Progetto Schematico di Tesi

Andrea Napoletano*
Supervisor: Prof. Luciano Pietronero

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

January 19, 2016

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 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 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.

*andrea.napoletano1990@gmail.com

1 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. The data come in the form of patents with associated their technological codes. 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. The sketch of how we propose to do that is the following:

- First we define and construct the network of codes by saying that two codes are linked together if they appear in the same patent. The strength of the link is a function of the number of such cooccurrencies.
- We then construct a randomized network of codes to which compare the real network.
- This allows us to calculate the Pvalues for the occurrence of couple of codes using the random network as a null model.
- Once we got the Pvalues, we can do different things: community detection in the real network by defining a threshold for the Pvalues , study of anomalous couples of codes to see weather they are of any significance...
- These operations are done for each year so that one may trace the evolutions of the communities year by year. The idea is that the realization of a new invention may be seen in the emergences of connections among codes that were not used together until that moment.
- Once all this have been done, we intend to generalise this study to triples or even in general n-ples of codes too look for even more detailed information than those provided by couples.

2 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[2]. Cancer is more and more seen as a disease of pathways, rather than a disease of genes as originally believed, [5], 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 bipartite networks of genes. There are two types of networks which are natural candidates for this study: one is the gene regulatory network

[1] http://www.epo.org/searching/subscription/raw/product-14-24.html

[2] A pathway is a portion of the gene regulatory network (GRN) which start from a particular gene and ends with all its possible products. GRN are introduced further in the text.

(GRN), where genes act on the transcriptome[3] and regulate the gene expression, the second is the network of microRNAs acting on mRNAs as post-transcriptional suppressors of gene expression [6]. These two networks, although they share the same basket of products, namely the transcriptome, 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 of our partners, we will focus on that part of the GRN and microRNAs network containing the Wnt and p53 pathways, they are both extremely relevant in cancer formation cancer progression and cell senescence, however how exactly their interaction works is not yet understood [7]. 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:

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

[6] 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)

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

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