ISSN 2281-4299

DIPARTIMENTO DI INGEGNERIA INFORMATICA
AUTOMATICA E GESTIONALE ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

Inference for Nonparametric Productivity Networks: A Pseudo-likelihood Approach

Moriah B. Bostian
Cinzia Daraio
Rolf Fare
Shawna Grosskopf
Maria Grazia Izzo
Luca Leuzzi
Giancarlo Ruocco
William L. Weber

Technical Report n. 6, 2018

Inference for Nonparametric Productivity Networks: A Pseudo-likelihood Approach

Moriah B. Bostian[1], Cinzia Daraio[2], Rolf Färe[3,4], Shawna Grosskopf[4], Maria Grazia Izzo[2,5], Luca Leuzzi[6,7], Giancarlo Ruocco[5,7], and William L. Weber[8]

[1]*Department of Economics, Lewis & Clark College, Portland, OR USA*

[2]*Department of Computer, Control and Management Engineering A. Ruberti (DIAG), Sapienza University of Rome, Italy. Correspondence to: Cinzia Daraio, DIAG Sapienza University of Rome, via Ariosto, 25, I-00185 Rome (Italy), Tel. (+39) 06 77274 068, Fax (+39) 06 77274 074, E-mail:* daraio@diag.uniroma1.it.

[3]*Department of Applied Economics, Oregon State University, Corvallis, OR USA*

[4]*Department of Economics, Oregon State University, Corvallis, OR USA*

[5]*Center for Life Nano Science, Fondazione Istituto Italiano di Tecnologia (IIT), Rome, Italy*

[6]*CNR-NANOTEC, Institute of Nanotechnology, Soft and Living Matter Lab, Rome, Italy*

[7]*Department of Physics, Sapienza University of Rome, Italy*

[8]*Department of Economics and Finance, Southeast Missouri State University, Cape Girardeau, MO USA*

Dated: July, 2018

Abstract

Networks are general models that represent the relationships within or between systems widely studied in statistical mechanics. Nonparametric productivity networks (Network-DEA) typically analyzes the networks in a descriptive rather than statistical framework. We fill this gap by developing a general framework -involving information science, machine learning and statistical inference from the physics of complex systems- for modeling the production process based on the axiomatics of Network-DEA connected to Georgescu-Roegen funds and flows model. The proposed statistical approach allows us to infer the network topology in a Bayesian framework. An application to assess knowledge productivity at a world-country level is provided.

Key Words: Network DEA, Bayesian statistics, Generalized multicomponent Ising Model, Georgescu Roegen flows and funds model, knowledge production, interactions, Input-output models.

JEL Classification: C1,C11, C13, C14;

1 Introduction

The economic systems are more and more conceived as complex eco-systems whose operating models have to be developed considering the interplay of many dimensions. In economics we are witnessing the development and expansion of networks (Schweitzer et al. 2009a,b; Elsner, et al. 2015) up to the point that Kirman (2016) posed the question about networks as a potential paradigm shift for economics. Social networks are a commonplace since the rapid development of social networking and the growth of social media. In sociology networks are widely used to study the structure and patterns of social interactions (Scott, 2017). In sociology there is a far longer history (see e.g. Granovetter, 1973) and a much broader meaning of networks including digital and online networks but also face-to-face relationships, political associations and connections, economic transactions and geopolitical relations.

There is a rich literature on the nonparametric estimation of efficiency based on networks, the so called Network DEA (Data Envelopment Analysis) or NDEA (Färe and Grosskopf (2000), Färe, Grosskopf and Whittaker (2007, 2014), Cook et al. (2010, 2014), Bogetoft et al (2009), Kao, 2014)). All these models aim to assess the performance of complex systems. A guiding thread among all the different Network DEA models is the consideration about the limit of the "independent" structure and the inclusion in the Network model of a structure of dependence (or interdependence) among the Decision Making Units (DMUs) or their branches. Indeed, "*Many production systems (technologies) may be conceptualized as the joint, interacting action of a finite number of production subtechnologies called Activities*" (Shephard and Färe, 1975 p. 43).[1]

In the concluding section of his recent book, Kao (2017) highlights some extensions and open questions, in particular the choice of the network model to apply in given empirical contexts, highlighting that it is not possible *a priori* to choose what is the best network model to apply.

However, most of the existing works analyse the productivity networks in a descriptive way, without considering them in a statistical framework. This means that the network structure is generally assumed and not estimated.

An exception is Trinh and Zelenyuk (2015) that propose a bootstrap-based comparison between average DEA-NDEA efficiency scores and their distributions but without questioning the network structure, that is assuming the NDEA structure. Indeed there is a rich literature on the statistical inference on nonparametric productivity frontier models (see Simar and Wilson, 2013, 2015). Bayesian inference is currently applied in the Stochastic Frontier Analysis framework (see Van Den Broeck et al., 1994, Kumbhakar et al. 2012,

[1] Quotation reported at the introduction of Färe, Grosskopf and Whittaker (2007).

Parmeter and Kumbhakar, 2014, Sickles and Zelenyuk, 2018). Bayesian inference has been also applied in a stochastic DEA framework (Tsionas and Papadakis, 2010). However, to the best of our knowledge, none of the existing studies have analysed the problem of how to infer the network structure in Network DEA models.

On the other hand, the structure and function of complex networks is widely studied in the statistical mechanics. See e.g. the classical reviews by Albert and Barabasi (2002) and Newman (2003) and the recent book by Barabasi (2016).

The main aim of this paper is to fill this gap. We propose a statistical-based approach to infer (reconstruct) the network's topology for nonparametric productivity frontier models. The approach is developed in a Bayesian framework and relies on some recent pseudo-likelihood techniques introduced in the physics of complex systems (Ravikumar et al. 2010, Aurell and Ekeberg 2012, Tyagi et al. 2016; Marruzzo et al. 2016). Our approach is *Semi-Parametric* because it proposes an inferential approach based on a *Parametric* Bayesian approach (*Generalized* multicomponent spin model) to make inference for *Nonparametric* Productivity (DEA) Networks.

Indeed, recent developments in econometrics of information (Golan, 2008; Judge and Mittelhammer, 2011), statistical inference and machine learning (Barber, 2012; Murphy, 2012) allow us a new use of general production processes representation such as the Georgescu Roegen's Flows and Funds production model (GRFF model hereafter).

We illustrate the approach with an application in the knowledge production area (Fukuyama et al. 2016, Weber, 2017). We estimate the interaction of disciplinary productivity at the world-country level, by using data extracted from the Scopus database.[2]

About the estimation of the productivity of scientific knowledge, Daraio (2018) shows that the complexity of research productivity and the expansion of network in economics imply the search for new and more general models to represent the production process.

The economic model underlying this paper will be described in the next section. It combines Input-Output analysis with network DEA and accounts for the organizational aspects of production by showing that network DEA axiomatics represents an implementation of GRFF model.

The paper unfolds as follows. In the next section we illustrate the main features of the economic model. Section 3 shows the connection of the statistical approach proposed with the GRFF model and after that illustrates it. Section 4 describes the data, some descriptive analysis of the productivity models we estimated and the main results of the application in

[2]Scopus is a bibliometric database owned by Elsevier. For more information, see *https* : *//www.elsevier.com/data/assets/pdf_file/*0007/69451/0597−*Scopus*−*Content*−*Coverage*−*Guide*−*US*−*LETTER*−*v*4−*HI*−*singles*−*no*−*ticks.pdf*, last accessed 20^{th} *July*, 2018.

the knowledge production area. The final section concludes the paper.

2 The economic model

In this paper we propose a general economic model that combines Input-Output analysis with network DEA axiomatics by accounting for the organizational aspects of production according to the Georgescu-Roegen's model of flows and funds (see Vittucci Marzetti (2012) for a critical survey), as detailed by Morroni (1992 and 2006) and Fioretti (2007).

The axiomatic production theory beyond this paper, that will be described in Section 2.1, may be found in Färe (1988), Färe and Grosskopf (1996), Shephard (1970) as extended to Network DEA by Färe and Grosskopf (2000), the first introduction of the concept of Network DEA in the literature.

Section 2.2 highlights the correspondence of the axiomatics of NDEA with the representation of the production process according to Georgescu Roegen.

In light of this, we can see all the different productivity network models (see the encyclopedic summary in Kao, 2017) as *implementation* of the Georgescu Roegen's Flows and Funds model. To the best of our knowledge, in the network DEA literature, this connection has not been noted up to now.

We end up with a new more general framework for modeling the production process which includes the representation of the production process based on GRFF model, estimation of production functions, information theoretic approaches to econometrics, machine learning and statistical inference from the physics of complex systems.

In this paper, we address a problem that has not been addressed so far. It is about the estimation, within an inferential setting, of the network structure of a complex Input-Output model which includes DEA network models that in turn implement the GRFF model.

As described by Prieto and Zofio (2007), Network DEA within an Input-Output model (for an introduction and a deep overview see Miller and Blair, 2009) allows us to gauge potential productivity gains by comparing technologies corresponding to different "economies". Input-output models represent a network where different sectoral nodes use primary inputs (endowments) to produce intermediate input and outputs (according to sectoral technologies).

In graph theory terms, in an Input-Output model, each sector (industry) is represented by a node and each flow of intermediate inputs and outputs is represented by a link.

Within an input-output model hence, it is possible to optimize primary inputs allocation, intermediate production and final production by means of Network DEA (NDEA). NDEA allows us to model the different sub-technologies corresponding to alternative production

processes, to assess efficient resource allocation among them, and to determine potential output gains if inefficiencies were dealt with.

Statistical inference in this setting, allows us to estimate the *chains* or *paths connections* with and between nodes. This is helpful to reveal the underlying *characteristic structure* of an input-output system (see Miller and Blair, 2009, p. 675).

The economic model we propose to adopt is then based on underlying multi-stage technologies that are embraced in an input-output system. In our knowledge production application (see Section 4), the sub technologies are the scientific disciplines, whose productivity/efficiency will be assessed with an appropriate Network DEA. Altogether, the interdependencies of productivity of scientific disciplines, will be investigated and the underlying network structure inferred.

By adopting the Georgescu Roegen's flows and funds productive model, we are able to build a bridge between the axiomatics of Network DEA and recent estimation techniques proposed in the field of complex system physics.

Georgescu-Roegen's 1971 book: *Entropy Law and the Economic Process*, describes the production process as a process governed by the rules of physics and in particular by the second law of thermodynamics (the Entropy law) and constitutes the justification of the statistical model we apply in this paper to "infer" the network structure between productivity networks estimated using nonparametric Network DEA techniques. Georgescu-Roegen in the *Preface* of his book (p.xiii) *The Entropy Law and the Economic Process* states that: "The thought that the economic process, too, must be intimately connected with the Entropy Law [*n.o.w.*, not the mechanistic dogma of classical physics] is the origin of the inquiry that forms the object of this book".

In the following we describe the axiomatics of NDEA and afterwards highlight its connection with GRFF model through the works of Morroni (1992, 2006 and 2014).

In Section 3.2 we introduce the entropy concept according to Georgescu Roegen (1971) and explain our inferential approach.

2.1 Axiomatics of DEA Network Models

2.1.1 Basic axioms

Network DEA models are a generalization of the basic DEA or activity analysis models of technology and efficiency; they are often referred to as looking inside the black box technology assumed in static DEA efficiency models. They provide a means of describing more complex production processes in the DEA framework, including intertemporal models. We follow Färe and Grosskopf (1996), hereafter FG (1996), and show that the axiomatic underpinnings are

similar to those of the static DEA technology. The notation is the following: inputs are $x \in \Re_+^N$, outputs are $y \in \Re_+^M$ and are used to represent technology in terms of the Graph of technology $GR = \{(x, y) : x \text{ can produce } y\}$, the Input Set $L(y) = \{x : (x, y) \in GR\}$, and the Output Set $P(x) = \{y : (x, y) \in GR\}$.

The activity analysis models have the common feature of having a set of linear constraints used to construct the so-called piecewise linear frontier of technology, in accordance with the basic axioms of production theory (listed below). Assume that there are $k = 1, \ldots, K$ activities, each with input output vector $(x^k, y^k) = (x_{k1}, \ldots, x_{kN}, y_{k1}, \ldots, y_{kM})$. There are then K Decision Making Units (DMUs). Kemeny, Morgenstern and Thompson (1956) proposed the following conditions on these 'coefficients', namely

$$\begin{array}{lll} i) & \sum_{m=1}^{M} y_{km} > 0, k = 1, \ldots, K & \text{each DMU produces some output} \\ ii) & \sum_{k=1}^{K} y_{km} > 0, m = 1, \ldots, M & \text{each output is produced by some DMU} \\ iii) & \sum_{k=1}^{K} x_{kn} > 0, n = 1, \ldots, N & \text{each input is used by some DMU} \\ iv) & \sum_{n=1}^{N} x_{kn} > 0, k = 1, \ldots, K & \text{each DMU uses some input.} \end{array} \tag{2.1}$$

If these are satisfied, then FG (1996) show that the basic activity analysis model:

$$\begin{aligned} P(x) = \{y : \quad & y_m \leqq \textstyle\sum_{k=1}^{K} z_k y_{km}, m = 1, \ldots, M, \\ & \textstyle\sum_{k=1}^{K} z_k x_{kn} \leqq x_n, n = 1, \ldots, N, \\ & z_k \geqq 0, k = 1, \ldots, K\} \end{aligned} \tag{2.2}$$

satisfies the axiom set below.

The basic axioms include

A.1 $0 \in P(x), \forall x \in \Re_+^N, y \notin P(0), y \geq 0.$

A.2 $x \in L(y), \lambda \geqq 1 \Rightarrow \lambda x \in L(y)$ (weak disposability of input).

A.2S $x \geqq x^o \in L(y) \Rightarrow x \in L(y)$ (strong disposability of input).

A.3 $y \in P(x), 0 \leqq \theta \leqq 1 \Rightarrow \theta y \in P(x)$ (weak disposability of output).

A.32S $y \leqq y^o \in P(x) \Rightarrow y \in P(x)$ (strong disposability of output).

A.4 $\forall x \in \Re_+^N, P(x)$ is bounded.

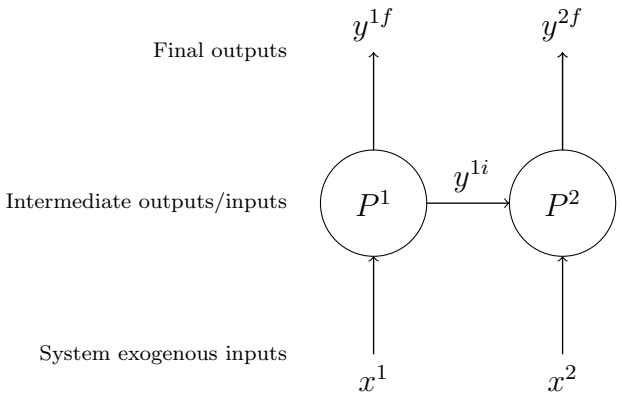

Figure 1: A Network Technology

A.5 The graph is a closed set.

In addition, it is often convenient to assume convexity of the input and output sets. The general result is that if each subtechnology in the network satisfies the Kemeny et al (1956) conditions then the network satisfies the axioms above. Similarly, if each subtechnology exhibits constant returns to scale, then the network also exhibits constant returns. We note that this holds for directed networks.

2.1.2 What makes a network a network?

We introduce the basic notions of networks by supposing that we have two subtechnologies, P^1 and P^2, where P^1 uses inputs x^1 to produce both a final output, y^{1f}, and an intermediate output y^{1i}. This intermediate output, together with x^2, serves as an input into the second subtechnology P^2 which produces a final output y^2. This simple network can be illustrated as in Figure 1.

Note that this is an example of the additional structure that can be put inside what we think of as the black box of technology.

This may also be given a multi–equation formulation, namely[3]

$$\begin{aligned} y^1 &= F^1(x^1) \\ y^2 &= F^2(y^1, x^2), \end{aligned} \tag{2.3}$$

where y^1, y^2 are non-negative scalars.

[3]This model is a special case of the two stage model in Danø (1966, p. 150).

This and other multi–equation models may be reformulated as a single equation model

$$y^2 = F^2(F^1(x^1), x^2) \tag{2.4}$$

or

$$F(x^1, x^2, y^2) = y^2 - F^2(F^1(x^1), x^2) = 0, \tag{2.5}$$

which we may think of as a standard transformation function, see Førsund (2018, p. 14).

Next we introduce a generic network model below due to Shephard and Färe (1975) which illustrates several types of networks. See also Danø (1966).

Assume that there are three sub-technologies

$$P^1, P^2, P^3$$

organized as in Figure 2, where outputs from P^1 and P^2 enter P^3 as inputs.

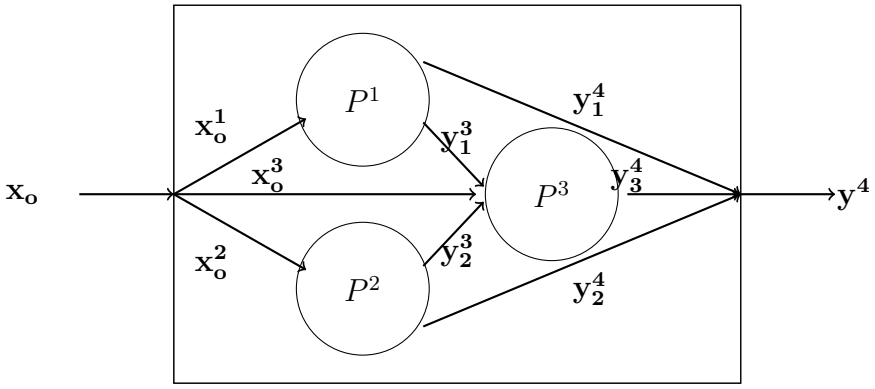

Figure 2: Three node network with input source and output sink

We extend this model to include a source ,'o', which distributes inputs to the network and a sink,'4', which collects the network outputs. The notation identifies the source of the variable with a subscript and the destination with a superscript. So $x_o^i, i = 1, 2, 3$ means that input x_o is distributed to the three subtechnologies, and we have

$$x_o \geqq x_o^1 + x_0^2 + x_o^3.$$

Similar notation is used for outputs, so y_2^3 means that outputs from P^2 are inputs into P^3. The final output is

$$y = y_1^4 + y_2^4 + y_3^4,$$

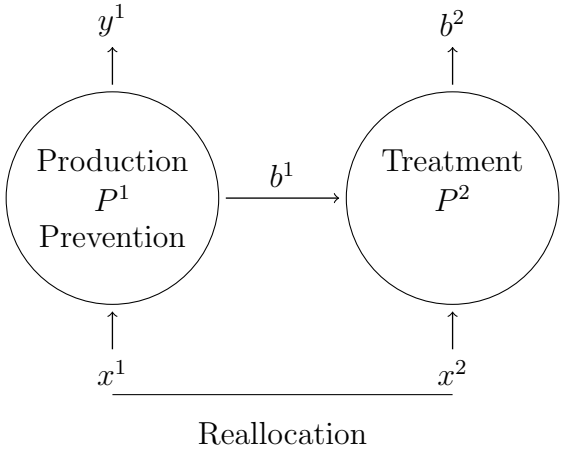

Figure 3: The Production Network Technology for Prevention and Treatment

with the appropriate choice of dimension of the output vectors. This schematic includes the possibility of parallel subtechnologies or processes like P^1 and P^2, as well as sequential sub-technologies which could be linked through time providing a basis for a dynamic network or supply chain.

One can also introduce a budget or reallocation constraint to the network. An example is illustrated in Figure 3 which models the choice between prevention and abatement in a production model with good and bad outputs. Here, some inputs (whose total is given, ie., $x^1 + x^2 \leqq \overline{x}$) may be allocated to either or both subtechnologies, with the goal of optimizing the net outputs of the network, namely y^1 and b^2.

2.2 Connection with Georgescu Roegen's Flows and Funds model

The schematic representation of network DEA just described above and the possibility of including parallel sub-technologies as well as sequential sub-technologies may be directly linked to the GRFF model.

Network DEA modelling may be conceived as an implementation of different sub-production processes in agreement with Georgescu Roegen's representation of the production process.

Generally (see e.g. Morroni, 1992), the Georgescu Roegen's flows and funds model takes into account the actual characteristics of production elements and processes, such as, indivisibility, complementarity, tacitness and heterogeneity of productive knowledge. It brings to the forefront the analysis of the sequence of operations performed and the time dimension of production processes in relation to specific organisational settings.

A *flow* is an input or an output that enters or exits from a process (for example, energy, water, software, loom, computer, etc.). A *fund* provides its services in several processes that

occur over time (for example, worker, software, land, loom, computer, etc.). A distinction is made between the agents of production processes and the services that they provide.

Activities consist of different operations which require the performance of one or more elementary tasks. An elementary task is an operation which, by definition, is not further divisible (for instance, loading or unloading an intermediate product or cutting a piece of fabric).

GRFF model allows the *analytical* representation of the *organisation* of production processes. For instance, we may analyse the differences between artisan and industrial production or within industrial production between various degrees of division of labour. See Figure 4 taken from Morroni (2014). The figure shows the organizational aspects of production inside the black box. It shows that the GRFF analytical description of an *organised elementary process* (Morroni, 1992, p. 54) is the basic component (unit) of NDEA different models (see e.g. Fig. 11.5 (Kao, 2017, p. 252) and Fig. 12.4 (Kao, 2017, p. 284)).

The schematic representation of network production technologies described in the previous section (see Figure 2 and Eq. 2.4 and 2.5), as a matter of fact, is based on this basic analytical description of the GRFF model. It therefore connects the NDEA to GRFF model.

GRFF model can be implemented both at the *microeconomic* level considering individual case studies, and at the *macroeconomic* level, analysing a set of production units in different sectors of activity.

Morroni (2006) shows that *capabilities*, *transactions* and *scale-scope* considerations are not rivals, but that they interplay in explaining the *boundaries* and the competitiveness of the firm. Under the conditions faced by the DMUs that are, radical uncertainty, complementarity between inputs, indivisibility of inputs, three-dimensionality of space, and set-up processes, the *weight* and the *interplay* of the three aspects is significant. A Firm's competitiveness is linked to its ability of coordinating the development of capabilities, the arrangement of transactions and the design of the scale of production. By including the dynamics of the variables which compose the production function Fioretti (2007) operationalizes the organizational aspects of production typical of the GRFF model and links them to recent developments in neural network modeling.

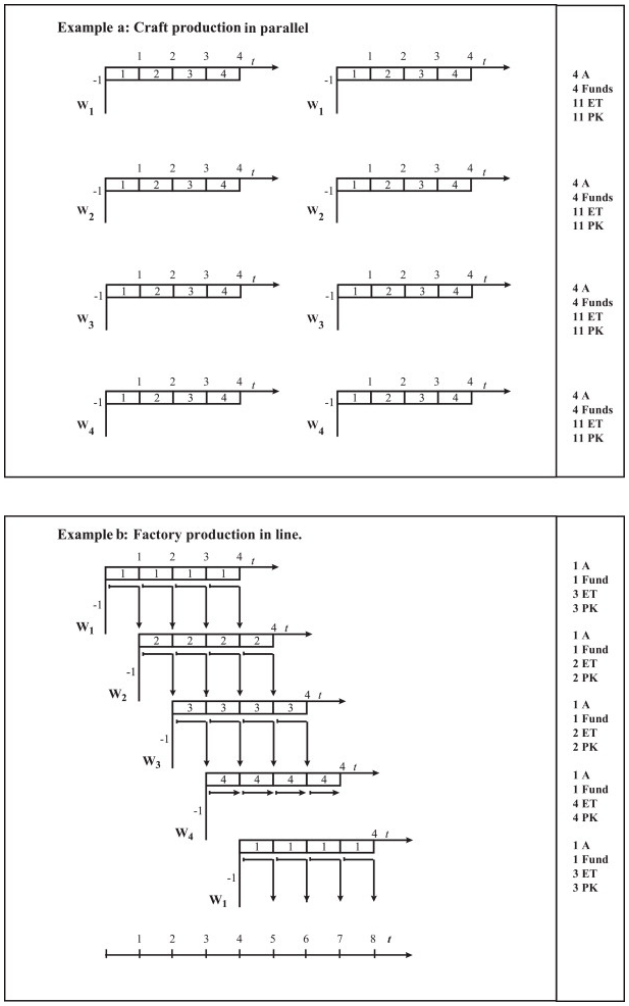

Figure 4: *Some examples of organization of the production process according to the Georgescu Roegen's flows and funds model. Source: Morroni (2014, p. 10 Figure 3).*

Summing up, the GRFF model allows an *analytical* representation of the *organization* of the production process that goes beyond the relationship between the inputs and the outputs, that is typical of production functions and input-output analyses. GRFF model allows deepening the analytical power of input-output models by including the *organization* and *time* dimension of production processes (Morroni, 1992). The formulation of the production network technology presented in the previous section is an implementation of the GRFF model. The GRFF model may be connected to the neo-Schumpeterian interpretative framework, of production of new processes by means of creation and diffusion of knowledge (Morroni, 2014) in which there is an interplay between capabilities, transactions and scale and scope to explain the boundary and the competitiveness of the analysed units (Morroni, 2006).

3 The statistical model

3.1 Introduction to Maximum Entropy

The principle of the Maximum Entropy is a cornerstone of complex dynamic systems analysis, statistical mechanics and constitutes a bridge between Complexity, Inference and Information theory. In this section we summarize some basic notions that can be skipped by the reader already familiar with the principle.

Real systems evolve spontaneously and possess stability characteristics at the equilibrium. The equilibrium is characterized by a maximum value of entropy. The principle of maximum entropy, which generalizes the statistical mechanics, is the foundation of the theory of inference (Jaynes, 1957) that is the reconstruction of probabilistic information from incomplete data. Maximizing the Entropy means taking into consideration all possible choices, without excluding any of them. On the contrary, the unjustified limitation of the field of choice is equivalent to the imposition of an arbitrary constraint. The key to the application of the principle (see Jaynes, 1957) is assigning to a probability density function (pdf) an entropy function that measures the *dispersion* or *uncertainty* with which the occurrence of possible events are expected. The principle of Maximum Entropy allows us to introduce some constraints, on the base of our knowledge on the system, that can be treated with the formalism of Lagrange multipliers (see Section 3.4 for more details). All in all, it is a matter of *re-derive* the form of a pdf from partial information, that is, from a finite number, usually small, of moments of the distribution itself.

"Entropy is the most influential concept to arise from statistical Mechanics (Sethna, 2006)". It was originally understood as a thermodynamic property of heat engines that inexorably increases with time. Entropy has become science's fundamental measure of disorder and information, quantifying everything from compressing pictures on the internet to the heat death of the Universe. According to Sethna (2006, p. 77), Entropy may be interpreted as a measure of:

- *disorder* in a system;
- our *ignorance* about a system;
- the *irreversible changes* in a system.

The most general interpretation of Entropy is as a measure of our *ignorance* about a system. In information theory it is referred as *uncertainty* which could be misleading for quantum mechanics because Heisenberg uncertainty has no associated entropy. Sethna (2006, p. 85) explains that the equilibrium state of a system maximizes the entropy because we have lost all information about the initial conditions except for the conserved quantities; maximizing the entropy maximizes our ignorance about the details of the system. The *second law of*

thermodynamics tells us that entropy increases presupposing some definition of entropy for systems that are *out of equilibrium*. In general we may describe our partial knowledge about such a system as a probability distribution, defining an *ensemble* of states.

Building upon Hayek (1967), Hinterberger (1994) argues that the economy is a complex system that is the outcome of uncoordinated individual behavior. In such systems, equilibria are modeled by the concept of attractors. As such, it becomes impossible to fully understand macro processes by examining individual behavior. Sergeev (2005) also indicates that, although representatives of the Austrian School had a skeptical regard to the use of mathematical tools in economics, their ideas can be expressed through the unit of statistical thermodynamics or the theory of information (see also Vozna 2016).

3.2 Entropy in Georgescu Roegen (1971)

In this section we report some introductory definitions taken from Georgescu Roegen (1971)'s book.

"According to Classical thermodynamics, energy consists of two qualities: (1) free or available and (2) bound or latent. Free energy is that energy which can be transformed into mechanical work. Like heat, free energy always dissipates by itself (and without any loss) into latent energy. The material universe, therefore, continuously undergoes a qualitative change, actually a qualitative degradation of energy. The final outcome is a state where all energy is latent, the Heat Death as it was called in the earliest thermodynamic theory. For some technical reasons, entropy was defined by the formula

$$Entropy = \frac{\text{Bound Energy}}{\text{Absolute Temperature}}, \Delta S = \Delta Q/T, \tag{3.6}$$

where ΔS is the entropy increment, ΔQ the increment of the heat transferred from the hotter to the colder body, and T the absolute temperature at which the transfer is made"(Georgescu Roegen, 1971, Ch. IV, pp. 129-130).

In chapter VI of his 1971 book, Georgescu Roegen introduces the changes in physics and the introduction of the new thermodynamics that is *statistical mechanics*. He highlights the second law of thermodynamics, the Entropy law, as an evolutionary law and describes the two laws of thermodynamics according to Clausius 1865 formulation:

- *The energy of the universe remains constant;*
- *The entropy of the universe at all times moves toward a maximum.*

"In nature there is a constant tendency for order to turn into disorder.

Disorder, then, continuously increases: the universe thus tends toward chaos, a far more forbidding picture than the Heat Death. Within this theoretical framework, it is natural that entropy should have been redefined as a measure of the degree of disorder.

But as some philosophers and physicists alike have pointed out, disorder is a highly relative, if not wholly improper, concept: something is in disorder only in respect to some objective, nay, purpose. A heap of books, for instance, may be in perfect order for the billing clerks but not for the cataloguing department of a library. The idea of disorder arises in our minds every time we find an order that does not fit the particular purpose we have at that moment. From the viewpoint advocated in this book, we associate the random order with disorder because it does not correspond to the analytical order we expect to find in nature. Nature is ordered only to the extent to which its modes of being can be grasped analytically, by our understanding. All the less can we see how disorder can be ordinally measurable.

Statistical mechanics circumvents the difficulty by means of two basic principles:

A. *The disorder of a microstate is ordinally measured by that of the corresponding macrostate.*

B. *The disorder of a macrostate is proportional to the number of the corresponding microstates.*

A microstate is a state the description of which requires that each individual involved be named.

In general, if there are m states and N particles, the measure of the disorder of the macrostate $(N_1, N_2, ..., N_m)$, $\sum_{i=1}^{N} N_i = N$, is given by the familiar formula of combinatorial calculus

$$W = \frac{N!}{N_1!N_2!...N_m!}. \tag{3.7}$$

Boltzmann's epoch-making formula for entropy viewed as a measure of disorder is

$$Entropy = S = k_B \ln W, \tag{3.8}$$

where $k_B = 1.38\dot{1}0^{-16}$ ergs per degree of temperature is a physical constant known as Boltzmann's constant.

For large values of N_i, with the aid of Stirling's asymptotic formula, Eq. 3.7 becomes:

$$\ln W = -\sum_{i=1}^{N} N_i \ln(N_i/N). \tag{3.9}$$

Putting $fi = N_i/N$, we can write Eq. 3.8 as follows:

$$S = -k_B N H, \tag{3.10}$$

where

$$H = \sum_{i=1}^{N} f_i \ln f_i, \tag{3.11}$$

is the famous H-function used by Boltzmann in his statistical approach to thermodynamics.

Clearly, $-k_B H$ represents the average entropy per particle. Note that H and S vary in opposite directions" (Georgescu Roegen, 1971, pagg. 142-145).[4]

3.3 The Ising spin glass model

The Ising spin glass model is made up of a lattice, where each node of the lattice is associated with a vector variable $\mathbf{s}_i$ at the $i-th$ site in the $D-$dimensional space, that represents the spin of a particle. See Figure 5 for an illustration.

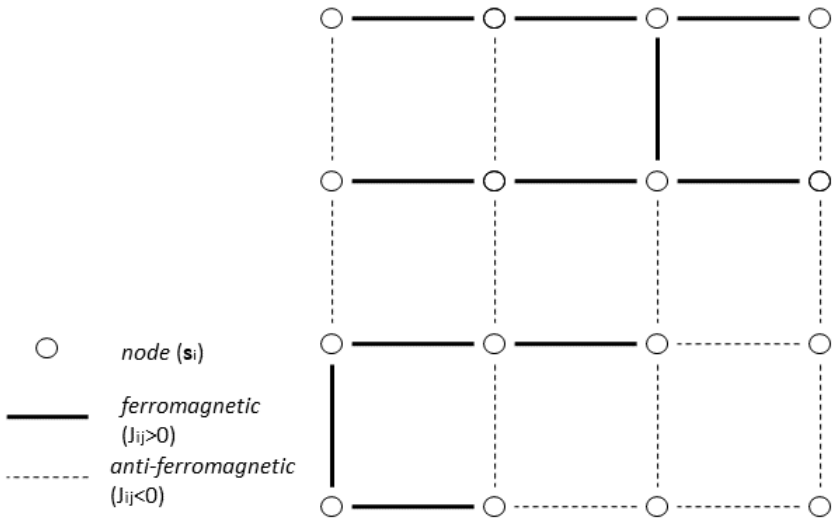

Figure 5: *Illustration of an Ising Model.* $J_{i,j} > 0$ *correspond to ferromagnetic couplings;* $J_{i,j} < 0$ *correspond to anti-ferromagnetic couplings.*

To set up a simple model able to describe the stationary states of the set of variables, we can associate to each couple of units i and j an 'energy' given by

$$J_{ij}\mathbf{s}_i \cdot \mathbf{s}_j, \tag{3.12}$$

where the symbol "$\cdot$" states for a scalar product and $\mathbf{s}_i \cdot \mathbf{s}_j = \sum_{\gamma=1}^{D} s_i(\gamma)s_j(\gamma)$.

If the system is in equilibrium at a given 'temperature' T, then the energy distribution of the units follows the Boltzmann law given by

$$F(E) = \frac{1}{Z}e^{-E/k_B T}, \tag{3.13}$$

where Z, the so called *partition function*, is given by $Z = \sum_{\{s\}} e^{-E(\{s\})/k_B T}$, where $\{s\}$ refers to a possible configuration and the sum is extended to all possible configurations, $e^{-E/k_B T}$ is called *Boltzmann factor* and k_B is *Boltzmann's constant.*

[4] A further discussion on the H-function is reported in Georgescu Reagan (1971) at pag. 149-150 and in Appendix B.

This assumes the ergodicity of the system, that is, the time average of functions on these random variables equals the average of these same functions over their probability distributions. This hypothesis is assumed in many processes that involve human systems, including econometrics of time series.

The whole system is therefore described by a Hamiltonian, a kind of total social energy, formed by two terms: a first term representing the coupling of the spin with an external magnetic field h_i and a second term representing the reciprocal interaction between all the nodes of the lattice. The parameters of the system are the intensity of the external magnetic field h_i, the intensity of the J_{ij} interaction energy and the temperature (β is the inverse of the temperature). In formula:

$$H = -\frac{1}{2}\beta \sum_{i,j=1}^{N} J_{ij}\mathbf{s}_i(t) \cdot \mathbf{s}_j(t) - \sum_{i=1}^{N} \mathbf{s}_i(t) \cdot \mathbf{h}_i \tag{3.14}$$

The J_{ij} in physics measure a direct and reciprocal effect (the interaction) of one entity on another entity. The nature and the strength of this effect can be measured by applying tools developed by the statistical physics of complex systems. The concept of interaction in physics can find its correspondence in the one of *interdependency* in Input-Output economic analysis. The latter means the existence of a mutual influence between sectors (industries).

Assuming the generalized multicomponent spin model holds, it allows us to estimate the *interaction parameters* J_{ij}, that are effective parameters embedding several effects. These *coupling* parameters J_{ij} generate the configurations of the system that may be characterized by the *correlations* between the spin variables, the so called *overlap* measures (see also Bongioanni, Daraio and Ruocco, 2014), defined as follows:

$$Q_{ij} = 1/T \sum_{t=1}^{T} \mathbf{s}_i(t) \cdot \mathbf{s}_j(t), \tag{3.15}$$

where $t = 1, ..., T$ is time.

As it is well known, a *correlation* measures the association between two variables. It shows a tendency of one variable to change with some regularity when the other changes, but this tendency may be moderated (influenced) by other factors, and depends on the whole configuration, including indirect effects. Correlation does not mean a direct effect or relation. On the other hand, *interactions* or *interdependencies* refer to *strict* relationships between variables which allow us to describe the *impact* of the variation of one variable on another. As already observed, interdependencies are also investigated in Input-Output economic models to analyse the strict interconnections between the different economic sectors. In physics the interdependencies correspond to the interactions between particles and are investigated to analyse direct and reciprocal effects between variables.

Assuming this model to make inference permits us to consider *beyond correlations* the interdependencies among the units of analysis. As we will see in the application (see Section 4), two disciplinary productivity's may be correlated because they tend to be associated in their variation, but they may not interact.

The Hamiltonian is obtained by a generalization of the Ising model, originally introduced to describe the behavior of ferromagnetic systems. In its more general formalism the Ising model can also account for a (parameter independent) weight, β, and external biases, h_i. When referred to magnetic systems, β is the inverse of temperature and h_i the magnetic external field. For sake of simplicity we fix here $\beta = 1$ and $h_i = 0$, $\forall i \in (1, N)$. $J_{ii} = 0$, $J_{ij} = J_{ji}$. Here we impose $J_{ij} \geq 0$ without loss of generality.[5]

The Ising model has been largely applied in different fields, such as modelling the behaviour of magnets in statistical physics (Brush, 1967), image processing and spatial statistics (Besag 1986, Geman 1984, Greig 1989), modelling of neural networks (Parisi, 1986) and social networks (Banerjee, 2008). The Ising model is a classical example of a graphical model in exponential form, to which the Boltzmann-Gibbs machine learning (or approximation method related to it) can be applied.

Given the Hamiltonian (H) related to this model, a positive interaction between two spin variables actually would lead to their convergence in order to satisfy the principle of minimum energy, when the system is in the *ground state* at $T = 0$.

However, since the Hamiltonian represents a disordered many-body system with pairwise interactions, competition can arise between them giving rise to so-called *frustration*, that is a spin blocked between two opposite configurations, which is not able to choose the pattern to follow grounding on the principle of minimum energy. The underlying hypothesis of our model is that a *positive interaction* would led to *alignment* between the spin variables, but *not vice versa*, that is the observation of alignment does not necessarily imply a positive interaction as well as the observation of misalignment does not exclude it. Our choice of the generalized multicomponent spin model is led by simplicity and because it guarantees the possibility to borrow the rigorous methodology developed by Boltzmann machine learning. It is supported by the Georgescu Roegen's book "The entropy law and the economic process". The aim of the present analysis is to derive the level and the structure of these interactions. It is an *inverse problem* because the inference of the interactions is drawn from a set of

[5]There is a link between the assumption of equilibrium underlying a Boltzmann-Gibbs distribution, and symmetry of the pairwise interactions. Symmetric couplings lead to a steady state described by the Boltzmann-Gibbs distribution while asymmetric ones to a non-equilibrium state (Krapivsky, 2010). We can assign to the system a particular dynamics, which leads it to a given steady state distribution. Recent developments achieved for dynamical inverse Ising model (Decelle, 2016, Nguyen, 2017) could represent an interesting extension of the present work, which is left for future research.

observed data. According to Judge and Mittelhammer (2011), inverse problems arise when one want to recover information on model parameters, i.e., coupling constants, by means of measurements of observable data. The solution of an inverse problem offers a connection between the data directly observed and the unknown information on model parameters. In an inverse problem, the model which generates the observed data is an input of the theory. It is important hence the support of Georgescu Roegen's (1971) book for the justification of the parametric assumptions underlying our statistical model.

3.4 Maximum Entropy estimates

The Shannon (1948) theorem stated the entropy defined in statistical mechanics as a measure of the 'amount of uncertainty' related to a given *discrete* [6] probability distribution. The latter quantity turned out to be (Shannon, 1948):

$$S[p] = -K \sum_{\{\mathbf{s}\}} p(\mathbf{s}) \log[p(\{\mathbf{s}\})], \tag{3.16}$$

where K is a positive constant and $p(\{\mathbf{s}\})$ is the probability distribution function (pdf) of the configuration $\{\mathbf{s}\}$. This quantity is positive, additive for independent sources of uncertainty and it agrees with the intuitive notions that a uniform (or broad) distribution represents more uncertainty than does a sharply peaked one. It is immediate to verify the latter observation in the one-dimensional case by considering Eq.3.16 and taking into account the property of the discrete distribution of probability, $p_i \leq 1$.

We recognize in Eq. 3.16 the entropy defined in statistical mechanics once the constant K is identified with the parameter β, the inverse of the temperature times the Boltzmann constant, k_B. In making inference on the basis of partial available information we must use that probability which maximizes the 'amount of uncertainty' or entropy subject to whatever is known (Jaynes, 1957). This permits to obtain an unbiased assignment, avoiding arbitrary assumption of information which by hypothesis we do not have (Jaynes, 1957). Since we know some empirical expectation values, formally this means that we find $p(\{\mathbf{s}\})$ as a solution of a constrained optimization problem, i.e. we maximize the entropy of the distribution subject to conditions that enforce the expectation values to coincide with the empirical ones. We will refer to the quantities whose averages are constrained as 'features' of the system, $\mathbf{f} \doteq \{f_1, ..., f_j, ...f_N\}$, where each feature is a function of the state $\{\mathbf{s}\}$, $f_j = f_j(\{\mathbf{s}\})$, $\forall j \in [1, N]$. By describing our system as an Ising model we observe the variables singly (interacting with a magnetic field) or in pairs. Because we furthermore imposed $\mathbf{h}_i = 0$, $\forall i$,

[6] Sethna (2006, p. 86 eq. 5.20 and ff.) presents the case of entropy to systems out of equilibrium with continuum distributions.

the features of the system we are considering are only the two-variable combinations. Our optimization problem thus reduces to

$$Max_{p(\{\mathbf{s}\})}S[p], \tag{3.17}$$

with the constraints

$$\sum_{\{\mathbf{s}\}} p(\{\mathbf{s}\}) = 1 \text{ and } < \mathbf{s}_i \cdot \mathbf{s}_j >_{p(\{\mathbf{s}\})} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{s}_i(t) \cdot \mathbf{s}_j(t), \tag{3.18}$$

where $< \cdot >$ means the average on the pdf $p(\{\mathbf{s}\})$, and the features are $f_{ij} = \mathbf{s}_i \cdot \mathbf{s}_j$. The solution of Eq. 3.17 with the constraints 3.18 can be solved by using the Lagrange multipliers λ_0, $\{\lambda_{ij}\}$. We obtain

$$p(\{\mathbf{s}\}) = ce^{-\lambda_0} e^{-\frac{1}{2}\sum_{i\neq j}\lambda_{ij}\mathbf{s}_i\cdot\mathbf{s}_j}, \tag{3.19}$$

by taking into account that $\mathbf{s}_i \cdot \mathbf{s}_j = \mathbf{s}_j \cdot \mathbf{s}_i$ The constant $ce^{-\lambda_0}$ can be determined by exploiting the constraint $\sum_{\{\mathbf{s}\}} p(\{\mathbf{s}\}) = 1$, obtaining $\sum_{\{\mathbf{s}\}} ce^{-\lambda_0} e^{-\sum_{i\neq j}\lambda_{ij}\mathbf{s}_i\cdot\mathbf{s}_j}$ and finally $ce^{-\lambda_0} = [\sum_{\{\mathbf{s}\}} e^{-\sum_{i\neq j}\lambda_{ij}\mathbf{s}_i\cdot\mathbf{s}_j}]^{-1}$. The quantity inside the brackets coincides with the partition function $Z(\{J\})$ defined in Eqs. 3.22 and 3.23, once we assume $J_{ij} = \lambda_{ij}$. The constants $\{\lambda_{ij}\}$ can be determined by the constraint: $< \mathbf{s}_i\cdot\mathbf{s}_j >_{p(\{\mathbf{s}\})} - \frac{1}{T}\sum_{t=1}^{T} \mathbf{s}_i(t)\cdot\mathbf{s}_j(t) = 0$. Because, as specified in Eq. 3.25 the quantity in the left side of the previous equation turns out to be the gradient of the (concave) Log-Likelihood function and such a quantity should be zero to satisfy the constraint, we can reformulate the problem of maximizing the entropy subject to constraint on the expectation values, as searching the maximum of the Log-Likelihood function within the class of models defined by the Boltzmann distribution (to which belongs Eq.3.19), as stated in the next section.

It is worth noting that if we know only the average energy, $E =< H >$, the maximum entropy problem subject to the constraint $E- < H >= 0$ is solved by the Boltzmann distribution.

We finally highlight the fact that in information theory the maximum entropy problem can be reformulated as a minimum cost of coding, which is actually a function defined as the opposite of the entropy (Shannon, 1948).

After having described the correspondence existing between maximum entropy and maximum likelihood estimation, we illustrate our proposed approach to make inference in the next section. Based on the maximization of the Shannon's entropy in Eq. 3.17, the technique that we will introduce in the next section, will allow us to recover the interactions or coupling parameters, J_{ij}, that are used afterwards to infer the network structure.

3.5 Maximum-Likelihood and Pseudo-Likelihood estimates

We consider the set of variables $\{\mathbf{s}\}$, whose i-th element is a vector of N components, $\mathbf{s}_i = (s_i^1, ..., s_i^N)$ and the set of data $\{\mathbf{s}(t)\}$, where t identifies a given realization of the set of variables $\{\mathbf{s}\}$ (t is related to a given time). We assume that each realization of the set $\{\mathbf{s}\}$ is drawn independently. We furthermore assume that the data have been generated by a (known) model, which depends on the set of (unknown) pairwise parameters $\{J\}$, with generic element J_{ij}. We aim to find the optimal value of these latter given the data. In other words, we want to find the values of J_{ij} that maximize the conditional probability $p(\{J\}|\{\mathbf{s}\})$. From the Bayes theorem (see e.g. Barber, 2012) it follows that

$$p(\{J\}|\{\mathbf{s}\}) = \frac{p(\{\mathbf{s}\}|\{J\})p(\{J\})}{p(\{\mathbf{s}\})} = \frac{p(\{\mathbf{s}\}|\{J\})p(\{J\})}{\int_{\{J\}} p(\{\mathbf{s}\}|\{J\})p(\{J\})}. \tag{3.20}$$

The probability $p(\{J\}|\{\mathbf{s}\})$ is called *posterior*, $p(\{J\})$ *prior*, $p(\{\mathbf{s}\})$ *evidence* and $p(\{\mathbf{s}\}|\{J\})$ *Likelihood*. If the prior is the uniform distribution, as we assume here, the most probable a posteriori set of variable is, as a consequence of Eq. 3.20, that one which maximises the Likelihood function.

We assume that the Likelihood function can be expressed in the functional form of the Boltzmann distribution, i.e.

$$p(\{\mathbf{s}\}|\{J\}) = \frac{1}{Z(\{J\})} e^{-H(\{\mathbf{s}\}|\{J\})}. \tag{3.21}$$

The function $Z(\{J\})$ whose presence ensures the correct normalization for the Boltzmann distribution, is

$$Z(\{J\}) = \sum_{\{\mathbf{s}\}} e^{-H(\{\mathbf{s}\}|\{J\})}, \tag{3.22}$$

the sum is extended to all possible configurations in the phase space of the set of variables $\{\mathbf{s}\}$. The Hamiltonian or cost function, defined by exploiting a generalization of the Ising model, is

$$H(\{\mathbf{s}\}|\{J\}) = -\frac{1}{2} \sum_{i,j=1}^{N} J_{i,j} \mathbf{s}_i \cdot \mathbf{s}_j, \tag{3.23}$$

Given the hypothesis of independence of the data set $\{\mathbf{s}(t)\}$ and the functional form of the Likelihood function, the Log-Likelihood function, $l(\{J\})$ is

$$l(\{J\}) = \log(L(\{J\}) = \sum_{t=1}^{T} -H(\{\mathbf{s}(t)\}|\{J\}) - T \log(Z(\{J\})), \tag{3.24}$$

where $t = 1, ..., T$. The inference problem consists in determining the set of parameters $\{J\}$ which maximizes the function in Eq. 3.24. The Log-Likelihood function is concave, thus

the maximum corresponds to the point of zero gradient. The gradient of the Log-Likelihood function with respect to the parameters of the set $\{J\}$ is given by

$$\frac{\partial}{\partial J_{ij}} l(\{J\}) = \frac{1}{2}\sum_{t=1}^{T} \mathbf{s}_i(t)\cdot\mathbf{s}_j(t) - T\frac{1}{Z(\{J\})}\frac{\partial}{\partial J_{ij}} Z(\{J\}) = \frac{1}{2}T\big[C_{ij} - < \mathbf{s}_1\mathbf{s}_j >_{\{J\}}\big], \tag{3.25}$$

where $C_{ij} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{s}_i(t)\cdot\mathbf{s}_j(t)$ is an empirical correlation function calculated on the observed data set, whereas it is

$$-\frac{1}{Z(\{J\})}\frac{\partial}{\partial J_{ij}} Z(\{J\}) = \frac{\sum_{\{\mathbf{s}\}} \frac{1}{2}\mathbf{s}_i\mathbf{s}_j e^{-H(\mathbf{s}|\{J\})}}{\sum_{\{\mathbf{s}\}} e^{-H(\mathbf{s}|\{J\})}} = = \frac{1}{2} < \mathbf{s}_i\cdot\mathbf{s}_j >_{\{J\}}, \tag{3.26}$$

where $< \ >_{\{J\}}$ states for ensemble average calculated by using the probability distribution obtained with the set of parameters $\{J\}$. When the ensemble average is calculated with the 'true' set of parameters, i.e. the ones whose associated distribution generated the data, in the limit $T \to \infty$, $C_{ij} \to < \mathbf{s}_i\cdot\mathbf{s}_j >$ and $\frac{\partial}{\partial J_{ij}} l(\{J\}) \to 0$. We thus can conclude that in this limit the maximum of $l(\{J\})$ is obtained for those values of $\{J\}$ which generates the correlation function $< \mathbf{s}_i\cdot\mathbf{s}_j >_{\{J\}}$ equal to the empirical one, C_{ij}, under the hypotheses i) of independence of the realizations of the variables $\mathbf{s}$ and ii) that the Likelihood function belongs to the class of the Boltzmann distribution function.

To solve the inverse problem described in Section 3.3 we have to maximize the Log-Likelihood with respect to the set of parameters J:

$$l(\{J\}) = log(L(\{J\})) = \sum_{t=1}^{T} -H(\{s\}_t|\{J\}) - T\log(Z(\{J\})). \tag{3.27}$$

This is computationally hard to solve and for this reason a pseudo-likelihood approach, which involves Bolzmann-Gibbs machine learning, is applied. The Pseudo-Log-Likelihood function is defined as the sum over all the nodes of the local conditional probability at a given node, i.e.

$$\lambda(\{J\}) = \sum_{\mu=1}^{M}\sum_{i=1}^{N} l'(\mathbf{s}_i^\mu|\{\mathbf{s}_{\setminus i}^\mu\}|\{J\}) \equiv \sum_{i=1}^{N} l'_i, \tag{3.28}$$

where $l'(\mathbf{s}_i|\{\mathbf{s}_{\setminus_i}\}|\{J\}) = \log[p(\mathbf{s}_i|\{\mathbf{s}_{\setminus i}\},\{J\})]$. The local conditional probability at the i-th node is

$$p(\mathbf{s}_i|\{\mathbf{s}_{\setminus i}\},\{J\}) = \frac{1}{Z_i(\{J\})} e^{-H_i(s_i|\{\mathbf{s}_{\setminus i}\},\{J\})}, \tag{3.29}$$

and the local partition function is $Z_i(\{J\}) = \sum_{\{\mathbf{s}_i\}} e^{-H_i(s_i|\{\mathbf{s}_{\setminus i}\},\{J\})}$. Further details about the Pseudo-Likelihood approach can be found in Daraio et al. (2018). The solution of the inverse problem consists in finding the optimal values of the set of parameters J, which are supposed to generate the observable set of data. Given a set of parameters J, a zero value of the parameter J_{ij} between the pair ij means that the two variables are not interacting, a positive value indicate a tendency to align towards the same pattern.

4 Application in Knowledge production

We consider our system as a *disordered* system at equilibrium, described by a generalized multicomponent spin model. To this aim, our variables $s_i^{(\gamma)}(t)$, which are the equivalent of the spin configurations, are defined as follows:

$$\begin{aligned} s_i^{(\gamma)}(t) &= \frac{\Delta_i^{(\gamma)}(t)}{\sqrt{\sum_{\gamma=1}^{D} \Delta_i^{(\gamma)}(t)^2}}; \quad \Delta_i^{(\gamma)}(t) = \pi_i^{(\gamma)}(t) - \overline{\pi}_i(t); \\ \overline{\pi}_i(t) &= \frac{1}{D}\sum_{\gamma=1}^{D} \pi_i^{(\gamma)}(t); \quad \gamma = 1, ..., D; i = 1, ..., N; \ t = 1, ..., T. \end{aligned} \tag{4.30}$$

where $\gamma = 1, ..., D$ are the countries, $\pi_i^{(\gamma)}$ is the productivity of country γ in a subject category i, for $i = 1, ..., N$, over the period $t = 1, ..., T$ (here 1996-2012). Finally, $\overline{\pi}_i(t)$ is the world average of productivity in subject i. They have the property that $\overline{s_i} = 0$ and $(\overline{s_i})^2 = \frac{1}{N}$. In this way we account for the recent trend of increasing scientific productivity all over the world. By the normalization reported above in fact we define as variables only instantaneous fluctuations of productivity around the world average in each given discipline at one data sample recording. Though the average scientific productivity is not constant over time the distribution of the deviations around the means can be considered as such. The Hamiltonian associated to such a system is $H = -\frac{1}{2}\sum_{i,j=1}^{N} J_{ij} \sum_{\gamma=1}^{D} s_i^{\gamma} s_j^{\gamma}$.

4.1 The network structure

Table 1 describes the main components of our model and the correspondence between the physics of complex systems and (economic) productivity analysis.

Figure 6 illustrates the structure of the network that we analyse in this application. This network models the interdependencies existing between disciplinary productivity/efficiency, where disciplines $i = 1, ..., N$ are the Scopus subject categories listed in Table 3.

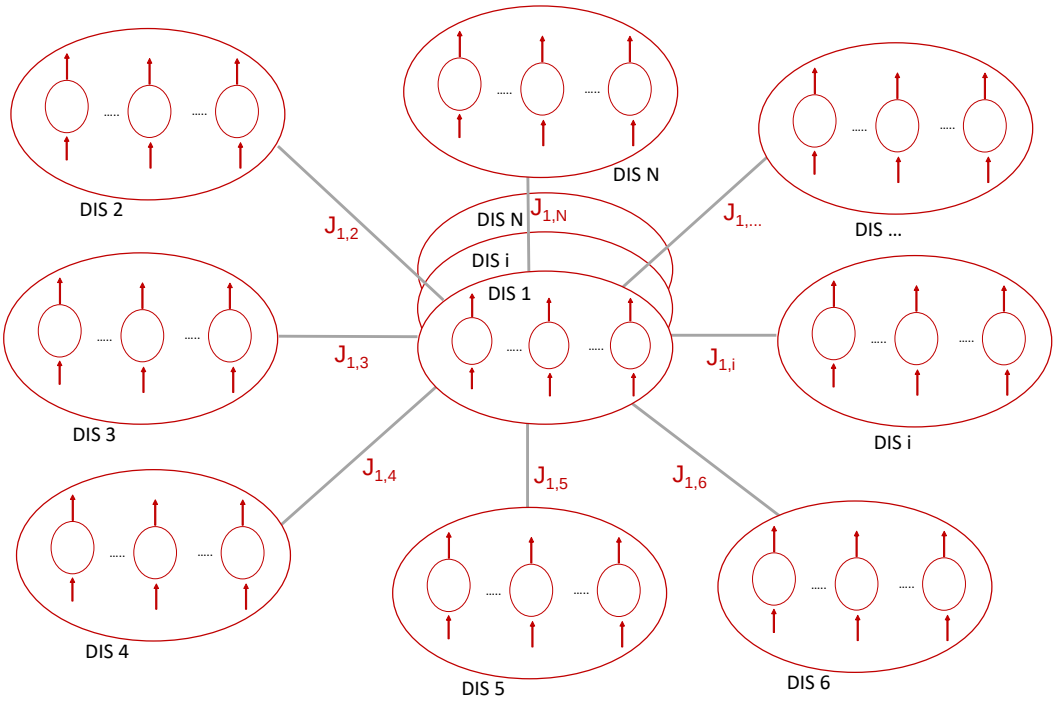

Figure 6: *An illustration of the network model. The J are the interrelations between disciplines (DIS in the figure) corresponding to the subject categories listed in Table 3.*

Table 1: *Model's components*

Statistical Physics	**Productivity Analysis**
Generalized Multicomponent spin model with arbitrary network	Network-DEA in an Input-Output model *interrelations*
node vectorial variable: spin $\mathbf{s}_i$	productivity of discipline i
node variable components: $\mathbf{s}_i = \{s_{i,\gamma}\}$	productivity of country γ in the discipline i
Node *interactions* or *couplings*: $J_{i,j}$	*interdep.* or *interrelations* between productivity/efficiency of disciplines
Hamiltonian, generalized cost function $H = -\frac{1}{2}\beta\sum_{i,j=1}^{N} J_{ij}\mathbf{s}_i(t)\cdot\mathbf{s}_j(t) - \sum_{i=1}^{N}\mathbf{s}_i(t)\cdot\mathbf{h}_i$	social energy to be minimized
β: inverse of the temperature	β external global parameter
$\mathbf{h}_i$: external magnetic field on i	contextual environmental variables of discipline i
J_{ii}: chemical potential of i	

4.2 Data and descriptive analysis

Data was extracted from the Scopus database and relate to the scientific production of world countries for 27 Scopus subject categories from 1996 to 2012.

Data problems in bibliometric studies are well known. A common way to reduce them is to analyse macro-level bibliometric data. According to Nederhof (1988) comparative analysis are more reliable when the unit of analysis is more aggregated because in a larger sample size, micro random errors mutually compensate. Another issue of concern is given by the changes in coverage from inclusion or exclusion of journals which affect more small countries with less publications. This may lead to unreliable values when a country only has a small number of scholarly outputs (see e.g. Schubert et al. 1989). To avoid this problem, we consider the 54 most productive countries (in terms of scientific productivity), which account for more than 95% of the world scientific production in the considered period.

Table 2 gives an overview of all the indicators available for the present study. The list of the Scopus 27 subject categories is reported in Table 3.

Luwell (2004) and Aksnes et al. (2017) report the main issues related to the integration of bibliometric data with inputs data in particular R&D expenditures. Methodological problems in measuring productivity at the macro level are mainly due to the fact that countries do not always standardise measurements of resources and outcomes. Moreover, the methodologies for collecting input and output data have been developed largely independently from each other. In this paper we use an input measure NA, the number of publishing authors which is coherent with the output given by P (number of publications) and HCP (number of articles in top 10% of most highly cited articles). The data on NA have been provided as the other data within the framework of an Elsevier Bibliometric Research Project (EBRP).

In the elaborations carried out to estimate the interactions (J_{ij}) and infer the network structure, to increase the number of available data we transformed yearly data into weekly data by means of a linear interpolation. The final number of observations considered is 833 and refers to weekly production of the number of articles and highly cited publications over the period 1996-2012.

Table 2: *List of Indicators*

Ind	**Description**
P	Number of articles (integer count)
P_f	Number of articles (fractional count, based on authors affiliations)
C	Total citations (4 years window, i.e., for articles in 2006, citations are from 2006-2009)
CPP	Total citations per paper (4 years window, i.e., for articles in 2006, citations from 2006-2009)
HCP	Number of articles in top 10 per cent of most highly cited articles in a discipline
$PINT$	Number of internationally co-authored papers
$PNAT$	Number of nationally (but not internationally) co-authored papers
$PINST$	Number of papers co-authored by members of different institutions within a country
PSA	Number of non-collaborative (single address) papers
NA	Number of publishing authors

Table 3: *List of the 27 Scopus' subject categories*

asjc code	**Subject Category**	**Description**
10	GENE	General
11	**AGRI**	Agricultural and Biological Sciences
12	ARTS	Arts and Humanities
13	**BIOC**	Biochemistry, Genetics and Molecular Biology
14	BUSI	Business, Management and Accounting
15	**CENG**	Chemical Engineering
16	**CHEM**	Chemistry
17	**COMP**	Computer Science
18	DECI	Decision Sciences
19	**EART**	Earth and Planetary Sciences
20	ECON	Economics, Econometrics and Finance
21	**ENER**	Energy
22	**ENGI**	Engineering
23	**ENVI**	Environmental Science
24	**IMMU**	Immunology and Microbiology
25	**MATE**	Materials Science
26	**MATH**	Mathematics
27	**MEDI**	Medicine
28	**NEUR**	Neuroscience
29	NURS	Nursing
30	**PHAR**	Pharmacology, Toxicology and Pharmaceutics
31	**PHYS**	Physics and Astronomy
32	PSYC	Psychology
33	SOCI	Social Sciences
34	VETE	Veterinary
35	DENT	Dentistry
36	HEAL	Health Professions

Table 3 reports in bold the 16 subject categories considered in the analysis. We excluded social sciences and humanity disciplines that present zeros and whose coverage of their scholarly outputs in indexed journals is much lower than the subject categories considered.

Table 4 shows some descriptive statistics for selected disciplines, namely: BIOC, COMP, ENGI, MEDI and PHYS.

Table 4: Summary Statistics Performance Variables, by Discipline (ASJC code) (54 Countries, 1996 - 2012)

BIOC (13)	Mean	Std. Dev.	Min	Max
Articles	4,521.5	9,902.0	60	85,295
Fractionalized	3,471.5	8,005.4	32	66,662
Highly Cited	552.5	1,611.3	0	15,480
Number of Authors	11,088.3	24,859.7	139	229,139
COMP (17)				
Articles	2,682.7	7,255.1	7	80,276
Fractionalized	2,245.1	6,436.3	4	75,956
Highly Cited	349.6	1,005.4	0	10,123
Number of Authors	4,092.3	11,452.4	11	128,273
ENGI (22)				
Articles	5,345.8	13,962.6	31	156,349
Fractionalized	4,552.3	12,614.8	21	149,094
Highly Cited	727.3	1,790.0	0	19,830
Number of Authors	8,832.7	23,263.8	54	293,605
MEDI (27)				
Articles	7,559.8	17,035.5	95	165,181
Fractionalized	6,022.4	14,055.7	52	133,744
Highly Cited	1,012.7	2,904.2	0	28,743
Number of Authors	15,529.0	34,013.8	211	351,702
PHYS (31)				
Articles	4,325.9	8,210.4	19	58,244
Fractionalized	3,212.6	6,674.1	13	52,985
Highly Cited	550.8	1,199.1	0	10,591
Number of Authors	6,830.0	14,282.1	26	127,209

4.3 Models for Estimating Knowledge Production

Our network models borrow from Fukuyama, Weber and Xia (2016) and Weber (2017), while the production variable choices are modeled after Georgescu-Roegen, where we include both flow variables and funds variables. We think of knowledge production in a production axiomatic framework, in which we include author count as a flow type input variable, and cumulated previous own publications as a funds or knowledge stock variable, which produces a flow of publication outputs.

What makes it a network is the fact that the cumulated publications are available to others in other countries in the same discipline, and previous cumulated publications from other countries in that discipline also are available to the discipline in the 'home' country as fund-type input variables. These proxy the public good/externality nature of publications as well as their role in contributing to the stock of knowledge.

As (final) output, our base model (Model 1) uses current period own publications as output for each country within the discipline being investigated. This is our 'quantity of knowledge' model. We estimate four versions of Model 1, a basic model which only considers the stock of previous publications within each country (1.1), a network version which adds to that the stock of previous publications from other countries (1.2), and two fractionalized versions (1.1f,1.3), which replace the raw quantity of publications with quantity weighted by the number of authors on each paper. Model 2 we refer to as our 'quality of knowledge model' and instead of publication counts as outputs and funds variables, we use the count of publications that are in the top 10% of cited papers in the relevant discipline and country and time period, in static (2.1) and network form (2.2). See the table.

Table 5: Network DEA Models of Knowledge Production

	Models	
Variable	Quantity Model(1)	Quality Model (2)
flow input	Author count(NA)	Author count (NA)
fund input 1	Own prev pub (Basic) Mod 1.1	Own prev highly cited pub, Mod 2.1
fund input 2 (externality)	Other prev pub (NW) Mod 1.2	Other prev highly cited pub, Mod 2.2
Outputs	Own current pub count (P) Fractional current pub (F)	Own current highly cited pub (HCP)

We illustrate the network for the case of two countries, γ and γ', both for discipline i in period t. Their flow inputs are the author counts denoted by x and their final outputs are denoted by y. The previous publication fund variables that provide the network connection are denoted as z.

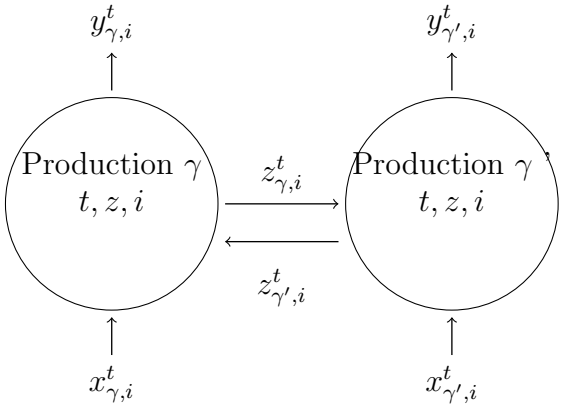

Figure 7: The Network Technology for Knowledge Production

4.4 Productivity estimated using NDEA

To fix ideas, and following Fukuyama, Weber and Xia (2016), let each country be indexed by $\gamma = 1, \ldots, D$, for periods $t = 1, \ldots, T$. We will augment their model by including disciplines $i = 1, \ldots, N$, here $N = 16$. Denote flow input as x^t_γ (here a scalar, but possibly a vector). Fund variables include own country cumulated past publications denoted as $z^t_{\gamma,i}$, and other country cumulated previous publications as $Y^t_{\gamma,i}$. Final output is own country current period publications denoted as $y^t_{\gamma,i}$.

Again following Fukuyama, et al, (2016) the fund variable inputs are defined as the sum of the previous 3 periods publications, which we follow here, i.e., $z^t_{\gamma,i} = \sum_{\tau=1}^{3} y^{t-\tau}_{\gamma,i}$,where $z^t_{\gamma,i}$ where $\tau = 1, 2, 3$ represents the sum of the previous 3 years publications of the home country. Similarly, spillover knowledge from other countries' previous publications is represented as $Y^t_{\gamma,i} = \sum_{\tau=1}^{3} \sum_{\gamma' \neq \gamma}^{D} y^{t-\tau}_{\gamma'}$.

In a DEA setup, the reference technology (output set) for period t, $P^t(x^t_{\gamma,i}, z^t_{\gamma,i}, Y^t_{\gamma,i}) = \{y : (x^t_{\gamma,i}, z^t_{\gamma,i}, Y^t_{\gamma,i})$ can produce $y\}$ may be written

$$
\begin{aligned}
P^t_{\gamma,i}(x^t_{\gamma,i}, z^t_{\gamma,i}, Y^t_{\gamma,i}) \quad &= \quad \{y^t : \sum_{\gamma=1}^{D} \lambda^t_{\gamma,i} y^t_{\gamma,i} \geqq y, \qquad (4.31)\\
\sum_{\gamma=1}^{D} \lambda^t_{\gamma,i} x^t_{\gamma,i} \leqq x^t_{\gamma,i}, \qquad & \sum_{\gamma=1}^{D} \lambda^t_{\gamma,i} z^t_{\gamma,i} \leqq z^t_{\gamma,i}\\
\sum_{\gamma=1}^{D} \lambda^t_{\gamma,i} Y^t_{\gamma,i} \leqq Y^t_{\gamma,i}, \qquad & \lambda^t_{\gamma,i} \geqq 0, \gamma = 1, \ldots, D; i = 1, \ldots, N; t = 1, \ldots, T.\}
\end{aligned}
$$

We can employ a Shephard output distance function as a scalar performance measure for each country in each period within a discipline relative to the technology specified above. Define the distance function for country γ (in discipline i) as

$$D^t_{\gamma,i}(x^t_{\gamma,i}, z^t_{\gamma,i}, Y^t_{\gamma,i}, y^t_{\gamma,i}) = \inf\{\theta : y^t_{\gamma,i}/\theta \in P^t_{\gamma,i}(x^t_{\gamma,i}, z^t_{\gamma,i}, Y^t_{\gamma,i})\}. \quad (4.32)$$

This function scales observed country output to the frontier of the output set $P^t_\gamma(.)$ and takes a value of unity for observations on the frontier.

For models 1.2 and 2.2, the productivity estimates $\pi_i^{(\gamma)}(t)$ are given by $D^t_{\gamma,i}$, while for the Basic model (whose results are reported in the following together with the results of the Models 1.2 and 2.2) $\pi_i^{(\gamma)}(t) = P/NA$, that is Number of articles/Number of publishing Authors.

4.5 Results

We estimate our two models for each of 54 countries using annual data from 1996-2012 for each of 16 disciplines reported in bold in Table 3. To summarize the results, we present the annual output share-weighted geometric means of the efficiency values by discipline, for the alternate versions of Models 1 and 2: Total Pubs (Basic and NW, 1.1 and 1.2), Fractionalized (Basic and NW, 1.1f and 1.3), Highly Cited (Basic and NW, 2.1 and 2.2).

For comparison, we add two alternative measures of performance: Pubs/Authors and Pubs/Authors (Fractionalized). These measures base performance on the ratio of own-country publications to number of authors, using raw and author-fractionalized quantities of publications, respectively. We present the annual output share-weighted geometric means of these in Table 6. We present the performance measure rank order correlations by discipline in Table 7.

Table 6: Geometric Mean Efficiency Results, by Discipline (ASJC code) (54 Countries, 1996 - 2012)

	COMP (17)	ENGI (22)	MEDI (27)	PHYS (31)
Model 1.1	0.995	0.995	0.996	0.992
Model 1.2	0.999	0.999	0.999	0.999
Model 1.1F	0.996	0.996	0.997	0.994
Model 1.3	0.999	0.998	0.999	0.998
Model 2.1	0.992	0.992	0.995	0.991
Model 2.2	0.998	0.998	0.999	0.998
P/A	0.992	0.991	0.987	0.992
P/A F	0.989	0.988	0.983	0.986

Table 7: Performance Measure Rank Correlations, by Discipline

COMP	P/A	P/A F	D11	D12	D11F	D13	D21	D22
P/A	1.000							
P/A F	0.824	1.000						
D11	0.532	0.476	1.000					
D12	0.389	0.455	0.635	1.000				
D11F	0.456	0.628	0.668	0.620	1.000			
D13	0.282	0.541	0.488	0.818	0.810	1.000		
D21	0.417	0.341	0.314	0.262	0.253	0.138	1.000	
D22	0.229	0.270	0.104	0.416	0.149	0.284	0.719	1.000
ENGI	P/A	P/A F	D11	D12	D11F	D13	D21	D22
P/A	1.000							
P/A F	0.765	1.000						
D11	0.545	0.550	1.000					
D12	0.437	0.556	0.646	1.000				
D11F	0.588	0.769	0.690	0.628	1.000			
D13	0.431	0.737	0.562	0.807	0.843	1.000		
D21	0.386	0.268	0.430	0.337	0.306	0.249	1.000	
D22	0.139	0.216	0.222	0.555	0.239	0.388	0.655	1.000
MEDI	P/A	P/A F	D11	D12	D11F	D13	D21	D22
P/A	1.000							
P/A F	0.823	1.000						
D11	0.605	0.523	1.000					
D12	0.548	0.618	0.805	1.000				
D11F	0.562	0.662	0.828	0.789	1.000			
D13	0.464	0.681	0.698	0.903	0.873	1.000		
D21	0.544	0.337	0.413	0.418	0.274	0.256	1.000	
D22	0.350	0.287	0.279	0.502	0.264	0.380	0.729	1.000
PHYS	P/A	P/A F	D11	D12	D11F	D13	D21	D22
P/A	1.000							
P/A F	0.380	1.000						
D11	0.536	0.422	1.000					
D12	0.165	0.369	0.371	1.000				
D11F	0.290	0.860	0.496	0.327	1.000			
D13	-0.063	0.723	0.191	0.655	0.708	1.000		
D21	0.390	0.148	0.382	0.235	0.117	0.042	1.000	
D22	-0.032	0.003	-0.033	0.452	0.013	0.256	0.619	1.000

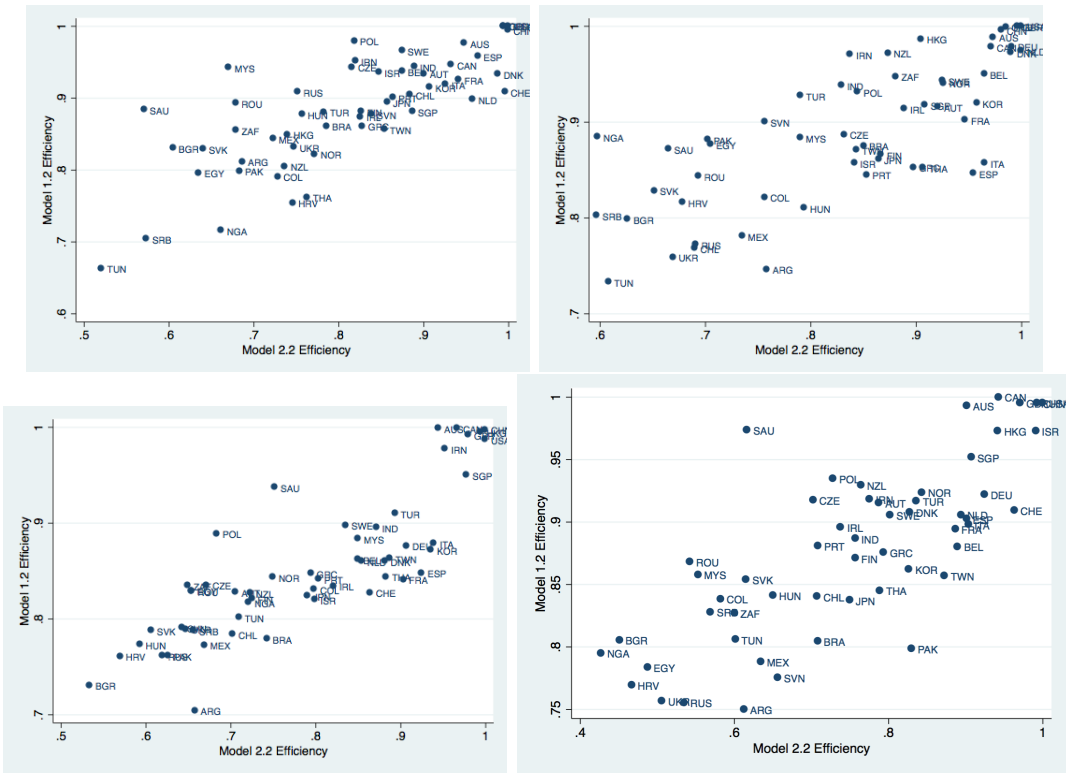

Figure 8: *Quality (2.2) vs Quantitative (1.2) Efficiencies, geometric means, for selected disciplines: PHYS, MEDI, ENGI and COMP.*

Figure 8 shows that the Iran scientific dominance in the Gulf area observed in Moed (2016) translates in a good position of Iran in the different subjects reported. Also Korea appears well positioned, after the intensive R&D investment plan (higher than 3%) introduced in the last decades.

We are interested in studying the interdependencies between scientific productivity at a macro-level (country level). The pattern of scientific productivity is different for different disciplines, in different countries, but because of the "interactions" that take place among different disciplines, one can expect that productivity patterns themselves tend to align towards a common productivity organization for different disciplines or on the contrary, a differentiated productivity configuration could emerge. The interactions between scientific productivity may be of different type, can have a different strength and can produce different effects. For example, a large exchange of collaborations between disciplines that are very productive could lead to a noticeable increase in their individual productivity, while collaborations between less productive disciplines could further reduce their productivity. On the contrary, very productive disciplines in different scientific fields could reduce their respective productivity due to the mutual learning of the basic vocabulary and tools of each discipline. There may also be cases in which different disciplines with different levels of productivity

are subject to "compensative effects" on overall scientific productivity for which the most productive disciplines slow down while the less productive ones increase their productivity. For this reason, analyzing the correlations or *indirect connections* (the *overlap* measures $Q_{i,j}$ introduced above, see Eq 3.15) between (disciplinary) productivity levels can be interesting, but going further and estimating their *interdependencies* (J_{ij}) can provide more useful information to analyze the way in which scientific knowledge is produced and the organization of the knowledge production worldwide. This information could be useful for policy makers who have to choose on which disciplines or topics to allocate research projects, or how to distribute research funds among disciplines, considering among other factors, also the impact that this allocation may have on the overall scientific productivity.

Figure 9 shows the estimated couplings parameters (J_{ij}) -left panels- and the inferred networks -right panels for the three productivity models. Top panels refer to the Basic productivity model, middle panels show the Model 1.2 (Quantitative model) and the bottom panels report Model 2.2 (Qualitative model) results. In left panels of Figure 9, the darker squares indicate higher J_{ij} and the NW to SE diagonal is made by all white squares since $J_{ii} = 0$.

The reconstructed networks reported in the right panels of Figure 9 are derived from the estimated J_{ij} obtained by the maximization of the Pseudo-Likelihood function. The J_{ij} are the edges. The diameter of the node representing the $i-th$ discipline is proportional to the number of interactions J_{ij}. The thickness of the edge depends on the intensity of the related interaction.

Figure 10 shows the calculated *overlap* measures ($Q_{i,j}$) and the estimated *interdependencies* (J_{ij}) for the different productivity models.

For instance, let's analyse the correlations ($Q_{i,j}$) and interdependencies (J_{ij}) between the productivity of CHEM with IMMU and MATE. In the basic productivity model (Figure 10 - top panel) CHEM and IMMU and CHEM and MATE show the same overlap measure $Q_{i,j} = 0.16$, meaning that their productivity tend to be positively associated. On the other hand, their respective interdependencies are different. In fact, J_{ij} between CHEM and IMMU is zero, while the interdependency between CHEM and MATE is 0.970 meaning that the productivities of CHEM and MATE present an high level of interdependency (mutual interaction). Analogously, the correlations (indirect connection) between PHYS with IMMU and MATE are respectively 0.10 and 0.12 while their interdependencies are respectively 0 and 0.49 meaning that the scientific productivity of PHYS interacts with MATE but not with IMMU although the respective productivities are correlated.

Inspecting the middle and the bottom panels of Figure 10 we note that the interaction (J_{ij}) between the productivities of PHYS and MATE is 0.79 in the Quantity model (Mod 1.2), it becomes 0.52 in the Quality model (Mod 2.2) while the respective correlations are 0.15 in Mod 1.2 and 0.1 in Mod 2.2.

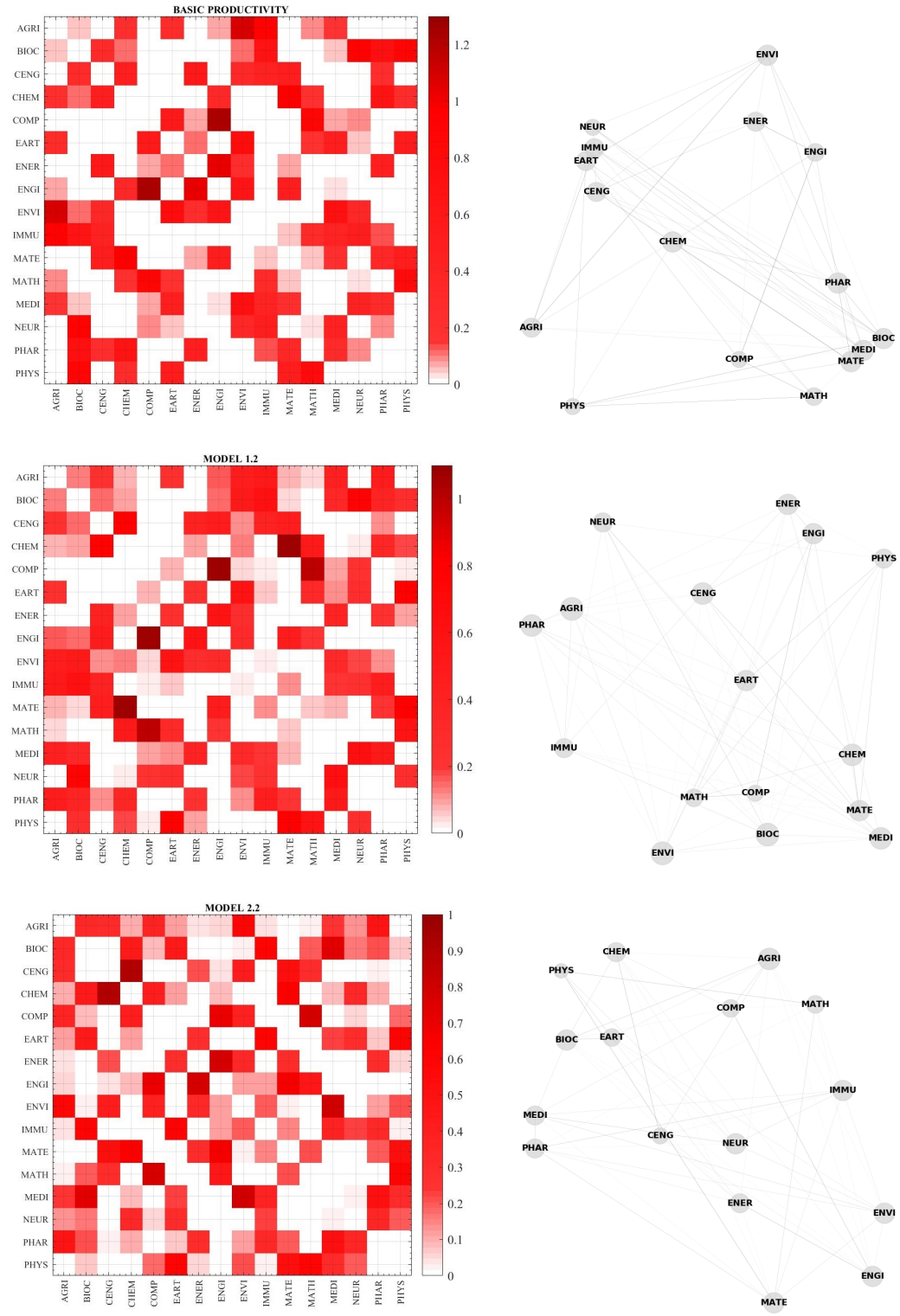

Figure 9: *Estimated J_{ij} (left panels) and inferred networks (right panels) for the three productivity models. Top panels refer to the Basic productivity model, middle panels show the Model 1.2 (Quantitative model) and the bottom panels report Model 2.2 (Qualitative model) results.*

	AGRI	BIOC	CENG	CHEM	COMP	EART	ENER	ENGI	ENVI	IMMU	MATE	MATH	MEDI	NEUR	PHAR	PHYS
AGRI	-	**0.055**	**0.000**	**0.210**	**0.000**	**0.256**	**0.000**	**0.067**	**1.080**	**0.895**	**0.000**	**0.084**	**0.162**	**0.012**	**0.000**	**0.004**
BIOC	0.182	-	**0.293**	**0.119**	**0.000**	**0.000**	**0.000**	**0.000**	**0.102**	**0.650**	**0.000**	**0.000**	**0.060**	**0.907**	**0.695**	**0.874**
CENG	0.147	0.167	-	**0.466**	**0.000**	**0.000**	**0.584**	**0.000**	**0.318**	**0.422**	**0.472**	**0.004**	**0.000**	**0.003**	**0.233**	**0.000**
CHEM	0.141	0.144	0.132	-	**0.000**	**0.000**	**0.000**	**0.316**	**0.000**	**0.000**	**0.970**	**0.188**	**0.000**	**0.000**	**0.645**	**0.253**
COMP	0.148	0.145	0.121	0.111	-	**0.551**	**0.068**	**1.223**	**0.000**	**0.000**	**0.000**	**0.867**	**0.077**	**0.089**	**0.000**	**0.000**
EART	0.167	0.149	0.122	0.106	0.163	-	**0.118**	**0.000**	**0.768**	**0.000**	**0.000**	**0.166**	**0.467**	**0.053**	**0.000**	**0.518**
ENER	0.140	0.144	0.136	0.116	0.139	0.134	-	**1.025**	**0.236**	**0.000**	**0.072**	**0.000**	**0.000**	**0.000**	**0.460**	**0.000**
ENGI	0.130	0.136	0.120	0.106	0.155	0.086	0.143	-	**0.613**	**0.000**	**0.481**	**0.000**	**0.038**	**0.000**	**0.000**	**0.000**
ENVI	0.182	0.176	0.149	0.125	0.145	0.177	0.140	0.127	-	**0.000**	**0.000**	**0.000**	**0.690**	**0.333**	**0.002**	**0.000**
IMMU	0.193	0.206	0.161	0.156	0.144	0.162	0.142	0.130	0.181	-	**0.042**	**0.302**	**0.401**	**0.481**	**0.135**	**0.000**
MATE	0.179	0.174	0.156	0.158	0.152	0.145	0.143	0.141	0.167	0.191	-	**0.060**	**0.212**	**0.000**	**0.339**	**0.491**
MATH	0.098	0.100	0.102	0.089	0.134	0.144	0.094	0.080	0.097	0.111	0.119	-	**0.000**	**0.034**	**0.013**	**0.775**
MEDI	0.218	0.220	0.157	0.169	0.174	0.193	0.153	0.154	0.210	0.254	0.213	0.103	-	**0.425**	**0.269**	**0.000**
NEUR	0.158	0.163	0.145	0.130	0.137	0.145	0.119	0.099	0.156	0.179	0.150	0.098	0.203	-	**0.084**	**0.000**
PHAR	0.152	0.171	0.144	0.136	0.115	0.105	0.125	0.119	0.147	0.166	0.151	0.084	0.173	0.132	-	**0.000**
PHYS	0.111	0.093	0.092	0.067	0.114	0.171	0.095	0.071	0.116	0.100	0.122	0.127	0.080	0.087	0.064	-

	AGRI	BIOC	CENG	CHEM	COMP	EART	ENER	ENGI	ENVI	IMMU	MATE	MATH	MEDI	NEUR	PHAR	PHYS
AGRI	-	**0.127**	**0.234**	**0.083**	**0.000**	**0.238**	**0.000**	**0.158**	**0.458**	**0.508**	**0.080**	**0.051**	**0.414**	**0.000**	**0.471**	**0.000**
BIOC	0.151	-	**0.150**	**0.098**	**0.000**	**0.000**	**0.000**	**0.144**	**0.485**	**0.602**	**0.049**	**0.000**	**0.344**	**0.762**	**0.362**	**0.249**
CENG	0.119	0.169	-	**0.785**	**0.000**	**0.000**	**0.410**	**0.480**	**0.112**	**0.410**	**0.453**	**0.000**	**0.000**	**0.003**	**0.104**	**0.000**
CHEM	0.141	0.192	0.182	-	**0.000**	**0.000**	**0.098**	**0.000**	**0.125**	**0.000**	**1.081**	**0.495**	**0.000**	**0.027**	**0.357**	**0.186**
COMP	0.115	0.161	0.144	0.139	-	**0.081**	**0.002**	**1.066**	**0.039**	**0.034**	**0.000**	**0.995**	**0.097**	**0.210**	**0.000**	**0.033**
EART	0.088	0.115	0.080	0.098	0.120	-	**0.227**	**0.000**	**0.572**	**0.062**	**0.000**	**0.326**	**0.110**	**0.267**	**0.000**	**0.783**
ENER	0.085	0.145	0.153	0.143	0.134	0.096	-	**0.583**	**0.252**	**0.000**	**0.000**	**0.000**	**0.385**	**0.006**	**0.214**	**0.096**
ENGI	0.129	0.187	0.179	0.160	0.211	0.100	0.170	-	**0.306**	**0.000**	**0.475**	**0.205**	**0.000**	**0.000**	**0.000**	**0.000**
ENVI	0.133	0.168	0.126	0.134	0.140	0.130	0.125	0.162	-	**0.020**	**0.000**	**0.000**	**0.326**	**0.174**	**0.104**	**0.000**
IMMU	0.135	0.207	0.152	0.167	0.133	0.103	0.129	0.137	0.138	-	**0.109**	**0.000**	**0.198**	**0.227**	**0.459**	**0.000**
MATE	0.145	0.205	0.188	0.223	0.168	0.098	0.155	0.199	0.145	0.178	-	**0.059**	**0.085**	**0.000**	**0.202**	**0.792**
MATH	0.081	0.136	0.109	0.153	0.174	0.108	0.119	0.170	0.108	0.094	0.167	-	**0.000**	**0.000**	**0.000**	**0.583**
MEDI	0.166	0.234	0.181	0.189	0.171	0.117	0.165	0.191	0.186	0.217	0.216	0.134	-	**0.627**	**0.482**	**0.000**
NEUR	0.135	0.177	0.146	0.160	0.131	0.093	0.105	0.140	0.129	0.160	0.158	0.097	0.181	-	**0.000**	**0.245**
PHAR	0.135	0.177	0.134	0.163	0.105	0.070	0.122	0.155	0.128	0.139	0.159	0.100	0.181	0.144	-	0.000
PHYS	0.087	0.130	0.096	0.127	0.137	0.132	0.110	0.126	0.095	0.103	0.151	0.144	0.115	0.106	0.081	-

	AGRI	BIOC	CENG	CHEM	COMP	EART	ENER	ENGI	ENVI	IMMU	MATE	MATH	MEDI	NEUR	PHAR	PHYS
AGRI	-	**0.323**	**0.299**	**0.105**	**0.346**	**0.115**	**0.039**	**0.051**	**0.586**	**0.047**	**0.000**	**0.022**	**0.248**	**0.135**	**0.475**	**0.000**
BIOC	0.127	-	**0.000**	**0.425**	**0.082**	**0.432**	**0.011**	**0.008**	**0.026**	**0.594**	**0.000**	**0.197**	**0.729**	**0.164**	**0.213**	**0.075**
CENG	0.097	0.081	-	**0.896**	**0.000**	**0.000**	**0.210**	**0.036**	**0.407**	**0.014**	**0.516**	**0.290**	**0.000**	**0.000**	**0.018**	**0.000**
CHEM	0.109	0.125	0.134	-	**0.395**	**0.116**	**0.000**	**0.094**	**0.000**	**0.000**	**0.640**	**0.000**	**0.080**	**0.321**	**0.108**	**0.000**
COMP	0.140	0.135	0.104	0.137	-	**0.000**	**0.000**	**0.691**	**0.387**	**0.000**	**0.000**	**0.766**	**0.000**	**0.047**	**0.010**	**0.172**
EART	0.118	0.120	0.078	0.075	0.135	-	**0.275**	**0.000**	**0.000**	**0.603**	**0.000**	**0.000**	**0.223**	**0.286**	**0.077**	**0.600**
ENER	0.081	0.097	0.077	0.083	0.092	0.067	-	**0.772**	**0.325**	**0.000**	**0.289**	**0.000**	**0.000**	**0.000**	**0.287**	**0.050**
ENGI	0.092	0.104	0.093	0.127	0.155	0.071	0.115	-	**0.114**	**0.114**	**0.665**	**0.462**	**0.000**	**0.000**	**0.000**	**0.000**
ENVI	0.122	0.138	0.088	0.107	0.141	0.108	0.084	0.114	-	**0.188**	**0.026**	**0.000**	**0.788**	**0.000**	**0.121**	**0.216**
IMMU	0.108	0.137	0.073	0.091	0.114	0.108	0.085	0.090	0.123	-	**0.147**	**0.000**	**0.367**	**0.227**	**0.322**	**0.029**
MATE	0.102	0.119	0.142	0.141	0.134	0.097	0.113	0.144	0.119	0.113	-	**0.216**	**0.000**	**0.000**	**0.194**	**0.524**
MATH	0.108	0.105	0.112	0.098	0.141	0.102	0.077	0.102	0.113	0.070	0.121	-	**0.000**	**0.000**	**0.000**	**0.596**
MEDI	0.126	0.142	0.088	0.104	0.124	0.110	0.085	0.098	0.146	0.129	0.118	0.106	-	**0.025**	**0.503**	**0.303**
NEUR	0.094	0.107	0.082	0.117	0.114	0.063	0.072	0.073	0.103	0.099	0.076	0.063	0.102	-	**0.331**	**0.201**
PHAR	0.117	0.125	0.103	0.120	0.118	0.071	0.085	0.103	0.121	0.118	0.120	0.074	0.136	0.119	-	0.000
PHYS	0.105	0.101	0.072	0.073	0.119	0.110	0.072	0.081	0.106	0.078	0.099	0.112	0.091	0.061	0.054	-

Figure 10: *Overlaps ($Q_{i,j}$) and Interdependencies (J_{ij}) of the Basic productivity model (top panel), of the Quantity Model (Mod. 1.2, middle panel) and of the Quality Model (Mod. 2.2, bottom panel). The North-East values reported in bold are the J_{ij} while the South-west values correspond to the $Q_{i,j}$.*

On the basis of the obtained results, we can do a comparative qualitative analysis between the three different models to explain the results and how the technique works. We solved an indirect problem to estimate the J_{ij} values for each dataset and each model separately, and so the inferred network structure is the best one for each dataset. We compared the Basic Productivity Model and Model 1.2 (Quantity Model) for checking if the Network DEA results would be affected by the *curse of dimensionality* (we have around 50 observations and more than three dimensions). As we can see, from the results it seems that they are quite similar: the only differences are due to the DEA modelling and so we should prefer the results of the Model 1.2 which accounts for the Georgescu Roegen's fund's modelling of the knowledge production.

We then compared the Model 1.2 which is based on number of papers (Quantity) with the Model 2.2 which has the same modelling of the knowledge production but considers Highly Cited Publications (Quality), to see whether the obtained estimates were consistent. We found for instance that the interaction between Physics (PHYS) and Computer Science (COMP) in the Quantity Model (1.2) is 0.033 while their interaction in the Quality Model (2.2) increases up to 0.172. In contrast, the interaction between Physics and Chemistry that is 0.186 in the Quantity Model goes down to zero in the Quality Model.

As we may expect, the interactions between CHEM and CENG are quite high in all the three productivity models (0.466 in the basic model, 0.785 in Mod. 1.2 and 0.896 in Mod 2.2); other expected results are the high interactions of COMP with ENGI and MATH because these disciplines share the same community. A striking result is the interactions we observe between PHYS and MEDI which is quite high in the Quality model (Mod. 2.2, with a value of 0.303) but absent in the Quantity model (Mod. 1.2). A policy implication of this result could be made in the discussion about supporting Great Societal Challenges that are focused on Medical Sciences and the important role that Physics could play in this context.

5 Conclusions

The economics of science (Stephan, 2012) reminds us that researchers do research for different reasons including their interest in "puzzle solving", reputation based on the priority of their discovery, awards and recognition for their achievements, also through publications which have a key role for funding and promotion. Research is a public good. This means that it has a non-excludable nature. This offers the possibility of free-riders behaviours and difficulty in capturing economic returns. When free-riders exist competitive markets tend to under produce public goods. In research, reward is not for effort but for achievement. What the economics of science tells us about the production of scientific research is that the production of research involves multiple inputs, including knowledge, time materials and equipment. Some inputs are embedded in people (knowledge and time in particular) and most of these inputs are expensive. As observed by Stephan (2012, p.228), incentives and cost matter for science and economics, but they are also about the allocation of scarce

resources across competing wants and needs; economics is also about whether resources are allocated efficiently. Stephan (2012, p.235) in concluding her book *How Economics shapes Sciences*, states three more general and difficult efficiency questions which require further research, and includes one about the definition of the *most efficient mix* in terms of budget allocated to the different disciplines.

In this paper, we propose a first step in the direction of facing this challenging issue. By applying a semi-parametric Bayesian approach, based on recently introduced pseudo-likelihood techniques developed in the context of complex systems, we assessed the interdependencies of disciplinary scientific productivity at the macro level.

The obtained results can be of great relevance for the policy-makers, for example for the distribution of resources among disciplines. For instance the non trivial high interaction value found here, measuring the productivity in terms of Highly Cited Publications, between Physics and Medicine could be useful to include physical sciences (or their interdisciplinary convergence with medical sciences) in the topics for the analysis of Great Societal Challenges of which medical sciences are the first and the most important one.

This approach could be applied also to estimate the interdependencies of structural *industrial profiles* (industrial value added) and structural *innovative/technological profiles* (based on patents) and infer the underlying network topology without assuming it.

Summing up, statistical mechanics of complex system can be exploited for making inference about productivity networks. The obtained results seems promising. The general framework developed here is based on complex systems behavior modeling and estimation which is strictly connected with the economic and organizational model of production of Network DEA seen as an implementation of the Georgescu Roegen's economic model (GRFF).

More generally, the proposed statistical inferential framework may be applied in a variety of productivity network problems (Kao, 2017), to infer the underlying topology of the network.

There are many possible extensions, left for further studies, including the implementation of out-of-equilibrium time-dependent Ising model.

6 Acknowledgements

The data used in this paper have been provided by Elsevier within the EBRP Project framework.
The financial support of the Italian Ministry of Education and Research (through the PRIN Project N. 2015RJARX7), of Sapienza University of Rome (through the Sapienza Awards no. 6H15XNFS), and of the Lazio Region (through the Project FILAS-RU-2014-1186) is gratefully acknowledged.
Previous versions of this paper have been presented at the 10^{th} International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2017), 16-18 December 2017, University of London (UK); at the International Conference on Data

Envelopment Analysis, DEA40 Aston University, Birmingham (UK), 16-18 April 2018, at the X North American Productivity Workshop (NAPW) Miami (USA) 12-15 June 2018, and at the 12th Asia-Pacific Productivity Conference (APPC 2018), Seul, Korea, 4-6 July 2018. Helpful discussions and comments by conferences' participants and in particular by Mario Morroni, Robin Sickles and Léopold Simar are gratefully acknowledged.

References

[1] Aksnes, D., Sivertsen, G., van Leeuwen, T. N., Wendt, K. K. (2017). Measuring the productivity of national R&D systems: Challenges in cross-national comparisons of R&D input and publication output indicators. Science and Public Policy, 44(2), 246-258.

[2] Aurell, E., Ekeberg, M. (2012). Inverse Ising inference using all the data. Physical Review Letters, 108(9), 090201.

[3] Barber,D. (2012), *Bayesian Reasoning and Machine Learning*, Cambridge University Press.

[4] Banerjee O., L. El Ghaoui, and A. dAspremont, (2008) 'Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data', Journal of Machine Learning Research, vol. 9, pp. 485516.

[5] Besag J., (1986), 'On the statistical analysis of dirty pictures', Journal of the Royal Statistical Society, Series B, vol. 48, no. 3, pp. 259279.

[6] Bogetoft, P., Färe, R., Grosskopf, S., Hayes, K., Taylor, L. (2009). Dynamic Network Dea: An Illustration. Journal of the Operations Research Society of Japan, 52(2), 147-162.

[7] Bostian, M. B., Färe, R., Grosskopf, S., Lundgren, T. (2016). Environmental Investment and Firm Performance: A network approach. Energy Economics 57, 243-255.

[8] Bostian, M.,Färe, R., Grosskopf, S. and Lundgren, T. (2017). Prevention or cure? Evaluating the tradeoffs between emissions abatement measures. mimeo.

[9] Cook, W. D., Liang, L., Zhu, J. (2010). Measuring performance of two-stage network structures by DEA: a review and future perspective. Omega, 38(6), 423-430.

[10] Cook, W. D., Zhu, J. (Eds.). (2014). *Data envelopment analysis: A handbook of modeling internal structure and network* (Vol. 208). Springer.

[11] Danø, S. (1966). Industrial Production Models: A theoretical study. Springer Verlag, New York.

[12] Daraio C. (2017), A Framework for the assessment of research and its impacts.Journal of Data and Information Science, Vol. 2 No. 4, 2017 pp 742.

[13] Daraio C., Fabbri F., Gavazzi G., Izzo M.G., Leuzzi L., Quaglia G., Ruocco G. (2018), Assessing the interdependencies between scientific disciplinary profiles, *Scientometrics*, https://doi.org/10.1007/s11192-018-2816-5.

[14] Daraio C. (2018), Econometric Approaches to the measurement of research productivity, in *Springer Handbook of Science and Technology Indicators* edited by Glänzel, W. , Moed H.F., Schmoch H. and Thelwall M., *forthcoming*.

[15] Daraio, C., Simar, L., Wilson, P. W. (2017). Central limit theorems for conditional efficiency measures and tests of the "separability" condition in nonparametric, two-stage models of production. The Econometrics Journal.

[16] Elsner, W., Heinrich, T., Schwardt, H. (2015). *The microeconomics of complex economies: Evolutionary, institutional, neoclassical, and complexity perspectives*. Academic Press. Oxford, UK: Academic Press.

[17] Färe, R. (1988). Fundamentals of production theory. Berlin: Springer-Verlag.

[18] Färe, R. and Grosskopf, S. (1996). *Intertemporal Production Frontiers: With dynamic DEA*. Kluwer Academic Publishers, Boston.

[19] Färe, R. and Primont, D. (1995). *Multi-output production and duality: theory and applications*. Norwell, MA: Kluwer Academic.

[20] Färe, R., Grosskopf, S. (2000). Network DEA. *Socio-economic planning sciences*, 34(1), 35-49.

[21] Färe, R., Grosskopf, S. (2006). New directions: efficiency and productivity (Vol. 3). Springer Science Business Media.

[22] Färe, R., Grosskopf, S., Whittaker, G. (2007). Network DEA. In Zhu and Cook(Eds.)(2007). *Modeling data irregularities and structural complexities in data envelopment analysis*. Springer Science & Business Media, 209-240.

[23] Färe, R., Grosskopf, S., Whittaker, G. (2014). Network DEA II. In *Data envelopment analysis: A handbook of modeling internal structure and network* (pp. 307-327). Springer US.

[24] Fioretti, G. (2007), The production function. Physica A: Statistical Mechanics and its Applications, 374(2), 707-714.

[25] Førsund, F. R. (2018). Economic interpretations of DEA. Socio-Economic Planning Sciences, 61, 9-15.

[26] Frisch, R. (1965). *Theory of production*. Dordrecht, D. Reidel.

[27] Fukuyama,H., Weber, W.L., Xia, Y. (2016). Time substitution and network effects with an application to nanobiotechnology policy for US universities, The International Journal of Management Science 60: 34-44.

[28] Georgescu-Roegen, N. (1970), The economics of production.The American Economic Review,60:2. 1-9.

[29] Georgescu-Roegen, N. (1971), *Entropy law and the Economic process*, Cambridge.

[30] Georgescu-Roegen, N. (1972), Process analysis and the neoclassical theory of production, *American Journal of Agricultural Economics*, 279-294.

[31] Georgescu-Roegen, N. (1979), Methods in economic science, Journal of economic issues (1979): 317-328.

[32] Geman S. and Geman D., (1984), 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 6, pp. 721741.

[33] Granovetter, M. S. (1973). The Strength of Weak Ties. *The American Journal of Sociology*. 78 (6): 13601380.

[34] Greig D. M., B. T. Porteous, and A. H. Seheuly,(1989) 'Exact maximum a posteriori estimation for binary images', Journal of the Royal Statistical Society B, vol. 51, pp. 271279.

[35] Golan, A. (2008). Information and Entropy Econometrics. A Review and Synthesis. Foundations and Trends (R) in Econometrics, 2(12), 1-145.

[36] Hayek, F.A. (1967), *Studies in Philosophy, Politics and Economics*. London: Routledge and Kagan Paul.

[37] Hinterberger, F. (1994), *Self-organizing systems*. Chapter 28 in The Elgar Companion to Austrian Economics edited by Peter J. Boettke. Edward Elgar, pp. 187-191.

[38] Hyvarinen A., (2006), Consistency of pseudolikelihood estimation of fully visible Boltzmann machines, *Neural Computation*, vol. 18, pp. 2283-2292, .

[39] Jaynes E.T. (1957). Information theory and statistical mechanics, Phys. Rev., 106, 620-630.

[40] Judge, G. G., Mittelhammer, R. C. (2011). *An information theoretic approach to econometrics*. Cambridge University Press.

[41] Kao, C. (2009). Efficiency decomposition in network data envelopment analysis: A relational model, European Journal of Operational Research 192, 949-262.

[42] Kao, C. (2014). Network data envelopment analysis: A review. European Journal of Operational Research, 239(1), 1-16.

[43] Kao, C. (2017), *Network data envelopment analysis; Foundations and extensions*. Springer.

[44] Kemeny,J., Morgenstern, O., and Thompson, G. (1956). A generalization of the von Neumann model of an expanding economy. Econometrica 24:2, 115-135.

[45] Kirman, A. (2016). Networks: A Paradigm Shift for Economics? In Bramoullé, Y., Galeotti, A., Rogers, B. W. (Eds.) *The Oxford handbook of the economics of networks*. Oxford University Press.

[46] Kumbhakar, S. C., Parmeter, C. F., Tsionas, E. G. (2012). Bayesian estimation approaches to first-price auctions. Journal of Econometrics, 168(1), 47-59.

[47] Luwel, M. (2004). The use of input data in the performance analysis of R&D systems. In *Handbook of quantitative science and technology research* (pp. 315-338). Springer, Dordrecht.

[48] Marruzzo, A., Tyagi, P., Antenucci, F., Pagnani, A., Leuzzi, L. (2016). Inverse problem for multi-body interaction of nonlinear waves. arXiv preprint arXiv:1607.08549.

[49] Miller, R. E., Blair, P. D. (2009). *Input-output analysis: foundations and extensions*. Cambridge University Press.

[50] Moed, H. F. (2016). Iran's scientific dominance and the emergence of South-East Asian countries as scientific collaborators in the Persian Gulf Region. Scientometrics, 108(1), 305-314.

[51] Morroni, M. (1992), Production Process and Technical Change, Cambridge University Press, repr. 2009.

[52] Morroni, M. (2006), Knowledge, Scale and Transactions in the Theory of the Firm, Cambridge University Press, repr. 2009.

[53] Morroni, M. (2014), Production of commodities by means of processes. The flow-fund model, input-output relations and the cognitive aspects of production, Structural Change and Economic Dynamics.

[54] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[55] Parisi, G. (1986). Asymmetric neural networks and the process of learning. Journal of Physics A: Mathematical and General, 19(11), L675.

[56] Parmeter, C. F., Kumbhakar, S. C. (2014). Efficiency analysis: a primer on recent advances. Foundations and Trends in Econometrics, 7(34), 191-385.

[57] Prieto, A. M., Zofio, J. L. (2007). Network DEA efficiency in input-output models: with an application to OECD countries. European Journal of Operational Research, 178(1), 292–304.

[58] Ravikumar, P., Wainwright, M. J., Lafferty, J. D. (2010). High-dimensional Ising model selection using a regularized logistic regression. The Annals of Statistics, 38(3), 1287-1319.

[59] Shannon, C. E. (1948). A mathematical theory of communication (parts I and II). Bell System technical journal, 379-423.

[60] Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., White, D. R. (2009a). Economic networks: The new challenges. Science, 325(5939), 422-425.

[61] Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., White, D. R. (2009b). Economic Networks: What do we know and what do we need to know?. Advances in Complex Systems, 12(04-05), 407-422.

[62] Scott, J. (2017). *A Social network analysis*. Sage. First edition in 1999, 4th edition in 2017.

[63] Sergeev, V.M. (2005), The thermodynamic approach to market. arXiv:0803.3432 [physics.soc-ph]

[64] Sethna, J. (2006). Statistical mechanics: entropy, order parameters, and complexity (Vol. 14). Oxford University Press.

[65] Shephard, R. W. (1953). *Cost and production functions*. Princeton NJ: Princeton University Press.

[66] Shephard, R.W. and Färe, R. (1974). The law of diminishing returns. Journal of Economics 34:1, 69-90.

[67] Shephard, R.W. and Färe, R. (1975). The dynamic theory of production correspondences, Berkeley operations research center.

[68] Sickles, R. C. and Zelenyuk, V. (2018), *Measurement of Productivity and Efficiency: Theory and Practice*, Cambridge University, forthcoming.

[69] Simar, L., Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. Journal of econometrics, 136(1), 31-64.

[70] Simar, L., Wilson, P. W. (2013). Estimation and inference in nonparametric frontier models: Recent developments and perspectives. Foundations and Trends in Econometrics, 5(34), 183-337.

[71] Simar, L., Wilson, P. W. (2015). Statistical approaches for nonparametric frontier models: a guided tour. International Statistical Review, 83(1), 77-110.

[72] Stephan, P. E. (2012). *How economics shapes science*. Cambridge, MA: Harvard University Press.

[73] Trinh, K. and Zelenyuk, V. (2015). Bootstrap-based testing for network DEA: Some theory and applications. Working Paper No. WP05/2015, School of Economics, University of Queensland.

[74] Tsionas, E. G., Papadakis, E. N. (2010). A Bayesian approach to statistical inference in stochastic DEA. Omega, 38(5), 309-314.

[75] Tyagi, P., Marruzzo, A., Pagnani, A., Antenucci, F., Leuzzi, L. (2016). Regularization and decimation pseudolikelihood approaches to statistical inference in X Y spin models. Physical Review B, 94(2), 024203.

[76] Van den Broeck, J., Koop, G., Osiewalski, J., Steel, M. F. (1994). Stochastic frontier models: A Bayesian perspective. Journal of Econometrics, 61(2), 273-303.

[77] Vittucci Marzetti, G. (2012), The flow-fund approach: A critical survey. *Journal of Economic Survey.*

[78] Vozna, L.Y. (2016), The Notion of Entropy in Economic Analysis: The Classical Examples and New Perspectives. *Journal of Heterodox Economics* 3(1): DOI 10.1515/JHEEC-2016-0001

[79] Wainwright M. J., Jordan, M. I. (2008) 'Graphical Models, Exponential Families, and Variational Inference', Foundations and Trends in Machine Learning, Vol. 1, pp. 1305.

[80] Weber, W.L. (2017). Network Production and Shadow Prices of Knowledge Outputs, mimeo, Economics, Southeast Missouri State University.