Institutions
Markets
Technologies
IMT
INSTITUTE
FOR ADVANCED
STUDIES
LUCCA

#05
2013

ISSN 2279-6894
IMT LUCCA EIC WORKING
PAPER SERIES 05
September 2013

RA **Economics and institutional change**

Missing Links in Multiple Trade Networks

Rachele Foschi
Massimo Riccaboni
Stefano Schiavo

ISSN 2279-6894
IMT LUCCA EIC WORKING PAPER SERIES #05/2013

Piazza San Ponziano 6, 55100 Lucca

Research Area
Economics and institutional change

Missing Links in Multiple Trade Networks

Rachele Foschi
IMT Institute for Advanced Studies Lucca

Massimo Riccaboni
IMT Institute for Advanced Studies Lucca

Stefano Schiavo
School of International Studies and Department of Economics and Management, University of Trento

Missing Links in Multiple Trade Networks

Rachele Foschi* Massimo Riccaboni† Stefano Schiavo‡

September 9, 2013

Abstract

In this paper we develop a network model of international trade which is able to replicate the concentrated and sparse nature of trade data. Our model extends the preferential attachment (PA) growth model to the case of multiple networks. Countries trade a variety of goods of different complexity. Every country progressively evolves from trading less sophisticated to high-tech goods. The probability to capture more trade opportunities at a given level of complexity and to start trading more complex goods are both proportional to the number of existing trade links. We provide a set of theoretical predictions and simulative results. A calibration exercise shows that our model replicates the same concentration level of world trade as well as the sparsity pattern of the trade matrix. Moreover, we find a lower bound for the share of genuine missing trade links. We also discuss a set of numerical solutions to deal with large multiple networks. **JEL Codes**: F14, F43. **Keywords:** Multiple Networks, Trade Networks, Preferential Attachment, Missing Trade, Innovation and Trade

1 Introduction

The paper develops a network description of international trade that accounts for the large fraction of zero trade flows one observes across countries. The work combines two stream of the recent literature: the former involves the study of the empirical regularities characterizing international trade flows, especially focusing on those stylized facts that are puzzling for standard economic modelsBaldwin & Harrigan (2007), Helpman et al. (2008), Armenter & Koren (2010), Easterly & Reshef (2009); the latter relates to the increasing use of network representations and concepts to

*IMT Advanced Studies Lucca - Laboratory of Innovation Management and Economics
†IMT Advanced Studies Lucca - Laboratory of Innovation Management and Economics
‡University of Trento - School of International Studies and Department of Economics and Management; OFCE-DRIC

describe economic systems, particularly international tradeHidalgo et al. (2007), Chaney (2010).

The sparse nature of trade data, resulting in a large proportion of zero trade flows has received a good deal of attention in recent years. By analyzing trade among 158 countries over the years 1970–1997 it has been shown that just around a half of all possible country-pairs links are ever activated (either in one or the other direction)Helpman et al. (2008). The pervasiveness of zeros increases the higher the degree of disaggregation: 82% of potential product-partner trade flows are actually zero for US trade data at 10-digit Harmonized System (HS)Baldwin & Harrigan (2007). The share goes up to 92% for imports. Similarly for product-destinations pairs, the share of zeros relative to the number of potential flows that range between 69% and 99.5%, with a mean value of 96% based on UN-Comtrade data at the HS-6 levelEasterly & Reshef (2009).

In the network literature, very skewed connectivity distributions are found to characterize many real-world applications beside trade (the internet, world-wide air transportation, mobile communication, interbank payments to quote just a few),so that a network approach appears well-placed to account for the two features of international trade data discussed so far. The simplest null model to account for the power-law connectivity distribution of real-world networks is the preferential attachment (PA) growth model Barabási & Albert (1999). However, to generate skewed connectivity and sparse network structures, the PA regime must be complemented by a constant inflow of new nodes. Such a model does not apply to the international trade network where the set of nodes (i.e. countries) is almost constant in time. To solve this puzzle we propose a simple generalization of the PA model to describe the topological structure of bilateral trade flows across countries. Given the large number of products that are exchanged internationally relative to the number of countries, for the process to match the large number of zeros observed in the data the aggregate adjacency matrix has to be decomposed in nested sub-matrices of different dimensions, representing various trade networks in which specific product are traded and not all countries are simultaneously operating. This formal treatment of the problem suggests a learning process whereby many countries trade the most basic products, whereas a small minority manages to produce and export the most sophisticated manufactured goodsHidalgo et al. (2007). In other words, we keep fixed the set of countries and consider multiple trade networks sorted by the complexity of traded goods. The PA regime is active across all networks and we model the entry probability of a country in high-tech trade as proportional to the total number of trade relationships it has already activated. The cheaper computational costs of this procedures also reflect in an inferior mathematical complexity, in writing the distributions of the simulated quantities of interest. This fact allows us to derive useful analytical properties of the process we want to reproduce.

This paper is structured as follows. Section 2 is devoted to the descriptive analysis of data and to illustrate the stylized properties of the real-world trade network that we aim at reproducing. Section 3 contains the main results of the paper: it illustrates the decomposition procedure and the criteria to establish the dimensions of the product networks and the number of links to be allocated in each of them. In Section 4 the procedure is applied to trade data. Section 5, analyzes from the mathematical point of view the topics of Section 3 and shows some regularity properties of the involved processes, such as the Markov property, and providing closed formulas for probability distributions. Finally, Section 6 contains discussions, about the role of cross-product dependence in our study, further research directions and conclusions.

2 Descriptive data analysis

From an empirical point of view, we refer to the BACI dataset maintained by CEPII, and reporting bilateral trade flows among a large number of countries over the years 1995–2009.[1] The data are organized according to the Harmonized System classification, at 6-digit level which is the finest level of disaggregation comparable internationally. Hence, each trade flow is defined by the source and destination country, the product code and the dollar amount shipped. Since we are mainly interested in the number of zeros and connectivity distribution, we disregard the information on the value of bilateral flows and re-aggregate the data at the country level by counting how many 6-digit products are exported from country i to country j. After dropping some small countries and territories, for each year we end up with a 189×189 matrix whose (i, j)-th entry K_{ij} represents the number of products exported from i to j.

From the data, we calculate the share of zeros in each trade matrix: the average over the whole 1995–2009 period is about 47 percent, ranging from 42 percent in 2008 to 57 percent in 1995.[2] Most countries export a small number of product to few destinations, while only a few players are extremely connected. Indeed, this is consistent with previous findings pointing to a core-periphery structure of world tradeFagiolo et al. (2009) .Data for 2001, for instance, tell that the number of (product-destination) links for each country N_r ranges between 35 and 322 064 (mean 30 075, standard deviation 59 144). We also notice that leading countries tend to dominate trade in every product category (see Figure 1). This evidence supports the view that most central nodes tend to be the same in all product

[1] www.cepii.fr/anglaisgraph/bdd/baci.htm

[2] Since the structural properties of the resulting network are rather stable over time (see for instance Fagiolo et al. (2009)), the specific year analyzed is not crucial for the results.

networks.

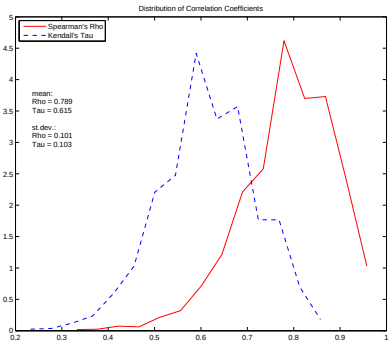

Figure 1: The distribution of the correlation between the number of trade destinations of a country for different HS2 products. For every country, we obtain 3,741 values, ranging between .2 and .9. Average correlations are: .62 (Sperman) and .79 (Kendall).

3 A generalized preferential attachment model for multiple networks

To replicate the sparsity patter of world trade we start from a pure PA model: countries establish new trade links based on the number of connection they already have. Hence, more active exporters are more likely to export new products and/or reach new markets. In our network model we start form a given set of active players (countries) w_0, each trading one product to a single destination: in our application we have $w_0 = 189$, thus there is no entry of new countries. Starting from this 189×189 matrix, N_{tot} trade links are allocated (each representing a product-destination pair), one at each step, according to the following procedure: the outgoing (incoming) link is assigned proportionally to the *export* (*import*) connectivity of countries, that is the probability to catch a new outgoing (incoming) connection is proportional to the node outdegree (indegree). This pure PA mechanism with no entry fills up the trade matrix too rapidly (the share of zeros is too low).[3]. Figure 2 shows that the zeros' percentage decreases at the increasing of the number of allocated links. It decreases very quickly for small values of N_{tot}, while it tends to stabilize for large N_{tot}.

In reproducing trade data, we are interested in $P(K = 0) \in (.42, .57)$ which we can obtain by keeping bounded the number of links to be allocated:

[3]Conversely, if we set $w_0 = 0$ and let the country enter with a constant probability α the share of zeros is too high

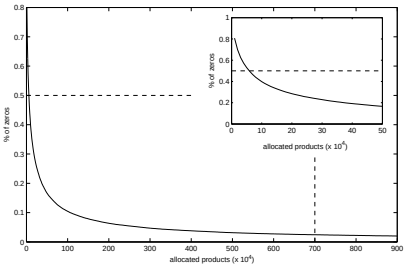

Figure 2: Probability of missing links in international trade, 189x189 country matrix

$N_{tot} \in (55000, 130000)$. However, this is much smaller than the real number of links observed in the data, which are in the order of $N_{tot} \approx 5 - 7 \cdot 10^6$.

To solve this puzzle, we propose a generalized PA mechanism for the growth of multiple networks. Our approach consists in allocating products to different networks, i.e. we group products into different categories. More precisely, to implement our method we must decide: (a) how to split total trade network in product specific trade sub-networks; (b) how to establish the number of products to be allocated in each sub-network; (c) how to re-aggregate product networks to obtain the aggregate world trade network. In choosing the different sub-matrices, the idea is reproducing the allocation dynamics related to different types of products, due for instance to their different level of technological intensity. In this context, products with the lowest complexity are those exported, or generally traded, by all countries, whereas the most sophisticated goods are produced and sold by a small number of countries. Those papers employ a methods of reflections that looks at the number of countries exporting one product to infer its complexity, and the ubiquity of a country's export mix to infer its capabilitiesHausmann & Hidalgo (2010). The idea is that the more products a country exports, the more capabilities it has.

Formally, differentiating among goods based on their complexity implies a progressive narrowing of the dimension of the matrices where we allocate products. Our model generalized the PA model by differentiating two dimensions: (1) the probability to catch a new trade opportunity for a given product is proportional to the number of links a country already has and (2) the probability to start trading a new product is proportional to the total number of connections a country has across all products. Once we have chosen number and dimensions of the sub-matrices of the original $N \times N$ matrix, we have to establish the number of products to be allocated in each sub-matrix. The number of zeros is decreasing with respect to the aggregation operation, as the latter adds items to cells, but cannot remove them. This

fact offers a first criterion for establishing an upper bound for the number of objects to be allocated in the low-tech matrix.

Among the possible criteria for choosing number and dimensions of the sub-matrices and the numbers of products to be allocated in each of them, we chose a decomposition method, based on $P(zeros)$ quantiles, keeping (as much as possible) constant the percentage of zeros in the different sub-matrices. To this aim, we simulate the number of allocations required in order to obtain a given percentage of zeros α in a matrix of a given dimension n (number of countries), $n = 10, \ldots, 190$.

For any given group of countries, we count the number of HS6 products they export. To obtain the number of links to be allocated in each sub-network, we have to complete the computation to the number of trade links (HS6 product-destination pairs) by any given number of countries.

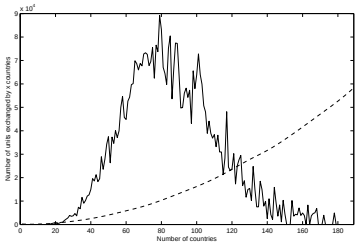

Figure 3: Number of units of products exported by countries, overlapped to the graphs of $Prod_\alpha(dim)$, the number of links to be allocated, needed for having a share of zeros $\alpha = 1/2$.

The idea is the following: looking at Fig. 3, starting from the right of the horizontal axis, we sum the numbers of units till such a sum reaches the curve corresponding to the wanted zeros' percentage α. In formulas, we can put

$$N_1 = N,\ N_{h+1} = \\ = \inf\left\{x| \sum_{c=x+1}^{N_h} ex\_un(c) < Prod_\alpha(N_h)\right\}, \tag{1}$$

and

$$n_h = \sum_{c=N_{h+1}+1}^{N_h} ex\_un(c),\ h = 1, \ldots, b-1. \tag{2}$$

We obtain therefore a decomposition of the matrix A into b sub-matrices $M_1, \ldots, M_b$, whose dimensions are respectively $N_1, \ldots, N_b$, and, in each sub-matrix M_h, we have to allocate n_h objects.

Product networks must next be aggregated to obtain the world trade matrix. Since the probability to enter new trade networks is proportional to total connectivity, we aggregate the matrices be means of the operation $\oplus$, consisting in a sum operation after having ordered columns and rows of the sub-matrices according to their connectivity, i.e. in summing up the matrices, resulting from a decreasing sorting both of totals of rows and columns. Since we observe that connectivity is strongly correlated across product networks, this aggregation procedure is both theoretically and empirically grounded. Such an ordering in particular concentrates the non-empty entries in the left upper corner and minimizes the probability that a non-empty entry of M_{i+1} is summed up to an empty entry of the matrix $M_1 \oplus \cdots \oplus M_i$. This method allows us to describe and recreate a stronger concentration than the classical PA model without entry, that actually leads to a too low percentage of zeros an to too uniformly filled up matrices. This aggregation procedure instead makes the most connected countries, i.e. the ones in the left upper corner, benefit of the PA in any technology level where they are active and of a sort of PA of levels themselves. In other words, such countries 'attract' products and, in any market, they 'attract' trade opportunities.

Let be $N_1, \ldots, N_b$ the dimensions of the sub-matrices $M_1, \ldots, M_b$, with b the number of the sub-matrices and $N_1 > \cdots > N_b$, and $\alpha_1, \ldots, \alpha_b$ the proportions of zeros in any sub-matrix. We define the aggregated matrix by

$$M = M_1 \oplus \cdots \oplus M_b.$$

We denote by $P_M[0](\alpha)$ and $\mathbb{E}_M[0](\alpha)$ respectively the zero probability and the expected number of zeros in M. We provide here some formulas allowing us to compute $P_M[0], \mathbb{E}_M[0]$ in terms of given $b, N_1, \ldots, N_b$, $\alpha_1, \ldots, \alpha_b$. Conversely, these formulas can be used for calibrating the zeros' percentages $\alpha_1, \ldots, \alpha_b$ (or, for sake of simplicity, $\alpha = \alpha_1 = \cdots = \alpha_b$), in order to obtain an acceptable percentage of zeros in the aggregated matrix M. With the following propositions, we provide the analytical expression of the function linking the disaggregate zeros' percentages to the aggregate one.

Proposition 1.

$$\mathbb{E}_M[0](\alpha) = \alpha_1 N_1^2 \prod_{i=2}^{b} \left[1 - (1 - \alpha_i) \left(\frac{N_i}{N_1} \right)^2 \right]. \tag{3}$$

Proof. Let us separately consider the matrices $M_1, \ldots, M_b$. The expected number of zeros any M_i contains is $\alpha_i N_i^2$ and therefore any M_i contains $(1 - \alpha_i) N_i^2$ non-zero entries. By aggregating by using the coordinate-wise matrices' sum + (i.e. without reordering in advance the elements of the sub-matrices) the first two matrices, M_1, M_2, any non-zero entry of M_2 has a probability α_1 to occupy a zero entry of M_1. The expected number of zeros of $M_1 + M_2$ is

$$\mathbb{E}_{M_1+M_2}[0](\alpha) = \alpha_1 N_1^2 - (1-\alpha_2)N_2\alpha_1 =$$
$$= \alpha_1 N_1^2 \left[1 - (1-\alpha_2)\left(\frac{N_2}{N_1}\right)^2\right].$$

A non-zero entry of M_3 has a probability $\frac{\mathbb{E}_{M_1+M_2}[0](\alpha)}{N_1^2}$ to occupy a zero entry of $M_1 + M_2$, so that

$$\mathbb{E}_{M_1+M_2+M_3}[0](\alpha) =$$
$$= \alpha_1 N_1^2 \left[1 - (1-\alpha_2)\left(\frac{N_2}{N_1}\right)^2\right]$$
$$\left[1 - (1-\alpha_3)\left(\frac{N_3}{N_1}\right)^2\right].$$

By iteration, we obtain the thesis. □

In this context the probability that a given entry contains a zero can be seen as a binomial probability, therefore it can be simply obtained by its expected value by dividing by the number of trials, in this case N_1^2, so to obtain

$$P_M[0](\alpha) = \alpha_1 \prod_{i=2}^{b} \left[1 - (1-\alpha_i)\left(\frac{N_i}{N_1}\right)^2\right]. \tag{4}$$

The advantage of dealing with expected values in place of probabilities is due to the linearity of the expected value. Eq. (4) can be used both to compute $P_M[0](\alpha)$, by assigning $\alpha_1, \ldots, \alpha_b$, and, conversely, to obtain, numerically, the $\alpha_1, \ldots, \alpha_b$ to be assigned in order to get a fixed $P_M[0](\alpha)$. In our application, we consider, for any $i = 1, \ldots, b$, $\alpha_i = \alpha$. Notice that $1 - (1-\alpha_i)\left(\frac{N_i}{N_1}\right)^2 < 1$. By replacing $\alpha_i = \alpha$ in eq. (4) we get $P_M[0](\alpha) < \alpha$. Heuristically, this fact implies that, in order to obtain an aggregated matrix with percentage of zeros $P_M[0](\alpha)$, we have to assign to the disaggregated matrices a higher percentage of zeros α. We provide a formula for the needed percentage α in the following two cases, that is useful to find a suitable α to be inserted in the simulation.

The relation

$$N_i \leq (1-\alpha_{i-1})N_{i-1} \text{ for any } i = 2, \ldots, b \tag{5}$$

in particular implies that, for any $i = 2, \ldots, b$, $(1-\alpha_i)N_i < (1-\alpha_{i-1})N_{i-1}$. In this case, since M_2 is contained in the upper $[(1-\alpha_1)N_1]^2$ block of M_1, any result of the allocations in the sub-matrices $M_2, \ldots, M_b$ does not affect the number of zeros of M_1. Therefore it is sufficient computing these last ones, amounting to αN_1^2 and trivially giving

$$P_M[0](\alpha) = \alpha. \tag{6}$$

Proposition 2. *Let be* $\alpha_i = \alpha_1 = \alpha$ *and* $N_i > (1-\alpha)N_{i-1}$ *for any* $i = 2, \ldots, b$. *Then*

$$P_M[0](\alpha) = 2\alpha - \alpha^2 - \frac{1-\alpha}{N_1^2} \sum_{i=1}^{\bar{b}} \frac{N_i}{2^{i-1}} (N_i - (1-\alpha)N_1). \tag{7}$$

Proof. Let us consider the increasing sequence $\mathfrak{N} = \{(1-\alpha)N_b, (1-\alpha)N_{b-1}, \ldots$ $\ldots, N_b, \ldots, (1-\alpha)N_1, N_{\bar{b}} \ldots, N_2, N_1\}$ where $\bar{b} = \max\{i | N_i > (1-\alpha)N_1\}$, and the partition of the $N_1 \times N_1$ matrix determined by the cartesian product $\mathfrak{N}^2$.
All the squares contained in the upper square $(1-\alpha)N_1 \times (1-\alpha)N_1$ are non-zero with probability 1. All the squares contained in the lower square $\alpha N_1 \times \alpha N_1$ are zero with probability 1.
We focus on the squares generated by the cartesian product $\{(1-\alpha)N_b, (1-\alpha)N_{b-1}, \ldots, N_b, \ldots, (1-\alpha)N_1\} \times \{(1-\alpha)N_1, N_{\bar{b}} \ldots, N_2, N_1\}$. The rectangle $\alpha N_1 \times (1-\alpha)N_1$, belonging only to the largest matrix, M_1, has non-zero entries with probability $\frac{1}{2}$. The rectangle $N_2 - (1-\alpha)N_1 \times (1-\alpha)N_2$ is also contained in the matrix M_2. We already observe therein a probability $\frac{1}{2}$ of non-zero entries due to the allocation of objects in the matrix M_1. The allocation process in the matrix M_2 generates in such a rectangle a non-zero with probability $\left(\frac{1}{2}\right)^2$. By iteration, we obtain that the number of non-zeros of the aggregate matrix M is given by

$$(1-\alpha)^2 N_1^2 + (1-\alpha) \sum_{i=1}^{\bar{b}} \frac{N_i}{2^{i-1}} (N_i - (1-\alpha)N_1).$$

□

Remark 1. *Eq.* (7) *does not hold any more, if we drop the condition* $\alpha_i = \alpha$. *In this case, in fact, we are no more able to order the terms* $(1-\alpha_i)N_i$ *for the different i's.*

Actually, objects falling in the upper square occupy an empty entry with a probability greater than 0.

In this case, however, given a greater complexity and variety of situations, we can provide for $P_M[0](\alpha)$ only an upper bound, reflecting on a lower bound for α.

Proposition 3. *For any fixed* $\alpha \in [0,1]$, *let* M_1 *be the connectivity ordered (square) matrix, with zeros' percentage* α. *Let* $1-\beta$ *be the zeros' percentage in the upper* $[(1-\alpha)N_1]^2$ *block and* $1-\tilde{\beta}$ *be the zeros' percentage in the rectangular blocks* $(1-\alpha)N_1 \times (N_1 - (1-\alpha)N_1)$.

$$P_M[0](\alpha) \leq 1 - \frac{2}{N_1^2}(1-\alpha)\Big\{ \sum_{i=1}^{\bar{b}} (1-\tilde{\beta})^{i-1} N_i^2 \left[\tilde{\beta} + (1-\alpha)(\beta - \tilde{\beta}\frac{N_1}{N_i})\right] +$$

$$(1-\alpha)\tilde{\beta} \sum_{i=2}^{\bar{b}} N_i \sum_{h=2}^{i} (1-\tilde{\beta})^{h-2} (1-\beta)^{i-h+1} (N_{i-h+1} - N_{i-h+2}) \Big\}. \tag{8}$$

Proof. We divide M_1 in two regions: the square where the probability of non-zeros is β and the rectangles where the probability of non-zeros is $\tilde{\beta}$. In each rectangle, the number of non-zeros added at any step i, by adding the (ordered) matrix M_i, is proportional to the number of entries of the rectangle, amounting to $(1-\alpha)N_i(N_i-(1-\alpha)N_1)$; each of these entries is "filled" with probability $\tilde{\beta}$, conditionally on this entry having kept empty till the i-th step; this conditioning event has probability $(1-\tilde{\beta})^{i-1}$. The already filled entries have been counted recursively at the previous steps.
The behavior in the square block is more complicated. The sub-matrices M_i, $i=2,\ldots,\bar{b}$, contain the block and each of them adds, in the (nested) square $[(1-\alpha)N_i]^2$ non-zero entries with probability β, conditionally on this entry having kept empty till the i-th step; this conditioning event has probability $(1-\beta)^{i-1}$.
Out of this block, we divide the rectangular block of each matrix M_i, where the non-zeros' percentage is $\tilde{\beta}$, in strips of width $(1-\alpha)(N_{i-h+1}-N_{i-h+2})$, $i=2,\ldots,\bar{b}$, $h=2,\ldots,i$. In each strip, the non-zeros probability is $\tilde{\beta}(1-\tilde{\beta})^{h-2}(1-\beta)^{i-h+1}$. Summing on the two indices, the contribution of non-zeros the sub-matrices M_i, $i=2,\ldots,\bar{b}$, give in the rectangular blocks is

$$2\tilde{\beta}(1-\alpha)^2\sum_{i=2}^{\bar{b}}N_i\sum_{h=2}^{i}(1-\tilde{\beta})^{h-2}(1-\beta)^{i-h+1}\}(N_{i-h+1}-N_{i-h+2}).$$

By summarizing the amounts of zeros in the different regions and dividing by the total number of entries of the aggregate matrix, we obtain Eq. (8).

For $i,j=\bar{b}+1,\ldots,b$, we do not know any more the relation between N_j and $(1-\alpha)N_i$ and therefore are no more able to compute the expected number of added non-zeros. Thus Eq. (8) comes from an under-estimation of the non-zeros' probability, that is an over-estimation of the zeros' probability and therefore it gives a lower bound for α. □

An analog of Eq. (6), when $\beta\neq 1\neq 2\tilde{\beta}$, can be obtained by dropping by (8) the part corresponding to the non-zeros added in the rectangular blocks of M_1 and modifying the part corresponding to its upper square block, in the light of the condition $(1-\alpha_i)N_i<N_i<(1-\alpha_{i-1})N_{i-1}$.

4 Model calibration

Since it is not possible to analytically nor computationally invert Eq. (7), in order to compute a suitable α for our simulation, we have to: (a) start from an arbitrary value of α, representing the percentage of zeros in each sub-matrix; (b) find suitable number and dimensions of the sub-matrices given α; (c) check which condition, between $N_i=(1-\alpha_{i-1})N_{i-1}$ and $N_i>(1-\alpha)N_{i-1}$ for any $i=2,\ldots,b$, our data satisfy; (d) compute the share of zeros, as it results from Eq. (7); (d) accept or not such a value and possibly repeat the procedure.

Applying it to the data of world trade in 2001, we obtain a decomposition in 156 sub-matrices, each with a percentage of zeros $\alpha=.575$. Actually,

the zeros' percentage in the aggregate matrix amounts to .437, instead of the expected $P_M[0](.575) = 0.4$. This is a consequence of the fact that, contrarily to what happens with the real matrices, in the simulated ones, the connectivity reordering concentrates non-zeros in the upper square of each matrix, but is not able to fully move zeros in other blocks. However, as the simulated data show, we can assume in a good approximation $\beta = 1$, $\tilde{\beta} = \frac{1}{2}$ and use, for a lower bound on α, Eq. (7). Actually Eq. (7) provides not precisely a lower bound for α, but a locally optimal solution. In fact, the zeros' percentage further decreases to .34 for $\alpha = .625$, while, for $\alpha = .75$, it increases again to .35.[4]

5 Probabilistic aspects and analytical properties of the allocation process

In this section we want to study some analytical properties of the processes involved in the simulation. To this purpose, we give first of all some definitions that will be useful in the following.

Definition 1 (Sufficient statistic)**.** *Let* $\mathbf{X}$ *a sample on* $(\Omega, \mathfrak{B}, \mathbb{P})$*, taking values in* $\mathfrak{X}$*, and let* $\mathcal{F} = \{f_{\mathbf{X}}(\cdot, \theta) : \theta \in \Theta\}$ *be a family of probability densities for* $\mathbf{X}$*, depending on the parameter* $\theta \in \Theta$*. A statistic* $T = T(\mathbf{X})$ *is* sufficient *if two functions* g, h *exist, such that, for any* $\theta \in \Theta$ *and for almost any* $\mathbf{x} \in \mathfrak{X}$*,*

$$f_{\mathbf{X}}(\mathbf{x}, \theta) = h(\mathbf{x}) g(T(\mathbf{x}), \theta).$$

Intuitively a sufficient statistic is a function of data containing all the information they can give. For further details and examples, see e.g. Spizzichino (2001).

Definition 2 (Counting process)**.** *A stochastic process* $\{N_t\}_{t \in [0, +\infty)}$ *is a* counting process *if, for any* t*,* N_t *satisfies the following properties:*

- $N_t \in \mathbb{N}$*;*
- $P(N_s \leq N_t) = 1$ *for any* $s < t$*;*

[4] We compared the two solutions in terms of the distribution of the deviation of the resulting export connectivity distribution from the empirical one. In terms of relative frequencies, y_r^i, $y_r^i(.575)$, $y_r^i(.75)$, computed on a histogram with 100 bins, we have

$$\sum_{i=1}^{189} |y_r^i - y_r^i(.575)| = \sum_{i=1}^{189} |y_r^i - y_r^i(.75)| = 66;$$

The reason for such a similarity may be found in the fact that the smaller number of objects to be allocated in each sub-matrix in the decomposition for $\alpha = .75$ is counter-balanced by a more refined partition of the matrix (in 173 sub-matrices instead of 156).

- *for any $s < t$, $N_t - N_s$ is the number of events occurred during the interval $(s, t]$.*

A counting process is said to be simple *if $N_0 = 0$ and*

$$\lim_{\delta \to 0} P(N_{t+\delta} - N_t > 1) = 0.$$

Definition 3 (Markov process)**.** *A stochastic process $\{X_t\}_{t \in [0,+\infty)}$ with discrete state space E is a* Markov process *if, for any $0 \leq s_1, \ldots, s_k < s < t$ and for any $i_1, \ldots, i_k, i, j \in E$,*

$$P(X_t = j | X_s = i, X_{s_1} = i_1, \ldots, X_{s_k} = i_k) = P(X_t = j | X_s = i). \quad (9)$$

Definition 4. *A simple counting process satisfying Eq.* (9) *(Markov property) is called a* pure birth *process.*

Intuitively, $\mathfrak{F}_n$ represents the information status at time n, that is all the information generated by the variables observed till n. Let $\mathfrak{M}_t = (\mathfrak{M}_t^{(1)}, \ldots, \mathfrak{M}_t^{(N)}) = (m_t^{(1)}, \ldots, m_t^{(N)})$ be the observed configuration of the countries' masses at time $t \in \mathbb{N}$. $\mathfrak{M}_t^{(c)}$ is the r.v. counting the number of products allocated to country c until time t. The process $\{\mathfrak{M}_t^{(c)}\}_{t \in \mathbb{N}}$, by construction, is a simple counting process (i.e. it cannot have multiple jumps at a time). The value $m_t^{(c)}$ is a realization of the r.v. $\mathfrak{M}_t^{(c)}$. According to the implemented procedure, we suppose $m_0^{(c)} = 1$ for any $c \in \{1, \ldots, N\}$.

Let us consider trade flows directionally, that is consider only export or import flows, that is $\mathfrak{M}_t^{(i)} = N_c^{(i)}(t)$ or $\mathfrak{M}_t^{(i)} = N_r^{(i)}(t)$, where $N_c^i(t)$ is the sum of the values in the i-th column, i.e. the import connectivity of the i-th country, at time t and $N_r^i(t)$ is the sum of the values in the i-th row, i.e. the export connectivity of the i-th country, at time t. In such a way, assigning a product to the country c means inserting it in some entry on the c-th row of the exchange matrix. Let $P_c(+1, t)$ be the probability that, at the instant t (i.e. at the t-th iteration of the procedure), the country c gets the newly inserted product and let $P_c(+k, [t_1, t_2])$ denote the probability that the country c gets k new products in the time interval $[t_1, t_2]$.

Proposition 4. *For any c and t,*

$$P_c(+1, t) = P_c(+1, 1) = \frac{1}{n}. \quad (10)$$

Proof. Since allocations are proportional to the initial masses, $P_c(+1, 1) = \frac{1}{n}$ obviously follows.

We prove the second equality, $P_c(+1, t) = P_c(+1, 1)$, by induction.

For $t = 2$, Eq. (10) holds. In fact

$$\begin{aligned} P_c(+1, 2) &= P_c(+1, 2| + 0, 1) P_c(+0, 1) + P_c(+1, 2| + 1, 1) P_c(+1, 1) \\ &= \frac{1}{n+1} \cdot \frac{n-1}{n} + \frac{1}{n} \cdot \frac{2}{n+1} = \frac{1}{n}. \end{aligned}$$

Let us now suppose Eq. (10) to hold for $t-1$ and let us prove it for t.

We want to use the total probabilities formula, by conditioning on events of the kind

$$\bigcap_{s=1}^{t-2}\{(+\omega_s, s)\},$$

where $\omega_s \in \{0,1\},\ \omega_0 = 1$ for all the countries. We have a different event for each different disposition $\omega = (\omega_1, \ldots, \omega_{t-2})$, amounting to a number of 2^{t-2}. Actually we are interested in the sufficient statistic of $\omega,\ \mathcal{S}(\omega) = \sum_{s=1}^{t-2}\omega_s$, consisting in the number of 1's contained in the vector ω. In other words, it is not important to know the times when a country "gets" a product, i.e. it is not relevant the product's age, but only the mass of the country. We will condition then on events of the kind $\{\mathcal{S}(\omega) = v\}$. By the inductive hypothesis,

$$P(\mathcal{S}(\omega) = v) = \left(\begin{array}{c} t-2 \\ v \end{array}\right)\frac{1}{n^v}\left(\frac{n-1}{n}\right)^{t-2-v}$$

and

$$\begin{aligned} P(+1, t-1) &= \sum_{v=0}^{t-2} P(+1, t-1|\mathcal{S}(\omega) = v)P(\mathcal{S}(\omega) = v) \\ &= \sum_{v=0}^{t-2}\left(\begin{array}{c} t-2 \\ v \end{array}\right)\frac{1}{n^v}\left(\frac{n-1}{n}\right)^{t-2-v}\frac{v}{n+t-2} = \frac{1}{n}. \end{aligned}$$

$$P(+1,t) = \sum_{v=0}^{t-2}[P(+1,t|+0,t-1,\mathcal{S}(\omega)=v)P(+0,t-1|\mathcal{S}(\omega)=v)$$

$$+P_c(+1,t|+1,t-1,\mathcal{S}(\omega)=v)P(+1,t-1|\mathcal{S}(\omega)=v)]P(\mathcal{S}(\omega)=v) =$$

$$\sum_{v=0}^{t-2}\left(\begin{array}{c} t-2 \\ v \end{array}\right)\frac{(n-1)^{t-2-v}}{n^{t-2}}\left[\frac{v+1}{n+t-1}\cdot\frac{v}{n+t-2} + \frac{v}{n+t-1}\cdot\left(1-\frac{v}{n+t-2}\right)\right]$$

$$= P(+1,t-1) = \frac{1}{n}.$$

Hence $P_c(+1,t) = P_c(+1,1)$ □

Corollary 1. $\{\mathfrak{M}_t^{(c)}\}_{t\in\mathbb{N}}$ *has stationary increments.*

Proof. It follows by Proposition 4. Notice that, for any c, $\mathfrak{M}_t^{(c)} = \sum_{s=1}^{t}\omega_s$ is a sufficient statistic of the history until time t of the attributions of products to the country c.

$$P(\mathfrak{M}_t^{(c)} = k+1) = P_c(+k,[1,t]) = \left(\begin{array}{c} t \\ k \end{array}\right)\frac{1}{n^k}\left(1-\frac{1}{n}\right)^{t-k}.$$

$$P_c(+k,[s+1,t+s]) = P(\mathfrak{M}_{t+s}^{(c)} - \mathfrak{M}_s^{(c)} = k) = P\left(\sum_{q=1}^{t+s}\omega_q - \sum_{q=1}^{s}\omega_q = k\right)$$

$$= P\left(\sum_{q=s+1}^{t+s}\omega_q = k\right) = \left(\begin{array}{c} t+s-(s+1)+1 \\ k \end{array}\right)\frac{1}{n^k}\left(1-\frac{1}{n}\right)^{t-k}.$$

□

Remark 2. *Since at any instant exactly one product is drawn, for any t and for any $k > 1$, $P_c(+k,t) = 0$. Consequently $P_c(+k,[s,t]) > 0$ only if $k \leq t - s + 1$.*

Let us assume now that a general initial configuration $\mathfrak{M}_0$ is allowed. In general, $P_c(+1,t)$ depends now on the country c for which it has to be computed and on $\mathfrak{M}_{t-1}$, only through $m_{t-1}^{(c)}$. We have

$$P_c(+1,1|\mathfrak{M}_0) = \frac{m_0^{(c)}}{\sum_{j=1}^n m_0^{(j)}}.$$

Furthermore it can be proven

Proposition 5. *For any $s < t$,*

$$P_c(+1,t|\mathfrak{M}_s) = P_c(+1,t|\mathfrak{M}_s^{(c)}) = \frac{m_s^{(c)}}{\sum_{j=1}^n m_s^{(j)}} = \frac{nm_s^{(c)}}{\sum_{j=1}^n m_s^{(j)}}P(+1,t).$$

Remark 3. *The previous results hold by replacing in them $P_c(+1,t)$ with $P_c(+1,t|\mathfrak{M}_s)$, that is by multiplying $P_c(+1,t)$ by $\frac{nm_s^{(c)}}{\sum_{j=1}^n m_s^{(j)}}$, where $\mathfrak{M}_s$ is the last observed mass configuration.*

Corollary 2. *For any k,t,s,s' such that $k < t - s'$, $s < s'$,*

$$P(+k,[s',t]|\mathfrak{M}_s) = \prod_{i=0}^{k-1}\frac{m_s^{(c)}+i}{\sum_{j=1}^N m_s^{(j)}+i}\prod_{i=k}^{t-s'}\frac{\sum_{j=1,j\neq c}^N m_s^{(j)}+i-k}{\sum_{j=1}^N m_s^{(j)}+i}. \quad (11)$$

Remark 4. *$\{\mathfrak{M}_t^{(c)}\}_{t\in\mathbb{N}}$ has not independent increments.*
In fact, $P(\mathfrak{M}_{t+s}^{(c)} - \mathfrak{M}_s^{(c)} = k|\mathfrak{M}_s^{(c)} = h) = \frac{h}{n+s}P(+k,[1,t]) \neq P(+k,[1,t]) = P(\mathfrak{M}_{t+s}^{(c)} - \mathfrak{M}_s^{(c)} = k)$. Therefore

$$P(\mathfrak{M}_{t+s}^{(c)} - \mathfrak{M}_s^{(c)} = k, \mathfrak{M}_s^{(c)} = h) \neq P(\mathfrak{M}_{t+s}^{(c)} - \mathfrak{M}_s^{(c)} = k)P(\mathfrak{M}_s^{(c)} = h).$$

Theorem 1. *For any $c \in \{1, \dots, N\}$, $\{\mathfrak{M}_t^{(c)}\}_{t\in\mathbb{N}}$ is a homogeneous Markov process, i.e.*

$$P_c(+1, t|\mathfrak{M}_{t-1}^{(c)}, \dots, \mathfrak{M}_0^{(c)}) = P_c(+1, t|\mathfrak{M}_{t-1}^{(c)}).$$

Proof.

$$P_c(+1, t|\mathfrak{M}_{t-1}^{(c)}, \dots, \mathfrak{M}_0^{(c)}) = P_c(+1, t|m_{t-1}^{(c)}, \dots, m_0^{(c)}) =$$

$$\frac{P_c(m_{t-1}^{(c)} - m_{t-2}^{(c)} + 1, [t-1, t]|m_{t-2}^{(c)}, \dots, m_0^{(c)})}{P_c(m_{t-1}^{(c)} - m_{t-2}^{(c)}, t-1|m_{t-2}^{(c)}, \dots, m_0^{(c)})}.$$

By iteration, we obtain

$$\frac{P_c(m_{t-1}^{(c)} - m_0^{(c)} + 1, [1, t]|m_0^{(c)})}{P_c(m_{t-1}^{(c)} - m_0^{(c)}, [1, t-1]|m_0^{(c)})}.$$

By Corollary 2, we get

$$\frac{\prod_{i=0}^{m_{t-1}^{(c)}-m_0^{(c)}} \frac{m_0^{(c)}+i}{\sum_{j=1}^N m_0^{(j)}+i} \prod_{i=m_{t-1}^{(c)}-m_0^{(c)}+1}^{t-1} \frac{\sum_{j=1, j\neq c}^N m_0^{(j)}+i-m_{t-1}^{(c)}-1+m_0^{(c)}}{\sum_{j=1}^N m_0^{(j)}+i}}{\prod_{i=0}^{m_{t-1}^{(c)}-m_0^{(c)}-1} \frac{m_0^{(c)}+i}{\sum_{j=1}^N m_0^{(j)}+i} \prod_{i=m_{t-1}^{(c)}-m_0^{(c)}}^{t-2} \frac{\sum_{j=1, j\neq c}^N m_0^{(j)}+i-m_{t-1}^{(c)}+m_0^{(c)}}{\sum_{j=1}^N m_0^{(j)}+i}}.$$

By changing the index in the product in the denominator, we finally obtain

$$\frac{m_{t-1}^{(c)}}{\sum_{j=1, j\neq c}^N m_0^{(j)} + m_{t-1}^{(c)}} \cdot \frac{\sum_{j=1, j\neq c}^N m_0^{(j)} + m_{t-1}^{(c)}}{\sum_{j=1}^N m_0^{(j)} + t - 1} = \frac{m_{t-1}^{(c)}}{\sum_{j=1}^N m_{t-1}^{(j)}}.$$

□

Theorem 2. *For any $c \in \{1, \dots, N\}$, $\{\mathfrak{M}_t^{(c)}\}_{t\in\mathbb{N}}$ is a sub-martingale, i.e.*

$$\mathbb{E}[\mathfrak{M}_{t+1}^{(c)}|\mathfrak{M}_t^{(c)}, \dots, \mathfrak{M}_0^{(c)}] \geq \mathfrak{M}_t^{(c)}.$$

Proof. First of all, in view of the Markov property,

$$\mathbb{E}[\mathfrak{M}_{t+1}^{(c)}|\mathfrak{M}_t^{(c)}, \dots, \mathfrak{M}_0^{(c)}] = \mathbb{E}[\mathfrak{M}_{t+1}^{(c)}|\mathfrak{M}_t^{(c)}].$$

Therefore, we can more easily compute

$$\begin{aligned} \mathbb{E}[\mathfrak{M}_{t+1}^{(c)}|\mathfrak{M}_t^{(c)}] &= \mathfrak{M}_t^{(c)} P(+0, t+1|\mathfrak{M}_t^{(c)}) + (\mathfrak{M}_t^{(c)} + 1) P(+1, t+1|\mathfrak{M}_t^{(c)}) \\ &= \mathfrak{M}_t^{(c)} + P(+1, t+1|\mathfrak{M}_t^{(c)}) \ > \ \mathfrak{M}_t^{(c)}. \end{aligned}$$

□

Some consequences of the formulas provided in theorems and propositions above concern

- the probability of assigning to a same country products in a certain number of categories, i.e. allocating products in a same row of a certain number of sub-matrices;
- the analytical expression for the zeros' probability.

Let us suppose again to start with a uniform configuration (one product per country in any matrix). The probability that one product for each of the r lowest levels of technology are assigned to a same fixed country is

$$\left(\frac{1}{N}\right)\left(\frac{1}{N_2}\cdot\frac{N_2}{N}\right)\left(\frac{1}{N_3}\cdot\frac{N_2}{N}\cdot\frac{N_3}{N_2}\right)\cdot\ldots=\frac{1}{N^r}, \tag{12}$$

where, in any factor $\frac{1}{N_l}\cdot\frac{N_2}{N}\cdot\frac{N_3}{N_2}\cdots\frac{N_l}{N_{l-1}}$, $\frac{1}{N_l}$ is the probability that, in the sub-matrix M_l, the product is assigned to the given country, while any term $\frac{N_h}{N_{h-1}}$ is the probability that that country is active in the technology level h, given that it is active in the technology level $h-1$. Since $\frac{1}{N_l}\cdot\frac{N_2}{N}\cdot\frac{N_3}{N_2}\cdots\frac{N_l}{N_{l-1}}=\frac{1}{N}$, Eq. (12) is also the probability that one product for each of r given levels of technology are assigned to a same fixed country.

For what concerns an analytical expression for the zeros' probability, it can be computed in two different (consistent) ways.

At the first step of the allocation procedure,

$$P(K_{ij}=0)=[Prob(\text{selecting a row } \neq i)+$$

$$Prob(\text{selecting row } i)Prob(\text{selecting a column } \neq j|\text{row } i)]P(K_{ij}(0)=0)$$

$$=\left(\frac{N-1}{N}+\frac{1}{N}\frac{N-2}{N-1}\right)\left(1-\frac{1}{N}\right).$$

Such a probability can also be computed as the one of the complementary event of observing a product to be allocated in the entry (i,j), i.e.

$$P(K_{ij}=0)=\left(1-\frac{1}{N}\right)\left(1-\frac{1}{N}\cdot\frac{1}{N-1}\right).$$

$$P(K_{ij}=0|[0,t])=\left(1-\frac{1}{N}\right)\left(1-\frac{1}{N}\cdot\frac{1}{N-1}\right)^t.$$

6 Final discussion

In this paper we presented a generalized version of the PA model Barabási & Albert (1999) which is able to account for the sparsity structure of the world trade network. Our model is based on the idea that almost every country takes part in the trade of low-tech products, while only a few of them have the capabilities to export sophisticated goods. In this paper we define a lower bound for the share of zeros in trade networks by considering perfect correlation between countries' capabilities in trading different products. However, since we know that trade in different products is not perfectly correlatedHidalgo & Hausmann (2009), in future work most realistic assumptions about the product space should be considered. We also contribute to the literature a new method to generate large trade networks. Our methodology allows to generate in parallel multiple product-specific trade networks. Product networks are then aggregated to obtain the world trade web. Therefore, different assumption about cross-product correlation could be implemented by modifying the aggregation function.

References

Armenter, R. & Koren, M. (2010), 'A balls-and-bins model of trade', (7783).

Baldwin, R. & Harrigan, J. (2007), Zeros, quality and space: Trade theory and trade evidence, Working Papers 13214, NBER.

Barabási, A.-L. & Albert, R. (1999), 'Emergence of scaling in random networks', *science* **286**(5439), 509–512.

Chaney, T. (2010), The network structure of international trade. mimeo.

Easterly, W. & Reshef, A. (2009), Big hits in manufacturing exports and development. mimeo.

Fagiolo, G., Reyes, J. & Schiavo, S. (2009), 'World-trade web: Topological properties, dynamics, and evolution', *Physical Review E* **79**(3), 036115.

Hausmann, R. & Hidalgo, C. (2010), Country diversity, product ubiquity and economic divergence, CID Working Paper 201, Harvard Univeristy.

Helpman, E., Melitz, M. & Rubinstein, Y. (2008), 'Estimating trade flows: Trading partners and trading volumes', *Quarterly Journal of Economics* **123**(2), 441–487.

Hidalgo, C. A. & Hausmann, R. (2009), 'The building blocks of economic complexity', *Proceedings of the National Academy of Sciences* **106**, 10570–10575.

Hidalgo, C. A., Klinger, B., Barabasi, A.-L. & Hausmann, R. (2007), 'The product space conditions the development of nations', *Science* **317**(5837), 482–487.

Spizzichino, F. (2001), *Subjective probability models for life-times*, Chapman and Hall/CRC, Boca Raton, FL.

INSTITUTE FOR ADVANCED STUDIES LUCCA