable embeddings obtained from the supervised machine learning algorithms introduced above. We used TSNE [48] and Variational Auto-Encoders [49] for the dimensionality reduction.
- Graph Embedding: embedding of graph topologies in Euclidean spaces. Here we used BiNE [50], an embedding technique specific for bipartite networks.

We test the predictive models on two link prediction tasks, namely (1 - All Links) to predict all the country–product links at time $t+\delta$ and (2 - Activations) to predict the links at time $t+\delta$ that had $RCA^t < 0.5$. The first task is much less interesting from an economic point of view because of the very strong auto-correlation of the country-activity structure: a model that trivially predicts that $M_{cp}^{t+\delta} = M_{cp}^t$ is typically able to achieve very high scores with every classification metric (See Supplementary Information, SI, tables 1,2, and 3). On the other hand, the second task measures the ability of the predictive models to forecast new links that are not due to small fluctuations of RCA, i.e. that are more likely to represent genuine economic development. Here we, therefore, focus on the results of the experiments on task 2. For the sake of completeness, we report in the SI also the results for task 1. We perform the experiments using 23 years of data, from 1996 to 2018. In particular, for task 2 we compute a prediction score for each element of the M_{cp}^t matrix where the corresponding RCA_{cp}^t was below 0.5 in the single year t, where t ranges from 1996 to 2013; then we validate the predictions by checking whether those c, p links are present 5 years later ($RCA_{cp}^{t+\delta} \geq 1$), i.e. in the 2001–2018 data. The same procedure is repeated for all the datasets. Note that in order to compare different prediction methodologies, we have to define a common predicting and testing framework. We choose the above definition of activation for its simplicity and interpretability. Other choices are however possible [44].

We evaluate the quality of the predictions with several standard classification metrics. In Fig. 2 we show the results for some of these metrics, chosen to be important for practical applications and to capture different aspects of the prediction task. A complete table of the results for all the metrics is available in the SI. The metrics presented in Fig. 2 are (see Methods section for their definitions):

- *BestF1*: the F1-Score computed at the optimal decision threshold. This score is computed across all the predicted links for task 2.
- *Precision@1000*: the precision of the top 1000 predicted links of the M_{cp} matrix in 2018
- *mAP@20*: the mean Average Precision of the top 20 predicted links for each country in 2018

BestF1 and Precision, therefore, are computed by considering the predicted links of the whole graph, and so we refer to them as *global metrics*, while AP@20 is computed country by country and then averaged (mAP@20) and so it is, in this sense, *local*.

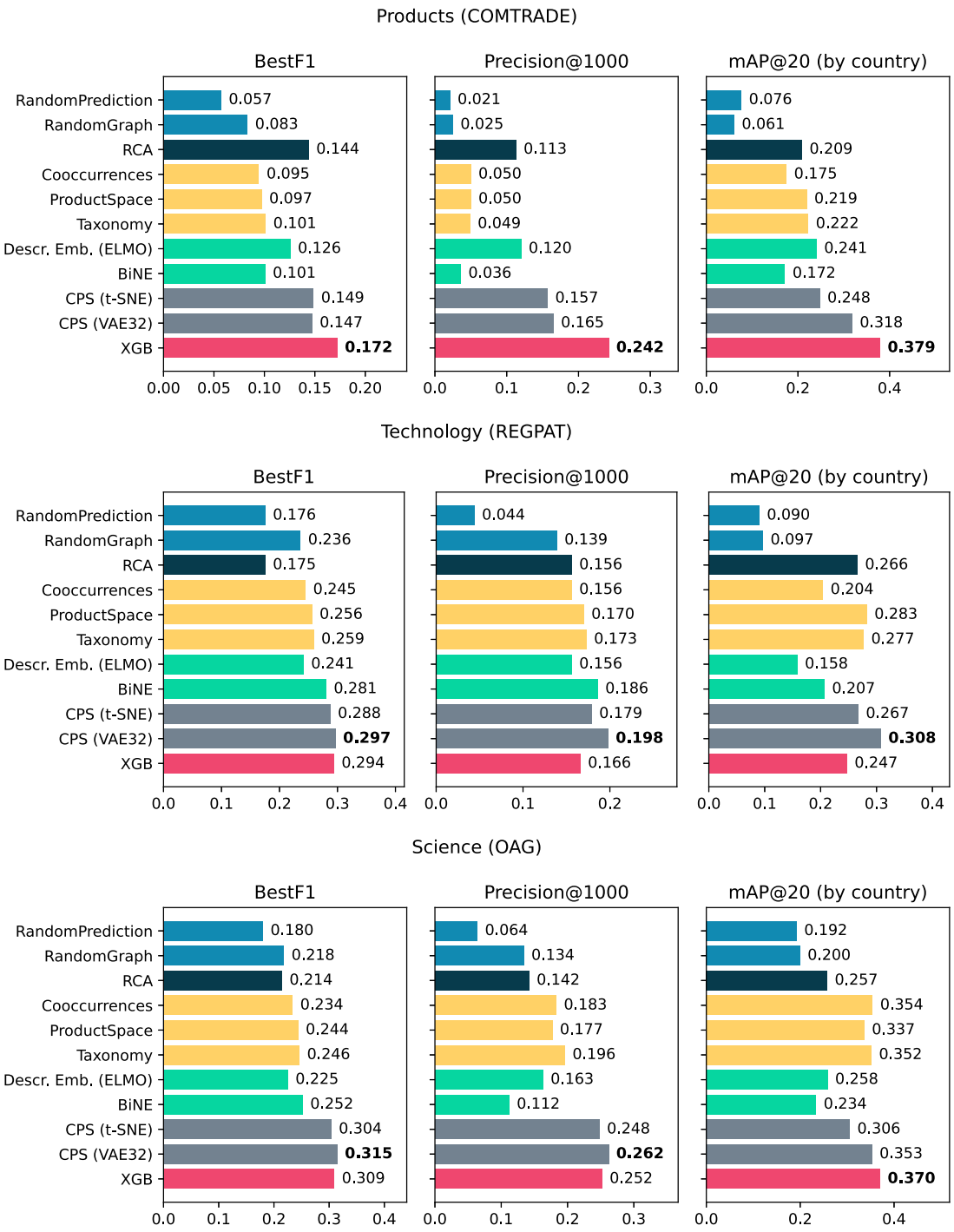

Fig. 2. Comparison of the prediction performance of different algorithm typologies using different databases and evaluation metrics. Tree-based machine learning techniques (XGBoost and/or CPS) outperform all other approaches. Description-based and CPS perform better than the RCA auto-correlation baseline, which outperforms co-occurrences for the predictions of new products.

2.2. Prediction results

In order to scientifically test the predictive power of the different approaches we applied our validation scheme to three country-level databases: UN-COMTRADE (export of products), REGPAT (patenting activity), OAG (scientific production), which are described in the Methods section and in [51]. Generally, for all datasets and all metrics, the best performing method is always either XGB or CPS-VAE32 (numbers in bold in Fig. 2).

The three datasets (see Table 1) have important differences in size, density, and activation probabilities. These differences reflect clearly in the baseline scores, and this allows us to gauge the difficulty of the prediction task in each dataset. E.g., in COMTRADE the Random Prediction scores 0.057, 0.021, and 0.076 in BestF1, Precision@1000, and mAP@20 respectively, but it scores 0.180, 0.064 and 0.192 in OAG, i.e. approximately a threefold improvement. This is easily explained by the fact that the activation probability of OAG is about 3 times higher than that of COMTRADE.

All the co-occurrence-based methods have particularly poor performances in the COMTRADE dataset. In the BestF1 and Precision@1000 metrics, they perform considerably worse than the trivial RCA prediction. A possible explanation for this is the number of activities that are

Table 1
Main features of the three datasets considered. The starting year is 1996 and the final year is 2018 for all datasets.

Database	Countries (C)	Activities (A)	C/A Ratio	Density (P(RCA>1))	Activation probability P(RCA(t+Δ)≥ 1| RCA(t)<0.5)
COMTRADE	169	5040	0.034	0.102	0.030
REGPAT	48	667	0.072	0.185	0.096
OAG	169	313	0.540	0.245	0.099

present in the COMTRADE dataset, which is one order of magnitude larger than in the other datasets. The complexity of estimating the co-occurrence matrix grows quadratically with the number of activities, and therefore a comparable growth in the number of available observations (i.e. countries) would be needed. In this scenario, the advantage of tree-based models is extremely evident (respectively +77%, +384%, and +73% in BestF1, precision@1000, and mAP@20 over Product Space), as they are better suited to learn complex relations with relatively scarce data. The CPS methods stand in between, demonstrating a noticeable improvement in performance over the co-occurrence methods, but still not reaching the performances of XGB. This implies that the Relatedness topology is a relevant concept in the COMTRADE data, but co-occurrence methods are not able to correctly infer such topology.

In REGPAT and OAG the difference between co-occurrence-based topologies and CPS is less pronounced (between 5% and 48%, comparing CPS-VAE32 and Product Space) but still clear and systematic across all datasets and metrics.

The performances of XGB are generally high, except in REGPAT. A possible explanation is the much smaller number of examples (countries) that are available in this dataset, which might be too limited for XGB to be able to learn proper patterns.

When compared to BiNE, CPS performs consistently better, especially in COMTRADE and OAG. We are confident that this is an indication that the CPS approach is a promising and very general network projection and embedding technique. Its performance with respect to standard benchmarks (for bipartite and monopartite networks) will be explored in a subsequent paper.

Finally, we have explored the effect of country-wise z-scoring on the prediction scores, as suggested e.g. in [7]. The idea is that, as clear from Eq. (3), more diversified countries tend to have generally higher prediction scores and this affects the link prediction when it is performed across different countries (this relevant for bestF1 and prec@1000 in this paper). While this bias is certainly present, we argue that it is balanced by a corresponding higher likelihood of activating links for highly diversified countries. In fact, as shown systematically in the SI, we find that z-scoring the forecasts has mixed effects on the metrics for all density-based predictors with no general trend of improvement in the COMTRADE dataset, while it is generally detrimental, with few exceptions, in OAG and REGPAT (see SI, tables 4,5,6). This effect does not affect the mAP@20 metrics, as they are computed per country and subsequently averaged.

2.3. The continuous projection space

One of the advantages of models that predict growth on the basis of pairwise distances, however, is that they allow for the visualization of development paths in low-dimensional representations of the Relatedness topology. In Fig. 3 we provide a visualization of the CPS-TSNE embeddings for the COMTRADE dataset, with the diffusion dynamics of 3 countries on that relatedness topology. For the sake of visualization, the CPS embeddings of Fig. 3 are computed without cross-validation on data ranging from 1996 to 2018, with products codified in the 1992 version of the Harmonized System. In Fig. 3 A we label the largest clusters of products that we find, to guide the interpretation of the dynamics highlighted in panels B to C. In those panels, we show the diffusion process of 3 countries: Ethiopia, which focused on clothing; China which has strongly diffused towards heavy industries with a net decrease in RCA in clothing; Vietnam has increased its RCA on some heavy industries, although much less than China, and on textiles, with much lower RCA gains in agrifood sectors, that appear to have been deprioritized in Vietnam's strategy. It is interesting to notice how CPS visualizations show at a glance a striking complementarity between the diversification strategies of China and Ethiopia, in a time frame where China has very strongly increased its influence and economic interests in Ethiopia.

3. Methods

3.1. Data

In this section, we describe the three datasets used in this work. The main statistical features of the datasets are summarized in Table 1.

3.1.1. Trade data

The UN-COMTRADE database (https://comtrade.un.org) provides the monetary volumes of the trade flows between countries, at the product level (approximately 5000 products classified according to the Harmonized System at 6 digits). Since importers' and exporters' declarations do not coincide, suitable reconstruction algorithms are needed in order to achieve a coherent and sanitized dataset. By using a global Bayesian optimization approach, we produced a denoised dataset [52] that permits, by the way, to increase the GDP prediction performance in a considerable way [53]. In this way, we obtain a set of export matrices E which correspond to as many country–product bipartite networks. We have one such export matrix for each year between 1996 and 2018. The matrix element E_{cp} represents the volume (in constant US dollars) exported by country c and relative to the product p.

Revealed Comparative Advantage (RCA). The exported volumes E are highly correlated with both the size of the country (total GDP, population, etc.) and the industrial sector. In order to obtain an *intensive* indication about the competitiveness of a country in a specific market, a suitable normalization procedure is used in the Economic Complexity literature [3,8,9]: one divides the export by both the total export of country c and the total volume of product p. Finally, a multiplicative factor given by the total exported volume assures the presence of a natural threshold equal to 1 to determine whether the given country exports that product in a competitive and relevant way. This formulation was introduced by Balassa [45] and it is called Revealed Comparative Advantage (RCA). In formula, the RCA of country c in product p is computed as:

$$\mathrm{RCA}_{cp} = \frac{E_{cp}/\sum_a E_{cp}}{\sum_c E_{cp}/\sum_{cp} E_{cp}}. \quad (1)$$

We finally define M as the matrix such as $M_{cp} = 1$ if $RCA_{cp} \geq 1$, and 0 otherwise.

3.1.2. Patent data - REGPAT

The Regpat database [54] is a publicly available resource about patents, published by the OECD on a yearly basis that covers the applications filed at the European Patent Office. Patents are localized using the residence of the inventors. Regpat classifies the technological content of patent applications according to the Cooperative Patent Classification (CPC), which has a hierarchical structure spanning all the way from very coarsely defined technological fields to very detailed ones.

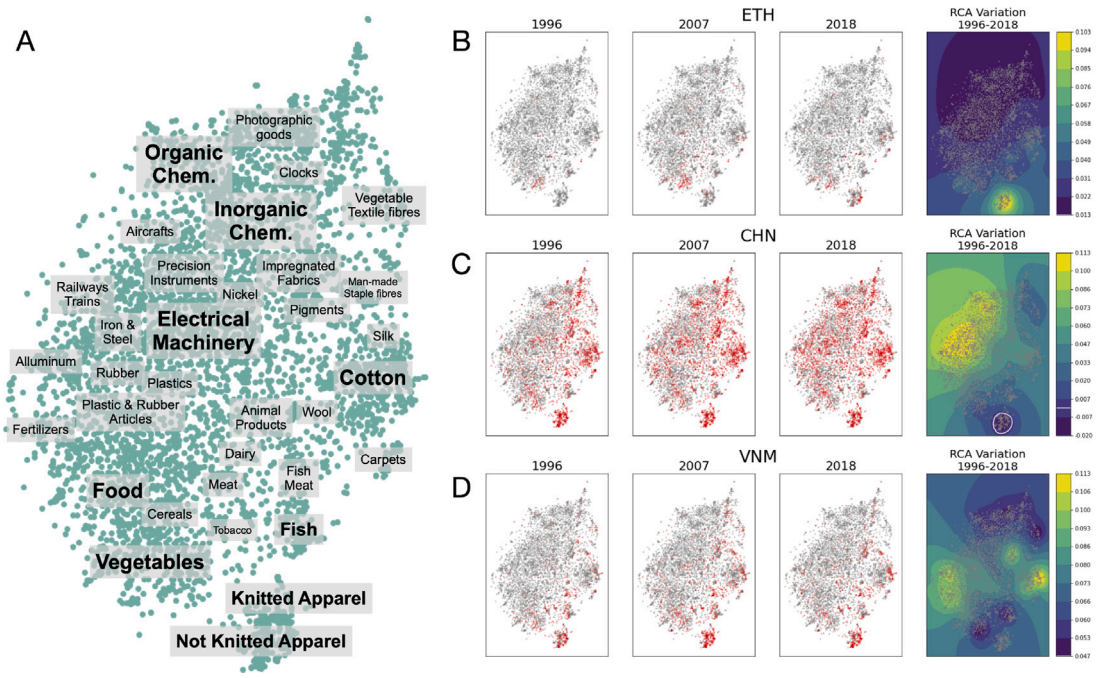

Fig. 3. A. Two-dimensional CPS/t-SNE embeddings of the Harmonized System 1992 6-digits products. Labels indicate the position of the main clusters. Panels **B**, **C** and **D** show the time evolution of the RCA of Ethiopia, China, and Vietnam projected on the CPS, with shades of red proportional to the RCA value. The rightmost panels show the average RCA variation averaged with a Gaussian kernel. Different development strategies appear very clearly through the lens of CPS. Ethiopia has strongly focused on the Apparel sector, with a considerably lower increase of RCA in all other sectors. China has focused on heavy industries and had lower RCA growth in other sectors, with a negative sign in the Apparel sector. Vietnam shows clearly delimited areas of focus in some heavy industries and textiles, with much lower growth in agrifood.

We use technology codes defined at 4-digit aggregation, corresponding to around 600 codes. The time interval covers the years 1996–2018. Here the activity of a country is quantified by counting the number of patents filed in a given year. Following [51], we then compute Balassa's RCA and apply a threshold equal to 1 to obtain a binary matrix M.

3.1.3. Science data - OAG

The database collecting the scientific performance of countries and geographical units is the Microsoft Academic Graph [55]. It is the aggregation of the scientific production collected by Microsoft, and we specifically consider the second and most updated dumps available for free usage (OAG [56]). The database is composed of a list of approximately 55 million scientific entries (article journals, books, conference proceedings, reviews, etc.) and 800 million citation counts, starting from 1800 until the end of 2018. The scientific production is categorized using 294 *Field of Study* (FoS), which are generated automatically by the similarity of each document with respect to the previous literature. Countries are assigned using the authors' affiliation. We select the 169 nations considered in the export database. In the same spirit, we select the temporal range available from the same database, therefore from 1996 to 2018. Following [51], as a proxy of the scientific competitiveness we use the logarithm of the number of citations obtained by a country in a given field. In order to have a binary matrix, we compute the Balassa index and we apply a threshold equal to 1.

3.2. Leave-k-countries-out cross-validation

In this paper, we perform link prediction in a temporal bipartite network. The most straightforward way of validating the results out-of-sample would be to learn the forecasting models using the network configurations up to a given time and then evaluate the quality of the forecast on future, unseen data. In the present case however, this approach suffers from a shortcoming: the networks object of this study are extremely auto-correlated in time, as each country tends to change only a very small portion of its product basket from year to year. For this reason, including the past of a country in the training set provides a great amount of information on how that country will look like in the future, that can be directly learned by the model; this is undesired as we want the models to be able to represent the general Relatedness patterns between products, that should not be country-dependent.

To overcome this limitation we adopt a *leave-k-countries-out* approach. We select a set of k countries and we exclude them from the training data. We learn the models from all the available data from the remaining countries. We use the models to predict all the country–product links of the countries left out starting from the year $t_0+\delta$, where $t_0 = 1996$ in our data and $\delta = 5$. We repeat the process by excluding other subsets of countries until we have a prediction for all countries. Unless otherwise stated, all the cross-validations in this paper have been performed with $k = 13$. The reason for this choice is only related to 169 (the number of countries in our COMTRADE network) being divisible by 13. We have explored other values of k and we found no significant differences as long as k is large enough, i.e. approximately greater than 5.

3.3. Models

Notation. In the following sections, we define the models $\mathcal{M}$ as functions that map the activities basket of a country c at time t, to a score S proportional to the estimated probability that the country will have $RCA > 1$ in a given activity p at time $t + \delta$. We define such models as

$$\mathcal{M}_p^{(\bar{c})}(\vec{M}_c^t) = S_{c,p}^{t+\delta} \tag{2}$$

where $\bar{c}$ indicates that the models' parameters have been tuned on training data where country c was excluded, and $\vec{M}_c^t$ is an array representing its activity basket.

Density based predictions. Following the literature [3], we define predictive models from the relatedness topologies by considering the density of activities in which a country has $RCA > 1$ around the target activity p weighted by the relatedness scores. More precisely, given a

relatedness matrix $B^{(\tilde{c})}$ computed on data that excludes country c, we define

$$\mathcal{M}_p^{(\tilde{c})}(\vec{M}_c^t) = S_{c,p}^{t+\delta} = \frac{\sum_{p'} B_{pp'}^{(\tilde{c})} M_{cp'}^t}{\sum_{p'} B_{pp'}^{(\tilde{c})}}. \quad (3)$$

The relatedness matrix can be computed using networks of co-occurrences, description embeddings, or the CPS described below.

Co-occurrence based topologies. In a seminal paper, Teece et al. [6] introduced the concept of *coherence* between neighboring activities (in their case, products) of firms. In order to measure the relatedness between two activities, they proposed to count the relative co-occurrences, that is the number of firms that are active in both. Using the language of the networks used in Economic Complexity, this corresponds to projecting the bipartite country-activity network into a monopartite network of activities, computing the activity–activity adjacency matrix as

$$B_{pp'} = \sum_c M_{cp} M_{cp'}. \quad (4)$$

This similarity measure can be normalized in different ways, to take into account trivial effects due to the degree structure of the bipartite network. A more general definition can be written as

$$B_{pp'} = \frac{1}{A} \sum_c \frac{M_{cp} M_{cp'}}{B} \quad (5)$$

In this work, we consider two specifications:

- **Product Space:** Setting $A = \max(u_p, u_{p'})$, where $u_p = \sum_c M_{cp}$ is the *ubiquity* of product p, and $B = 1$ Eq. (5) becomes the Product Space [3]. This normalization controls for the fact that more ubiquitous products have more co-occurrences.
- **Taxonomy** Setting $A = \max(u_p, u_{p'})$ and $B = d_c$ where $d_c = \sum_p M_{cp}$ is the *diversification* of country c, Eq. (5) becomes the Taxonomy network [4]. This choice of A and B controls again for the ubiquity of products, but also gives a smaller weight to co-occurrences happening in countries with high diversification, as those are more likely to be random.

The resulting networks can be filtered using suitable algorithms, such as the Minimal Spanning Tree, used in [3] for visualization purposes, or null models [6,25,34,38], able to filter spurious effects and obtain statistically validated projections; these approaches are however beyond the scope of this paper. When, like in the present case, more than one year of data is available, the relatedness matrix B can be computed by stacking vertically the M_{cp} matrices.

Description embeddings. A relatedness signal can be extracted from the textual description of the activities as defined in the respective standard classifications (see Data section). We use this signal as a control for other relatedness measures based on actual data from the country-activity networks. We use the Elmo technique [57] to obtain a similarity score between couples of textual descriptions of activities. This defines a $B_{pp'}$ relatedness matrix, in analogy with what is done with the co-occurrence-based topologies. The forecasting model is built following Eq. (3). More details are provided in the SI.

Boosted trees. All the predictive models that we consider in this paper are based on some form of Relatedness topology and ultimately make use of Eq. (3) to build their link prediction scores, with all the difference being in how the Relatedness matrix $B_{pp'}$ is built. All these methods, therefore, are based on an independent sum of 2-activities relations (Fig. 1A). With decision tree-based methods we can learn more general relations between patterns of presence/absence of sets of activities in the basket of a country, and this can radically improve the link prediction quality (Fig. 1B). To learn these complex relationships, however, we are still subject to the same scarcity of data that makes it difficult to learn the co-occurrence-based Relatedness matrices. To minimize the risk of learning spurious correlations and to improve the out-of-sample prediction quality, modern decision-tree learning algorithms make use of two ideas: *bagging* and *boosting* (Fig. 1C). The term *bagging* refers to a data-augmentation strategy where the original training data is randomly sub-sampled many times. In each sub-sample, some rows (examples, here countries) and columns (features, here activities), randomly chosen, are removed from the training data, and a *weak* model is trained on the remaining data. When the weak models are trees, as in the present case, the column sub-sampling is usually performed randomly every time a new split in the tree is learned. The resulting trained models are defined *weak* as they are learned on incomplete data. However, repeating this operation many times and finally building a meta-model that aggregates the predictions of several weak models is shown to significantly reduce overfitting [47,58], which would be a significant problem in learning complex models with scarce data.

The term *boosting* refers to algorithms that learn weak models in sequence, with each new model trained by giving focus to the training examples that generated the largest losses in the previously learned models.

The bagging and boosting paradigms can be applied in a variety of ways. Here we make use of the XGBoost [39,47] framework to train boosted decision forests. More specifically, we train one boosted forest for each activity to perform a binary classification task, i.e. we build the input-label pairs as:

$$(input, label) = (\overrightarrow{RCA}_c^t, M_{cp}^{t+\delta}) \quad (6)$$

that is, the model learns to associate a given activity basket to the possible presence of a given activity after δ years.

The parameters are the default provided by the XGBoost library (version 1.2.0) except for the *n_estimators* parameter that has been set to 30. To stabilize the results, we repeat the leave-k-countries-out cross-validation 3 times, on 3 different randomizations of the hold-out sets, and average the scores across the 3 runs. In the COMTRADE dataset, the computational cost of the leave-k-countries-out cross-validation exercise is considerable: in total we train $N_p * N_c/k * 3$ models, i.e. with $N_p = 5053$, $N_c = 169$ and $k = 13$ we train 197067 XGB models. The boosted trees models perform generally better than the co-occurrence-based models considered in this paper, the main reason being their increased functional complexity. While all the other non-trivial models that we consider are based on binary activity–activity relationships that are independently summed or averaged together, the boosted trees models explicitly consider higher-order relationships between groups of products. This increased complexity translates into a better capability of the models to represent complex patterns and ultimately to produce better forecasts, but comes at the price of much lower interpretability. The only exception comes from the CPS models that manage, in some cases, to outperform all other models including boosted trees, but still rely on a Relatedness-based topology.

Continuous projection space. We introduce the Continuous Projection Space (CPS) approach in order to recover interpretability from the tree models while maintaining a compelling capability to measure relatedness. The CPS procedure allows to translate the trained tree-based models into an activity–activity relatedness matrix that retains at the same time the interpretability of co-occurrence-based relatedness models and a vastly improved forecasting ability. The CPS procedure can be seen as a general self-supervised graph-embedding algorithm, along the lines of several others that gained popularity in the recent literature such as Node2Vec, BiNE, and others [59]. The idea behind CPS is to represent each node of a graph through a link-prediction model, i.e. a model that is trained to predict the set of other nodes to which that node is linked and, for weighted graphs, the weight of the link. Here we will use Random Forests as link prediction models. Once trained, such a link prediction model can be *sampled*, by using it to predict links in the graph. By using the prediction scores associated with each link we can generate a vector of numbers for each node, i.e. the predicted relation

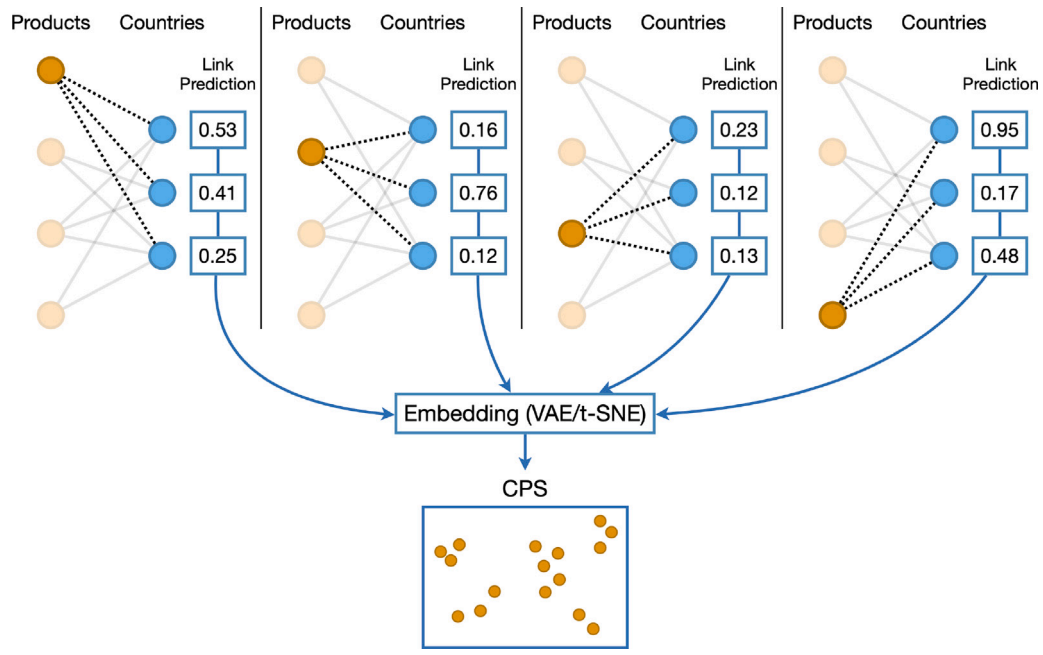

Fig. 4. Schematic explanation of the CPS methodology. First (top panel) each product is associated with a vector of predictions of the probability to be linked to each country. This provides a high-dimensional embedding. Then the dimension of these representations are reduced using standard techniques (in this case Variational Autoencoders and t-SNE).

towards each other node in the graph (or a subset of them). In practice, each activity can be represented as an embedding vector by using the associated prediction scores. This vector embeds information about the model, in terms of its outputs in relation to a fixed set of inputs (i.e. the features used for the link prediction). Intuitively, two nodes that share similar connectivity patterns will be associated with similar models that will, in turn, generate similar vectors of predictions. These vectors can be the CPS embedding per se or, as in the present case, can be further embedded into lower dimensional spaces with general dimensionality reduction techniques to improve interpretability. This technique is very general and its effectiveness depends greatly on the choice of the link-prediction models, their training procedure, and the sampling strategy. A general overview and comparison with other approaches in the literature is beyond the scope of this paper and will be discussed in upcoming work. Here we provide the specifications of the CPS implementation used in this paper. A schematic explanation of the procedure is provided in Fig. 4

To compute the CPS embeddings in the present bipartite dynamical case, we train Random Forests as the link prediction models to predict 5-year delayed links from each activity to the set of countries. That is, we use exactly the same setting that we use to train the boosted trees models, with the same leave-k-countries-out cross-validation scheme, the only difference being that we train plain Random Forests instead of Boosted Forests to reduce computational time. After the out-of-sample inference, we obtain predictions for all countries, all years, and all activities, i.e. a tensor S of the same shape of M^{y}_{cp}, where y indexes the years. Then, for each activity, we consider the vector of all the predictions for all the countries and all the years, i.e. a vector of $N_c * N_y$ entries that represent the (ordered) prediction scores. Each vector is a high-dimensional representation of the predicted activity. Intuitively, if our model predicts that the same countries will (and will not) engage in two activities, these will have similar vectors and so they will be related. We then reduce the dimensionality of such vectors in two steps: first, we train a 16-dimensional Variational Autoencoder (VAE32) [49], reducing the vectors to 32 dimensions (i.e. the 16+16 parameters of the VAE), then we perform a further dimensionality reduction from 32 to 2 dimensions with the t-SNE algorithm [48] (se SI for more details). The result is shown in Fig. 3.

To perform the predictions we first compute the $N_p \times N_p$ matrix $D_{pp'}$ of Euclidean distances between the embedding vectors. Then we transform D to a matrix B of Gaussian weights

$$B_{pp'} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(0.5\left(\frac{D_{pp'}}{\sigma}\right)^2\right) \tag{7}$$

Finally, we plug such matrix as the $B_{pp'}$ matrix in Eq. (3) to perform the forecasting. This is equivalent to a Nadaraya–Watson kernel regression with a Gaussian kernel. The resulting forecasting scores represent our CPS-based measure of relatedness, to be used to assess the feasibility of a target activity for a specific country.

It is to be noted that this procedure by itself would not imply a fully out-of-sample prediction that can be directly compared to the results presented for the other methods. This is due to the fact that the leave-k-countries-out cross-validation guarantees that the forecasts for each country are done without using any knowledge from that country, but since the CPS embeds each activity as a combination of the forecast for all countries then the resulting embeddings actually make use of all the data. For this reason, only for the CPS results, we implement one further step: we compute a set of embeddings and the resulting relatedness matrix $B^{\tilde{c}}_{pp'}$, for each country, by completely eliminating that country from the data, and then using Eq. (3) to produce forecasts for that country only. The σ parameter of Eq. (7) is chosen as the one that maximizes the in-sample Best F1 score in forecasting the links of the countries used to compute $B_{pp'}$. This procedure is repeated once for each country in the dataset, i.e. 169 times. For this reason, we are not presenting cross-validated CPS results based on Boosted Trees models, but only on the much faster Random Forest.

CPS represents a significant step forward in the economic complexity literature since it allows a better assessment of the relatedness of activities while preserving the possibility to visualize the reason behind such an assessment.

RCA baseline. As a baseline forecast, we consider the trivial model where

$$S^{t+\delta}_{c,p} = RCA^{t}_{c,p} \tag{8}$$

Given the strong auto-correlation of the countries-activities networks, this trivial model provides relatively good predictions. Strikingly, these

results outperform the ones obtained from the network of co-occurrences in the COMTRADE network.

Random graph baseline. The forecasting in Eq. (3) depends on two terms: one is the Relatedness matrix $B_{pp'}$, and the other is the M_{cp} matrix. In the Random Graph baseline model, we define $B_{pp'}$ as a random matrix i.e. a fully connected graph with random weights with a homogeneous distribution in $(0, 1)$ and with the weight of all self-loops set to 1. In this way, we completely destroy the relatedness signal, but still, observe a forecasting power that is better than a completely random forecast and, for some tasks, not far off the co-occurrence-based Product Space and Taxonomy matrices. This is due to the known [60] stylized facts of the countries-activities networks, and in particular to its nested structure: more diversified countries are more likely to become even more diversified than non-diversified countries. In Eq. (3), even when $B_{pp'}$ is random, more diversified countries get generally higher scores than non-diversified ones, even though on random activities. This bias is enough to produce forecasts that are better than random and, for some tasks, comparable to co-occurrence-based Relatedness topologies.

3.4. Evaluation metrics

In order to compare the different prediction methodologies we make use of a series of performance indicators, usually adopted for classification tasks. It is important to point out that our results (see for instance Fig. 1) are highly consistent across different indicators, even if we choose them for covering different aspects of the prediction exercise. Let us now discuss in detail how the prediction performances can be quantitatively evaluated. The specific instance to predict can be *positive* or *negative*, if the corresponding element of the activity matrix M_{cp} is equal to 1 or 0, respectively. A *true positive* is a correctly predicted positive instance. Let us focus on the top k scores of a given algorithm, that is, the k (c, p) couples that are predicted to have the higher likelihood $S_{cp} = P(M_{cp}^{(t+\delta t)} = 1)$. The indicator *prec@k* is defined as the fraction of these k elements for which $M_{cp}^{(t+\delta t)} = 1$. This is a measure of the global precision of the algorithm. In the paper we consider $k = 1000$, in the SI we report results also for $k = 10000$. However, this measure takes into account all matrix elements together, while we might be interested in evaluating the prediction performance on a country basis: on average, how much are we precise when predicting specific activities within a country? To do so we first evaluate the precision country by country, and then we average. This is called mean precision. Moreover, it is also important to weigh our success or failure using the scores rank: we want the highest scores to predict better than the lowest scores. So we have to compute a weighted average. In practice, we use the mean Average Precision *mAP@n*. Let us focus on a single country c first. The Average Precision *AP@n(c)* is defined as

$$AP@n(c) = \frac{\sum_{k=1}^{n} prec@k \times rel(k)}{P} \tag{9}$$

where *prec@k* is the precision at k; rel(k) is equal to 1 if the product at rank k is positive and zero otherwise; P is the total number of positives; and everything above is referred to country c. Then, the *mAP@n* is simply the country average:

$$mAP@n = \frac{\sum_{c=1}^{C} AP@n(c)}{C} \tag{10}$$

where we have chosen n = 20.

Precision-related measures deal with the minimization of false positives FP. In order to take into account also the problem of false negatives FN, *recall* is usually considered. In general, precision is defined as the ratio between the number of true positives TP and the number of predicted positives, while recall is the ratio between true positives TP and true instances, the real positives P. Since we want a global and balanced measure, we average the two with a harmonic mean, called F1-score. The harmonic mean aggravates the impact of possible small values of one of the two indicators. Since binary classifiers usually provide a continuous set of scores, one has to specify a threshold t above which the score is associated with a positive prediction; as a consequence, precision, recall, and so the F1 score will depend on t. We point out that in the computation of *prec@1000* the threshold choice is derived from the arbitrary choice $k = 1000$. To have a nontrivial and nonarbitrary threshold we decided to take the one that maximizes the in-sample F1 score, as suggested by [61]. Summarizing the above considerations, the best F1 score shown in Fig. 1 is defined as

$$\text{BestF1score} = \max_t \frac{2}{\text{prec(t)}^{-1} + \text{rec(t)}^{-1}} \tag{11}$$

where

$$\text{prec(t)} = \frac{TP(t)}{TP(t) + FP(t)} \qquad \text{rec(t)} = \frac{TP(t)}{TP(t) + FN(t)} = \frac{TP(t)}{P}. \tag{12}$$

These prediction performance measures can be computed for any test set. In order to show the replicability and the extent of our results, we show in the SI that they do not change if the test set and the indicators are reasonably changed. In Fig. 1 we compute the *prec@1000* and the *mAP@20* for the last year of our dataset (2018): this bears the interpretation of our result as a recommender system, in which products are suggested as feasible to countries, and countries actually start exporting them. Instead, the best F1-score is computed on all the available years in cross-validation, to show that our results are stable and comparable across different periods. Finally, in the SI we report the results also for the Area Under the Receiver Operating Characteristic curve (AUC ROC) computed in the same setting as the Best F1.

4. Discussion

The concept of Relatedness can be an extremely powerful tool to understand development dynamics and to inform policy decisions. By quantifying the proximity between activities in terms of knowledge, inputs, and infrastructures, it can help design paths to diversification or specialization strategies, based on empirical evidence. However, while the idea is extremely appealing, we show that current methods to estimate Relatedness from data perform poorly in predicting the actual trajectories of countries, even when compared to trivial alternative approaches such as using RCA as a prediction score (note that using RCA in a multivariate regression setting would lead to an aggregate forecast which prevents any comparison). If such methods cannot forecast future activities, then their assessment of relatedness has insufficient usefulness for policymakers and new methods should be introduced.

In order to overcome these limitations we have introduced a novel network embedding technique called Continuous Projection Space (CPS) for the computation of Relatedness from data. CPS performs up to 230% better than current approaches (comparing CPS-VAE32 precision@1000 against ProductSpace's in COMTRADE) while retaining the same overall properties and interpretability. Moreover, we have shown that moving from one-to-one product Relatedness to many-to-one relationships using suitable Machine Learning algorithms allows to achieve performances up to 384% better than current approaches, despite losing some of their interpretability. Globally, we can state that machine learning approaches provide a better assessment of Relatedness, as measured by their ability to forecast future activities. XGB or CPS can represent the best approach for specific databases but, in any case, they are shown here to outperform the mainstream co-location metrics. This improvement allows for a safe passage to a 6-digit relatedness, while in literature the signal-to-noise ratio forces one to work with 4 digits. We repeat our analysis also at 4 digits obtaining similar results, that are reported in the Supplementary Information.

We believe that these results can have a huge impact on the applicability of these ideas in policy-making as the quality and confidence of

the recommendations are dramatically improved. The magnitude of the improvement, especially in many-to-one models as in the COMTRADE dataset, is enough to concretely move these methods from a research idea to an applied tool to inform policy decisions. The adoption has already started in large institutions such as the World Bank [15] and the European Commission [17]. We point out that our results do not imply that the more related activity is the one a country should enter into. This would represent an easy choice, but not necessarily the best one. However, quantifying the feasibility of a possible strategy is key in a policy perspective.

CPS, on top of its ability to perform in some cases even better than Tree-Based models, has the advantage of providing a fully explainable prediction in terms of pairwise Euclidean distances, which can be of practical interest for policymakers. Besides allowing to visualize and explore the Relatedness space, this fact can have the added benefit of generalization. It would be technically possible to use the same Relatedness metrics learned at the national level to inform development strategies in regions, cities, firms, and all entities for which obtaining a consistent worldwide dataset would be extremely harder. This is due to the linear, additive form of Eq. (3). In contrast, this generalization is much less likely to work in tree-based methods. Using the terminology of modern data science, CPS is an effective *representation-learning* tool, that allows to generalize its applications to tasks different than those it was trained for.

The work presented here opens to various further research ideas. First, the CPS approach is a general Network Embedding technique that can in principle be applied to monopartite networks as well as bipartite or multipartite networks, as in the present case. We plan to systematically explore the capabilities of the CPS technique and compare it with the existing literature (in the present case, a comparison with the state-of-the-art method for embedding bipartite networks, BiNE, is reported in Fig. 2; its performance is generally much lower than CPS). Second, the tree-based prediction is suited to be generalized to multi-partite networks as well (such as, e.g., the Countries–Technologies–Products–Research network used in [25]), and we plan to explore if mixing information from multiple layers can indeed improve the quality of the predictions. Third, a CPS embedding is in theory feasible also for such multipartite networks, allowing to embed nodes from different layers in a common space. Finally, we could use graph neural networks [62,63] to perform our forecast: an essential step of this technique is the definition of a suitable embedding measure and this embedding could be provided by the CPS itself.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge the CREF project “Complessità in Economia” and the PRIN project “WECARE”, number 20223W2JKJ.

Appendix A. Supplementary information

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.chaos.2023.114071.

References

[1] Hidalgo CA, Balland PA, Boschma R, Delgado M, Feldman M, Frenken K, et al. The principle of relatedness. In: International conference on complex systems. Springer; 2018, p. 451–7.
[2] Balland PA, Boschma R, Crespo J, Rigby DL. Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification. Reg Stud 2019;53(9):1252–68.
[3] Hidalgo CA, Klinger B, Barabási AL, Hausmann R. The product space conditions the development of nations. Science 2007;317(5837):482–7.
[4] Zaccaria A, Cristelli M, Tacchella A, Pietronero L. How the taxonomy of products drives the economic development of countries. PLoS One 2014;9(12):e113770.
[5] Lü L, Medo M, Yeung CH, Zhang YC, Zhang ZK, Zhou T. Recommender systems. Phys Rep 2012;519(1):1–49.
[6] Teece DJ, Rumelt R, Dosi G, Winter S. Understanding corporate coherence: Theory and evidence. J Econ Behav Organ 1994;23(1):1–30.
[7] Hidalgo CA. Economic complexity theory and applications. Nat Rev Phys 2021;1–22.
[8] Hidalgo CA, Hausmann R. The building blocks of economic complexity. Proc Natl Acad Sci 2009;106(26):10570–5.
[9] Tacchella A, Cristelli M, Caldarelli G, Gabrielli A, Pietronero L. A new metrics for countries’ fitness and products’ complexity. Sci Rep 2012;2:723.
[10] Tacchella A, Cristelli M, Caldarelli G, Gabrielli A, Pietronero L. Economic complexity: conceptual grounding of a new metrics for global competitiveness. J Econom Dynam Control 2013;37(8):1683–91.
[11] Cristelli M, Gabrielli A, Tacchella A, Caldarelli G, Pietronero L. Measuring the intangibles: A metrics for the economic complexity of countries and products. PLoS One 2013;8(8):e70726.
[12] Sbardella A, Pugliese E, Zaccaria A, Scaramozzino P. The role of complex analysis in modelling economic growth. Entropy 2018;20(11):883.
[13] Bardoscia M, Barucca P, Battiston S, Caccioli F, Cimini G, Garlaschelli D, et al. The physics of financial networks. Nat Rev Phys 2021;1–18.
[14] Smolyak A, Levy O, Shekhtman L, Havlin S. Interdependent networks in economics and finance—A physics approach. Physica A 2018;512:612–9.
[15] Lin J, Cader M, Pietronero L. What african industrial development can learn from east Asian successes. EMCompass 2020.
[16] Pugliese E, Tübke A. Economic complexity to address current challenges in innovation systems: A novel empirical strategy linked to the territorial dimension. Ind R&I –JRC Policy Insights 2019.
[17] Pugliese E, Tacchella A. Economic complexity for competitiveness and innovation: a novel bottom-up strategy linking global and regional capacities. Ind R&I –JRC Policy Insights 2020.
[18] van Dam A, Frenken K. Vertical vs. Horizontal policy in a capabilities model of economic development. 2020, arXiv preprint arXiv:2006.04624.
[19] McNerney J, Li Y, Gomez-Lievano A, Neffke F. Bridging the short-term and long-term dynamics of economic structural change. 2021, arXiv preprint arXiv:2110.09673.
[20] Neffke F, Henning M. Skill relatedness and firm diversification. Strateg Manag J 2013;34(3):297–316.
[21] Tacchella A, Di Clemente R, Gabrielli A, Pietronero L. The build-up of diversity in complex ecosystems. 2016, arXiv preprint arXiv:1609.03617.
[22] Neffke F, Henning M, Boschma R. How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. Econ Geogr 2011;87(3):237–65.
[23] Alshamsi A, Pinheiro FL, Hidalgo CA. Optimal diversification strategies in the networks of related products and of related research areas. Nat Commun 2018;9(1):1–7.
[24] Saracco F, Straka MJ, Di Clemente R, Gabrielli A, Caldarelli G, Squartini T. Inferring monopartite projections of bipartite networks: an entropy-based approach. New J Phys 2017;19(5):053022.
[25] Pugliese E, Cimini G, Patelli A, Zaccaria A, Pietronero L, Gabrielli A. Unfolding the innovation system for the development of countries: coevolution of science, technology and production. Sci Rep 2019;9(1):1–12.
[26] Zaccaria A, Mishra S, Cader MZ, Pietronero L. Integrating services in the economic fitness approach. World Bank policy res working paper 8485, 2018.
[27] Stojkoski V, Utkovski Z, Kocarev L. The impact of services on economic complexity: Service sophistication as route for economic growth. PLoS One 2016;11(8).
[28] Patelli A, Pietronero L, Zaccaria A. Integrated database for economic complexity. Sci Data 2022;9(1):1–13.
[29] Bun J, Bouchaud JP, Potters M. Cleaning large correlation matrices: tools from random matrix theory. Phys Rep 2017;666:1–109.
[30] Mariani MS, Ren ZM, Bascompte J, Tessone CJ. Nestedness in complex networks: observation, emergence, and implications. Phys Rep 2019;813:1–90.
[31] Nesta L, Saviotti PP. Coherence of the knowledge base and the firm’s innovative performance: evidence from the US pharmaceutical industry. J Ind Econ 2005;53(1):123–42.
[32] Bottazzi G, Pirino D. Measuring industry relatedness and corporate coherence. 2010, Available at SSRN 1831479.

[33] Li MX, Palchykov V, Jiang ZQ, Kaski K, Kertész J, Miccichè S, et al. Statistically validated mobile communication networks: the evolution of motifs in European and Chinese data. New J Phys 2014;16(8):083038.
[34] Cimini G, Squartini T, Saracco F, Garlaschelli D, Gabrielli A, Caldarelli G. The statistical physics of real-world networks. Nat Rev Phys 2019;1(1):58–71.
[35] Tripodi G, Chiaromonte F, Lillo F. Knowledge and social relatedness shape research portfolio diversification. Sci Rep (Nature Publisher Group) 2020;10(1).
[36] Zhou T, Kuscsik Z, Liu JG, Medo M, Wakeling JR, Zhang YC. Solving the apparent diversity-accuracy dilemma of recommender systems. Proc Natl Acad Sci 2010;107(10):4511–5.
[37] Zhou T, Ren J, Medo M, Zhang YC. Bipartite network projection and personal recommendation. Phys Rev E 2007;76(4):046115.
[38] Cimini G, Carra A, Didomenicantonio L, Zaccaria A. Meta-validation of bipartite network projections. Commun Comput 2022;5(1):1–12.
[39] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd international conference on knowledge discovery and data mining. 2016, p. 785–94.
[40] Che NX. Intelligent export diversification: An export recommendation system with machine learning. Tech. Rep., International Monetary Fund; 2020.
[41] Tacchella A, Napoletano A, Pietronero L. The language of innovation. PLoS One 2020;15(4):e0230107.
[42] Palmucci A, Liao H, Napoletano A, Zaccaria A. Where is your field going? A machine learning approach to study the relative motion of the domains of physics. PLoS One 2020;15(6):e0233997.
[43] Fan J, Cohen K, Shekhtman LM, Liu S, Meng J, Louzoun Y, et al. Topology of products similarity network for market forecasting. Appl Netw Sci 2019;4(1):1–15.
[44] Albora G, Pietronero L, Tacchella A, Zaccaria A. Product progression: a machine learning approach to forecasting industrial upgrading. Sci Rep 2023;13(1):1481.
[45] Balassa B. Trade liberalisation and "revealed" comparative advantage 1. Manch School 1965;33(2):99–123.
[46] Zhelezniak V, Savkov A, Shen A, Moramarco F, Flann J, Hammerla NY. Don't settle for average, go for the max: Fuzzy sets and max-pooled word vectors. 2019, arXiv preprint arXiv:1904.13264.
[47] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist 2001;1189–232.
[48] Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(11).
[49] Kingma DP, Welling M. Auto-encoding variational bayes. 2013, arXiv preprint arXiv:1312.6114.
[50] Gao M, Chen L, He X, Zhou A. Bine: Bipartite network embedding. In: The 41st international ACM SIGIR conference on research & development in information retrieval. 2018, p. 715–24.
[51] Patelli A, Napolitano L, Zaccaria A, Cimini G, Gabrielli A, Pietronero L. Economic fitness and complexity: An inquiry into the innovation and competitiveness of world regions. Joint Research Centre, European Commission; 2021.
[52] Mazzilli D, Andrea T, Pietronero L. [in preparation].
[53] Tacchella A, Mazzilli D, Pietronero L. A dynamical systems approach to gross domestic product forecasting. Nat Phys 2018;14(8):861–5.
[54] European Patent Office. http://www.oecd.org/sti/inno/intellectual-property-statistics-and-analysis.htm).
[55] Tang J, Zhang J, Yao L, Li J, Zhang L, Su. Z. ArnetMiner: Extraction and mining of academic social networks. In: Proceedings of the fourteenth ACM SIGKDD international conference on knowledge discovery and data mining. 2008, p. 990–8.
[56] Microsof Academic Division. https://www.openacademic.ai/oag/.
[57] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. 2018, arXiv preprint arXiv:1802.05365.
[58] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55(1):119–39.
[59] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey. Knowl-Based Syst 2018;151:78–94.
[60] Bustos S, Gomez C, Hausmann R, Hidalgo CA. The dynamics of nestedness predicts the evolution of industrial ecosystems. PLoS One 2012;7(11):e49393.
[61] Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge University Press; 2008.
[62] Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 2020;32(1):4–24.
[63] Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: A review of methods and applications. AI Open 2020;1:57–81.