Progetto di Tesi

Andrea Napoletano*
Supervisor: Prof. Luciano Pietronero

Sapienza, Universitá di Roma
Piazzale Aldo Moro 5, 00185, Roma

April 4, 2016

1 Introduction: The Economic Complexity Framework

The Economic Complexity framework consists in different powerful tools developed ad hoc to study nested bipartite network. As the name suggests, it was born to give a more quantitative and scientific analysis of the macroeconomic scenario by introducing a new indicator, the Fitness of a country, to complement its GDP [1]. The main new idea of this framework is to study the data of the WTO (World Trade Organization) and organize it into a bipartite network of countries and products exported. Standard economic theory prescribes that most developed countries specialize in high tech products. Instead what was found is that they tend to produce all available products in the market, so that the country-product matrix assumes a triangular shape. To better gasp this nested structure, Fitness of countries and Complexity of products have been introduced. Fitness and Complexity may be calculated in a self-contained way as fixed points of a non-linear map acting on the data themselves. In its original meaning, Fitness may be understood as the total diversity of products produced by a country weighted by their Complexity, while Complexity is roughly determined by the Fitness of the least fit country that can produce it. Adding the new fundamental dimension of Fitness allows to study countries in the two dimensional space spanned by GDP and Fitness and consequently recognizing multiple regions of different behaviour of countries. This powerful tool has been used and has already proved useful in giving a new insight of the macroeconomic scenario [2], [3].

More recently, the successful application of the Economic Complexity framework, inspired a series of works into different fields of research and strengthened the idea that the emergence of nested bipartite network is a transversal phenomenon common to all those situations in which species competes over a common bucket of resources. The first natural interdisciplinary application of this framework has been in biology. The Fitness and Complexity mechanism has proven itself to be very effective in classifying different species in ranking of importance with respect to the survivability of the ecosystem. Once all species available have been ranked, the size of cascade of extinction generated by removing one of them, followed the ranking predicted. Different algorithms, like Google Page Rank among all, have been tested together but they were all outperformed and, furthermore, the ranking provided by the Fitness and Complexity algorithm was close to optimal ranking [4].

My PhD thesis naturally comes in this work flow, it aims to extend the range of applicability of the Economic Complexity framework and to identify different situations which have as common underlying feature a (nested) bipartite network. Two different scenario are currently under active research: on the one hand I'm working on the technological data concerning the patents registered in the United States, on the other I'm part of an interdisciplinary team working on genetic data of cancer evolution. Since this last project has to fully start yet, I will just describe its main features in the last section, focusing now on the first project.

*andrea.napoletano1990@gmail.com

2 Patents and Technological Codes

The EPO Worldwide Patent Statistical Database[1] (PATSTAT) is periodically updated with data coming from all the world that record year by year the patents registered in each country. We decided to focus on the United States since its database is very rich and most innovations comes from there. Furthermore different countries may have different criteria of recording data and we wanted to avoid the problem of mixing them up. To be more specific, we have all the data starting from 1920 up to 2008. However for most of the analysis already done, we focused on the period 1990-2008 because there are more patents each year. They naturally constitute a bipartite network because given the group of patents and the group of codes only links between the two groups are allowed and there are no links intragroup. The aim of my work is to dig out from the data the signal of the development of new relevant inventions, so that, maybe, we would be able apply what we learned to forecast the imminent arrival of inventions yet to be made.

2.1 Weighted Network of Codes

The first step is to construct the network of codes. Starting from the bipartite network, we assert that two codes are related only if they occur in the same patent. Notice that this is the only time we make use of patents. In all other steps of the analysis, we disregard them and work only with codes. We repeat this operation every year to have all the networks of codes together. On each of them, we define weights for each link between codes in such a way to take into account the multiplicity of the patent that contains the code and of the code.

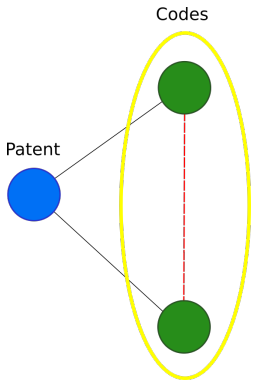

Figure 1: Be M_{pc} the binary matrix which yields 1 if the patent p has the code c or 0 otherwise. The strength of the link between codes c and c', as discussed in [2] is: $B_{cc'} = \frac{1}{\max(u_c,u_{c'})} \sum_p \frac{M_{pc}M_{pc'}}{d_p}$. Where u_c is the degree of the code c and d_p the degree of the patent p.

2.2 Community Detection

Once we posses the weighted networks of codes, on each of them we run a community detection algorithm. The one tested for now, is based on modularity maximization[2]. On average, this method identifies 50 communities per year, but the structure of communities is quite dynamical, going from 45 to 74.

[1] http://www.epo.org/searching/subscription/raw/product-14-24.html
[2] For a comprehensive treatment of community detection in graph, look at [5]

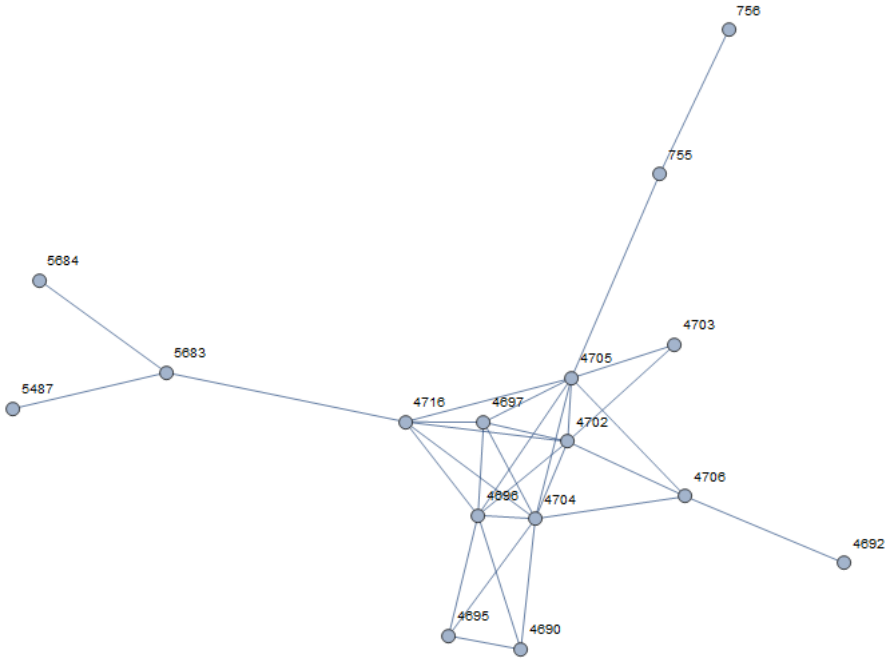

Figure 2: An example of a community detected by the algorithm, this is the community releted of whashing machines. In the table below some of the codes of this community with their description.

755	Washing or rinsing machines for crockery or table-ware
756	Apparatus or implements used in manual washing or cleaning of crockery, table-ware, cooking-ware or the like
4690	Washing machines having receptacles, stationary for washing purposes, with agitators therein contacting the articles being washed
...	...

This is a rather small well defined community. Our analysis shows that, however, there are a few large communities made by strongly connected sub-comminities.

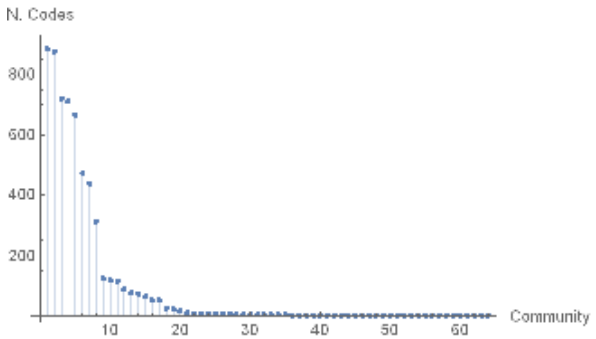

Figure 3: The distribution of communities. There are few huge macro-communities and more smaller better defined communities.

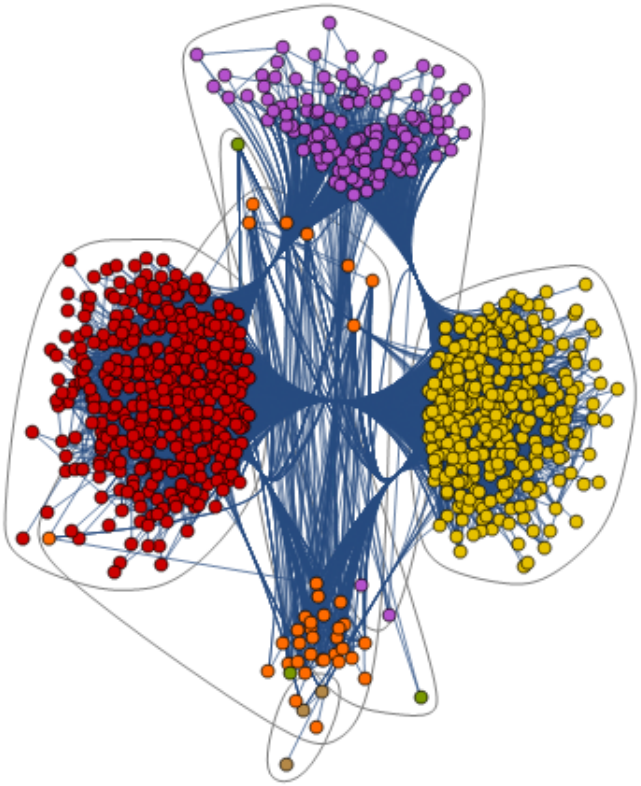

Figure 4: An example of how the biggest communities can be divided into sub-communities.

In Fig 4, we see how big communities can be divided into sub-communities and a deeper analysis actually shows that also those sub-communities are clusterized. This suggests that biggest communities are the result of merging of smaller well defined communities as the one showed in Fig 2. Our task is dual, on one side, we must identify the building blocks of larger communities, on the other side we want to follow the evolution of the structure of the network, see how these building block of codes are born or die, how they split and merge, shaping larger communities.

2.3 Evolution of Communities

To study the dynamic of the networks of codes, we must address a fundamental problem: how to identify communities in time. How to trace them and keep track of them. There is no canonical way to do that. As already discussed for example in [6], where a similar problem is faced, there are different function and measure of similarity one could introduce to decide whether two communities are related or not. If we want to trace the evolution of a community year by year, we find that there are two quantity of interest. One is the percentage of a certain community that is preserved and goes together as a block in the communities of the next year, the other is how much the old block constitutes of the new community. There are different ways one might think to use this information, for example, given a community one could simply look at all the communities it goes into and at how it contributes to them. If an appropriate threshold is not set, as showed in Fig 5 the tree branching out from a community becomes soon pretty big and quite hard to read, so we are are testing different settings according to the information we want to extract. Another possibility is to look how much of a community goes into how much of an another next year and multiply these two numbers. Upon renormalization, we get a number in the range $[0, 1]$ where 1 means that the community is perfectly preserved.

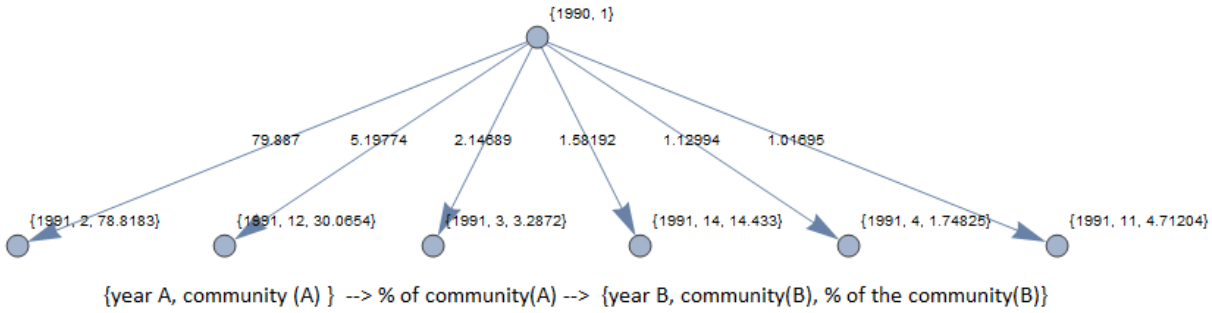

Figure 5: Here we see how the first community in year 1990 splits in different communities in 1991. The largest one is community 2. We see indeed that almost 80% of community 1 in 1990 goes into the 79% of community 2 in 1991. The rest gets split among the other communities.

We decided to set a threshold at 0.25 to see whether the community is preserved or not. In Fig 6 there is an example of the tree we can get. The current work in progress is to find criteria

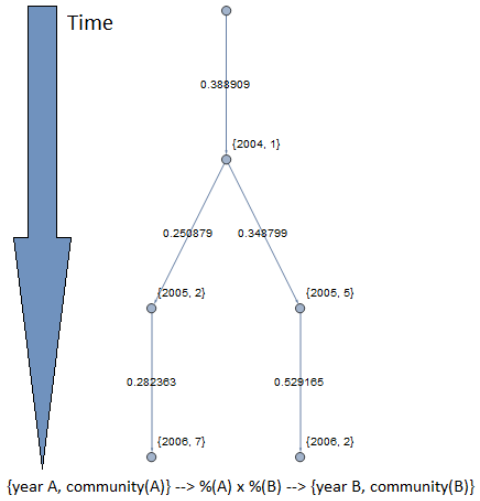

Figure 6: Here we see a window on the tree that start from community 1 in 1990 centred on an event of splitting. Having put a threshold, limits the number of branches.

of similarity most suited to study community inheritance and build all the trees. Our feeling is that, we may need different tools according to the scale at which we want to study the network of codes, be it macrocommunities or single links. Once that has been made, as is done in [6], we can focus on the event of splitting and merging of communities and building blocks, and study property as density and coherence to see if there are correlations and to seek for the footprints of innovations.

2.4 Patents and Technological Codes: Maching Learning Approach

By studying the network of codes, we want to find in its dynamic signs of innovations and inventions, so that we might be able to identify their precursors to great innovations, like a change in local and global properties of the communities[3]. A different and complementary approach on the other side would be to go back at patents and treat them as a context in which technological codes appear. The idea is to implement the same analysis done in [7]: represent

[3]Some of them discussed in [6] for example.

codes as vectors and by their occurrence in the same patent, trying to make them parallel. This system is frustrated because the algorithm tries to make two vector more parallel whenever they appear in the same patent, and does this operation for each couple of vector for each patent. The advantage of this approach with respect to the analysis of the network of codes is that instead of considering couples of codes, here we can consider all codes together and put a topology on the space of codes. We can then study the evolution of this topology and see how codes group together. We expect that major results should be stable and been seen with different methods. Google released a full library in python[4] suited for maching learning and for this kind of analysis, and we are actively working to make best use of it.

3 Statistical Physics of Cancer Evolution

An increasing amount of medical data has been made available on patients affected by cancer. With the advancement of microbiology and the introduction of new technologies and techniques for sequencing the DNA, recent years have seen an explosion in the amount of data that is now ready to researches on individual mutations, gene expression and epigenetic factors related to cancer. The main challenge one would like to address is how to integrate all these data into a coherent picture in which phenotypic traits of cancer cells emerge as the result of the interactions within genes and between pathways. A pathway is a portion of the gene regulatory network (GRN) which start from a particular gene and ends with all its possible products. A gene regulatory network (GRN) is a network composed by genes that act on the transcriptome[5] and regulate the gene expression. Therefore if a gene can affect another one through the proteins it codes, the two genes are related. Cancer is more and more seen as a disease of pathways, rather than a disease of single genes as originally believed, [8], and, within an interdisciplinary team of three work groups of genetists and physicists, I will work exactly on this idea, trying to study cancer as an emerging collective property of suitable networks of genes. One is the GRN just introduce, but there others, like, for example, the network of microRNAs acting on mRNAs as post-transcriptional suppressors of gene expression [9]. The microRNAs network act on the same products of the GRN, namely the transcriptome. The main difference is that the microRNAs network can only act as a suppressor, after a mRNA is coded by a gene, if a microRNA acts on it, it can only shut it down. So these two networks, although they share the same basket of products, are different on a fundamental level. On the one hand, in the GRN, one might expect to have nonlinear and feedback effects since the action on the mRNAs may either suppress or enhance the expression of a particular gene which then will have a different impact on the network, while on the other hand, the microRNAs network can only act as a further suppressive mechanism. Nevertheless we shuold consider both network together if we want to try to have a complete picture. Given the vastness of these networks, following the insight and the of our partners, we will focus on that part of the GRN and microRNAs network containing the Wnt and p53 pathways, because they are both extremely relevant in cancer formation cancer progression and cell senescence [10]. There are multiple ways leading to the same outcome: different tumors and different patients with the same tumor could correspond to different avenues in the network therefore we want to perform a systematic computational approach, integrated with experimental data gathered by the other work group, to fully depict and understand these two pathways (p53 and Wnt). The leading idea is to try to interpret cancer as a collective complex phenomenon emerging from the network and not as a result of local events. Roughly our program will consist on two tasks:

[4]TensorFlow https://www.tensorflow.org/

[5]The ensamble of mRNA to be assembled into proteins

- On the one hand to perform all the analysis on the two networks, starting from standard operation like measuring grade distribution, assortativity and centrality and quickly moving to a more ad hoc analysis inspired by the economic complexity framework.
- On the other to perform all the studies described above on two different samples, one of ill patients, and one on control group of sane patients. This way, we might be able gain some insight of cancer as an emerging phenomenon thanks to the comparative analysis.

References

[1] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, L. Pietronero; *A New Metrics for Countries' Fitness and Products' Complexity*, Scientific Reports 2, Article number: 723 doi:10.1038/srep00723 (2012)

[2] A. Zaccaria, M. Cristelli, A. Tacchella, L. Pietronero; *How the Taxonomy of Products Drives the Economic Development of Countries*, PLoS ONE 9(12): e113770. doi: 10.1371/journal.pone.0113770 (2014)

[3] M. Cristelli, A. Tacchella, L. Pietronero; *The Heterogeneous Dynamics of Economic Complexity*, PLoS ONE 10(2): e0117174. doi:10.1371/journal.pone.0117174 (2015)

[4] V. Domnguez-García, M. A. Munoz; *Ranking species in mutualistic networks*, Scientific Reports 5, Article number: 8182 doi:10.1038/srep08182 (2015)

[5] S. Fortunato; *Community detection in graphs*, Physics Reports, 486 75-174 (2010)

[6] D. Chavalaris, J.P. Cointet; *Phylomemetic Patterns in Science Evolution-The Rise and Fall of Scientific Fields*, PLoS ONE 8(2):e54847. doi:10.1371/journal.pone.0054847

[7] T. Mikolov, I. Sutskever, K. Chen G. Corrado, J. Dean; *Distributed Representations of Words and Phrases and their Compositionality*, Google Inc. Mountain view ()

[8] Y. Drier, M. Sheffer, E. Domany; *Pathway-based personalized analysis of cancer* PNAS 110(16):6388-93 (2013)

[9] B. N. Bossel, R. Avraham, M. Kedmi, A. Zeisel, A. Yitzhaky, Y. Yarden, E. Domany; *Context-specific microRNA analysis: identification of functional microRNAs and their mRNA targets*, Nucleic Acids Res. 40(21):10614-27 (2012)

[10] E. Caron et al; *A comprehensive map of the mTOR signaling network*, Mol Syst Biol. 6: 453. (2010)