DCICSS/2017/X

Dynamics of Cumulative Innovation in Complex Social Systems
DCICSS PROJECT

Technology Network

Directed network between patenting fields based on regional co-occurrence

Lorenzo Napolitano, Emanuele Pugliese
Institute for Complex System NRC (Italy), UOS Sapienza

February 4, 2018

1 Database

The analysis relies on the patent data contained in PATSTAT [3], a comprehensive database collecting information about applications filed at patent offices around the world. PATSTAT comprises several tables linking over fifty million patent applications (as of 2014) to information such as the applications' filing date, the patent families they belong to, and their technological content as described by the International Patent Classification (IPC) codes attributed by patent examiners to the claims contained in the documents. IPC codes define a hierarchical classification consisting of six levels (*sections, sub-sections, classes, sub-classes, groups, sub-groups*), including as little as 8 codes at the coarsest level (sections) and more than 70 thousand codes at the bottom of the tree.

We are also able to match patents to the location of their assignees (unfortunately not the inventors) through Orbis, a comprehensive database of firm level data maintained by the Bureau van Dijk. The Orbis data contain, for each firm active in patenting at a certain point in time, the list of patent applications granted to the company. There are other sources of patent data linking applications to assignees (and also inventors), the best known probably being the OECD's REGPAT database. The latter however features only for patent applications presented to the European Patent Office (EPO). ORBIS on the other hand does not select any particular patent office, so that the number of patent families we can include in the analysis is much greater through this channel than through REGPAT. In particular, due to the relatively high cost of the application to the EPO as compared to national patent authorities, the REGPAT sample is likely to select patents that are perceived as potentially high value by their assignees.

We match the technological codes associated to patent families taken from PATSTAT with the firm-level data so to unambiguously localize firms geographically through their country of residence and their zip code. This allows us to construct a spacial hierarchy by attributing each firm, thus its associated IPC codes, to an area on a geographical grid, which we define for a variety of disaggregation levels. We obtain the hierarchy by assigning postcodes to progressively coarser sub-national regions and, finally, to nations. Whenever

possible, we connect postal codes to the corresponding NUTS and *Local Administrative Unit* (LAU) regions for European Countries. Instead, when met with extra-European countries or European nations for which the postcode-NUTS correspondence is not available, we resort to national classifications or other on-line resources providing information about regional boundaries, which is built following in broad accordance with the definitions provided by the NUTS area and can hence assure the consistency of the geographical tree. Overall, our data define a spacial hierarchy comprising 39 countries worldwide and aggregating information about the location of around 500 thousand patenting firms.

The basic units of observation for the construction of the numerical matrices constituting the basis of our analysis, which we call *weighted geo-technological matrices* (wGTM), or $\mathbf{W}(t)$, are individual firms and the patents they own. We consider patent families as single inventions because of the strong relation between the documents included in such groupings. In constructing the matrices, we assume that each family that appears in year t counts as a unit and thus weighs accordingly within $\mathbf{W}(t)$. Moreover, we make the hypothesis that the technologies expressed within a patent family can be reasonably accounted for by considering the set of unique IPC codes they contain and that it is unnecessary to double-count the codes when they appear in more than one application belonging to the same patent family. Hence, for each family appearing in our dataset in a given year, we evenly split its unit of weight among all the technology codes and all the locations it maps to. With these caveats in mind, we therefore define the element $W_{cf}(t)$ of the matrix $\mathbf{W}(t)$ as the number of patent families, or shares thereof, in the field f filed by a corporate patentee located in region c.

In order to include the matrices in the analysis, we need to transform them into presence-absence matrices. In line with the literature [2], we assign a value of 1 to a location-technology pair if the value within the corresponding cell is compatible with a measure of revealed advantage. In particular, we use revealed comparative advantages [1] to produce a matrix $\mathbf{M}(t)$ starting from the corresponding $\mathbf{W}(t)$. $M_{cf}(t)$ is recorded as a presence, i.e. set equal to 1, if

$$\frac{W_{cf}}{\sum_c W_{cf}} > \frac{\sum_f W_{cf}}{\sum_{c,f} W_{cf}},$$

and an absence (i.e. $M_{cf} = 0$) otherwise.

2 Directed network

The aim is to measure the relationship between the patenting activity taking place within a geographical region in a technological field f at time t and the patenting activity performed in the same geographical region in a possibly different field f' at time $t + \delta$. We do so by counting how often patents in field f are present at time t in regions that produce patents in field f' at time $t + \delta$. Of course, we have to discount for regional diversification (d_c) – i.e. the number of fields in which region c is active – and the ubiquity of different fields (u_f) – i.e. the number of regions in which each field is represented – to establish a measure of the excess probability. In line with [7], with the same procedure provided in [5], we obtain:

$$B_{ff'}(t, \delta) = \frac{1}{u_f(t)} \sum_c \frac{M_{cf}(t) M_{cf'}(t + \delta)}{d_c(t + \delta)} \tag{1}$$

Notice that $B_{ff'}(t, \delta)$ can be represented as the probability that a country revealing a

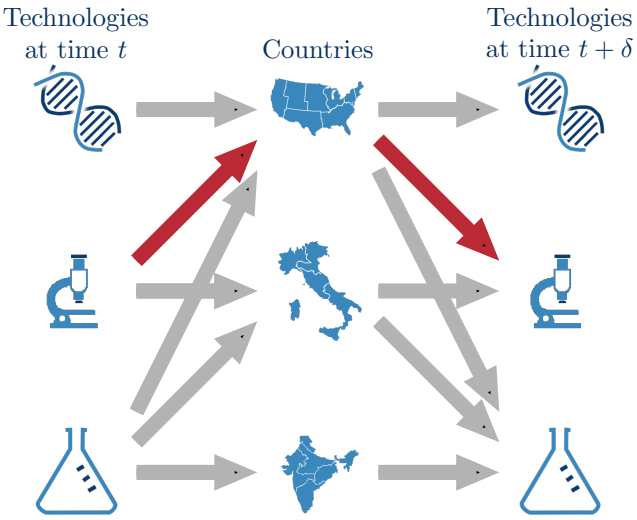


Figure 1: Model representation of the triple-layer technology-country-technology.

competitive advantage at time t in the field f will reveal a competitive advantage at time $t+\delta$ in the field f', i.e.

$$B_{ff'}(t,\delta) = Probability(f', t+\delta|f,t) = \sum_c Probability(f', t+\delta|c) Probability(c|f,t). \quad (2)$$

Of course, in equation 2 we assumed that the information about the capabilities linking pairs of technological fields is fully captured by their co-occurrence within each country, i.e. $Probability(f', t+\delta|c, f, t) = Probability(f', t+\delta|c)$.

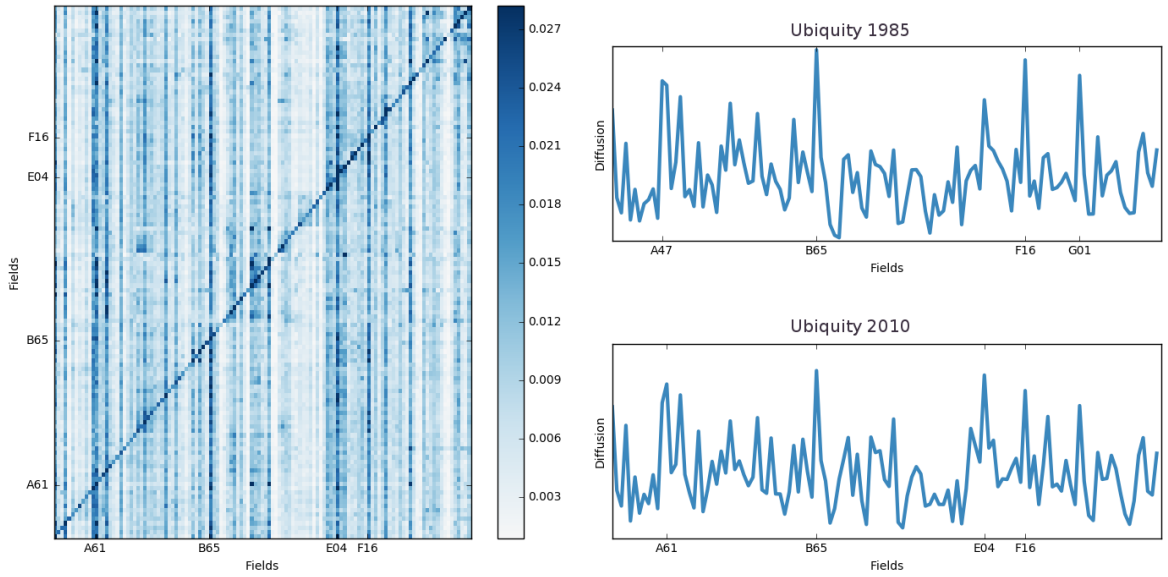
A further equivalent way of interpreting the same formula is obtained by constructing the tripartite directed network connecting (1) technological fields f at time t with (2) countries c and countries to (3) technological fields f' at time $t+\delta$, as in Figure 1. In this framework, $B_{ff'}(t,\delta)$ is equivalent to the probability that a random walk on the network starting from technology f reaches technology f' instead of a different one. This way, the interpretation as a technological spill-over is clearer.

Naturally the same formula can be applied to observe the system at different granularities at the technological level as well as the the product level and the geographical level.

An example of the results is visualized in Figure 2. It is interesting to note that,

- $\sum'_f B_{ff'}$ – *i.e.* the sum of the contributions *to* different fields from field f – is by definition equal to 1. This is an obvious consequence of 2: since $B_{ff'}$ is a probability distribution, the sum on all f' has to be 1.
- $D_f = \sum_f B_{ff'}$ – *i.e.* the sum of the contributions *from* different fields to field f' – is very heterogeneous. Equation 1 can be easily used to show that D_f is equal to the ubiquity of field f, which implies that the more a technological field is diffused, the more it has an impact on other fields.

Figure 2



A) Relation between patenting activity in technological field f (vertical axis) at time t (2010) and patenting activity in technological field f' (horizontal axis) at time $t+\delta$ (2011). The figure represents IPC classes (122 items) and has been computed by looking at technology concurrences at the province level (NUTS 3)

B) The sum of the effects of field f – *i.e.* its ubiquity – in 1985 and 2010. The general distribution is stable. The labeled sectors are the most diffused fields. *B65: Packing; Storing* and *F16: Engineering elements: general measures for producing and maintaining effective functioning of machines or installations* are among the most represented codes in both years.

3 Statistical Validation

While the measure is indeed normalized to be a probability, to assess if a link between two technologies is statistically significant a Null Model is required. The choices of the Null Model is not a trivial matter. Indeed some links could be very important because of the graph characteristics, without representing any real causal effect. For example, very advanced technological codes could be performed only by few regions in the World. On this basis, such codes could often concurre in the same regions, without any causal relationship. Following [5] we use as a Null Model a Bipartite Configuration Model [6]. This allows us to test our network against a random case that still has, in average, the same number of degrees: in the random graphs we generate each region has the proper expected diversification in term of technological codes and each technological codes has the proper expected ubiquity. The degrees is the only information we extract from the empirical matrices to generate the null models.

Generating many null matrices using the same null model, we can estabilish how significant is each link between technologies. This is done, and it allows us to generate a further matrix, $\mathbf{P}$, with the percentile of the null distribution in which the link t, t' fell. We can define therefore both the statistical significance of one link and the statistical significance of several aggregated indicators of the matrix.

In [4] we use these probabilities as a base for a Bonferroni-like method to validate at the same time all the edges.

4 Data Availability and Naming

The network, together with the scripts to replicate [4], is downloadable freely at: `http://people.bath.ac.uk/ee224/files/code_rsos.zip`

The network is freely usable for research puroposes. To acknowledge your use of the network, you can cite [4].

Each dataset is composed by three files:

- `labels_Y_G_T.tsv`
- `matrix_Y_G_T.tsv`
- `percentile_Y_G_T.tsv`

where `Y` is the year, `G` the geographical scale (see tables), `T` the technological scale (see tables). The first file lists the IPC codes corresponding to the fields in the rows and columns of the matrix, the second file presents the actual values of B in a *tab separated* format, the third file the percentile of the null distribution in which the corresponding element of B falls.

G	Geographical Scale
7	Countries
6	NUTS 1 - Macroregions
5	NUTS 2 - Regions
4	NUTS 3 - Provinces
3	Towns
2	ZIP codes
1	Firms

T	Technological aggregation
6	Section
5	Sub-sections
4	Classes
3	Sub-classes
2	Groups
1	Sub-Groups

References

[1] Bela Balassa. Trade liberalisation and "revealed" comparative advantage1. *The Manchester School*, 33(2):99–123, 1965.

[2] Sebastián Bustos, Charles Gomez, Ricardo Hausmann, and César A Hidalgo. The dynamics of nestedness predicts the evolution of industrial ecosystems. 2012.

[3] EPO. Epo worldwide patent statistical database - 2014 spring edition. Technical report, 2014.

[4] L Napolitano, E Evangelou, E Pugliese, P Zeppini, and Room G. Technology networks: the autocatalytic origins of innovation.

[5] Emanuele Pugliese, Giulio Cimini, Aurelio Patelli, Andrea Zaccaria, Luciano Pietronero, and Andrea Gabrielli. Unfolding the innovation system for the development of countries: co-evolution of science, technology and production. Technical report, arXiv preprint arXiv:1707.05146, 2017.

[6] Fabio Saracco, Riccardo Di Clemente, Andrea Gabrielli, and Tiziano Squartini. Randomizing bipartite networks: the case of the world trade web. *Scientific reports*, 5, 2015.

[7] Andrea Zaccaria, Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero. How the taxonomy of products drives the economic development of countries. *PloS one*, 9(12):e113770, 2014.

Dynamics of Cumulative Innovation in Complex Social Systems
DCICSS PROJECT

http://opus.bath.ac.uk/55142/9/DCICSS_project.pdf

- **Research Team at the University of Bath (UK)**
  - Professor Graham Room
    Professor of European Social Policy
    http://www.bath.ac.uk/sps/staff/graham-room/
  - Professor Alastair Spence
    Professor in Numerical Analysis
    http://www.bath.ac.uk/math-sci/contacts/academics/alastair-spence/
  - Dr Evangelos Evangelou
    Lecturer in Statistics
    http://www.bath.ac.uk/~ee224
  - Dr Paolo Zeppini
    Lecturer in Environmental Economics
    http://www.bath.ac.uk/economics/staff/paolo-zeppini/
- **Research Team at the Institute for Complex Systems National Research Council (Italy)**
  http://pilhd.phys.uniroma1.it/PILgroup_Economic_Complexity/Home.html
  - Dr Emanuele Pugliese
    Research Fellow
  - Dr Lorenzo Napolitano
    Research Fellow

- **Acknowledgements**
  We acknowledge gratefully the assistance kindly provided by PATSTAT at the European Patent Office, in supplying the datasets we are using for our patent case study.
  http://www.epo.org/searching-for-patents/business/patstat.html#tab1