ANDREA TACCHELLA

ECONOMIC COMPLEXITY

Supervisor: Prof. Luciano Pietronero

PilGroup
http://pilhd.phys.uniroma1.it/PILgroup_Economic_Complexity/Home.html
2014, October

WEBSITE:
http://pilhd.phys.uniroma1.it/PILgroup_Economic_Complexity/Home.html

E-MAIL:
andreatacchella@gmail.com

ACKNOWLEDGEMENTS

All the ideas, methods and innovations proposed in this thesis are the result of a team work, under the supervision of Luciano Pietronero, for which I need to thank and acknowledge my friends and colleagues: Matthieu Cristelli, Riccardo Di Clemente, Andrea Gabrielli, Emanuele Pugliese, Fabio Saracco and Andrea Zaccaria.

PUBLICATIONS

Some of the contents of this thesis have been published in peer reviewed journals. In particular:

- Economic Complexity: conceptual grounding of a new metrics for global competitiveness– PlosOne (2013) - Authors: A. Tacchella, M. Cristelli, A. Gabrielli, G. Caldarelli, L. Pietronero.
- How metrics for countries' competitiveness and products' complexity are affected by noise – Complexity Economics (August 2014) - Authors: F. Battiston, M. Cristelli, A. Tacchella, L. Pietronero.
- Heterogeneous Dynamics of Economic Complexity -To appear in PlosOne (2014) - Authors: M. Cristelli, A. Tacchella, L. Pietronero.
- How the Taxonomy of Products Drives the Economic Development of Countries -To appear in PlosOne (2014) - Authors: A. Zaccaria, A. Tacchella, M. Cristelli, L. Pietronero.

CONTENTS

GENERAL INTRODUCTION

Nowadays economic systems have evolved to become globalized, largely financialized and interconnected. This observation leads directly to another: any approximation that regards them as the sum of many independent parts is going to fail. In other words nowadays economic systems are Complex Systems and they are much more than the sum of their single parts, for the fact that they show emergent chaotic behaviours, that we are not able to predict nor to explain within the mainstream framework. As the set of relevant interactions among economic entities is much wider than what has been used to build present-day's economic theories, we can't build up on these theories anymore. What we need is probably more similar to a brand new start. If we want our theories to be predictive in such a complex scenario we need to link them tightly to data and let our models go beyond overly simplified linear approaches. More importantly, we need to state clearly the limits of these theories and to discriminate situations where we can actually make predictions from those where we can't.

The methods, approaches, data and ideas proposed in this work and the way these ingredients are combined together, embody our view on how a brand new start in Macroeconomic modeling should look like, for two reasons. The first is our concern in remaining constantly linked to quantitative data. Of course we make assumptions about some underlying mechanisms and general principia, but we constrain their validity to the extent of the ability of our methods to perform quantitative predictions. The second is the fact that we always try to use the right method for the right problem. Examples of this are provided in chapters 2 and 4 where methods derived from dynamical systems and machine learning are used and tested in the field of economic prediction. Mainstream economic modeling is often rigid and simplistic from a methodological point of view, with a predominant and pervasive use of linear regressions and descriptive statistics. These tools have strict validity limits and, from a conceptual point of view, provide a low level of falsifiability, unless a comprehensive theoretical framework is available. Thus, since reductionism is made hard in economics by the impossibility to repeat experiments and control the environment, our goal is always to look at concrete predictions and we value our results on the basis of their accuracy.

The philosophy that drives our approach is the similarity that exists between economic and ecologic systems. As discussed in chapter 3 economy and ecology are close not only for sharing the same root as a word, but also in the very essence of the interactions they describe: sets of individual competing and cooperating for the allocation of common resources. Once this connection is put in focus it is easy to move concepts and ideas among the two fields and start to observe similarities in the data that we have about economic networks and ecologic ones. In particular we explore the concept of nestedness in interaction networks, observed in ecology and economics: namely when specialized entities only interact with generalist ones. This means that only generalists have access to exclusive relationships. Both in economics and ecology this provides an obvious advantage and leads to the identification of the concept of diversification with a measure of fitness, intended as the ability to perform well and survive in the ecosystem. The presence of a nested structure also suggest the presence of a precise underlying dynamics in the formation of such networks, and allows us to discover very informative taxonomic structures, as described in the final chapter.

We open our discussion in chapter 1 with an empirical observation about the structure of the bipartite network defined by countries and the products they export. The nestedness of this network is exploited to build a quantitative measure of complexity for countries' national economies and for products, in a self-consistent way. This measure of complexity is then used in chapter 2 to introduce a framework in which we are able to predict the Macroeconomic development of a set of countries. By using an approach derived from dynamical systems we are able to define quantitatively the level of predictability of a country's development, similar in spirit to a weather forecast, where turbulent areas are much harder to predict. The parallelism with biological systems is studied more in detail in chapter 3, where we observe how the build-up of diversity seems to follow some universal features. We propose a minimal scheme to model this dynamics and we introduce the concept of *Usefulness* which reveals to be a crucial feature if we want to picture real-world diversity as the combination of smaller Building Blocks. Finally in chapter 4 we look at how the export data can be used to infer technological relations among products. These relations are relevant for countries' development, as we show that countries move in recurrent paths when developing. The accuracy of our methods in predicting new links in the countries-products network is quantitatively assessed and results to be as high as 16% in a-priori selected subsets.

ABOUT THE DATA

Throughout this work we make use of two main datasets on international trade. Both are based on raw data provided by the United Nations COMTRADE database. The first is the BACI World Trade Database [1]. This dataset contains trading data about more than 200 countries and 5000 products classified according to a six digit code (categorization: Harmonized System 2007). It is possible to reduce the number of different product categories by dropping couples of digits from the classification: we use the 4-digit nomenclature accounting for a total of about 1131 product categories. This dataset, as documented in [1], is the result of a reconciliation procedure performed on the annual reports from countries customs offices, gathered by COMTRADE. This database is used in chapter 1, 2 and 4.

The second database is the World Trade Flows database, containing analogous international trade data, but spanning on a much longer time window, from 1962 to 2000. Also for this database a reconciliation procedure was applied on the basis of the COMTRADE data [2]. This database is used in chapter 3 and chapter 4. It is to be noticed that these data are normally used mainly for statistical purposes: in such applications small errors or inconsistencies in the final database are not of crucial importance since they are likely to be of microscopic order with respect to total trades. In our case of application however, since non-linear iterative procedures are involved, any small error in the data, like a missing or fictitious flow of goods, may in principle propagate and have a large effect. In order to deal with this kind of issues we have operated a cautious cleaning procedure on the BACI data. Moreover we have performed an extensive analysis of noise effects on our methodology, which shows that our results are robust even with significant levels of noise (see Appendix 1.C)

1 FITNESS AND COMPLEXITY

CHAPTER ABSTRACT

We observe that countries tend to produce all the possible products they can, given their level of technology and development. Less competitive countries tend to produce combinations of products that are nested subsets of the production of more developed ones. This observation provides a stimulus to look for a metric to characterize the competitiveness of a country in term of the diversification and complexity of its industrial production. We propose a non-linear metric based on the topology of the countries-products (*c-p*) bipartite network. Our metrics is defined as the fixed point of two coupled maps operating in a simplex. We present a detailed comparison of the results of this approach directly with those of the Method of Reflections[3] by Hidalgo and Hausmann, showing the better performance of our method and a more solid, scientific and consistent economic foundation.

1.1 INTRODUCTION

The increasing complexity and interconnectedness of economic systems cannot be anymore neglected by Economics and call for a paradigm change in economic thinking. These aspects must be effectively addressed and incorporated in economic theory.

In this perspective, recent data-driven works [3–5] have proposed a *complexity* approach to measure the intangible elements which drive the competitiveness of countries starting from the dataset of international trade. These works have pointed out that countries commonly considered as *rich* and *competitive* are also characterized by high diversification of their export basket, differently from what expected from the Ricardian economic paradigm [6]. We consider explicitly export data, even if our reasoning would suggest that is the production that needs to be diversified, not necessarily what is exported. But we expect export to be a good proxy of actual production. Moreover data about export is available in a homogeneous standardized format for almost any country in the world, which makes comparisons possible.

It is traditionally supposed in the Ricardian paradigm [6] that the wealthiest countries specialize in economic niches characterized by the production of only few products with a high degree of specialization. This hypothesis can take a simple mathematical representation: if we introduced a binary country-product matrix where entries are equal to 1 if the country exports (under a fixed criterion) the product and 0 otherwise, it would be possible to rearrange rows and columns in a “mostly" block diagonal shape. However, this is not the shape obtained when considering real data: rather by listing countries in increasing order of specialization and products in decreasing order of ubiquity, we obtain an approximately triangular shape (see Fig. 1). This is a clear indication that countries tend to produce all the possible products they can, given their level of technology and development. The fundamental challenge arising

from this observation is therefore how to characterize the competitiveness of a country in term of the diversification and complexity of its exports.

A first attempt in this direction has been recently presented by Hidalgo and Hausmann (HH)[3]. In the present work we study in detail an affine, but substantially different method, both conceptually and mathematically speaking, self-consistent and with a strong economic grounding, to evaluate the competitiveness of countries and the complexity of products. Indeed, as shown below and also in Ref. [7], the HH method suffers from a number of problems both conceptual and practical.

The main differences between our approach and the HH algorithm consist in the non-linearity of our approach and in the fact that the diversity of export basket is explicitly taken into account in our scheme. While the HH method is based on the hypothesis of a linear relation (more precisely an arithmetic average) between the ubiquity of a product and the competitiveness of its exporters at a given order of iteration, our metric is based on a highly non-linear and almost extremal relationship between the *complexity* of products and the *fitness* of countries producing them. Such an approach proves to be much more effective in reflecting the ideas underlying the arguments of a capability driven economic competitiveness with respect to the HH method. In particular, the approximate triangular structure of such a matrix implies that the information that a product is made by a diversified country conveys little information on the complexity of the product itself; indeed these countries export almost all products. Conversely, if we know that a poorly developed country is able to export a given product, it will be very likely that this product requires only the low level of sophistication which characterizes the poor technological development of such a country.

These observations on the fundamental feature of the country-product matrix lead us to formulate the main argument behind our mathematical approach: from one side it is reasonable to measure the competitiveness and development of a country as the sum of the product complexity of its exports. On the other hand, it is misleading to keep such a linear approach to measure the complexity of products in terms of the competitiveness of the respective producers. In other words, the structure of the international exports represented by the country-product matrix does not permit to consider *the complexity of a product as the average of the fitnesses of its producers*[8]. By the above consideration it is instead natural to write a relation such that the complexity of a product is mainly determined by the fitness of the less competitive exporters. This requires the introduction of a strongly non-linear relation, implying that the only possibility for a product to have a high level of sophistication (or complexity) is to be produced only by highly competitive countries. As shown below, these changes with respect to the approach of HH, determine a crucial improvement in the results of the algorithm both from a conceptual and economic point of view.

In summary our method consist in the introduction of two coupled non linear maps. These define two sets of variables: the countries' fitness (F_c), namely the sum of the complexities of the products of a country, and the products' complexity (Q_p), a non-linear function of the fitnesses of the countries producing it. We iterate the two maps until we reach a fixed point.

Each iteration of the algorithm adds higher order information on these quantities up to reach broad Pareto-like distributions for the two metrics at the fixed point.

Given the non-linear features of the algorithm, we extensively test the robustness of our results by numerical simulations. We show that the so found metrics for country competitiveness and product complexity is the unique asymptotic solution (i.e. fixed point) of our non-linear map for any economically meaningful initial condition. Therefore our metrics is measuring a

genuine feature of the country-product matrix and it is not dependent on the initial conditions (Appendix 1.A).

Detailed analyses of these metrics for countries and products allow to verify that they are conceptually consistent and well-grounded from an economic point of view. Moreover they can be used to produce a wealth of new information in various directions both on the economies of countries and on the "zoology" of the space of products. We argue that this scheme also provides a new approach to the fundamental analysis of the productive system of countries and permits the introduction of a non-monetary and non-income based classification of product complexity. One the most important implications is that their direct comparison with standard monetary or income-based indices as GDP of countries can be interpreted as the potential for future growth as discussed in Chapter 2.

Binary country-product (c-p) matrix

In order to define a suitable economic metrics to compare the trades of different countries in different products, taking into account the difference in sizes and total export, as in [3], we use Balassa's Revealed Comparative Advantage (RCA)[9]. Using its definition, we consider a country c to be a competitive exporter of a product p if the value RCA_{cp} of its RCA for such product overcomes some minimal threshold value R^*. We take here this value to be $R^* = 1$ as in standard economics literature[1].

We can therefore construct the *binary* country-product matrix M whose generic element is:

$$M_{cp} = \begin{cases} 1 & \text{if } RCA_{cp} > R^* = 1 \\ 0 & \text{if } RCA_{cp} < R^* = 1 \end{cases} \tag{1}$$

saying that country c can be considered an exporter of product p if and only if (iif) $M_{cp} = 1$. If we represent countries and products as nodes of a network we can pictorially say that the node of the country c is linked to the node of the product p *iif* $M_{cp} = 1$. Since links are not permitted between two countries or two products, the matrix M defines a *bipartite* country-product network. This means that the nodes are divided into two sets: $\{c\}$ of N_c nodes (countries) and $\{p\}$ of N_p nodes (products). Connections (links) are permitted only between couples of nodes belonging to different sets.

In what follows we analyze also the effects of the possibility of including weights in the country-products matrix. In particular, this can be done by defining the *weighted* country-product matrix $\hat{M}$ through

$$\mathcal{M}_{cp} = \frac{q_{cp}}{\sum_{c'} q_{c'p}} \tag{2}$$

with q_{cp} giving the total export (e.g. in US dollars) of country c for product p in the considered year.

The fundamental information about the structure of the international export of products is encrypted in the matrix M. It is however a matrix with some hundreds of thousands of entries and extracting useful information on the status of the single economies is a non-trivial task. A

1 A statistical argument in favor of this choice is presented in [3] (Supplementary Material)

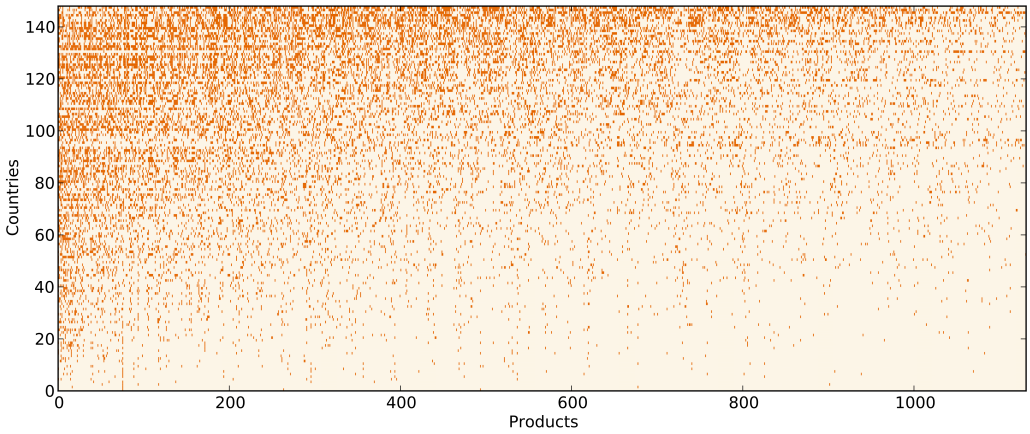


Figure 1: Graphical representation of the experimental matrix Mcp for the year 2010 after reordering of rows and columns by respectively decreasing K_c and K_p. It is evident the substantial triangular structure of the matrix.

first insight is obtained by reordering the rows and columns of the matrix respectively by the total number of exported products by each country

$$k_c = \sum_{p=1}^{N_p} M_{cp} \tag{3}$$

and by the number of exporting countries

$$k_p = \sum_{c=1}^{N_c} M_{cp} \,. \tag{4}$$

The quantities k_c and k_p are the *degree* or *coordination numbers* of the nodes c and p in the bipartite network and are called respectively *diversification* of c and *ubiquity* of p [3]. As shown by Fig. 1, through this procedure, M takes a quite marked triangular structure [3, 5] which is very far from what happens for instance with the same reordering of rows and columns starting from a completely random distribution of the binary entries M_{cp} (compare Figs. 1 and 2). Such an organization of the international trade of products looks very far from the standard view of Ricardian or Heckscher-Ohlin theories which predict as an optimal situation a high degree of specialization of national economies for which it would be possible to rearrange rows and columns so that the matrix M would result almost diagonal or block-diagonal.

This structure makes clear that in the international trade we find countries exporting a large fraction of all products (highly diversified countries), and some others exporting a very small fraction of products (poorly diversified countries). At the same time the products exported by a small number of countries (less ubiquitous products), which are presumably of high complexity value as produced only by few countries, are exported only by highly diversified countries. It is therefore plausible that such structure is related to a ranking in terms of development and competitiveness among the economies of different nations.

The fact that the matrix can be arranged to get a substantially triangular shape rather than block-diagonal, suggests that the dynamical evolution of advanced economies is quite different

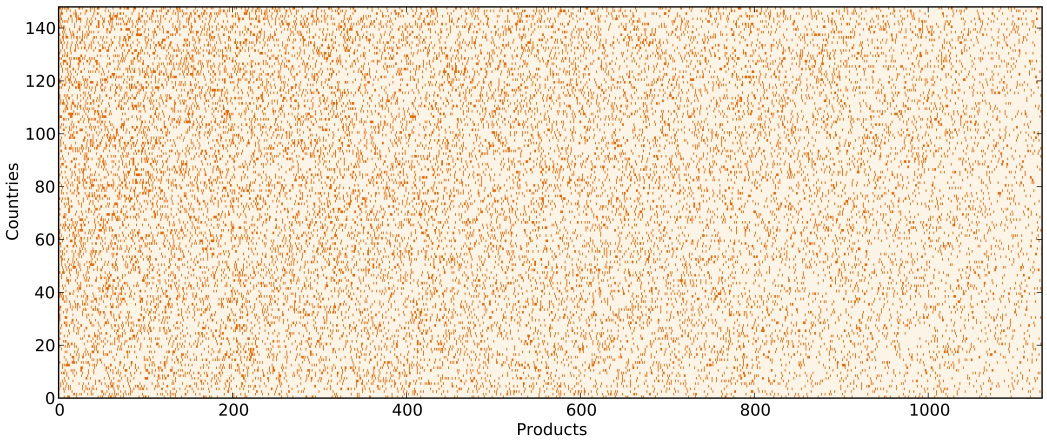


Figure 2: Graphical representation of an artificial Mcp matrix with random binary entries (same number of entries of the matrix of Fig. 1) after reordering of rows and columns by respectively decreasing K_c and K_p. It is clear that even after such a reordering the matrix does not acquire a triangular structure as instead empirical data show.

from the standard view: as countries evolve becoming more and more complex, they acquire a higher degree of diversification rather than specialization. This marks a sort of analogy with the evolution of biological organisms in complex and varying ecosystems. The best adaptation is achieved when organisms can rely on a broad set of resources, rather than being dependent on very specific environmental conditions. In the same way diversified nations are not dependent on very specific market conditions. Moreover the structure of the matrix M suggests that the larger is the present basket of products for a given country the more likely will be in the future to make new and innovative products for it. This analogy will be investigated further in Chapter 3.

1.1.1 The theory of hidden capabilities

These observations about the information contents of the structure of the country-product matrix have motivated a series of recent works [3, 10, 11] aiming at going beyond the limits of the standard economic theories. In these articles the authors propose a new conceptual framework in order to explain how and why the increase of diversification of production and export is a manifestation of optimal strategies to keep and increase the economic wealth of a country in a complex and transforming economic environment. On the same ground, such an approach aims also at explaining why the country-product matrix is basically triangularly shaped.

The key point of this complexity approach is the following: each country is characterized by special fundamental endowments, called *capabilities*, which represent all the resources of the economy of the given country and the features of the national social organization making possible the production and the export of the basket of tradable goods by the same country. Capabilities are usually non-tradable goods and are very difficult to measure and compare from country to country (e.g. infrastructures, educational system, technological transfer, climate, geography, political stability). In other words, the capabilities are supposed to be all

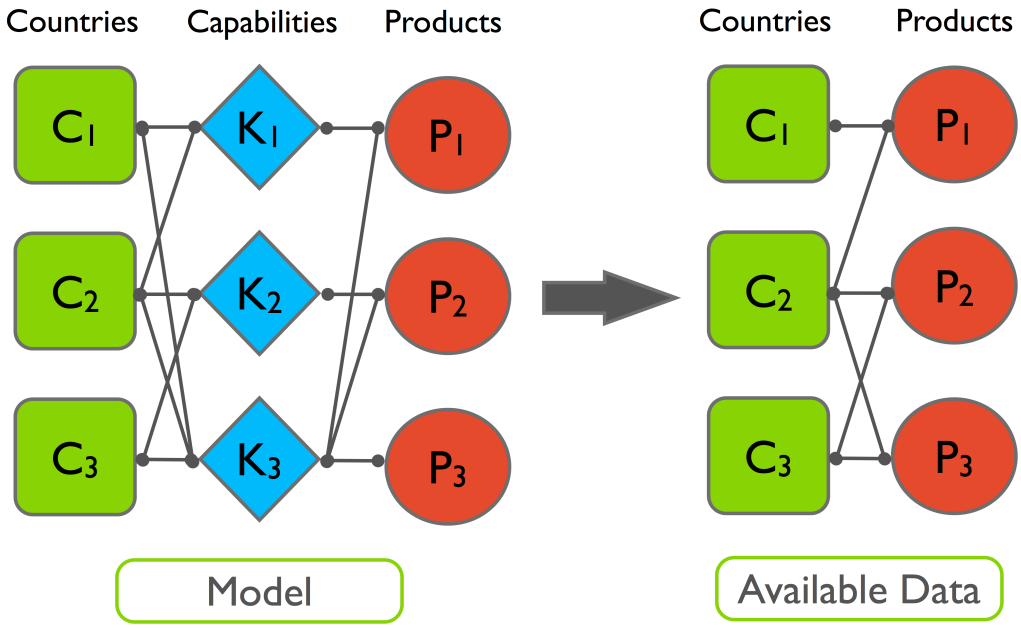


Figure 3: A schematic representation of the hidden capabilities layer. The real observable data is the contraction of the tripartite network Countries-Capabilities-Products: each country is connected to all and only those products for which owns all the necessary capabilities.

the intangibles assets which drive the development, the wealth and the competitiveness of a country. However, listing all the capabilities is impossible. Furthermore they vary enormously from country to country depending on political organization, history, geography etc. and we cannot define a universal standard measure for them.

In [3] HH consider them as the fundamental bricks behind the economy of each country determining their fitness to compete in the international market. In practice they determine the complexity of a productive system as each product requires a specific set of necessary capabilities which must be owned by a country in order to produce and then to export it. In this perspective, we can draw an analogy with biological systems: in an evolving economic environment for a country it is much more convenient to accumulate capabilities than specializing in a particular sector of production selecting and preserving only a limited and particular set of capabilities.

Due to the difficulty in categorizing, quantitatively analyzing and comparing capabilities, exported products by each country become in such a scenario the main proxy to infer the level of complexity of a productive system, that is the endowment of capabilities. In some sense the basket of exported products of a country contains encrypted information about its fundamental capabilities, i.e., the peculiar social and economic substrate on which the complexity of the national economic system is built.

It is possible in principle to represent schematically this conceptual framework in terms of a tripartite country-capability-product network in which capabilities are the intermediate layer between countries and products (see Fig. 3).

A tripartite network is in general a network in which nodes can be grouped into three classes ($\mathcal{C}$, $\mathcal{K}$ and $\mathcal{P}$ in the present case) such that links are permitted only between nodes belonging to two different classes. In the particular present case a node in the classes $\mathcal{C}$ (countries) and $\mathcal{P}$ (products) can only be connected to nodes in the class $\mathcal{K}$ (capabilities). The fact that we cannot define and thus observe capabilities directly, means that we can only access to the "contraction" of this tripartite network into the bipartite country-product network which is an equivalent description of the binary export matrix M. We argue that the from the structure of this contracted network we can extract information about the hidden layer. This information, as we will show in the rest of this work, can be highly predictive about future developments in countries industrial diversification and economic performance.

We can put these relations and structure in formulas to properly highlight the strongly non-linear relationship between capabilities and diversification of the production basket(see also [12]). Here we discuss the simplest modeling of the tripartite network, namely the one with maximally entropic random links. In chapter 3 we will discuss more general implementations of this scheme and analyze how they can relate to some non-trivial stylized facts present in a wide class of bipartite networks, not only in economics, but more in general in biological competitive ecosystems. Anyway the simple implementation that we anticipate here is able to provide strong indications towards the need of non-linearity in the metrics.

Let us call $\mathcal{C} \equiv \{c\}$ the set of countries, $\mathcal{K} \equiv \{k\}$ the set of capabilities, and $\mathcal{P} \equiv \{p\}$ the set of products. We can define the following two binary matrices:

- $\hat{S}$ connecting countries to capabilities, whose element S_{ck} is 1 if the country c owns the capability k and 0 otherwise. The c^{th} row of this matrix provides in this way the whole set of capabilities owned by country c, while the k^{th} column gives the set of countries having capability k.
- $\hat{T}$ connecting capabilities to products, whose element T_{kp} is 1 if the capability k is a necessary "ingredient" to produce the product p. The p^{th} column of this matrix gives all the necessary capabilities to produce and export p. The k^{th} row gives instead the set of products for which capability k is a fundamental ingredient.

A product is exported by a country only if it owns all the necessary capabilities to produce the given product. We can consequently define the matrix M as

$$M_{cp} = \prod_{k} [1 - T_{kp}(1 - S_{ck})] \tag{5}$$

which is 1 *iif* c owns all the capabilities to produce p and 0 otherwise. It is important to note the high non-linearity of the relation (5), which implies that the acquisition of a new capability k by a country produces an effect which strongly depends on the basket of capabilities already owned by country c, and therefore by the basket of products that such a country already exports. This can be illustrated by the following approximated argument. Let us assume that the country c acquires the capability k_0, so that S_{ck_0} switches from 0 to 1. The impact on the basket of exports of country c will be given by the difference δk_c of $k_c = \sum_p M_{cp}$ after and before the acquisition of the capability k_0. It is simple to show that

$$\delta k_c = \sum_{p} T_{k_0 p} \prod_{k}^{\neq k_0} [1 - T_{kp}(1 - S_{ck})] = \sum_{p} T_{k_0 p} \prod_{\{k\}_p}^{\neq k_0} S_{ck}\,, \tag{6}$$

where $\{k\}_p$ indicates the set of capabilities necessary to produce the product p. Let us see what happens in the case in which all the entries in the matrix T_{kp} are independent identically distributed binary random variables with mean $q \in (0,1)$. In this case, taking the average of the second expression in (6) we can say that

$$\overline{\delta k_c} \simeq q\,\overline{k_c}\,, \tag{7}$$

where the average is taken over the possible values of T_{kp}. This simple calculation shows that even in a maximally random case the higher is the number of capabilities owned by the country c, and therefore k_c, the higher will be the average advantage in productivity and export by the acquisition of a new capability. This suggests that if a country owns a small amount of capabilities, and therefore a small basket of "simple" (i.e. requiring only few capabilities owned by almost all countries) products, it is almost impossible for such a country to improve its economic performance in the international trade of products by a simple "step by step" acquisition of new capabilities. This is instead, by the simple combinatorial argument behind Eqs. (6) and (7), an efficient way of evolving the economic system in order to keep the good performance for rich and "complex" countries (i.e. owning already many capabilities and consequently exporting many different products from simple to complex ones). This would indicate a difference in the evolution of economies of respectively developing countries, which are rapidly increasing the basket of exports, and already developed countries which are already in the set of top exporters. While countries in the first group are expected to rapidly accumulate known capabilities already owned by the best exporters, for top countries, with already advanced economies, one should observe a slower step by step addition of new and more and more complex capabilities with a high impact on the economy, basically by developing new technologies. One could also conclude that poorly diversified countries can only improve their situation by a radical change of economic/political system and not by slow acquisition of new capabilities. This heterogeneity in the developing paths of countries with different level of fitness is indeed observed and will be thoroughly discussed and quantified in Chapter 2.

1.2 RESULTS AND DISCUSSION

1.2.1 New metrics from a non-linear algorithm: motivations and mathematics

Previous sections suggest that there is a strongly non-linear entanglement between the competitiveness (in terms of owned capabilities) of a country and the complexity of its products and that this non-linear relation is strongly related to the set of capabilities that the country owns, i.e. to the "complexity" of its economic/political organization. In order to translate into appropriate mathematical form this entanglement we introduce an iterative non-linear algorithm. The reasons underlying such an iterative approach is that we are looking for a self-consistent complexity measure based on the empirical country-product matrix. As we are going to see, this self-consistent metrics can be found and is given by the unique fixed point of the method we propose. Being the fixed point non-trivial and corresponding to the only attractor of the coupled equations, iterating is an effective strategy to determine the fixed point.

On such a basis we propose (see [5]) and study below an iterative algorithm able to capture efficiently the intrinsic link between the export basket of different countries, the complexity of products and implicitly the set of owned capabilities.

In order to formulate such an iterative algorithm, we start from the simple aforementioned observations on the relation between diversification of countries and ubiquity of products. Ubiquitous products, in the "capabilities" picture, should have a low degree of complexity requiring only a small amount of capabilities to be produced so that even countries with few simple capabilities can produce them. On the other side, most exclusive products are exported only by the most diversified countries. The most diversified countries show in this way to own so many capabilities to be able to produce a large variety of goods from very simple (i.e. low quality/value, requiring few capabilities) to very complex (i.e., high quality/value requiring the *ad hoc* mix of many advanced capabilities).

Let us consider, in the light of the triangular structure of the matrix M, the importance of the following pieces of information: (i) a randomly chosen product is produced by a diversified country; (ii) a randomly chosen product is produced by a poorly diversified country; (iii) a randomly chosen country produces a widely diffused product (i.e. simple product); (iv) a randomly chosen country produces an exclusive or non-ubiquitous product (i.e. complex product).

Since diversified countries are expected to produce a large fraction of all products from very simple to very complex, information (i) does not give any insight into the quality/complexity of the product. On the contrary, information (ii) is very important. Indeed, due to the triangular shape of M, the fact that a product is exported by a poorly diversified (and presumably scarcely differentiated in the spirit of capabilities) country makes very likely that this product has a low complexity, requiring few common capabilities to be produced. In a similar way information (iii) is completely irrelevant to determine the quality (i.e. economic development) of the country, as ubiquitous products are exported by definition by most of countries and presumably requires few and simple capabilities to be produced. Instead situation (iv) is very informative on the quality of the country as the triangularity of the matrix M implies that almost only highly diversified and presumably developed countries can export un-ubiquitous products.

All these observations suggest a non-linear and quasi-extremal relation between the complexity of an exported good and the competitiveness of its producers. In particular, in order to predict the quality of a product, it is much more informative to know if among its exporters there are poorly diversified and presumably non-competitive countries than knowing the mean quality of all producers as it happens in the HH method. On the other side the sum of the complexities of the exports of a country is expected to be a good tracer of its competitiveness in the global market. Indeed this sum is expected to increase with the development of a country, i.e. with the basket of its capabilities. The need of a non-linear relation is also strongly suggested in [7] by exploring the possibility of ranking countries and products through a linear algorithm obtained by generalizing the PageRank method [13] to the case of the country-product bipartite network in which the presence of asymmetric biases is permitted. The need of strong non-linear biases warmly suggests to move directly to a non-linear approach.

We can therefore introduce a non-linear relation, based on the structure of the matrix M, relating the quality and complexity of products Q_p^* to the fitness (i.e. competitiveness and development) of countries F_c^* (in [3] an iterative scheme is proposed too; however, as will be discussed in Sections 1.2.3 and 1.2.4, we argue that this method suffers from several mathematical and conceptual problems and is conceptually different from the present approach).

The precise definition of the algorithm is based on the introduction of two sets of variables $\{F_c^{(n)}\}$ and $\{Q_p^{(n)}\}$ measuring respectively the estimate of the Fitness of all countries $\{c\}$ and the

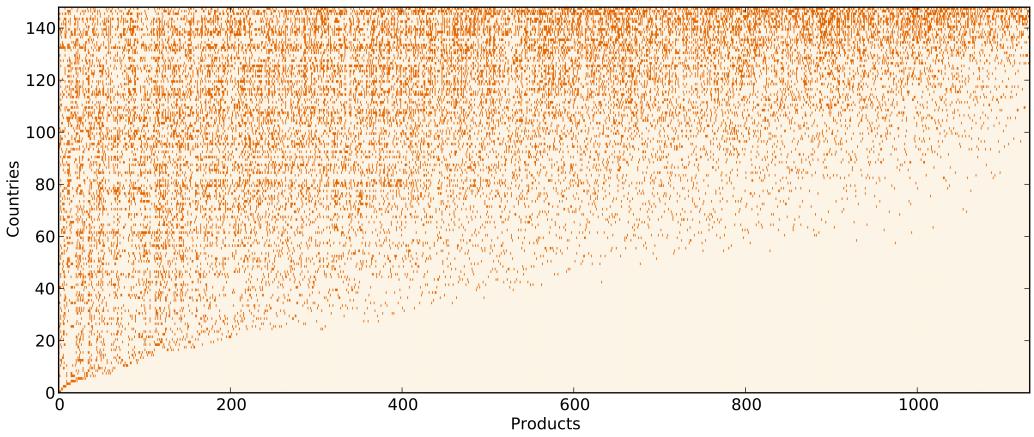


Figure 4: Graphical representation of the experimental M_{cp} matrix for the year 2010 after reordering of rows and columns by respectively decreasing F_c^* and increasing Q_p^*. It is evident the substantial triangular structure of the matrix even more pronounced than in the case of a reordering of rows and columns in terms of k_c and k_p.

Complexity of all products $\{p\}$ after n iterations. The algorithm is defined by the following formulas reflecting the essence of the above considerations. We first compute the intermediate variables $\tilde{F}_c^{(n)}$ and $\tilde{Q}_p^{(n)}$ and then normalize them so that to have a standard measure of these properties:

$$\left\{\begin{array}{rcl} \tilde{F}_c^{(n)} & = & \sum_p M_{cp} Q_p^{(n-1)} \\ \tilde{Q}_p^{(n)} & = & \dfrac{1}{\sum_c M_{cp} \dfrac{1}{F_c^{(n-1)}}} \end{array}\right. \rightarrow \left\{\begin{array}{l} F_c^{(n)} = \dfrac{\tilde{F}_c^{(n)}}{\langle \tilde{F}_c^{(n)} \rangle_c} \\ Q_p^{(n)} = \dfrac{\tilde{Q}_p^{(n)}}{\langle \tilde{Q}_p^{(n)} \rangle_p} \end{array}\right. \tag{8}$$

with the initial conditions $\tilde{Q}_p^{(0)} = 1 \; \forall p$ and $\tilde{F}_c^{(0)} = 1 \; \forall c$. The normalization can be interpreted as the fact that the total amount of Fitness and Complexity is conserved through the iteration. Mathematically it corresponds to constraining the dynamics into an highly dimensional simplex. The iteration converges after a few tens of step to a stable fixed point, $\{F_c^*\}$ and $\{Q_p^*\}$, which defines the metrics.

The main idea is, as aforementioned, that while the fitness of a country is indeed defined by the sum of the complexities of its products, the complexity of a product is bounded by the development of the poorly diversified producers. This idea originates from the triangular structure (as shown in Fig. 4 where we ordered countries according to the fitness we compute) of the country-product matrix M.

Note that Eqs. (8) can be seen as a mathematical realization of economic concepts about the relation between the complexity of products and developments of countries. As we show below, this non-linear method uncovers the hidden capability distribution of countries; indeed the ranking and metrics of countries and products, as given by the fixed point of Eq. (8), well describe the complexity of the economic status of countries and the complexity of products.

Unweighted vs. weighted algorithm

In Eq. (8) we can use as matrix M both the binary (unweighted) one defined in Eq. (1) and the weighted one defined in Eq. (2). Clearly using the former or the latter will give different quantitative information, even if partial and qualitatively overlapping features are present.

The choice of using the unweighted and binary version of the country-product matrix is motivated by the following consideration: we believe that it represents better than the weighted one the potential of growth of a country. For instance, if an emerging country starts the export of a new product, the information about the export given by switching M_{cp} from 0 to 1 is more important in many aspects, for the evolution of that economy, than to know the volume of the export.

On the other side, the approach based on the weighted matrix determines the effect of the information about the relative importance of the different exporters of the same tradable good. In this way it can, for instance, better detect most influent countries in the global market dynamics in different product sectors. As mentioned in Sect. there are in principle different possible choices for the weights in the matrix M_{cp}.

A first possible attempt towards an extensive generalization of our metrics is represented by the direct use of the RCA matrix which is the matrix defined by the RCA coefficients. However, such RCA coefficients suffer from a number of disadvantages. In fact in order to measure a very large RCA (> 100), a country typically must own a very large share of the export of a product and, at the same time, this product must have a much lower average share of the world wealth. This usually happens for exporters of natural resources (especially raw materials such as crude oils, metals, coal, etc.) which are in general characterized by a small diversification. As examples of such a phenomenon, we can list Chile which owns about 30% of copper export and Saudi Arabia for crude oil. On the other hand most diversified countries, which include the richest and most advanced countries, on average are characterized by a more homogeneous set of RCA values which appear to be not dominated by a single product. In this way, the choice of RCA coefficients for weighting M_{cp} would favor those countries with a low diversification which, by chance, have a large amount of natural resources. For such reasons RCA has been discarded for a weighted version of our method.

It is much more reasonable and effective to define a weighted country-product matrix as in Eq. (2). This is a direct generalization of the binary M matrix where the entries of the matrix can assume a value ranging from 0 to 1. We want to stress that the definition adopted is still an *intensive* version of the matrix M from the product point of view. Indeed, given a product, each exporter of this product is weighted according to the owned share of that product, however the sum over all exporters of each products is normalized to one. That is, products are considered intensively. In other words we are not taking into account the fact that different products have in general a different share of the global export.

The reasons for such a choice are twofold. We believe that the complexity of products is intrinsically independent on the total exported volume. In fact by keeping products as an intensive quantity we are still able to filter purely *monetary* effects linked to market prices, price inefficiencies, raw materials value, out of our method. At the same time, fixed a product, we can still consider the scale of each country which export the product.

As a final remark, it has to be observed that such weighted metrics behaves as an extensive economic indicator (for instance the total GDP of a country), but it does not trivially coincide with the GDP information. Similarly the binary/unweighted case follows the behavior of a *per capita* indicator as shown in Fig. 5. Starting from this observation we indicate from now on as

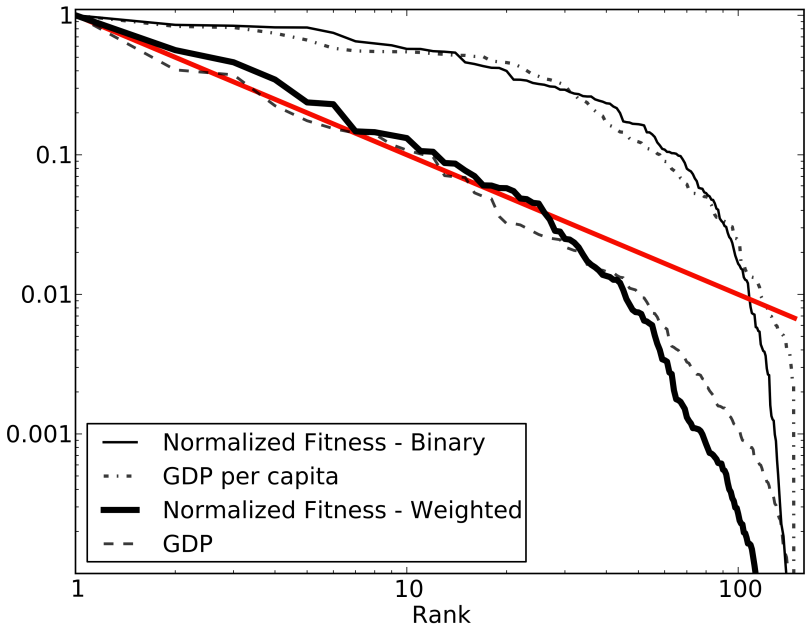


Figure 5: The weighted metrics behaves as an extensive economic indicator (for instance the total GDP of a country), but it does not trivially coincide with the monetary information. Similarly the binary/unweighted case follows the behavior of a per capita indicator, in that case the GDP per capita.

intensive fitness the one resulting from the unweighted matrix and as *extensive* the one from the weighted case. The intensive/extensive feature must be only referred to their different economic behavior as discussed in Fig. 5. There is no reference to the properties of the matrix adopted to estimate the two metrics.

1.2.2 Economic implications of the metrics

Country analysis: BRIC and PIIGS countries:

Different economic analyses can be carried out in the framework of our approach. In this section we propose some relevant results to show the potential applications. On one hand the two metrics introduced in the method for ranking countries and products by themselves can provide important and new information on the analysis of the growth of countries. Here we analyze the correlations of the value of fitness with the outcome of traditional economic analysis. In Chapter 2 we will show how the deviations from the assesments of standard indicators (GDP per capita) triggers dynamics that are even more informative.

The importance of the Complexity of products will be discussed at the end of this section and then further analyzed in Chapter 4.

Standard economics consider BRIC countries, namely Brazil, Russia, India and China, as emerging economic systems which have a high rate of growth. These four countries are considered similar from a GDP point of view, i.e. in respect of their GDP growth rate. However, following the indications that arise from the Fitness calculations, we can argue that from a

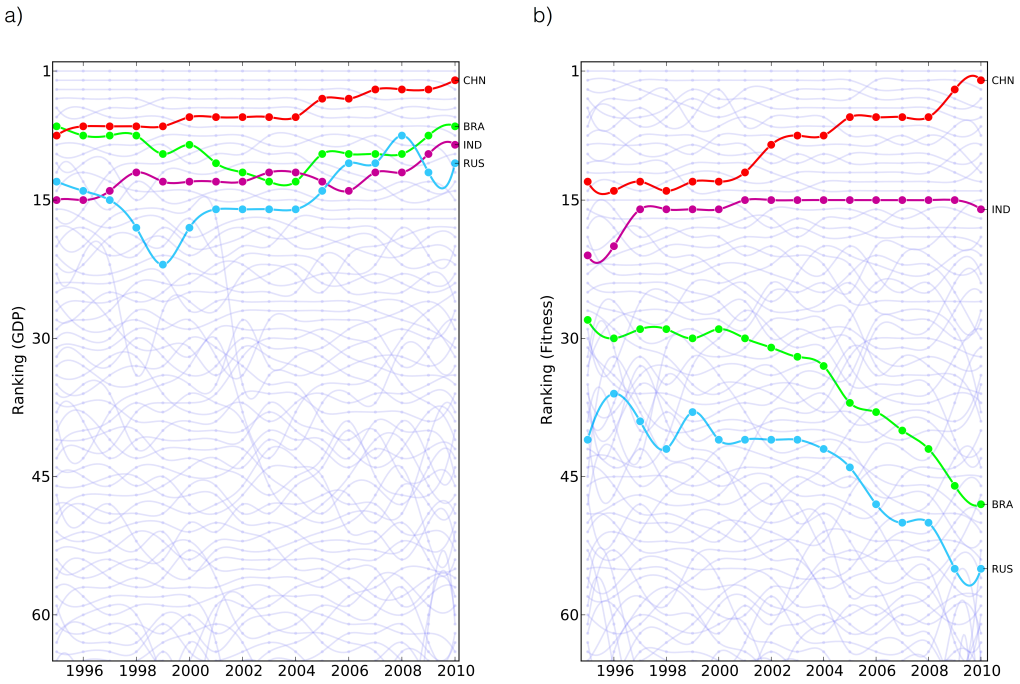


Figure 6: Fundamental analysis of the BRIC countries according to our metrics. We argue that India and China undergo a real economic development characterized by accumulation of new and more and more complex capabilities. Therefore the GDP growth corresponds to a real increase of the competitiveness of these two countries. Conversely we observe that the GDP growth of Brazil and Russia appears to be mainly fueled by the price bubble of the raw material sector and these countries are not using these extra richness to develop and accumulate new capabilities in order to settle a solid basis to their productive system.

fundamental point of view these four countries undergo a very different development: while India and China appears to have a well-grounded economic development characterized by a complex basket of exports, it is not the case for Brazil and especially Russia. In fact as shown in Fig. 6 panel a, by analyzing BRIC countries in standard GDP terms, we find that in the last fifteen years all these countries appear very similar and are characterized by high rate of growth of their GDP (mostly above the world growth rate). However, looking at panel b of the same figure, our metrics reveals a strong heterogeneity among these four countries which a conventional analysis is not able to capture. The evolution of the fitness, which as aforementioned we interpret as the degree of competitiveness of a productive system, reveals that, while India and especially China have strongly increased their competitiveness in the global economic playground, Brazil and in particular Russia, despite a growing GDP, have lost many positions according to the fitness ranking. The economic interpretation of such difference, on the basis of our metrics, is the following:

(1) India and China (*IC*) reflects a genuine economic and industrial development characterized by accumulation of new, more and more complex capabilities. Therefore the GDP growth corresponds to a real increase of the competitiveness of these two countries.

(11) Brazil and Russia (*BR*) are very important raw material exporters. We therefore argue that their GDP growth is mainly fueled by the price bubble which characterizes this sector. In this sense we interpret the decreasing competitiveness of Brazil and Russia in terms of the fact that they are not using their extra richness deriving from raw materials to develop and accumulate new capabilities in order to settle a solid industrial and technological basis to their productive system.

It is worth noticing that the idea that Brazil's GDP growth is mainly depending on commodities is becoming popular only in the last three years and the consensus on such feature is not at all uniform (see Ref. [14] and [15] as examples of two different points of view on Brazil). If one would have used the new metrics one could have seen a significant loss of complexity of Brazil economic system years in advance. In fact from 2002 there is a clear and steady decrease of the Fitness Ranking of Brazil. This anticipation of the trend is a characteristic of this innovative methodology which measures the hidden potential and not just the present status. We argue that the situation for Russia is somewhat similar. We can therefore conclude that the development of IC countries is well-grounded from a productive point of view differently from BR countries. We believe that the most interesting result concerns Brazil, indeed its growth is usually considered of the same kind of the one of India, China and other emerging Asian countries (e.g. Vietnam, Thailand, etc). Our analysis implies instead the opposite, Brazil growth is closer to the Russian case where the development is dominated by the market price of fossil fuels. We are aware that there may also exist macro-economic and political reasons that could determine lower export for a country given a level of capabilities and therefore our method would measure a lower level of competitiveness than what expected. In fact in the case of Brazil, besides being an important raw material exporters, there exists a strong state planning of the production, sectoral incentives and a strong boost of internal production against exports. However, the great advantage of our fundamental analysis with respect to conventional ones consists in clear quantitative statements that can be extensively tested

Let us now consider a different set of countries, the so-called PIIGS, i.e. Portugal, Italy, Ireland Greece and Spain. They are European developed countries which are usually considered the most fragile economies from a financial point of view among European Union. Indeed, the rating of PIIGS' sovereign debt is on average lower than the other members of EU.
Let us move to the analysis of fitness evolution for the PIIGS as shown in Fig. 7 (as a benchmark of a non-PIIGS we choose Netherlands).
The fundamental analysis of the competitiveness points out a scenario in which Greece, Portugal have an increasing fitness, Spain and Italy a stable competitiveness ranking (and a behavior very similar to Netherlands) and Italy is even always ranked in the top 5 position, very close to the level of Germany. In addition Spain, Portugal and Italy in 2010 are above the average world fitness ($\langle F \rangle = 1$). We want to recall that we are considering the intensive metrics which measures the intrinsic level of complexity that each country has developed. We stress that in the weighted analysis Italy is well below China as expected. Only Ireland exhibits a decreasing fitness in the intensive scenario. We also report the evolution of Iceland's fitness as a prototype of a developed non-PIIGS country which has gotten in big financial troubles in the last decade.

The reasons for this apparent discrepancy between standard rating or evaluation of these countries and our results is twofold:

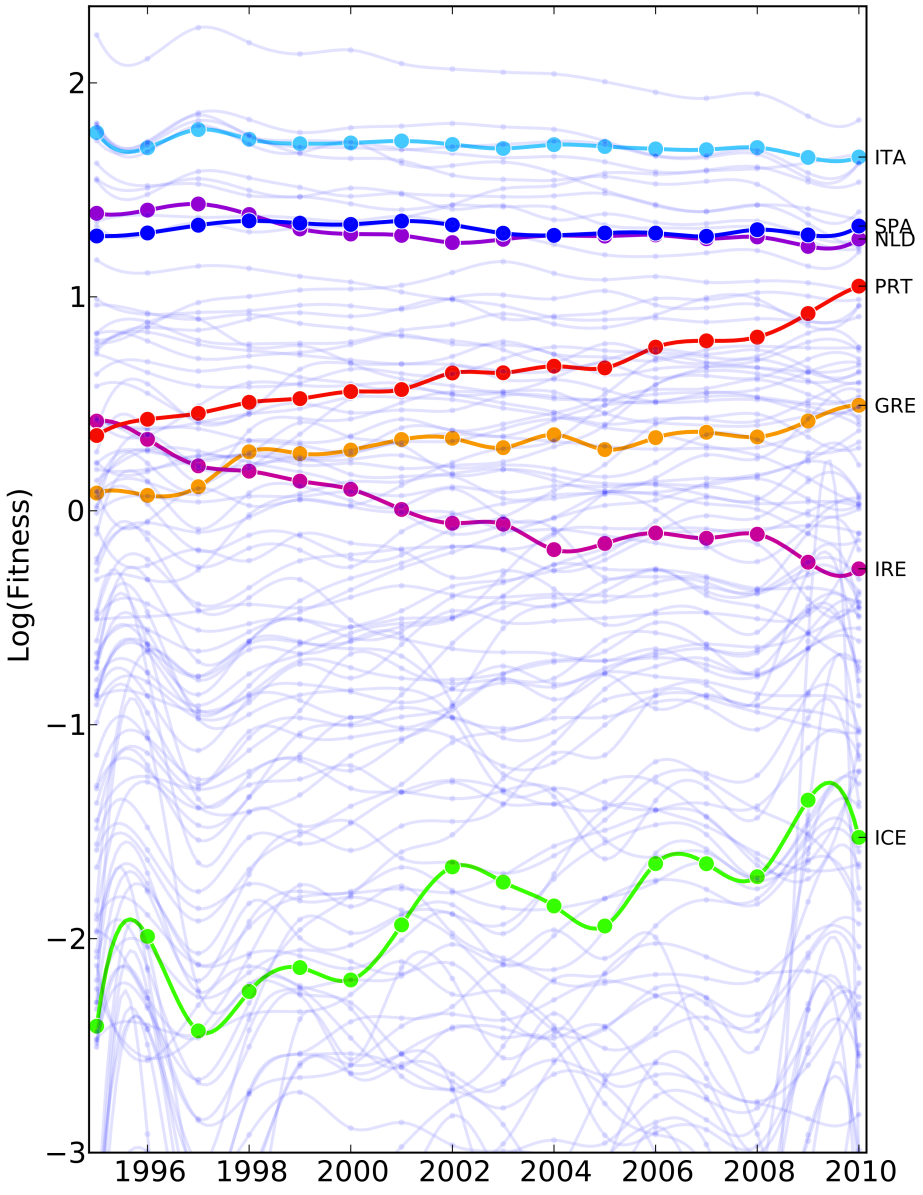


Figure 7: Fundamental analysis of the PIIGS countries (Portugal, Italy, Ireland Greece and Spain) according to our metrics. We find a scenario which seems to be apparently in contrast with the rating of the sovereign debt of these countries. For instance we find that Greece, Portugal have an increasing fitness and Italy is always ranked in the top 5 positions along the time period considered. The main reason of this apparent discrepancy, in our opinion, relies on the fact there exists different regimes for the economic complexity.

(1) on one side it seems that the main source of the fragility of PIIGS countries has only a financial origin independently on the competitiveness of the productive systems, except Ireland for which both analysis give similar results;

(II) on the other side, the main reason, in our opinion, stands in the fact that different regimes exist for the dynamics of economic complexity, as will be discussed in Chapter 2

On this account it is clear that different factors concur to the economic development of a nation, both endogenous and exogenous: development of industrial capabilities indeed, but also international policies, wars, geo-political instabilities, the aggregate international demand, etc. In the present framework we develop a metrics to assess mainly the endogenous factors. As we will show in Chapter 2 the endogenous factors that the Fitness is able to summarize are a main driving force in some regimes such as the one of emerging countries, but this is not the case for developed ones. In fact PIIGS are all developed countries and somehow they almost saturated their phase space of capabilities: in fact these countries are among the most diversified ones, especially Italy, Portugal and Spain. In this sense they are in a completely different economic regime in respect of emerging countries. We therefore argue that the main driving force of the economic growth of developed countries is no more the fast development or acquisition of new capabilities and the following invasion of the product space (see Chap. 4). Instead in mature developed countries, politics, in particular economic ones, and in general non-capabilities driven exogenous features appear to dominate the growth and the evolution of these countries.We want to point out that this does not imply that the acquisition of new capabilities has no impact for this type of countries. Instead we believe they play a different economic role due the fact that they almost saturated the space of capabilities, hence the space of products. In fact the development of new, and generally of high technological value, capabilities in developed countries usually triggers bursts of new high complexity products on the market. However, these events tend to be rarer with respect to the acquisition of already established capabilities as it happens for emerging countries. We try to give some insights on the dynamics of this innovation process in Chapter 3

It is worth noticing that this second explanation calls for the concept of heterogeneity in economic growth dynamics and prediction. On this account the result discussed in 2 clearly points in this novel direction: the dynamics of the development of countries shows a high degree of heterogeneity, consequently a novel approach is required and new concepts like selective predictability must be considered.

Extensive vs Intensive Metrics

In section 1.2.1 we have introduced a generalization of our iterative method by considering suitable weights which partially take into account the export volumes. We now want to interpret, from an economic point of view, the kind of information carried by the two cases and spot the differences of the two analyses. Let us focus on some specific countries: Germany, China, Italy, USA, United Kingdom, Austria, India and Poland. In Fig. 8 we report the evolution of the intensive fitness (panel a) and of the extensive one (panel b). Focusing first our attention on Germany, China, Italy, USA, we find that the intensive fitness ranking does not reflect the traditional economic prediction. In fact Italy's fitness is higher than USA's one and almost equal to the China's one. In 2010 Italy is the most diversified country with respect to the export basket with more than 500 products (for which the RCA coefficient is above the threshold) and our metrics correctly grasp this feature. Once the weights are taken into account, we find instead (see also Table 1 for details) a ranking closer to the one provided by GDP even if significant differences persist. For instance, from a GDP-oriented analysis China results to be the 2nd-3rd economic power, in our framework, China is already the most competitive country in extensive

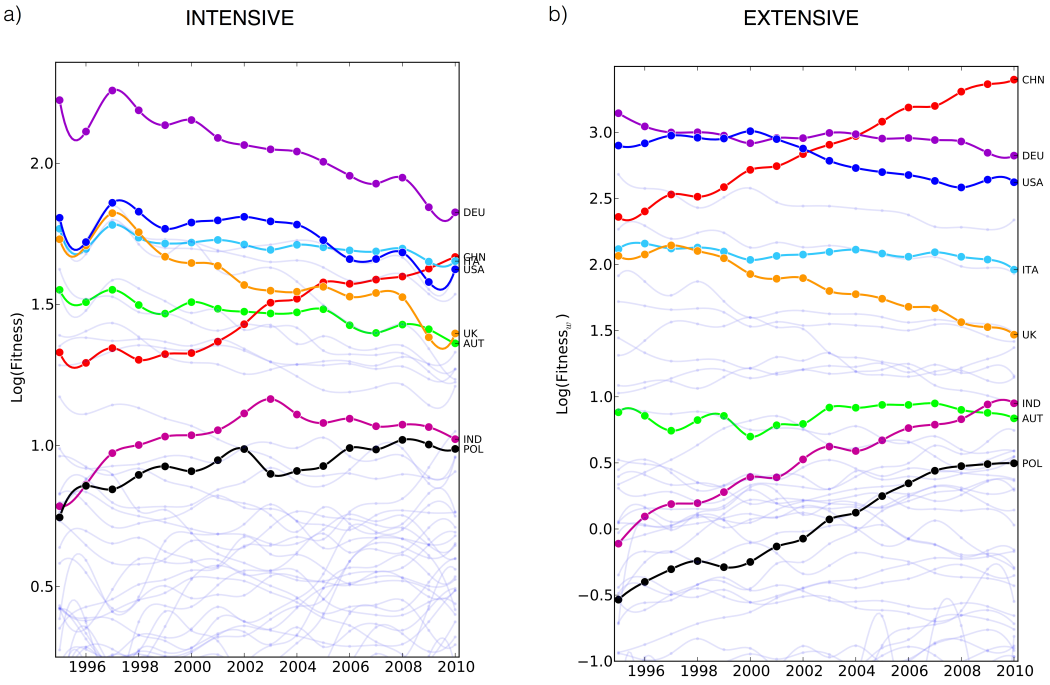


Figure 8: Economic interpretation of evolution of the fitness in the intensive and extensive case. The intensive fitness gives a medium-long term information of the development of countries, in this sense, we can consider it as informative on the growth potential of a country. On the other hand the extensive analysis complements the information carried by the intensive fitness conveying a short term perspectives and giving a stronger emphasis to the monetary aspects.

terms.

As a second point let us compare the two pairs United Kingdom-Austria and India-Poland. From an intensive point of view these pairs appear almost degenerate while extensively we observe that, as expected, bigger countries in both pairs have larger fitness. However, it is worth noticing that even if we consider the export volumes, the weighted fitness does not simply reproduce the GDP ranking or the relative monetary distance among these countries: the fitness ratio of two countries is not trivially the ratio of their GDP. In some sense, intensive analysis is able to spot *niche* of competitiveness, while extensive metrics moves the focus of the analysis to the scale of the economic system.

We argue that the intensive fitness conveys long-term information of the competitiveness of a country. The intensive metrics is a measure of potential of growth and somehow a measure of resilience and recovery features of economic systems (especially for developed countries). In this sense the results of Italy in the top 3 position of the intensive fitness ranking is not surprising since historically Italy is known as a very resilient system. In the light of our fundamental analysis and neglecting specific economic policies and exogenous aspect (which could become dominant as discussed in the previous section and may enhance or contract the recovery from the recent global crisis), Italian productive system has an intrinsic strength

Table 1: Countries' Fitnesses

Country	Int. Ranking	Int. Fitness	Ext. Ranking	Ext. Fitness	GDP_{2010} (bill. of US$)
Germany	1	6.21	2	16.84	3400
China	2	5.30	1	29.92	5800
Italy	3	5.23	5	7.11	2100
USA	5	5.08	3	13.77	14600
UK	7	4.04	8	4.35	2150
Austria	8	3.90	15	2.31	380
India	16	2.78	14	2.59	1700
Poland	17	2.69	21	1.64	470

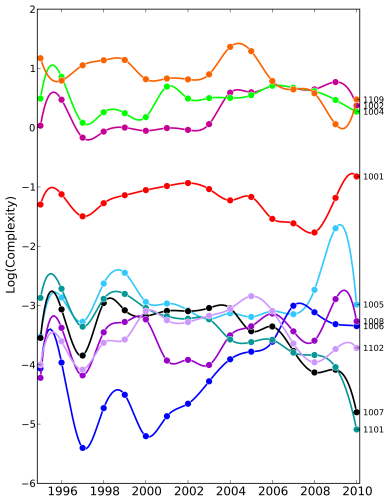
Intensive and extensive fitness for a selection of countries.

and recovery capacity, much higher that other European countries, say Spain, Ireland, Greece, which is probably connected to its diversity.

We point out once again, that our metrics provides undoubtedly new information (for instance the Brazil analysis), but the novelty of our method relies on the fact that it gives a quantitative assessment which can be tested with respect to standard economic indicators. On the other hand the weighted fitness complements the information carried by the intensive fitness since it gives a present and short term perspectives of the country analysis giving a stronger emphasis to the monetary aspects. Russia and Brazil are paradigmatic cases in this sense. In a short term horizon or, more precisely, in the monetary horizon set by the availability of raw materials in these two countries, they are competitive (monetary information) but in terms of diversification, resilience, adaptability and, in general, competitiveness of their productive system (intensive information) they appear weak, or, at least, much weaker than other emerging countries.

Products

Similarly to countries, our method defines a metrics for the complexity of products. A part from the MR of HH, this is a completely novel measure because we are not aware on the existence of economic indicators for the complexity of product which do not rely on monetary estimate. In fact a standard measure adopted is the market value of products, however, this quantity suffers from strong bias due to market speculation, labor cost, etc. While it is reasonable to believe that products characterized by a high complexity are likely to have high market prices, it is very easy to find striking counterexamples where *simple* products have anomalously high price, for instance the *Tulip mania* of XVII century. Therefore we propose the Complexity of products as a new synthetic indicator which permits to quantitatively assess the complexity of products in a non-monetary and non-market oriented way.
In this respect a large spectrum of analysis can be performed: detailed analysis of the export basket of countries, relative strength/weakness of countries with respect to export of specific products, indices to quantify the complexity of economic sectors, etc. In addition, in analogy to the evolution of country fitness, it is possible to investigate the evolution of complexity of products year by year, in such a way, in principle, we may track the evolution of the economic cycles and the development or the technological contraction of specific sectors.



HS2007 code	Description of the product
1001	*Wheat and meslin*
1002	*Rye*
1004	*Oats*
1005	*Maize (corn)*
1006	*Rice*
1007	*Grain sorghum*
1008	*Buckwheat, millet and canary seed; other cereals*
1101	*Wheat or meslin flour*
1102	*Cereal flours other than of wheat or meslin*
1109	*Wheat gluten, whether or not dried*

Figure 9: Time evolution of the product complexity from 1995 to 2010 for a selection of cereals which result to be organized into two main groups. The former group has an average complexity around the average complexity of all products, $Q \sim 1$, the latter one is composed of cereals whose level of sophistication is much lower than the previous as measured by our metrics, $Q \sim 10^{-3}, 10^{-4}$). By analyzing the typical typical usage of oats and rye we find that these two cereals are not typical of a substance economic system since they are used in livestock industry and brewed-product industry.

As an example, in Fig. 9 we show the time evolution of the complexity for a selection of cereals from 1995 to 2010. Cereals result to be organized into two main groups: the former has an average complexity around the average complexity of all products (i.e. $Q \sim 1$), while the latter is formed of cereals whose level of sophistication is much lower than the previous as measured by our metrics (i.e. $Q \sim 10^{-3}, 10^{-4}$). Given this observation, among cereals, our method reveals two different complexity regimes for cultivation. In order to verify if the two classes correspond to a real difference in the level of technology of the country exporting them we analyze the typical usage of oats and rye. Supporting the finding that these two cereals are not typical of a substance economic system, we find that they are used in livestock industry and brewed-product industry.

In general, the time evolution of the product complexity must be carefully analyzed because of the specific structure of the non-linear coupled maps defining the metrics. In fact, while the country fitness is very robust with respect to errors in the database, the complexity is very sensitive to changes of the exporters of a given product, especially when the variations are due to low-fitness countries. On one hand we verified that the cleaning procedure of data is able to fix the wide anomalous oscillations of several orders of magnitude of some product complexity due to wrong custom reports - especially from small African countries. On the other hand, on average, the complexity of products shows an intrinsically higher degree of volatility with respect to fitness of countries even in a perfect dataset. In fact, given the economic assumptions underlying the metrics, a new (real and not due to errors) exporter can produce a significant variation of the complexity of a product while the addition of a product to the export basket of a country very likely will have a small effect on its fitness.
A hand-waving argument for this aspect is obtained by simply observing that since the fitness

of a country is given by the sum of the complexities of its products, if we assume that products have the same degree of volatility of their complexity and are statistically independent, the volatility of the fitness of the country will by roughly $\sqrt{k_c}$ times smaller. The opposite is not true because the complexity of a product is not at all the sum of the fitnesses of its producers, but a highly non-linear combination of them. For instance, for high complexity products we expect that very likely a new exporter (i.e. producer) will have a lower fitness than the typical fitness of the exporters of that product and therefore a short term decrease of complexity is, on average, expected. In this sense we argue that general trends and cycles are the meaningful analysis rather than short term variations of the level of technology in the case of products.

As a final remark, It is worth noticing that the knowledge of the intrinsic value of a product (i.e. the complexity) is critical for goods like commodities which are subject to strong speculative bubbles and whose market prices, differently from stock prices, are affected by strong inefficiencies, for instance the agricultural sector and in particular cereals. A systematic analysis of the metrics for product complexity and the features of product space will be discussed in Chapter 4.

1.2.3 Critical analysis of the state of the art (the Method of Reflections).

In [3, 11] the authors have tried to obtain a measure of competitiveness of countries and of products from the binary matrix M by introducing an iterative linear algorithm very different from ours, called *Method of Reflections* (MR). Through this method the authors rank countries and products in the international market and measure the difference in in their competitiveness by using only the information contained in the country-product matrix M. However, as shown below, the MR leads to very different results with respect to our approach and is affected by a series of conceptual problems. In this section we give a short *resumé* of this approach in order to make clear the mathematical and theoretical flaws.

In the MR algorithm an infinite set of variables, iteratively related, $\{k_c^{(n)}\}$ and $\{k_p^{(n)}\}$ with $n = 0, 1, 2, ...$ are introduced respectively for each country c and for each product p so that the information is considered more and more refined at increasing order n. At zero order the values are fixed by the initial condition $k_c^{(0)} \equiv k_c$ (diversification of c) and $k_p^{(0)} \equiv k_p$ (ubiquity of p) defined in Eqs. (3) and (4). In agreement with the previously exposed theory of capabilities, k_c has to be considered a first rough measure of the competitiveness of country c, as it is assumed that a large diversification corresponds roughly to the development and storage of a large set of capabilities. In an analogous way k_p provides a rough measure of the "dis-value" of product p, as in principle a very ubiquitous products will require a small number of capabilities to be exported reflecting a low level of economic complexity behind its production.

At higher orders $k_c^{(n)}$ and $k_p^{(n)}$ are defined by the following iterative equations:

$$\begin{cases} k_c^{(n+1)} = \frac{1}{k_c} \sum_{p=1}^{N_p} M_{cp} k_p^{(n)} = \left\langle k_p^{(n)} \right\rangle_c \\ k_p^{(n+1)} = \frac{1}{k_p} \sum_{c=1}^{N_c} M_{cp} k_c^{(n)} = \left\langle k_c^{(n)} \right\rangle_p \end{cases}, \tag{9}$$

where $\left\langle k_p^{(n)} \right\rangle_c$ means the arithmetic average of $k_p^{(n)}$ for all products exported by country c and $\left\langle k_c^{(n)} \right\rangle_p$ the arithmetic average of $k_c^{(n)}$ for all countries c exporting the product p. In

the idea of the authors of [3, 11] these equations define the iterations to a higher level of non-monetary and trade related information about countries and products leading to a better and better description of the competition in the global trade market. However, as we show below, this algorithm suffers of different important flaws which led us to introduce other iterative observables and a non-linear iteration algorithm which is better founded both mathematically and conceptually, and leads to a deeper comprehension of the international competition in the export market.

In the following we discuss in a schematic way the conceptual and mathematical flaws of the HH scheme.

Conceptual and mathematical problems

- We observe that the nature of the k_c and k_p completely changes from the starting iteration to the following one: while the starting point of the iteration is extensive in the number of products and countries, the following order are intensive with respect to products and countries because of the average considered. This fact derives from the expression

$$k_c^{(2)} = \left\langle k_p^{(1)} \right\rangle_c \tag{10}$$

  considering that $\left\langle k_p^{(1)} \right\rangle_c$ is the mean diversification of the countries exporting all products pexported by country c which therefore is a first order measure of the complexity of the product. Equation (10) makes clear a fundamental difference between our non-linear algorithm and the MR; it basically states that at first order the successfulness of a country is given by the *average* of the "complexity" of its products. This is very different from Eq. (8) for which instead the fitness of a country is given by the *sum* of the "complexity" of its products. This implies that, while in the MR two countries having the same mean complexity of the exports are supposed to have the same competitiveness independently of the relative diversification, in our method both the mean complexity of products and the diversification are, as natural, important in determining the fitness of a country in the global competition. Let us make an example to make this crucial point clear. In the HH scheme, paradoxically, a greatly diversified country (say about 500 products given a total of about 1000 products) with average complexity of its export set equal to $\tilde{k}$, whatever is $\tilde{k}$, would have the same competitiveness at the following iteration step of a country exporting only one product with $k_p = \tilde{k}$. Therefore the HH scheme is not consistent with respect to the assumptions underlying the capability arguments implying the importance of the concept of diversification.

- The highly non-linear (*quasi*-extremal) relation between competitiveness of countries and complexity of products, required by the triangular structure of the country-product matrix, cannot be implemented through an average as discussed by HH. As explained above, the triangularity of the matrix M implies that the information that some countries with small competitiveness (or development) export a product must bound the complexity from below, regardless of the competitiveness of the most developed exporters. Therefore one would expect a strongly non-linear and almost extremal relation between the complexity of a product and the competitiveness of the producers. Instead in the MR model at each order $2n+1$ the complexity of a product $k_p^{(2n+1)}$ is given basically by

the average of the $k_c^{(2n)}$ of its producers, so that the information about the most complex countries exporting this product is as important as the information about the less complex ones.

- As shown in the next section through an appropriate toy model, it is simple to see that the variables describing the competitiveness of countries in the MR rapidly loose correlation with the capabilities of the countries when iterated.

- The MR changes the economic meaning of the iteration at each iteration. It can be shown that $k_c^{(2n)}$ can be linearly related directly to $k_c^{(2n-2)}$ by substituting the second equation of (9) into the first one. A similar argument can be made for even k_p and for odd ones. However, it looks quite strange that in an *empirical* and *phenomenological* iterative approach to the ranking of countries and products the iterated quantities have different economic "dimensions" (averages of averages of ubiquities or diversification, respectively) depending on the parity of the order of iteration. Even iterations with the same parity change their economic meaning throughout the iteration procedure (as the number of averages increases). The economic and statistical interpretation of these quantities is rapidly lost when increasing the order n. In our framework the variables are simply the refinement of the ones of previous iteration and the iteration procedure has to be seen as an algorithm to solve the self-consistent fixed point equation.

- In [3] the authors consider at the end of the iterations for the economic analysis the rescaled quantity

$$\delta_c^{(2n)} = \frac{k_c^{(2n)} - \overline{k_c^{(2n)}}}{\sigma_c^{(2n)}}, \tag{11}$$

where $\overline{k_c^{(2n)}}$ is the arithmetic mean of $k_c^{(2n)}$ over all countries and $\sigma_c^{(2n)}$ is the standard deviation of $k_c^{(2n)}$ over the same set.

By using an algebraic approach, it is possible to show[7] that the MR makes all $k_c^{(2n)}$ to converge to the same constant k^* independent on the index c, which is therefore a trivial fixed point of the transformation relating $k_c^{(2n)}$ to $k_c^{(2n-2)}$. This is basically due to the fact that, writing in vectorial form these linear equations, the linear operator characterizing the linear transformation is the transposed of an ergodic Markov transition operator.

This explains why the authors of [3] subtract the mean value $\overline{k_c^{(2n)}}$ from $k_c^{(2n)}$ before any economic analysis. Indeed this accounts for the subtraction of the fixed point k^* from all $k_c^{(2n)}$. In a similar way it is possible to see that the division by the standard deviation $\sigma_c^{(2n)}$ to obtain $\delta_c^{(2n)}$ in Eq. (11) basically accounts for the contraction factor of the distribution of the set $\{k_c^{(2n)}\}$ around k^* at increasing order n due to the asymptotic convergence to such a single value for all c. The fact that the authors of the MR stop the analysis at $n = 18$ in [3] can be explained by the fact that this convergence is exponentially fast and at the value of n the numerical limits of resolution of different $k_c^{(2n)}$ are reached.

In an empirically defined algorithm the quantities involved in its formulation, and not to a vanishing component of them, should be directly related to observables.

Two critical issues emerge from this mathematical observation. On one hand the MR produces a shrinkage of the k_c and k_p distributions. Even if they are rescaled at the $n = 18$ iteration, the behavior of the algorithm is conceptually wrong because we would expect that differences among countries are in general magnified by one iteration step and not reduced. The reason of such expected magnification is that if we compare two similarly diversified countries but with different complexity of their export, once the information about the complexity of products is inserted in the method through the iterations, the distance between the variables measuring the successfulness of these to countries should increase.

On the other hand, the previous mathematical arguments show that the correct way to extract the rescaled k_c, k_p is to consider the eigenvector associated to the second largest eigenvalues of a fixed point equation (see [7]). Even if the method is presented as an iterative method, the HH complexity index (i.e. the k_c, k_p variables) cannot be self-consistently obtained iteratively in the form in which the MR is presented in [3] because their index is the eigenvector associated to the second largest eigenvalue of the transposed Markov operator [7], thus is vanishing when $n \to \infty$.

In summary it is possible to see that, extending the analysis in [7], the MR suffers of different critical aspects, which in our opinion make necessary a deep revision of the approach to the measure of the complexity of countries and products towards a non-linear approach. We recall that the non-linearity of the method, before even testing the metrics on economic benchmarks, is a key element to properly address the conceptual and economic consistency of a method based on the complexity/capabilities arguments which are intrinsically non-linear as extensively discussed in this chapter.

1.2.4 Comparison between our Metrics and the Method of Reflections.

In this section we provide a direct comparison of the results obtained with the non-linear Fitness-Complexity algorithm (FQ) we are proposing and from the MR method.

First of all in the next Subsection 1.2.4 we show, through the use of a minimal but significant toy model, that while in the MR method the correlations between the competitiveness of countries and their capabilities are rapidly lost when increasing the order of the iteration, with the FQ method they are kept constant at all order.

In the Subsection 1.2.4 we give a direct comparison of the ranking of countries coming out from the FQ method and the MR. In particular we highlight the most meaningful examples of the countries with a rapid economic development as eastern Asian countries and countries whose economy is basically determined, not by a development of advanced technological capabilities, but by the monopolistic export of natural resources as oil.

Toy model

It is instructive to analyze a simple toy model where we can explicitly introduce capabilities and test how the two metrics are able to extract information from M_{cp}. Actually, in the real world it is impossible to directly access the vector of capabilities that each country owns. Nevertheless, it is possible to study a simple model (originally proposed in [3]) in which we may explicitly define the capabilities that each country owns and how they combine to produce

products. To this end we need to define two matrices already introduced in Sect. 1.1.1: a country-capability matrix, whose entries S_{ck} specify which capabilities are owned by a country, and a capability-product matrix whose elements T_{kp} specify which capabilities are required to make a product. The model is completed by introducing the simple rule to build the M_{cp} matrix: each country exports a product if and only if it has all the capabilities needed to produce it. In formulas M_{cp} is defined exactly as in Eq. (5).

In this way we now have access to the set of information on which the theory of hidden capabilities is based, i.e. the endowment of capabilities of a country, and we can now compare the asymptotic results of the two different iterative procedures with the real number of capabilities assigned to each country.

We implement the model by extracting random binary numbers, 0 or 1, to fill the $\hat{S}$ and $\hat{T}$ matrices: the entries are equal to 1 respectively with probability 0.7 and 0.05. We consider 200 capabilities, 120 countries and 800 products (following exactly [3]).

In Fig. 10 we show the result of the two methods performed on the artificial M_{cp} matrix obtained from this toy model. Clearly, in this extremely simple framework, the best information about the capabilities is given by the diversification, which corresponds to the first order of iteration of both measures (up to a normalization factor): this is due to the fact that there is no difference whatsoever in the importance of different capabilities, and they are randomly linked to countries and products. We also show the Pearson's correlation coefficient between the two different measures and the assigned capabilities, with respect to the iteration order. Fitness obtained by our approach correctly grasps the relevant information present in the M_{cp} matrix and does not significantly change with the iteration (the reason why these correlations do not improve has to be found in the relative simplicity and randomness of the model, as discussed below). Conversely $k_c^{(n)}$ obtained by the MR seems to be loosing its meaning when the equations are iterated and it is not possible to observe an asymptotic correlation value before the machine precision breaks down.

Economic playground: is China 2nd or 34th?

So far we tested our method and the MR with respect to theoretical aspects and toy models designed to verify the conceptual consistency of the two approaches. We now move our attention to real economic data.

A first striking observation is the anomalously low competitiveness of China that results from the MR method Indeed MR ranks China in the 29th position in 2010 (see [16] pag. 64), just below Romania which is 27th. This result appears rather odd as it would imply that nowadays competitiveness of China is very similar to the one of Romania and far below the one of western countries. Standard economic analyses show instead that China is significantly eroding the competitiveness gap with respect to developed countries and always appears in the very top positions whatever economic indicator is adopted. Therefore, in order to test the economic consistency of the two methods, we are interested in comparing a set of countries which undergo a large variation of ranking in the two frameworks (i.e. China, India, Cyprus, Qatar, see panel *a* of Fig. 11). In the view of standard analysis, they represents respectively two well-established emerging countries whatever economic criterion we consider, an European country with low GDP *per capita* and an oil exporter.

We do not make a direct comparison between our metrics and the ranking of [16] because our dataset is slightly different from the one used in [16] and for a consistent test we prefer to perform the MR on our dataset. In addition we perform a further step of data cleaning. A

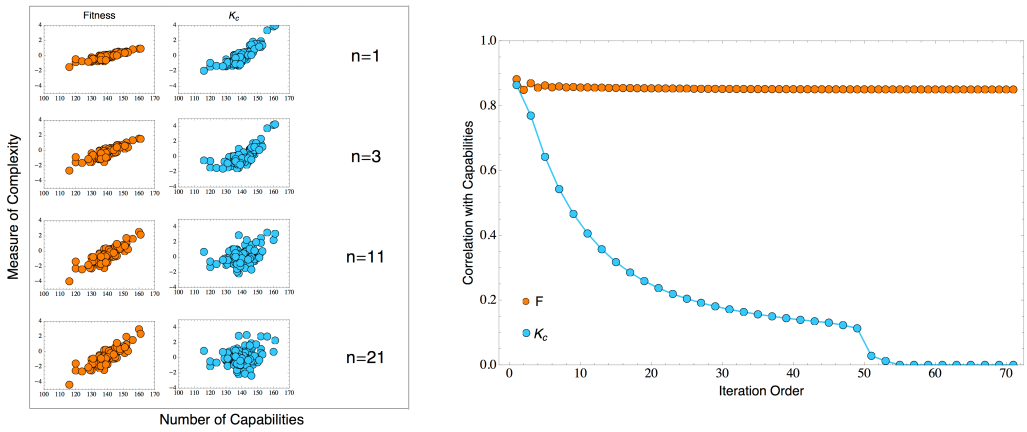


Figure 10: Testing the capability information content measured by the two methods. (Left) Results of the iterations on the toy model matrix. Fitness preserves the first order information, while $k_c^{(n)}$ appears to be rapidly destroying any correlation with the assigned capabilities. We plot the logarithm of Fitness and the rescaled $K_c = (k_c^{(n)} - \langle k_c^{(n)} \rangle)/\sigma(k_c^{(n)})$ at four different orders of iteration. (Right) Pearson's correlation between the measures of complexity and the number of assigned capabilities vs. the iteration order. While Fitness maintains the same level of correlation of the first step, iterating the $k_c^{(n)}$ measure leads to a destruction of information. It is to be noted that in this trivial model the M_{cp} matrix does not contain more information than the simple diversification. Again, the logarithm of Fitness and the rescaled $K_c^{(n)}$ are considered.

second difference stays in the number of countries: 128 in [16], 148 in the present analysis. In spite of some minor differences, the results of the MR on our datasets appear to be similar with respect to the one found in [16] - in fact, as shown in Fig. 11, the MR on our datasets ranks China in 33rd position and Romania in 34th (compare panel *b* and [16] pag. 64).

The anomalous position of China is even more striking when we follow the evolution of the variable $k_c^{(2n)}$ of the MR method from 1995 to 2010 (panel *b* of Fig. 11) where we find that the competitiveness of China follows a growth pattern which does not at all reflect the fact China is now the second GDP power behind USA. We surprisingly find that in MR framework Cyprus and Romania overcome the growth of China in the last years of our analysis. In other words, according to the MR, China, Romania and Cyprus result to be countries characterized by a very *similar* competitiveness and a similar pattern of growth. This scenario appears to be inconsistent with almost all economic analysis of these three countries.

Conversely our method (panel *c* of Fig. 11) on one side spots the spectacular growth of the Chinese productive system in the last fifteen years which was ranked in 13rd position in 1995 and is now in the 2nd position just below Germany which is the country with the highest fitness. On the other hand it depicts Romania and Cyprus as economies of a completely different kind with respect to China: they are growing economies, but we do not spot, as in the Chinese case, the tremendous erosion of competitiveness against most developed countries.

The economic scenario for India is even worst according to MR. India is ranked far below China, Cyprus and Romania and *competes* with Qatar which is a country with a very low

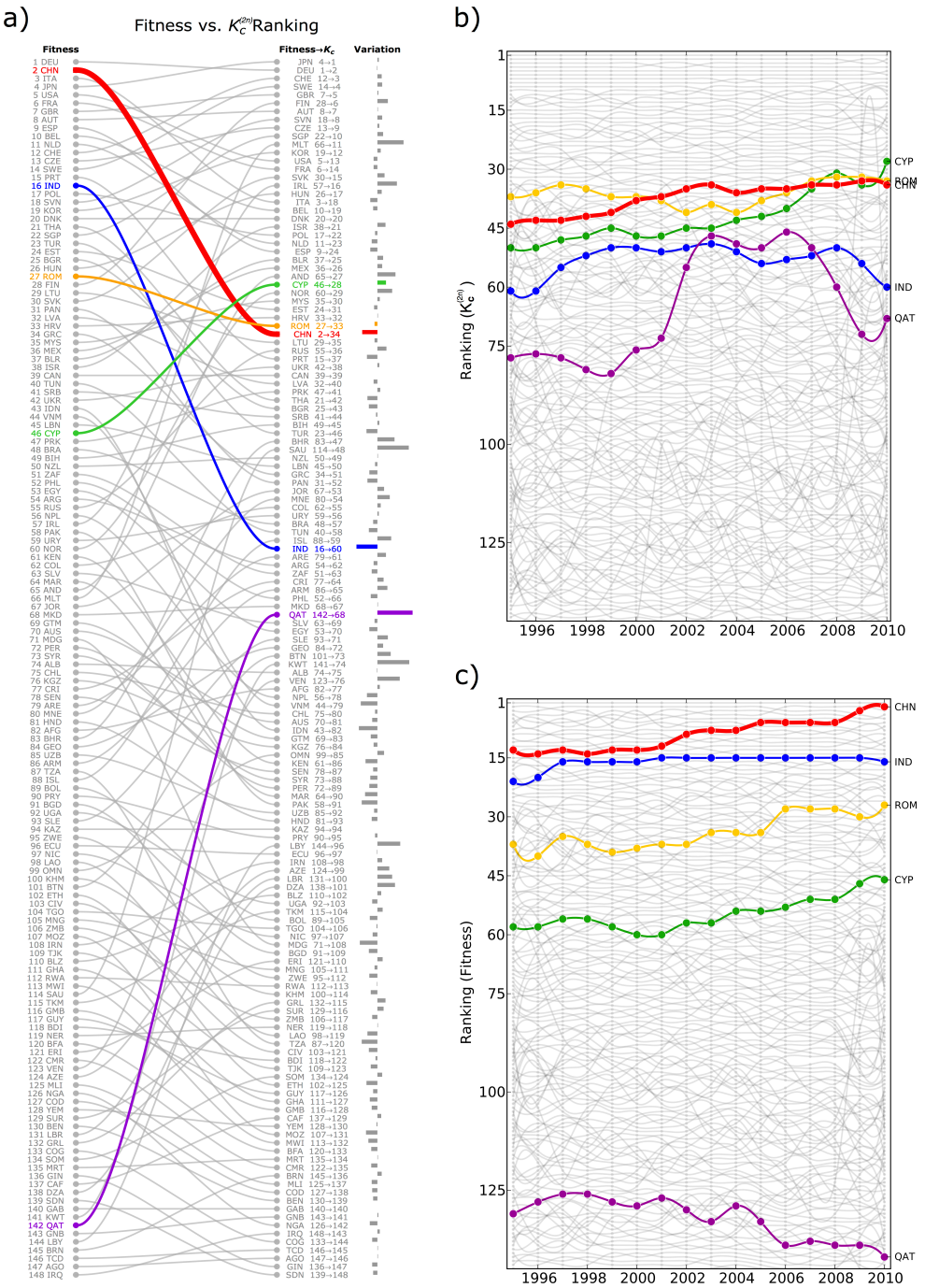


Figure 11: Comparison of the country ranking between FQ method and the MR. In panel *a* we compare the two rankings while in panel *b* and *c* we respectively show the time evolution of the k_c and the fitness. China, India, Romania, Cyprus and Qatar are highlighted for their anomalous behavior in the MR framework and are paradigmatic with respect to the conceptual weakness of this method.

diversification (as it happens for almost all oil exporters). Instead in our method India is an emerging country, above Romania and Cyprus, while Qatar is one of the countries with lowest fitness and with a decreasing competitiveness.

In general in the framework of MR all oil exporters (Kuwait, Saudi Arabia, Iraq, Venezuela, etc), which are paradigmatic of poorly diversified systems, are characterized by relatively high level of competitiveness and tremendous oscillations (compare in [17] for instance Kuwait ranking in 2007 and 2008. Kuwait drops in 1 year from position 66, a relatively high position for a very poorly diversified countries, to position 113. For an explanation of the instability of the HH ranking see 1.C). By consequence the MR also predicts that raw materials are not among those products with very low complexity as it is expected from the observation that a country owns raw materials reserves only by a matter of chance.

The complexity of raw materials such as crude oil is largely overestimated by the MR, for the fact that a large and very diversified country like the US enters linearly in the estimation of oil complexity. Conversely, in our framework, the non-linearity exploits the information that even very non-diversified countries are able to export crude oil, thus it must be a very simple product.

Countries with very large diversification are systematically penalized and medium sized countries tend to be favored by the MR algorithm. As previously observed, the reasons for such a behavior are in the fact that the variables representing the competitiveness of a country in the MR method are linear averages. It follows that the MR ranking is set by the average complexity of the products exported by a country, with an unclear dependence on the level of diversification. This explains why China and India are so poorly ranked and why poorly diversified countries are often over-ranked by the MR: even though China and India have a very diversified export basket, the average complexity of their export is very close to countries much less diversified as Romania, Cyprus and oil exporters.

Instead, in our framework, the fitness of a country is an extensive variable with respect to the number of products exported and properly takes into account both aspects: the average complexity of the products and the diversification of a country.

To sum up, the conceptual flaws of MR produce inconsistent economic results because, differently from the spirit of the theory of capabilities, in the mathematical expression of MR the diversification does not represent a competitive advantage.

1.3 CONCLUSIONS

In this chapter we have presented a framework to define a data-driven non-monetary and non-income based metrics to assess quantitatively and self-consistently the level of competitiveness of a country and the complexity of its products.

We argue that a key element to properly cope with this issue is the non-linearity of the algorithm defining the metrics, inspired by the triangular structure of the countries-products matrix M. The economic observation that developed countries export most of the products implies that the information on the complexity of a product is mainly due to the less competitive countries among all its exporters. The translation in mathematical terms implies that the fitness (i.e. competitiveness) of countries and the complexity of products must interact in a non-linear, almost extremal way.

Differently from previous attempts [3], we are able to correctly grasp the economic essence of the triangular structure of the matrix M and to consistently translate the theory of capabilities in mathematical terms. On one hand we show why the linear method of reflections of [3] is in disagreement with the complexity of economics. On the other hand, by presenting a series of results we spot the consistency of our findings with respect to relevant economic benchmarks. In the following chapter we will focus on how the insights that this methodology provides can be used to make a step forward in introducing new paradigms in economic forecasting.

APPENDICES TO CHAPTER 1

APPENDIX 1.A UNIQUENESS OF THE METRICS' FIXED POINT

Given the rather complex structure of Eq. (8) it is not immediately clear whether a non-trivial fixed point exists and, if so, under which conditions on the country-product matrix (in the trivial case of $M_{cp} = 1\ \forall(c,p)$ a fixed point of course exists and is given by $F_c = 1\ \forall c$ and $Q_p = 1\ \forall p$).

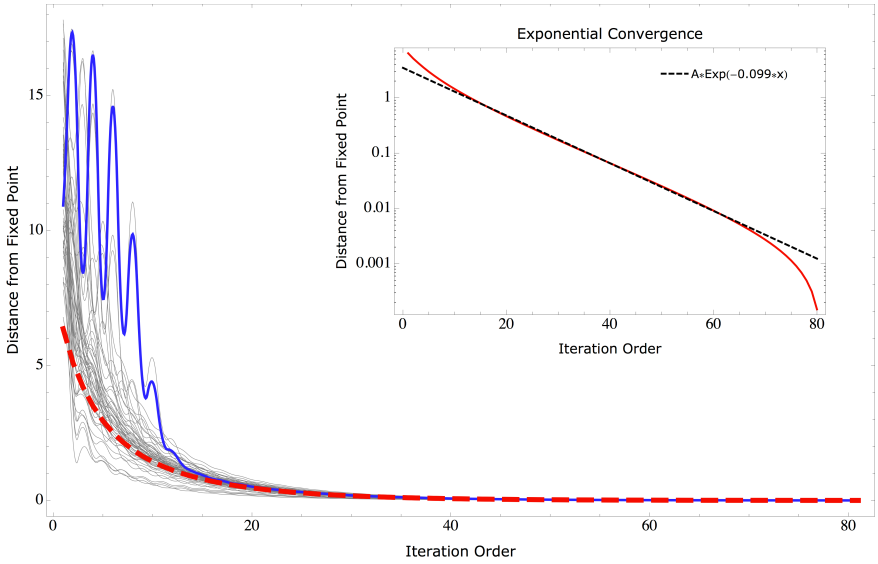


Figure 12: Euclidean distance from the 80th iteration (fixed point) for a particular realization of M_{cp} with $N_c = 5$, $N_p = 15$, $P_h = 0.6$ and $P_l = 0.05$. The red line shows the path obtained with the standard initial conditions given by $\tilde{Q}_p^{(0)} = 1\ \forall p$ and $\tilde{F}_c^{(0)} = 1\ \forall c$. In grey the paths of a set of randomly sampled initial condition. In blue the particular path analyzed in fig.13. The inset shows the exponential nature of the convergence.

For our purposes, being the metrics defined as the fixed point of Eq. (8), we need this fixed point not only to exist, but also to be unique, since we want our result to be independent from the choice of the initial conditions. An analytical proof, due to the strong non-linearity, to the fact that the normalization step constrains the maps to be inside an high dimensional simplex and of course to the dependency on the shape of M_{cp}, if at all possible, is a hard task, out of the scope of the present work. For this reason we perform a numerical analysis of the map defined by Eq. (8). Our analyses are performed for a large number of randomly generated matrices of different sizes but with a triangular shape in analogy to what is observed in the real case. In our random model, by introducing $r = N_p/N_c$ (N_c and N_p are the number

of countries and products respectively), the N_p elements of the $i-th$ row of the matrix are defined as

$$\begin{cases} M_{ij} = 1 & \text{with probability } P_h \\ M_{ij} = 0 & \text{with probability } (1 - P_h) \end{cases} \tag{12}$$

if $j \leqslant ri$ and

$$\begin{cases} M_{ij} = 1 & \text{with probability } P_l \\ M_{ij} = 0 & \text{with probability } (1 - P_l) \end{cases} \tag{13}$$

if $j > ri$, and with $P_l < P_h$. The results presented here are obtained with $P_h = 0.6$ and $P_l = 0.05$ but changing these values even significantly doesn't seem to change the qualitative features of the convergence to the fixed point. We choose a value for r comparable to the ratio of the real matrix, i.e. $r \approx 8$ but also this parameter does not seem to be relevant. We analyze a sample of 300 matrices for 5 values of N_c, i.e: 5, 10, 75 and 150. For each matrix obtained from this model we sample uniformly the $(N_c) - dimensional$ simplex where F_c is defined and the $(N_p) - dimensional$ simplex where Q_p is defined. This correspond to extract random vectors from Dirichlet Distributions of vector parameter $\overrightarrow{\alpha}$ with all unitary components and with proper dimensionality. In order to use these vectors as initial conditions for the iterations we normalize them so that $\langle F \rangle = 1$ and $\langle Q \rangle = 1$. These randomly sampled vectors are used as initial conditions for the maps defined in Eq. (8). For each realization of M_{cp} 1000 initial conditions are tested. Convergence is always observed to a unique fixed point, which only depends on M_{cp}, for all values of N_c and for all the single initial conditions tested.

We present the example of a simple random bipartite network with with $N_p = 5$ and $r = 3$ in order to be able to visualize it. The results are qualitatively similar in all the explored combinations of parameters. In Fig.12 the typical convergence process is shown for the corresponding particular realization of M_{cp}. In the vectorial space defined by the Cartesian product of the two simplexes where $\overrightarrow{F}$ and $\overrightarrow{Q}$ are defined, the Euclidean distance $\Delta^{(n)}$, where n is the order of the iteration, from the point reached at the 80th iteration is evaluated. The red line represents the convergence process with the initial conditions given by $\tilde{Q}_p^{(0)} = 1 \ \forall p$ and $\tilde{F}_c^{(0)} = 1 \ \forall c$. A subset of the paths originated from the randomly sampled initial conditions are shown in grey. All the paths converge around iteration 40 and all the oscillations are damped. As shown in the inset the convergence is exponential $\Delta^{(n)} \sim e^{-\eta n}$. The exponent η depends on the size of the matrix, with bigger matrices converging faster (for $N_c = 150$, $\eta = 0.28 \pm 0.04$). In order to understand the meaning of the peculiar oscillations shown in Fig.12, we plot in Fig.13 the bipartite network relative to that particular realization of M_{cp}, and, considering the trajectory highlighted in blue, we draw the nodes with size (weight) proportional to $F^{(n)}$ and $Q^{(n)}$ at each iteration. The oscillations in n are due to the fact that the weight is being moved from one side to the other of the bipartite network, but these oscillations are damped by the normalization. Notice that this mechanism has the ability of leading to a fixed point also the completely disconnected sub-network formed by the 5th country (in red) and the 13th product. To conclude we can state that, given the observation that the fixed point of Eq.(8) does not depend on the initial condition, the metrics proposed are measuring an intrinsic property of the M_{cp} matrix.

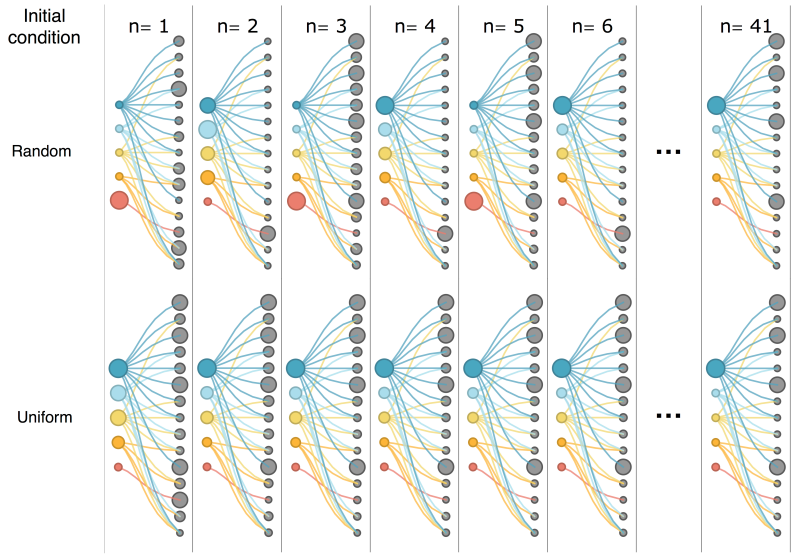


Figure 13: Representation of the non-linear iterations on a simple bipartite network with different initial conditions. Colored nodes represent countries, grey nodes represent products. A random initial condition (top) may give rise to oscillating behaviors (blue line in fig.12) which are dumped by the normalization step. It should be noticed that even disconnected pieces of the network (red "country" node) are brought to a fixed point. The standard uniform initial condition follows a much smoother path (red dashed line in fig.12) and converges to the same fixed point.

APPENDIX 1.B CONVERGENCE AND PROPERTIES OF THE FIXED POINT OF THE ITERATION: ANALYTIC APPROACH.

For the sake of completeness of this dissertation, in this appendix we reproduce, with minor adaptations, the content of [18], with the permission of the authors.

Here we analyze the characteristics of M required to have convergence of the algorithm to numbers strictly greater than zero, for both every F_c and Q_p. We will first show in section 1.B a class of M for which it is possible to solve analytically the algorithm in closed form and find these characteristics; we therefore produce an *ansatz* for the general case. We will then show in section 1.B.1 that the insight is numerically supported for any analyzed random matrix.

A theoretical example

The Matrix Class

We will consider a specific class of matrices. An example of the class is presented in equation 14. The generic matrix in the class is composed of 4 blocks, one of which of zeros. The other three blocks will not be composed only of zeros but they will have some 1s. The density of

1s in the block can be in principle any value between 0 and 1, but to allow for a closed form solution each block has to have the same number of 1s and 0s in all its columns and rows. It is interesting to note that, since the algorithm holds for reordering of rows and columns, a matrix is also a member of the class if it is possible to reorder its rows and columns in such a way to obtain a matrix in the class.

We will number the four blocks from 1 to 4 anti-clockwise starting from the top right corner. In the following we will always assume that the block 4, the bottom right block, is the block with only zeros. Alternatively, will also refers to the block 4 as the *external area*. We will also call the blocks 1 and 3 the *frontier*, block 2 the *internal area*.

In the following example of this class of matrices the four densities are, anti-clockwise from the top right corner, 1/2, 1, 1/3, 0.

$$\begin{array}{c} \\ R_2 \\ \\ \\ \\ R_1 \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc|ccccccc} & C_1 & & & & C_2 & & & \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \end{array} \tag{14}$$

The key variables will be the sizes of the four blocks: their height R_1 and R_2 and their width C_1 and C_2. It will be revealed instead as almost unimportant the density of 1s in the non-empty blocks. To make the notation easier in the following we will assume that all the densities of the three non-empty blocks are equal to 1: the three non empty blocks (1, 2 and 4) will be composed only of 1s. In subsection 1.B we will explain what changes when we relax this assumption.

Computations

The symmetry of the matrix is particularly useful because we can state, for symmetry reasons, that all the countries in the R_2 upper rows have the same fitness, and the same is true for the countries in the R_1 lower rows. This is obvious since their rows are equal or can be made equal with a simple rearrangement of the columns. The same is also true for the complexity of the first C_1 products and the complexity of the last C_2 products.

Therefore the fitness at the iteration n of the algorithm can be represented with just two numbers, with an obvious change of notation

$$F^{(n)} = \begin{pmatrix} a \\ b \end{pmatrix} \tag{15}$$

It is easy to compute the next iteration of complexities from equation 8,

$$\tilde{Q}^{(n+1)} = \begin{pmatrix} \frac{ab}{R_2 b + R_1 a} \\ \frac{a}{R_2} \end{pmatrix} \tag{16}$$

Since we are not interested in the Complexities, to avoid unnecessary computations we will not normalize accordingly to 8. In the next steps we will compute the fitnesses and we will

normalize them: any multiplicative factor to all the complexities would be removed in any case at that step. Thanks to this property we can rewrite the Complexities in an easier way for the next computation, in particular as

$$\tilde{Q}'^{(n+1)} = \begin{pmatrix} R_2^2 b \\ R_2^2 b + R_1 R_2 a \end{pmatrix} \tag{17}$$

where $\tilde{Q}'$ is proportional to $\tilde{Q}$.

We can now compute the iteration $n+1$ for the fitness from equation 8. Since we are using $\tilde{Q}'$ instead of Q, the values obtained by equation 8 will be proportional to $\tilde{F}$. Defining these as $\tilde{F}'$, we obtain

$$\tilde{F}'^{(n+1)} = \begin{pmatrix} C_1 R_2^2 b + C_2 R_2^2 b + C_2 R_1 R_2 a \\ C_1 R_2^2 b \end{pmatrix} = \begin{pmatrix} C_1 R_2^2 b + C_2 R_2^2 b + C_2 R_1^2 + C_2 R_1 R_2 - C_2 R_1^2 b \\ C_1 R_2^2 b \end{pmatrix} \tag{18}$$

where the second step is due to imposing the normalization condition:

$$R_1 a + R_2 b = R_1 + R_2. \tag{19}$$

This constraint allows us to reduce the number of variables from 2 to 1 and to derive a close solution.

Finally, normalizing accordingly to equation 8 we have

$$F^{(n+1)} = \begin{pmatrix} \frac{C_1 b R_2^2 + C_2 R_2^2 b + C_2 R_1^2 + C_2 R_1 R_2 - C_2 R_1^2 b}{C_1 b R_2^2 + C_2 b R_2 (R_2 - R_1) + C_2 R_2 R_1} \\ \frac{C_1 b R_2^2}{C_1 b R_2^2 + C_2 b R_2 (R_2 - R_1) + C_2 R_2 R_1} \end{pmatrix}. \tag{20}$$

Let's focus only at the second component of the Fitness vector, which in the following we will call $F_2^{(n)}$, and which is referred to the fitnesses of the R_1 lowest rows. Comparing 15 and 20 we see that, in one algorithmic step, we had

$$b \to \frac{C_1 b R_2^2}{C_1 b R_2^2 + C_2 b R_2 (R_2 - R_1) + C_2 R_2 R_1} \equiv \frac{b}{A_1 b + A_2}, \tag{21}$$

where

$$\begin{aligned} A_1 &= 1 + \frac{C_2 (R_2 - R_1)}{C_1 R_2} \\ A_2 &= \frac{C_2 R_1}{C_1 R_2}, \end{aligned} \tag{22}$$

from which the closed form solution, by induction, is trivial

$$F_2^{(n)} = \frac{1}{A_1 \sum_{i=0}^{n-1} A_2^i + A_2^n}. \tag{23}$$

The solution for the other component can be obtained from the normalization condition given by Eq. 19

Convergence

How does $F_2^{(n)}$ behave when n goes to infinity? It trivially depends on A_2:

- if $A_2 > 1$, F_2 converges to 0 exponentially fast while n goes to infinity;
- if $A_2 = 1$, $F_2^{(n)} = 1/(nA_1 + 1)$ and therefore converges to 0 as fast as n^{-1};
- if $A_2 < 1$, F_2 converges to a non zero number, $(1 - A_2)/A_1$.

From 22 we notice now that A_2 has also a particular geometric meaning: it is the ratio of the areas of the bottom right block (block 4) and the top left block (block 2). Remembering our definitions of the blocks, A_2 is therefore greater than 1, and the second fitness component $F_2^{(n)}$ converges to 0, if the *external area* is bigger than the *internal area*. Or, visually, if the belly of the non-zeros area is inward.

$A_2 = 1$, when A_2 is defined through equation 22, defines the diagonal line, or its obvious geometric extension to rectangular matrices,

$$R_1 = (R_1 + R_2)\frac{C_1}{C_1 + C_2}. \tag{24}$$

Heterogeneous Density

Interestingly the result does not change if the densities of 1s in the blocks 1, 2 and 3 are different than 1, if the block 4 is still composed only of 0s and the block 2 has a positive density[2]

Only the shape matters, and not the density. In particular doing the same computations with densities different from 1 (and potentially heterogeneous) we find that:

- equation 23 does not change,
- A_2 does not change,
- A_1 does depend on the densities of the blocks 1 and 3, but not from the density of block 2 if it is not 0 (see ahead for the case in which the density of block 3 is zero).

The specific value of the density of 1s and 0s in the internal area is therefore completely irrelevant, while the density inside the frontier only determines the specific value of the convergence point (if there is convergence to a number greater than 0).

Zeros outside the frontier

The case in which the density in the internal area is 0 is a peculiar case, since in this case it is not possible to really define an inside and an outside. In fact it is possible to rearrange the rows and columns to switch inside and outside. In this case it is still possible[3] also to define A_2. If A_2 is different than 1 the fitness of one of the two sets of countries is always converging to zero. In the case in which A_2 is equal 1 instead, any value of fitness is a stationary point

2 this condition is obviously needed to clearly define an internal area and an external area.

3 Of course it is possible to rearrange rows and columns to switch the blocks, i.e. there are two different ordered M, causing A_2 to go to A_2^{-1}. Therefore A_2 is defined but for a possible multiplicative inversion.

of the algorithm and therefore the fitnesses produced by the algorithm will be equal to the starting conditions.

This can be said in a different way. In this case the blocks 1 and 3 are not connected through a product common among countries belonging different blocks. Therefore the algorithm in this case does not have any direct information to compare fitnesses and complexities of countries and products in different blocks. To compare the two blocks it is therefore only possible to use their densities and shape. However the density turns out to be irrelevant, since the more products in a block the more is diluted their influence on the Fitnesses of the countries in their block: as it is possible to compute, this cancel out the additional effect of having more or less products. Only the relative shape among the blocks, measured by A_2, is left to compare the two blocks. If A_2 is equal 1, and therefore also the shape does not help to compare the blocks, any fitness and complexity is consistent. If A_2 is different than 1, it will determine the convergence to 0 of the fitnesses and complexities of one of the two blocks.

Ansatz

The ansatz that will be then corroborated by the numerical investigation in the next sections is therefore simple, at least using loose definitions:

Ansatz 1. *If the belly of the matrix is outward all the fitnesses and complexities converge to numbers greater than* 0*. If the belly is inward, some of the fitnesses will converge to* 0*.*

While in the case defined in 1.B the conversational definition of "inward belly" - corresponding to $A_2 > 1$ - is unambiguous, for the general cases that we will investigate in the sections 1.B.1 and 1.B.3 there are many possible definitions.

Defining

- *ordered* matrix, as the matrix M after rearranging the columns and rows such that all the countries are ranked accordingly to their fitnesses and all the products are ranked accordingly to their complexities[4];
- *diagonal* line, as the line, in the previously defined matrix, going from the least fit country and least complex product to the most fit country and most complex product; if the matrix is squared, this definition is the usual definition of diagonal of the matrix; if the matrix is rectangular, this definition is the trivial geometric extension of the previous one.
- the *external* area of the matrix is the joint set of zeros in the ordered matrix including the corner corresponding to the 0 for the lowest fitness and highest complexity product;

we will see that a proper definition, with relative guess, is

Ansatz 2. *A matrix* M *have all the fitness and complexity different from zero if, after ordering said matrix, the diagonal line does not pass through the external area.*

4 while in same cases, as we have seen and we will see, the fitness could converge to 0 for many countries, for any finite number of iteration the fitness values will be greater than 0. Moreover, since the convergence speed after some iterations is constant for each country and products, there will be a number of iterations such that, for any following iteration, the ranking of fitnesses and complexities will be constant.

It is worth to note that when a country's fitness converges to zero, from equation 8, we know that all the exported products' complexity will converge to zero too. At the same time, the effect on the fitnesses of countries which export a product whose complexity converges to zero, is negligible. Therefore, we can consider that, but for multiplicative common factors coming from equations normalization factors in eq. 8 removing from the matrix M a line corresponding to a country converging to zero fitness and all the columns corresponding to its exported products, the output will be not affected.

It is therefore reasonable to make a further guess:

Ansatz 3. *If the diagonal line does pass through the external part of a ordered matrix* M*, some countries and products will converge to zero. If we progressively remove them from the analysis, defining new ordered matrices* M*, we will have a convergence to finite values of fitnesses and complexities for the remaining countries and products for which the the new* M *have a diagonal line not passing through the external area.*

These ansatz implicitly define the *crossing country* of a matrix as the lowest fitness country that converges to a nonzero fitness. In other words, it is the first country that does not need to be removed in the removal process described above, defining the largest matrix such that all its countries have nonzero fitness. Note that in general searching instead from the top countries does not give the proper result in a situation in which the diagonal cross multiple times the external area. If one starts the check from the top countries and keeps adding countries with lower fitnesses, the crossing country is not the first one met. In this case of multiple crossing, the crossing country is the first country met in the process starting from the bottom, as in ansatz 3.

The guesses 2 and 3 will be tested numerically in the sections 1.B.1 and 1.B.3.

1.B.1 A numerical investigation

In this section we present the results of our numerical simulations, in which we study some peculiar behaviors of sample matrices which we did not treat analytically. In particular, by means of synthetic examples of possible M matrices we study the cases in which more than two blocks are present, the importance of producing rare products and we investigate the ansatz proposed in the previous section. The reader will notice that the matrices investigated in the first two subsections correspond to the case $A_2 = 1$ discussed above.

1.B.2 More than two blocks

In the case of more than two blocks it is not possible to write the components of $F^{(n)}$ in the form given by Eq.15, because one would have more than one unknown variable. However, what one can do is progressively reduce the given matrix to submatrices, each one made of two blocks. In the following example we present a block diagonal matrix:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

We have found that if off block diagonal elements are absent (that is, both the internal and the external areas are empty) all countries will remain in the respective initial conditions. As we expected from the two block analysis presented in Sec.1.B, this result is stable even if density is lower: only the relative dimensions of the blocks counts.

The importance of oligopolies

The presence of common products does not add much information other than the fact that those countries which produce only very common products can not be very high in the fitness ranking. As a consequence, we can expect *monopolies* of already diversified countries to be relevant. To investigate the consequences of having common ubiquitous products we take into account a synthetic situation in which three countries produce only very common products, two countries have a *duopoly* and one country, which has the same diversification of these two countries, has a monopoly. The results are the following:

$$\begin{matrix} c \\ n^{-1} \\ n^{-1} \\ n^{-2} \\ n^{-2} \\ n^{-2} \end{matrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

On the left side of the matrix we show the speed of the fitness decays with the iteration n, with c meaning convergence to a finite value of fitness[5]. For example, the notation $n^{-\alpha}$ means that the fitness $F_c^{(n)}$ of the country c corresponding to that row is converging towards zero with a power law of the number n of the algorithm iterations, that is, $F_c^{(n)} \sim n^{-\alpha}$ for a sufficiently larger n. As one can notice, the presence of a monopoly makes the first country the only one to converge to a nonzero value of F, while the advantage given by the extra products makes the second and third country perform better with respect to the last three countries. In any case, this advantage is not striking, because we are still in the case in which $A_2 = 1$, so the difference is only in the exponent of the decay. We stress that the presence of a different decay means that there exist a number N so that from iteration $n = N$ the ranking will stay constant, and so the ranking will be given by the decay exponent.
A further, self-illustrative situation of the consequences of having common products is presented below.

5 In this section and in the following we will not report the complexity decays, which follow trivially from the ones of the countries.

$$\begin{array}{c} c \\ n^{-1} \\ n^{-2} \\ n^{-3} \end{array} \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Let us now consider the following matrix:

$$\begin{array}{c} c \\ c \\ n^{-0.6} \\ n^{-1} \\ n^{-1} \\ n^{-1} \end{array} \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

here we have three blocks of nations, with one monopoly (country 3) and two duopolies (countries 1 and 2). On the basis of the previous considerations, we know that the 3x3 submatrix constituted by countries 1,2 and 3 and products 4, 5 and 6 is degenerate, in the sense that every initial condition $F^{(0)}$ is stable under the iterations (also the whole matrix would be degenerate if country 1 would not have produced product 2). However, in this case, the presence of an external product changes the situation, giving an advantage to the duopoly, which converges to a finite value of the fitness and makes the monopoly tend to zero, even if with a lower exponent (0.6 instead of 1). By means of extensive numerical simulation we have found that the specific value of the decay is given by the relative size of the worst block (the one which decays with n^{-1}) with respect with the total size of the externally connected block, that is, the sum of the sizes of the blocks which are connected by the external 1. In other words, the decay is given by the ratio between the worst block and the sum of the sizes of all the blocks but the one whose decay we are computing. In the situation presented above, the worst block has size 3 and the converging block has size 2, from which we have $3/(3+2) = 0.6$. To make another example we can imagine a different case, in which we give an advantage to the monopoly with respect to the block of size 3; in this case, the duopoly will decay with an exponent equal to 3/4.
Another interesting phenomenon is evident. Let us consider the set of countries 4, 5 and 6. Even if the second product is owned only by countries 4 and 5, also country 6 decays in an equally faster way. This is due to the fact that the last block is *connected*, that is, there exist a path of products that connects all the countries in the block, and the absence of monopolies in the block.

Exponential decays

Until now we have seen that different power law decays come out from, in general, N-polistic competitions whose symmetry is broken by the presence of products external with respect to the competitors. In the case in which one or more N-poly has a further N-poly which is not shared, the decay of the competitor will be *exponential*. Here it is an example:

$$\begin{matrix} c \\ e^{-n} \\ e^{-n} \\ e^{-n} \end{matrix} \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

In this case the first country has two monopolies; country 3 one monopoly and countries 2 and 4 no monopolies at all. All countries tend to zero exponentially faster, but the first one. This is a simple example of an inward bellied matrix, that is, a matrix such that the diagonal crosses the external empty part. Because in this case the crossing country is the first one, all the other countries will converge to zero exponentially, as stated in the ansatz 3. This phenomenon will be investigated in detail the next subsection.

Numerical verification of the ansatz

In Sec.1.B we suggested a link between the convergence properties of countries and the shape of the belly of the ordered matrix. In this section we verify the given ansatz by means of simple matrices, whose behavior is nevertheless analogous to the one we will find in real world applications. Let us start with a set of 5x5 square matrices. We point out that we use square matrices only because the diagonal and, as a consequence, the shape of the belly is evident by eye. As we will see in the following, these results can be applied also to rectangular matrices by means of the generalized definition of diagonal discussed above.

A

$$\begin{matrix} c \\ n^{-1} \\ n^{-2} \\ n^{-3} \\ n^{-4} \end{matrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

B

$$\begin{matrix} c \\ e^{-n} \\ e^{-n} \\ e^{-n} \\ e^{-n} \end{matrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

C

$$\begin{matrix} c \\ c \\ c \\ c \\ c \end{matrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

D

$$\begin{matrix} c \\ n^{-1} \\ n^{-1} \\ n^{-1} \\ n^{-1} \end{matrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Matrix A represents a borderline case, in which the same lines of reasoning adopted in Sec.1.B.2 can be applied. All countries but one are converging to zero, preserving a well defined ranking.
Matrix B shows a clear inward belly. This means that, according to our ansatz, in order to know which countries will converge to a nonzero value of fitness we have to eliminate countries starting from the ones with a lowest ranking until we find a submatrix with an outward belly. In this case this is not possible, in the sense that the only fully convergent submatrix would be the trivial 1x1 matrix containing only country 1 and product 5. As a consequence, only the country 1 (which, in this case, is the trivial crossing country) and product 5 converge to a nonzero value of fitness and complexity, while the other countries and products decay

exponentially.
On the contrary, matrix C has a clear outward belly: all countries have products beyond the diagonal and so converge to nonzero values of fitness. An interesting situation is shown in matrix D, which is equal to C but for one product. By removing product 4 from country 2 is it possible to make all countries but one converge to a zero value of fitness. In fact, in this case the diagonal crosses the external part of the matrix in correspondence with an high ranking country. This situation highlights the importance, in general, of a careful data sanitation when empirical M matrices are built, not only with respect to low fitness countries, but also for high fitness ones. In any case, we point out that when one passes from matrix C to D the ranking is preserved.
In order to show a case in which the M matrices are rectangular and the crossing country is nor in the top, neither in the lowest position in the matrix, but somewhere in the middle of it, we consider the three following matrices:

$$\begin{array}{c} \\ c \\ c \\ c \\ c \end{array}\!\!\begin{array}{c} \text{E} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ c \\ c \\ n^{-1} \\ n^{-1} \end{array}\!\!\begin{array}{c} \text{F} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ c \\ c \\ e^{-n} \\ e^{-n} \end{array}\!\!\begin{array}{c} \text{G} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

While matrix E shows full convergence of its countries, it is enough to remove one product to make the external area large enough to be crossed by the diagonal and, as a consequence, to make two countries converge to zero (matrix F). The removal of one more product turns the convergence to zero exponential with the number of iterations. We stress that, even if the number and the category of products of country 4 is the same for all the three situations, its value of fitness changes. However, this fact is not paradoxical as long as the values of the matrices are calculated in a correlated way, by considering, for example, the Revealed Comparative Advantage.

1.B.3 Real cases

In this section we study how this variety of convergence behaviors affects real M matrices. Typical M matrices are bigger than the ones seen in chapter 1.B.1 and present specific characteristics of nestedness and density; nevertheless, our convergence ansatz turns out to be relevant for real M matrices.

hs2007

We consider the 1995-2010 BACI database. In all the years the matrix is outward bellied, accordingly to the definition implicit in 2. In other words, the diagonal does not cross the external (i.e., empty) part of the matrix, and so the algorithm converges for most countries to a non zero value of fitness. While the general behavior is clear, some exceptions are there for the least fit countries, as it is visible in figure 14. In the figure we show a pictorial representation of the ordered M matrix for the year 1995. The yellow (red) dots mean that the country corresponding to the given row is (not) exporting (in the RCA sense) the product corresponding to the given column, that is, the matrix element is equal to 1 (or zero, respectively). The horizontal line corresponds to the *crossing country* defined in 1.B, and the vertical one to the product with

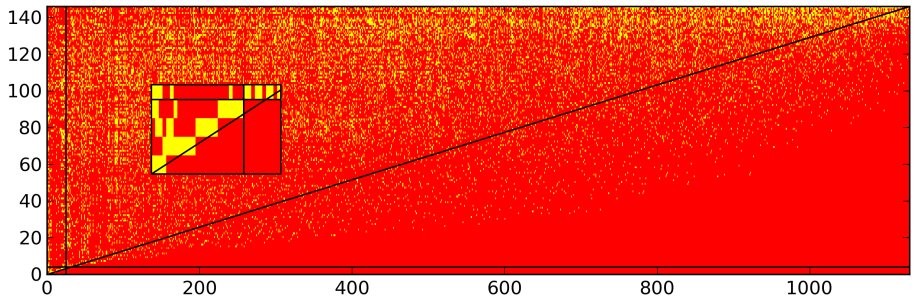


Figure 14: Ordered M matrix for the BACI dataset, year 1995. The diagonal is consistently above the external part. However, there is an exception for the low fit countries, zoomed in the picture inset. The horizontal and vertical lines represent the number of countries and products that converge to zero.

higher complexity exported by it. The part of the matrix which converges to nonzero values of fitness and complexity is defined by these lines (in particular, the top right portion of the original matrix). The dashed line is the diagonal of the fully convergent matrix. Obviously, after the removal process the matrix is outward bellied in all its parts.

It is interesting to note that the number of countries and products that need to be removed to have a matrix converging to non-zero values is reduced for the later years. This behavior is general, for all the datasets checked. The causes of this phenomenon will be discussed in the conclusions in section 1.B.4.

Sitc2

Here we consider the World Trade Flows Database, covering the 1963-2000 period. The dataset is particularly interesting because of the long time span of data presented in a consistent form. In this dataset the convergence ansatz proposed is more relevant, since big parts of it is consistently above the diagonal. An example, that highlight also an application of ansatz 3, is presented in figure 15.

The example also highlights a peculiar graphical feature of the ordered matrix. Indeed, for the countries and products that are converging to zero it is possible to observe how the frontier between the internal and external part is smooth and continuous, while a more rough behavior can be observed for the areas converging to positive numbers. This behavior is due to the fact that, since the fitnesses and complexities of the countries and products below the crossing country are converging exponentially to zero, for any given finite iteration the ratio between two consecutive fitnesses and complexities is also exponentially growing. Therefore the fitnesses are mostly determined by the highest complexity product and, similarly, the complexities are given by the lowest fitness country. As a consequence there is a clear correspondence between best products and worst countries, determining the smooth frontier that it is clearly visible in figure 15. At the opposite, for the countries above the crossing country, their relative fitnesses converges to fixed ratios. Therefore the fitnesses of the countries are determined by all the products not converging to zero, causing their ranking to be not uniquely determined by their best product. A similar line of reasoning can be used for the products. Therefore after the crossing country there is not a clear frontier.

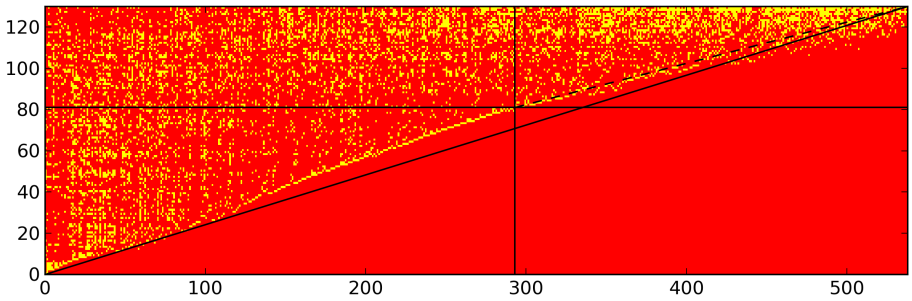


Figure 15: Ordered M matrix for the 1965-2000 dataset, year 1965. The matrix frontier is consistently above the diagonal line. The horizontal and vertical lines represent the number of countries and products that converges to zero. The dashed line shows the diagonal of the new block remaining after removing those countries and products: the diagonal of the new M matrix now does not cross the external part of the matrix.

A similar behavior can be observed in all the matrices that we studied. Even in the case of figure 14 it is possible to recognize the phenomenon for the last four countries (zoomed in the inset).

Patents

The convergence to zero of a sizable set of countries is very evident for matrices particularly empty, like the one produce from the patents dataset. It is a dataset organized by OECD starting from data produced by the European Patent Office[6] joining countries and technological sectors if a firm in a country has patents granted in that sector. It is then manipulated in the usual way, through consideration of comparative advantage to remove the size of the countries from the argument, generating a matrix that has a 1 or 0 if it produce more patents than expected in a particular technological sector. Since to patent a discovery in a sector the country has to be on the frontier of the technological progress in that sector, the matrix comes out with an inward belly, as it is visible in figure 16.

As for guess 3, a possible situation leading to a global convergence to zero is for a single country being diversified in too many sectors in which no other countries is interested in. In this case we can have all the other countries' fitnesses converging to zero. An example is in figure 17.

1.B.4 Conclusions and discussion

We have shown that the intrinsic non linear structure of the algorithm proposed in this chapter has highly non trivial consequences on its convergence properties. In particular, we have linked the shape of the ordered matrix on which the algorithm is based to the possible presence of countries whose fitness tends to zero. Obviously, also all the complexities of their exported products will converge to zero. In synthetic but general cases it is possible to show analytically that the presence of a large empty part of the matrix makes some countries converge to zero.

6 For more informations, see the OECD Patent Statistics Manual available at http://browse.oecdbookshop.org/oecd/pdfs/free/9209021e.pdf

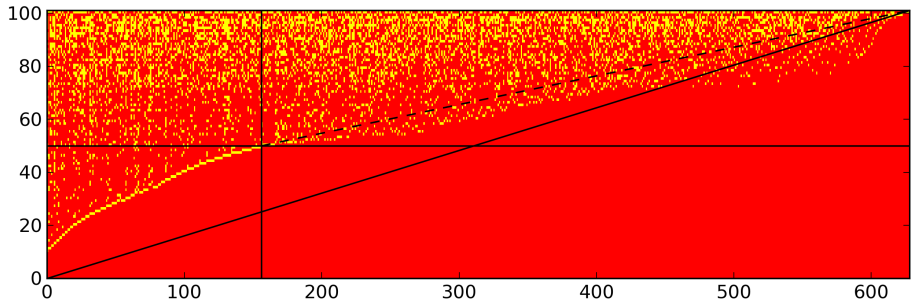


Figure 16: Ordered M matrix for the patents dataset, year 2002. The external part is very big, so diagonal line. crosses it and a consistent fraction of countries converges to zero. The dashed line shows the diagonal of the new block remaining after removing those countries and products.

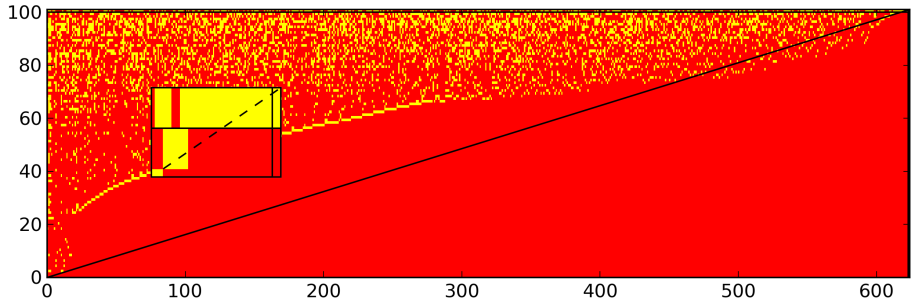


Figure 17: Ordered M matrix for the patents dataset, year 2000. The diagonal line is consistently inside the external area. More importantly, this is true even for just the top two countries, as it is visible in the zoomed in the picture inset. As a consequence, all the countries and products converges to zero but the first one, as shown by the horizontal and vertical lines. The dashed line shows the diagonal of the block that would be left removing all countries and products but the first two: the diagonal would still cross the external part. Therefore the only remaining non-zero-fitness country is the first one, and the only remaining non-zero-complexity products are the ones produced only by it.

In particular, the condition to check is if the matrix diagonal crosses or not the empty part of the matrix. We study numerically some simple cases which confirm our results also when the analytical approach was not possible and in some real world cases, such as two different country-product databases and a country-patent database.

A striking result of our analysis is that a large variety of situations is present. For example, if one considers the hs2007 BACI 1995-2010 database, almost all countries have a finite fitness, while in many years of the sitc2 1963-2000 database more than a half of the countries have a zero fitness. On the other hand, in some patents datasets all the countries fitnesses but one tend to zero.

A practical application of our results is that the usual convergence conditions are not suitable for this algorithm. In particular, the condition $|F^{(n)} - F^{(n-1)}| < \epsilon$ does not imply that the single components of the fitness vector stop to decrease and, in general, the resulting fitnesses and complexities could depend on ϵ. This could be particularly relevant if one is interested in non linear expressions of fitness or complexities (e.g. the logarithm of the countries' fitness), also given the fact that convergence rates can be very slow.

In principle, one may be interested in the rankings or in the cardinal value of fitnesses and complexities. Obviously, different criteria can be assessed for the ranking convergence. We suggest, as a practical recipe, to set a suitable threshold at a very small value and record the order of the crossings. Then one should remove the countries and products which have crossed the threshold and run the algorithm with the reduced matrix, in the spirit of ansatz 3. This procedure should be stopped when a fully convergent matrix is obtained, and all the remaining countries and products have a finite value of fitness and complexity, respectively, as we have shown in sections 1.B.1 and 1.B.3. Obviously, in principle nothing prevents ranking changes to happen after the threshold has been crossed. However, we believe that this procedure, if used with a suitable small value of the threshold, for example the machine precision, leads to a ranking which is more stable with respect to the ones that could be obtained using the usual convergence criteria.

APPENDIX 1.C HOW NOISY DATA CAN AFFECT THE ANALYSIS

A (very) Basic Model for Economic Complexity

Along with the present idea of economic complexity, in a dynamic competitive environment the export adjacency matrix M of the real diversified world shows a triangular profile. These observation reflects the idea that disposing of an increasing number of capabilities means enjoying increasingly higher advantages in terms of productive capacity. Trade's network structure and the economic forces which shaped export data into a triangular matrix are presently under investigations. In order to study the robustness with respect to noisy data of HH and NL metrics, according to the phenomenology of trade data we propose to consider a perfectly triangular matrix (see fig. 18). We call respectively N_p the number of products and N_c the number of countries. Given the rectangularity coefficient of the matrix $\beta = N_p/N_c$, we set

$$M_{cp} = 1 \qquad \textit{if} \quad p \leqslant \beta c \tag{25}$$

$$M_{cp} = 0 \qquad \textit{otherwise}. \tag{26}$$

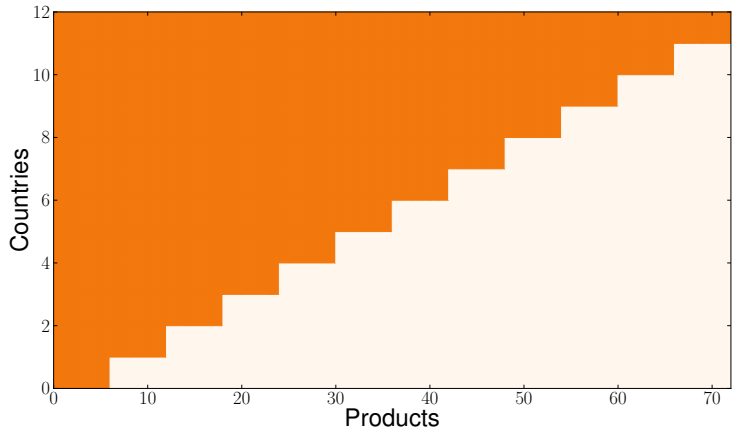


Figure 18: Basic Model for Economic Complexity for an export matrix M with $N_p = 72$ and $N_c = 12$. All the bits whose coordinate satisfy the relation $p \leqslant (N_p/N_c)c$ are set equal to one. All other bits are equal to zero.

Given $k_c^{(o)}$, the most simplistic hypothesis consists in assigning to each country c a number of capabilities A_c proportional to its diversification. Hence we set $A_c = \gamma k_c^{(0)}$, being γ an unknown multiplicative factor. This is the fundamental assumption of our Basic Model for Economic Complexity (BM). Moving from this assumption, we are now able to test the two metrics' robustness properties.

Robustness and sensitivity analysis of the linear and nonlinear metrics

The export values of the different countries registered in US dollars are inevitably subject to a certain number of errors. Part of these errors can be corrected by comparing it with the M_{cp} value in question with that of the immediately previous and following years and by generally revising the structure of the *Harmonized System*. Inevitably, a certain fraction of errors remains and affects the quality of measures of economic complexity on C-P network. In order to be effective, a metric must produce a ranking which shows proper robustness and low sensitivity with respect to corrupted data. The main goal of this paper is then to undertake a scrupulous analysis of the two metrics' performances in presence of known amount of errors in the export matrix. From here on, we refer to the fraction of altered data in the C-P network as the quantity of noise η. The introduction of a phenomenological model such as the BM is fundamental to test the robustness of HH and NL metrics. We call noisy a bit for which the value of M does not satisfy the prescription given in eq. 25 and 26. Since the information about the existence of a link in the bipartite network C-P is binary, a noisy bit is a bit subject to inversion.

Given the definition of a vector of capabilities such as the one provided by the BM and this definition of noise, we claim that it is possible to reckon the robustness' properties of HH and NL metrics by calculating the correlation between the two measures of economic complexity ($HH : \vec{k}_c, \quad NL : \vec{F}$) and the countries' capabilities with the respect to the level of noise in the system.

Spearman's correlation coefficient ρ_s is a non-parametric measure of statistical dependence between two variables. It is defined as the Pearson correlation coefficient between the ranked variables. For a sample of size n, the n raw scores X_i, Y_i are converted to ranks x_i, y_i and r_s is computed from these:

$$\rho_s = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{27}$$

Maximum correlation between the measured values of economic complexity and the number of capabilities of each country is obtained in absence of noise, where we set $\gamma = 1$ without loss of generality. In order for the two metrics to be robust, it must be

$$\rho_s(\vec{A}, \vec{k}_c^{\eta \neq 0}) \approx \rho_s(\vec{A}, \vec{k}_c^{\eta = 0}) \tag{28}$$

$$\rho_s(\vec{A}, \vec{F}^{\eta \neq 0}) \approx \rho_s(\vec{A}, \vec{F}^{\eta = 0}) \tag{29}$$

for $\eta > 0$ and small enough, where η is the ratio between the number of inverted bits and the total number of bits. In fig. 19 we show $\rho_s(HH)$ and $\rho_s(NL)$ as a function of η. Although theoretically parameter η runs from 0 to 1, we cannot compute ρ_s for too high values of noise. In fact, in order to be defined, the two metrics must operate on a matrix without null rows or null columns. However, a complete inversion of the bits of the BM export matrix leads exactly to a situation of this kind. To avoid problem about metrics' definition, we decided to stop at $\eta_{max} = 0.9$ for all the considered volumes.

As shown in fig. 19, the non-linear iterative process of our metrics preserves the ranking of countries' economic complexity from the introduction of a certain amount of noise. The value of Spearman's rank coefficient displays a plateau for approximately $\eta < 0.35$: the measure of fitness is robust even in presence of a level of noise in the system of 35%. Contrary, correlation of HH metrics on matrices with noisy data decreases almost linearly with η. For $\bar{\eta} = 1/2$ the matrix is completely disordered and $\rho_s \approx 0$ (slightly greater than 0 because of the little asymmetry in the number of 1 and 0 for $\eta = 0$). For $\eta > \bar{\eta}$ the matrix is inverted. Correctly, $\rho_s(NL)$ is an odd function of η with respect to the center of symmetry $\bar{\eta}$. Analogous trends are found for Q and k_p. Our results are consistent with sampling of M_{cp} with different volumes. As fig. 20 shows, the extent of the plateau increases with greater matrices. In other terms, the same fraction of noise is less effective in greater systems. Contrary, as shown in fig. 21, measures of k_c do not improve for greater volumes.

Our results are basically not affected by the parameter β (in other words, the shape of M).

Convexity of the adjacency matrix M

From a phenomenological point of view, in BM we qualitatively considered as perfectly triangular the profile of the connected part of the adjacency matrix M. However, if we undertake

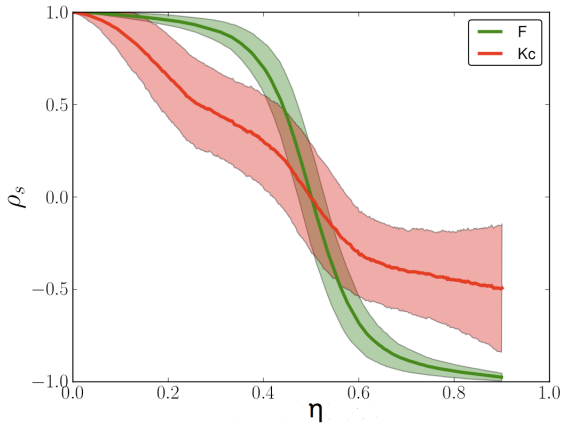


Figure 19: Computation of Spearman's rank correlation coefficient between the number of capabilities $\vec{A}$ of each country and the two measures of economic complexity with respect to the level of noise in M. In green are shown results for the non-linear metrics F, in red for the linear metrics k_c. Fitness' correlation shows a plateau for around $\eta < 0.35$ for a matrix with $N_p = 720$ and $N_c = 120$. This indicates that NL measure of economic complexity is robust even in presence of a level of noise in the system of 35%. Contrary, the fast decline of the value of ρ_s for k_c means that HH measure is sensitive to the introduction of even a small amount of noise.

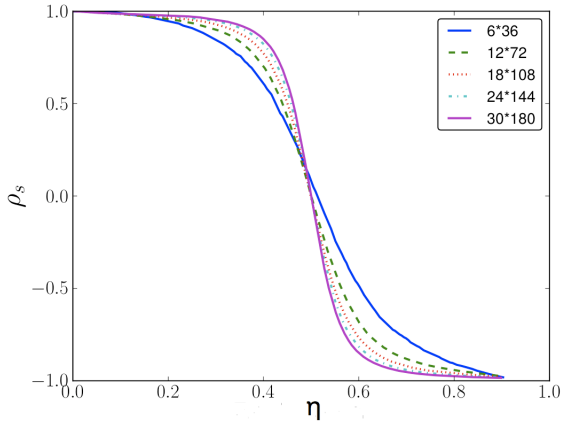


Figure 20: Computation of Spearman's rank correlation coefficient between A and F with respect to the level of noise for different volumes of M. Legend shows the considered values of $N_c * N_p$. Measures are consistent for different volumes and the extent of the plateau increases with greater matrices.

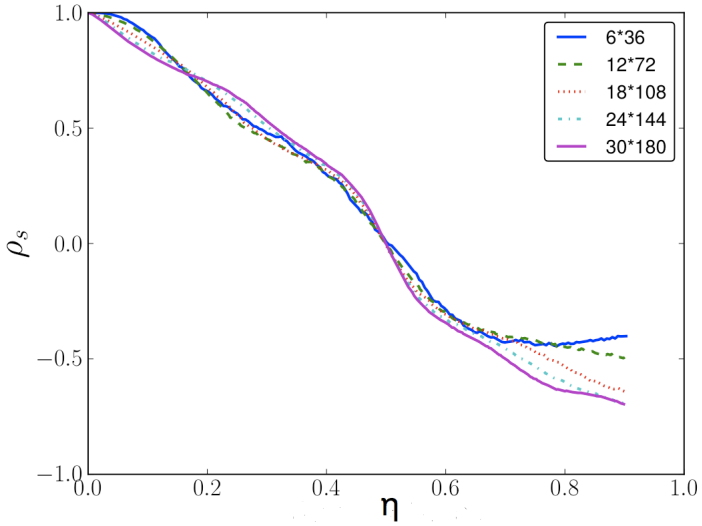


Figure 21: Computation of Spearman's rank correlation coefficient between A and k_c with respect to the level of noise for different volumes of M. Differently from NL metrics, measures of k_c do not become more robust for greater volumes.

a more precise analysis, M shows on average a slightly convex profile. A robust metric must preserve its properties of noise sensitivity for matrices with non-null concavity. We define $\Pi_c = M_{c,p_{max}}$ the most complex product exported by the country c. From the formal point of view, in the zero model we can parameterize the profile Π through the concavity parameter α:

$$\Pi(x_c, \alpha) = N_p e^{-\alpha \frac{\beta x_c}{N_p}} - a\frac{\beta x_c}{N_p}, \tag{30}$$

where higher α values correspond to more concave M and a is a fixed parameter such that $\Pi(x_c = 0) = N_p$ and $\Pi(x_c = N_c) = 0$, then $a = N_p e^{-\alpha}$. We associate α values with the same absolute value but with negative sign to specularly convex profiles (see fig. 22).

Figg. 23 and 24 show results for NL metrics and HH metrics. In the case of nonlinear metrics, for small positive and negative α values the ρ_s parameter does not show significant deviations compared to the null concavity case. The results of the linear metrics, definitely noise sensitive in case of perfectly triangular profile already, also show strong dependence to the specific concavity value of M.

Finally, our ranking process shows remarkable properties of robustness to noise with low dependence from the system volume and the specific shape of the network. On the contrary, HH metrics lacks these properties and the economic complexity ranking it produces can depend strongly on possible database errors.

Effects of noise in specific ranges of the matrix M

Now that the greater robustness of NL metrics is determined, let us study the incidence in the fitness ranking of noisy data localized in specific ranges of the M matrix. At first, we may think

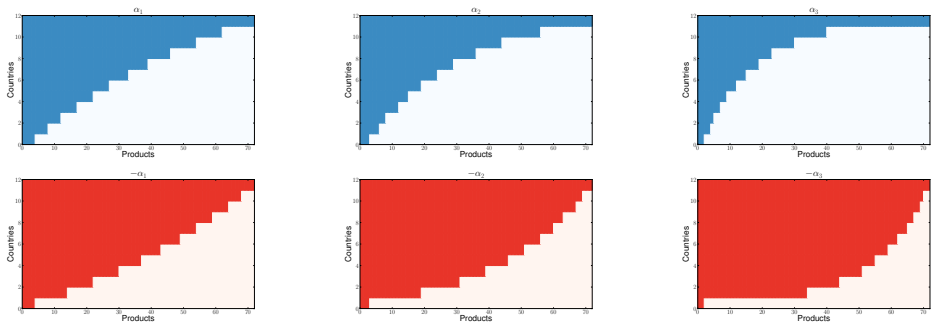


Figure 22: BM matrices for non-null concave and convex profiles in absence of noise with $N_p = 72$ and $N_c = 12$. α is the concavity parameter. Higher α values correspond to more concave profiles and α values with the same absolute value but with negative sign to specularly convex profiles, with $\alpha_3 > \alpha_2 > \alpha_1 > 0$. Real data show a profile similar to $-\alpha_2$.

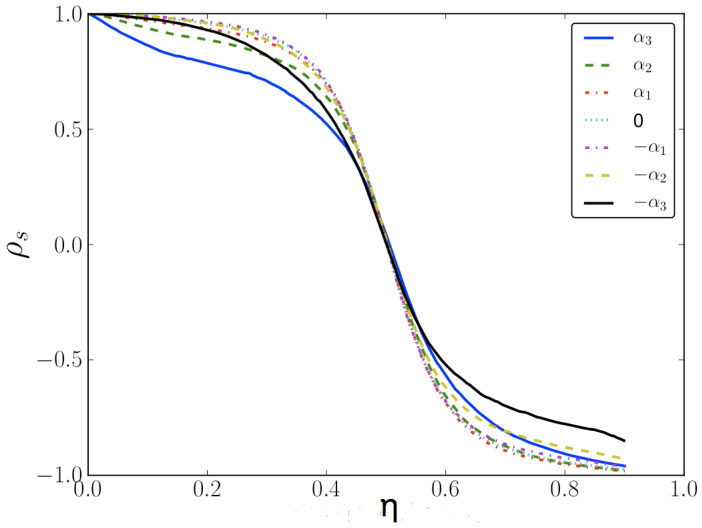


Figure 23: Computation of Spearman's coefficient for F with respect to the level of noise for adjacency matrix with convex and concave profiles. Concavity of M does not affect significantly the fitness' ranking.

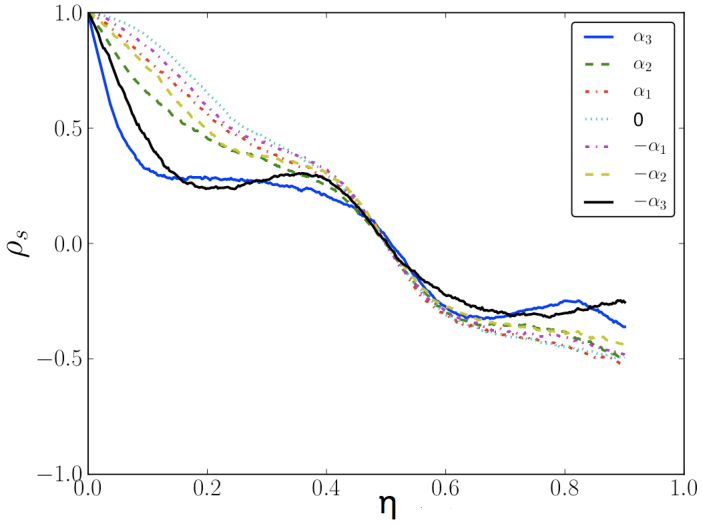


Figure 24: Spearman's coefficient for k_c varying noise for adjacency matrix with convex and concave profiles. k_c shows sensibility to the exact shape of the non-empty region's profile.

of testing the sensitivity of the metrics in the presence of noise in every single bit. However, the nonlinearity of the metrics preserves the economic complexity ranking from single errors of the kind. Instead, it is interesting to study what happens if noise is able to unhinge the structure of a sector that is spatially contiguous into the C-P network. Let us consider a BM matrix and let us divide M in 4 sectors (see fig. 25).

We then study noise sensitivity for each sector separately. ρ_s is valued up to a maximum noise corresponding to 50% of the dimensions of the sector itself. Fig. 26 reports the trends of the 4 sectors and compares them with the case where noise is uniform all over the matrix.

In accordance with the phenomenological criteria that determined the mathematical form of the nonlinear iterative process in order to calculate economic complexity, sector A (complex countries and non-complex products) is the least sensitive to the presence of noise. Although some low complexity products do disappear from the export basket of more complex countries, this is substantially insignificant to determine the fitness of these countries. The other three sectors, on the contrary, show that the presence of compound noise in certain M zones can produce rankings with a more significant number of errors. Coherently with what we would expect, sector D (non-complex countries and complex products) is the most sensitive to the presence of corrupted data. The two border sectors B (complex countries and complex products) and C (non-complex countries and non-complex products) show an analogous trend, sector B being slightly more robust due to the asymmetry induced in C-P network by the nonlinearity of the metrics. This difference between B and C is even more accentuated if we consider matrices with concavity greater than 0 instead of a perfectly triangular one (see fig. 27).

In this case in fact border sector B is characterized by complex countries separated by wide gaps of diversification. The presence of a small level of noise in this sector is thus unlikely to mix up the ranking of countries. On the contrary, sector C, which now presents many poorly

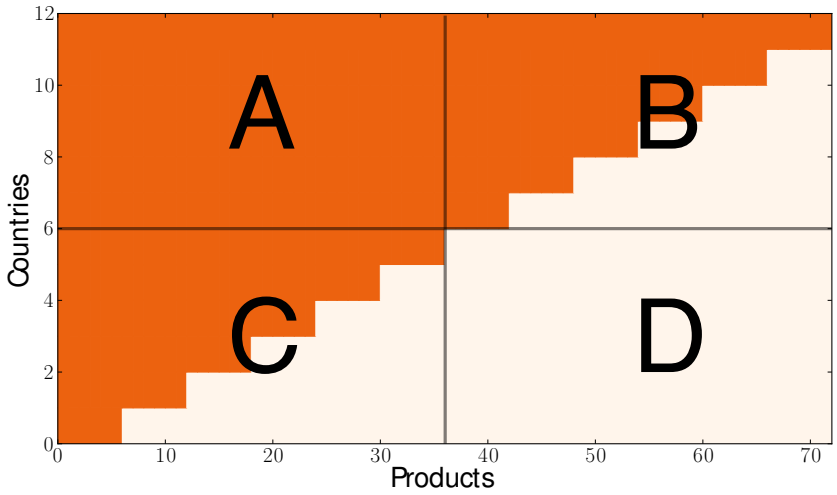


Figure 25: Division of M in four sectors. Sector A: complex countries and non complex products. Sector B: complex countries and complex products. Sector C: non complex countries and non complex products. Sector D: non complex countries and complex products.

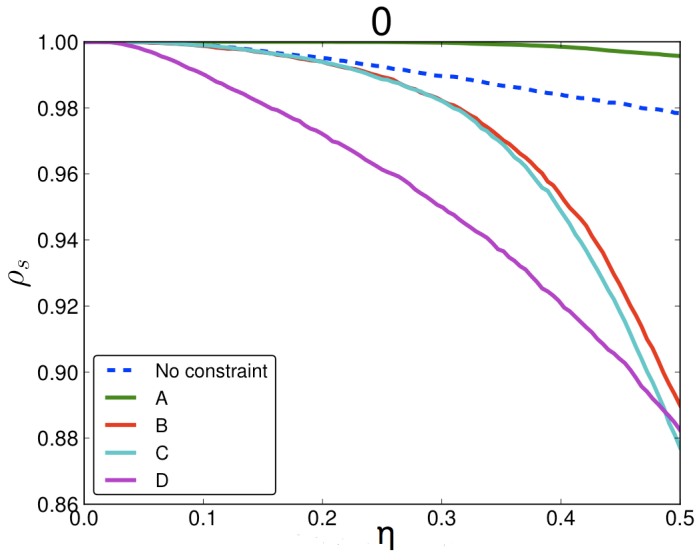


Figure 26: Computation of Spearman's coefficient for F by varying the fraction of noise separately in sectors A, B, C and D and comparison with the case where noise is uniform all over the matrix. Sector A is the least sensible to the introduction of a small amount of noise since less complex products do not weigh much in determining the fitness of complex countries.

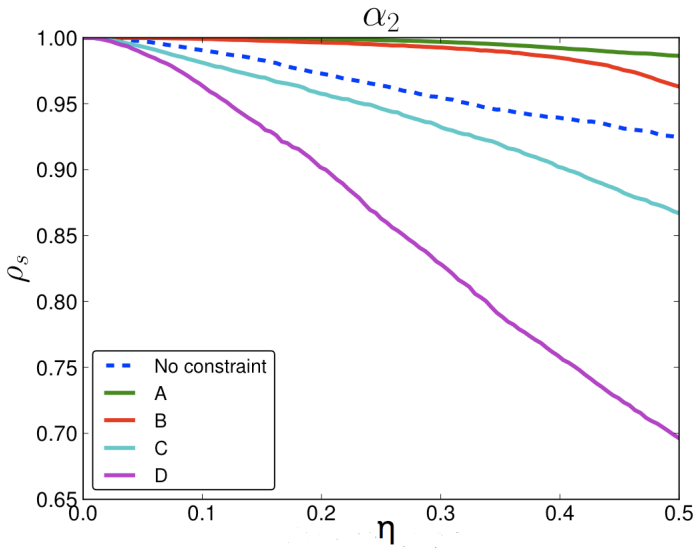


Figure 27: Computation of Spearman's coefficient by varying the fraction of noise separately in sectors A, B, C and D and comparison with the case of noise not subject to spatial ties for a matrices with slightly concave profiles. Sector B becomes less sensible to noise since complex countries are separated by wide gaps of diversification.

complex countries with a very close fitness value, became more sensitive to small perturbations. Sectors A and D, not directly involved in the M profile alteration, remain respectively the most robust and the least robust. Specularly, a convex bulge makes sector C more robust to the detriment of sector B (see fig. 28). Now that this sector (complex countries and complex products) is characterized by many countries with minimum levels of diversification, noise turns out to be even more incident than in sector D.

Generally speaking we can conclude that the spatial localization of noise is a significant factor and, when dealing with economic complexity measures, it is necessary to consider the reliability of data concerning exportations of countries and products in the C-P network.

Noise estimation in real data

Previously we saw that the sensitivity trend of the NL metrics when $V_{M_{c,p}} \to \infty$ is the reaching of a plateau with values of $\rho_s \approx 1$ for $\eta < \bar{\eta}$. Instead, correlation falls dramatically for noise values greater than 50%. Hence, we must estimate the fraction of noisy data within the database currently used in the C-P network analysis. To do so we must compare the measures that are operable on the C-P network with analogous measures made on the toy-matrices of the BM for economic complexity as η changes. One possible way is to intersect the two measures of complexity calculating the correlation between the two asymptotic values of F and k_c. Data on M have been in general significantly adjusted and corrected in order to be completely consistent with the import-export data of countries. We then are able to produce these results for both raw and cleaned data. These measures turn out to be uncorrelated to market trend

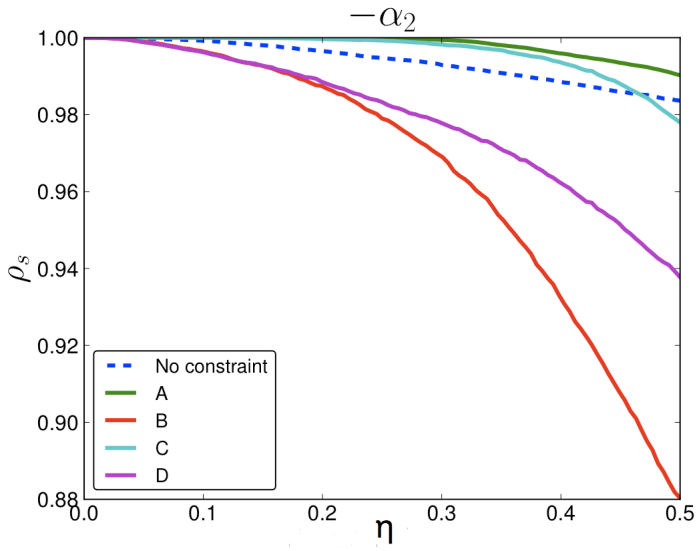


Figure 28: Computation of Spearman's coefficient by varying the fraction of noise separately in sectors A, B, C and D and comparison with the case of noise not subject to spatial ties for matrices with slightly convex profiles. Sector C becomes less sensible to noise since complex countries are separated by wide gaps of diversification.

and do not show signs of the various economic conjunctures crossed by global economy in the data concerning the years we dispose of (1995-2010, fig. 29).

For the reasons highlighted in this appendix, the correlation ρ_s between F e k_c decreases very quickly as η increases in the matrices of the zero model for economic complexity. This property allows us to estimate the fraction of noisy bits within the adjacency matrix.

As fig. 30 shows, we can estimate the fraction of noisy bits within the C-P network by calculating for which level of noise the Spearman's rank correlation coefficient between F and k_c computed for the BM matches the value from the real matrix. We can estimate the level of noise across years 1995-2010 to be approximately 7% for raw data and 5% for cleaned data (see fig. 31). Moreover, these results allow us to be confident in our operation of adjusting and cleaning raw data. As far as the analysis of the robustness of the two metrics goes, these value are well in the plateau for NL metrics, while they affect significantly the accuracy of HH metrics' measures.

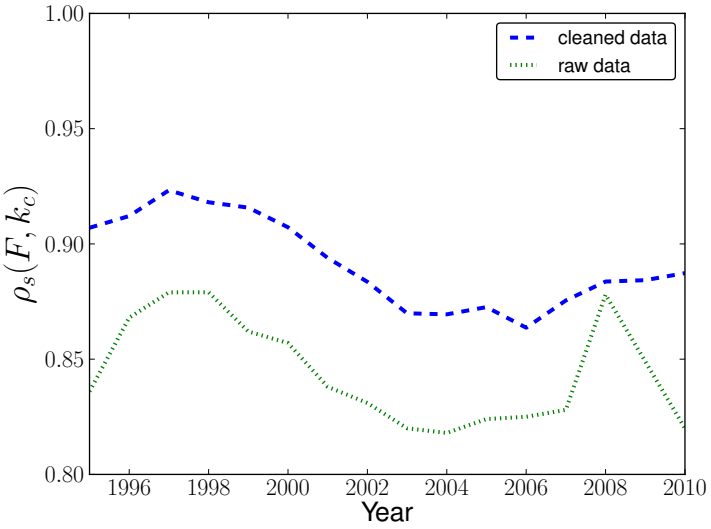


Figure 29: Correlation between asymptotic measures of F and k_c since 1995 to 2010 for cleaned and raw data.

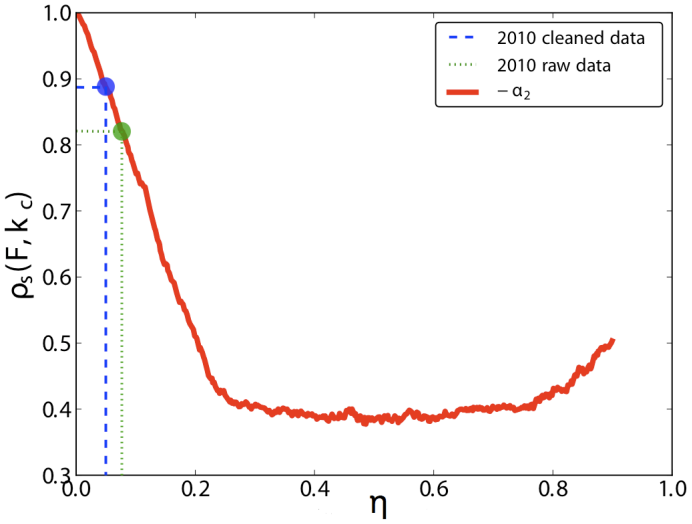


Figure 30: Computation of Spearman's coefficient between F and k_c for matrices with slightly convex profiles by varying the level of noise and comparison with 2010 data. By intercepting the measures on real data with the one on the noisy toy-matrices we can estimate the level of noise in the real dataset.

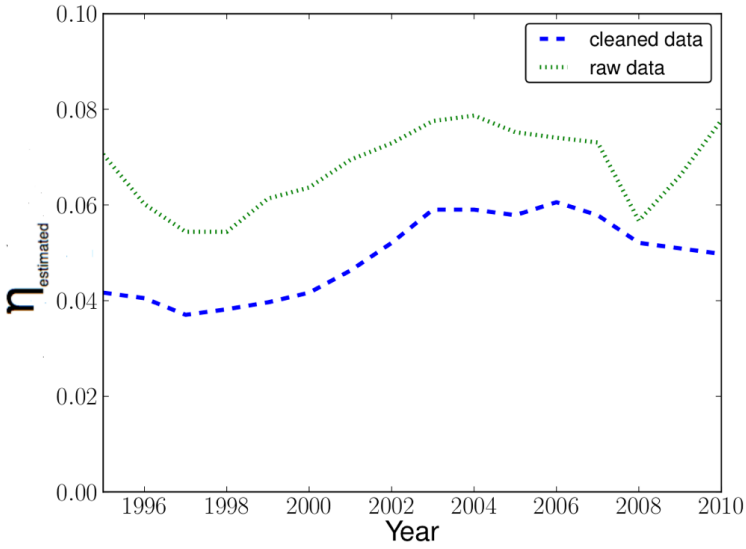


Figure 31: Estimated level of noise on real matrices since 1995 to 2010 for cleaned and raw data.

2 HETEROGENEOUS DYNAMICS IN ECONOMICS.

CHAPTER ABSTRACT

A fundamental test for the validity of the conceptual and technical innovations that we propose in this work, is their actual ability to make predictions. This is the main topic of this chapter.

To be clear, the kind of questions that we want to address are very basic from an economic point of view: e.g. "What will be the growth of the Gross Domestic Product (GDP[1]) of China, United States, and Vietnam in the next 3, 5 or 10 years?" Despite this kind of questions has a large societal impact and an extreme value for economic policy making, providing a scientific basis for economic predictability is still a very challenging problem.
The standard approach used in economics usually relies on the accumulation and display of large datasets and time-series of standard indicators. While data-oriented approaches are for sure at the basis of a scientific grounding, what is lacking in standard economics is a unified view on tools and models that allow a univocal interpretation of the gathered data in terms of predictions. The metrics described in Chapter 1 overcome this issue by summarizing all the data into a single number for each country. The limitations of such a synthetic picture make rise an obvious question: what is the amount of information content that this single number is able to carry? A standard approach to this kind of question, used in economics, would be to check what is the differential increment in the R^2 of a multivariate linear model when the new variable is added to the picture. Simplifying, the role of the exceeding variables in the linear model is to discriminate the effect that the new indicator has on the dependent variable. In other words it is an attempt to take into consideration the heterogeneity of the responses in terms of the dependent variable of the sample to the value of our new regressor.

The approach that we propose here is much more radical because, while being generalized to any kind of response, not just linear, it takes into consideration that the heterogeneity is a dynamical feature: the heterogeneity is *in* the dynamics and *is* dynamic. To get a flavor of what the idea is one can think of weather forecast. It is a fact that weather forecasts have a much better reliability in some geographical areas than in others (heterogeneity of the dynamics). So we might not be able to make good predictions of the dynamics of a set of clouds while it lays in a turbulent area, but we might be able to forecast its evolution quite well once it enters a laminar-flow area (dynamic heterogeneity).

To make this ideas quantitative, in this Chapter we explicitly study the dynamics of countries in the *fitness-income per capita* plane. We observe that country dynamics presents strongly heterogeneous patterns of evolution. The flow in some zones is found to be laminar while in others a chaotic behavior is instead observed. These two regimes correspond to very different predictability features for the evolution of countries: in the laminar regime, we find strong predictable patterns while the latter scenario exhibits a very low predictability. The fact that a

1 For the sake of brevity and readability we will often refer to GDP even when talking about Per Capita GDP. When a distinction is needed it will be made clear by the use of the pedix: GDP_{pc}. However most of the times, since we will be dealing with growth rates, this distinction will be superfluous, given the fact that demographic changes are mostly irrelevant in the time scales that we consider.

laminar regime exists in the GDP-Fitness dynamics means that, for the countries living in that regime, the knowledge of fitness (and obviously GDP) on a given year.

Regressions, the usual tool in economics, are no more the appropriate strategy to deal with such a heterogeneous scenario and new concepts, borrowed from dynamical systems theory, are mandatory. We therefore propose a data-driven method - the *selective predictability scheme* - in which we adopt a strategy similar to the *methods of analogues*, firstly introduced by Lorenz, to assess future evolution of countries.

2.1 INTRODUCTION

Which are the key ingredients determining the economic performance of a country and its future development? Economists traditionally measure the performance with monetary figures such as the gross domestic product (GDP) reflecting, at most, the actual status of a country. The assessment of the evolution and growth of countries is instead highly controversial. Many intangible elements are therefore invoked such as good education, financial status, labor cost, high tech industry, energy availability, quality of life, etc. However, these concepts are usually discussed in a qualitative way.

A longstanding objective of macro-economic theories is the development of predictive schemes in order to give a quantitative assessment of the future evolution of economic indicators, such as income, Gross Domestic Product (GDP), inflation rate, etc and to provide criteria and indications for economic interventions, stimuli, growth incentives, etc. It may appear surprising that country growth forecast, even if crucial for the wealth of nations and people, is still a controversial open question, as witnessed by the recent critical analysis of [19, 20].

In this Chapter we discuss the non-monetary metrics introduced in Chapter 1 can provide a fundamental analysis of the hidden potential of growth for countries. This approach is based on the idea that the productive basket of a country is able to discount and reflect (almost) all the information encoded in the intangible assets, usually hardly modelable, driving competitiveness.

When we compare monetary figures, as the GDP *per capita*, with this metrics of country intangibles, forecasting economic growth faces issues very similar to those of weather forecasts and, in general, conceptually resembles the challenge of forecasting the evolution of a dynamic system. In this perspective, we are going to show that there exists a strong evidence of a high degree of heterogeneity in the dynamics of countries in the plane defined by the fitness and the GDP *per capita* (hereafter we use income and GDP *per capita* in an equivalent way). This observation calls for a completely new framework for the predictability criteria making regression-based approaches inappropriate to address the heterogeneous dynamics of the country growth. We propose to call this new approach *selective predictability scheme*. Loosely speaking, this predictive scheme recalls concepts from dynamical systems such as effective dimensions of the phase space and the *methods of analogues* [21, 22] which are typical tools in scenarios in which the laws of evolution are unknown and only a set of empirical observations of the evolution is known. Depending on whether a country has a lower or higher level of income compared to expectations based on its level of fitness, we are able to detect strong and stable evolution trends of countries in specific regimes.

The emerging picture from the *selective predictability scheme* also suggests that, rather than simply overcoming money-based measure of wealth and development (namely GDP) by substituting them with new indicators [23], a more scientific approach would consist in complementing GDP-based measures with new dimensions. In such a way it would be possible to compare purely monetary information with non money-driven indicators to detect informative content about intangible features. Mathematically speaking, this corresponds to project the economic dynamics of country evolution onto a suitable multi-dimensional problem, as done in the case of the fitness-income plane, rather than simply projecting on a one-dimensional indicator alternative to the GDP. It is also worth noticing that such substitutive approaches usually face the issue of mixing heterogeneous indicators whose commensurability is often questionable and problematic.

The features of the economic evolution of countries in the fitness-income plane are the starting point of our analysis. As shown in Fig. 32 and more in detail in Section 2.A, a first valuable result is that countries in this plane, defined by fitness and income, are not evolving to an equilibrium situation - the equilibrium would correspond to the case in which all countries are on a straight line in the fitness-income plane - at least on the time horizon investigated, i.e. 1995-2010. As a consequence of this fact, the first observation is that the distance between the real income of countries and the expected one is not directly a measure for the potential of growth of countries. This means that both the fitness and this distance are not good candidate as regressors in a standard approach. Further considerations about the poor performances of a regressive approach with these variables will be made in Sec. 31.

However, as argued, the features of the evolution of countries in the fitness-income plane call for a scheme in which regressions are no longer the appropriate tool to address the issue of GDP growth forecast. Standard approaches traditionally assume the existence of overall trends to uncover. The assumption underlying a regression-based approach as the correct analysis tool, is that the system responds homogeneously to a specific set of (independent) variables, explaining a certain amount of the variance of the dependent variable. In other words, regressions may be appropriate tools to analyze systems whose dynamics is homogeneous.

To investigate how income depends on fitness and vice-versa, we must move from a static picture to a dynamical investigation of the countries in the income-fitness plane. In Fig. 32b we represent the income-fitness plane with the trajectories of all countries from 1995 to 2010. A better understanding of the dynamics in this plane can be achieved by a coarse graining of the trajectories. We build a vector-like representation of the movements in the income-fitness plane by dividing the plane in a grid, as shown in Fig. 32c. Each arrow represents the average of all 1-year displacements whose starting point belongs to the box of the grid corresponding to the arrow.

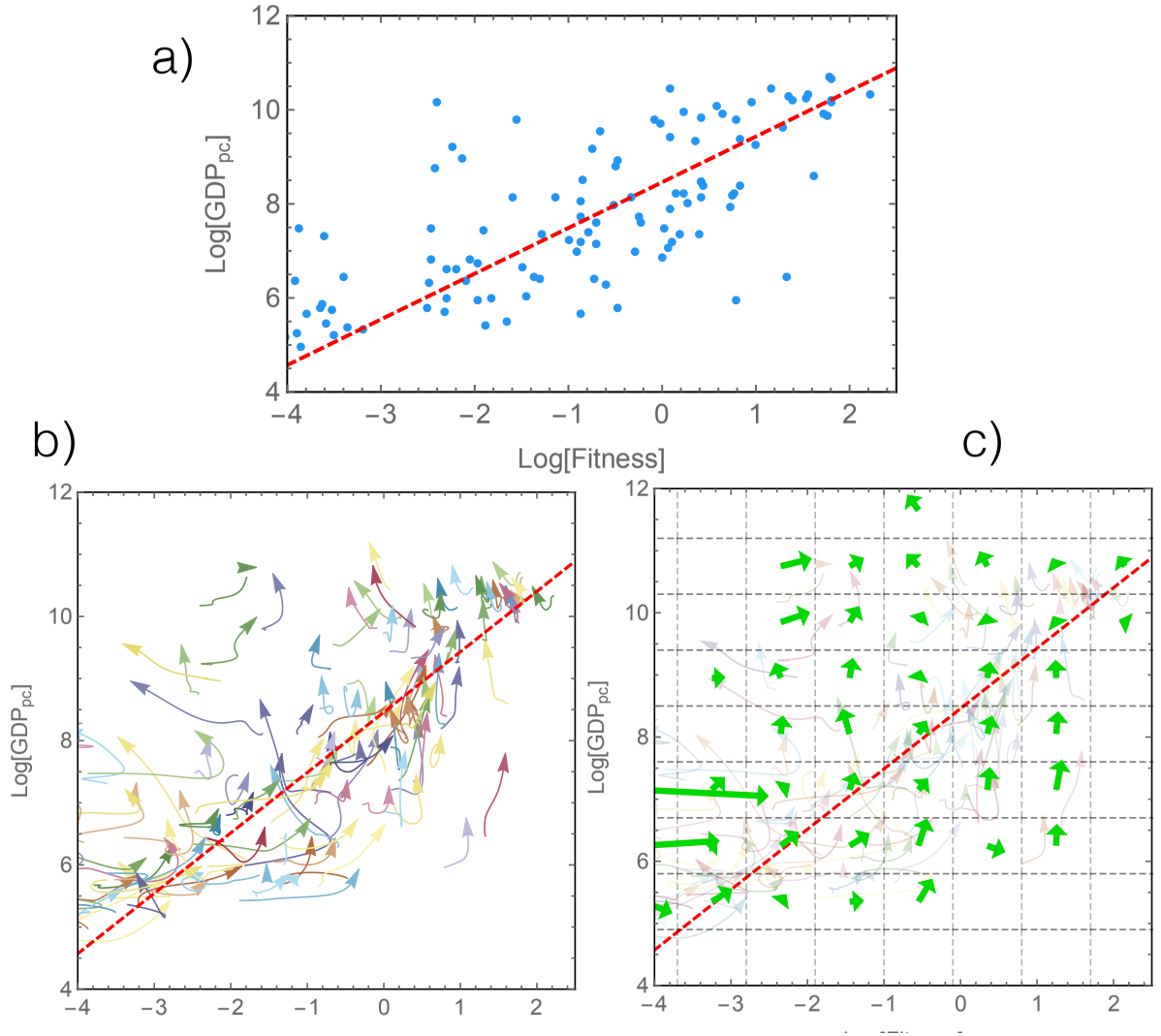


Figure 32: Heterogeneous dynamics of economic systems in the fitness-income plane. Panel a: we report in the fitness-income plane the position of the countries in 1995. The red line indicates the expected level of income, given the level of fitness of a country, and it is the result of the minimization of the Euclidean distance of the countries from the line weighted by the country GDP. Panel b: evolution in the fitness-income plane from 1995 to 2010. We observe a strongly heterogeneous dynamics of the countries in this plane. In order to point out emergent trends in this dynamics, we perform a coarse graining of the trajectories, as shown in panel c. A laminar-like regime is observed. With respect to the evolution of the countries with intermediate/large fitness, this regime is characterized by a regular flow and an income lower than what expected from the average red line (top left corner).

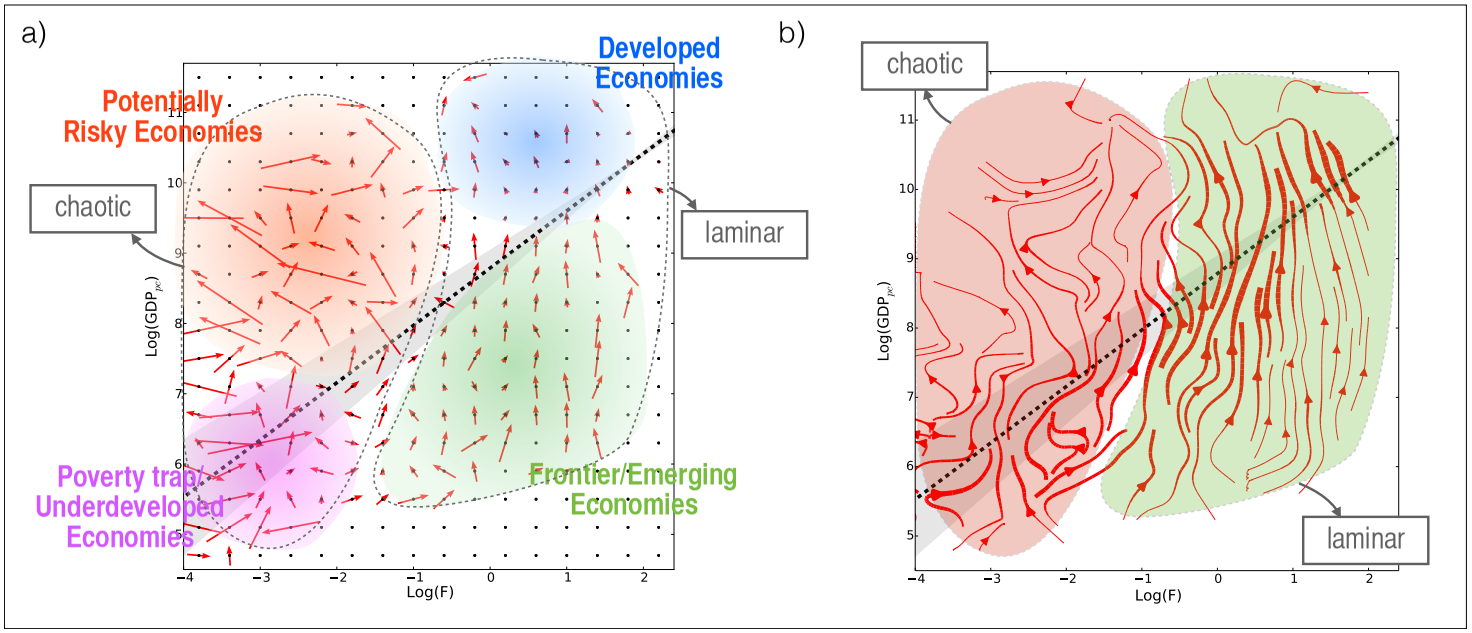


Figure 33: The four regimes of the heterogeneous dynamics of economic complexity: Panel a) A finer coarse graining of the dynamics highlights two regimes for the dynamics of the evolution of countries in the fitness-income plane. There exists a laminar region in which fitness is the driving force of the growth and the only relevant economic variable in order to characterize the dynamics of countries. We argue that the evolution of countries in this region is highly predictable. There is also a second regime, which appears to be chaotic and characterized by a low level of predictability. In the laminar regime we also find two different kinds of evolution patterns for the emergent countries and developed ones respectively. In this heterogeneous scenario for the economic dynamics of countries, regressions are no more the appropriate tool to develop a predictive scheme, which instead must face issues which are very close to the problems of predictability for dynamical systems (i.e. atmosphere, climate, wind, ocean dynamics, and weather forecast, etc). Panel b) we report a continuous interpolation of the coarse grained dynamics to better illustrate the two regimes of predictability

While the motion start to become clearer at this level of coarse graining, interestingly by reducing the size of the boxes an evident heterogeneity in the dynamics regimes start to emerge. The analysis of country evolutions in the income-fitness plane reveals a strongly heterogeneous behavior (Fig. 32). We observe the emergence of different regimes of country evolution and development, depending on the relative value of fitness and wealth. On this account, according to the position in the income-fitness plane, we are able to distinguish two main regimes: a laminar-like regime, in which the dynamics appears to be predictable and informative of the country growth (green and blue shades) and a chaotic-like regime in which no clear emergent pattern is observable (purple and red shades). A more careful analysis reveals an even richer ecology and at least four different types of dynamics can be uncovered:

- **Very low fitness regime (Fig. 32a purple area)**: countries are stuck into a poverty trap. Their industrial competitiveness is irrelevant with respect to many other exogenous factors.
- **Low/intermediate fitness regime (Fig. 32a red area)**: for these countries industrial competitiveness is still scarcely relevant with respect to other exogenous factors, which, in this case, have a positive effect. Most of the exporters of heavy natural resources lie in this area.
- **Intermediate fitness regime and income lower than expected (Fig. 32a green area)**: it appears that the fitness is the main driving force for economic growth.
- **High fitness and high income regime (Fig. 32a blue area)**: this region corresponds to developed countries, the flow is still laminar in the income-fitness plane but the dynamics, even if predictable, is of a different kind with respect to the previous regime.

The heterogeneity of the income dynamics in relation to the value of the fitness has several conceptual and practical consequences on how to carry out a predictive scheme for the trajectory evolution. As mentioned, regressions are appropriate tools when we observe homogeneous responses to specific variables but they are not effective in such a heterogeneous scenario. On one hand, in the chaotic-like region of the income-fitness plane there is no clear dynamic relationship between the two variables. On the other hand, even in the laminar-like regime - the intermediate fitness regime with income lower than expected and the high fitness regime - we observe two different types of behavior across emerging and developed countries.

We therefore propose to define a *selective predictability scheme* in which the degree of predictability of the economic dynamics depends on the specific position in the income-fitness plane. The sharp observations made in Fig. 32 suggest that this scheme faces issues which are very close to the problem of forecasting the evolution of dynamical systems (i.e. atmosphere, climate, wind, ocean dynamics, and weather forecast, etc) when the laws of motion are unknown. We believe that the present framework opens new paths towards providing more scientific basis for economic predictability.

Finally the existence of a laminar flow area tells us that the Fitness is a very informative measure of the economic status of a country. While for low fitness countries we need further variables and elements in order to be able to make predictions on the dynamics of single countries, in the high Fitness area the motion is smooth in the income-Fitness plane: in this case the knowledge of only two variables, fitness and GDP, allow to make much more accurate predictions.

2.2 RESULTS

How to predict the heterogeneous dynamics of the economic complexity: the *Selective Predictability Scheme*.

Let us now discuss the results of Fig. 32, borrowing concepts and methods from the theory of dynamical systems. In terms of the jargons of dynamical systems [21, 22, 24], in the laminar regimes (green and blue shades of Fig. 34), we argue that the effective dimension of the phase space of the dynamics is approximately two and, in this perspective, we are looking at the right two dimensions for the dynamics. In other words, the economic dynamics is, in general, a highly dimensional problem but in the laminar regime the effective dimension of the phase space of this dynamics is much lower.

For the chaotic-like regime we can find two different explanations.On one hand, it could be that the effective dimension of the phase space is still two and the dynamics is indeed chaotic.

On the other hand, it might be that in this region the dimension of the phase space of the dynamics is much larger than two. Therefore the trajectories in the fitness-income plane are the projection of the $d-$dimensional dynamics onto a two-dimensional space. The dynamics appears to be chaotic-like, as the result of the large dimension of this space, because we are not able to see any real recurrences, as intended by Poincaré in his theory [25]. According to this second interpretation, trajectories which appear close in the projected two-dimensional space are instead, differently from points and trajectories in the laminar-like regime, very far in the real $d-$dimensional ($d \gg 2$) space. Therefore these trajectories are only apparently good candidate to be *analogues*, that is close points in the whole space of the evolution. Roughly speaking, the higher is the dimensionality of the space phase, the harder is to find a good analogue - a point close enough in the phase space - to infer the future form the past.

Translating these arguments in economical terms, we believe that the latter interpretation better fits our scenario. In the laminar-like regime, the fitness is the relevant economic variable in order to understand the dynamics of the income and in general of the growth of the GDP. In the other regime, instead, the dynamics is ruled by several exogenous factors which compete with the fitness in driving the evolution of countries.

It results that the predictive scheme, required by the dynamics of economic complexity, is analogous to the problem of predicting the future of a dynamical system in the case in which we do not know the equations of motion (i.e. the rule of the evolution). The best strategy is therefore to try to predict the future from the knowledge of the past: this method is called *method of analogues* and was firstly introduced by Lorenz [21, 22].

As a first step to implement a predictive scheme inspired by the method of analogues, we investigate the 10-years growth of the GDP *per capita* in the income-fitness plane (in sec 2.C we also discuss the 5-years growth case for which similar results hold).
Following the line of reasoning of Fig. 32, we can perform a coarse graining of the trajectories dividing the plane into square boxes. Let us now suppose that in a specific year we find a given number of countries in a given box and we record where these countries evolve in the following $n-$years – as mentioned, we now discuss the case for $n = 10$ years.

Some more detailed discussions on methods, calculations, definitions and robustness of the findings about the *selective predictability scheme* are reported in the Appendices to this chapter.

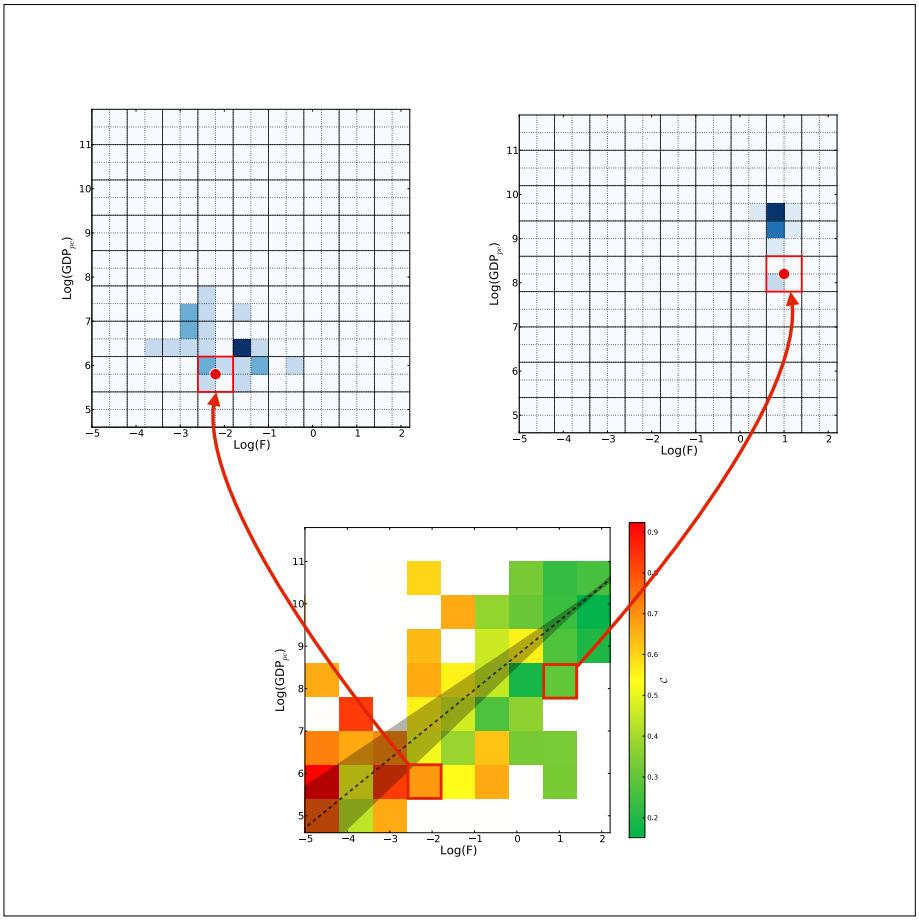


Figure 34: The *selective predictability scheme*: the scheme is based on the *method of analogues* [21, 22] which is a strategy to predict the future from the knowledge of the past. We track the evolution of the countries in a given box and we record where these countries have evolved in the following 10 years. By carrying out systematically this procedure for each box, we build the empirical distribution of the 10-years evolution of countries. In this figure we report the empirical 10-years distributions (ED) for two boxes: the former from the chaotic-like regime and the latter from the laminar-like one, as defined in Fig. 33. A visual inspection reveals that the EDs from the chaotic-like regime tend to have a larger dispersion than the ones from the laminar one, which are instead very concentrated in few boxes. To quantify this effect we introduce a measure of concentration for the EDs: $\mathcal{C} = (n^{(i)}_{boxes}/N^{(i)} - 1/N^{(i)})(1 - 1/N^{(i)})^{-1}$, where $n^{(i)}_{boxes}$ and $N^{(i)}$ are respectively the number of occupied boxes by the ED associated to the box i and the number of points giving rise to ED. This measure confirms the existence of two regimes characterized by two very different levels of predictability: a laminar regime (green boxes) for which the flow is regular and tends to be concentrated in few boxes and a chaotic regime characterized by very dispersed distributions of the country evolution. We argue that in the first regime the fitness is the key ingredient to understand the evolution of the economic systems.

By carrying out systematically this procedure for each box, we can build the empirical distribution of the 10-years evolution of the countries, box-by-box. We report the 10-years empirical distributions (ED hereafter) for two boxes from the chaotic-like regime and the laminar-like regime respectively in Fig. 34 (top panels) (a larger sample of EDs are reported in the Appendices to this Chapter). Confirming the heuristic observations drawn in Fig. 33, we observe that the dispersion of the EDs tends to be larger in the chaotic-like regime than in the laminar-like one. Let us try to quantify the degree of predictability of the 10-years evolution in this plane for each box by measuring the concentration of the EDs we have obtained. The simplest way to define such a concentration is by counting the number of occupied boxes, using the ED associated to a given box i and normalizing the results with the total number of points in the box i. We would therefore obtain the highest concentration, and therefore the highest predictability from our point of view, if all the countries in a box would evolve in the same arrival box. As for information theory entropy which is zero when the system is completely predictable, we then define our concentration measure as

$$\mathcal{C} = \frac{\frac{n_{boxes}^{(i)}}{N^{(i)}} - \frac{1}{N^{(i)}}}{1 - \frac{1}{/N^{(i)}}}$$

where $n_{boxes}^{(i)}$ and $N^{(i)}$ are respectively the number of occupied boxes by the ED associated to the box i and the number of points giving rise to ED. In Fig. 34 bottom panel, we show the predictability for each box as measured by $1 - \mathcal{C}$ (see Supporting Information for further considerations on this point). In general, we find that all the concentration/predictability measures we have tested confirm the heuristic visual grouping discussed in Fig. 33: there exists a region of the fitness-income plane in which the dynamics is laminar and is characterized by a high degree of predictability, by contrast the chaotic region exhibits broader EDs with a large dispersion corresponding to a lower degree of predictability of the dynamics.

At this stage, considering the 1995-2010 evolution, we have confirmed that there exist regions for which the EDs are very concentrated. Now we want to test if the EDs built from past evolution of countries are a good tool to assess the issue of the growth forecast. Back into the jargons of dynamical systems, if we have a box in a region for which the effective dimension of the economic dynamic is close to 2 – fitness is the only driving variable for the dynamics of country development – we expect that the ED of this box is a good proxy for evaluating the evolution of a country. In fact in such a case, the ED would be estimated from a set of points which are good candidates to be analogues. These points are not only close in the fitness-income plane but also in the real space of the economic dynamics and therefore they are ruled by a similar economic regime, even in the projected space. It is worth noticing that we are also assuming that the EDs vary on a time horizon longer than the one under investigation. We expect that the boxes with highest predictability, as measured by the concentration $\mathcal{C}$ in Fig. 34, are also the regions for which we expect EDs to be reliable instruments to predict growth of countries. Such a framework does not merely forecast the GDP growth (with an appropriate uncertainty) but the future evolution of the country's trajectory in the fitness-income plane. To verify this hypothesis, we perform a back-test analysis of our methodology. Given the time length of our dataset, we cannot perform the training of our predictive scheme on a 10-years horizon because we would not have enough years in order to consider an independent out-of-the-sample set of 10 years. We therefore back-test our method on the 5-years time horizon and we use our series in the following way: we train the method on the period from 1995

to 2005 and we build the 5-years EDs for each box. Once the 5-years EDs are estimated, we test how successful they are in forecasting the evolution of countries in the period 2005-2010. The results, shown in Fig. 34a, confirm our arguments and the existence of high predictable regimes, where our scheme is effective and of low predictable regimes, where the fitness is no more the key driving ingredient of the development of countries. This overall picture for the economic dynamics and its heterogeneity could not be properly grasped by a regression-based approach. In panel b) of Fig. 34, we report the relative error of 2010's GDP *per capita* forecast. We observe a systematic underestimation of the growth. We argue that the reasons are twofold: on one hand, the training period is shorter than a complete economic cycle and therefore we are estimating future growth with an incomplete set of information. On the other hand, we expect the systematic over or under estimation to be less likely observed on longer time-horizon (around 10 years). The same histogram for fitness (not reported) instead tends to be substantially bell-shaped and peaked around 0. See Supporting Information for further details on the method used to estimate the 2010's GDP *per capita*.

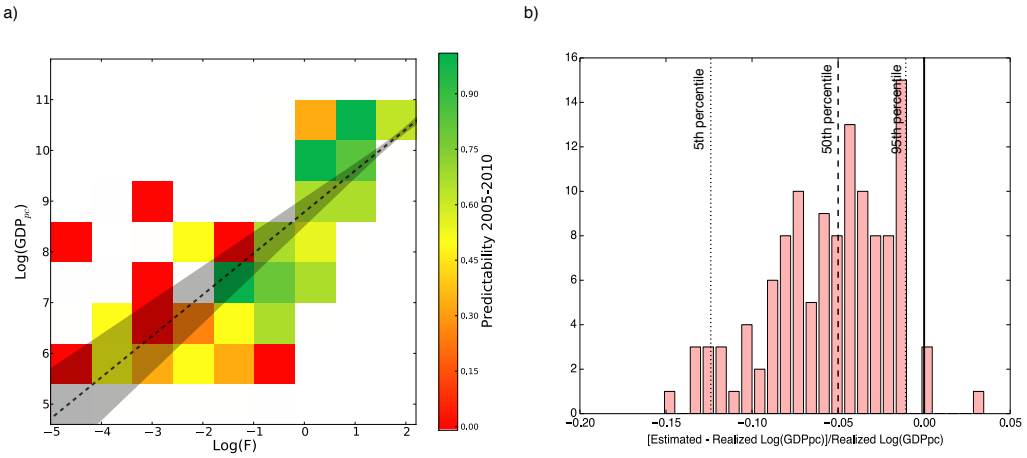


Figure 35: Backtesting of the 5-years *Selective Predictability* Scheme. Panel a), we train the EDs on the time interval 1995-2005. We test the rate of success of the forecast of the position of a country in 2010, according to the ED associated to the departure box of the country in 2005. We compute how many times we are able to guess correctly the box in which the country will be after 5 years. Although the very limited statistics of this back testing procedure, the results confirm the fact that, not only the forecasted area in which we expect to find the country is smaller in the laminar regime, as demonstrated by Fig. 34, but even the dynamics appears to be more predictable because of a general higher rate of success of the *selective predictability scheme* in the laminar region. Panel b), we report the histogram for the relative error of the forecast of 2010's GDP *per capita*. Differently from the fitness histogram (not reported), the distribution is not peaked around 0. We believe this systematic under evaluation of growth is due to a training set shorter than the typical length of an economic cycle.

2.3 CONCLUSIONS

Are development, wealth and growth only a matter of GDP? In the last decade, a growing literature [23] is trying to overcome a description of economic systems in purely GDP-oriented terms, by substituting GDP with new economic indicators. However, here we argue that economic systems are unavoidably *also* a matter of monetary information, given the organization and the rules of modern economies. It would be therefore *naif* to simply neglect monetary dimension in the attempt of assessing the development and wealth of countries. We believe a more scientific grounding of a new economic thinking would consist in a line of reasoning close to the one proposed in this chapter: comparing monetary information with measures of intangible assets of countries. In this chapter we have shown that the growth dynamics of countries in the fitness-income plane exhibits a high degree of heterogeneity and that regression-based analysis consequently are not the appropriate tools for developing a predictive scheme. We argue instead that techniques and methods deriving from the dynamical system theory appear as natural candidates to explain and model the complex dynamics in this plane. One of the main consequences of this approach is that we observe a heterogeneous degree of predictability, depending on the region of the plane considered.

The scheme that we propose for the prediction of the heterogeneous dynamics of the economic complexity resembles the so-called method of analogues, which is a method developed to predict the evolution of a system (typically a dynamical system) given the observation of the past and without the knowledge of the equation of the dynamics. This conceptual framework is also able to give some insights at the base of a regime-dependent predictability of economic evolution. We know from dynamical systems theory that the limit of application of this method relies on the dimension of the phase space of the dynamics. We argue that only in the laminar-like region such analysis can be effective since the effective dimension of the space is approximately 2. In this regime, economically speaking, the fitness is the driving and dominant variable for understanding the growth of countries.

As a final point, we stress the generality of the proposed approach, which can be extended to the analysis of the dynamics of many economic and demographic indicators. This work points towards the development of forecasting methodologies and techniques which have a stronger scientific grounding than standard approaches used in economics.

Last but not least, we also observe that the formulation in terms of dynamical systems solves the problem of estimating causality relations from the the observation of simple correlations in a scenario where, differently from physics and other natural sciences, it is hard to pinpoint cause-effect relations among variables.

APPENDICES TO CHAPTER 2

APPENDIX 2.A FITNESS–INCOME CLOUD FROM 1995 TO 2010

The deviation of the metrics from the monetary information is the key point to uncover the hidden potential of the growth of countries. The natural candidate for this study is the fitness-income scatter plot (i.e. GDP per capita-fitness plane), as shown in Fig. 36 where we report the static plot for all the available 16 years of our dataset from 1995 to 2010. The red line represents an estimation of the expected income of a country given its level of complexity. This line is not a regression in the form $GDP = \alpha F + \beta$ but it is the result of the minimization of the Euclidean distance from the line weighted by the country GDP. At this stage, this line does not represent a statement of cause-effect relationship between fitness and income. However, we will show that by looking at the dynamics in this plane we are able to develop a predictive scheme in some specific regimes.

As also shown by the evolution of the residuals of the minimization in Fig. 37, we do not observe any convergence to an equilibrium situation, that is the convergence of the cloud to a straight line in the fitness-income plane. On the opposite, the variance of the cloud tends to increase in the period under investigation. One of the reason of such increase can be traced back in the fact that in the last years of our dataset we observe that China and other emerging countries are eroding the fitness of western countries and overcoming almost all of them. We are somehow in a sort of changing of the drivers or barycenter of the world economy, which is shifting from western developed countries to Asia. In forthcoming works, we plan to deepen the analysis of the overall dynamics of the cloud on a longer time window.

It is worth noticing that the increase of the residuals is not merely due to the increase of the number of countries considered (from 146 to 148) in the time window investigated; we observe, in fact, the same increasing patterns for the residuals even considering constant the number of countries (146). We also want to stress that, in such a framework, we do not expect that the fitness asymptotically will converge to the value of the GDP *per capita* as in an equilibrium scenario. Furthermore, if a static picture and a convergence of the fitness towards GDP *per capita* are assumed, a non trivial issue of the type *why now?* arises, i.e. why in the last 20-30 years the world went out of equilibrium?

APPENDIX 2.B FITNESS *VS* POPULATION.

In this section, we briefly discuss how and whether the fitness correlates with the population of a country. In Refs. [5, 26], we have already noticed that the *correct* counterpart of the fitness appears to be a monetary intensive measure, such as the GDP *per capita*, thanks to the observation of the scaling properties of the distribution of the fitness. Here, we confirm and support this observation and interpret the non-trivial residual dependence of fitness on population.

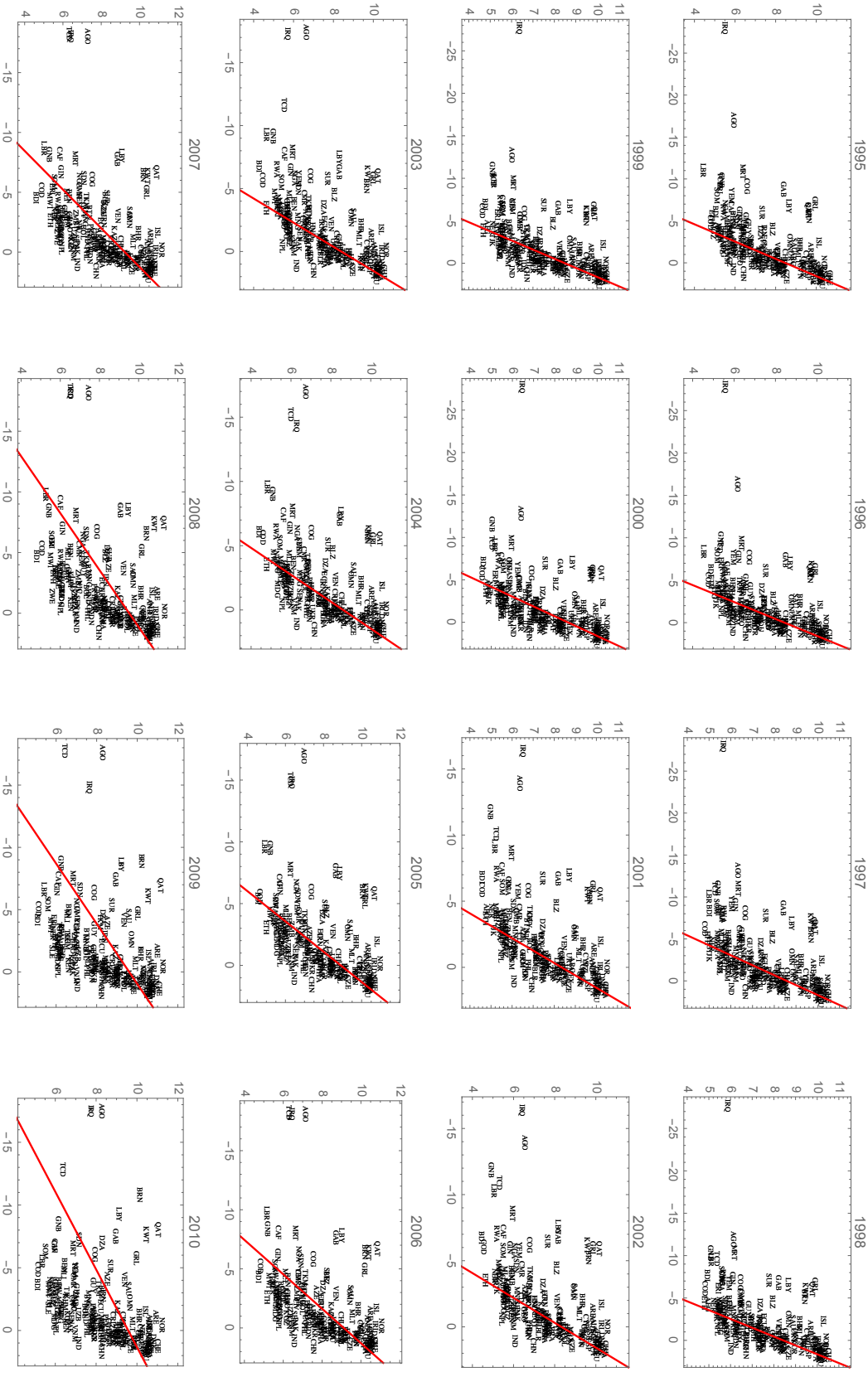


Figure 36: Countries in the fitness-income plane from 1995-2010

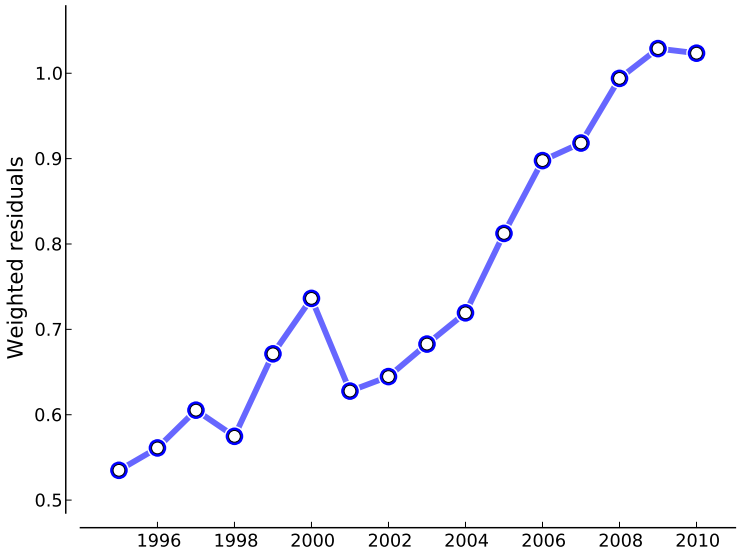


Figure 37: Time evolution of the residuals from the average line from 1995 to 2010: we do not observe, over the period considered, a tendency towards an equilibrium situation in the fitness-income plane, i.e. a straight line. On the contrary, we observe a general increase of the residuals.

We find that population of a country accounts for a small fraction of the variance (around $3-7\%$ in the range of years here investigated) of the fitness. In Fig. 38 we show the scatter plot in 2004 (the shape appears very similar across years), in that year, the logarithm of the population explains approximately the 6% of the variance of the Fitness.

On one hand, we observe that the proposed algorithm removes almost all trivial correlations among the fitness and the size/population of countries allowing for direct comparison of countries and supporting the observation that the monetary counterpart of the fitness is indeed the GDP *per capita* since the fitness almost uncorrelated with the number of inhabitants of a country. For comparison, we find that the population of countries accounts for $11-15\%$ of the variance of the diversification (the zero order of the fitness) in the same range of years.

On the other hand, we can give an economic interpretation in terms of capabilities of the small residual fraction of the fitness' variance explained by the population of countries. We argue that population is a kind of capability and, even once the size effect of a country is removed by our method, it emerges that a residual part of the competitiveness of a country, as measured by the fitness, can be explained in terms of its population. The inspection of the sign of the correlation between the two variables indicates, coherently with the interpretation in terms of capabilities, a positive dependence of the fitness on the population. This results is also supported by the general belief that demographic aspects are one of the key factors necessary for the growth.

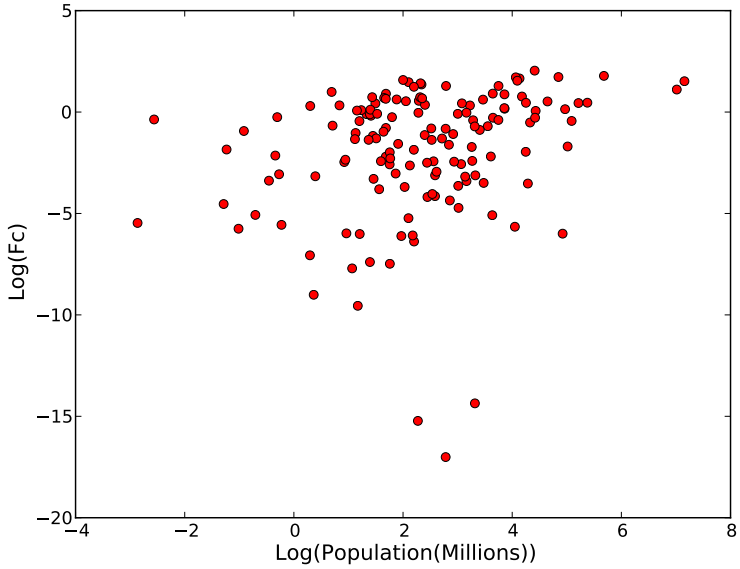


Figure 38: Log(Population) vs Log(Fitness) - 2004. For the regression in which Log(Fitness) is the dependent variable and the Log(Population) the independent one, we obtain $R^2 = 0.06$. The sign of the coefficient estimated by the regression indicates a positive relationship of the residual dependence between these two variables.

APPENDIX 2.C STANDARD REGRESSIVE APPROACH AND HETEROGENEITY.

We now discuss the results which would be obtained from a standard regressive approach in which we test a series of variables and their combination as regressors for the growth of GDP and of the GDP *per capita* over a period of 5 and 10 years.

The regressions performed will be in the following form:

$$\log(GDP_{t+\Delta}) - \log(GDP_t) = c + \sum_i b_i x^{(i)} \tag{31}$$

where $\log(GDP_{t+\Delta}) - \log(GDP_t)$ is the log-return of the GDP, $\Delta = 5, 10, 15$ years, c and $\{b_i\}$ are the coefficients estimated in the regressions and $\{x^{(i)}\}$ are the set of regressors used.
In Table 2 we report the description of the variables used as regressors in the following sections.

We report in the following the results of the regressions. Each line of the tables represents a different regression, the first column indicates the dependent variable of the regression, i.e. the log-return of the GDP, in the last column we report the percentage of variance of the log-returns explained by the regressors used and in the remaining columns we report the coefficients associated to the regressors. A missing value in one of the columns associated to a regressor means that this specific variable is not used in the regression.

Following standard symbols about significance of regression results, we adopt the following convention for the p-values of the coefficients estimated by the regression: $^{*} = p - value < 0.05$, $^{**} = p - value < 0.01$ and $^{***} = p - value < 0.001$.

Table 2: Description of variables used as regressors in Eq. 31.

Name	Description
GDP	GDP at time t expressed in Trillions of USD
GDP_{pc}	GDP per capita at time t expressed in 10^3 USD
log(F)	Logarithm of the fitness of countries
d	Signed vertical distance of the countries from the average line shown in Fig. 36 at time t. Countries above the line get a negative distance, countries below a positive one.

5-YEARS PREDICTION On such time horizon we have three non-overlapping periods, namely 1995-2000, 2000-2005 and 2005-2010 to perform the regression. This corresponds to 439 observations considering the three periods together.

Table 3: Regressions for 5-years GDP growth forecast.

5 years	c	log(F)	d	GDP(Trillions)	R^2
Return GDP	0.180	−0.0744**			0.023
Return GDP	0.328**	−0.0493*		−0.399***	0.069
Return GDP	0.440***			−0.441***	0.059
Return GDP	0.337***		0.00241		0.000
Return GDP	0.453***		0.0136	−0.443***	0.059
Return GDP	0.166	−0.212***	0.234***	−0.312***	0.108

Table 4: Regressions for 5-years GDP per capita growth forecast.

5 years	c	log(F)	d	GDP_{pc}	R^2
Return GDP_{pc}	0.178*	−0.0423*			0.009
Return GDP_{pc}	0.663***	0.0116		−0.0487***	0.126
Return GDP_{pc}	0.630***			−0.0475***	0.126
Return GDP_{pc}	0.296***		0.0348		0.004
Return GDP_{pc}	0.639***		0.0135	−0.0471***	0.126
Return GDP_{pc}	0.654***	0.00717	0.00557	−0.0481***	0.127

10-YEARS PREDICTION In that case, given the range of years investigated, we cannot find more than one period non-overlapping. Therefore we perform two separate regressions on the periods 1995-2005 and 2000-2010 composed of 146 observations each.

15-YEARS PREDICTION For this time horizon, we have only one possible set of data (1995-2010) corresponding to 146 observations.

Table 5: Regressions for 10-years GDP growth forecast.

$95-05$	c	log(F)	d	GDP(Trillions)	R^2
Return GDP	0.411^{**}	-0.0935^{**}			0.055
Return GDP	0.608^{***}	-0.0627^{*}		-0.626^{***}	0.154
Return GDP	0.767^{***}			-0.700^{***}	0.131
Return GDP	0.557^{***}		-0.0546		0.013
Return GDP	0.716^{***}		-0.0376	-0.684^{***}	0.137
Return GDP	0.521^{***}	-0.175^{*}	0.144	-0.555^{***}	0.172

Table 6: Regressions for 10-years GDP per capita growth forecast.

$95-05$	c	log(F)	d	GDP_{pc}	R^2
Return GDP_{pc}	0.368^{***}	-0.0456^{*}			0.026
Return GDP_{pc}	0.591^{***}	-0.0209		-0.0257^{**}	0.070
Return GDP_{pc}	0.659^{***}			-0.0288^{**}	0.066
Return GDP_{pc}	0.460^{***}		-0.0106		0.000
Return GDP_{pc}	0.645^{***}		-0.0110	-0.0289^{**}	0.067
Return GDP_{pc}	0.329	-0.166	0.165	-0.00399	0.087

Table 7: Regressions for 10-years GDP growth forecast.

$00-10$	c	log(F)	d	GDP(Trillions)	R^2
Return GDP	0.532^{*}	-0.167^{**}			0.052
Return GDP	0.612^{*}	-0.154^{*}		-0.242	0.059
Return GDP	0.960^{***}			-0.372	0.017
Return GDP	0.855^{***}		-0.0358		0.001
Return GDP	0.939^{***}		-0.0286	-0.367	0.018
Return GDP	0.232	-0.487^{***}	0.495^{***}	-0.0507	0.135

Table 8: Regressions for 10-years GDP per capita growth forecast.

$00-10$	c	log(F)	d	GDP_{pc}	R^2
Return GDP_{pc}	0.513^{*}	-0.115^{*}			0.028
Return GDP_{pc}	1.558^{***}	-0.00291		-0.124^{***}	0.227
Return GDP_{pc}	1.566^{***}			-0.125^{***}	0.227
Return GDP_{pc}	0.786^{***}		0.0444		0.002
Return GDP_{pc}	1.56^{***}		-0.0119	-0.125^{***}	0.227
Return GDP_{pc}	1.73^{***}	0.0690	-0.0926	-0.137^{***}	0.227

The conclusions which can be derived from these results are twofold. On one hand, all regressions performed appear to have a very poor predictive power on country growth. On

Table 9: Regressions for 15-years GDP growth forecast.

15 years	c	log(F)	d	GDP(Trillions)	R^2
Return GDP	0.579*	−0.182**			0.069
Return GDP	0.665*	−0.168**		−0.275***	0.075
Return GDP	1.09***			−0.471	0.019
Return GDP	0.889***		−0.0850		0.010
Return GDP	0.992***		−0.0740	−0.440	0.027
Return GDP	0.349	−0.575***	0.526***	−0.0165	0.152

Table 10: Regressions for 15-years GDP per capita growth forecast.

15 years	c	log(F)	d	GDP_{pc}	R^2
Return GDP_{pc}	0.496*	−0.128*			0.038
Return GDP_{pc}	1.461***	−0.0214		−0.112***	0.202
Return GDP_{pc}	1.531***			−0.114***	0.201
Return GDP_{pc}	0.771***		−0.0172		0.000
Return GDP_{pc}	1.50***		−0.0188	−0.115***	0.202
Return GDP_{pc}	1.37**	−0.0718	0.0575	−0.104**	0.203

the other hand, the coefficients found are not consistent across different regressions. In the 10-years growth case, the predictive power of the regression is dramatically dependent on the time interval considered and the signs of the coefficients estimated show inconsistencies (see also [3] for sign inconsistencies in the use of regressions for forecasting GDP growth). As discussed in the main text, a regression-based approach answers to the question of unveiling a general homogeneous behavior of the system. Making a parallel with weather forecast, regressions, in such context, would correspond to ask how the weather is in the world tomorrow. Clearly this question is ill-defined, the correct question for the atmosphere dynamics is how weather will be in a specific region/city. In a similar way, we argue that the correct question in the assessment of the growth forecast is the expected growth of a country in a specific economic regime. The expected evolution of countries is dependent on the economic regime in which the country is found to be, as the atmospheric dynamics is dependent on the region we are considering. In this sense, we are in a scenario – the heterogeneous dynamics of economic complexity – where we face issues similar to those encountered in dynamical systems.

Robustness of the heterogeneous regime in the Fitness-Income plane

A necessary condition to obtain a meaningful and successful forecast scheme is that the heterogeneity of the dynamics must show a stability in time. As confirmed in Figs. 39-40, the two regions of the economic dynamics – laminar and chaotic – result to be robust in time. In both figures, we report in red the coarse-grained dynamics discussed in the main text and obtained using the full dataset from 1995 to 2010 for comparison. As a future extension of the present work on a longer dataset, we plan to investigate the evolution of the boundary between the laminar and the chaotic regime.

It is also interesting to observe the small discrepancies among the economic dynamics in the three time windows considered in Figs. 39. This figure supports the idea that the *proper* scale to evaluate the EDs is in the range $10-15$ years in order to average over short time effects due to specific moments of economic cycles. In this sense, the *selective predictability scheme* has a natural time horizon for country evolution forecasting around 10 years, as also confirmed in the following sections.

We also tested the robustness of all analyses with respect to shifts of the grid and different size of the coarse graining and found a substantial independence of our findings on the details of the coarse graining procedure.

Measures of concentration

A natural candidate to measure the dispersion of the EDs would be the entropy, which can be indeed seen as a measure of the concentration of the information of a distribution. However, the correct estimation of the entropy of the EDs critically relies on a robust estimation of the empirical frequencies of the EDs. Simple numerical simulations on toy models reveal that, given the small *typical* level of statistics of the present analysis, the measure of the entropy of the ED would strongly depends on the finite size effects, which affects the empirical frequencies.

We therefore define an *average* measure of concentration, which does not rely on the estimation of the empirical frequencies of the ED as it follows

$$\mathcal{C} = \frac{n_{boxes}^{(i)}/N^{(i)} - 1/N^{(i)}}{1 - 1/N^{(i)}} \tag{32}$$

where $N^{(i)}$ and $n_{boxes}^{(i)}$ are respectively the number of events giving rise to the $i-th$ ED and the number of boxes in which these $N^{(i)}$ evolved after a given time lag. The $\mathcal{C}$ is a normalized concentration measure since it can range from 0 to 1. In addition, the present measure has the advantage to estimate the concentration of the EDs independently on the features of the ED. Instead entropy mixes these two aspects. As a second step of our analysis, to measure how broad or peaked are the distributions arising from the boxes, we use a standard measure of concentration in Economics, the normalized Herfindahl index H^*:

$$H^* = \frac{H - 1/N}{1 - 1/N} \tag{33}$$

which ranges from 0 to 1 and Herfindahl index H is defined as:

$$H = \sum_i p_i^2, \qquad \sum_i p_i = 1 \tag{34}$$

As shown in Fig. 41, although the EDs of boxes from the laminar regime have very similar level of concentration as measured by $\mathcal{C}$, the Herfindahl index of the boxes from the laminar regime shows, in its turn, a non-trivial degree of heterogeneity of the features of the EDs. We recall once again that, independently on the value of the Herfindahl index, these EDs are characterized by an evolution in which the number of final occupied boxes is very small compared to the number of events.

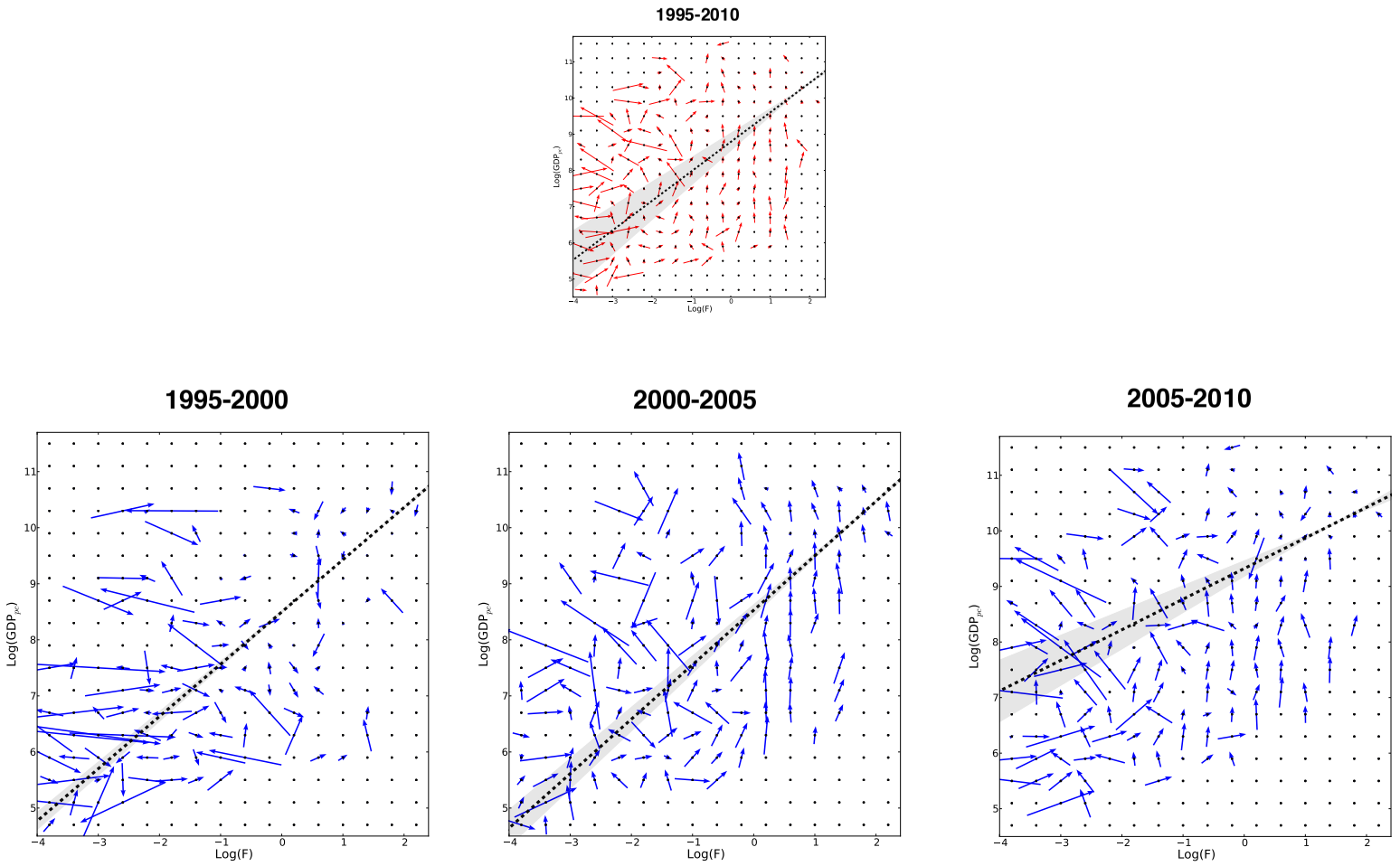


Figure 39: Coarse grained dynamics in the Fitness-Income Plane performed on time windows of 5 years. Top panel: for comparison, coarse grained dynamics from 1995 to 2010.

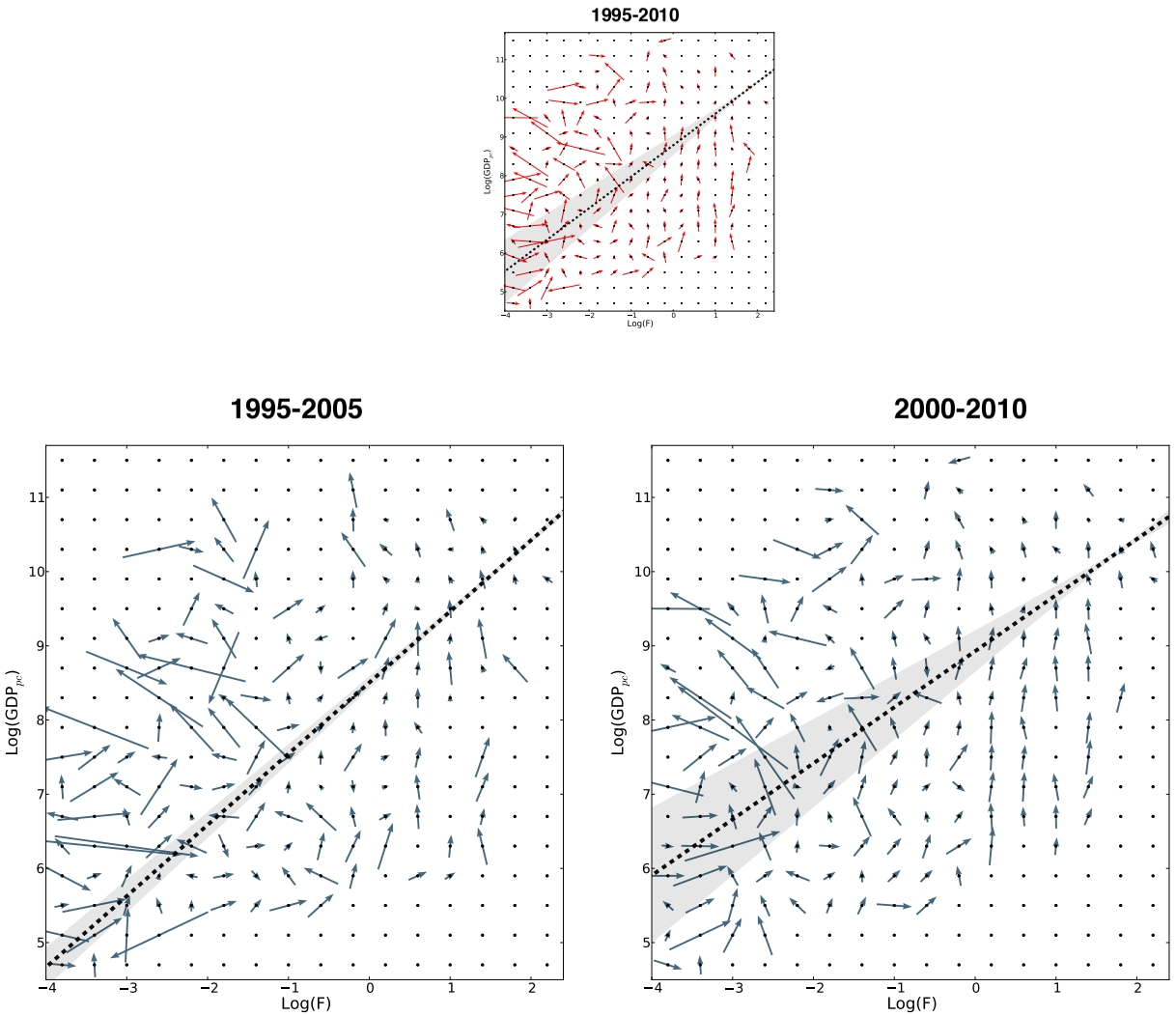


Figure 40: Coarse grained representation of the dynamics in the Fitness-Income Plane performed on time windows of 10 years. Top panel: for comparison, coarse grained dynamics from 1995 to 2010.

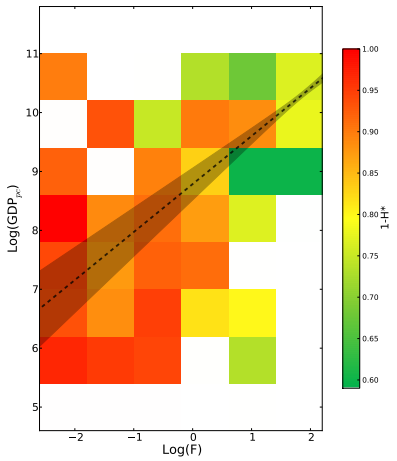


Figure 41: Normalized Herfindahl index of the EDs in the laminar regime, time lag= 10 years. The dynamics of economic complexity of countries in the fitness-income plane also exhibits a non-trivial heterogeneity in the features of the EDs in the laminar regime. $H^* = 1$ corresponds to the case in which all the events are concentrated in a single box while $H^* = 0$ when the distribution is uniform. We plot $1 - H^*$ in order to obtain a quantity which, as the entropy and $\mathcal{C}$, is 0 when all events are in a single box. We report the measure for boxes with at least 4 events.

Estimation of the EDs

As illustrated in Fig. 42, each ED is obtained by considering all countries originating from a box and recording their positions after a certain time lag (in our case 5 and 10 years). For a finer forecast resolution, the grid in which the positions of the evolved countries are recorded has a smaller box size. In our case, we consider a grid for evolved distributions whose box size is the half of the grid defining the starting box of our scheme.

Selective predictability scheme: 5-years prediction

We report in this section the *selective predictability scheme* in the case in which the EDs are built tracking the 5-years evolution in the fitness-income plane. In Fig. 43, we show a selection of EDs, the red squares and dots indicate the starting box, while in Fig. 44 we show the concentration as measured by $\mathcal{C}$ (left panel) of the EDs for all boxes. In the right panel, we also report the entropy for each ED since we have a larger statistics and the entropy estimation is less biased by the finite size effects occurring in the reconstructions of the empirical frequencies. We observe that entropy behavior confirms all the conclusions based on the analysis of $\mathcal{C}$ and Herfindahl index.

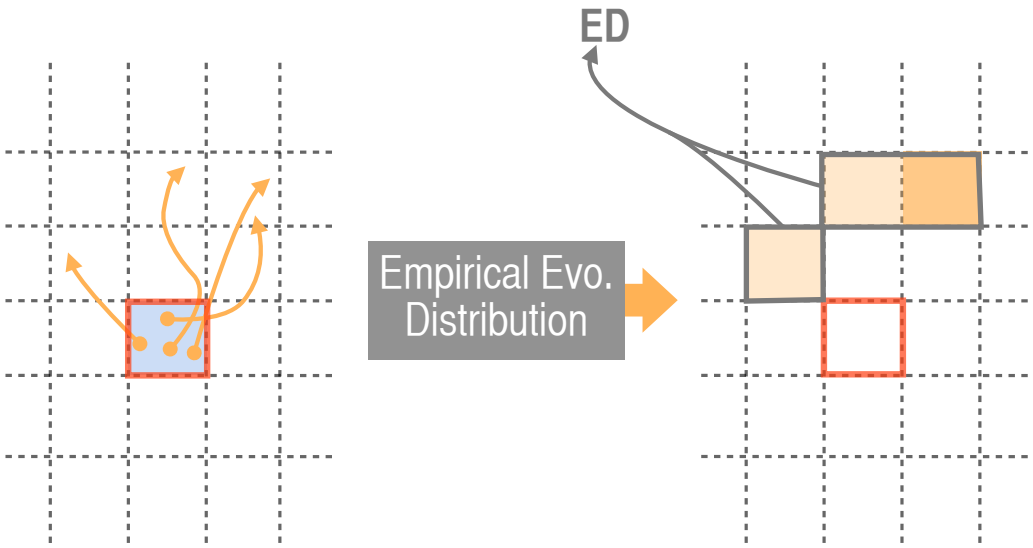


Figure 42: Illustration of how the EDs are estimated .

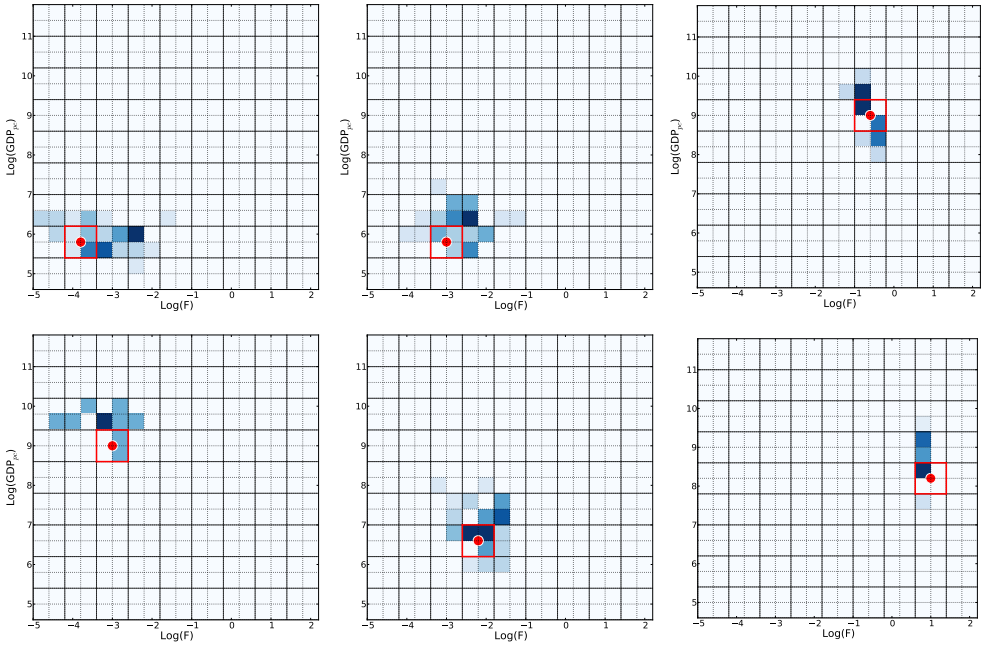


Figure 43: Selection of EDs from both the chaotic and laminar regime. The time lag is equal to 5 years.

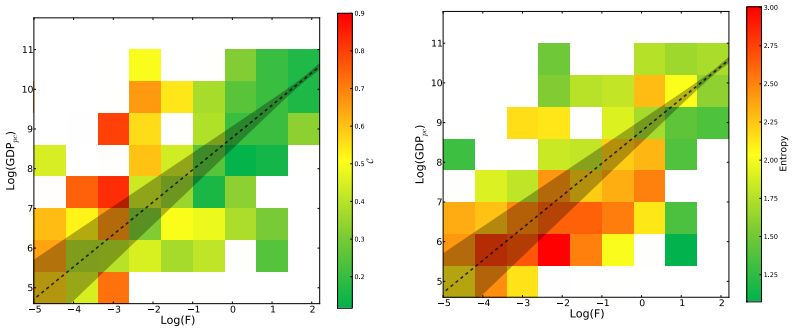


Figure 44: Measures of concentration for EDs. The time lag is equal to 5 years. (Left panel) The 5-years case of the *selective predictability scheme* appears to be similar to the 10-years case: we can divide the dynamics in two regions, laminar characterized by EDs poorly dispersed and chaotic where EDs are very broad. (Right panel) In that case we have enough statistics to show significant estimation of the entropy which confirms the existence of two regions with very different features of the EDs. The entropy analysis also confirms the heterogeneity inside the laminar regime where, even if poorly dispersed, two kinds of EDs are observed: very peaked on only one box (high value of normalized Herfindahl index) and more uniformly distributed (low value of normalized Herfindahl index).We report the concentration measure for boxes with at least 5 events.

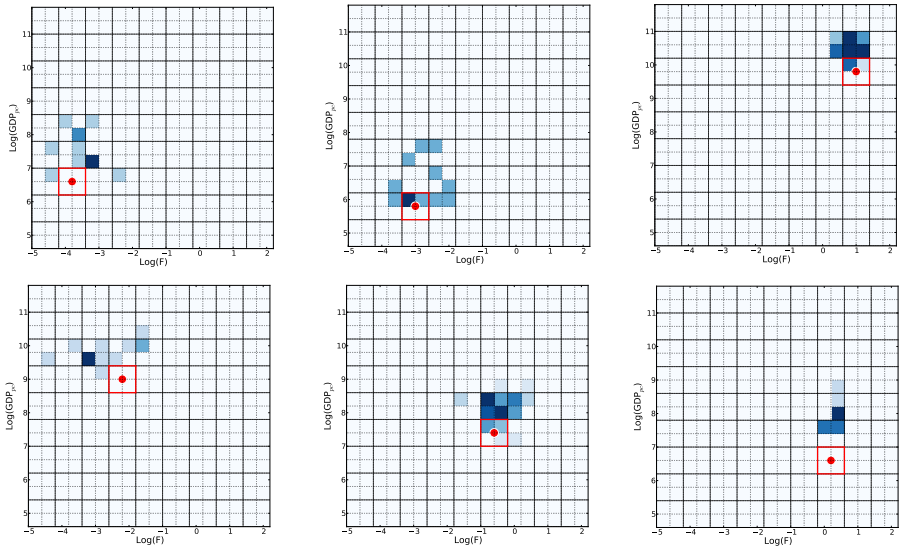


Figure 45: Selection of EDs from both the chaotic and laminar regime. The time lag is equal to 10 years.

Selective predictability scheme: 10-years prediction

We report in this section the *selective predictability scheme* in the case in which the EDs are built tracking the 10-years evolution in the fitness-income plane. In Fig. 45 we show a selection of EDs. As previously discussed, we do not have a reliable estimation of the entropy of the EDs due to the small statistics we are considering, however, for the sake of completeness we report the entropy for each box in Fig. 46 .

APPENDIX 2.D BACKTESTING ED-BASED FORECASTING SCHEME

The stability in time of the coarse grained dynamics and the robust patterns observed measuring the concentration of EDs ground the existence of two kinds of regime for the dynamics of the economic complexity and, consequently, the *selective predictability scheme*. As a final analysis to support our forecasting scheme for economic growth, we perform a backtesting of our method – backtesting represents a standard way to test a forecasting scheme, see for instance [27] for financial application.

Given the limited time window under investigation, we can only perform a backtest of the 5-years *selective predictability scheme*. We estimate the 5-years EDs using the evolution of countries from 1995-2000 and then test the rate of success of the prediction of the position of countries in 2010 given their position in 2005 according to the EDs obtained in training time period.

In Fig. 47 we report the rate of success measured as the ratio of predicted events and the total cases. In the left panel, we report the case in which we consider boxes with at least 2 events, while in the left with at least 3. Despite the small statistics of the test, it appears that the ED from the laminar regime has a significant and systematically higher rate of success. We

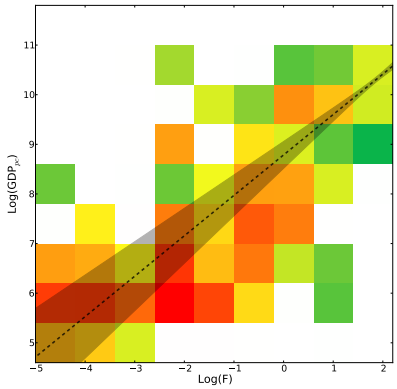


Figure 46: Entropy of the EDs for 10-years *selective predictability scheme*. Differently from the 5-years case, entropy estimation may be strongly biased by the small statistics we are considering.

stress once again that, even if the ratio of success of two EDs, one from the laminar regime and one from the chaotic regime, were the same, the forecast of the evolution of the country in the first case would correspond to indicate a much smaller area in the fitness-income plane in which we expect to observe the country.

2010's fitness and GDP *per capita* forecast

Given the ED estimated from the training set from 1995 to 2005 and given 2005's fitness and GDP *per capita*, the forecast of 2010's evolution is illustrated in Fig. 48. We calculate the center of mass of starting points for each ED (B_1 in Fig. 48) and the center of mass of the evolved points (B_2 in Fig. 48). For each ED, we then compute the vector associated to the displacement of the center of mass as shown in panel a) of Fig. 48. For each country, we apply to the 2005's position in the fitness-income plane the displacement vector previously calculated depending on the box in which the country is (Fig. 48 (panel b)) . The relative error reported in Fig. 35 is simply the difference between this forecast and the realized 2010's GDP *per capita* normalized with the realized GDP *per capita*. As discussed in the main text of the chapter, we believe that the systematic under estimation is due to a training set shorter than the length of an economic cycle.

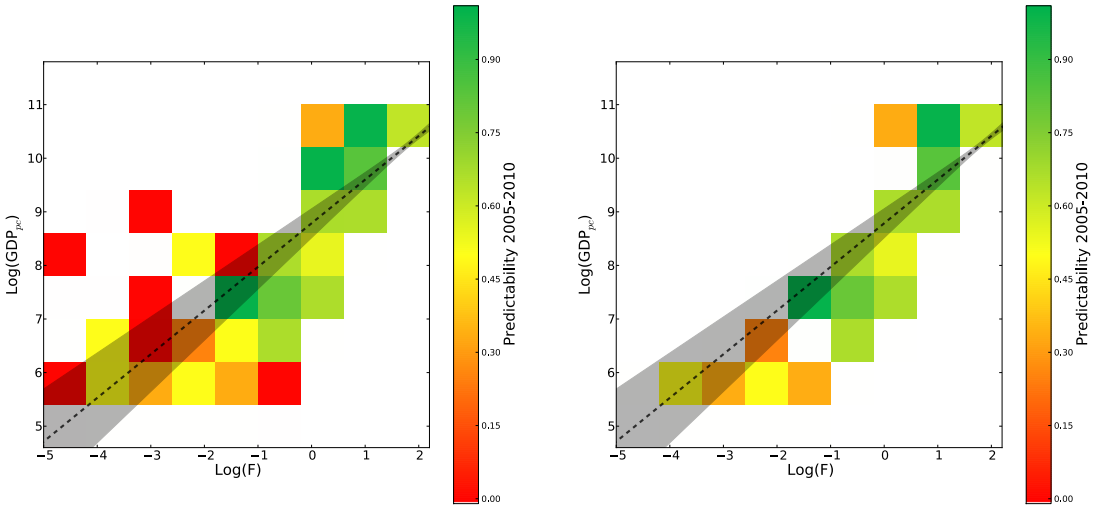


Figure 47: Backtest of the 5-years *selective predictability scheme*. (Left panel) We consider boxes with at least 2 events, (right panel) events with at least 3 events. In both cases the conclusion that the laminar regime exhibits a much higher degree of predictability is confirmed.

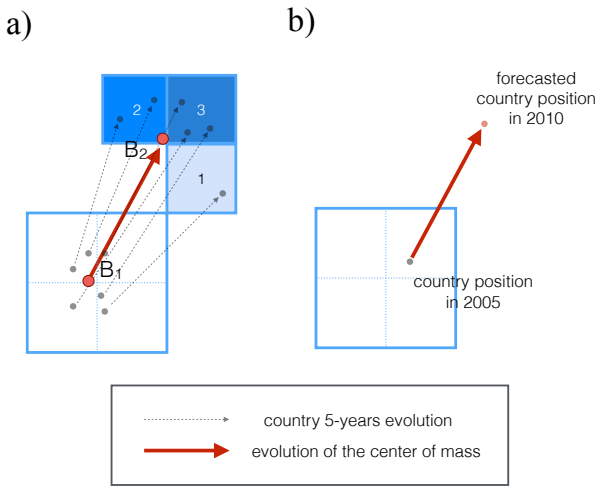


Figure 48: Illustration of the forecast method for 2010.

3 THE BUILD UP OF DIVERSITY IN COMPLEX ECOSYSTEMS: PATHS TOWARDS NESTEDNESS

ABSTRACT

Diversity is a fundamental feature of ecosystems, even when the concept of ecosystem is extended to sociology or economics. Diversity can be intended as the count of different items, animals, or, more generally, interactions.

There are two classes of stylized facts that emerge when diversity is taken into account. Diversity explosions are the first stylized fact: evolutionary radiations in biology, or the process of escaping "Poverty Traps" in economics are two well known examples. The second stylized fact is nestedness: entities with a very diverse set of interactions are the only ones that interact with more specialized ones. In a single sentence: specialists interact with generalists. Nestedness is observed in a variety of bipartite networks of interactions: Biogeographic (Islands-Animals), macroeconomic (countries-products) and mutualistic (e.g. Pollinators-Plants) to name a few. This indicates that entities diversify following a pattern.

For the fact that they appear in such very different systems, these two stylized facts seem to point out that the build up of diversity might be driven by a fundamental mechanism of probabilistic nature, and in this chapter we try to sketch its minimal features. Namely we show how the contraction of a random tripartite network, which is maximally entropic in all its degree distributions but one, can reproduce stylized facts of real data with great accuracy which is qualitatively lost when that degree distribution is changed.

We base our reasoning on the combinatoric picture that the nodes on one layer of these bipartite networks (e.g. animals, or products) can be described as combinations of a number of fundamental building blocks. We propose the idea that the stylized facts of diversity that we observe in real systems can be explained with an extreme heterogeneity (a scale-free distribution) in the number of meaningful combinations (*usefulness*) in which each building block is involved. We show that if the *usefulness* of the building blocks has a scale-free distribution, then maximally entropic baskets of building blocks will give rise to very rich behaviors in accordance with what is observed in real systems.

3.1 INTRODUCTION

The study of complexity in ecosystemic interactions has a long history. The seminal paper of May[28] disputed the intuitive view that a complex network of interactions would tend to stability when its size is increased[29]. May's formal results regarded networks with random interactions, but in real cases interactions are far from being random. In particular in this work we focus on the case of bipartite networks of interactions. In this case we can separate nodes

in two layers such that nodes from one layer only have direct interactions with nodes from the other. This representation is useful when one layer can be interpreted as a set of possible resources for the nodes on the other one (with this relation being reciprocal in the case of mutualistic networks).

By enlarging the scope of the analysis, we can notice that bipartite networks of interactions are common in many fields such as biology (plants-pollinators[30], islands-species[31–33]), economics (countries-products[3, 5, 26], advertisement (customers-purchased items[34, 35])), sociology (sexual partners[36]). The idea of considering economic or social interactions as entities affine to proper biological ecosystems isn't new. At the root of this association stands the fact that in both these two contexts there are entities competing and interacting for resource allocation. In such contexts being able to rely on a diversified set of resources is of course a great advantage, because it improves resiliency. At the same time exclusivity, namely having access to resources which are challenged by a small number of rivals, can boost this advantage. In a dynamic ecosystem fitter entities explore a phase-space of features (phenotypes in biology, or capabilities in economics) that allow them to make use of a possibly increasing range of exclusive resources. The result of such a dynamics can be a situation in which specialists (exclusive) resources are only accessible to fitter, generalists agents: a concept known as nestedness.

The idea that the same dynamics is taking place in such different contexts is strengthened by the observation that some stylized facts are present in observative data related to both economics and biologic ecosystems.

The first stylized fact is related to a dynamic phenomena observed in complex ecosystems, namely sudden diversity explosions that are in sharp contrast with the previous rate of innovation of the system. In ecology this phenomena is known as evolutionary radiation. Examples of radiations are the well known Cambrian Explosion[37], or the evolution of insect-eating placental mammals into a wide variety of herbivores, flying mammals and marine mammals just after the Cretaceous[38]. The mechanisms that drive such very fast increases in biodiversity are still object of discussion. While many ecologists focus on environmental causes[39, 40], some others indicate that possible causes should be related to the appearance, by chance, of some novel functional traits (like eyes[41]) that trigger a chain of evolutions by opening new possibilities. More in general some point out that a *complexity threshold* might have been crossed[42].

In economics something closely related has happened in relatively recent times.As a matter of fact, for most of human history the ratio between population and wealth remained constant (see fig. 49). This phenomena is known as Malthusian Trap[43]. Malthus' theory was that population was merely limited by available resources and that at whatever point in time a technological progress was made, allowing access to greater resources, population would rapidly grow accordingly, thus keeping the wealth per capita in a trapped state. In the last years of Malthus' life the industrial revolution was beginning in England. The process lead to what is today known as The Great Divergence[44]: the industrialized countries were able to escape the Malthusian Trap and the wealth per capita has since then grown tremendously in these countries. In other words the industrial revolution coincided with a real shift of regime.

If we consider the industrial revolution from the point of view of technological diversity it is not very different from an evolutionary radiation. In an incredibly short amount of time a variety of new technologies, resources, scientific advances and consumable products has stemmed, as the result of the introduction of a single new idea, the motor, in an ecosystem of technologies that weren't combined with that efficiency before. It was likely not the first time in history that such kind of revolutions happened (one can think of the invention of the

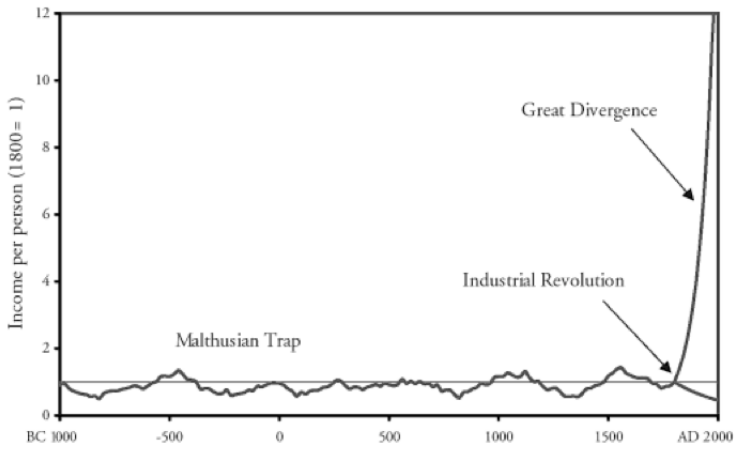


Figure 49: The income per capita of the world population for the past 3000 years. The value remained more or less stable until the industrial revolution. Economy was so simple that whenever an improvement in productivity was made, the population simply grew accordingly, keeping the ratio between income and population constant. The industrial revolution marked a shift of regime. Interestingly it coincided with a divergence in which some countries remained stuck in the trap, while others escaped.

wheel, or the development of agriculture) and the interesting fact is that it was still happening in recent times. In fact, in the recent past, just after the second World War, many countries were still living in a Malthusian Trap (or Poverty Trap). Some of these countries (e.g. China, India or South Korea) were able to escape the trap and move on to an industrialized rather than subsistence economy. In the process the diversity and complexity of their production exploded and this anticipated[45] the later observed GDP growth.

It must be stressed that these topics are treated in a very qualitative way in literature for what concern the dynamics of diversity, both in economics and ecology. This is of course related to the fact that such revolutions happen rarely. Thus their characterization is hard to formalize and digging into the details of such dynamics is more a philosophical exercise rather than scientific. But yet the observation of these dynamics poses interesting scientific questions when one is interested in the general features. A good example is shown in fig. 50. The dynamics is characterized by some interesting qualitative features. When we look at the general trend the growth is logistic, with an exponential increase which seems to be saturating. But at a finer level the dynamics shows a peculiar "bumpy" behavior, with bursts of diversity followed by periods of null growth or even extinctions. Given the rather qualitative level of such analysis, the introduction of a second class of stylized facts could help us settle the problem and identify some fundamental features of the build up of diversity.

The second stylized fact is known as nestedness in ecology and has been studied for a long time. In bipartite networks of interactions, nestedness is a peculiar degree correlation between nodes in the two layers: entities with a very diverse set of interactions are the only ones that interact with more specialized ones. In other words, specialists interact with generalists. It has been suggested that this organized structure of interaction may be beneficial for the stability of the ecosystem[30, 46]. Interestingly nestedness is also a very clear feature of international trade bipartite networks: in this case non-ubiquitous products are produced only by diversified

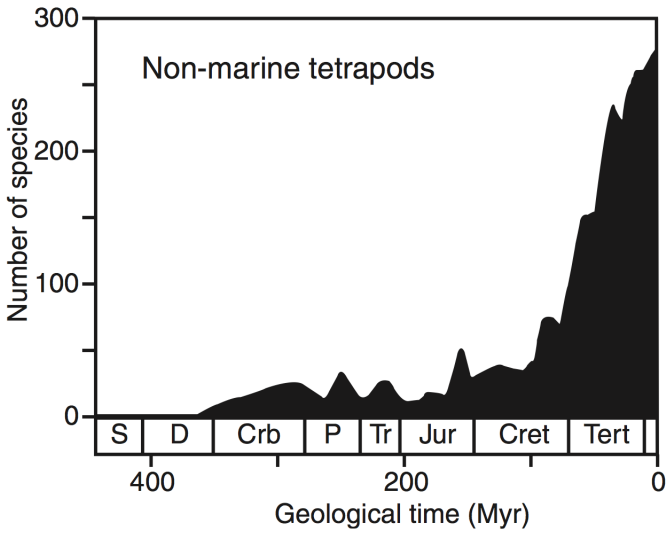


Figure 50: The evolution of biodiversity of non-marine tetrapods. The dynamics is characterized by some interesting qualitative features. When we look at the general trend the growth is logistic, with an exponential increase which seems to be saturating. But at a finer level the dynamics shows a peculiar "bumpy" behavior, with bursts of diversity followed by periods of null growth or even extinctions.

countries, and non diversified countries only produce ubiquitous products. This feature of the countries-products network motivated the introduction and definition of the metrics described in [26]

Remarkably the same metrics can be applied to biological networks and yield a measure of the importance of a given node in relation to cascades of extinctions much more accurate than any other standard measure of centrality[47]. Interestingly the rankings given by the metrics seem to solve the long standing problem of the optimal nested ordering of a matrix in an almost optimal way, much more efficiently than the standard "nestedness temperature" approaches[48], which were already known to be problematic[49]. Nestedness emerges when we consider the diversity associated with multiple entities that evolved in the same ecosystem. The fact that diversity is organized suggests that the process of emergence of diversity follows a pattern. This consideration combined with the observation of similar stylized facts in systems of very different nature seem to point out that the build up of diversity might be driven by a fundamental mechanism of probabilistic nature.

3.2 A MODEL FOR THE DYNAMICS OF DIVERSITY

We picture diversity as the number of meaningful combinations of small pieces, or building blocks, that combine together to create meaningful associations. An ecosystem is thus a basket of such small pieces and the environment (and the competition) define which combinations are meaningful. The active agents of the ecosystem collect some of these building blocks from the basket and their fitness is larger the larger the number of meaningful combinations they

can make, being larger the set of possible interactions they can have. Thus, in the economic framework, a country will be able to produce all the products for which it owns all the needed building blocks, or capabilities. The concept of capability was introduced by Lall[10] and is at the basis of the idea of Economic Complexity. In this view capabilities are all the technical, political, geographical, infrastructural and social requirements that allow the production of a given product in a country. From a biological point of view we can consider the islands of an archipelago. An island will contain all the life forms that can emerge out of the genetic traits that are present in its ecosystem. In an archipelago, the fact that these genetic traits all come from a common pool (the basket) generates nestedness. In the case of mutualistic networks animals will be able to gather resources from all the plants that the combinations of their phenotypic traits will allow, and interestingly, the same can be said for plants in the opposite direction.

In this view, the observed diversity is the result of two processes. The first is the random appearance of novel traits (or technologies, or ideas), with diversity increasing as the number of new combinations made possible. The second is natural selection: non-meaningful o non-fit combinations are removed from the system, thus decreasing diversity. In this work we focus on the first process and we schematize the second in a statical way. In particular we assume that the natural selection is always at equilibrium in our model, and we simply impose a set of acceptable combinations and discard all the others, with the requirement that the number of acceptable combinations is much smaller than the number of possible combinations.

3.2.1 Explosions of diversity and the concept of *Usefulness*

First we try to characterize a minimal model that is able to reproduce qualitatively dynamics close those shown in figg. 49 and 50. The framework of the model is easily understood by looking at fig. 51. Building blocks and combinations form a bipartite network. The collector is endowed with an increasing number of building blocks, one at a time. This correspond to the disputable hypothesis of a constant rate of exploration, which is in any case the simplest assumption in this framework. The diversity of the collector at any given time is the number of combinations for which it has all the building blocks.

What is left to define are the properties of the topology of the bipartite network of building blocks and combinations. Again, to keep the model minimal, we define this network to be random and we only focus on its degree distributions. In these degree distributions stands the key feature that embodies our view of the fundamental mechanism driving the build-up of diversity: the concept of *Usefulness*. We define *Usefulness* as the number of meaningful combinations in which a building block is involved. Previous implementations of similar models such as those used in Chap. 1 and in [3] did not make use of this concept and implicitly imposed a binomial distribution for the *Usefulness* of the building blocks. We argue that in nature this isn't the case. We can think of a number of practical situations in which this distribution would be substantially different than a unimodal exponentially decaying one: in technology ideas such as the transistor have clearly a much larger number of applications than, say, a particular metal working technique; in biology, as already noted, phenotypic features such as eyes find place on a much larger set of fit life forms than, say, the pouch of marsupials. One could also notice that the *Usefulness* of wings would be somewhere in the middle among eyes and the pouch, but yet with orders of magnitude of distance from each of the two. A mathematical

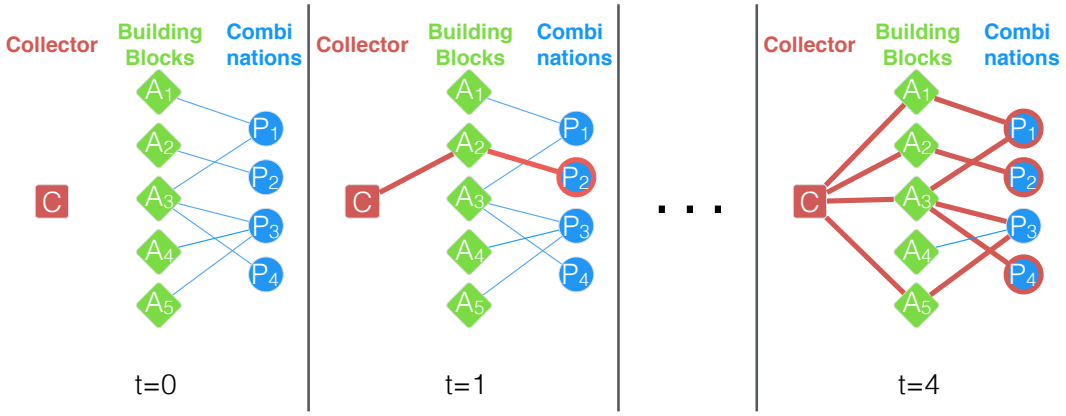


Figure 51: A schematics representation of the framework of the model: a bipartite network connecting Building Blocks and combinations is defined and a collector gathers Building Blocks through time. The dynamics is implemented as follows: at each time step we add a link between the Collector and one randomly chosen Building Block. The diversity at that time step is the number of Combinations that the Collector is able to make given the Building Blocks it has collected until that time step.

formalization of these ideas corresponds to a "fat-tailed" distribution of *Usefulness*, and as we show this assumption allows for a clear qualitative shift in the output of our models.

In detail we build the bipartite Building Blocks-Combinations network as follows:

- First we draw from a power-law distribution $P(n) \propto n^{-\alpha}$ a number n_i for each building block, that is its *Usefulness*.
- Then for each Building Block i we choose randomly n_i Combinations to which the Building Block will be connected. In this way the expected value of the length of each Combination is the same.

As mentioned the dynamics is implemented in the simplest possible way: at each time step we add a link between the Collector and one randomly chosen Building Block. The diversity at that time step is the number of Combinations that the Collector is able to make given the Building Blocks it has collected until that time step. About the numerosity of the nodes of the bipartite network we only request that the number of possible combinations of Building Blocks is much larger than that of the actually allowed ones. In practice all the results shown here are obtained with a fixed number of Building Blocks N_a and Combinations N_p with $N_a = N_p = 1131$. Changing these numbers and the ratio between N_p and N_a only causes quantitative changes in the behaviors, as long as the two numbers are large enough and as long as the number of possible combinations of Building Blocks remains much larger than that of the allowed Combinations.

In fig. 52 we show the resulting dynamics for three possible distributions of *Usefulness*. The first two behaviors show an essentially exponential increase in diversity. While the loss of the unimodality for the exponentially decaying distribution causes the trajectory to be a bit rougher, it is clear the qualitative shift that is obtained with the slowly decaying distribution. The global "logistic-like" shape of the curve and the bursts of activity shown in the biological example of fig. 50 are qualitatively well reproduced. What is missing is of course the extinction

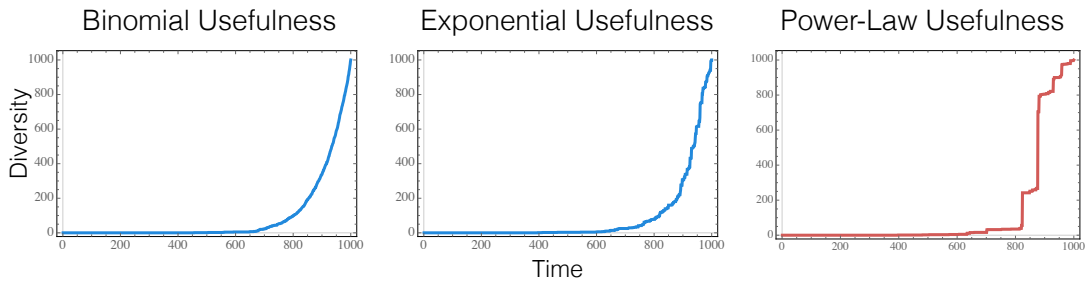


Figure 52: A schematics representation of the framework of the model: a bipartite network connecting Building Blocks and combinations is defined and a collector gathers Building Blocks through time. The dynamics is implemented as follows: at each time step we add a link between the Collector and one randomly chosen Building Block. The diversity at that time step is the number of Combinations that the Collector is able to make given the Building Blocks it has collected until that time step.

part, which is out of the scope of this model, since, as we stated, we consider the natural selection to be always at equilibrium, thus unfit Combinations are automatically suppressed.

We can interpret the dynamics with the presence of two regimes: at the beginning the system lives in the "poverty trap", when for a long time no relevant increase in diversity is observed. Then, once the system has accumulated a large enough number of Building Blocks, or complexity, the dynamics shift to a different regime of fast growth. The sudden increases in diversity corresponds to the discovery of a very *Useful* idea, that allows to exploit a large part of the Building Blocks already owned but that were missing a piece that could tie them together.

3.3 MANY COLLECTORS AT THE SAME TIME: EMERGENCE OF COMPLEX NESTEDNESS

The bipartite networks in which we observe nestedness can be put in correspondence with a generalization of the dynamical model proposed in the previous section, but observed at a fixed time step. In particular it correspond to consider many independent collectors that are endowed with random sets of Building Blocks, drawn from the same basket. Thus all the collectors are in principle the same, and no heterogeneity, other than different random choices of the capabilities is introduced. As we will see this will nevertheless result in a significant heterogeneity of the resulting diversity among the Collectors, when we consider the contracted bipartite networks of Collectors with the Combinations that they can make

For the fact that we developed quantitative methods to describe the properties of the adjacency matrix of the bipartite Collectors-Combinations projection of such network, we can describe in more detail some stylized fact present in real data. From a qualitative point of view we can begin with the trivial observation of the shape of the matrix: once rows and columns are ordered by Fitness and Complexity the shape is triangular-like (see fig. 4). While this is itself a symptom of nestedness it is worth to notice that there is something more about the structure of these matrices that we can observe. We can, for example, build a random binary matrix in which a given density of 1's is concentrated in one of its two triangles while

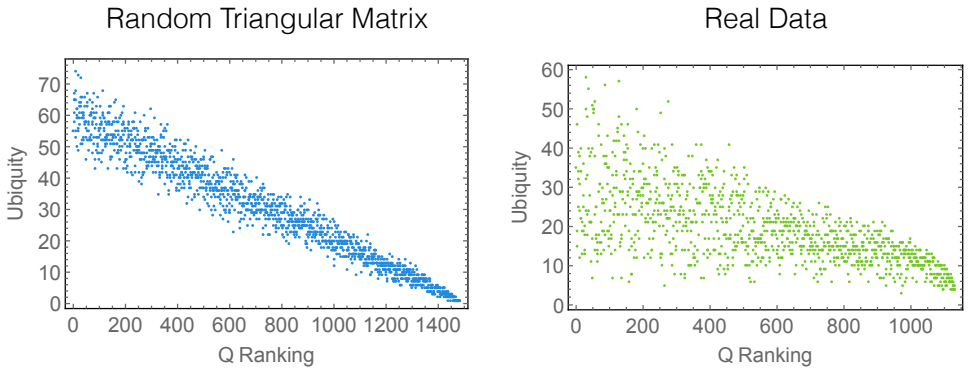


Figure 53: Ubiquity versus Complexity ranking for a random triangular matrix and the real M_{cp} for year 2010. In the random case the knowledge of ubiquity is almost the same as the ranking of complexity. In the real case the situation is much different. Ubiquity is not a good proxy for complexity since even a non ubiquitous product can be of low complexity if a non-diversified country is able to export it. This complexity is not present in the random case.

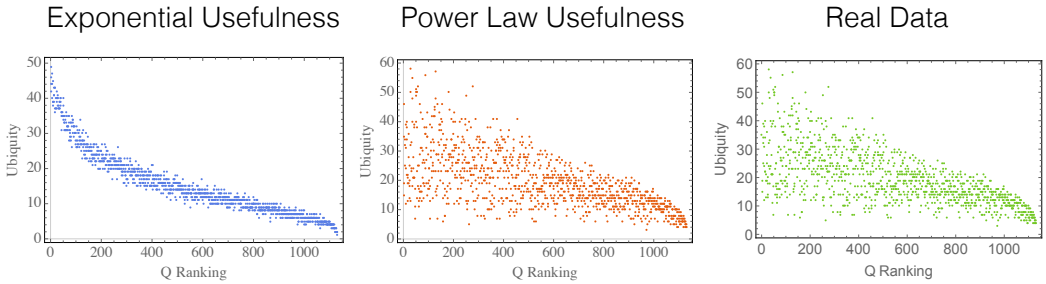


Figure 54: Ubiquity versus Complexity ranking for the cases of different distributions of *Usefulness* and the real data. As it is clear, the power-law distribution gives results in good accordance with the real case.

the other is left empty. The shape is now qualitatively similar to that of the real matrix. Nevertheless, we can use Fitness and Complexity to spot that this randomness does not reproduce the full complexity of the real data. In fig. 53 is shown a comparison of the Ubiquity versus Complexity Ranking plot for a random triangular matrix and the real M_{cp} for year 2010. In the random case the knowledge of ubiquity is almost the same as the ranking of complexity. In the real case the situation is much different. Ubiquity is not a good proxy for complexity since even a non ubiquitous product can be of low complexity if a non-diversified country is able to export it. This complexity is not present in the random case. Thus we start to understand that the matrix is not only nested, but some non trivial structure, or complexity, is present in its fine details.

We can then try to use Fitness and Complexity to assess the accuracy with which our models are able to reproduce not just the shape, but also the finer features of the nested matrices that we observe. In fig. 54 we show a comparison of the same plot for two different distributions of *usefulness* and the real matrix.

As it is clear from the figure, the introduction of a "fat-tailed" distribution for the *Usefulness* introduces a qualitative change in the fine structure of the M_{cp} matrix.

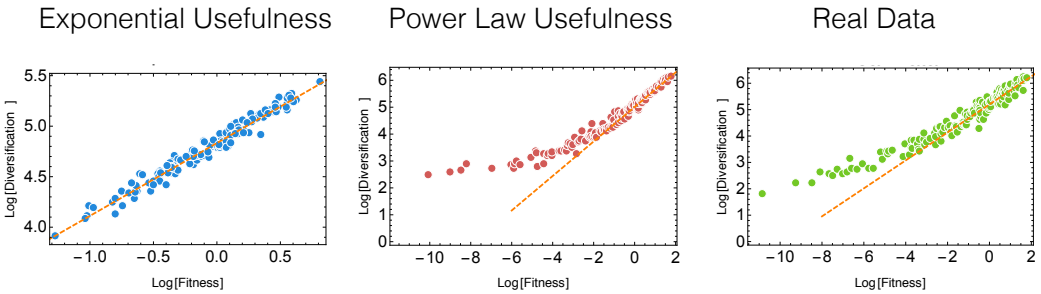


Figure 55: Fitness versus diversification for two different distributions of *Usefulness* and the real data. From the real data two regimes emerge in a clear way: collectors with low fitness live in a "Poverty Trap" where a given increase of complexity leads to a small increase in diversity; collectors with higher fitness (along the dashed trend line) have a much larger benefit from the same increase in complexity. Their efforts are thus much more rewarded. Interestingly the same two regimes are present in the power-law case but not in the exponential case. This is also reflected by the features of the dynamics shown in fig. 52.

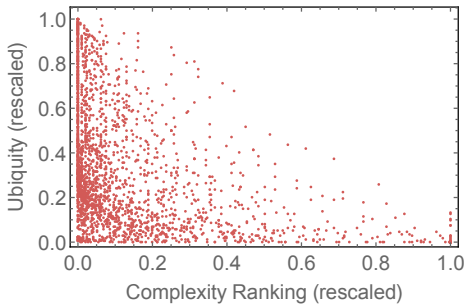


Figure 56: Ubiquity vs. Complexity ranking relation for 59 Plant-Pollinators networks. The points are rescaled to collapse in a 1*1 box, being the size of the networks very heterogeneous.

Moreover we can use the Fitness to see what is the relation between the complexity of collectors and their diversification when different distributions of *Usefulness* are considered. In fig. 55 we show the Fitness versus diversification plot for two different distributions of *Usefulness* and the real data. From the real data two regimes emerge in a clear way: collectors with low fitness live in a "Poverty Trap" where a given increase of complexity leads to a small increase in diversity; collectors with higher fitness (along the dashed trend line) have a much larger benefit from the same increase in complexity. Their efforts are thus much more rewarded. Interestingly the same two regimes are present in the power-law case but not in the exponential case. This is also reflected by the features of the dynamics shown in fig. 52.

We can also see how nested matrices from biological datasets display the same properties. As an example we plot the Ubiquity vs. Complexity ranking relation for 59 Plant-Pollinators networks[1], and the results are shown in fig. 56. Again the ubiquity is substantially different from complexity, in a way that only a fat-tailed distribution of *Usefulness* is able to explain in this framework.

1 Source: Web of Life database (www.web-of-life.es)

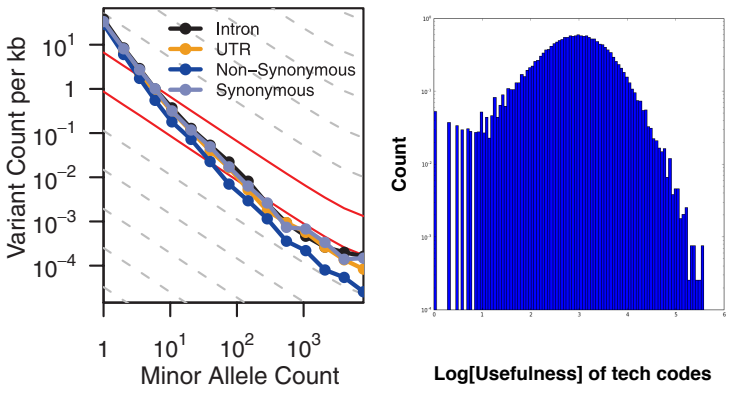


Figure 57: a) Frequency of appearance of genetic variants in a population of 15000 humans(From [50]). b) Frequency of appearance of technological codes in a dataset of patents (From [51])

Even when comparing more quantitative features for real nested matrices the power-law distribution of *Usefulness* proves to be able to give results in striking accordance with the real observations. Since all the other features of the tripartite Collectors-BuildingBlocks-Combinations network are random, we think that from this simple model we can learn something interesting about the mechanisms governing the build-up of diversity in complex ecosystems.

3.4 FAT TAILED DISTRIBUTIONS OF USEFULNESS IN REAL DATA

It would be interesting to observe something similar to a distribution of *Usefulness* in a real system. It is not easy to check experimentally if such kind of distribution do exist in nature, mostly because giving a precise definition of the Building Blocks is a hard task. In the technological case the qualitative observation that some technological ideas have found a much wider application than many other is readily made. A quantitative approach can be made with respect to the frequency with which we observe technological codes in patents: this seems to follow a roughly log-normal distribution. This "fat tailed" distribution corresponds to a non negligible number of technology codes appearing on a very large amount of patents. From a biological perspective some hints that these distribution are present in nature can come from genetics. Recent studies have demonstrated that the frequency with which genetic variants appear in a population follow a scale-free distribution[50]. These findings are summarized in fig. 57.

3.5 CONCLUSIONS

By imposing very general conditions on a very simple model we try to highlight how the concept of *Usefulness* might be crucial to understand the mechanism that drive the build-up of diversity in complex ecosystems. These fundamental mechanisms seem to be very general, as the stylized facts that they produce are observed in a wide class of systems, driven in principle

by very different laws. The ideas that we propose in this work are anyway general enough to be transposed to different contexts with an obvious correspondence of the variables.

4 FORECASTING OF TECHNOLOGICAL DEVELOPMENT

CHAPTER ABSTRACT

The data about countries' export can be exploited to reconstruct technological relations among products. In our framework these relations can be thought as something proportional to the size of the intersection of the sets of capabilities needed to produce the products. In this chapter we approach the problem of determining the strength and direction of these relations in three different ways. First we use a static approach in which the relations among products are inferred by using the statistics of cooccurencies of couples of products in the same export basket. We select the relevant links from the resulting matrix of similarities in such a way that we can identify a taxonomy: a directed tree in which janitor nodes represent necessary steps to get to produce more refined ("leaves") products. We compare this approach with a dynamic one, in which the "necessity" of a product for the production of another one is estimated by looking at how often a given product was exported prior to the addition of the other. We observe how the two approaches provide very similar results. Finally we move from one-to-one relations to a more general framework. We use a machine-learning approach to test whether the presence of a particular set of products in an export basket is likely to trigger the appearance of another one after a given time lag. We use part of our dataset as a training set to build decision trees, and test their performance in pointing out situations of possible development with the remaining part of the dataset. The predictive power of the different approaches is compared and results up to three times better than a random choice in predicting the growth of exported volumes, even if the test is performed in the 2007-2010 portion of the dataset, plagued by the global economic crisis.

INTRODUCTION

The study of how countries develop has a central role in Economics and major consequences in political, industrial and financial analyses and evaluations. Historically, a number of approaches have been applied to this problem. According to the model introduced by Heckscher and Ohlin [52], which is based on Ricardo's comparative advantage [53], the possible pattern of progress of a country is a direct consequence of its endowments, namely the presence of productive factors such as land, labor and capital. This approach has been challenged by Leontief [54, 55], who found a striking empirical counterexample, now known as Leontief's paradox (but see also [56] for a contrary view) and, again on an empirical basis, by Bowen et al. [57]. Another approach has been proposed by Aghion and Howitt [58], whose model is inspired on the concept of creative destruction, originally introduced by Schumpeter [59], which focuses on endogenous factors such as technology. This perspective has originated from the seminal

paper by Romer [60]. As pointed out, for example, by Hausmann and Rodrick [61], these views can be summarized in the assumption that the basic needs for a sustained growth are tradable technology and good local institutions. Hausmann and Rodrick in the same paper give two examples, the growth of some Asian countries and the recession of the Latin American ones, in which the opposite was true. For example, South Korea and Taiwan experienced an impressive growth even if they retained high levels of protection, while Latin American countries performed better in the decades 1950-1980, with poorer institutions, than in the 90's, when their governments adopted the long awaited structural reforms. Lall [10] suggested a third line of reasoning, which he calls the "capabilities approach". A capability of a country can be, in its wider sense, anything which makes the country able to produce a given product, from infrastructures to efficient scholar and administrative institutions, from a mild fiscal policy to demography issues. According to Lall, the crucial point is not the simple knowledge of a technology, but the ability to exploit its potential, that is to be able to use it efficiently given the intrinsic properties of the specific country. As a consequence, each country has to find its own path towards development, focusing on its learning system in order to add capabilities to the ones it already owns. This line of reasoning, in which each country has to learn first what one is good at producing and then which technology can be best adapted to its case, has been modeled in a general equilibrium framework in [61]. By contrast, our approach is closer to the concept of *adjacent possible*, introduced by Kauffman [62] and originally applied to biological systems [63]. Finally, we mention the evolutionary approach [64], in which the optimizing role of the market is substituted by a natural selection process which assures a ceaseless change, in general, of any economic process. This ideas gave rise to new fields of research, such as the ones regarding innovation [65].

The time series of exports give a fundamental insight to understand countries' development, and can help define an empirical framework to assess the validity of theoretical paradigms. If we suppose that products are defined by means of the set of capabilities which are needed for a country to be able to produce it, the presence or the absence of a product in a country's export basket represent a hint on the capability basket of the country itself. By extending this reasoning it is easy to see how one can build a network of products in which two products are connected if they share some capabilities; in practice, if many countries produce the same couple of products. In this way, one can avoid to study the capabilities structure of countries, which is, at best, very hard to represent or even define from a quantitative point of view. This network, called the Product Space, has been introduced and studied by Hidalgo et al. [11] (see also [66], and [7] for a different approach). In this chapter we propose different approaches to the problem of technological development of countries, all based on the capabilities framework. The first two strategies are based on the idea of building a network of products, in which two nodes are connected by a directed link which represents the causality relationship between them. For example, two products a and b will be connected not if they are just similar, but if one of them, say a, makes more probable that b will be produced in the future. In this case, the directed link will go from a to b. We propose two algorithms that make use of the information contained in the empirical export data and permit to build a hierarchical network whose nodes are products and the directed links are given by the necessity relationship between products. In this structure, in which the link between capabilities and development emerges in a clear way, the number of edges is reduced with respect to the almost fully connected network which can be obtained by a simple projection of the bipartite country-product network; namely, we reduce the number of links from about N^2 to order N by selecting the most informative ones from the point of view of economic progress. The two proposed approaches differ for the

fact that in the first case we use a static picture of the export baskets of countries, namely the strength of a link is determined by the probability that two products are exported at the same time. The second instead introduces a time lag in the relation. Interestingly the resulting networks are quite similar and, even if in the first case we don't add any dynamic content to the network the algorithm is able to determine not only relevant links among the products but also a meaningful direction for those links, that relates very tightly with the results given by the dynamic approach.

Finally we introduce a third approach in which not only the dynamics is explicitly considered, but that also takes advantage of the available information in a more complete way, by exploring causality links not only among couples of products but looking at the probability that particular *sets* of products trigger the production of a new one. This is done via a machine-learning approach that defines so-called decision trees in a training set consisting in a relevant part of our full 1995-2010 dataset. The performance of the approach is then assessed "out-of-the-sample" in the remaining part of data and compared with the dynamic "one-to-one" approach.

Each of the tree approaches, despite adding important methodological aspects one to the other, provides a marginal improvement in terms of predictive power to the previous one. We interpret this finding as a strong signal in favor of the theory of hidden capabilities. In fact the first, static approach, is designed explicitly to uncover similarities in terms of capabilities, and it performs very well in toy models based on capabilities (as will be shown in this chapter). The fact that adding direct observations of past dynamics to the story gives results strongly correlated with the network obtained via the static analysis means that most of what is needed to explain development in the medium term is already contained in the theory of capabilities.

Aside from the methodological aspects, a different way of looking at the findings presented in this chapter is to observe that while the development of many countries is mostly driven by their initial conditions, most of them walk through recurrent paths, suggesting the presence of mandatory steps in the industrial progress of nations. This has of course direct implications in industrial programming and policy making, in particular in the case of developing countries. While this might to some extent be already known, the approaches that we propose here make these concepts quantitative in way that wasn't attempted before to the best of our knowledge.

4.1 STATIC APPROACH: A TAXONOMY FOR PRODUCTS.

4.1.1 Taxonomy and Proximity

As we anticipated above, differently from the approach described by Hidalgo et al.[11], we want to build a hierarchically ordered network, whose structure is inferred from the M_{cp} matrix. The idea can be easily understood by means of the concept of capability [3, 10]. Let us define the products in terms of the capabilities which are needed to conceive and produce them. For example, the capability 1 corresponds to a basic product. A country equipped with a second capability, 2, can export the "12" product. Capabilities 1,2 and 3 could simply not lead to a product, while "134" can be a product, and so on. A hierarchy naturally arises, in which some products are mandatory intermediate steps to be able to produce more complex technologies, and the *sons* are connected to the *father* by a directed edge. In Fig.58(a) we show a

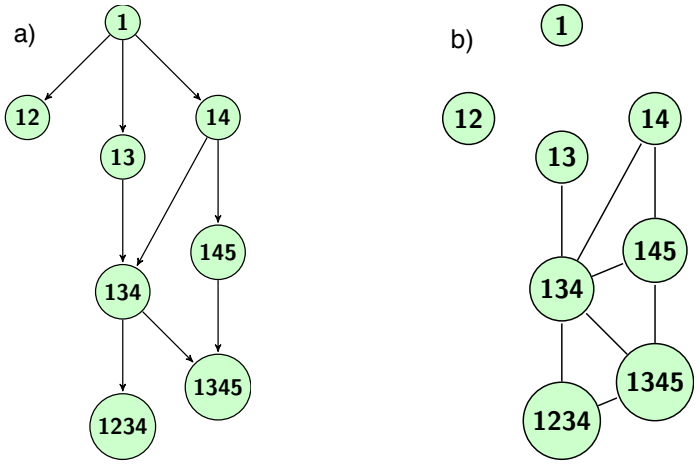


Figure 58: Two different ways to connect the same products, which are characterized by the capabilities needed to produce them. On the left, we consider a hierarchical relationship; on the right, we join similar products.

possible example of this kind of structure, that we call a *taxonomy* network. On the other hand, Fig.58(b) shows an example of a *proximity* network, in which the same products are connected if they share a fraction, in this particular example more than one half, of their composing capabilities. In this case, one will have an undirected network, because the products are connected if they are similar, and so they are at the same level.
We want to stress that, when one takes into consideration real data, we expect that a country will likely move from basic products to more complex ones when it develops new capabilities. Thus, the time evolution of the technological progress should be closer to a taxonomy than to a proximity network.

4.1.2 Algorithm description

Let us define the *diversification* d_c as the number of products exported by the country c, as measured by the Revealed Comparative Advantage:

$$d_c = \sum_p M_{cp} \tag{35}$$

and the ubiquity u_p as the number of countries which export the product p:

$$u_p = \sum_c M_{cp}. \tag{36}$$

In order to obtain a product-product matrix we project M_{cp}:

$$B_{pp'} = \frac{1}{\max(u_p, u_{p'})} \sum_c \frac{M_{cp} M_{cp'}}{\sqrt{d_c}} \tag{37}$$

this way of normalizing the projection is similar to the one introduced by Zhou et al.[35]. The $\sqrt{d_c}$ factor takes into account the different contribution given by countries of different diversifications, by dividing the corresponding terms by the expected values in a random binomial case. Nevertheless, since the exponents of ubiquity and diversification are somehow arbitrary, we have checked a posteriori their goodness by means of the toy model and the sample matrices discussed in the next section. Moreover, in order to obtain the adjacency matrix of a network with number of edges of the same order of magnitude of the number of products we select only the maximum entry of each row, excluding the diagonal elements. In other words, for each product p we look for the product $p' \neq p$ which maximizes the normalized probability $B_{pp'}$ to be exported in a pair. Possible degeneracies are removed by looking at which product contributes the most with respect to its column; in other words, we pick the product whose column has the smallest elements. This same criteria, applied to the two products that are finally connected, defines the direction of the link. As we will show in the following, this filtering procedure is able not only to discard redundant and noisy information but also to define a set of preferred patterns for industrialization and development policies.
We point out that this procedure, in principle, could be applied using only the data which refers to one year, while we actually have 38 different matrices in the 1963-2000 dataset and 16 in the 1995-2010 dataset. While it would be natural to apply this algorithm for each matrix of every year, we preferred to aggregate them in a single matrix with the same columns (the 538 or 1131 exported products) and, as rows, all countries, including repetitions due to different years. In this way, most of the fluctuations are averaged out.
In the following section we will describe the properties of the Taxonomy network we obtained by applying our algorithm on the complete M_{cp} matrix.
For the sake of completeness we mention that, using our notation, the Product Space introduced by Hidalgo et al. is based on the proximity $\phi_{pp'}$ between the products p and p', which is defined as[66]:

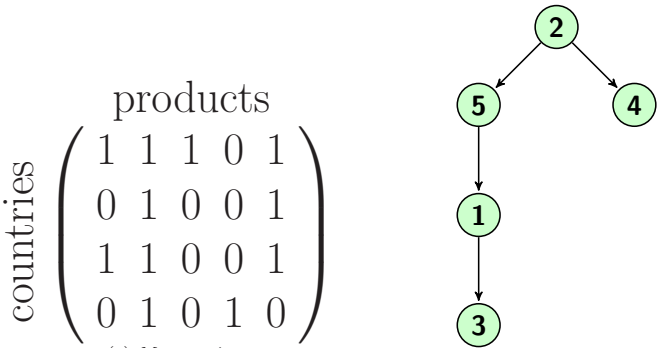
$$\phi_{pp'} = \min\left(\frac{\sum_c M_{cp}M_{cp'}}{u_p}, \frac{\sum_c M_{cp}M_{cp'}}{u_{p'}}\right). \tag{38}$$

This expression is quite similar to Eq.37 and, when used without any further filtering process, leads to an almost complete weighted network. The purpose of our maximum picking procedure is to enhance the signal to noise ratio in such a way to build a conceptually different network, whose links are directed and related to necessity instead of proximity.
In summary, the differences between the Taxonomy Network and the Product Space are i) the presence of directed links, with a clear causality meaning; ii) the reduction of the number of link from order N^2 to order N and iii) the different normalization, which takes into account the different diversifications o countries.

4.1.3 Tests of the algorithm

Sample matrices

Now we give an example of the output of our algorithm, starting from a simple M_{cp} matrix. We show both the matrix and the resulting taxonomy network in Fig.59. Here countries are ordered in rows and products in columns; for example, the second country produces the second and the fifth product. Now we focus on the relationship between the structure of the matrix and the one of the network.



(a) M_{cp} matrix

(b) Taxonomy Product Network

Figure 59: On the left, a sample M_{cp} matrix. On the right, the resulting Taxonomy Network. One can notice how the presence of products in the export baskets of the countries influences their position in the network. For example, the ubiquitous product 2 becomes the root.

Product 2, which corresponds to the second column, is made by all the countries: this means that, probably, the capabilities needed are relatively few or simple to achieve. On the contrary, products 3 and 4 are exported by only one country, so we can argue that very specific features, which has been developed only by countries 1 and 4 respectively, are required by these products. Products 5 and 1 lay somehow in the middle. The resulting network is depicted in Fig.59. The ubiquitous product 2 results to be the root, and it is needed to make all the other products. In particular, the fact that country 4 (fourth row) exports only products 2 and 4 suggests that the capabilities needed to produce 2 are a mandatory condition to produce 4. The left branch is constituted by a chain of products built following the same line of reasoning.

Capabilities-based model

In order to further test our algorithm we use a simplified version of the model proposed in Chapter 3. This allows us to have access both to a realistic M_{cp} matrix, that embodies the main stylized facts of the real data on which we will apply the algorithm, and to an artificial taxonomy that is used to generate it. By applying the algorithm to this artificial M_{cp} we come out with a taxonomy that can then be compared to the one we used to build the matrix. For completeness and readability we give here a precise definition of the model we use for this test, even if this is a partial of the contents of Chapter 3.

Definition of the model. The construction of the product taxonomy starts with R root products. Each of these products needs only one capability in order to be produced. At this stage we intend a capability as the *minimal* and *non-trivial* endowments needed in order to produce a product. By *non-trivial* we mean that a given capability is not owned by all the countries by default (in a real-world example a trivial capability could be water or sunlight). By *minimal* we mean that a capability is the smaller set of endowments which makes the difference between being able or not to produce a new product in at least one case (in a real world example a single oil well will not make a country an oil exporter while a vast oilfield

can).
The product taxonomy is then built as follows:

1. At each time step a new capability is introduced.
2. The new capability defines a new product p' by being added at random to one of the existing products p with a uniform probability.
3. A directed link is inserted from p to p'.

Then the M_{cp} matrix is built as follows:

1. A diversification d_c is assigned to each country c; the specific value is extracted from a real-world distribution.
2. The country chooses randomly d_c products from the taxonomy; the probability of choosing a particular product is inversely proportional to the number of capabilities (i.e. the distance from the root) associated with that product.
3. All the products that are on the shortest path from the root of the corresponding tree to any chosen product are assigned to the country c.

The values of the d_c are chosen such that the distribution of the diversification in the model is similar to the one coming from the real data.
For more detailed discussions about the meaningfulness of this model and its ability to reproduce the observed data refer to Chapter 3.

4.1.4 Analysis of the taxonomy network

In this section we present a study of the two taxonomy networks built starting from empirical data.
The network we obtain from the 1995-2000 data has 1131 vertices (this number is, obviously, equal to the number of products) and 985 edges, while the 1963-2000 network has 538 vertices and 456 edges. So they are quite sparse and not fully connected (this is due to both the intrinsic heterogeneity of the products and to our filtering procedure, which selects at most one link per row. As we will see in the following, this filtering permits to identify the most relevant links from the point of view of the observed time evolution). In both networks we have about one hundred components with heterogeneous sizes. However, most of these components have a well defined economical and technological meaning. In Fig.60 we show the largest component of the taxonomy network built from the 1995-2010 export matrices. Green filled nodes represent products that are exported by Sweden in the year 2010, while the red ones have $M_{cp} = 0$. The diameter of the vertices is proportional to the logarithm of the product complexity, whose measure has been defined in [5, 26]. One can notice a clear tendency to have products of large complexity on the border of the network, while more basic products lay in the center and have a higher degree, that is, centrality tends to be anticorrelated with complexity. This behavior is in agreement with our hypothesis that the few capabilities needed to produce *low* complexity products represent a necessary condition to be able to produce *high* complexity products, in the spirit of the Taxonomy Network concept we introduced in the previous sections. A zero-order validation of this idea can be found in the fact that for both networks about the 70% of the edges point from a high to a low complexity product. In a purely random framework we

would expect this value to approach one half, given the presence of hundreds of edges. On the contrary, we observe a situation with a negligible p-value, and so we can conclude that the direction of links is not given by a fair coin flipping.

Now we want to turn our attention to how countries occupy the Taxonomy Network. In particular, we would like to study the possibility to link macroeconomic features of the countries with the properties of the vertices corresponding to the products they export. We have noticed that developed countries tend to occupy outlying vertices. In order to better study this feature we need a measure of the centrality of a given vertex which takes into account not only its degree but also the direction of the links, in such a way to pass the received authority following the links. One possible measure is the PageRank [13]. In order to evaluate the degree of development of a given country we count its products weighting more the ones that lie away from the center, that is, vertices with a low PageRank. We study the sum of the inverse PageRank of the exported products of a given country c:

$$D_c = \sum_p M_{cp} PR_p^{-1} \tag{39}$$

which we call *disposition* of the country. In Fig.61 we plot this quantity versus the so-called fitness [5, 26], that is a measure of the growth potential of a country, for all the countries in our database, referring to the year 2000, finding an impressive correlation between the two ($R^2 = 0.92$). For clarity purposes we have taken the logarithm of both variables. This is an interesting link between a network based quantity and the fitness, which is the result of an algorithmic interplay between the countries and the complexity of the products they export.

Study of countries' development

One of the most important features of this approach is the visualization of countries' economic development. In order to show how clearly patterns emerge when studying specific countries through time, we focus on a specific example of the development of one of the so-called "Asian Tigers", South Korea, which is often reported as a case study for a successful industrialization process. In particular, in Fig.62 we show a technological component of the taxonomy network. The root product is *radio broadcast receivers*, while on the border we find *automatic data processing machines*, that is, computers. An evident exception is *umbrellas*, a product which obviously has nothing in common with the others and remains connected to this component despite the filtering procedure which, on the contrary, seems to perform well for the other vertices. In Fig.63 we show the time evolution of the South Korean export for this component. The colors are proportional to Balassa's Revealed Comparative Advantage (RCA) [9]: light blue means that the product is not exported, while the different shades of red are proportional to an RCA increase. In 1963 this country did not export any product of this component in a significant way. After three years, the root starts to be produced together two close products. In the following years South Korea explores the network, reaching in 1993 an impressive level of diversification. In 2000 South Korea focused its exports on borderline products, as expected from an already developed country from the disposition analysis we presented above. The presence of a meddlesome product (in this case, *umbrellas*) is due to noise, but it can be spotted thanks to its RCA behavior, which is uncorrelated with the other nodes. So, even if the probabilistic approach we use to define the network can lead to spurious results, like the presence of unexpected products in otherwise well defined clusters, one can see that a careful

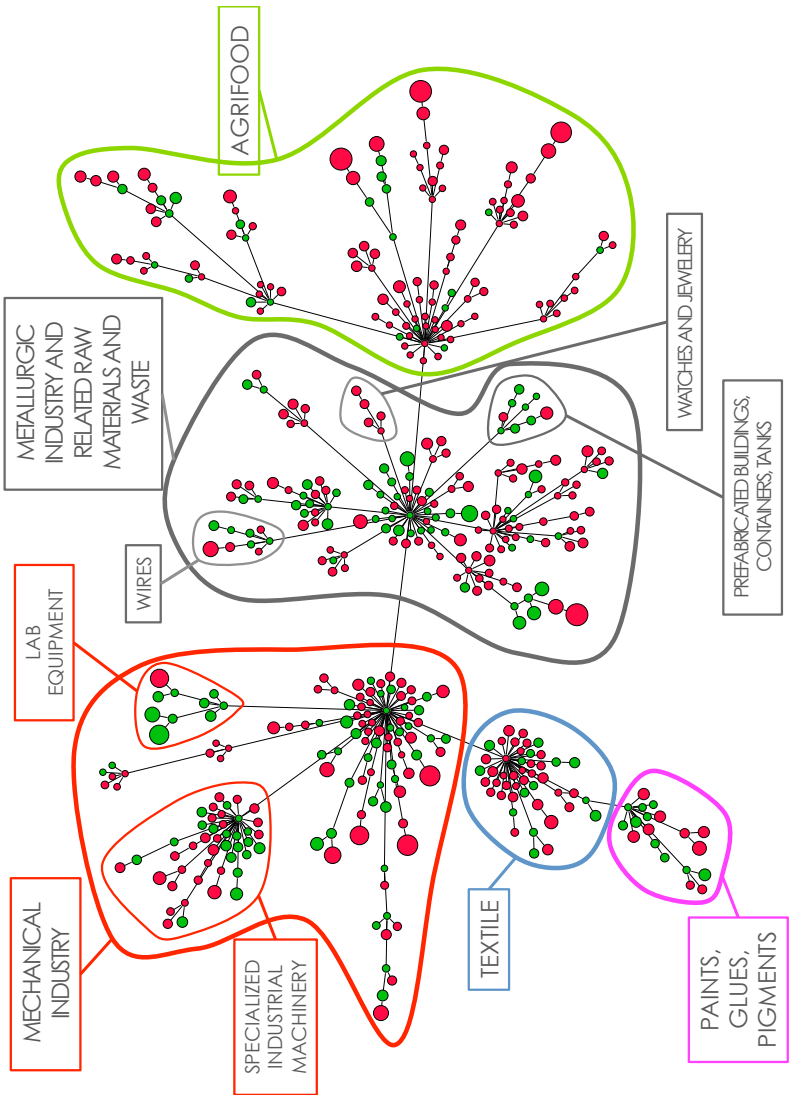


Figure 60: The largest component of the taxonomy network built from the 1995-2010 database. The colors refer to the value of the M_{cp} matrix for Sweden, year 2010: green is 1, red is 0. The diameter of the vertices is proportional to the logarithm of the product complexity, as defined in [26]. Already from a visual inspection one could argue that a good strategic move for Sweden could be to produce the red, high complexity product in the Lab Equipment community.

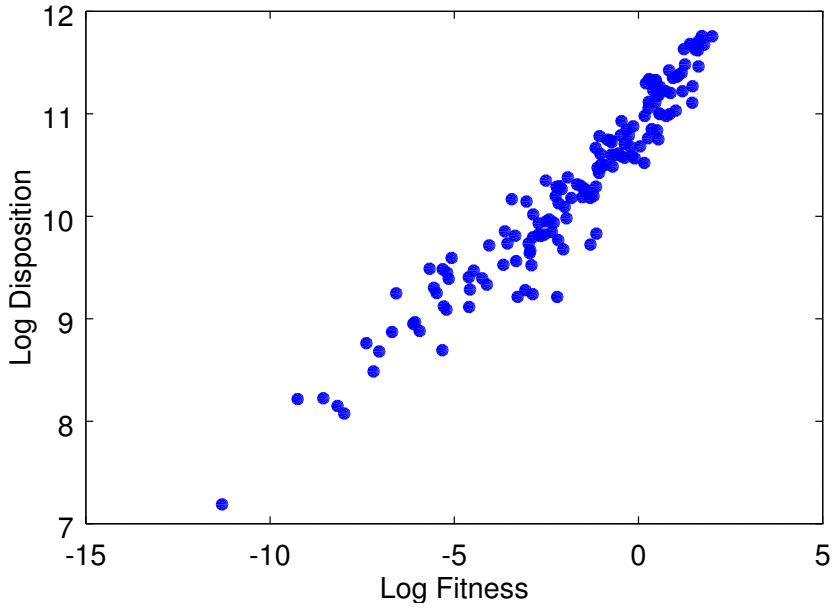


Figure 61: The disposition and the fitness for each country. There is a clear correlation between the two variables, indicating a link between the growth potential of a country and its disposition on the taxonomy network.

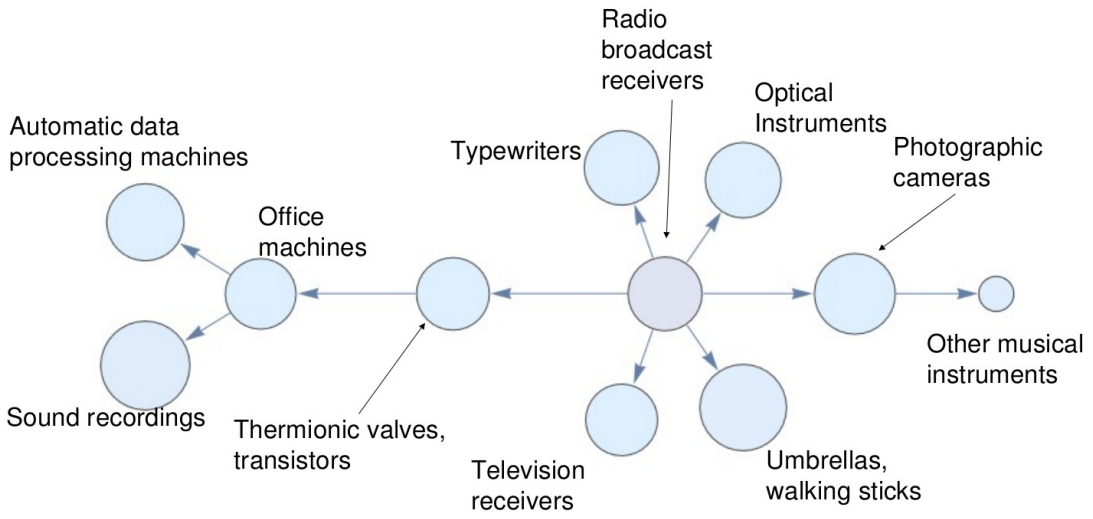


Figure 62: A component of the Taxonomy Network. All nodes are clearly member of the same technological community, but *umbrellas*, whose presence is due to noise.

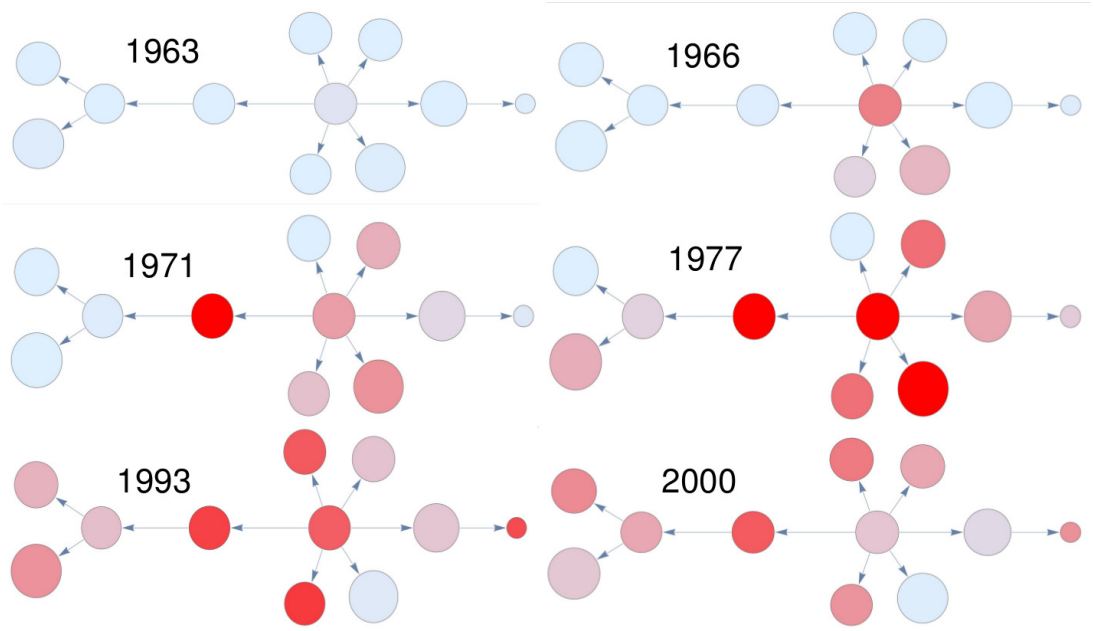


Figure 63: An example of the time evolution of a component of the Taxonomy Network. The studied country is South Korea. The red fillings represent an Increase of the RCA value. One can notice the diffusion from the center (root product) towards the borders of the component.

analysis of the dynamics clearly points to the fact that this site is anomalous with respect to the cluster considered.

4.2 DYNAMIC APPROACH: THE ENABLING MATRIX

Now we try to tackle the problem from a dynamical perspective. Namely we want to quantify the correlations between the presence of a product in a country's export basket and the appearance of a different one at a later time. In order to do this we try to quantify how much the presence of a product p influences the possible turning on of another product p'. One possible measure of this helpfulness is the frequency of the activations given the presence of an already activated product. In practice, first one has to calculate the three dimensional Activation Matrix

$$Z_{cpy} = M_{cpy} - M_{cp(y-1)} \tag{40}$$

where $y \in [1964, 2000]$. We focus only on the activations of products ($Z_{cpy} = 1$) and so we ignore the cases in which $Z_{cpy} = -1$, which corresponds to the dismission of a certain production. In order to evaluate the frequency of activation of p' given the presence of another product p, we calculate the Enabling Matrix

$$C_{pp'} = \frac{\sum_{c,y} Z_{cp'y} M_{cp(y-1)}}{\sum_{c,y} Z_{cp'y}} \tag{41}$$

where, in the previous formulas, the matrix operations are intended as element by element operations. The elements of this matrix represent an empirical proxy of the strength of the directed link from p to p'. Obviously, this could be a rough approximation, because in principle one can think that the more a product is present, the more it will appear to be necessary even if it could be not. For this reason we checked the weight of products' ubiquity, finding that, even if ubiquitous products tend to be more necessary, once that this effect is removed our results are substantially left unchanged. Another possibility is that it is the presence of a set of products that changes the probability that a country has to produce a new product, and not only one as supposed above. In this case it is not straightforward to calculate the relative usefulness of the products and a suitable approach will be described in the next section. Nevertheless as a rough approximation, one can implement a "mean field" version of the previously described enabling matrix. In practice we give for every activation a score $1/n$ to each product which was already exported during the previous year, where n is the number of the products exported by the studied country, and so, in general, a function of p, c and y. Using this approach the empirical strength of the link from p to p' will be given by the sum of the scores collected by the different countries through the years. In this way, the Enabling Matrix is calculated supposing a mean field interaction, in contrast with the previous approach, in which the interaction was assumed to be pairwise.
Once that we defined an empirical benchmark regarding the time evolution of countries' export, it is interesting to assess its connection with the taxonomy network. To do so we sort the rows of the two enabling matrices from the largest to the smallest element and we check the position of the matrix element that we would have picked following the taxonomy network. In other words, we check how often a path suggested by the Taxonomy Network is actually followed, with respect to the other possible links. The results are depicted in Fig.64. The taxonomy network correctly identifies most of the top empirical temporal connections between products, both in the pairwise and in the mean field approach. In particular, about one hundred of the links are in the top 2.5% and about half of them are in the top 10%. The taxonomy network performs slightly better in the mean field case.
This result points out a clear connection between the taxonomy network, which is built up without considering the time evolution of the exports, and the properties of countries' development in terms of the temporal connections among the products they are exporting and the ones they will export.

4.3 MACHINE LEARNING APPROACH.

In this section we describe how to apply a general Machine Learning approach to the problem of predicting the appearance of new links in the M_{cp} matrix. First we define as a *state* the binary vector that defines the export of a given country at a given time. We want to assess empirically the probability of a new product to appear in a future *state*, given the present *state*. In principle one should look at all the possible combinations of exported products and correlate them with the appearance of each new one. But besides the number of these combinations being overwhelmingly large (2^{N_p}, with N_p of order 10^3), we can only observe a finite and much smaller number of such combinations that can be used to tune our expected values. Moreover each of the observed combinations is in turn very "rarely" observed (each combination is observed a number of times which is of order 1), being the space of possible

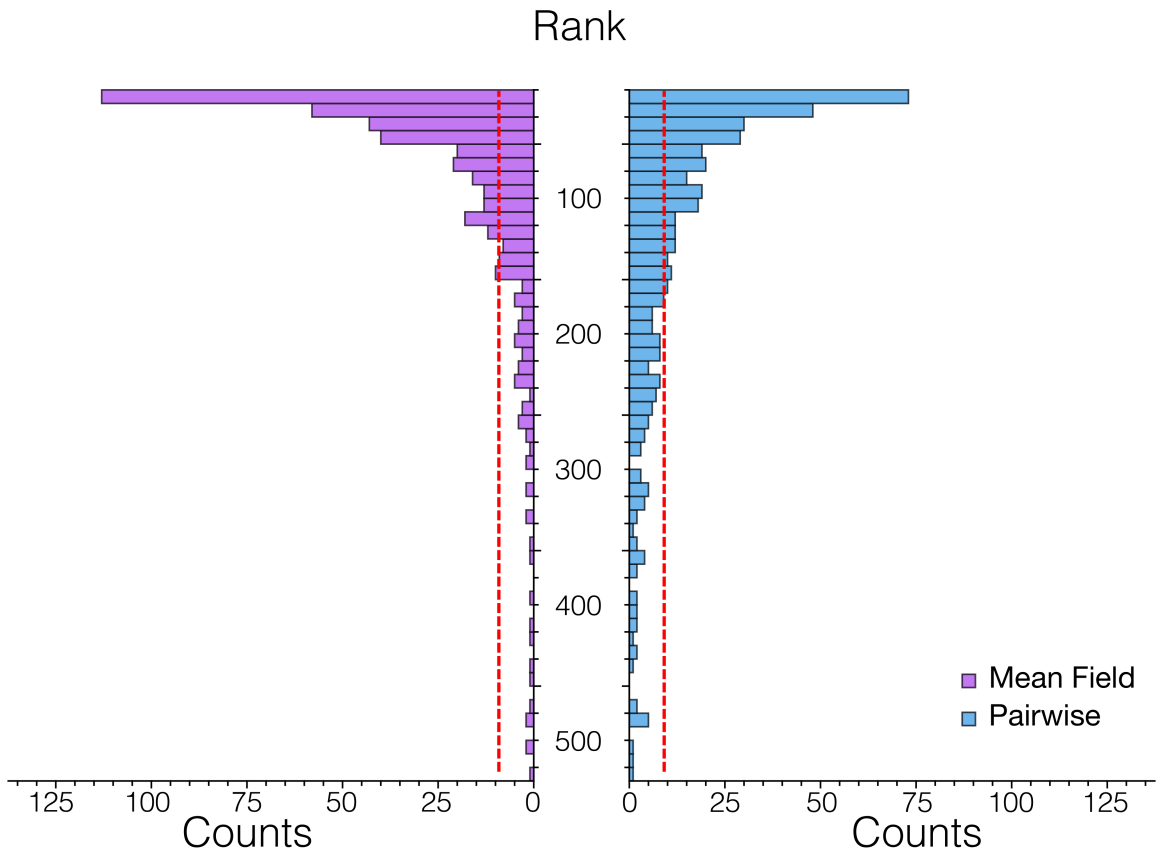


Figure 64: The Taxonomy Network correctly finds the empirically most active edges, as empirically calculated with the Enabling Matrix, using two different approaches which are described in the text. Here we show the distribution of the rankings of the Enabling Matrix elements selected by the Taxonomy Network. The rankings are calculated ordering the rows of the Enabling Matrix from the largest element to the smallest one. The resulting distribution is peaked around small values, implying that a large fraction of the links suggested by our network correspond to the largest elements of the Enabling Matrices.

combinations so large. Thus for each product that we want to predict we need to select a small subset of relevant affine products. This is precisely the aim of a *Decision Tree*.

First of all we divide our dataset (this analysis is performed on the 1995-2010 data) in two parts. The 1995-2007 part is used as a training dataset, while the remaining data are used to benchmark our results. For each product we collect all the *states* in which the corresponding bit was set to 0, in other words all those cases in which a nation was not exporting the product in a given year. We associate to each *state* an *answer* which may be 1 or 0: 1 if the product "turns-on" after a given time lag L in the export basket of the corresponding country, 0 otherwise. A decision tree can be built by selecting those particular bits of information in the collection of *states* associated to a given product that permit the best discrimination of the corresponding answers.

4.3.1 Construction of the Decision Trees

A possible practical approach is described in [67], of which we give here a synthetic exposition.

To build a decision tree for product i first we select from our training set all the *states* in which a particular country was not exporting product i. To each *state* we assign the corresponding answer, i.e. either 1 if product i is exported by that particular country after L years or 0 otherwise. We then compute the information entropy $E_i = -\sum_{k=\{0,1\}} p_k^{(i)} \log(p_k^{(i)})$ of the distribution of the collected answers, where $p_k^{(i)}$ is the fraction of the bits with value k in the set. The next step is to look at all the possible $N_p - 1$ partitions of the set of the answers by means of the bits of the corresponding *state*. So for each $j \neq i$ we split the set of the answers in two sets: the first with all the answers corresponding to *states* where the j-th bit was equal to 1, the other with the ones corresponding to *states* with j-th bit equal to 0. Finally we compute the information gain for each partition as follows:

- First we compute the information of the partition as the average of the information entropies of the two sets, weighted by the number of elements in each set
- Then we define the *information gain* as the difference between E_i and the new average entropy

Two new branches are added to the decision tree, corresponding to the splitting with the highest *information gain*. The procedure is then iterated recursively on the two resulting subsets, and new branches are attached below the previous ones. The iteration stops when for a given subset is not possible to find a splitting which gives an information gain higher than a given threshold. At the end of the iteration we are left with a binary tree, i.e. a tree in which each node except the leaves have out-degree equal to 2. This tree is the *Decision tree* for product i. Each node corresponds to a product $j \neq i$ and the two outgoing links correspond to the two possible states (0 or 1) of product j.

The *Decision Tree* acts as a classifier: following the branches to recursively split sets of *answers* one should be able to group coherent subsets of answers as a function of some bits of the states. This approach can lead to undesirable results when the splittings are to fine. Thus we perform some statistical tests in order to determine the quality of the classification given by each node.

4.3.2 Calibration of the trees

We use our training set also to calibrate each of our $N_p = 1131$ decision trees and to perform some statistical tests on the significance of our results. Namely we assign to each node some statistics related to the states that are selected by the branches until that point. For any node of each tree we evaluated (in the training set) the following statistics:

- N: Number of events selected
- f: Fraction of answers with value 1 selected
- P_b: An indicator of the significance of the selection under a Binomial Null Hypotesis. More precisely we first compute p, i.e. the fraction of positive answers in the full set of answers for the product i. Then we define

$$\tilde{P}_b = \binom{N}{N \cdot f} p^{N \cdot f} (1-p)^{N(1-f)}$$

  Finally we compute $P_b = -\log(\tilde{P}_b)$. The higher P_b the higher the significance.

An example of a decision tree together with it's calibration statistics is shown in fig. 65

4.4 ACCURACY OF PREDICTIONS

We perform some statistical tests on how the three approaches that we proposed are able to predict new links in the M_{cp} matrix. Namely we use the 2007 data as present states to be used to infer new links in the 2010 countries-products network. Thus we build the *Enabling Matrix*[1] in the 1995-2007 dataset. Next we rank all the possible new links in the countries-products network (i.e. all those link that weren't present in the 2007 states) by the likelihood of being active in 2010 as given by the two different approaches. More in detail the two rankings are built as follows:

- **Enabling Matrix**: for each nation we compute the sum of the "enabling signals" given by all the products present in the export basket in 2007 to all the absent products. Namely, if $\mathcal{P}^c$ is the set of active products for country c, we compute $S_{cp} = \sum_{p' \in \mathcal{P}^c} C_{p'p}$ for all $p \notin \mathcal{P}^c$. Then the links are ranked by the value of S_{cp}
- **Decision Trees**: we project the state of each nation on all the decision trees corresponding to unactive products in 2007. Each nation c thus performs a path D_{cp} on all the decision trees corresponding to products p not owned in 2007. On each path we choose the node with the highest P_b. We use the value of f for the selected node, as obtained from the calibration set, as a proxy to rank the likelihood that the corresponding c-p links will be active in 2010

In fig. 66 we show the fraction of correct predictions for a time lag $L = 3$ in a subset of increasing size of the set of all the possible new links, ordered with the two methods. The *Decision Trees* are able to outperform the Enabling Matrix by a factor 2. For the top 500 links the fraction of correct prediction is around 15-16% for the Decision Trees and around 8% for the Enabling Matrix. As a reference a Random Choice would result in a 3.7% success rate.

1 In this section we discuss only the results for the "mean-field" version. The pairwise implementation leads to similar but slightly worse results.

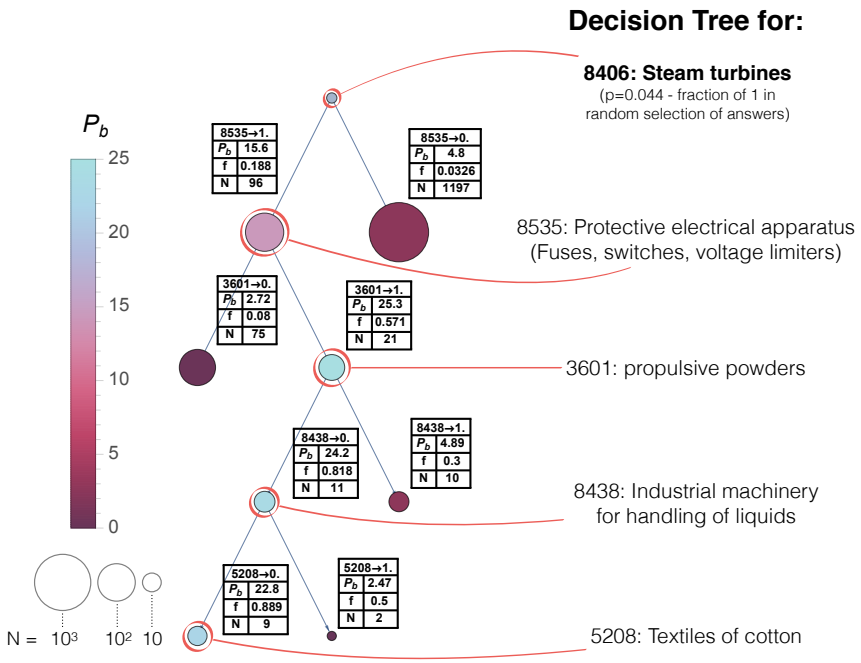


Figure 65: Decision Tree for product 8406: Steam Turbines. Here is shown the resulting decision tree for product 8406. The root vertex corresponds to the analyzed product itself. The two nodes in the next layer correspond to states with opposite value for product 8535 and so on. The tables show the statistics for the answers corresponding to the states selected by the nodes. A color/size coding is provided to guide the eye. Going down in the tree the nodes individuate sets with increasing fraction of positive answers but with a smaller size. The P_b indicator increases in the first layers and then starts to decrease, indicating that the statistical value of the classification is getting worse.

4.5 CONCLUSIONS

In this chapter we have shown how the nested structure of the countries-products network contains useful information about the technological relations among products. In particular an analysis of co-occurrences reveals a network structure which is significantly related with the paths of development followed by national economies. Moreover we have shown that these paths are robust and an analysis of past dynamics can be fruitfully used to predict the development of nations which are found in similar situations as those already observed. This approach is conceptually similar to the "method of analogues" used in chapter 2 and its success strengthens the idea that the development of nations can, at least in certain regimes, be regarded as something non very different from a deterministic dynamical system which we are able to describe to some extent. The methods developed in this chapter are a precise quantification of these ideas and the fact that these are connected with a defined network topology can open the field of a large number of practical applications. First of all these quantitative methods can be of great help in the industrial planning of developing nations from a policy making perspective. Moreover the approach is general enough to be applied

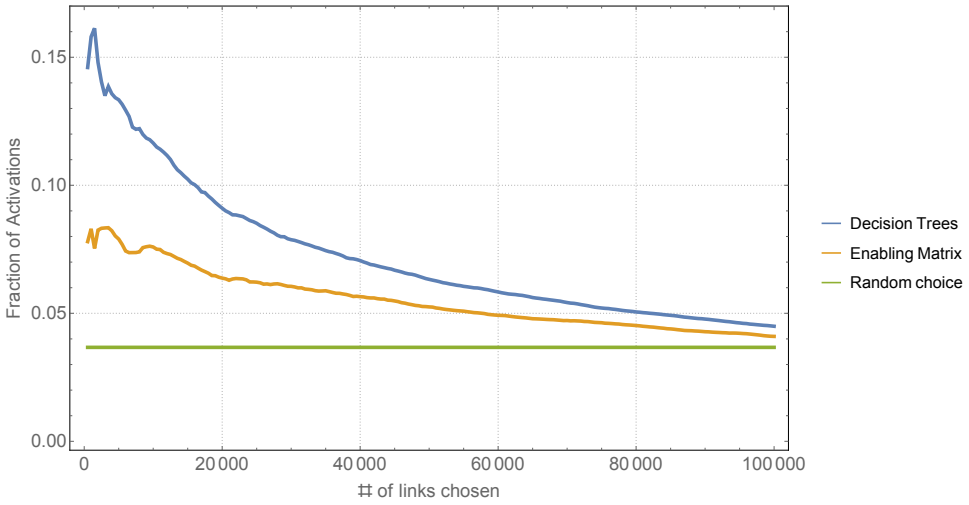


Figure 66: Accuracy of predictions for new links in the countries-products network. We rank the set of possible new links based on the strength of the signal given by the two methods (Decision Trees in blue and Enabling Matrix in orange) when applied to the 2007 states. We look at the fraction of links that are found active in 2010, when choosing subsets of increasing size of the ranked prediction set. The first point corresponds to a choice of the top 500 links. For comparison the expected value (0.037) of a random selection of links is reported (Green). The accuracy (in terms of improvement with respect to the random case) of the Decision Tree approach is greater than that of the Enabling Matrix by a factor 2.

to different bipartite networks. To mention one, we are exploring applications to systems like consumer-purchased products. We argue that our methods could allow an automatic pairing of products that are similar on the basis of non-observable features that are relevant for the consumer (the analogue of capabilities). It is supposed that consumers also follow precise paths in the discovery of new niches (one could think of the first time one listens to the most famous song of a singer, and is then stimulated to explore its discography) and being able to predict these paths could be profitable for many categories of online shops and media distributors.

BIBLIOGRAPHY

[1] Guillaume Gaulier and Soledad Zignago. "BACI: International trade database at the product-level." In: (2009).

[2] Robert C Feenstra et al. *World trade flows: 1962-2000*. Tech. rep. National Bureau of Economic Research, 2005.

[3] César A Hidalgo and Ricardo Hausmann. "The building blocks of economic complexity." In: *Proceedings of the National Academy of Sciences* 106.26 (2009), pp. 10570–10575.

[4] Ricardo Hausmann, Jason Hwang, and Dani Rodrik. "What you export matters." In: *Journal of economic growth* 12.1 (2007), pp. 1–25.

[5] Andrea Tacchella et al. "A new metrics for countries' fitness and products' complexity." In: *Scientific reports* 2 (2012).

[6] David Ricardo and Ronald Max Hartwell. *On the principles of political economy, and taxation*. Vol. 165. Penguin Books Harmondsworth, 1971.

[7] Guido Caldarelli et al. "A network analysis of countries' export flows: firm grounds for the building blocks of the economy." In: *PloS one* 7.10 (2012), e47278.

[8] Andrea Tacchella et al. "Economic complexity: conceptual grounding of a new metrics for global competitiveness." In: *Journal of Economic Dynamics and Control* 37.8 (2013), pp. 1683–1691.

[9] Bela Balassa. "Trade liberalisation and "revealed" comparative advantage1." In: *The Manchester School* 33.2 (1965), pp. 99–123.

[10] Sanjaya Lall. "The Technological structure and performance of developing country manufactured exports, 1985-98." In: *Oxford development studies* 28.3 (2000), pp. 337–369.

[11] César A Hidalgo et al. "The product space conditions the development of nations." In: *Science* 317.5837 (2007), pp. 482–487.

[12] Ricardo Hausmann and César A Hidalgo. "The network structure of economic output." In: *Journal of Economic Growth* 16.4 (2011), pp. 309–342.

[13] Lawrence Page et al. "The PageRank citation ranking: Bringing order to the web." In: (1999).

[14] *Goldman Sachs Website*. URL: http://www.goldmansachs.com/our-thinking/view-from/a-view-from%5C-brazil/index.html.

[15] *The Economist Website*. URL: http://www.economist.com/blogs/freeexchange/2012/10/%5Cgrowth?zid=305%5C&ah=417bd5664dc76da5d98af4f7a640fd8a.

[16] Ricardo Hausmann and César A Hidalgo. *The atlas of economic complexity: Mapping paths to prosperity*. MIT Press, 2014.

[17] *The Atlas of Economic Complexity Website*. URL: http://www.http://atlas.media.mit.edu/rankings/.

[18] Emanuele Pugliese, Andrea Zaccaria, and Luciano Pietronero. "On the convergence of the Fitness-Complexity Algorithm." In: *The European Physical Journal Special Topics* 225.10 (2016), pp. 1893–1911.

[19] C. M. Reinhart and K. S. Rogoff. "Growth in a Time of Debt." In: *American Economic Review* 100 (2010), pp. 573–78.

[20] P. Krugman. "The Excel Depression." In: *New York Times* 100 (2013).

[21] E. N. Lorenz. "Atmospheric predictability as revealed by naturally occurring analogues." In: *J. Atmos. Sci* 26 (1969), pp. 636–646.

[22] E. N. Lorenz. "Three approaches to atmospheric predictability." In: *Bull. Am. Meteorol.* 50 (1969), pp. 345–349.

[23] Robert Costanza et al. "nature. com." In: *Policy* (2014).

[24] F. Cecconi et al. "Predicting the future from the past: An old problem from a modern perspective." In: *Ann. J. Phys.* 80 (2012), p. 1001.

[25] H. Poincaré. "Sur le probleme des trois corps et les l'equations de la dynamique." In: *Acta Math.* 13 (1890), pp. 1–270.

[26] Matthieu Cristelli et al. "Measuring the Intangibles: A Metrics for the Economic Complexity of Countries and Products." In: *PloS one* 8.8 (2013), e70726.

[27] C.-H. Park and S. H. Irwin. "What do we know about profitability of technical analysis." In: *Journal of Economic Surveys* 21 (2007), 786Ð826.

[28] Robert M May. "Will a large complex system be stable?" In: *Nature* 238 (1972), pp. 413–414.

[29] Robert MacArthur. "Fluctuations of animal populations and a measure of community stability." In: *ecology* 36.3 (1955), pp. 533–536.

[30] J. Bascompte and P. Jordano. "Plant-animal mutualistic networks." In: *Annu. Rev. Ecol. Evol. Syst.* 38 (2007), pp. 567–93.

[31] Bruce D Patterson and Wirt Atmar. "Nested subsets and the structure of insular mammalian faunas and archipelagos." In: *Biological Journal of the Linnean Society* 28.1-2 (1986), pp. 65–82.

[32] Wirt Atmar and Bruce D Patterson. "The measure of order and disorder in the distribution of species in fragmented habitat." In: *Oecologia* 96.3 (1993), pp. 373–382.

[33] JR DARLINGTON. "Jr.(1957): Zoogeography: the geographical distribution of animals." In: *john Wiley* 8 (), pp. 488–510.

[34] Tao Zhou et al. "Solving the apparent diversity-accuracy dilemma of recommender systems." In: *Proceedings of the National Academy of Sciences* 107.10 (2010), pp. 4511–4515.

[35] Tao Zhou et al. "Bipartite network projection and personal recommendation." In: *Physical Review E* 76.4 (2007), p. 046115.

[36] Güler Ergün. "Human sexual contact network as a bipartite graph." In: *Physica A: Statistical Mechanics and its Applications* 308.1 (2002), pp. 483–488.

[37] Robert Riding. *The Ecology of the Cambrian Radiation*. Columbia University Press, 2001.

[38] Richard Fortey. *Life: An unauthorized biography*. HarperCollins UK, 2010.

[39] David C Catling et al. "Why O2 Is Required by Complex Life on Habitable Planets and the Concept of Planetary" Oxygenation Time"." In: *Astrobiology* 5.3 (2005), pp. 415–438.

[40] Xavier Fernàndez-Busquets et al. "Self-recognition and Ca2+-dependent carbohydrate–carbohydrate cell adhesion provide clues to the cambrian explosion." In: *Molecular biology and evolution* 26.11 (2009), pp. 2551–2561.

[41] Andrew Parker. *In the blink of an eye: how vision sparked the big bang of evolution*. Basic Books, 2009.

[42] Ricard V Solé, Pau Fernández, and Stuart A Kauffman. "Adaptive walks in a gene network model of morphogenesis: insights into the Cambrian explosion." In: *arXiv preprint q-bio/0311013* (2003).

[43] Tomas Kögel and Alexia Prskawetz. "Agricultural productivity growth and escape from the Malthusian trap." In: *Journal of Economic Growth* 6.4 (2001), pp. 337–357.

[44] Kenneth Pomeranz. *The great divergence: China, Europe, and the making of the modern world economy*. Princeton University Press, 2009.

[45] Pietronero L Cristelli M Tacchella A. "The heterogeneous dynamics of economic complexity." In: *Plos One - Under Review* (2014).

[46] Samir Suweis et al. "Emergence of structural and dynamical properties of ecological mutualistic networks." In: *Nature* 500.7463 (2013), pp. 449–452.

[47] Virginia Dominguez-Garcia and Miguel Muñoz. "Species importance ranking in Mutualistic Networks - *in preparation*." In: ().

[48] Wirt Atmar and Bruce D Patterson. "The nestedness temperature calculator: a visual basic program, including 294 presence-absence matrices." In: *AICS Research Incorporate and The Field Museum* (1995).

[49] Joern Fischer and David B Lindenmayer. "Treating the nestedness temperature calculator as a "black box" can lead to false conclusions." In: *Oikos* 99.1 (2002), pp. 193–199.

[50] Matthew R Nelson et al. "An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people." In: *Science* 337.6090 (2012), pp. 100–104.

[51] L. Napolitano. "*Forthcoming..*" In: ().

[52] Eli Filip Heckscher and Bertil Gotthard Ohlin. *Heckscher-Ohlin trade theory*. The MIT Press, 1991.

[53] David Ricardo. *Principles of political economy and taxation*. G. Bell and sons, 1891.

[54] Wassily Leontief. *Domestic production and foreign trade: The American capital position re-examined*. 1954.

[55] Wassily Leontief. "Factor proportions and the structure of American trade: further theoretical and empirical analysis." In: *The Review of Economics and Statistics* (1956), pp. 386–407.

[56] Edward E Leamer. "The Leontief paradox, reconsidered." In: *The Journal of Political Economy* (1980), pp. 495–503.

[57] Harry P Bowen, Edward E Leamer, and Leo Sveikauskas. "Multicountry, multifactor tests of the factor abundance theory." In: *The American Economic Review* (1987), pp. 791–809.

[58] Philippe Aghion and Peter Howitt. *A model of growth through creative destruction*. Tech. rep. National Bureau of Economic Research, 1990.

[59] Joseph Alois Schumpeter. *The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle*. Vol. 55. Transaction Publishers, 1934.

[60] Paul M Romer. "Endogenous technological change." In: *Journal of political Economy* (1990), S71–S102.

[61] Ricardo Hausmann and Dani Rodrik. "Economic development as self-discovery." In: *Journal of development Economics* 72.2 (2003), pp. 603–633.

[62] Stuart A Kauffman. *Investigations*. Oxford University Press, 2002.

[63] Stuart A. Kauffman. *The origins of order: Self-organization and selection in evolution*. Oxford university press, 1993.

[64] Richard R Nelson and Sidney G Winter. *An evolutionary theory of economic change*. Harvard University Press, 2009.

[65] Jan Fagerberg and Bart Verspagen. "Innovation studies: The emerging structure of a new scientific field." In: *Research policy* 38.2 (2009), pp. 218–233.

[66] César Hidalgo. "The dynamics of economic complexity and the product space over a 42 year period." In: *Center for International Development at Harvard University. Working Paper* 189 (2009).

[67] J. Ross Quinlan. "Induction of decision trees." In: *Machine learning* 1.1 (1986), pp. 81–106.

[68] Federico Battiston et al. "How metrics for economic complexity are affected by noise." In: *Complexity Economics* 1.1 (2014), pp. 1–22.

[69] Adam Smith and Joseph Shield Nicholson. *An inquiry into the nature and causes of the Wealth of Nations...* T. Nelson and Sons, 1887.

[70] Robert C Feenstra et al. *World trade flows: 1962-2000*. Tech. rep. National Bureau of Economic Research, 2005.

[71] M Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. "Patterns of dominant flows in the world trade web." In: *Journal of Economic Interaction and Coordination* 2.2 (2007), pp. 111–124.

[72] Tiziano Squartini, Giorgio Fagiolo, and Diego Garlaschelli. "Randomizing world trade. I. A binary network analysis." In: *Physical Review E* 84.4 (2011), p. 046117.

[73] Tiziano Squartini, Giorgio Fagiolo, and Diego Garlaschelli. "Randomizing world trade. II. A weighted network analysis." In: *Physical Review E* 84.4 (2011), p. 046118.

[74] Gene M Grossman and Elhanan Helpman. "Quality ladders in the theory of growth." In: *The Review of Economic Studies* 58.1 (1991), pp. 43–61.