Supporting Online Material for

The Product Space Conditions the Development of Nations

C. A. Hidalgo,* B. Klinger, A.-L. Barabási, R. Hausmann

*To whom correspondence should be addressed. E-mail: chidalgo@nd.edu

Published 27 July 2007, *Science* **317**, 482 (2007)
DOI: 10.1126/science.1144581

This PDF file includes:

Materials and Methods
SOM Text
Figs. S1 to S18
Table S1
References

The Concept of Proximity

The Intuitive Definition

The concept of proximity formalizes the intuitive idea that the ability of a country to produce a product depends on its ability to produce other ones. For example, a country with the ability to export apples will probably have most of the conditions suitable to export pears. They would certainly have the soil and the climate, together with the appropriate packing technologies, frigorific trucks and containers. They would also have the human capital, particularly the agronomists that could easily learn the pear business. However, when we consider a different business such as mining, textiles or appliance manufacture, all or most of the capabilities developed for the apple business render useless. Unfortunately this intuitive definition of proximity is, very cumbersome to measure. It requires quantifying the overlap between the set of markets related to each product. Thus, we measure proximity by using an outcome based method founded on the assumption that similar products are more likely to be exported in tandem.

Details of the measurement

First, a stringent measure of exports is needed. We do not want to consider marginal exports, and thus we say that a country exports a product whenever they have Revealed Comparative Advantage (RCA) in it. We use the Balassa[1] definition of RCA which is given by

$$RCA(c,i) = \frac{\frac{x(c,i)}{\sum_i x(c,i)}}{\frac{\sum_c x(c,i)}{\sum_{i,c} x(c,i)}}$$

where ***x(c,i)*** is the value of the exports of country ***c*** in the ***i***'th good. Basically RCA is larger than one when the share of exports of country on a given product is larger than the share of that product on the global trade. This definition of RCA allows us to set a hard threshold for a countries export. When RCA(c,i) is greater or equal to 1 we say that

country ***c*** exports product ***i***, and when RCA(c,i)<1 that country is not an effective exporter of that product.

Using RCA as an indication of a country effectively exporting a good, we define the proximity between goods ***i*** and ***j*** as:

$$\phi_{ij} = min\left\{ P(RCA_i|RCA_j), P(RCA_j|RCA_i) \right\}$$

where ***P(RCAi|RCAj)*** is the conditional probability of exporting good ***i*** given that you export good ***j***. In this definition we consider the minimum between both conditional probabilities because in the case that a country is the sole exporter of a particular good we would have that the conditional probability of exporting any other good given that one would be equal to one for all other goods exported by that country. The converse is not true and by taking the minimum we get rid of this problem and at the same time symmetrize the proximity matrix.

More details about the motivation of proximity and the option value associated with it were covered in the work of Hausmann and Klinger [2].

References

[1] B. Balassa, *The Review of Economics and Statistics*, **68**, 315 (1986).
[2] Ricardo Hausmann and Bailey Klinger, *Structural Transformation and Patterns of Comparative Advantage in the Product Space*, CID Working Paper No. 128, August 2006, -abstract-

Source Data

International trade data is taken from Feenstra, Lipsey, Deng, Ma, & Mo's "World Trade Flows: 1962-2000" dataset. This dataset consists of imports and exports both by country of origin and by destination, with products disaggregated to the SITC revision 4, four-digit level. The authors build this dataset using the United Nations COMTRADE database. The authors cleaned that dataset by calculating exports using the records of the importing country, when available, assuming that data on imports is more accurate than data from exporters. This is likely, as imports are more tightly controlled in order to enforce safety standards and collect customs fees. In addition, the authors correct the UN data for flows to and from the United States, Hong Kong, and China. We focus only on export data, and do not disaggregate by country of destination. More information on this dataset can be found in NBER Working Paper #11040, and the dataset itself is available at www.nber.org/data. and http://cid.econ.ucdavis.edu/data/undata/undata.html

Basic Statistics

We define the product space as the set of all proximity measures. We now concentrate on the product space built using trade flow data from 1998-2000, which consists of a 1006x1006 matrix whose entries are the proximity between products. Each row and column of this matrix represents a particular product and each off-diagonal element represents the proximity between a pair of products. Figure S1A shows the proximity matrix where columns are sorted using its sitc4 code name. Figure S1B shows the same matrix sorted using an average linkage clustering algorithm, revealing its modular structure and the many empty rows and columns which belong to untraded products. In fact only 775 products conforms the actual product space.

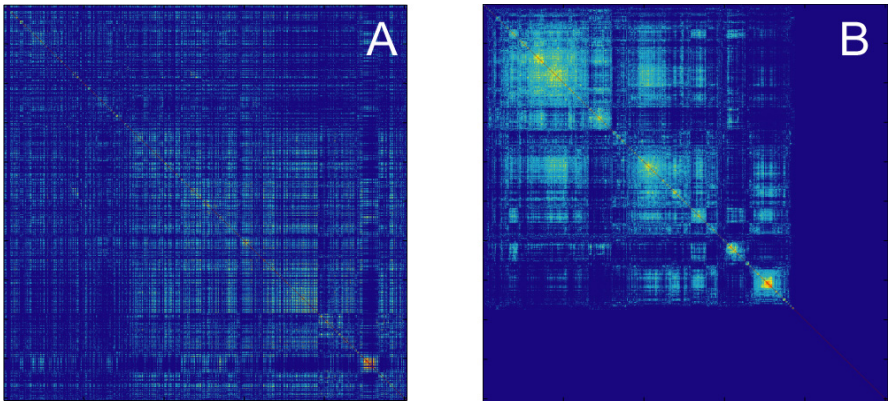

Figure S1. The product space matrix representation. A. The product space matrix sorted in incresing order of the Sitc4 numerical code. B. The product space hierarchically clustured shows a modular structure and also revelas that only 775 products actually belong to the space.

Proximity values follow a broad, approximately log-normal distribution. Figure S2A shows the number of links below a certain threshold. Figure S2B. shows the frequency distribution of proximities. This broad, heterogeneous distribution evidences that the product space has a few strong links and many marginal links, which are not significant and represent the background of the proximity measure.

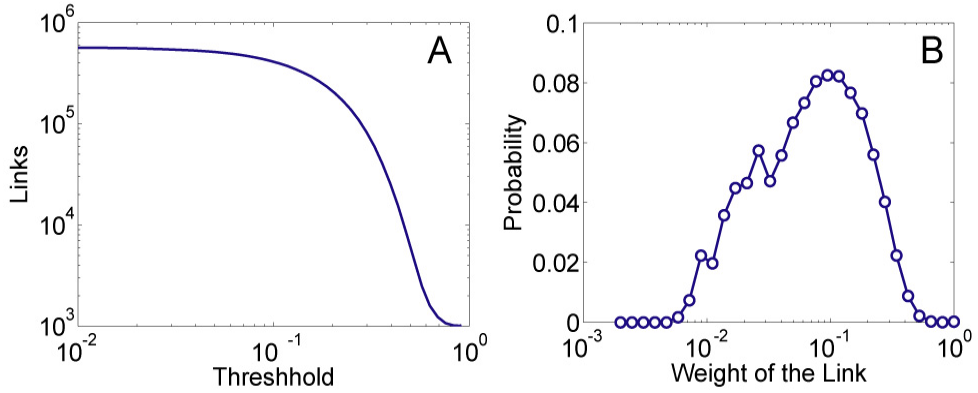

Figure S2. Distribution of Proximity Values. A. Cumulative Distribution of Proximity Values. B. Density Distribution of Proximity Values.

At a value of 0.5 the proximity matrix has a giant connected component. Figure S3 shows a rough network representation of the proximity matrix in which only links above 0.5 have been considered. This visualization is not the most suitable and shows that a simple threshold criterion does not reveal much of the structure. We invite you to look at our more sophisticated visualization technique in the next section.

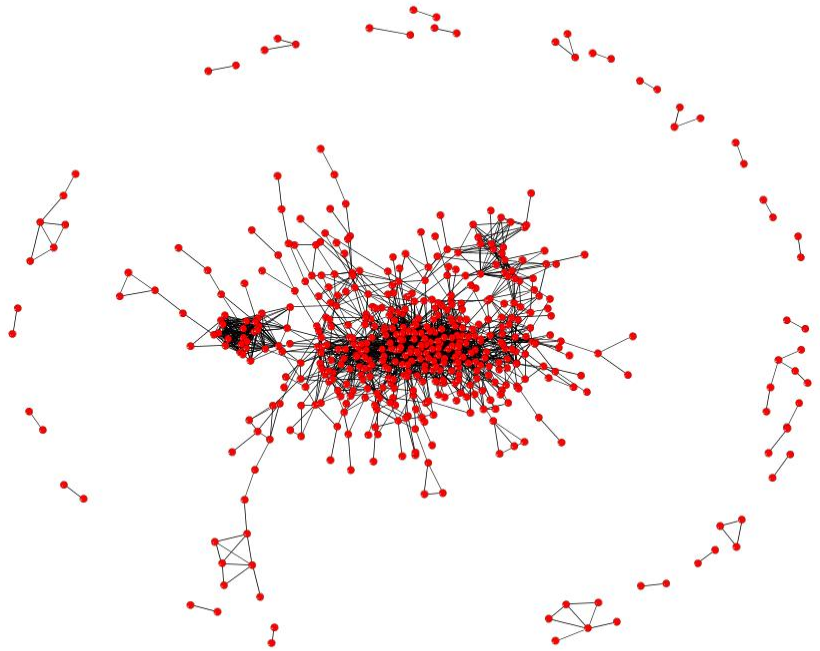

Figure S3. Network representation of the proximity matrix that considers only the proximity values above 0.5.

Network Representation of the Product Space

We generated a network representation of the proximity matrix to help us develop intuition about its structure as well as to visualize and study the dynamics of countries on it. The matrix representing the product space has many small values which represent weak connections between products. That is why a network representation becomes an adequate way to layout the products, giving us a quick visual way to show the relevant links and to determine were countries are located and where they could be headed.

Maximum Spanning Tree (MST)

To include all products in our network we generated a "skeleton" for it: the Maximum Spanning Tree (MST). This is nothing more but the tree containing a sum of weights which is maximal. In other words, it is the set of N-1 links (N being the number of nodes) that connect all nodes in the network and maximizes the sum of the proximities in it.

We generated the MST by considering the strongest non-diagonal value of the proximity matrix and then considered the strongest link connected to that dyad. We then picked up the strongest link connecting a new node to our triad and continued adding links until all the nodes on the network were considered (Figure S4).

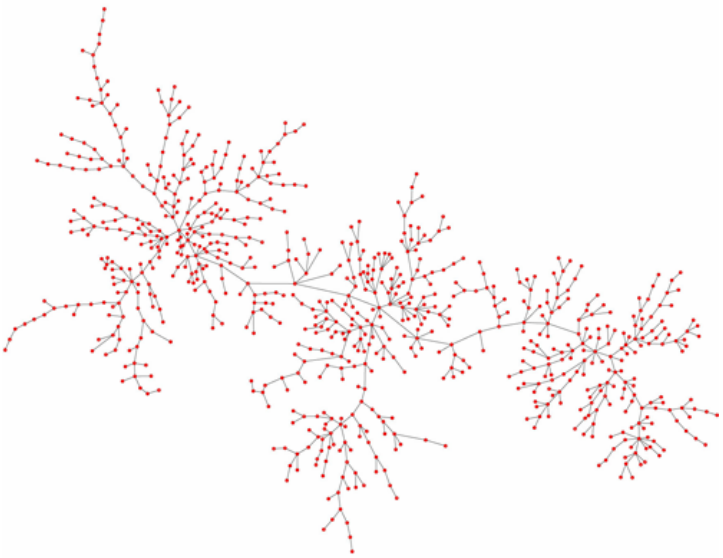

Figure S4. Earliest version of the MST representing the "skeleton" of the product space.

We also wanted to consider the strongest links which are not necessarily in the MST. We did this by considering the MST plus all the links above a certain threshold. A suitable visualization was obtained by keeping all links with a proximity value of 0.55 or larger (Fig. S5). This resulted in a network with 775 nodes and 1525 links. Lower proximity values gave rise to crowded network representations while higher values resulted in sparse networks. As a rule of thumb, a good network visualization can be achieved with an average degree equals to 4. This is when the number of links is twice the one of nodes, which is the case for the 0.55 threshold.

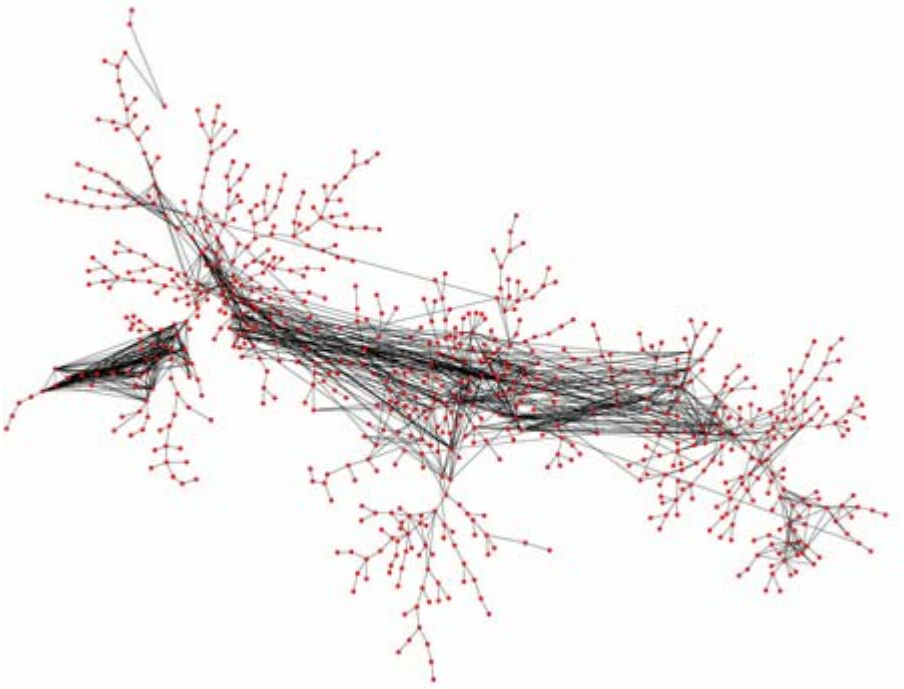

Figure S5. Representation of the product space based on the MST plus all links with a proximity above 0.55.

Network Layout

Good network visualization requires an appropriate layout. This is why we lay out the network using a force spring algorithm. Here nodes are represented as equally charged particles and links are assumed to be springs. The layout is determined by the relaxed positions.

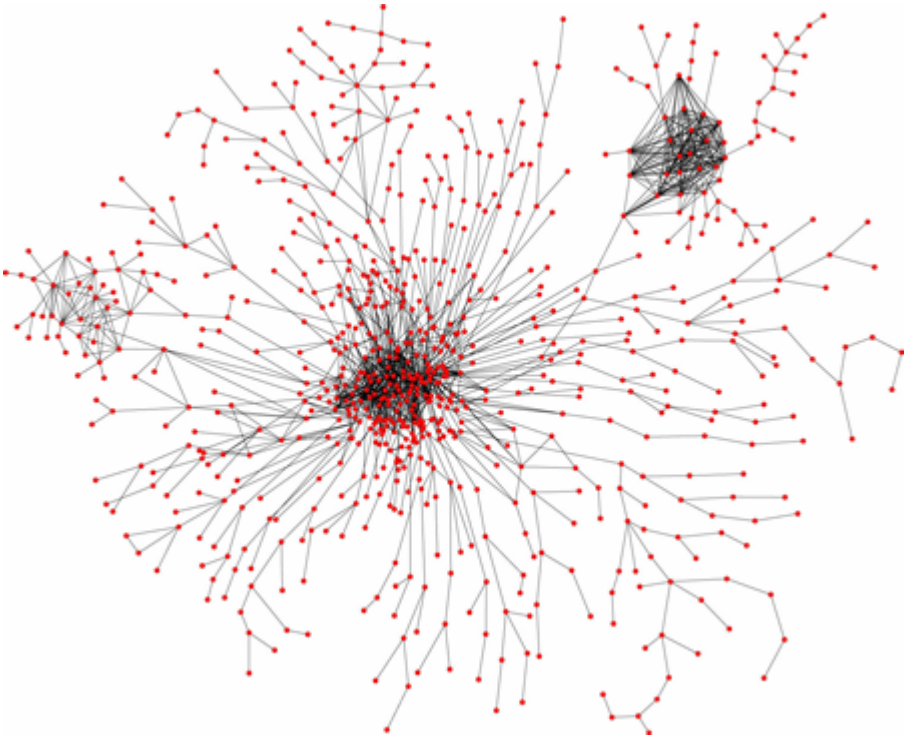

Figure S6. Network representation of the product space. Layout uses a force spring algorithm.

The force spring layout is not the ultimate solution, but it brings us close to a good one. That is why we retouched the layout manually to avoid overlapping links and untangle dense clusters.

Node Sizes and Colors

An advantage of using a network representation is that we can simultaneously look at the structure of the space and other covariates. In our case we painted the network using the product classifications performed by Leamer[1], and made the size of the nodes proportional to the money moved by that particular industry or World Trade. To give a sense of the proximity of the links involved in our network representation we color coded them by using dark red and blue for strong links; and yellow and light blue for weaker ones.

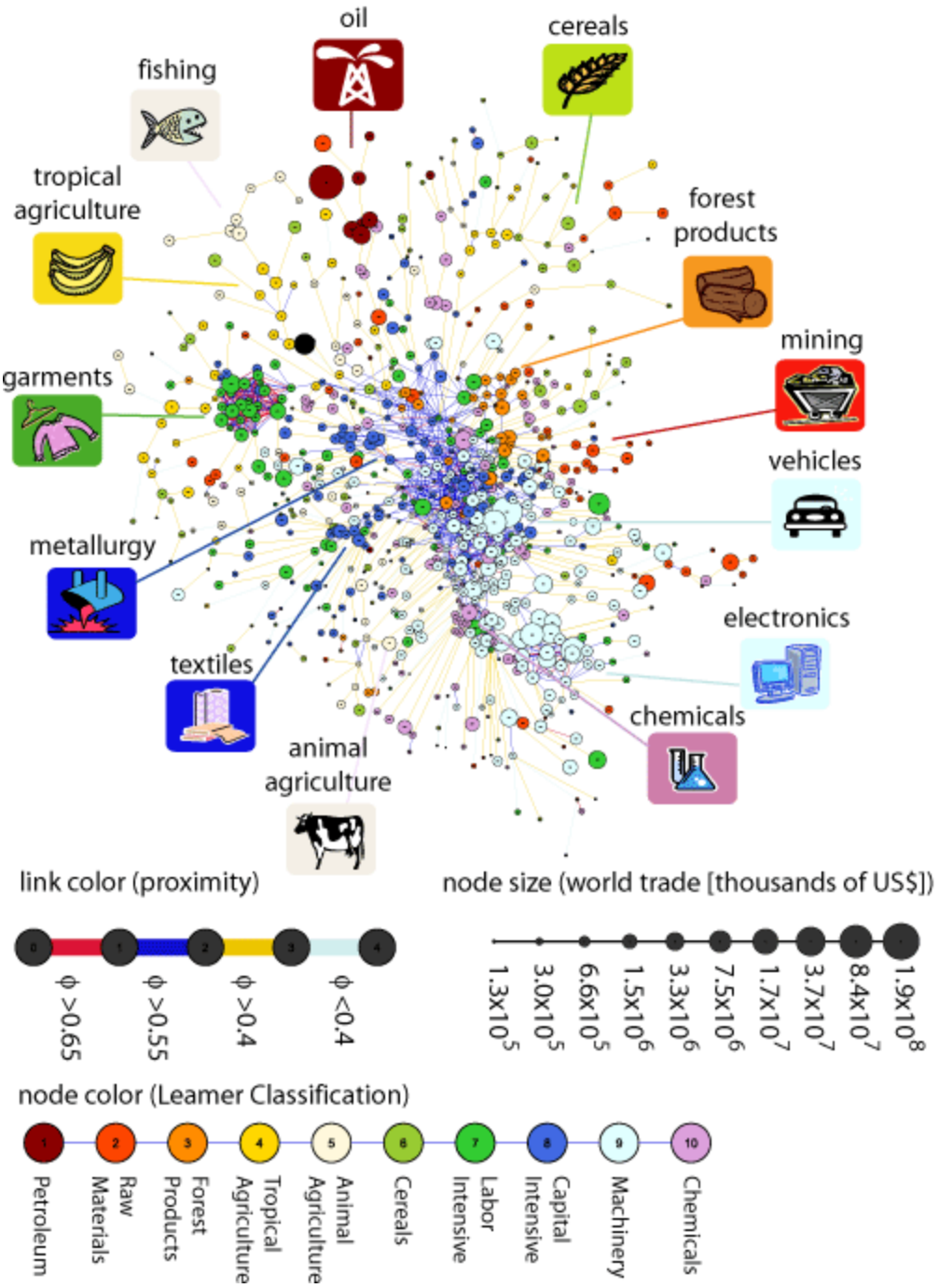

Figure S7. Final version of the product space in which node size represents its world trade, node color shows its classification as proposed by Leamer and link color indicates a range in the proximity values.

(1) E. Leamer , *Sources of Comparative Advantage: Theory and Evidence* (MIT Press, Cambridge MA, 1984).

Product Space Properties

Using a network representation for the products space we can not only see which products are close to each other and the groups they form, but also their classifications and values. However, the network representation is nothing more than a powerful visualization technique and we still need to study the space properties using the entire proximity matrix complemented.

The Product Space Can Classify Products

The first property we study is the ability of the product space to classify goods into different classes. We compare our network representation with the clusters introduced by Leamer, as it is shown in figure 1, by using a different color for each product class. We see that the product space is not colored at random. Products in the same classes lie close to each other and tend to form clusters.

Although the classification performed by Leamer was done used a different methodology, the agreement between it and the structure of the product space is striking. Beyond the intuitive proof of Figure S7 we can tests the strength of these correlations by taking the average proximity between and within the products belonging to one of the clusters defined by Leamer (table S1).

1998	Petroleum	Raw Materials	Forest Products	Tropical Agriculture	Animal Products	Cereals	Labor Intensive	Capital Intensive	Machinery	Chemical	Within / Between
Petroleum	**0.28**	0.15	0.16	0.16	0.15	0.13	0.14	0.16	0.11	0.15	0.89
Raw Materials		**0.17**	0.15	0.13	0.14	0.13	0.13	0.15	0.12	0.14	0.24
Forest Products			**0.26**	0.16	0.17	0.13	0.18	0.21	0.17	0.17	0.57
Agriculture				**0.21**	0.18	0.15	0.17	0.17	0.11	0.14	0.57
Animal Products					**0.20**	0.15	0.16	0.17	0.12	0.16	0.29
Cereals						**0.14**	0.13	0.15	0.11	0.14	0.04
Labor Intensive							**0.22**	0.22	0.17	0.16	0.38
Capital Intensive								**0.26**	0.19	0.20	0.43
Machinery									**0.24**	0.21	0.62
Chemical										**0.24**	0.46

Table S1. Average strength of the links between and within products as classified by Leamer.

Table S1 shows that the average proximity of products belonging to the same cluster is always higher than the proximity for products belonging to different clusters. But not all clusters have the same size, thus we look at the distribution of proximities for all links connecting products with the same or different Leamer classifications. Figure S8 shows the distribution of proximity for links connecting nodes with the same Leamer classification (blue) and for links connecting nodes annotated differently. It is clear from

the figure that nodes with the same classification are connected by links with higher proximity values, and because of the large number of links present in the system (L>200'000), the difference between these two distributions is highly significant (log(P-value)<-300 ANOVA)

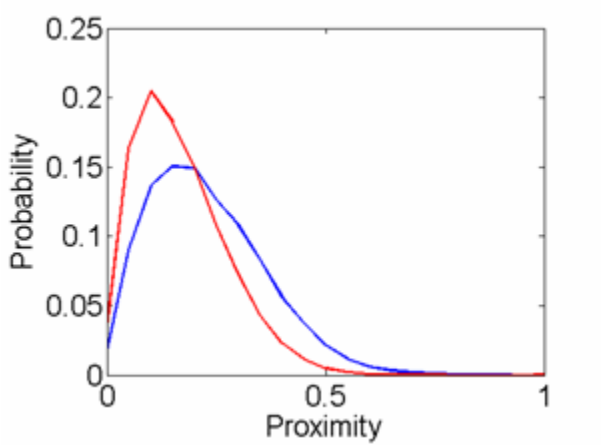

Figure S8. Distribution of proximity for links connecting products with the same Leamer classification (blue) and with a different one (red).

Correlations Between the Position and Value of the Goods.

All products have a value, which in this work we consider as the average income per-capita associated with that good or PRODY. It follows to ask: Are rich goods located in particular parts of the product space? By looking at its network representation and setting the size of the nodes proportional to the PRODY of a product (figure S9), we see that the largest nodes are located either in the center or the down most portion of the network. At a first glance, we can say that there is a rich region of the product space, composed by machinery, electronics and chemicals, and a poor, peripheral region, made of some agricultural and labor intensive goods.

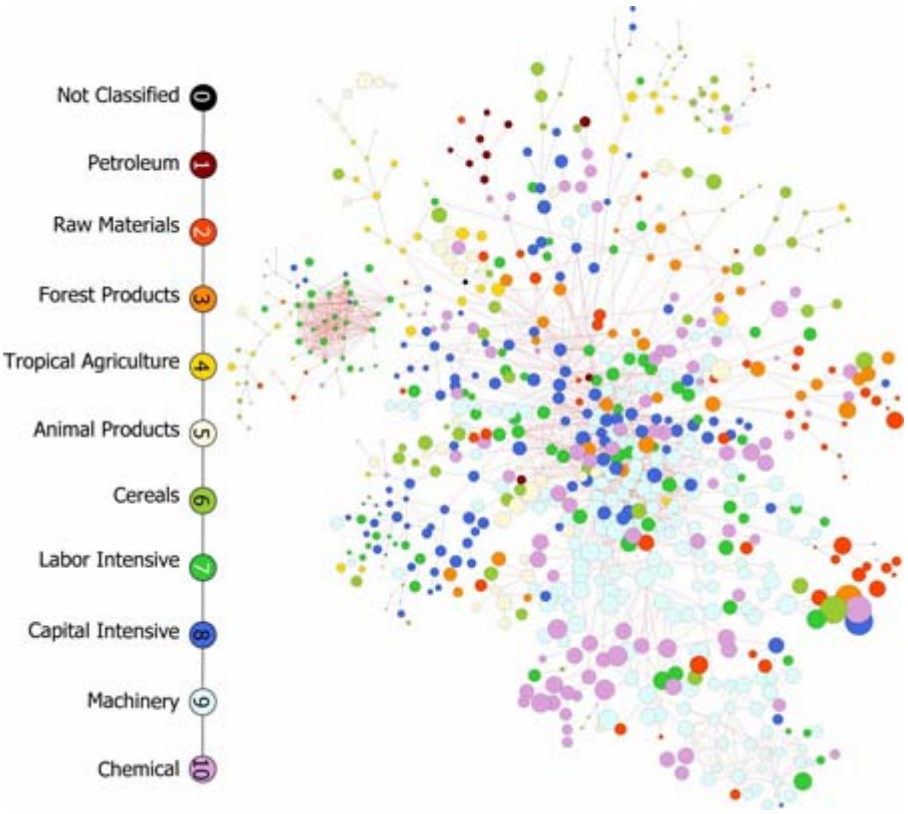

Figure S9. Network representation of the product space in which node sizes are proportional to PRODY.

We con look beyond the actual value of products and study the value of goods as a function of their distance between them. Basically we ask: Is this particular product at the top or at the bottom of the PRODY sophistication scale? To answer this we study the average PRODY of products at a given distance of a particular node. We define distance as *-log(Proximity).* Figure S10 shows six examples of products, three of them at the bottom of the sophistication scale (Footwear, Cotton Undergarments and Coats and Jackets) which belong to the labor intensive cluster and thus products far from them are richer or more attractive. On the other hand, chemicals such as organo sulphur compounds, phenols and cyclic alcohols appear at the top of the sophistication scale and see all other products as less sophisticated.

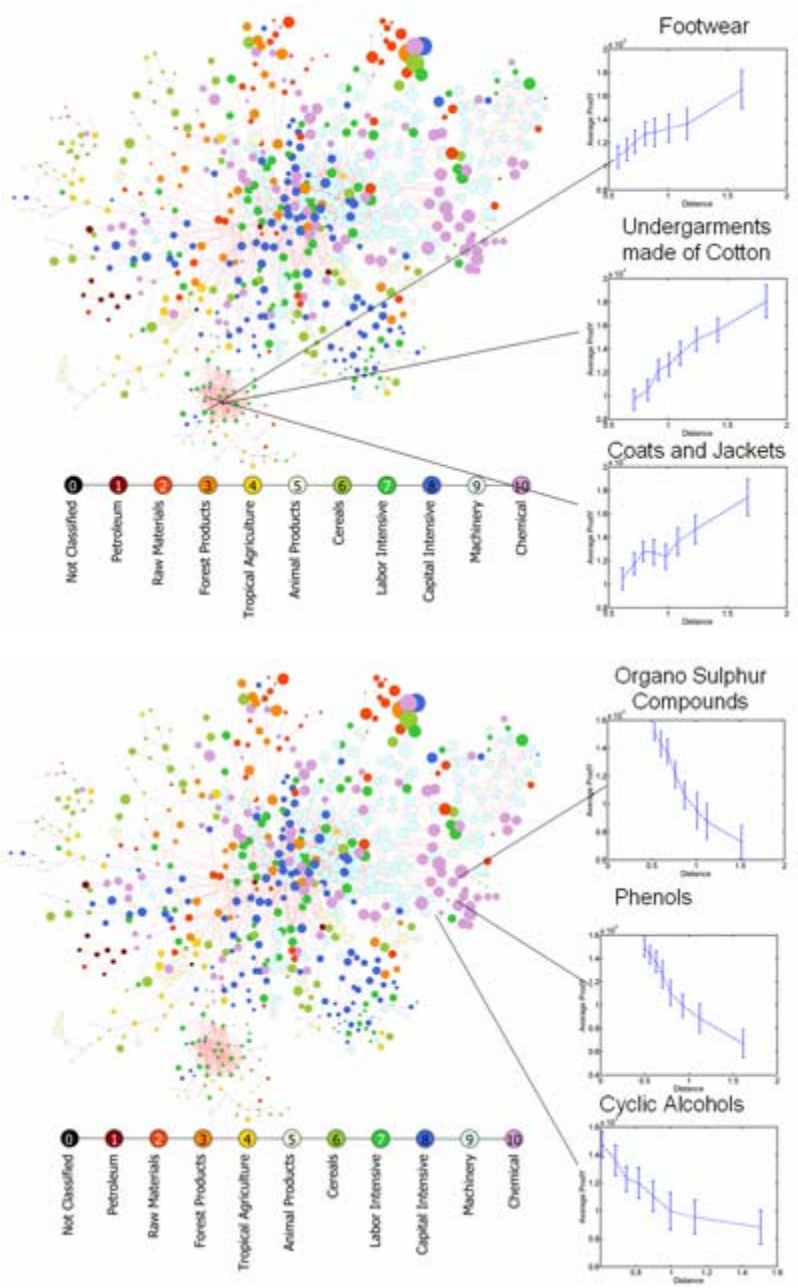

Figure S10. Prody as a function of distance for six different products in the space. Plots were calculated using the full proximity matrix.

We performed the same analysis for each product class and found that there are products at the top of the scale, at the bottom and in local maxima (Figure S11). If the structural transformation only moves countries to more sophisticated goods, a local maximum would trap countries. Examples of these are cereals and animal agriculture products which are goods located in the periphery of the product space but have a relatively large PRODY compared to their neighbors.

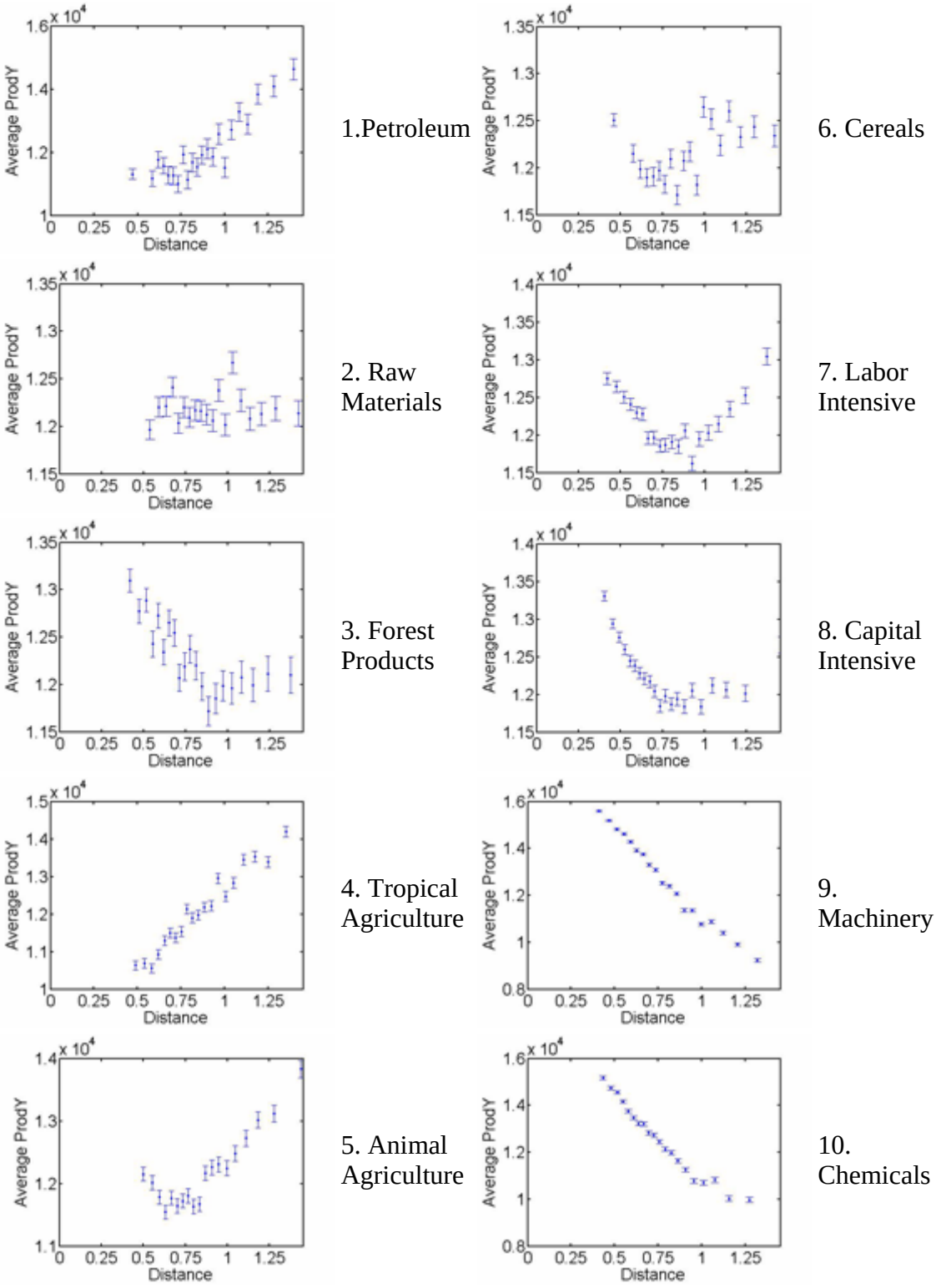

Figure S11. Average PRODY as a function of the distance for products with a given Leamer Annotation.

Changes in Time

How fast does the product space changes in time? We can take a simple look at these by calculating the Pearson's Correlation Coefficient (PCC) between the matrices representing the product space in 1985, 1990 and 1998. Table S2 shows that the structure of the product space appears to be stable and that although links do change in time, after 10 or 13 years strong links remain strong and weak links remain weak. Thus products that are close tend to remain close and the ones that are far tend to stay far. The correlation was calculated over each pair of corresponding proximities between different time periods. Proximity values equal to zero were excluded from the calculation.

PCC	**1985**	**1990**	**1998**
1985	**1**	**0.702**	**0.696**
1990		**1**	**0.616**
1998			**1**

Table S2. Pearson's Correlation Coefficient between the product spaces generated with data from 1985, 1990 and 1998.

Empirical Diffusion

Looking at pictures

Once the product space has been created and visualized, it becomes relatively easy to visualize the structural transformations of countries and how they are conditioned by the space itself. This study begins by visualizing where countries are located at different times, an amazing visual experience able to summarize the productive structure of a nation that preserves a significant level of detail. Figure S12 shows the products for which Malaysia has developed RCA with black squares. Figure S13 shows the same for Colombia. The versions presented here come from the beginning of our research, in which the layout was a slightly different, but still conserves the same color code and overall position. Node sizes are still proportional to world trade. We can appreciate that Malaysia had an impressive spread over the electronics and forest products cluster while during these same time period Colombia was able to spread through the garments sector.

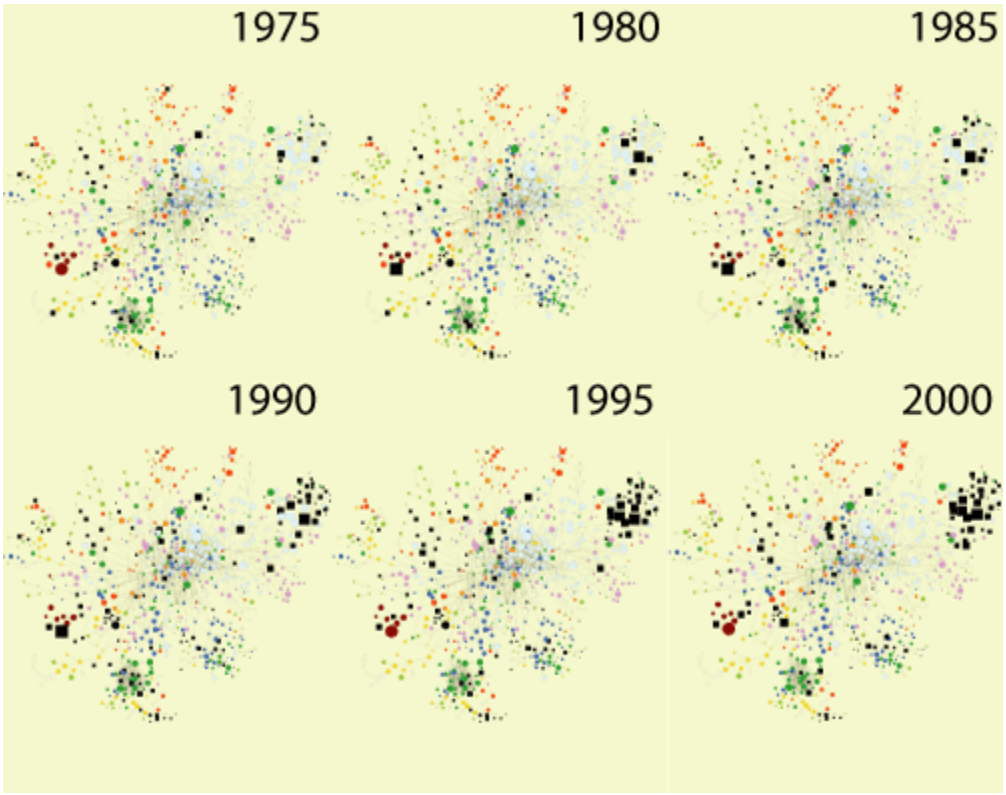

Figure S12. Evolution of Malaysia (High Resolution Image) The products for which Malaysia has developed RCA are shown with black squares. Vector Image .ai.

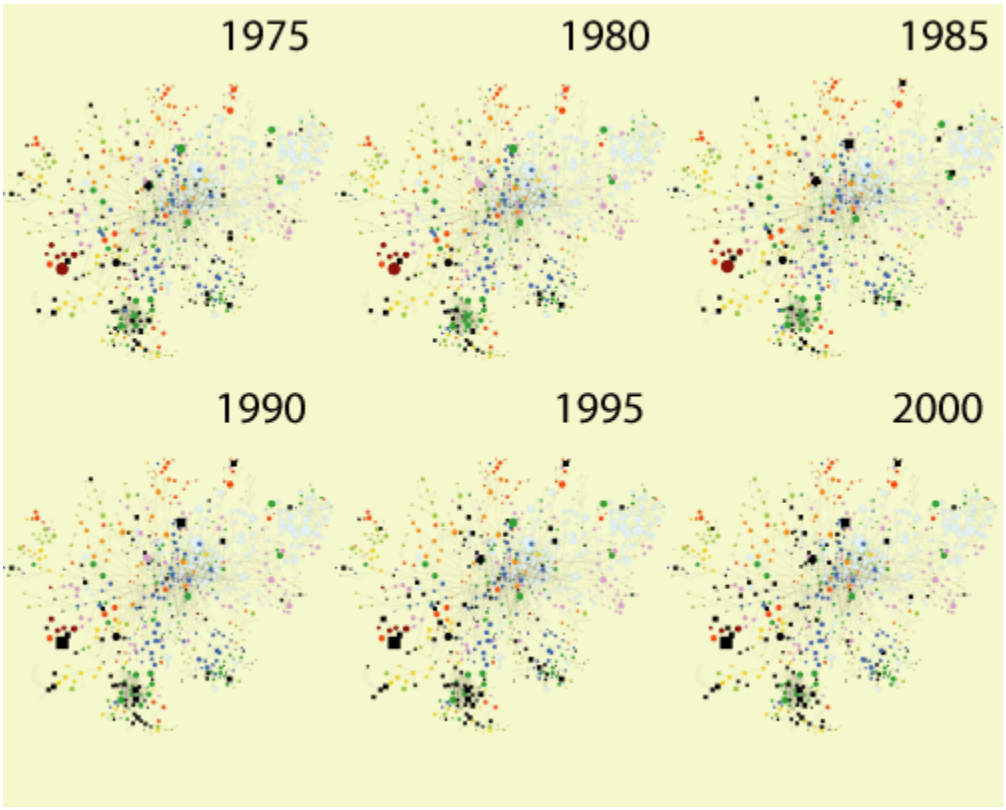

Figure S13. Evolution of Colombia (High Resolution Image) The products for which Colombia has developed RCA are shown with black squares. Vector Image. ai.

Additional data supporting this paper can be found at www.nd.edu/~networks/productspace. The pdf files show in black outlined squares all products in which a country has an RCA>1. The number on the top of each node indicates the sitc-4 classification associated to each node. File names are given by the countries ISO code and the year which they represent. All PDF files are currently available at www.nd.edu/~networks/productspace/Data/CountryMaps.rar. There are also gml files available for more time periods. To view gml files you need to download and install cytoscape, which is free and available at (http://www.cytoscape.org/). To view a file, once in cytoscape, go to File/Import/Network (multiple file types) and choose the .gml file you want to view. File names are given by ISO country codes and the year which they represent. Gml files are currently available at www.nd.edu/~networks/productspace/Data/netsgml.rar.

Possible transitions between products

The visual examples presented above help develop our intuition and explain in a simple way how countries undergo structural transformations. They show that there is a tendency for countries to develop RCA close to products for which RCA was already developed, but are not a proof of this. For simplicity, we call a product for which a country has developed RCA, an *occupied product* (O), and one for which it has not an *unoccupied product* (U). When we compare two time points there are 4 possible

transitions (U->U,U->O,O->U,O->O), and in our case, we are concerned with the second one which takes unoccupied products to occupied ones. Additionally, we call a product undergoing this particular transition: *transition products*. We now ask: are transition products closer to occupied products than to unoccupied ones? If this is significantly the case, it would be evidence supporting that countries perform structural transformations by *jumping* from occupied products to nearby ones.

To proof this we need to define some quantities. First we define density as the weighted fraction of the space which appears to be occupied from the point of view of a product in a particular country. Mathematically density ω can be written as:

$$\omega_j = \sum_i x_i \phi_{ij} / \sum_i \phi_{ij} \quad ,$$

where ϕ_{ij} is the proximity between the *i'th* and *j'th* product and x_i is 1 when the *i'th* product is occupied and zero otherwise. To measure this quantity empirically we consider as undeveloped products all those that had an RCA<0.5 on 1990. Starting from this definition, transition products are the ones that had an RCA >1 on 1995 and undeveloped products are the ones that remain with an RCA<0.5 on 1995. The ones that had an RCA between 0.5 and 1 in 1995 were regarded as inconclusive.

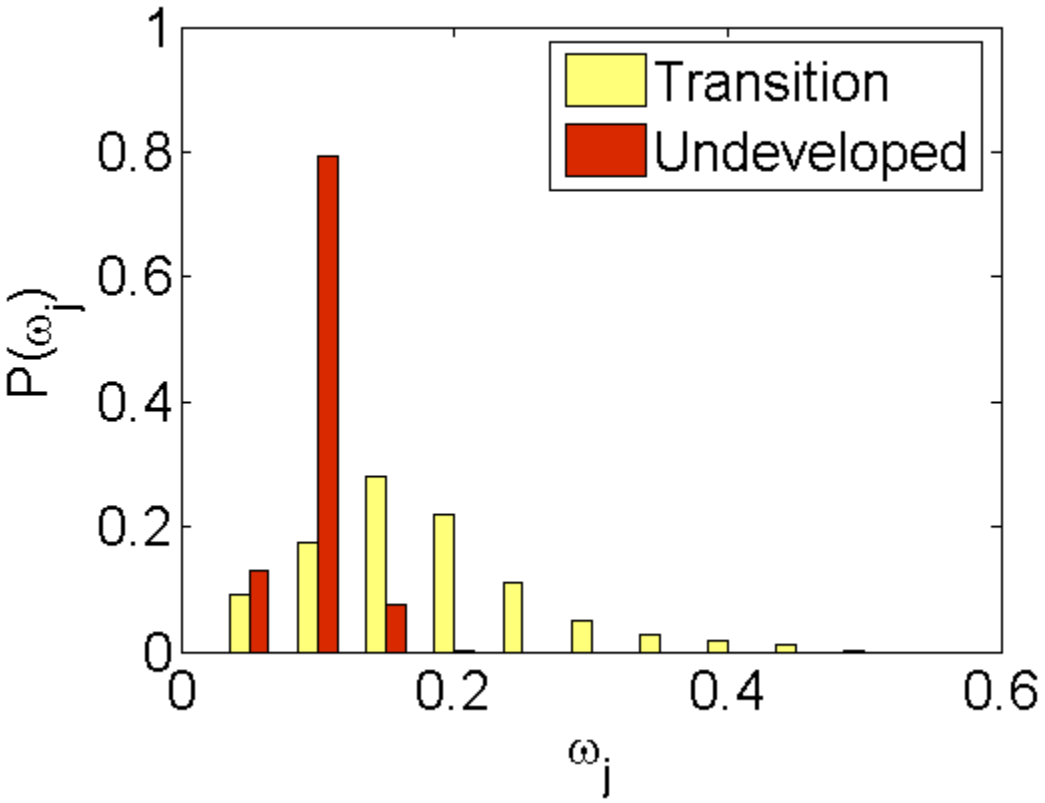

Figure S14. Density distribution for products that underwent a transition and those remaining undeveloped.

Figure S14 shows that the density distribution for transition products takes significantly larger values than for products that remained undeveloped, suggesting that density predicts a transition. To further characterize this we can take the ratio between the average density of products on countries were they underwent a transition and compare it to the average density for countries were they remain undeveloped. We call this the discovery factor (*H*) which can be written as

$$H_j = \frac{\sum_{k=1}^{T} \omega_j^k \Big/ T}{\sum_{k=T+1}^{N} \omega_j^k \Big/ (N-T)}$$

where the top summation goes over the *T* countries where the *j'th* product underwent a transition and the bottom one over the *N*-*T* countries were the product remain undeveloped. Figure S15 shows that in fact for more ~80% of the goods this ratio is larger than one, illustrating again that density tends to be higher for transition products.

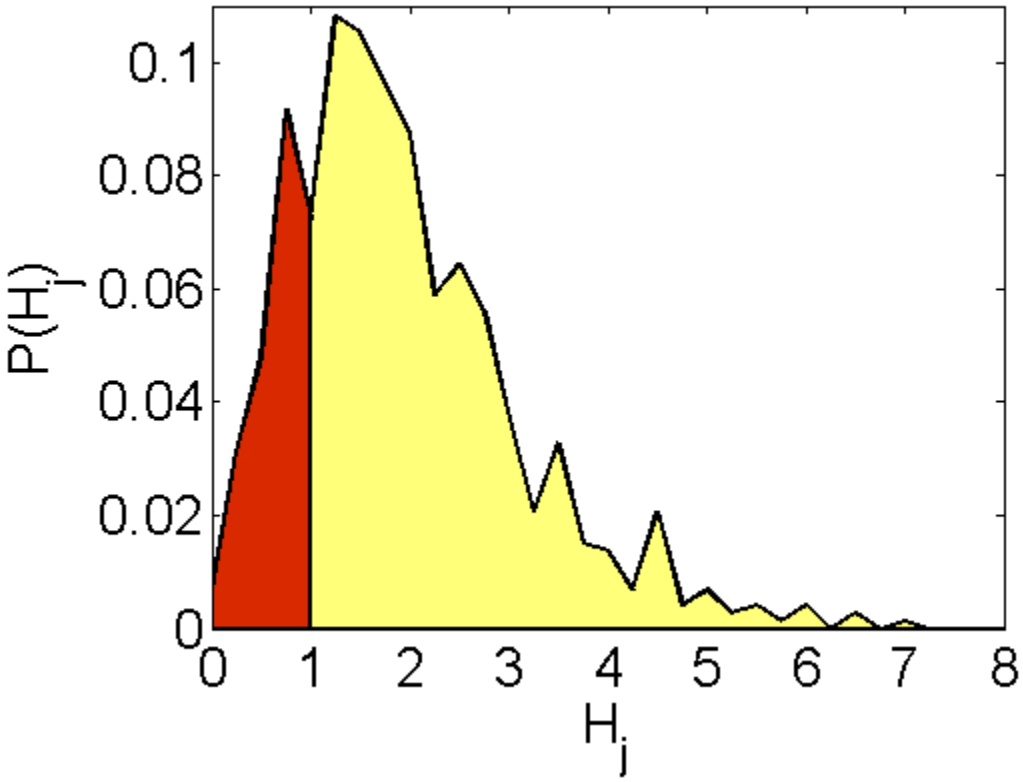

Figure S15. Distribution of the discovery factor H.

Yet another way to show this is to consider the probability for a product to develop given that the closest developed product is at proximity ϕ. Figure S16 shows that this is a monotonically increasing function of ϕ. In fact further inspection shows that P has a quadratic dependence on f. Thus the chances for a country to develop a product increase enormously when that product is close to an already developed one.

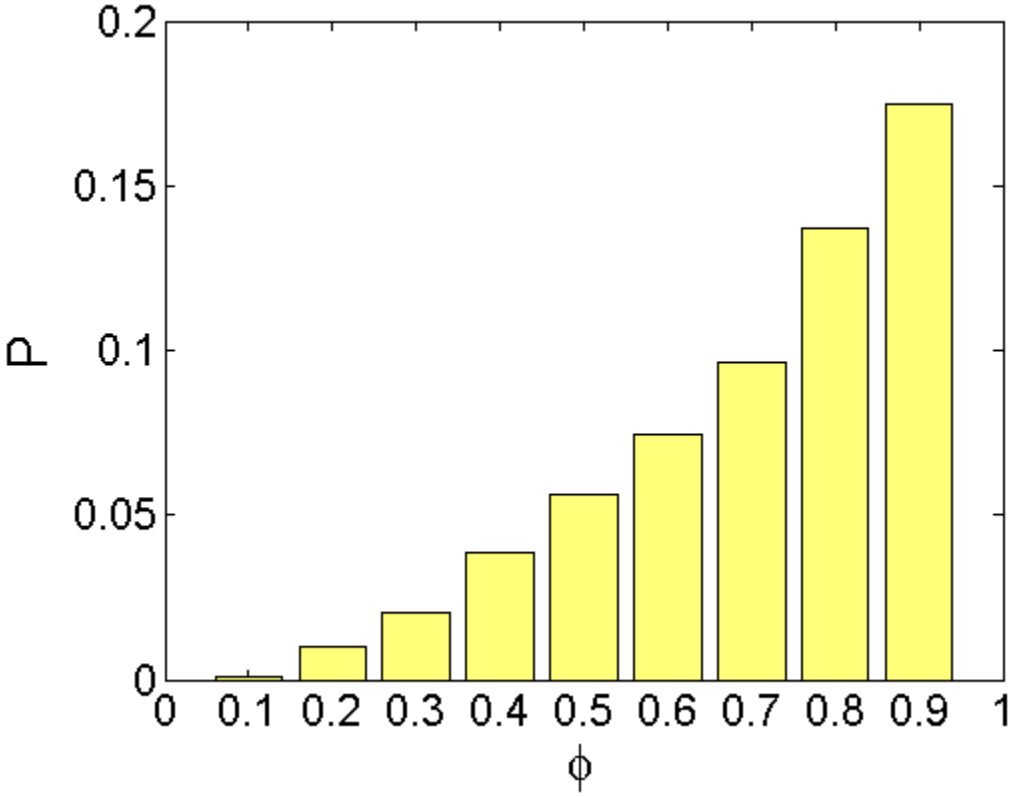

Figure S16. Probability of undergoing a transition given that the closest occupied product is at proximity ϕ.

Simulated Diffusion

One diffusion step

Empirically, we showed using examples and statistics that products in which countries develop RCA tend to lie close to other products for which these countries have already developed RCA.

Using these we try to anticipate how a country will diffuse across the product space. As an example, we show figure S17, in which we highlighted with black squares all products at a given proximity of the ones already developed by Chile and Korea. We refer to this example as one diffusion step.

In this case we tuned the proximity of the jump and show that for high proximities the set of options available is small while for low proximities is large, however different.

The available options are strongly conditioned by current exports. Korea is a country that has developed RCA in several branches of machinery and therefore can diffuse from the center of the space. At proximity of 0.5 its options include the entire core of the network plus the entire electronics and garments clusters, among other things. Chile diffuses from the periphery and to achieve a similar set of options needs to diffuse as far as proximities of 0.3.

In summary we find that the set of options available for a country are strongly conditioned by its position in the product space and its ability to diffuse into products up to given proximities.

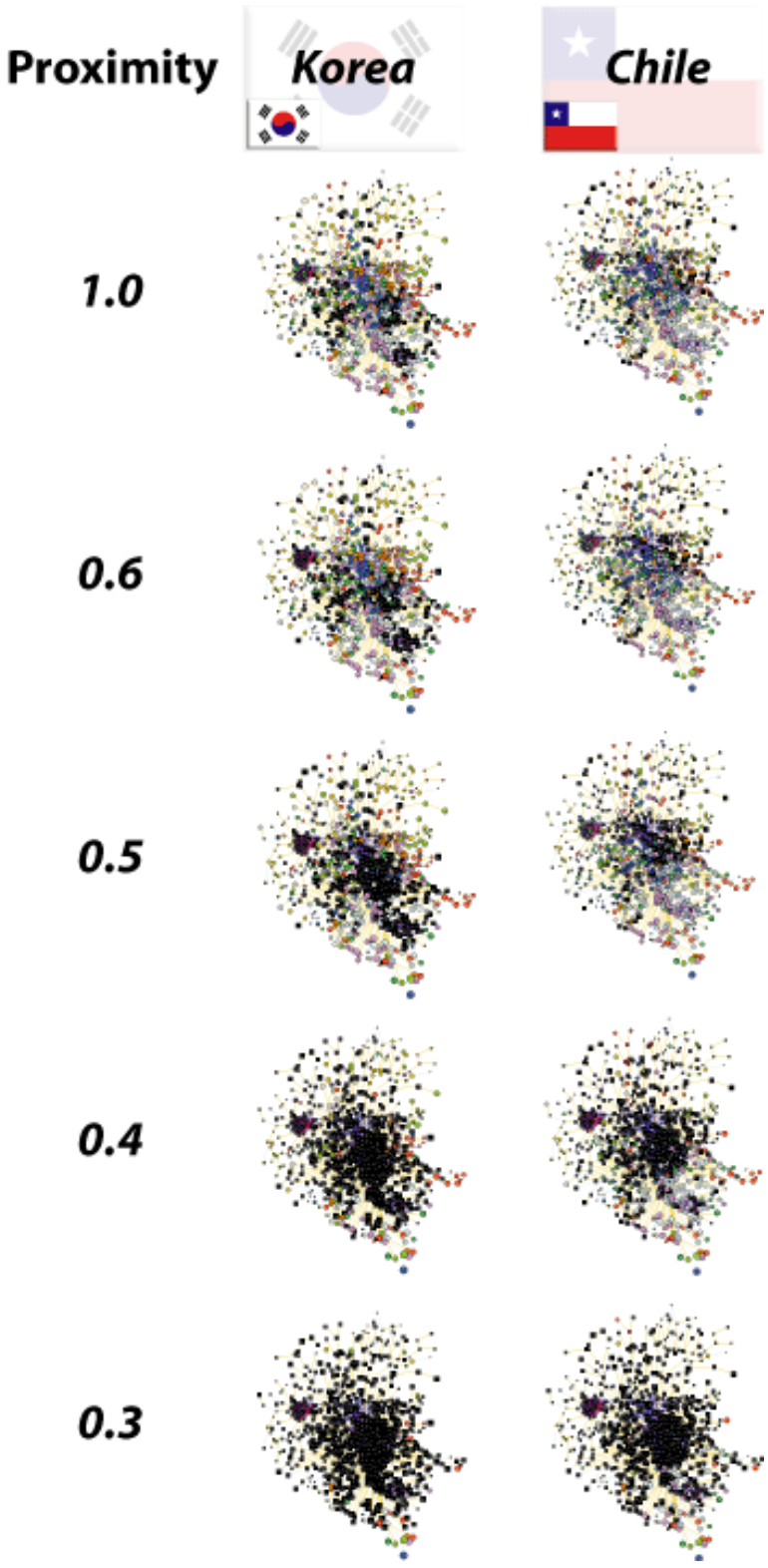

Figure S17. One step diffusion process for Korea and Chile. The black squares denote all products closer than a given proximity considering their exports baskets in the year 2000.

Iterated diffusion

We can refine the diffusion process presented above by choosing a particular proximity and iterate the one step diffusion process. This represents a set of products potentially available to countries after diffusing to close products iteratively. At this point we ask ourselves: Is there a critical value of proximity at which countries will be able to diffuse across the product space? To explore this question we simulate a diffusion process in which a country "jumps" to all goods reachable from its current export basket, such that the proximity to them is larger or equal than a given value. Figure S18 illustrates through a color code the products available to Chile and Korea after diffusing iteratively at different proximities for 4 time steps. We observe that at relatively low proximities (ϕ = 0.55) both countries are able to diffuse, however Chile does so much slower and reaches the core in the second and third rounds, compared to Korea which does so on the first and second. At larger proximities the diffusion process halts. At ϕ = 0.65 Chile is unable to diffuse at all, while Korea slowly does so close to the core of the product space.

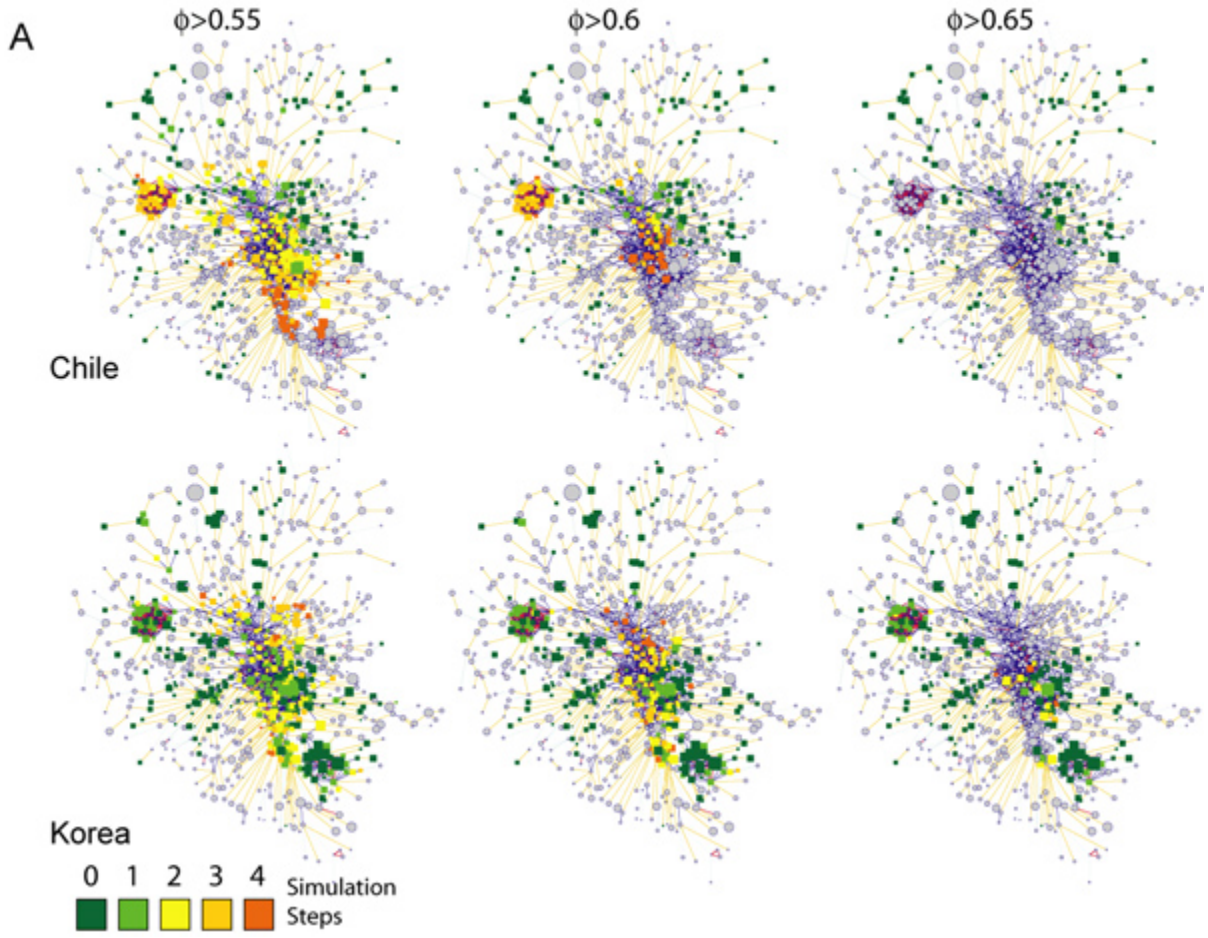

Figure S18. Iterated diffusion process for Chile and Korea.

Economic Convergence

We characterize the value of a certain configuration by considering the value of its top products. We can assign value to a good by following the work of Hausmann, Hwang and Rodrick in which the value or sophistication of a good is equal to the average GDP per capita associated with that good. This quantity is called PRODY and in our particular example we consider the average PRODY of the top N products of a countries export basket after M diffusion steps with proximity ϕ. We denote this quantity by $< PRODY >^{N}_{M\phi}$.

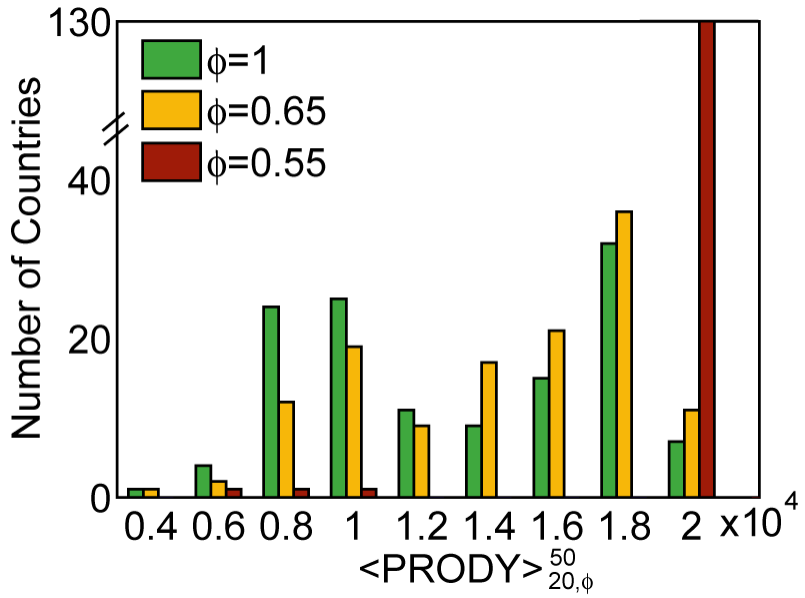

Figure S19. Distribution for the average PRODY of the top 50 products reached after 20 diffusion steps at three different proximities.

Figure 3 shows that the original distribution of $< PRODY >^{N}_{M\phi}$ is bimodal. Indicating a world in which countries are divided into those producing sophisticated goods and unsophisticated ones. If we allow countries to diffuse in this space to acquire only goods that are really close by (ϕ=0.65). This distribution remains practically unchanged evidencing the structural constrains imposed by the product space. Whereas, if we allow countries to diffuse into products at relatively large proximities (ϕ=0.55) we find that after a large number of rounds most countries are able to reach the most attractive parts of the space, except for a few of them that remain stuck in the lowest bracket of this distribution.