Capabilities’ substitutability and the “S” curve of export diversity

HONGMEI LEI and JIANG ZHANG

School of Systems Science, Beijing Normal University - Beijing 100875, China

received 19 April 2013; accepted 10 March 2014
published online 25 March 2014

PACS 89.65.Gh – Economics; econophysics, financial markets, business and management
PACS 89.75.-k – Complex systems

Abstract – Product diversity, which is highly important in economic systems, has been highlighted by recent studies on international trade. We found an empirical pattern, designated as “S-shaped curve”, that models the relationship between economic size (logarithmic GDP) and export diversity (the number of varieties of export products) on the detailed international trade data. As the economic size of a country begins to increase, its export diversity initially increases in an exponential manner, but overtime, this diversity growth slows and eventually reaches an upper limit. The interdependence between size and diversity takes the shape of an S-shaped curve that can be fitted by a logistic equation. To explain this phenomenon, we introduce a parameter called “substitutability” into the list of capabilities or factors of products in the tri-partite network model (i.e., the country-capability-product model) of Hidalgo et al. As we observe, when the substitutability is zero, the model returns to Hidalgo’s original model but failed to reproduce the S-shaped curve. However, in a plot of data, the data increasingly resembles an S-shaped curve as the substitutability expands. Therefore, the diversity ceiling effect can be explained by the substitutability of different capabilities.

Copyright © EPLA, 2014

Introduction. – Recent research on international trade has highlighted the diversity phenomenon, which is generally ignored by conventional economic studies. However, both the amount and the types of goods that a country produces affect the economic growth [1–4]. Important new facts have been uncovered by analyzing large amounts of high-quality data pertaining to international trade. For example, there is a negative relationship between the diversification of countries and the ubiquity of products [5–7]. To account for this phenomenon, Hidalgo et al. constructed a tri-partite network model and attempted to claim that the capabilities or non-tradable factors a country possesses are the “building blocks” of its economy and determine its diversification. To their credit, the negative correlation between diversity and ubiquity can be reproduced by their model.

However, Hidalgo et al. did not explain what ingredients determine the non-tradable capability of a country: although they tried to link the economic size or richness of a country with the number of these capabilities it has in paper [7], they did not give any empirical evidence because these capabilities are non-measureable. In contrast, because an economy’s size, as measured by its GDP, may be the most important datum in modern economics, it must have a correlation with a country’s diversification degree [8]. It is obvious that countries with large GDP always produce and export more diversified products and that countries with small GDP usually have more homogeneous products and markets [9–11]. This observation can be described quantitatively by an S-shaped curve that models a country’s logarithmic GDP and export diversity [1]. This interdependence between size and diversity is ubiquitous in global trade and economic systems.

To this point, the theoretical understanding of the S-shaped curve as a model of the relationship between economic size and export diversity is still deficient. Therefore, this paper tries to build a model to reproduce this size-diversity curve. Initially, we simply link the probability that a country may possess certain capabilities with its economic size. In this way, we can investigate how the economic size determines a country’s diversification. However, this interdependence between size and diversity in Hidalgo et al.’s model is exponential, as they have noted; thus, once a country’s economic size exceeds a certain threshold, the country will receive increasing returns. As a result, the type of products the country’s
businesses are able to export increases without any limitation; otherwise, the country’s economy cannot overcome the so-called “quiescence trap” [6,7]. However, the empirical data reveal that there is the ceiling effect of the diversity curve [1] that cannot be reproduced by the original tri-partite model. Therefore, we have introduced an important parameter into our model, namely, the substitutability $s$ between different capabilities; this parameter’s purpose is to relax the overly strict condition on the number of capabilities that a product requires. Substitutability, as an important concept in economics, has a long history of study [12-14]. However, in this paper, the substitutability is just a fraction of the replaceable factors among a set of capabilities, but not defined by the production function as the economic literatures usually do. Interestingly, in paper [7], Hidalgo et al. mentioned the idea of substitutability between factors but they did not develop it. In this paper, we report that our model that includes the substitutability $s$ can reproduce the S-shaped curve of economic diversity.

This paper is organized as follows: in the next section, we briefly introduce the S-shaped relationship between the export diversity and the economic size of countries into our model to simulate the empirical relationship between these factors and to achieve both accurate and approximate analytic solutions. In the third section, we show the simulation and analytic results (both exactly and approximately), which can resemble the empirical “S” curve. Furthermore, we discuss how the key parameters in our model affect the shapes of fitting curves.

Method. –

The S-shaped relationship. The S-shaped relationship between logarithmic GDP and export diversity can be derived from the empirical data we have collected. The world GDP statistics are from the World Bank’s web-site (www.worldbank.org) and the export diversity data are from the NBER-UN world trade database (www.nber.org/data). In the latter data set, the detailed bilateral trade flows of approximately 150 countries and 800 types of products (according to the SITC4 classification standard) during 1962-2000 are included. In this paper, we only show the “S” curve in 1995. A more detailed discussion of empirical “S” curves for other years can be gleaned from previous work [1]. The COMTRADE international trade data in 2000 with 6-digits codes for products (Harmonized classification) is downloaded from http://comtrade.un.org/ as a comparison.

The empirical data shows a strong dependence between logarithmic GDP and the types of exports in fig. 1(A) ($R^2 = 0.87$). The empirical data can be fitted by a logistic function:

$$D_i = \frac{A}{1 + e^{-(X_i - s_m)}},$$

(1)

where $D_i$ stands for the number of categories of goods (each category is represented by a distinct 4-digits code) that country $i$ export and $X_i$ represents the logarithmic GDP of country $i$. Furthermore, $A$, $k$ and $s_m$ are parameters of the logistic function. The estimated values are shown in the legend of fig. 1(A).

From fig. 1, we observe that all countries can be divided into three groups: small countries that are in the lower part of the “S” curve (e.g., Liberia), intermediate countries (e.g., Iran and New Zealand) that are in the “accelerate” part of the curve and large countries (e.g., the USA) that are at the top of the curve. The small countries in the first group can export very few varieties of goods if their sizes do not exceed a certain threshold (the “quiescence trap” [7]). The second segment of the “S” curve represents an exponential increase, in which, countries have very different types of export products. However, the accelerating-growth effect stops at the uppermost part of the “S” curve due to the ceiling effect of diversity, as large countries’ diversification levels are not as high as an exponential curve would indicate. This pattern in the relationship between economic size and export diversity is very stable in all of the years of our data set (see [1]), as well as in different classification standards.

One may conjecture that the S-shaped curve is just the artifact of the classification resolution limitation. If that was true, then the “S”ness may decrease as we increase the classification resolution. From fig. 2, we observe that the “S”ness of the curve does not disappear but slightly shift to the right when the resolution increases. While, when we compare the “S” curve of 4 digits (with the SITC4 classification) with 6 digits (with the Harmonized classification), we find that a higher resolution shows a much deeper S curve and the ceiling arrives earlier than the lower resolution. Therefore, we suppose that the “S”ness will

![Fig. 1: (Color online) The S-shaped relationship between log(GDP) ($X_i$) and the export diversity $D_i$ of countries. (A) The original data and the logistic fitting for all countries in 1995. The fitting logistic function is $D_i = 749.53/(1 + \exp(-1.932(X_i - 10.625)))$, $R^2 = 0.8693$. (B) The blue scatter points are the simulation results from our tri-partite model (the parameters are $N_i = 200, P = 765, q = 0.067, s = 0.4$). Whereas the red curve is the analytic result of eq. (11) (see the second section of main text), the black curve is the logistic fitting ($D_i = 784.56/(1 + \exp(-1.921(X_i - 10.403)))$, $R^2 = 0.8695$).]
not disappear and there must be other reasons to account for that.

**Model.** It is important to consider why this interdependence between size and diversity in the international trade exists. In paper [5], Hidalgo et al. proposed a tri-partite network model to account for various facts regarding export diversity and product ubiquity. In their model, the first and third layered nodes are the countries and their products, respectively, whereas the nodes in the hidden layer between countries and products are introduced to represent non-tradable elements, factors or capabilities, e.g., management skills, raw materials, regulation, property rights, etc. Thus, countries need to have these elements locally available to produce goods.

Following Hidalgo’s model, we hope to account for the S-shaped relationship by constructing a modified model which also assumes that each product requires some non-trade capabilities, and each country can export a product only if the required capabilities are possessed by this country. However, we initially link the logarithmic GDP with the degrees of the nodes that represent countries because our purpose is to explain the relationship between economic size and economic diversity. That is, the number of links to country $c$ is proportional to its logarithmic GDP $\log(GDP_c)$. However, this mechanism cannot reproduce the “S” curve because it resembles the sub-linear power law. And an existing study has showed that many macroscopic variables of countries scale with GDP [8].

In Hidalgo et al.’s model, country $c$ can export product $p$ if and only if the paths from $c$ to $p$ include all of the hidden nodes that connect $c$ to $p$. Thus, all of the capabilities that are devoted to producing $p$ are possessed by country $i$. However, this mechanism cannot reproduce the “S” curve because it models the relationship between size and diversity, and as a result, we must replace this mechanism with a new rule we have designed.

We borrow the “substitutability” concept from economics to revise Hidalgo et al.’s model. One important parameter called the “average substitutability rate” (or “substitutability”) $s$ is introduced to represent the proportions of the total capabilities that are required to produce product $p$; these capabilities can be replaced by other available capabilities. Country $c$ can export product $p$ if and only if the paths from $c$ to $p$ cover $(1-s)\times 100\%$ of the capabilities required by $p$ (which would imply that the hidden nodes are connected to $p$). Hence, among the requisite capabilities, only $1-s$ fractions are necessary and other $s$ fractions are substitutable in average.

**Fig. 2:** (Color online) Normalized diversity-log(GDP) curve with different classification resolutions. 4-digits SITC4 and 6-digits Hamonized classifications are highlighted to be compared.

**Fig. 3:** (Color online) A concept graph of a three-layer network and the analytic solution. (A) This three-layer network is designed to depict the mapping between countries and products. The nodes in the first layer represent countries and the nodes’ sizes represent their logarithmic GDPs. The nodes in the second and third layers represent the capabilities and products, respectively. A country can produce more products when the substitutability $s$ increases. (B) Provided that country $c$ has $L$ distinct capabilities (such capabilities are colored in red), the probability that this country produces product $p$ is determined by the mapping between these capabilities and product $p$. (C) The $m$ capabilities required by product $p$ can be divided into two groups based on whether they match the capabilities that are owned by country $c$. 

---

“S” Curve of export diversity

---

68003-p3
substitutability $s$ is 0, all of the capabilities are necessary, and, as a result, we recover Hidalgo et al.’s model. However, when $s$ increases, more countries can export diverse products to the same extent as the largest countries. As a result, an S-shaped curve between logarithmic GDP and diversity is obtained.

For example (see fig. 3(A)), suppose that country C2 can only export product P2 when $s = 0$ and that C2 cannot export P1 because all of the capabilities connected to P1 (namely, A1, A2, and A3) would have to be covered by the paths from C2 to P2, yet A1 is not covered. When $s$ increases to 0.5, only 50% of the capabilities must be covered by the paths. Therefore, C2 can export P1 because more than half of the capabilities required by P2 have been covered.

In general, we consider $N$ countries whose logarithmic GDPs are $(X_1, X_2, \ldots, X_N)$. The adjacency matrix between countries and capabilities is $C_{ik}$, and its elements are randomly assigned:

\[
C_{ik} = \begin{cases} 
1, & \text{with probability } r_i, \\
0, & \text{otherwise},
\end{cases}
\]

where the subscript $k$ is iterated from 1 to $N_a$ (which is the total number of capabilities we consider). To link the number of capabilities that a country has with this country’s GDP, we assume that the probability $r_i$ is proportional to the value of $X_i$ of country $c_i$, symbolically,

\[
r_i \propto X_i. \tag{3}
\]

Actually, any linear relationship between $r_i$ and $X_i$ can produce an S-shaped curve. In our model, we let $r_i = (X_i - X_m + d)/(X_M - X_m + d)$ to reduce the number of parameters as much as possible, where $X_M$ and $X_m$ are the largest and smallest value of log(GDP) in the list of countries, respectively. $d$ is a constant taking the value of 1.038, so that the minimum diversity is around 10 according to eq. (11) which is consistent with the empirical data. Additionally, we assign connections from product $p_j$ to the required capabilities with probability $q$. The matrix $P_{kj}$ represents the connections between these two layers:

\[
P_{kj} = \begin{cases} 
1, & \text{with probability } q, \\
0, & \text{otherwise}.
\end{cases}
\]

In the above equation, $j$ is iterated from 1 to $P$ (the total number of possible products). Suppose the adjacency matrix between countries and products is $M_{ij}$. We assume that country $c_i$ can export product $p_j$ (i.e., that $M_{ij} = 1$), if and only if the proportion of capabilities that are owned by the producing countries is at least a certain percent ($s$) of the total number of capabilities that are required by products:

\[
M_{ij} = \begin{cases} 
1, & \text{if } \sum_k C_{ik} P_{kj} \geq (1-s)\sum_k P_{kj}, \\
0, & \text{otherwise}.
\end{cases}
\]

Finally, the export diversity $D_i$ is defined as the total number of types of products that country $c_i$ exports, namely,

\[
D_i = \sum_j M_{ij}. \tag{6}
\]

In the simulations, all $X_i$’s are defined by the real-world log(GDP) data that we have collected, the number of countries is $N$, and $P$ is defined as the total number of products in our data set. The number of capabilities $N_a$, the link density between the capabilities and products $q$, and the substitutability rate $s$ are the parameters. In each simulation, we can generate a tri-partite network according to the rules we introduced, and as a result, the relationship between $X_i$ and $D_i$ can be derived.

**Analytic solution.** Before giving the simulation results, we will first derive the analytic relation between $D_i$ and $X_i$ to explain the mathematical essence of this model.

In our model, all of the connections among the countries, capabilities, and products are independent per se. Therefore, by analyzing the probability that a typical country $c_i$ exports a specific product $p_j$, we can derive its export diversity:

\[
D_i = P \pi_i, \tag{7}
\]

where $\pi_i$ is the probability that country $c_i$ can produce any specific product. Suppose that the capabilities that country $c_i$ possesses and that product $p_j$ requires are $k_{c_i}$ and $k_{p_j}$, respectively. Because the density of capability that a country has is proportional to $X_i$, the expectation value of $k_{c_i}$ is also proportional to $X_i$. To simplify our discussion, we treat $k_{p_j}$ as a predetermined value.

The probability $\pi_i$ can be decomposed into various other probabilities as follows:

\[
\pi_i = \sum_{m=1}^{N_a} \Pr[k_{p_j} = m] \cdot \Pr[c_i \to p_j | k_{p_j} = m], \tag{8}
\]

where $\Pr[c_i \to p_j | k_{p_j} = m]$ is the probability that country $c_i$ exports product $p_j$, which depends on the degree ($k_{p_j}$) of $p_j$ being $m$.

First, because a product has a probability $q$ of requiring one capability, the number of capabilities $k_{p_j}$ required by $p_j$ satisfies a binomial distribution. Hence, we know that the probability that node $p_j$ requires $m$ distinct capabilities is

\[
\Pr[k_{p_j} = m] = \binom{N_a}{m} q^m (1-q)^{N_a-m}. \tag{9}
\]

Second, we derive $\Pr[c_i \to p_j | k_{p_j} = m]$. If the number of connections of nodes $c_i$ and $p_j$ are given, then the situation can be as depicted by fig. 3(B). The probability $\Pr[c_i \to p_j | k_{p_j} = m]$ is the number of connection configurations satisfying that the number of elements in the set of capabilities that are connected with both $c_i$ and $p_j$ is larger than $(1-s)m$ over all of the possible connection
configurations. This number is computed by means of the following steps:

i) There are \( \binom{N_a}{m} \) ways that the product \( p_j \) is connected to \( m \) capabilities.

ii) All of the \( m \) capabilities that are required by product \( p_j \) can be divided into two groups based on whether they are owned by country \( c_i \). Without loss of generality, suppose that there are \( n \) capabilities in the first group (that are possessed by \( c_i \), i.e., A3 and A4 in fig. 3(B)), and \( m - n \) capabilities in the second group (i.e., A2 in fig. 3(B)). (See fig. 3(C)).

iii) There are \( \binom{k_{c_i}}{n} \) ways to match the capabilities in the first group.

iv) Similar to iii), the number of ways that the capabilities in the second group can be matched with the \( N_a - k_{c_i} \) capabilities that are not owned by country \( c_i \) is \( \binom{N_a - k_{c_i}}{m-n} \).

Indeed, because there must be at least \( \lceil (1 - s)m \rceil \) capabilities in the first group, we can obtain

\[
Pr\{c_i \to p_j | k_{p_j} = m\} = \sum_{n=n_1}^{m} \binom{k_{c_i}}{n} \binom{N_a - k_{c_i}}{m-n} \tag{10}
\]

In the above equation, the summation index \( n \) begins at \( n_1 = \max\left( \lceil (1 - s)m \rceil, k_{c_i} + m - N_a \right) \) because the number of capabilities \( (m - n) \) that are not required by \( p_j \) cannot exceed \( N_a - k_{c_i} \) (the number of capabilities that are not owned by \( c_i \)) and \( n \) must be larger than \( k_{c_i} + m - N_a \). By inserting eqs. (10) and (9) into eqs. (8) and (7), we can derive the following expression

\[
D_i = P \sum_{m=1}^{N_a} \binom{N_a}{m} q^m (1 - q)^{N_a - m} \sum_{n=n_1}^{m} \binom{k_{c_i}}{n} \binom{N_a - k_{c_i}}{m-n} \tag{11}
\]

Notice that \( k_{c_i} \propto X_i \), which implies that \( D_i \) is actually a function of \( X_i \). If we allow duplicate links to exist in the network, then the summands in the second summation of eq. (11) can be replaced by a binomial distribution. Furthermore, we can use \( \lceil (1 - s)m \rceil \) to approximate \( \max\left( \lceil (1 - s)m \rceil, k_{c_i} + m - N_a \right) \), and set \( s = 0 \), then eq. (11) becomes \( (k_{c_i} q + 1 - q)^N \) according to the binomial theorem. Because eq. (11) is the same equation as the relation between capability and diversity derived that was in paper [7], it is actually a general definition of \( D_i \) in terms of \( X_i \) in which substitutability between capabilities is allowed.

**Results.**

The S-shaped curve. In the previous sections we introduced our model. Here, we will give our simulation and numeric results. In fig. 1(B), the blue circles represent the simulation results and the red squares represent both the numeric results of eq. (11) and the logistic fitting. When we set the number of capabilities \( (N_a) \) at 200 [5], the number of products \( (P) \) at 765 (which is also the maximum product diversity of the countries in our empirical data), the link density of capabilities and products \( (q) \) at 0.067, and the substitutability at 0.4, we obtain an “S” curve that resembles the empirical curve of the best fit for the data recorded in 1995. Furthermore, we use the logistic eq. (1) to fit both the empirical and theoretical curves and to compare their fitting parameters. We found that the parameters are similar (fig. 1). Therefore, we conclude that our model can simulate the empirical S-shaped relationship very well.

Parameter space. Although there are several parameters, the most important ones are \( q \) and \( s \). In fact, we can fix the other parameters (specifically, we select \( P = 765 \) and \( N_a = 200 \)) and study how \( q \) and \( s \) affect the shape of the “S” curve. From the notions introduced above, the parameter \( q \) determines the capabilities that are required by products. Thus, we can understand \( q \) as the average complexity of all products. As \( q \) increases, countries find it more difficult to make products, and, as a result, the S-shaped curve is steeper and the diversity gap between rich countries and poor countries becomes large (see fig. 4(A) and [15,16]).

The parameter \( s \) represents the average substitutability degree of the products: one country must possess a proportion of \( 1 - s \) of the capabilities required by a product if this country wants to export that product. From fig. 4(B), no ceiling for the S-shaped curve can be observed when \( s \) is small because countries need to have locally almost all of the capabilities required by products in this case. Thus, when \( s \) is zero, the simulation result is the same as the result of Hidalgo’s model. In contrast, a ceiling for the export diversity emerges as \( s \) increases because more capabilities that are required by products can be replaced by other available capabilities that are owned by producing countries. Hence, the resources (for instance, labor, skills, fund and so on) required to produce the goods in question are more diverse.
reasonable fitting parameters. From fig. 5(B), (C), we can observe that whereas the slope ($k$) of $X_i D_i$ is influenced mainly by $q$ and not $s$, the center position of the curve is determined mainly by $s$.

**Conclusion.** – In general, this paper discusses how the revision of Hidalgo et al.’s tri-partite network model can generate the observed S-shaped curve of the global export diversity, which depends on the economic sizes of countries. In this model, we found that the substitutability $s$ is an important parameter that can account for the ceiling effect in the S-shaped curve. However, this work is only the first step toward a fuller understanding of the export diversity in the international trade. The S-shaped curve that models the relationship between diversity and economic size can only show the aggregate information regarding one country’s export diversity. Additional studies that investigate the distribution of different products in a given country are worth conducting in the future.

**REFERENCES**

[1] Hu L., Tian K., Wang X. and Zhang J., *Physica A*, 391 (2012) 731.
[2] Eagle N., Macy M. and Claxton R., *Science*, 328 (2010) 1029.
[3] Krugman P. R., *J. Int. Econ.*, 9 (1979) 469.
[4] Petersson L., *S. Afr. J. Econ.*, 73 (2005) 785.
[5] Hidalgo C. A. and Hausmann R., *Proc. Natl. Acad. Sci. U.S.A.*, 106 (2009) 10570.
[6] Hausmann R. and Hidalgo C. A., *J. Econ. Growth*, 16 (2011) 309.
[7] Hausmann R. and Hidalgo C. A., *CID Working Paper*, No. 201 (2010).
[8] Zhang J. and Yu T., *Physica A*, 389 (2010) 4887.
[9] Sozen A. and Arcaklioglu E., *Energy Policy*, 35 (2007) 4981.
[10] Mozumder P. and Marathe A., *Energy Policy*, 35 (2007) 395.
[11] Camba-Mendez G., Kapetanios G., Smith R. J. and Weale M. R., *Econom. J.*, 4 (2001) S56.
[12] Uzawa H., *Rev. Econ. Stud.*, 29 (1962) 291.
[13] Grossman J. B., *Rev. Econ. Stat.*, 64 (1982) 596.
[14] Syverson C., *Rev. Econ. Stat.*, 86 (2004) 534.
[15] Hidalgo C. A., Klinger B., Barabasi A.-L. and Hausmann R., *Science*, 317 (2007) 482.
[16] Ausloos M. and Lambiotte R., *Physica A*, 382 (2007) 16.

We acknowledge the support of the National Natural Science Foundation of China under Grant No. 61004107.