The Dynamics of Nestedness Predicts the Evolution of Industrial Ecosystems

Sebastián Bustos, Charles Gomez, Ricardo Hausmann, and César A. Hidalgo

CID Working Paper No. 236
April 2012

Working Papers
Center for International Development
at Harvard University

The Dynamics of Nestedness Predicts the Evolution of Industrial Ecosystems

Sebastián Bustos[1,2], Charles Gomez[3] & Ricardo Hausmann[1,2,4], César A. Hidalgo[1,5,6,†]

[1] Center for International Development, Harvard University
[2] Harvard Kennedy School, Harvard University
[3] Program on Organization Studies, Stanford University
[4] Santa Fe Institute.
[5] The MIT Media Lab, Massachusetts Institute of Technology
[6] Instituto de Sistemas Complejos de Valparaiso

[†] hidalgo@mit.edu

Abstract:

Decades of research in ecology have shown that nestedness is a ubiquitous characteristic of both, biological and economic ecosystems. The dynamics of nestedness, however, have rarely been observed. Here we show that the nestedness of both, the network connecting countries to the products that they export and the network connecting municipalities to the industries that are present in them, remains constant over time. Moreover, we find that the conservation of nestedness is sustained by both, a bias for industries that deviate from the networks' nestedness to disappear, and a bias for the industries that are missing according to nestedness to appear. This makes the appearance and disappearance of individual industries in each location predictable. The conservation of nestedness in industrial ecosystems, and the predictability implied by it, demonstrates the importance of industrial ecosystems in the long term survival of economic activities.

Introduction

One of the best-documented findings of macro ecology is that rare species inhabit predominantly diverse patches, while ubiquitous species tend to inhabit both, diverse and non-diverse locations[1-4]. In ecology, the term *nestedness* is used to refer to this pattern, which has been observed numerous times in geographic patterns[1-4] and mutualistic networks[5-8]

In the case of mutualistic networks, nestedness implies that ecosystems are composed of a core set of interactions to which the rest of the community is attached[5]. The nestedness of interaction networks also implies that specialist species interact mostly with generalist species, and because generalist are less fluctuating[9], nestedness can help enhance the survival of rare species [10]. Moreover, nestedness enhances biodiversity[11] and overall ecosystem stability[12], making nestedness an important feature of interaction networks.

Nestedness, however, is a general network measure that can be used to characterize non-biological ecosystems, such as global and local economies. In fact, in the past, the nestedness of economic systems has been described for interaction networks, connecting industries to other industries, such as Leontief's input-output matrices[13], or supply relationships in the case of the New York Garment industry[14,15]. Here, instead, we study the dynamics of the geographic presence of industries by looking at their physical locations and find nestedness to be conserved over time. We do this by using Atmar and Patterson's Temperature metric[16,17] and Almeida-Neto et al's NODF[18,19], together with dynamic null models. Finally, we use the conservation of nestedness to predict the appearance and disappearance of industries at locations by showing that deviations of nestedness are associated with increases and decreases on the probability that an industry will appear or disappear at a given location.

Data & Methods

At the international level we use yearly trade data connecting 114 countries to 772 different products between 1985 and 2009. Going forward, we refer to this as the country-product network. We consider a country to be connected to a product if that country's exports per capita are larger than 25% of the world's exports per capita in that product for at least five consecutive years. This thresholds reduce the noise in the country product data coming from re-exports and helps make sure that a country is connected to the products that they export substantially and consistently. In the Supplementary Material we check for the robustness of our results by using a different definition of presences and absences based on Balassa's[20] Revealed Comparative Advantage (RCA) and find the results to be robust to this change in the definition of presences.

We note two important limitations of international trade data. First, it does not include products that are produced and consumed domestically. This is because it only considers a product once it has crossed an international border. Second, trade data is limited to goods, and therefore does not include any data on services. Despite these limitations, trade data is good for international comparisons because it is collected in a standardized classification that makes data for different countries comparable.

At the domestic level we use information on the tax residence of Chilean firms collected by Chile's *Servicio de Impuestos Internos* (SII), which is the equivalent of the United States Internal Revenue Service (IRS). Going forward, we refer to this dataset as the municipality-industry network. The municipality-industry network contains information on 100% of the firms that filed value-added and/or income taxes in Chile between 2005 and 2008 and comprises firms from all economic sectors, whether they export or not, and whether they produce goods or services. This data set consists of the universe of Chilean firms (nearly 900,000) which are classified into 700 different industries and assigned to each of Chile's 347 municipalities. Here we consider an industry to be present in a municipality if one or more firms, filing taxes under that industrial classification, declare that municipality as their tax residency.

Finally, we note that the Chilean tax data has the limitation that the tax residency of a firm can differ from the location of all of its operations. Going forward, we take the fact that our results hold in both, international trade and domestic tax data, as an indication that they are not driven by the limitations of each of these data sets and that they represent a natural characteristic of the economic ecosystem that underpins each of them. For more details on both datasets see SOM.

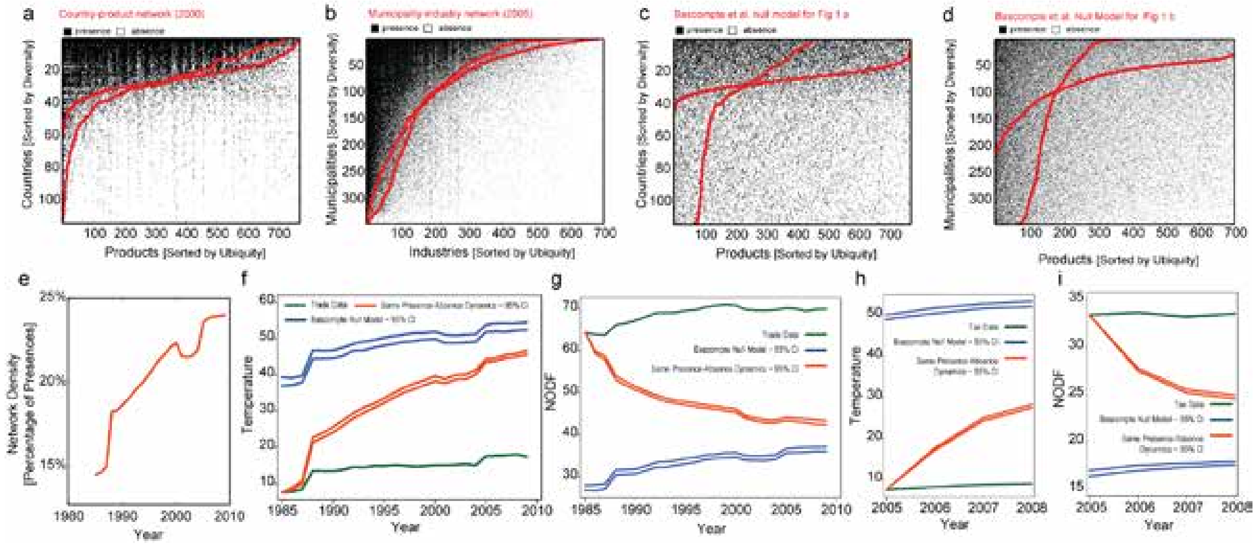

Figure 1 The nestedness of international and domestic economies. **a** Country-product network for the year 2000. **b** Municipality-industry network for the year 2005. **c** Bascompte el al. null model for the matrix shown in **a**. **d** Bascompte el al. null model for the matrix presented in **b**. In **a-d** red lines indicate the diversity of a location and the ubiquity of an industry. **e** Evolution of the density of the country-product network between 1985 and 2009. **f** Evolution of the temperature of the country-product network between 1985 and 2009 (green), its corresponding Bascompte et al. null model (blue, upper and lower lines indicate 95% conf. intervals), and that of a matrix that started identical to that for 1985, but that was evolved by considering an equal number of appearances and disappearances than in the original data (red, upper and lower lines indicate 95% conf. interval). **g** Same as **f** but using NODF. **h** Same as **f** but for the municipality-industry network **i** Same as **h** but using NODF.

Results

Figures 1 **a** and **b** show the country-product and the municipality-industry matrices (Respectively, T=15.08 and T=7.50 and NODF=70.81 and NODF=83.35. We note that T and NODF have opposite relationships to nestedness: T=0 and NODF=100 indicate perfect nestedness, whereas T=100 and NODF=0 indicate no nestedness)[21]. Here, the red lines indicate the diversity of each country and the ubiquity of each product -the number of locations where it is present- (see SM). These lines are used as a guide to indicate where presences would be expected to end if the nestedness of these networks were to be perfect. They can be thought of a simplified extinction line[17]. Figure 1 **c** and **d** show their corresponding Bascompte et al. null models[5]. In the Bascompte et al. null model, the probability to find a presence in that same cell of the matrix is equal to the average of the probability of finding it in that row and column in the original matrix. The figures show that nestedness of the original networks is clearly larger than that of their respective null models, showing that industrial ecosystems are more nested than what would be expected for comparable networks (respective null model temperatures of 51±1 and 54.5±0.7 and NODF of 35.0±0.6 and 46.5±0.3, errors are 9%% conf. intervals calculated from 100 implementations of the null model).

Next, we study the temporal evolution of nestedness. In the case of the country-product network, where a larger time series is available (1985-2009), the percentage of presences almost doubled during the observation period (Figure 1 **e**), going from less than 15% to nearly 25%. In the case of the municipality-industry network, presences went up from 22.9% to 25.7% between 2005 and 2008. The nestedness of both, the country-product and the municipality-industry networks, however, remained relatively stable during this period, as measured by both Temperature and NODF (green lines in Figure 1 **f-i**).

We test the constancy of these networks' nestedness by comparing them with two null models. The first one is an ensemble of null models[5] calculated for each respective year (blue lines in Figure 1 **f-i**). This shows that the nestedness of the empirical networks is always significantly higher than their randomized counterpart. Then, we show that a network subject to the same turnover dynamics would have lost its nestedness during the observation period. We do this by randomly adding and subtracting a number of links equal to the one gained or lost by the original network. We do this following the probability distributions defined by the Bascompte et al[5] null model to make sure that these additions and subtractions keep the degree sequence of the network close to the original. This dynamic null model preserves the exact density of the network and also the number of links that appeared and disappeared each year in each country and each product. Yet, it does not preserve nestedness. This shows that when the location of the links that appeared and disappeared is chosen differently from that observed in reality, the nestedness of the network quickly evaporates (red line in Figure 1 **f-i**). Hence, we conclude that the nestedness of these networks has remained much more stable than what would be expected from networks with the same general turnover dynamics.

Next, we show that nestedness predicts the appearance and disappearance of industries in locations. For the country-product network we consider as an appearance an increase in exports per capita from less than 5% of the world average to more than 25%. To make sure that we are capturing structural changes and not mere fluctuations, we ask the increase in exports per capita of a country to be sustained for at least 5 years. Hence, our final year of observation is 2005. Conversely, we count disappearances as a decrease in exports per capita of a country from 25% or more of the world's average to 5% or less (also sustained for at least 5 years). For the municipality-industry network we count appearances as changes from zero industries to one or more, and disappearances as changes from one or more industries to zero.

Figure 2 **a-d** visualizes the position in these networks' adjacency matrices of the industries that were observed to appear (green) and disappear (orange) in the intervening period. We predict these appearances and disappearances by fitting each observation in the industry-location network using a probit model that considers only information on the diversity of the location and the ubiquity of the industry for the initial year (see SM):

(1)

Here is the industry-location network's adjacency matrix, is the diversity of location *c* at time *t* (defined as its degree centrality or), is the ubiquity of product *p* at time *t* (defined as its degree centrality or), and where we have also added an interaction term taking the product between diversity (and ubiquity (. The error term is represented by .

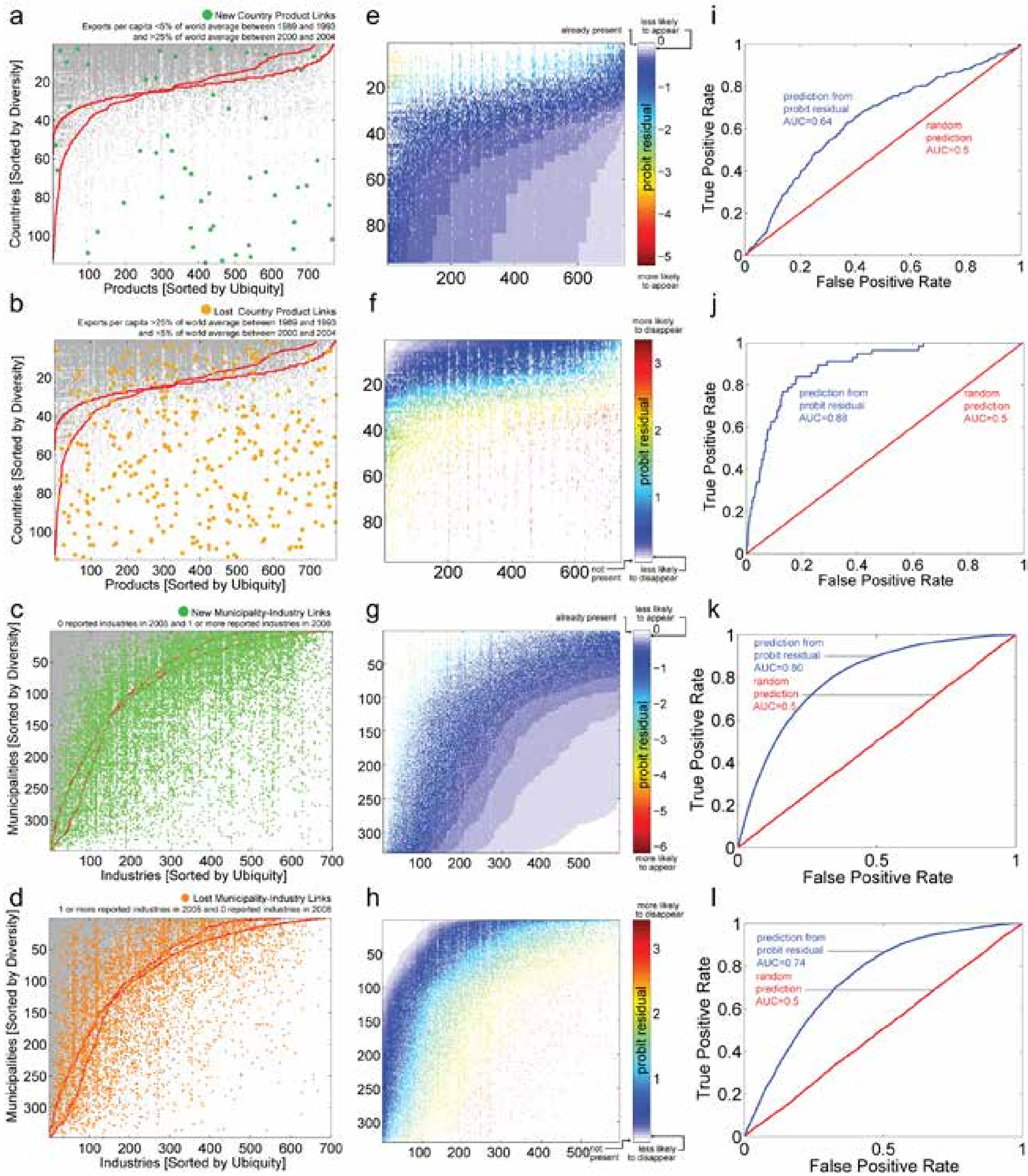

Figure 2 Nestedness predicts appearing and disappearing industries. **a** The country-product network for the year 1993 is shown in grey. Green dots show the location of industries that were observed to appear between 1993 and 2000. **b** Same as **a**, but with the industries that disappeared in that period shown in Orange. **c** The municipality-industry network is shown in grey and green dots show the location of industries that were observed to appear between 2005 and 2008. **d** Same as **c**, but with the industries that disappeared in that period shown in Orange. **e-h** Deviance residuals of the regression presented in (1) applied to the presences-absences shown in **a-d**. **i-l** ROC curves summarizing the ability of the deviance residuals shown in **e-h**, to predict the appearences and dissapearences highlighted in **a-d**.

In general, we find that the probit regression accurately explains presences and absences (average Efron's pseudo-R^2=0.53±0.02 for the country-product network and 0.54±0.01 for the municipality-industry network). This shows that, as it is expected from a nested network, the marginal sum of the adjacency matrix's rows and columns are enough to make an accurate guess of the presence or absence of that industry in that location. Here, however, we use the deviance residuals of this regression to predict future appearances and disappearances. Negative residuals, represent unexpected absences[2] and are used to rank candidates for new appearances. Positive deviance residuals, on the other hand, represent unexpected presences[2] and are used to rank the likelihood that an industry will disappear in the future. (Figures 2 **e-h**).

Finally, we quantify the accuracy of these predictions using the area under the Response Operator Characteristic curve or ROC curve[22,23]. An ROC curve plots the True Positive Rate of a prediction criterion as a function of its False Positive Rate. The Area Under the Curve, or AUC, is commonly used to measure the accuracy of the prediction criterion[22,23]. A random prediction will find true positives and false positives at the same rate, and therefore will give an AUC of 0.5. A perfect prediction, on the other hand, will find all true positives before hitting any false positive and will be characterized by an AUC=1. Figures 2 **i-l** show the ROC curves obtained when the appearances and disappearances shown in Figures 2 **a-d** are predicted using the deviance residuals obtained from (1) for data on the initial year. In all cases, the ROC curves of these predictions (in blue), have an area that is significantly larger than the one expected for a random prediction (in red), showing that nestedness can help predict which links in these industry-location networks are more likely to appear or disappear.

Finally, we extend this analysis to all pairs of years. Figures 3 **a** and **b** show the number of events (appearances or disappearances) for each pair of years for the international trade data. As expected, there are fewer events for pairs of years that are close by in time. Also, we note that the number of appearances is larger than that of disappearances, a fact that is consistent with the observed increase in the density of the network. Figure 3 **c** shows the AUC value obtained for each pair of years, showing that for the country product network, disappearances (Fig 3 **b**) are predicted much more accurately than appearances.

The time series data available for Chile's municipality-industry network is much more limited. Hence, we show the average number of events (Figure 3 **d**), and the average AUC for networks separated by a given number of years (Figure 3 **e**). Here, we find that predictions of appearances and disappearance are both remarkably strong, and that there is no statistically significant difference in the predictability of both kinds of events.

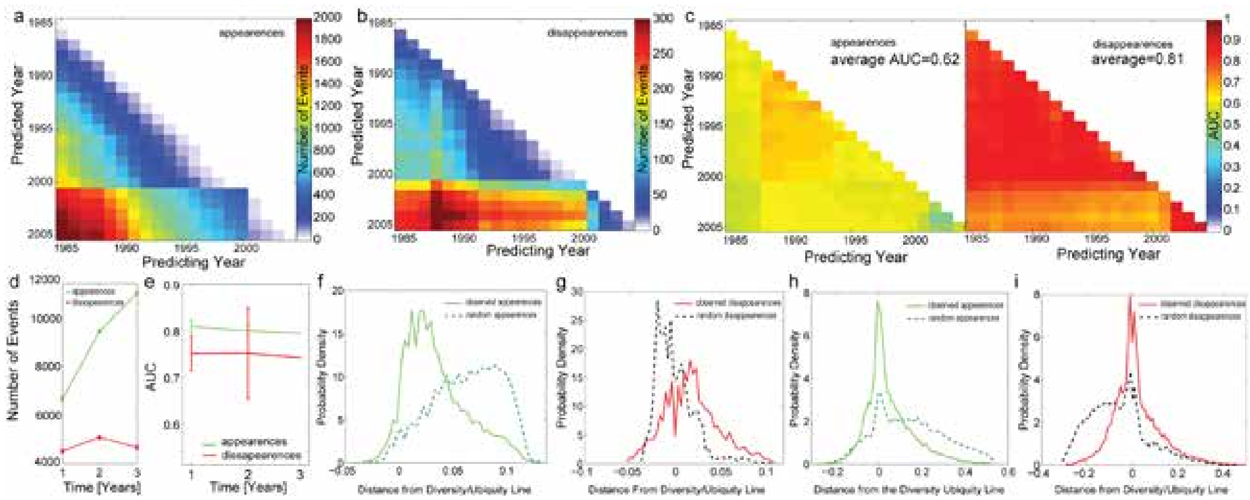

Figure 3 Predicting appearances and disappearances using nestedness. **a** Number of appearances for every pair of years in the country-product network. **b** Number of disappearances for every pair of years for the country-product network. **c** Accuracy of the predictions for each pair of years measured using the Area Under the ROC Curve (AUC). **d** Average number of appearances and disappearances for the Chilean data (error bar smaller than symbol). **e** Average accuracy of the predictions for the municipality-industry network. Error bars indicate 99% confidence intervals. **f** Distribution for the distance to the diversity-ubiquity line obtained for the observed appearances and for an equal number of random appearances. **g** Same as **f** but for disappearances. **h** Same as **f**, but for the municipality-industry network. **i** Same as **h** but for disappearances.

To conclude, we look at the position in the network's adjacency matrix of appearances and disappearances. If the stability of nestedness is related to the location in this matrix of industrial appearances and disappearances, then appearances should be closer to the diversity-ubiquity line than random appearances. By the same token, disappearances should be farther away. For each event, we estimate its distance to the diversity and the ubiquity lines illustrated in figures 1 **a-d** and figures 2 **a-d** using

(2)

Here and are respectively the lines of diversity and ubiquity (see SM), is the position in the adjacency matrix of the i^{th} event, and N_c and N_p are respectively the number of locations and industries in the network. We use N_c and N_p to normalize the maximum possible vertical and horizontal distances to 1 and thus make sure that the measure is less sensitive to the rectangularity of the different matrices. The || operator represents the Euclidean distance and = 1 if the position of the event is outside of the nested area defined by both and and -1 otherwise (see SM). As a benchmark comparison we consider an equal number of appearances and disappearances, but draw these from a random set of eligible positions in the adjacency matrix.

Figure 3 **f-i** compare the distributions of distances (D) with those associated with an equal number of random appearances or disappearances. We find that appearances tend to lie significantly closer to the diversity/ubiquity lines than what would be expected for an equal number of random events (ANOVA F=59,935, p-value=0 for the country-product network and ANOVA=10895 p-value=0 for the municipality-industry network). In the case of disappearances, the opposite holds true. The observed appearances tend to be mostly located outside of the nested area defined by the diversity/ubiquity lines. Our random expectation, however, would be for disappearances to come mostly from the highly populated area inside the diversity/ubiquity lines. Once again, differences between observations and null model expectations are highly significant for both networks (ANOVA F=6246 p-value=0 for the country-product network and ANOVA F=6463 p-value=0 for the municipality-industry network).

These results show that both, the global and domestic economic structures explored in this paper, evolve following a pattern in which unexpected absences are more likely to become the future locations of an industry and where unexpected presences signal industries that are more likely to disappear from a location. Overall, this dynamics helps keep the nestedness of these networks constant across time.

Discussion

In this paper we showed that industry-location networks are nested, just like industry-industry networks[13-15], or their biological counterparts[1-4,16,17]. Using time series data for both, international and domestic economies, we showed that the nestedness of these networks tends to remain constant over time and that this empirical conservation law can be used to predict the pattern of industrial appearances and disappearances.

The strong link between biological and industrial ecosystems opens a variety of questions. First, are the models currently used to explain nestedness able to explain its conservation across time? Second, is the geographical nestedness described in this paper a consequence of industry-industry nestedness, or are these independent phenomena? Third, are the mechanisms generating nestedness at the global level the same that generate nestedness at the national level?

In this paper we showed that the geographical nestedness of industries holds at both, the global and at the national scale. This is certainly not the case for biological ecosystems, since the biota of the artic is not a subset of that of the rain forest. The fact that the nestedness of industrial ecosystems holds at scales as large as that of the world economy suggests that the coupling between international economies is strong. This highlights the importance of understanding the global economy as a unified ecosystem, since after all, its nestedness suggests that it is working as one.

The predictability implied by nestedeness, on the other hand, has important implications in a world where income is connected to the mix of products that a country makes[24,25]. Ultimately, the dynamics implied by nestedness could represent a fundamental constraint to the speed at which international incomes could either converge or diverge.

More research will certainly need to be done on both, the causes of the structures and the time patterns that were uncovered in this paper. This will require strengthening the bridge between the natural and social sciences because, if there is something that the nestedness of economies show, is that humans tend to generate patterns in social systems that strongly mimic those found in nature[26,27].

ACKNOWLEDGEMENTS:

We acknowledge the support of the Santo Domingo Foundation, Standard Bank, and the ABC Career Development Chair at the MIT Media Lab. We also thank Frank Neffke, Daniel Stock & Muhammed Yildirim for their comments.

References:

1 Hulten, E. *Outline of the history of Artic and Boreal biota during the Quaternary Period*, Lund University, (1937).

2 Ulrich, W., Almeida, M. & Gotelli, N. J. A consumer's guide to nestedness analysis. *Oikos* **118**, 3-17, doi:10.1111/j.1600-0706.2008.17053.x (2009).

3 Darlington, P. J. *Zoogeography: the geographical distribution of animals.*, (wiley, 1957).

4 Daubenmire, R. Floristic plant geography of eastern Washington and northern Idaho. *Journal of Biogeography* **2** (1975).

5 Bascompte, J., Jordano, P., Melian, C. J. & Olesen, J. M. The nested assembly of plant-animal mutualistic networks. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9383-9387, doi:10.1073/pnas.1633576100 (2003).

6 Dupont, Y. L., Hansen, D. M. & Olesen, J. M. Structure of a plant-flower-visitor network in the high-altitude sub-alpine desert of Tenerife, Canary Islands. *Ecography* **26**, 301-310, doi:10.1034/j.1600-0587.2003.03443.x (2003).

7 Ollerton, J., Johnson, S. D., Cranmer, L. & Kellie, S. The pollination ecology of an assemblage of grassland asclepiads in South Africa. *Annals of Botany* **92**, 807-834, doi:10.1093/aob/mcg206 (2003).

8 Gilarranz, L. J., Pastor, J. M. & Galeano, J. The architecture of weighted mutualistic networks. *Oikos*, 001–009, doi:doi: 10.1111/j.1600-0706.2011.19592.x (2011).

9 Turchin, P. & Hanski, I. An empirically based model for latitudinal gradient in vole population dynamics. *American Naturalist* **149**, 842-874, doi:10.1086/286027 (1997).

10 Jordano, P. PATTERNS OF MUTUALISTIC INTERACTIONS IN POLLINATION AND SEED DISPERSAL - CONNECTANCE, DEPENDENCE ASYMMETRIES, AND COEVOLUTION. *American Naturalist* **129**, 657-677, doi:10.1086/284665 (1987).

11 Bastolla, U. *et al.* The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature* **458**, 1018-U1091, doi:10.1038/nature07950 (2009).

12 Bascompte, J., Jordano, P. & Olesen, J. M. Asymmetric coevolutionary networks facilitate biodiversity maintenance. *Science* **312**, 431-433, doi:10.1126/science.1123412 (2006).

13 Leontief, W. W. The Structure of the U.S. Economy. *Scientific American* **212**, 25-35 (1965).

14 Saavedra, S., Reed-Tsochas, F. & Uzzi, B. A simple model of bipartite cooperation for ecological and organizational networks. *Nature* **457**, 463-466, doi:10.1038/nature07532 (2009).

15 Saavedra, S., Stouffer, D. B., Uzzi, B. & Bascompte, J. Strong contributors to network persistence are the most vulnerable to extinction. *Nature* **478**, 233-U116, doi:10.1038/nature10433 (2011).

16 Atmar, W. & Patterson, B. D. THE MEASURE OF ORDER AND DISORDER IN THE DISTRIBUTION OF SPECIES IN FRAGMENTED HABITAT. *Oecologia* **96**, 373-382, doi:10.1007/bf00317508 (1993).

17 Patterson, B. D. & Atmar, W. NESTED SUBSETS AND THE STRUCTURE OF INSULAR MAMMALIAN FAUNAS AND ARCHIPELAGOES. *Biological Journal of the Linnean Society* **28**, 65-82, doi:10.1111/j.1095-8312.1986.tb01749.x (1986).

18 Almeida, M., Guimaraes, P. R. & Lewinsohn, T. M. On nestedness analyses: rethinking matrix temperature and anti-nestedness. *Oikos* **116**, 716-722, doi:10.1111/j.2007.0030-1299.15803.x (2007).
19 Almeida-Neto, M., Guimaraes, P., Guimaraes, P. R., Loyola, R. D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* **117**, 1227-1239, doi:10.1111/j.2008.0030-1299.16644.x (2008).
20 Balassa, B. COMPARATIVE ADVANTAGE IN MANUFACTURED GOODS - A REAPPRAISAL. *Review of Economics and Statistics* **68**, 315-319 (1986).
21 Guimaraes, P. R. & Guimaraes, P. Improving the analyses of nestedness for large sets of matrices. *Environmental Modelling & Software* **21**, 1512-1513, doi:10.1016/j.envsoft.2006.04.002 (2006).
22 Bradley, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition* **30**, 1145-1159, doi:10.1016/s0031-3203(96)00142-2 (1997).
23 Zweig, M. H. & Campbell, G. RECEIVER-OPERATING CHARACTERISTIC (ROC) PLOTS - A FUNDAMENTAL EVALUATION TOOL IN CLINICAL MEDICINE. *Clinical Chemistry* **39**, 561-577 (1993).
24 Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 10570-10575, doi:10.1073/pnas.0900943106 (2009).
25 Hausmann, R., Hwang, J. & Rodrik, D. What you export matters. *Journal of Economic Growth* **12**, 1-25, doi:10.1007/s10887-006-9009-4 (2007).
26 Bettencourt, L. & West, G. A unified theory of urban living. *Nature* **467**, 912-913 (2010).
27 West, G. B., Brown, J. H. & Enquist, B. J. A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122-126 (1997).

SUPPLEMENTARY MATERIAL FOR:

The Dynamics of Nestedness Predicts the Evolution of Industrial Ecosystems

Sebastián Bustos[1,2], Charles Gomez[3] & Ricardo Hausmann[1,2,4], César A. Hidalgo[1,5,6,†]
[1] Center for International Development, Harvard University
[2] Harvard Kennedy School, Harvard University
[3] Program on Organization Studies, Stanford University
[4] Santa Fe Institute.
[5] The MIT Media Lab, Massachusetts Institute of Technology
[6] Instituto de Sistemas Complejos de Valparaiso
[†] hidalgo@mit.edu

TABLE OF CONTENTS

DATA DETAILS:

International Trade Data:

The international data set is a merge of two data sources: The Feenstra et al. (2005) data set, which has data for the years prior to 2000, and the UN Comtrade database (comtrade.un.org), which we used for the period going from 2001 to 2009. Both dataset follow the product classification established by the Standard International Trade Classification (SITC) revision 2*. In the UN Comtrade dataset we associated countries to products according to what was reported as exports to the WLD category (World). For the products in which no exports to WLD (World) was found, exports were reconstructed using the reports from importing countries, when available, and by aggregating the reported bilateral exports of the exporting country as a last resource. We prioritize imports over exports because imports tend to be more tightly controlled than exports.

While the Feenstra et. al (2005) data set contains trade starting 1962, we chose 1985 as our starting year because there are several reclassifications of the data that affect their reliability for previous years (see SM2 Data Continuity). Since presences are averages over 5 years, the first year that is included in our dataset is 1981 (in the counting of presences for 1985).

We find, however, that international trade data is characterized by a nested matrix even for the years that we do not include in this paper. Figure SM1 shows the Temperature and NODF calculated for all years. Our choice to restrict the number of years in the dataset was performed to reduce the number of false appearances and disappearances that could be introduced by reclassifications of the SITC categories.

---

* For more information visit http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=8&Lg=1

Figure 1 - Temperature and NODF for all years

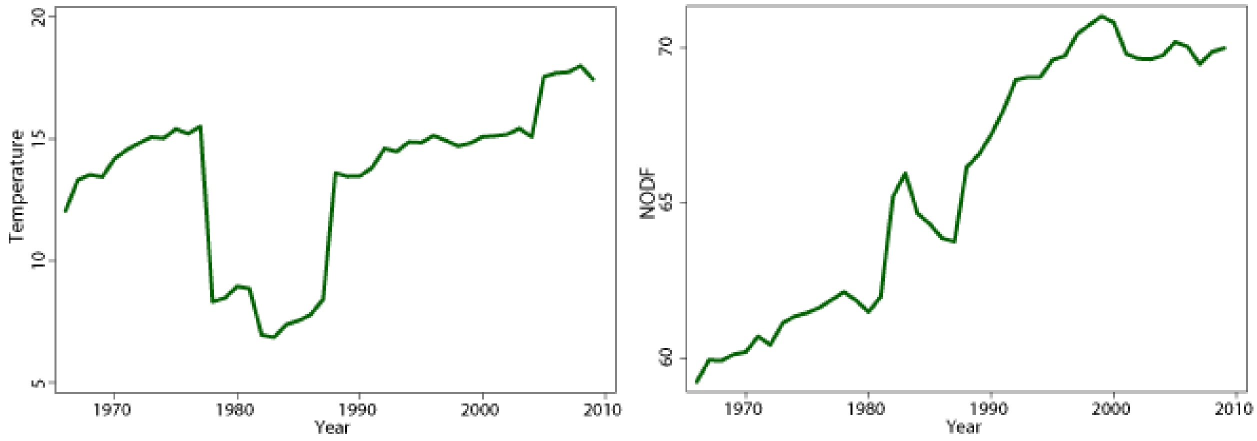

Figure SM4 The nestedness of international trade data from 1966 to 2009.

Finally, we restrict the sample of countries for all those that have a population of at least 1.25 million in the year 2000. We also remove countries from the Former Soviet Union (FSU), because this countries lack data for the 1980's, and have noisy data for the early 1990's. For Germany, we use data on West Germany for the years prior to 1992.

The final dataset consists of 114 countries and 775 products, classified according to the SITC4 rev2 classification (http://reportweb.usitc.gov/commodities/naicsicsitc.html). The datasets includes only tradable products, from raw materials and agriculture, to manufactures and chemicals.

Domestic Tax Data:

The domestic data for Chile consists of a matrix indicating the number of firms from a given industry in each municipality. The data has records for the year 2005, 2006, 2007 and 2008 and is based on the fiscal residence of each firm (it is hence a firm, and not a establishment level dataset). The number of firms reported for each year is shown in table 1.

Year	Number of Firms
2005	862,405
2006	876,948
2007	891,383
2008	899,156

Table 1: Number of Firms in the Chilean Tax Data

These data contains information on the universe of Chilean firms and includes firms from all economic sectors, from raw materials and manufacturing, to restaurant, retail and banking services. The data contains information for 347 municipalities and 700 industries classified according to the Código the Actividad Económica (CAE) (http://www.sii.cl/catastro/codigos.htm).

PRESENCE-ABSENCE MATRIX DEFINITION:

For the international trade data set, we define the presences of an industry in a country if that country has exports per capita that are at least 25% of the world average for 5 consecutive years. Formally, we do this following:

Where M_{cp} is the presence-absence matrix, EXP_{cp} are the exports of product p by country c, and P_c is the population of country c. For the domestic tax data, we define as a presence a municipality that has one or more firms filing taxes under that industrial classification. We use a single year in this case.

NESTEDNESS METRICS: TEMPERATURE AND NODF

We calculate the nestedness of the exports per capita absence-presence matrices using both, Atmar and Patterson's temperature metric and Almeida-Neto et al.'s NODF metric. Preparation of these matrices for both analyses is similar. For the temperature metric, the rows and columns of a matrix are sorted and rank-ordered to yield a nested matrix with the absolute minimum temperature possible for this matrix. For the NODF metric, the rows and the columns of a matrix are swapped and rank-ordered by the sum of the presences in each of these rows and columns, respectively. The transformed matrices are then ready to be processed by the following algorithms. For a more detailed explanation, please reference the respective works of Atmar and Patterson (1993) and Almeida-Neto et al. (2008). Also the review by Ulrich, Almeida and Gotelli (2009) is a good place to learn about both of these metrics.

ATMAR AND PATTERSON'S TEMPERATURE MEASURE

Atmar and Patterson's temperature metric calculates the number and the degree of unexpected presences and absences in an ordered adjacency matrix. Unexpected presences and absences are calculated with respect to an *extinction line* that separates the adjacency matrix into two areas: The top-left triangle, which we will call Section 1, where only presences are expected to appear, and the bottom right triangle, which we call Section 2, where only absences are expected (Figure SM2). In a perfectly nested matrix an ideal extinction line is a skew diagonal bisecting the matrix, where all of the presences are to one side of the line and all of the absences are to the other side.

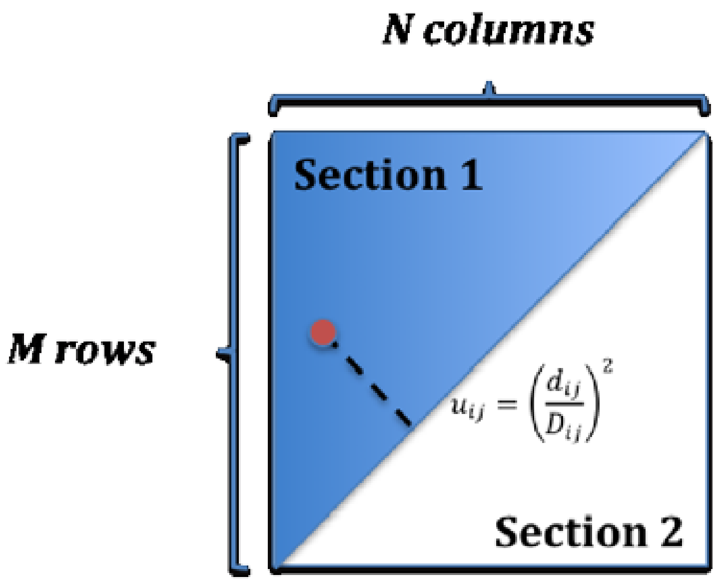

Figure SM2: A Perfectly Nested Matrix with M Rows and N columns

Presences in Section 1 that are closer to the extinction line are considered more likely to face extinction. A presence in Section 2, on the other hand, is considered an unexpected presence. Thus, the distance from the extinction line captures the degree of unexpectedness of presence. Conversely, absences in Section 1 are considered unexpected absences.

A perfectly nested matrix is characterized by a temperature of zero degrees. Alternatively, a fully disordered matrix is characterized by a temperature of 100 degrees.

The degree of "unexpectedness" for any presence or absence is the squared ratio of its distance, d_{ij}, to the ideal extinction line, D_{ij}. This local unexpectedness is expressed as:

The degree of unexpectedness for the matrix, *U*, is the sum of each of these local unexpectedness values. This sum is normalized by the number of rows (m) and columns (n), to ensure the measure is unaffected by the size or the shape of the adjacency matrix:

The total unexpectedness is transformed to a temperature scale using a normalization factor. The temperature scale goes from 0 degrees, corresponding to a perfectly ordered matrix, to 100 degrees, indicating a matrix full with unexpected values:

where U_{max} is 0.04145.

Almeida-Neto et al.'s NODF Measure

The Nested Overlap and Decreasing Fill, or NODF metric, measures the degree of overlap between an adjacency matrix's rows and columns. NODF is determined by comparing all row-row and all column-column pairs. A row-row pair *ij* is any row *i* paired with each row above it, row *j*, in an ordered matrix. Similarly, a column-column pair *ij* is any column *i* paired with each column behind it, column *j*, in an ordered matrix. This is first achieved by calculating the Paired Overlap, PO_{ij}, for each row-row and each column-column pair. PO_{ij} is calculated as the percentage of presences in row or column *i* that are also present in row or column *j*:

where MT_j is marginal total, the sum of presences in row or in column j, and O_{ij} is the number of presences overlapping between the row-row or the column-column pair.

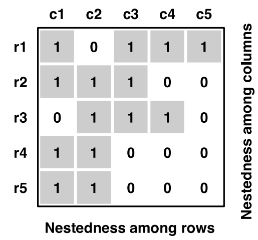

Figure **SM3**: An Ordered Matrix[†]

For example, consider rows r1 and r2 from Figure 2:

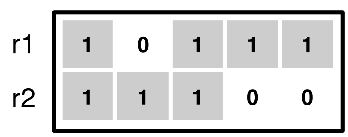

Figure SM3a: A Sample Row-Row Pair from the Matrix in Figure SM3

In figure SM3a, r2 – the less populated row with three presences – overlaps with two presences in r1 – the more populated row. The PO_{ij} for the r1-r2 pairing is thus two presences divided by three presences, or PO_{12} = 66.67%. Similarly for columns, consider figure 2b:

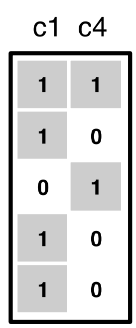

Figure SM3b: A Sample Column-Column Pair from the Matrix in Figure 2

Column c4 – the less populated column with only two presences – shares only shared presence with column c1 – the more populated column. Thus, the PO_{ij} for the c1-c4 pairing is PO_{14} = 50%.

[†] Figures SM3, SM3a, SM3b, and SM4 are taken directly from: *Almeida-Neto, M., Guimaraes, P., Guimaraes, P. R., Loyola, R. D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. Oikos* ***117****, 1227-1239, doi:10.1111/j.2008.0030-1299.16644.x (2008).*

With the paired overlap, we can now calculate both the decreased fill, $DF_{ij,,}$ for every row-row and column-column pair. The DF_{ij} takes one of two values depending on the marginal total, or *MT*, of the rows or the columns in the pair. Thus, in an ordered adjacency matrix, if the marginal total of row *i*, MT_i, is less than the marginal total of row *j*, MT_j, then DF_{ij} takes on the value of 100. Otherwise, if MT_i is greater than or equal to the MT_j, then DF_{ij} takes on the value of 0.

The penultimate variable is the paired nestedness, N_{ij}, for every row-row and every column-column pair. Similar to DF_{ij}, N_{ij} can take on only one of two values based on the DF_{ij} and the PO_{ij} of its row-row or its column-column pair. Thus, if $DF_{ij} = 100$, then $N_{ij} = \mathrm{PO}_{ij}$; otherwise, $N_{ij} = 0$.

The N_{ij} is calculated for every row-row and column-column pair in the matrix. The NODF score is the average of all N_{ij} values:

where and are the the total number of possible row-row and column-columns pairs in the matrix. Figure SM4 illustrates the entire NODF calculation for the matrix in figure 2.

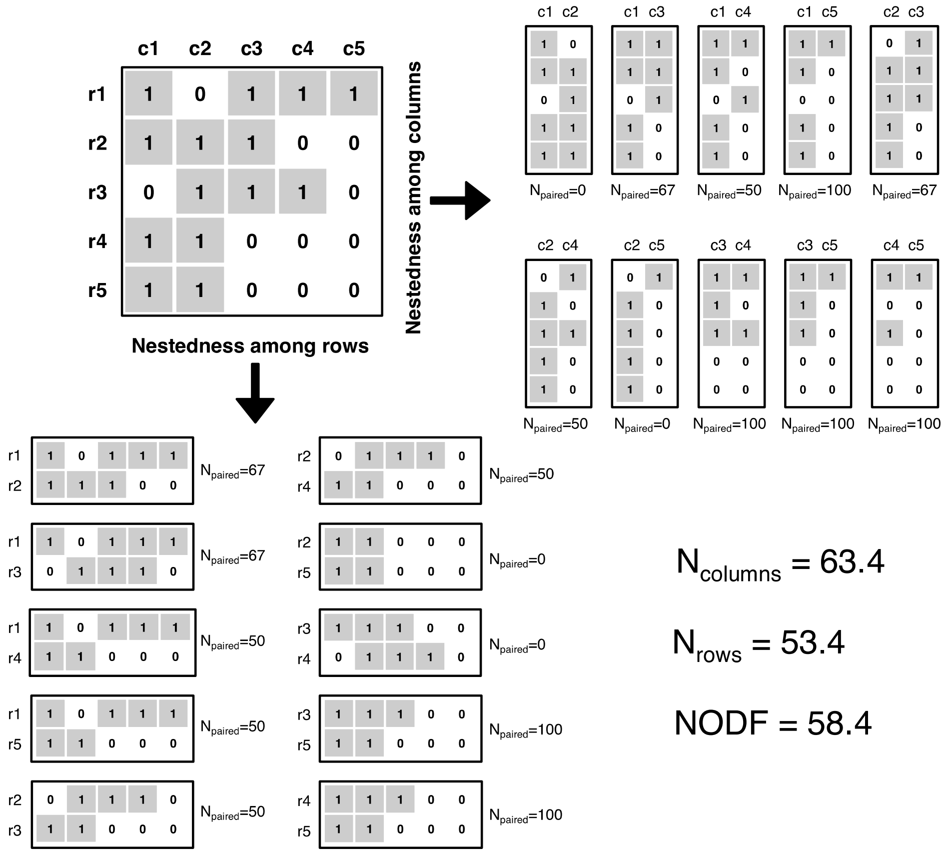

Figure SM4: The Entire NODF Calculation for the Matrix in Figure 2

NULL MODELS

Static Null Model (Bascompte et al.)

Bacompte et al. (2003) introduced a null model to show whether the nested order of the data is statistically meaningful. For this, they introduced a null model (M_{cp}*) in which the probability to find a presence in that same cell of the matrix is equal to the average of the probability of finding it in that row and column in the original matrix (M_{cp}).

Using this model we performed 100 random realizations of the matrix for each year. Then we calculated the Temperature and NODF of each realization of the resulting null matrices to obtain a distribution of possible outcomes. Figures 1f and 1g show the 95% confidence interval Temperature and NODF of these null matrices. Since both the Temperature and NODF of the matrices lie outside the confidence interval, we can say that the nestedness of the matrix is statistically significant.

Dynamic Null Model

To show that nestedness of the network connecting countries to the products is conserved over time we introduce a dynamic null model. This dynamic null model preserves the exact density of the network and also the number of links that appeared and disappeared each year in each country and each product. First, we calculate the number of links that appeared and disappeared for each year. Then, starting with data for the year 1985, we introduced the same number of appearances and disappearances that were observed in the transition between 1985 and 1986 with a location in the matrix determined by the Bascompte et al. null model explained above. The result is a matrix for year 1986 that has the same density of the real data. We continue this procedure to the last year of our data. The procedure was repeated 100 times, and for each matrix we calculated the Temperature and NODF. Figures 1f and 1g of the main text show the 95% confidence interval of the distribution of Temperature and NODF of these dynamic null matrices. The figures show that the dynamic null model does not keep the same level of order of the real data and disorders rather quickly. Hence, the order of the real-data remains highly nested despite large changes in the links of the network.

DIVERSITY AND UBIQUITY LINES, AND DISTANCE OF EVENTS

To gauge the position of appearances and disappearances in the presence absence matrix, we introduce the diversity and ubiquity lines as a line indicating where presences would be expected to end if the matrix were to be perfectly nested.

In an adjacency matrix sorted by the sum of its rows and columns, the diversity line is a line that goes through the column that is equal to the number of presences in that row. In the case of locations (countries or municipalities) this is equal to their diversity. For each column, the ubiquity line is one that goes through the row equal to its number of presences. In the case of an industry, this represents its ubiquity, or the number of locations where it is present. Figure SM5 illustrates the diversity and ubiquity lines, and the distance of an event to them.

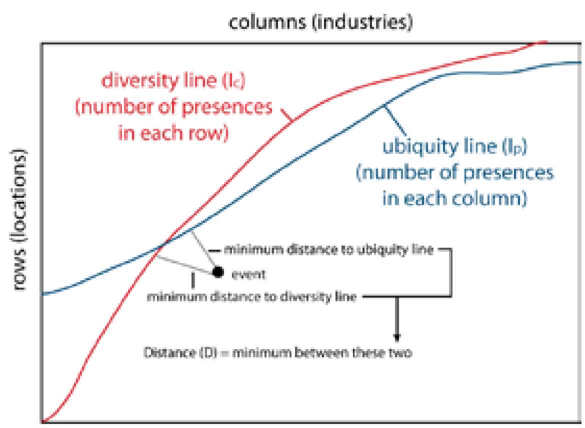

Figure SM5 Diversity and Ubiquity lines and the distance of an event to them.

ROBUSTNESS CHECKS FOR NESTEDNESS

In this section we show the robustness of some of the main stylized facts of the paper to a different definition of presences and absences. Here, we indicate presences and absences using Balassa's (1986) definition of Revealed Comparative Advantage (RCA). Moreover, we use data for all years (1962-2009).

Balassa's (1986) RCA compares the share of a country's exports that a product represents with the share of world trade represented by that same product. If that product represents a share of that country's export that is larger than its share of world trade, then we say that the country has RCA in that product. We define a presence as having RCA≥1 in a product for at least five consecutive years. Figure SM6 a shows the increase in the number of links in the presence-absence matrix of the RCA network between 1966 and 2009. Figure SM6 b shows the RCA country-product network and their respective diversity and ubiquity line for the year 2000. This matrix is characterized by a temperature of 12±2 and a NODF of 21±8. Figure SM6 c shows its respective Bascompte et al. (2003) null model. In this case, temperature is 12±2 and NODF is 21±8.

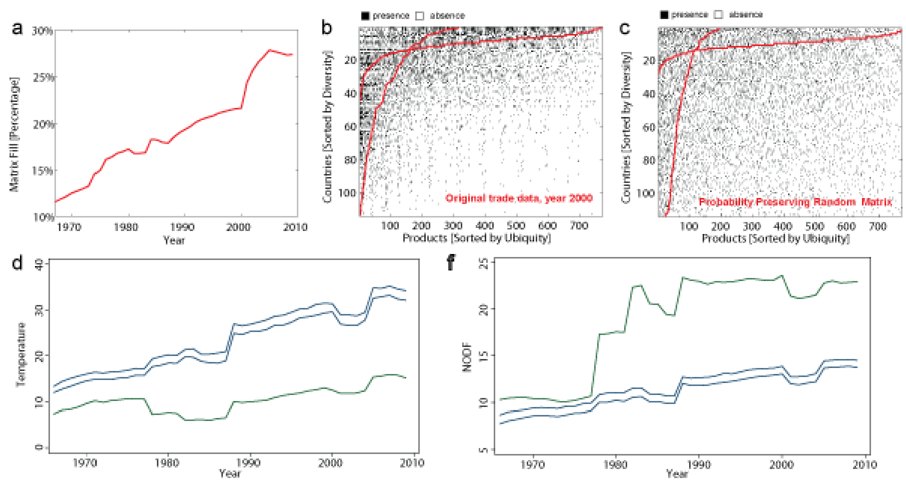

Figure SM6 The nestedness of international economies using RCA. **a** Evolution of the density of the country-product network between 1985 and 2009. **b** Country-product network for the year 2000. **c** Bascompte el al. null model for the matrix shown in **b**. **d** Evolution of the temperature of the country-product network between 1966 and 2009 (green), its corresponding Bascompte et al. null model (blue, upper and lower lines indicate 95% conf. intervals). **f** Same as **d** but using NODF.

Figure SM7 reproduces Figure 2 of the paper's main text using Balassa's (1986) definition of RCA. These figures illustrate the robustness of the analysis to the difference in definition. It is worth noting that using Balassa's (1986) definition of RCA, instead of the exports per capita definition used in the main text, provides slightly weaker, albeit statistically significant, predictions.

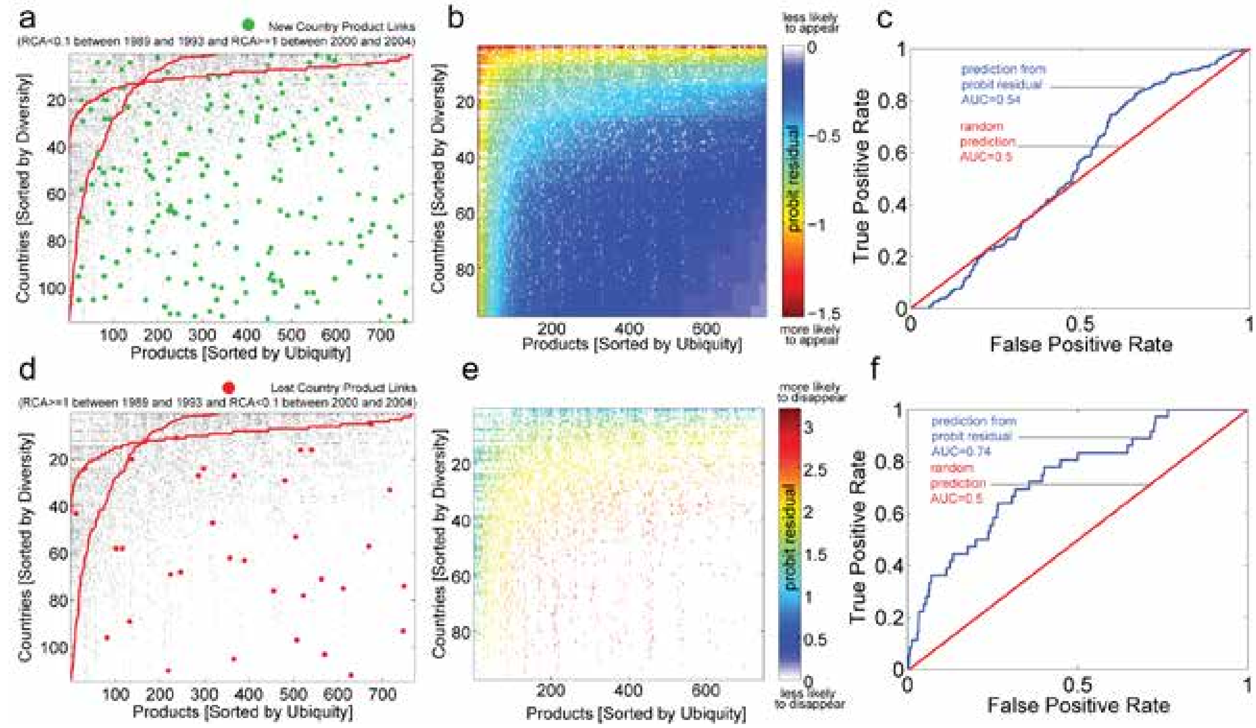

Figure SM7 Nestedness using RCA. **a** The country-product network for the year 1993 is shown in grey. Green dots show the location of industries that were observed to appear between 1993 and 2000. **d** Same as **a**, but with the industries that disappeared in that period shown in Orange. **b and e** Deviance residuals of the regression presented in (1) of the main text applied to the presences-absences shown in **a-d**. **c and f** ROC curves summarizing the ability of the deviance residuals shown in **b-e**, to predict the appearances and disappearances highlighted in **a and d**.

Finally, Figure SM8 reproduces figure 3 of the main text using Balassa's (1986) RCA to indicate presences and using data for all years. Here, we see that results hold except when the years 1974-1977 are used as predictors. This is because of a large discontinuity in the data classification introduced between 1973 and 1974. This is documented in the second supplementary material of the paper, which shows the fraction of countries that had >0 exports in each product category for all years for the 1006 product categories in the SITC4 rev2 classification.

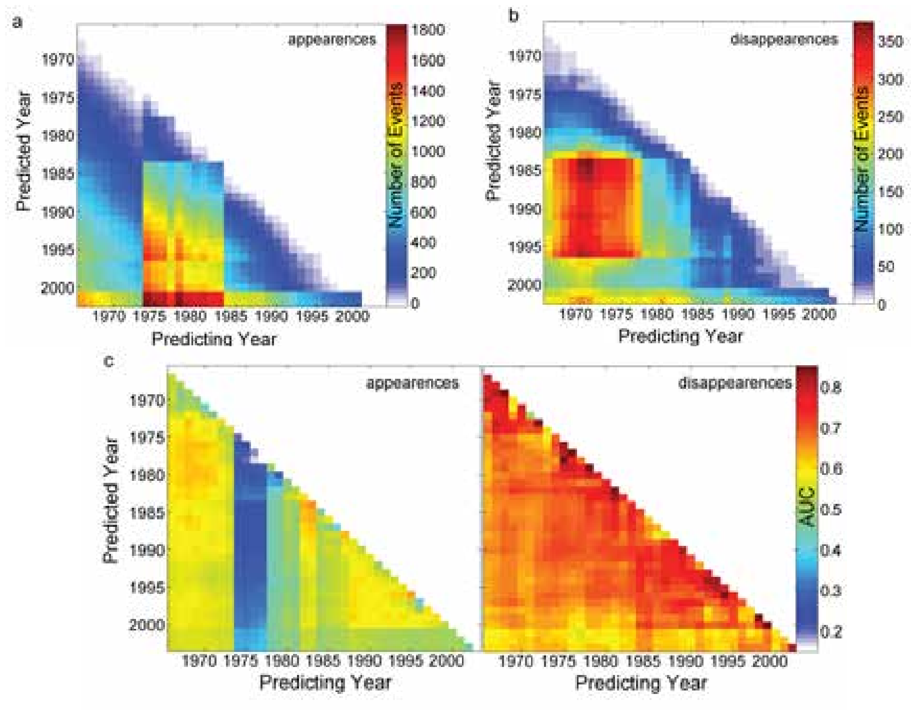

Figure SM8 Predicting appearances and disappearances using nestedness. **a** Number of appearances for every pair of years in the country-product network. **b** Number of disappearances for every pair of years for the country-product network. **c** Accuracy of the predictions for each pair of years measured using the Area Under the ROC Curve (AUC).

References:

Almeida-Neto, M., Guimaraes, P., Guimaraes, P. R., Loyola, R. D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* **117**, 1227-1239, doi:10.1111/j.2008.0030-1299.16644.x (2008).

Atmar, W. & Patterson, B. D. The Measure of Order and Disorder in the Distribution of Species in Fragmented Habitat. *Oecologia* **96**, 373-382, doi:10.1007/bf00317508 (1993).

B. Balassa, The Review of Economics and Statistics, 68, 315 (1986).

Bascompte, J., Jordano, P., Melian, C. J. & Olesen, J. M. The nested assembly of plant-animal mutualistic networks. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9383-9387, doi:10.1073/pnas.1633576100 (2003).

Feenstra, R. R. Lipsey, H. Deng, A. Ma and H. Mo. "World Trade Flows: 1962-2000" NBER working paper 11040. National Bureau of Economic Research, Cambridge MA (2005).

Ulrich, W., Almeida, M. & Gotelli, N. J. A consumer's guide to nestedness analysis. *Oikos* **118**, 3-17, doi:10.1111/j.1600-0706.2008.17053.x (2009).