Peer-reviewed version available at *Entropy* **2019**, *21*, 126; doi:10.3390/e21020126

Article

Measuring the impact of technological innovations on industrial products through a multilayer network approach

Martina Formichini[1], Giulio Cimini[2,3], Emanuele Pugliese[4], Andrea Gabrielli[3]

[1] Dipartimento di Fisica, "Sapienza" Università di Roma, Piazzale A. Moro 5, 00185 - Rome, Italy
[2] IMT School for Advanced Studies, Piazza S. Ponziano, 6 - 55100 Lucca (Italy)
[3] Istituto dei Sistemi Complessi (ISC) - CNR, UoS Sapienza, Dipartimento di Fisica, "Sapienza" Università di Roma, Piazzale A. Moro 5, 00185 - Rome, Italy
[4] EC - JRC, Calle Inca Garcilaso 3, E-41092 Seville (Spain)

Version August 4, 2018 submitted to

In this work we identify combinations of technological activities that signal the presence local capabilities in a country to successfully export a product. We use country-level patent and trade data to generate a multi-layer network, and we apply maximization of entropy to generate synthetic data to effectively divide signal from noise. We show that in several sectors the signal far exceed the noise. Our exercise provides robust evidence of the presence of synergies between technologies to explain trade performances in specific markets. This can be highly useful for policy makers, to inform industrial and innovation policies.

1. Introduction

Technological innovation is the main driver of modern economic growth [1,2]. It is therefore not surprising that measuring and predicting the potential impact of technological innovations on export competitiveness has been the central issue of many studies in the last forty years[3,4] and the focus of a general interest in the field of innovation systems. Such academic effort provided a theoretical framework and empirical stylized facts helpful to understand the crucial effect of innovation in shaping competitive advantage of countries in different markets. Policy makers are however more interested in identifying the specific technologies that are relevant for specific markets[5,6]. These questions are much harder to answer in an organic and objective fashion: the scientific approach to address the impact of specific technologies on specific markets has been through ad-hoc case studies of difficult comparability.

In a recent paper [7] a multilayer network approach to the innovation system has been proposed: starting from the three bipartite networks countries-products, countries-technologies and countries-scientific sectors, with arbitrary relative time delay, a three-layered network of innovation activities is derived in which it is possible to measure the conditional probability that a bit of information produced in an innovation activity (e.g. a scientific domain as for instance microelectronics) will impact after a given delay on another innovation activity (e.g. an industrial product category). By grounding on previous fundamental studies of Economic Fitness and Complexity [8,9], this is the first approach representing the innovation system as a complex system and using bipartite and multilayer network, diffusion theory on networks and maximum entropy network ensembles for its study.

In this paper we adopt a similar approach, by limiting our analysis to the relationships between technologies and products, but generalizing the method so that we can measure the potential impact of a group of technological sectors (in particular pairs of sectors) on single product categories. This is indeed a crucial aspect of technology: the interaction between different technological fields is often a crucial aspect of technological progress[10]

We use the following established databases to construct the bi-layered technologies-products space: for Technology, we consider the number of patents in different technological sectors extracted from Patstat (`www.epo.org/searching − for − patents/business/patstat`); and for Products, we

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2018 doi:10.20944/preprints201808.0081.v1
Peer-reviewed version available at Entropy 2019, 21, 126; doi:10.3390/e21020126

use export data collected by UN COMTRADE (`https://comtrade.un.org/`) - which are typically used as proxy of a competitive industrial production.

Results are then statistically validated by building appropriate advanced null models derived from the constrained maximum entropy Bipartite Configuration Models (BiCM) for both the bipartite countries-technologies and the countries-products networks.

2. General approach and strategy

Starting from Patstat and UN COMTRADE databases, we can define the Revealed Comparative Advantage of a country c on an activity a (which has to be meant either a technological sector t or and product category p) in a given year y as

$$RCA_{ca} = \frac{W_{c,a}(y)}{\sum_{a'} W_{c,a'}(y)} \bigg/ \frac{\sum_{c'} W_{c',a}(y)}{\sum_{c',a'} W_{c',a'}(y)}, \tag{1}$$

where in the case of a technology t, the quantity $W_{c,t}(y)$ is the number of patents of country c in the technological sector t in the year y, while in the case of products $W_{c,p}(y)$ we use, as a proxy of the activity of a country c in a product category p in the year y, the amount of dollars coming to the same country by the export of that product in that year. The years interval considered in this work $1995-2012$, the index c spans over all the considered countries whose total number N_c varies from 66 to 72 depending on the year, t over the total set of technological sectors whose number N_t veries between a minimum number 629 and a maximum 636 also depending on the year, and p spans over the total set of considered product categories N_p which varies between 1140 and 1176 (note that slight variations of N_c, N_t and N_p year by year are due to geographical changes and periodical recategorization of technologies and products).

It is possible to define for a given year y the binary bipartite networks countries-technologies and countries-products respectively represented by the binary biadjacency matrices $\mathbf{M}^{\mathcal{C},\mathcal{T}}(y)$ and $\mathbf{M}^{\mathcal{C},\mathcal{P}}(y)$ whose elements $M_{ct}(y)$ and $M_{cp}(y)$ are defined as follows:

$$M_{ca}(y) = \begin{cases} 1 & \text{if } RCA_{ca}(y) \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where again a refers to a technological sector t in the case of the bipartite countries-technologies network and to a product category p in the countries-product case.

Once we have the matrices $\mathbf{M}^{\mathcal{C},\mathcal{T}}(y_1)$ for the year y_1 and $\mathbf{M}^{\mathcal{C},\mathcal{P}}(y_2)$ for the year y_2, in analogy with [7], we can construct the *assist matrix* $\mathbf{B}^{\mathcal{T}\to\mathcal{P}}(y_1, y_2)$, whose generic element is defined as:

$$B_{tp}(y_1, y_2) = \sum_{c=1}^{N_c} \frac{M_{ct}(y_1)}{k_t(y_1)} \frac{M_{cp}(y_2)}{k_c^{(p)}(y_2)} \tag{3}$$

where $k_t(y_1) = \sum_c M_{ct}(y_1)$ is the number of countries having the technology t in their technological basket of year y_1, $k_c^{(p)}(y_2) = \sum_p M_{cp}(y_2)$ is the cardinality of the product basket of country c in the year y_2. As well explained in [7], $B_{tp}(y_1, y_2)$ with $y_1 \leq y_2$ gives the conditional probability that a bit of information produced in the technological sector t in the year y_1 arrives at the product category p in the year y_2 (through one of the countries having t in its technological basket of the year y_1 and p in the product basket of the year y_2). Note that the elements $B_{tp}(y_1, y_2)$ can be also seen as the weights of the links of the bipartite network technologies-products with the former at year y_1 and the latter at year y_2 whose probabilistic interpretation has just been given above.

Through the assist matrix $\mathbf{B}^{\mathcal{T}\to\mathcal{P}}(y_1, y_2)$, we define the Λ motifs of such bipartite network as:

$$\Lambda_{tt'}^{p}(y_1, y_2) = B_{tp}(y_1, y_2) B_{t'p}(y_1, y_2) \tag{4}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2018 doi:10.20944/preprints201808.0081.v1

Peer-reviewed version available at Entropy 2019, 21, 126; doi:10.3390/e21020126

which measures the joint and conditional probability of co-occurrence in a single country of a pair of technologies t and t' at year y_1 and of a product p at year y_2. As explicitly shown in [7] $\Lambda^p_{tt'}(y_1, y_2)$ can also seen as the joint and conditional probability that two bits of information produced in the technological sectors t and t' in the year y_1 arrives at the product category p in the year y_2 (through one of the countries having t and t' in its technological basket of the year y_1 and p in the product basket of the year y_2). The name $\Lambda-$motif comes from the fact that in the aforementioned bipartite network technologies-products between the years y_1 and y_2, Eq. (4) gives the weight of the a Λ shaped set of two links having different origin in the technology layer and same end in the product layer [11].

Thus in principle starting from the biadjacency matrices $\mathbf{M}^{\mathcal{C},\mathcal{T}}(y_1)$ and $\mathbf{M}^{\mathcal{C},\mathcal{P}}(y_2)$ extracted directly from the databases Patstat and UN COMTRADE at different respective years y_1 and y_2, one can measure the complex probability $\Lambda^p_{tt'}(\Delta y)$ of impact of pairs of technologies on a product category after an arbitrary time-delay Δy, by averaging up all $\Lambda^p_{tt'}(y_1, y_2)$ with fixes years difference $\Delta y = y_2 - y_1$. In principle this quantity can be even easily generalized to assess the impact of a wider group of technologies on a single product, but for aim of simplicity and statiostical significance we will limit here our analysis to pairs of technologies and single products.

Once we evaluate the probability $\Lambda_{tt'}(\Delta y)$ from real datasets, we statistically validate it by comparing its value with the probability distribution of the values taken by the analogous quantity in an appropriate null model.

Such a null model is schematically defined as follows (see Appendix for a thorough presentation): for each year y we build two statistical ensembles of biadjacency matrices $\tilde{\mathbf{M}}^{\mathcal{C},\mathcal{T}}(y)$ and $\tilde{\mathbf{M}}^{\mathcal{C},\mathcal{P}}(y)$ respectively for the bipartite networks countries-technologies and countries-products which are maximally random apart from fixing the ensemble mean values of the degrees of the nodes of such bipartite networks (i.e. the technological diversification of countries $\tilde{k}^{(t)}_c(y) = \sum_t \tilde{M}^y_{ct}$ and the technologies ubiquities $\tilde{k}_t(y) = \sum_c \tilde{M}^y_{ct}$ for the first bipartite network and the product diversification of countries $\tilde{k}^{(p)}_c(y) = \sum_p \tilde{M}^y_{cp}$ and the product ubiquities $\tilde{k}_p(y) = \sum_c \tilde{M}^y_{cp}$ for the second bipartite network) equal to the values in the real matrices. The probability measure of both statistical ensembles of binary bipartite networks is, as explicitly shown in Appendix through a constrained maximum entropy approach in the spirit of the information theory approach to the statistical mechanical ensembles [12].

These ensemble are called in literature Bipartite Configuration Models (BiCM) [13]. Finally, given the appropriate BiCMs for the all bipartite networks $\tilde{\mathbf{M}}^{\mathcal{C},\mathcal{T}}(y_1)$ and $\tilde{\mathbf{M}}^{\mathcal{C},\mathcal{P}}(y_2)$ with fixed $y_2 - y_1 = \Delta y$, we can derive, through Eqs. (3) and (4) appropriately applied to the BiCMs, the probability distribution of the value $\tilde{\Lambda}^p_{tt'}(\Delta y)$ of the joint probability that two bits of information, started a given year from the technologies t and t', arrive at the product p after a time delay Δy.

As explained in the section below, by comparing in terms of fixed statistical significance value $\alpha = 0.01$ the values of the probabilities $\Lambda^p_{tt'}(\Delta y)$ obtained by the real bipartite networks with respect to the distribution of the values of the same quantity in the null model, we can finally validate the real impact of the pair of technologies on single products after an arbitrary delay.

3. Results

As mentioned in the previous section, the empirical assist matrices $\mathbf{B}^{\mathcal{T}\to\mathcal{P}}(y_1, y_2)$ we consider are made up of a number between 629 and 636 of technological sectors (at a 4-digits resolution) and a number between 1140 and 1176 products (also at a 4-digits resolution) depending on the years y_1 and y_2. Years y_1 and y_2 which refer to matrices of countries-technologies $\mathbf{M}^{\mathcal{C},\mathcal{T}}(y_1)$ and of countries-products $\mathbf{M}^{\mathcal{C},\mathcal{P}}(y_2)$ respectively both range from 1995 to 2012. The null case for these matrices are built by numerically generating two ensemble of 10^3 matrices $\tilde{\mathbf{M}}^{\mathcal{C},\mathcal{T}}(y_1)$ and $\tilde{\mathbf{M}}^{\mathcal{C},\mathcal{P}}(y_2)$ using the BiCM, and then contracting each pair to generate a final ensemble of 10^3 null matrices $\tilde{\mathbf{B}}^{\mathcal{T}\to\mathcal{P}}(y_1, y_2)$. As explicitely described in the previous section, we define Λ motifs of the assist matrix by Eq. (4) and then we average the values of $\Lambda^p_{tt'}(y_1, y_2)$ over all pairs of years with the same difference $y_2 - y_1 = \Delta y$ in order to get $\Lambda^p_{tt'}(\Delta y)$ representing the co-occurrences of *two* technologies t, t' and *one* product p with a time delay

doi:10.20944/preprints201808.0081.v1

Peer-reviewed version available at *Entropy* **2019**, *21*, 126; doi:10.3390/e21020126

Δy between technologies and product in the capabilities of countries. We define the signal $\phi(\Delta y)$ as the fraction of significant $\Lambda^{p}_{tt'}(\Delta y)$ (at the $\alpha = 0.01$ significance level) for combinations of t, t' and p chosen for selected matrix regions. The exact procedure followed to generate a null model to determine the statistical significance of the observed $\Lambda^{p}_{tt'}(\Delta y)$ is explained in the Appendix 5. In general we consider a population of motifs equal to 5500 units (see below).

As a first test we report the mean signal ϕ, *i.e.*, the signal averaged over all combinations of t, t' and p. Since the total number of such motifs is extremely large, we choose 5500 motifs at random and take their signal as representative of the mean one. Figure 1 shows that ϕ basically remains within one standard deviation from the noise level α, indicating that the mean signal within the data is negligible.

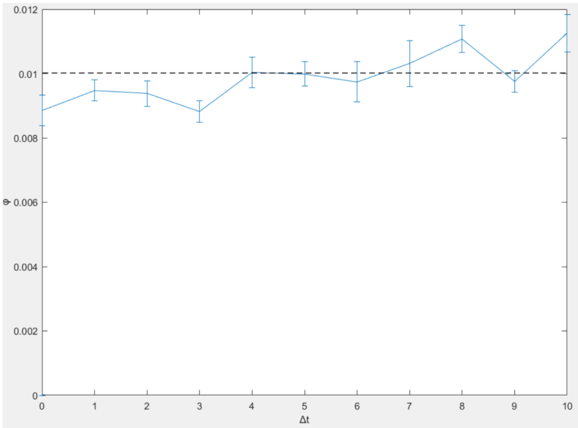

Figure 1. Mean signal ϕ of 5500 combinations of t, t' and p chosen at random, for different values of the time lag Δy. Error bars represent the standard deviation over the year pairs giving the same time lag, whereas, the dotted line is the significance level α.

We then report the signal relative to motifs within selected regions of the assist matrix. Specifically, we choose sub-regions of 11 technologies and 100 products (related to specific technology and production fields), whose total number of (unordered) Λ motifs is 5500. From Figure 2 we see that, by selecting coherent sets of technologies and products, the signal is much enhanced: the presence of a pair of technologies in the capability basket of a country can predict if that country can successfully export a product, and this happens almost independently on the time lag Δy.

As consistency checks we make two exercises, both reported in Figure 3. Firstly, we show that the results we just presented do not depend on the particular resolution used to choose the motifs. Secondly, we show that for incoherent technologies and products, we can indeed get a much lower signal—even lower than the significance level. In this latter case, a significant development of specific technologies corresponds to a low-level export of given products.

We finally want to provide a few examples of motifs with high signal. To do that, since the total number of motifs is extremely high so that a complete exploration is not efficient and surely not clever, we checked the motifs made up of the link pairs $B_{tp}(y_1, y_2)$ and $B_{t'p}(y_1, y_2)$ which are independently the most significant. Table 1 reports some instances of such motifs for the specific choice $\Delta y = 0$.

4. Conclusions

In this work we provided an effectual way of measuring the combined effect of a set of technology on one product. In particular in this work we highlight lambda motifs: the paired effect of two technology together.

In the process of finding relevant combinations, we highlight several results. First of all, we show how the combination of multiple technologies has a very different role in different industrial and technological sectors. This heterogeneity, while expected, is here quantitatively measured. Secondly,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2018 doi:10.20944/preprints201808.0081.v1

Peer-reviewed version available at Entropy 2019, 21, 126; doi:10.3390/e21020126

Table 1. Sample of top-significant Λ motifs. The p-value is averaged over all year pairs $y_1 = y_2$ giving $\Delta y = 0$. To make this selection, we picked the most significant pairs (individual links) (t, p) and then choose t' within the region where the average p-value was highest.

p-value	p, t, t'
$2 \cdot 10^{-4}$	4701: Wood pulp C05B: Lime; Magnesia; Slag; Cements C09K: Materials for applications not otherwise provided for
$5 \cdot 10^{-4}$	2605: Mineral products C21D: Modifying the physical structure of ferrous metals F04F: Pumping of fluid by direct contact of another fluid or by using inertia of fluid to be pumped
$5 \cdot 10^{-4}$	2605: Mineral products C21D: Modifying the physical structure of ferrous metals F04F: Working metallic powder
$5 \cdot 10^{-4}$	8443: Printing machine D02H: Mechanical methods or apparatus in the manufacture of artificial filaments G01T Measurement of nuclear or x-radiation
$7 \cdot 10^{-4}$	4703: Chemical wood pulp D21F: Decorating textiles B27C: Planing, drilling, milling, turning, or universal machines
$8 \cdot 10^{-4}$	2605: Mineral products C21D: Modifying the physical structure of ferrous metals FF15B: Systems acting by means of fluids in general
$2 \cdot 10^{-3}$	4703: Chemical wood pulp D21F: Paper-making machines F03D: Wind motors
$2 \cdot 10^{-3}$	4703: Chemical wood pulp D21F: Paper-making machines D06Q: Decorating textiles
$3 \cdot 10^{-3}$	8519: Sound recording or reproducing apparatus G10K: Sound-producing devices G01T: Capacitors; rectifiers, detectors, switching devices
$3 \cdot 10^{-3}$	8519: Sound recording or reproducing apparatus G10K: Sound-producing devices G04f: Time-interval measuring

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2018 doi:10.20944/preprints201808.0081.v1

Peer-reviewed version available at *Entropy* **2019**, *21*, 126; doi:10.3390/e21020126

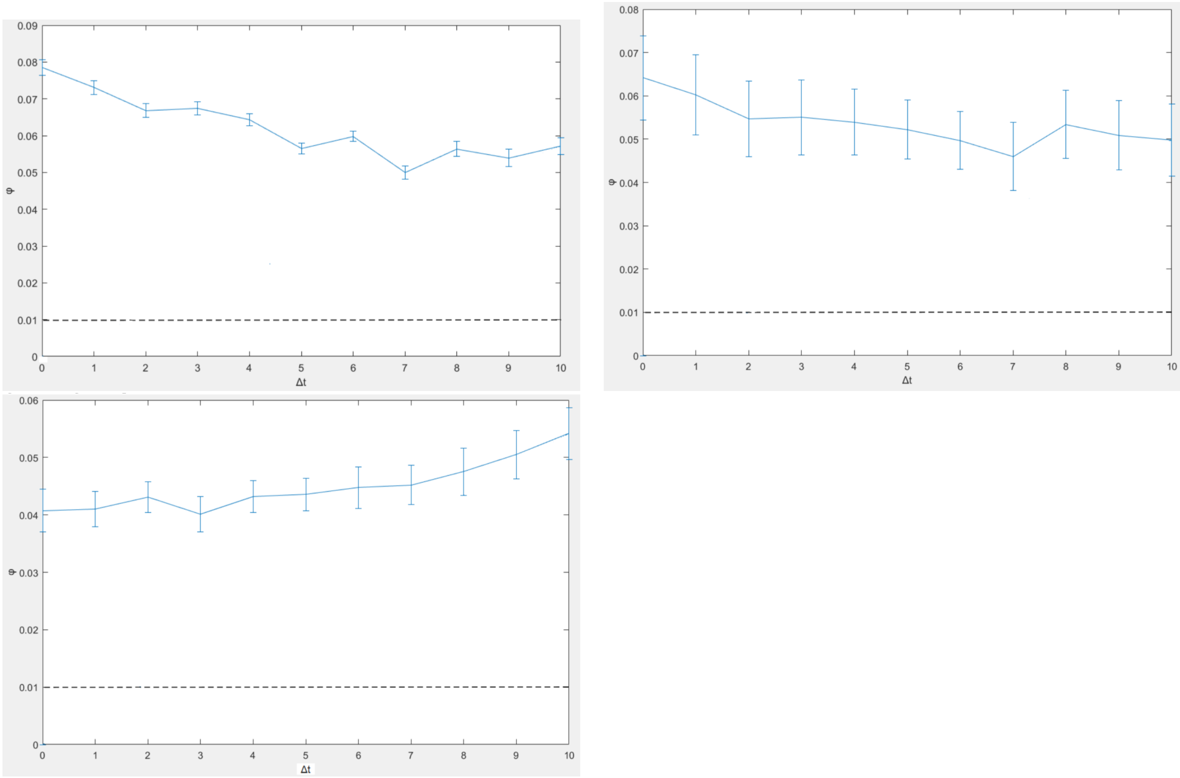

Figure 2. Mean signal ϕ of 5500 combinations of t, t' and p chosen for specific region of the assist matrix. Top left panel: technological codes related to sector "physics" – in the region G09F-G11B *instruments of communications, acoustics, optics*, and products in the region 8401-8518 *machinery and metals*. Top right panel: technological codes related to sector "engineering" – in the region F22G-F23N *various types of machines including steam and combustion*, and products in the region 8401-8518 *machinery and metals*. Bottom panel: technological codes related to sector "chemistry" – C09H e C10L *macromolecular and inorganic compounds, gas and petroleum*, products, and products in the region 2706-3104 *inorganic and organic chemicals, pharmaceuticals*. In all cases, error bars represent the standard deviation over the year pairs giving the same time lag, whereas, the dotted line is the significance level α.

we confirm that regional co-occurrences between activities are able to extract information on shared capabilities, to inform policy makers and stakeholders of relevant synergies.

The mapping provided by the techniques of the effects of pairs of technologies on products can be transformed in a powerful instrument to inform policies and industrial strategies. This operational step will be an important aspect of future research.

5. Appendix: the Bipartite Configuration Model (BiCM) and the null model

In order to assess the statistical significance of elements of the assist matrices, we resort to a null model for the bipartite matrices $\{\mathbf{M}^{\mathcal{C},\mathcal{T}}(y), \mathbf{M}^{\mathcal{C},\mathcal{P}}(y)\}$, built by randomly reshuffling their elements (*i.e.*, the network links connecting nodes in the layer $\mathcal{C}$ of countries with respectively nodes the layer $\mathcal{T}$ of technologies and $\mathcal{P}$ of products), but preserving country diversifications and activity (i.e., technologies or products) ubiquities in the two bipartite networks (*i.e.*, the degrees of the nodes in both layers of the bipartite networks countries-technologies and countries-products) . This means that we randomize the signal coming from the high order network connectivity patters, beyond that encoded by the node degree. In order to formulate the null model analytically, avoid relying on a conditional uniform graph test [14,15], degree constraints are imposed on average, in a formally similar way to what happens in the Canonical ensemble in Statistical Mechanics with the constraint of the energy [12]. This coincide to

 doi:10.20944/preprints201808.0081.v1

Peer-reviewed version available at Entropy 2019, 21, 126; doi:10.3390/e21020126

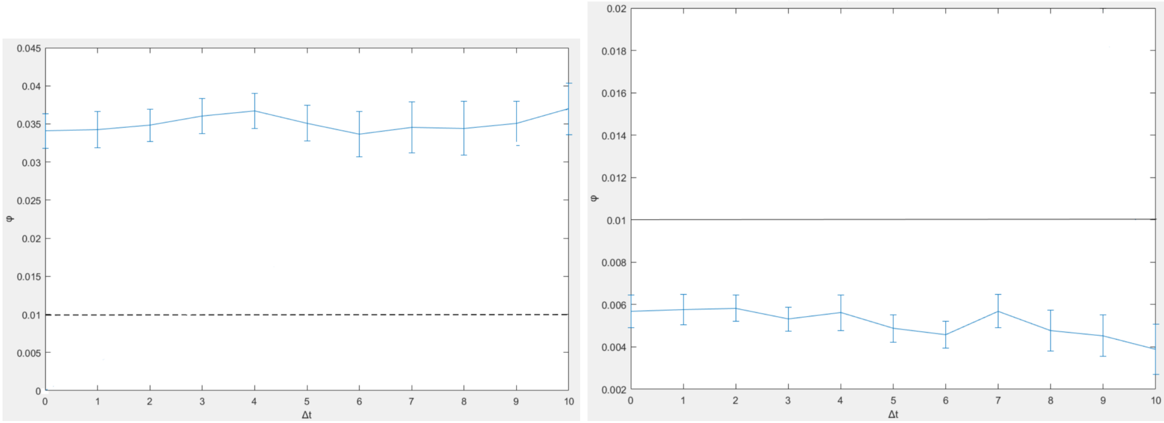

Figure 3. Mean signal ϕ of 5500 combinations of t, t' and p chosen for specific region of the assist matrix. Left panel: technological codes related to sector "chemistry" – C07 e C30 (*i.e.*, a much larger region than that represented in the bottom panel of Figure 2) and products in the region 2706-3104 *inorganic and organic chemicals, pharmaceuticals*. Right panel: technological codes related to sector "physics" – in the region H01P-H02M *electricity*, and products in the region 4411-5516 *textiles*. In all cases, error bars represent the standard deviation over the year pairs giving the same time lag, whereas, the dotted line is the significance level α.

set a null hypothesis described by the *Bipartite Configuration Model* (BiCM) [13], which is an extension of the *Configuration Model* [16], originally introduced by Park and Newman, to bipartite networks.

From a general point of view, the BiCM null model for any given (*i.e.*, real) binary biadjacency matrix $\mathbf{M}$, representing a bipartite network with two layers of node $a \in A$ and $b \in B$, is defined as the ensemble Ω of bipartite networks which are maximally random, apart from the ensemble average of the degrees of the nodes of both layers of the bipartite network which are constrained to appropriate values. As shown below, by a *maximal likelihood* argument, it is possible to show these appropriate values coincide with observed ones in the empirical network [17]: $\langle \tilde{k}_a \rangle_\Omega = k_a \; \forall a \in A$ and $\langle \tilde{k}_b \rangle_\Omega = k_b \; \forall b \in B$, where we have indicated with k the observed degrees in the real network and with $\tilde{k}$ the degrees in a generic configuration of the null model. We remind that $k_a = \sum_b M_{ab}$ and $k_b = \sum_a M_{ab}$ and analogously for "tilded" quantities. In the following, we use symbols with the tilde for quantities assessed on null model configurations, and without the tilde for observed values.

As we explicitly described below, such BiCM is operationally defined by two-steps: (i) by a constrained *maximum entropy* approach, we introduce the Exponential Random Bipartite Graph (ERBG) ensemble $\{\tilde{\mathbf{M}}\}$ with arbitrarily fixed mean values of the degrees of the nodes of the two layers; (ii) in order to choose among all possible ERBGs the best null model for our real bipartite network $\mathbf{M}$, we set the mean values of the node degrees equal to the values that they take in the real network $\mathbf{M}$ by a *maximum likelihood* argument.

Let us start by introducing the ERBG with fixed mean node degrees and let $\tilde{\mathbf{M}} \in \Omega$ be a network configuration in such ensemble and $P(\tilde{\mathbf{M}})$ be the probability of that graph. By implementing the prescriptions from Information Theory and Statistical Mechanics [12,18], the least biased choice of $P(\tilde{\mathbf{M}})$ is the one that maximizes the informational entropy

$$S = -\sum_{\tilde{\mathbf{M}} \in \Omega} P(\tilde{\mathbf{M}}) \ln P(\tilde{\mathbf{M}}), \tag{5}$$

subject to the normalization condition $\sum_{\tilde{\mathbf{M}} \in \Omega} P(\tilde{\mathbf{M}}) = 1$ plus the constraints:

$$\langle \tilde{k}_a \rangle_\Omega = \sum_{\tilde{\mathbf{M}} \in \Omega} P(\tilde{\mathbf{M}})\, \tilde{k}_a(\tilde{\mathbf{M}}) = k_a^* \quad \forall a \in A, \qquad \langle \tilde{k}_b \rangle_\Omega = \sum_{\tilde{\mathbf{M}} \in \Omega} P(\tilde{\mathbf{M}})\, \tilde{k}_b(\tilde{\mathbf{M}}) = k_b^* \quad \forall b \in B. \tag{6}$$

where $k_a^* \; \forall a \in A$ and $k_b^* \; \forall b \in B$ are arbitrarily fixed values for the mean degrees of nodes belonging to layers A and B. By defining the respective Lagrange multipliers ω, $\{\mu_a\}_{a\in A}$ and $\{\nu_b\}_{b\in B}$ (one for each node of the bipartite network), the probability distribution that maximizes the entropy satisfies satisfying at the same time the constraints, for all configurations $\tilde{\mathbf{M}} \in \Omega$ is given by the following variational equation:

$$0 = \frac{\delta}{\delta P(\tilde{\mathbf{M}})} \left[S + \omega \left(1 - \sum_{\tilde{\mathbf{M}}\in\Omega} P(\tilde{\mathbf{M}}) \right) + \right. \\ \left. + \sum_{a\in A} \mu_a \left(k_a^* - \sum_{\tilde{\mathbf{M}}\in\Omega} P(\tilde{\mathbf{M}})\, \tilde{k}_a(\tilde{\mathbf{M}}) \right) + \sum_{b\in B} \nu_b \left(k_b^* - \sum_{\tilde{\mathbf{M}}\in\Omega} P(\tilde{\mathbf{M}})\, \tilde{k}_b(\tilde{\mathbf{M}}) \right) \right]. \tag{7}$$

It is a matter of simple algebra to show that the solution of this equation is:

$$P(\tilde{\mathbf{M}} \,|\, \{\mu_a\}, \{\nu_b\}) = e^{-H(\tilde{\mathbf{M}} \,|\, \{\mu_a\},\{\nu_b\})} \Big/ Z(\{\mu_a\}, \{\nu_b\}), \tag{8}$$

where the function $H(\tilde{\mathbf{M}} \,|\, \{\mu_a\}, \{\nu_b\})$ is usually called the Hamiltonian of the graph configurations

$$H(\tilde{\mathbf{M}} \,|\, \{\mu_a\}, \{\nu_b\}) = \sum_{a\in A} \mu_a\, \tilde{k}_a(\tilde{\mathbf{M}}) + \sum_{b\in B} \nu_b\, \tilde{k}_b(\tilde{\mathbf{M}}), \tag{9}$$

and $Z(\{\mu_a\}, \{\nu_b\})$ is the corresponding partition function

$$Z(\{\mu_a\}, \{\nu_b\}) = e^{\omega+1} = \sum_{\tilde{\mathbf{M}}\in\Omega} e^{-H(\tilde{\mathbf{M}} \,|\, \{\mu_a\},\{\nu_b\})}. \tag{10}$$

The above Eqs. (8), (9) and (10) define the network ensemble known as the BiCM model.

Note that, as we have implemented only local constraints, i.e. the node degrees, Eq. (8) can be rewritten as the product of single link probability distributions over all pair of nodes belonging respectively to the two different layers [13]:

$$P(\tilde{\mathbf{M}} \,|\, \{\mu_a\}, \{\nu_b\}) = \prod_{a\in A} \prod_{b\in B} \pi_{ab}^{\tilde{M}_{ab}} \, (1 - \pi_{ab})^{\tilde{M}_{ab}}, \tag{11}$$

where π_{ab} is simply the probability of the link between nodes $a \in A$ and $b \in B$:

$$\pi_{ab} = \langle \tilde{M}_{ab} \rangle_\Omega = \sum_{\tilde{\mathbf{M}}\in\Omega} \tilde{M}_{ab}\, P(\tilde{\mathbf{M}} \,|\, \{\mu_a\}, \{\nu_b\}) = \frac{\eta_a\, \theta_b}{1 + \eta_a\, \theta_b} \tag{12}$$

with $\eta_a = e^{-\mu_a}$ and $\theta_b = e^{-\nu_b}$. In other words the existence of different links are independent events with respective probabilities related only to the Lagrange multipliers of the lateral nodes of the links. The values of the Lagrange multipliers are determined by the constraints Eqs. (6) which can be rewritten in terms of the derivatives of the partition function:

$$-\frac{\partial}{\partial \mu_a} \ln Z(\{\mu_a\}, \{\nu_b\}) \equiv \langle \tilde{k}_a \rangle_\Omega = k_a^* \quad \forall a \in A, \qquad -\frac{\partial}{\partial \nu_b} \ln Z(\{\mu_a\}, \{\nu_b\}) \equiv \langle \tilde{k}_b \rangle_\Omega = k_a^* \quad \forall b \in B. \tag{13}$$

Equations (8)-(12) define the generic EBRG with fixe mean degrees of nodes on both layers of the bipartite network.

At this point, in order to obtain the optimal null model for a given *real* biadjacency matrix (i.e. a bipartite network) $\mathbf{M}$ among all possible above defined EBRGs, we have to choose the best values for $\{k_a^*\}_{a\in A}$ and $\{k_b^*\}_{b\in B}$, or equivalently for the Lagrange multipliers $\{\mu_a\}_{a\in A}$ and $\{\nu_b\}_{b\in B}$, in relation to the connectivity properties of $\mathbf{M}$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2018 doi:10.20944/preprints201808.0081.v1

Peer-reviewed version available at Entropy 2019, 21, 126; doi:10.3390/e21020126

To this aim we write the log-likelihood function [17]

$$\mathcal{L}(\{\mu_a\},\{\nu_b\}) = \ln P(\mathbf{M}\,|\,\{\mu_a\},\{\nu_b\}) = \sum_{a\in A} k_a \ln \eta_a + \sum_{b\in B} k_b \ln \theta_b - \sum_{a\in A}\sum_{b\in B} \ln(1+\eta_a\,\theta_b), \tag{14}$$

where $P(\mathbf{M}\,|\,\{\mu_c\},\{\nu_a\})$ is the probability measure (11) evaluated for a configuration coinciding with the real network $\mathbf{M}$ and $\{k_a\}_{a\in A}$ and $\{k_b\}_{b\in B}$ are the node degrees of $\mathbf{M}$. The best values for $\{\eta_a\}_{a\in A}$ and $\{\theta_b\}_{b\in B}$ (or equivalently $\{\mu_a\}_{a\in A}$ and $\{\nu_b\}_{b\in B}$) are therefore obtained by maximizing such log-likelihood in these parameters. It is simple to show that this amounts in nsolving the system of $|A|+|B|$ equations in $|A|+|B|$ unknowns:

$$\begin{cases} \sum_{b\in B} \dfrac{\eta_a\,\theta_b}{1+\eta_a\,\theta_b} = k_a & \forall a \in A \\ \sum_{a\in A} \dfrac{\eta_a\,\theta_b}{1+\eta_a\,\theta_b} = k_b & \forall b \in B \end{cases} \tag{15}$$

which coincides to choose $k_a^* = k_a\ \forall a \in A$ and $k_b^* = k_b\ \forall b \in B$. Finally, the null model for a real bipartite network $\mathbf{M}$ is defined by Eqs. (11) and (12) with the Lagrange multipliers set by Eqs. (15). This recipe can therefore be applied to construct an appropriate null model for all empirical bipartite networks $\{\mathbf{M}^{\mathcal{C},\mathcal{T}}(y), \mathbf{M}^{\mathcal{C},\mathcal{P}}(y)\}_{y_{min}\le y\le y_{max}}$ obtained respectively by the databases Patstat and COMTRADE.

In order to build a null model for the assist matrices $\mathbf{B}^{\mathcal{T}\to\mathcal{P}}(y_1,y_2)$, and consequently for the Λ motifs defined by Eq. (4), obtained by Patstat and COMTRADE data at different years y_1 and y_2, we can now compose null models for the bipartite networks $\mathbf{M}^{\mathcal{C},\mathcal{T}}(y_1)$ and $\mathbf{M}^{\mathcal{C},\mathcal{P}}(y_2)$. This is done by contracting the two BiCMs for the matrices $\mathbf{M}^{\mathcal{C},\mathcal{T}}(y_1)$ and $\mathbf{M}^{\mathcal{C},\mathcal{P}}(y_2)$ along the country dimension [11, 19], as for Eq. (3). We have:

$$\tilde{B}_{tp}(y_1,y_2) = \sum_{c=1}^{N_c} \frac{\tilde{M}_{ct}(y_1)}{\tilde{k}_t(y_1)} \frac{\tilde{M}_{cp}(y_2)}{\tilde{k}_c^{(p)}(y_2)}, \tag{16}$$

where $\tilde{k}_t(y_1) = \sum_c \tilde{M}_{ct}(y_1)$ and $\tilde{k}_c^{(p)}(y_2) = \sum_p \tilde{M}_{cp}(y_2)$ are respectively the ubiquity of technology t and the product diversification of country c in the two single configurations for the BiCM null models for the two bipartite networks countries-technologies of year y_1 and countries-products of year y_2. In other words, starting by the two BiCM ensembles for $\mathbf{M}^{\mathcal{C},\mathcal{T}}(y_1)$ and $\mathbf{M}^{\mathcal{C},\mathcal{P}}(y_2)$ we build by composition an ensemble of configurations of bipartite networks $\Omega^{\mathcal{T}\to\mathcal{P}}(y_1,y_2)$ with link weights given by Eq. (16). The distributions of such $\tilde{B}_{tp}(y_1,y_2)$ values describing the null model can be in principle obtained using exact techniques [11,19]. However due to the non-Gaussianity of such distributions, we adopt a more practical sampling technique: starting from the BiCMs for $\mathbf{M}^{\mathcal{C},\mathcal{T}}(y_1)$ and $\mathbf{M}^{\mathcal{C},\mathcal{P}}(y_2)$, we use Eqs. (11), (12) and (16) to generate null Assist matrix bipartite networks, and populate the related ensemble $\Omega^{\mathcal{T}\to\mathcal{P}}(y_1,y_2)$ to estimate the full distributions. In a similar way, by using the composition Eq. (4) for the ensemble $\Omega^{\mathcal{T}\to\mathcal{P}}(y_1,y_2)$:

$$\tilde{\Lambda}_{tt'}^{p}(y_1,y_2) = \tilde{B}_{tp}(y_1,y_2)\tilde{B}_{t'p}(y_1,y_2) \tag{17}$$

and averaging over all pairs of years y_1 and y_2 with fixed delay $\Delta y = y_2 - y_1$ to get $\tilde{\Lambda}_{tp}(\Delta y)$, we can construct the null distribution of Λ motifs for each triple t, t', p and delay Δy.

The generic observed element $\Lambda_{tt'}^{p}(\Delta y)$ is then considered statistically significant depending on the p-value that we can infer from its distribution under the null hypothesis. The specific threshold for statistical significance and the size of the generated ensemble vary on the exercises performed (as highlighted in the text). In our case we fixed the statistical significance level at $\alpha = 0.01$. It is useful to recall that the two choices, the threshold and the size of the ensemble, are not unrelated: the higher the threshold we want to test, the bigger the sample we require. We consequently extracted for each couple of years y_1 and y_2 two ensemble of 1000 configurations/matrices for the two BiCM null models $\mathbf{M}^{\mathcal{C},\mathcal{T}}(y_1)$ and $\mathbf{M}^{\mathcal{C},\mathcal{P}}(y_2)$, and contracting one to one the configurations in the two ensembles, we get

1000 values of $\tilde{\Lambda}^{p}_{tt'}(y_1, y_2)$ to finally determine the statistical significance of the observed value $\Lambda^{p}_{tt'}(\Delta y)$ as explained in the Sect. 3.

Acknowledgement

This work was supported by the Italian PNR Project CRISIS-Lab. G.C. also acknowledge support from the EU project CoeGSS (676547).

1. C. Freeman, Technology policy and economic performance (1989).
2. P.M. Romer, Journal of political Economy **98**, 5.2 (1990).
3. P. Krugman, Journal of political economy, **87**, 2 (1979).
4. L. Soete, Output Measurement in Science and Technology (1980).
5. European Commission, (2009).
6. JC. Junker, (2015)
7. E. Pugliese, G. Cimini, A. Patelli, A. Zaccaria, L. Pietronero, and A Gabrielli, https://arxiv.org/abs/1707.05146 (2017).
8. A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, L. Pietronero, Scientific Reports **2**, 723 (2012).
9. M. Cristelli, A. Gabrielli, G. Caldarelli, A. Tacchella, L. Pietronero, PLoS ONE **8**(8), e7072 (2013).
10. H. Youn, D. Strumsky, L. M. Bettencourt, J. Lobo, Journal of The Royal Society Interface, 12(106), (2015).
11. F. Saracco, R. Di Clemente, A. Gabrielli, T. Squartini, Scientific Reports **6**, 30286 (2016).
12. E. T. Jaynes, Physical Review **106**(4), 620?630 (1957).
13. F. Saracco, R. Di Clemente, A. Gabrielli, T. Squartini, Scientific Reports **5**, 10595 (2015).
14. K. Zweig and M. Kaufmann, Social Network Analysis and Mining **1**, 187 (2011).
15. Z. Neal, Social Networks **39**, 84 (2014).
16. J. Park and M. E. J. Newman, Physical Review E **70**, 066117 (2004).
17. T. Squartini, G. Caldarelli, G. Cimini, A. Gabrielli, D. Garlaschelli, Physics Reports, doi.org/10.1016/j.physrep.2018.06.008 (2018).
18. K. Huang, Statistical Mechanics (Wiley, 1987).
19. S. Gualdi, G. Cimini, K. Primicerio, R. Di Clemente, D. Challet, Scientific Reports **6**, 39467 (2016).