IMT Institute for Advanced Studies, Lucca

Lucca, Italy

Essays on Economic and Social Complexity

PhD Program in Economics, Markets and Institution

XXVI Cycle

By

Riccardo Di Clemente

2014

The dissertation of Riccardo Di Clemente is approved.

Program Coordinator: Prof. Davide Ticchi,
IMT Institute for advanced studies Lucca

Supervisor: Prof. Luciano Pietronero,
ISC-CNR Instituto di Sistemi Complessi

Tutor: Prof. Massimo Riccaboni,
IMT Institute for advanced studies Lucca

The dissertation of Riccardo Di Clemente has been reviewed by:

Prof. Yi-Cheng Zhang,
University of Fribourg

Prof. Tiziana Di Matteo,
King's College London

IMT Institute for Advanced Studies, Lucca

2014

"Fatti non foste a viver come bruti
ma per seguir virtute e canoscenza"
Dante A. *Inferno* **XXVI** vv. 119-120

Contents

List of Figures

List of Tables

Acknowledgements

Chapter 2 is the reproduction, with some extensions, of the article "Statistical Agent Based Modelization of the Phenomenon of Drug Abuse", *(Nature) Scientific Reports* (2012) co-authored with **Prof. Luciano Pietronero** (ISC-CNR, Italy). It was two years work started during my master thesis and finished at the begin of my second year of doctorate. A special thank in this respect to **Riccardo Gatti**, director of **Prevo.Lab** Project for many interesting and important discussions. This work also received a great deal of attention from Italian press for the novelty of this research approach to forecast drug abuse.

Chapter 3 is the reproduction of the article "The Italian primary school-size distribution and the city-size: a complex nexus", *(Nature) Scientific Reports* (2014), presented with a poster to ***ECCS' 13***. It is a joint work with **Alessandro Belmonte** (IMTLucca, Italy) and co-authored with **Prof. Sergey V. Buldyrev** (Yeshiva University, USA). It was a two year work and I would like to special thank Alessandro for his contribution, dedication and friendship. I am really grateful to Prof. Sergey V. Buldyrev, which despite the geographical distance among us, believes in the project, supported it and stimulated our research. This work is already accepted to be presented with a talk to ***SigmaPhi2014*** *International Conference on Statistical Physics* (Rhodes, Greece 2014) and to ***ECCS'*** 14 *European Conference on Complex Systems* (Lucca, Italy 2014).

Chapter 4 is the reproduction of the article "Diversification versus specialization in complex ecosystems" submitted on 05/2014 to *PlosOne*. It is a joint work with **Guido Chiarotti**, **Matthieu Cristelli**, **Andrea Tacchella**, **Luciano Pietronero** (ISC-CNR, Italy). It was a one year and half work split be-

tween ISC-CNR and IMTLucca. I acknowledge in this respect CrisiLab @IMTLucca for the ®**Bloomberg Platform**.

I want to express my deepest gratitude to the reviewer of the manuscript. For giving me the time and the attention to review my thesis.

Collaboration & Support

First and foremost, I would like to express my sincere gratitude to **Prof. Luciano Pietronero** for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would like to express the deepest appreciation to **Prof. Massimo Riccaboni** to support and stimulate my research during my period at **IMTLucca**. Especially thank to his meticulous comments to my second research project during the **LIME group** meeting.

I am really grateful to **Prof. Eric Beihnocker** (INET@Oxford, UK) who gave me the possibility to spend nine months of my life in the beautiful city of Oxford, and being in contact with the group of **INET Economic-Complexity researchers**. I would like to thank him for the opportunity to work together and with **IPPr**, the Institute for Public Policy Research, on the challenging project to study and analyze the UK economic complexity.

I am indebted to my many of my IMTLucca colleagues to have supported me. Specially I would like to thank **Stefano Sebastio**, which his criticism stimulated me in drive my research though novel patterns, and **Olga Chiappinelli**, together with my **companion of classroom C**, has accompanied and supported my craziness during the three and half years of the

doctorate journey.

A special thank to my collaborators in Rome: **Matthieu Cristelli**, **Andrea Tacchella**, and all the **fellows of room 206 of PIL group** in the physics department at the University of Rome "Sapienza". They have created a place where it is possible to discuss and to share novel ideas with enthusiasm and passion.

I owe my deepest gratitude to my family and my girlfriend who have shown me unconditional love and support for this achievement.

Funding

1. *"Crisis-Lab"* PNR Project **@Instituto dei Sistemi Complessi ISC** CNR - UOS di Roma "Sapienza" to support my first publication.
2. *"Crisis-Lab"* PNR Project **@IMTLucca** to support my participation to Santa Fe Institute WorkShop: getting inside the black box: Technological Evolution and Economic Growth; and to have subscribed the ®Bloomberg platform.
3. *Dr. Bernard W. Gamson Computational Science Center* **@Yeshiva College** to support my second paper publication.
4. *"Talent At Work"* **EULifelongLearning Programme** Erasmus Consortia Placement to support my visiting period at **INET@Oxford** Institute for New Economic Thinking at the Oxford Martin School.
5. *GROWTHCOM "Growth and innovation policy-modelling: applying next generation tools, data, and economic complexity ideas"* (grant Agreement N. 611272) **@Instituto dei Sistemi Complessi ISC** CNR - UOS di Roma "Sapienza" to support my third publication submission, my last part of my Ph.D. and the support to the conferences.

Vita

Sept. 07, 1983 Born, Rome (Rome), Italy

Current Appointments

03/2011 Ph.D. Candidate in Economics, Market and Institution at **IMT-Lucca** Institute for Advanced Studies XXVI ciclo Supervised by: Prof. Prof. Luciano Pietronero and tutor: Prof. Massimo Riccaboni.

06/2013 Collaboration with **IPPr**, the Institute for Public Policy Research to analyze the UK economy and complexity.

04/2014 Researcher **@ISC-CNR** Institute of Complex System - UOS di Roma "Sapienza" (PIL group), funded with **FP7** *"GROWTHCOM"* (grant Agreement N. 611272)

Past appointments

2012-14 Affiliation to Instituto dei Sistemi Complessi **ISC-CNR** - UOS di Roma "Sapienza".

2012-14 Involvement in CNR-PNR National Project *"Crisis-Lab"* **@ISC-CNR**.

2010 Collaboration with a **Prevo.Lab** Project to elaborate a model to forth cast drug abuse and addiction

Research visit

2013-14 Visiting Scholar at **INET@Oxford**, Institute for New Economic Thinking at the Oxford Martin School supervised by Prof. Eric Beihnocker. (9 months)

Higher Education

2010 **M.Sc.** in **Theoretical Statistical Physics**, University of Rome "Sapienza".
Thesis: *"Statistical Agent Based Modelization of the Phenomenon of Drug Abuse"*.
Supervisor: Prof. Luciano Pietronero

2006 **B.Sc**. in **Physics**, University of Rome "Sapienza"
Thesis: *"Number of Nash equilibria of Minority game"*.
Supervisors: Prof. Miguel Angel Virasoro & Dr. Andrea De Martino

Ph.D. Schools

09/2013 Summer School of Mathematics for Economics and Social Sciences.(Information theory)
CRM - SNS Centro di Ricerca Matematica Ennio De Giorgi - Scuola Normale superiore (San Miniato, Italy).

08/2013 WorkShop: getting inside the black box: Technological Evolution and Economic Growth
Santa Fe Institute (Santa Fe, NM, USA).

04/2013 GOEX week on risk management
Universitá degli studi di Firenze (Firenze, Italy)

10/2012 FOC-CRISIS School on Complex Financial Networks
IMTLucca Institute for Advanced Studies (Lucca, Italy).

09/2012 Summer School of Mathematics for Economics and Social Sciences.(Bifurcation theory)
CRM - SNS Centro di Ricerca Matematica Ennio De Giorgi - Scuola Normale superiore (San Miniato, Italy).

06/2011 CIdE XXII Course of Econometrics for PhD students Summer School in Econometrics 2011.
CIdE - Centro Interuniversitario di Econometria (Bertinoro, Italy).

Certifications

Bloomberg Essential Training Program BESS
Commodity Essential, Equity Essential, Fixed Income Essential, FX Essential.

Department and University Service

2011-13 Representative of the IMTLucca Ph.D. students in the Executive Council of IMTLucca.

Honors

2013 *"Talent At Work"* EU Lifelong Learning Programme 9-months grant 2013/2014.

2013 Admission to the Summer School of Mathematics for Economics and Social Sciences2013 with tuition fee waiver and free lodging.

2012 Admission to the Summer School of Mathematics for Economics and Social Sciences2012 with tuition fee waiver and free lodging.

2011 Admission to the Ph.D. program in Economics, Markets and Institution relations at IMT Institute for Advanced Studies Lucca, with 3-years full scholarship, tuition fee waiver and free lodging.

Press Coverage

20/08/2012 The italian world news agency ANSA and some italian journals (e.g. Corriere della Sera, La Stampa) had published a news on the relevance of my paper *"Statistical Agent Based Modelization of the Phenomenon of Drug Abuse,* ***DOI:10.1038/srep00532***" for studying drugs Abuse phenomena. Read the ANSA (in italian).[1]

[1] http://www.ansa.it/web/notizie/rubriche/scienza/2012/08/20/-Scienza-complessita-contro-droga_7359345.html

Publications

1. Di Clemente, R., Chiarotti, G., Cristelli, M., Tacchella, A. & Pietronero, L. Diversification versus specialization in complex ecosystems. (Submitted 05/2014 on *PlosOne*)

2. Belmonte, A., Di Clemente, R. & Buldyrev, S. V. The italian primary school-size distribution and the city-size: a complex nexus *(Nature) Scientific Reports* **4**, 5301 (2014) **DOI :10.1038/srep05301**

3. Di Clemente, R. & Pietronero, L., Statistical Agent Based Modelization of the Phenomenon of Drug Abuse *(Nature) Scientific Reports* **2**, 532 (2012) **DOI :10.1038/srep00532**

Conference

Poster

1. Di Clemente, R., Belmonte, A., & Buldyrev, S. V. A Complexity Approach To The Italian Primary School-Size Distribution ***ECCS'*** 13 *Book of Abstract: European Conference on Complex Systems* (Barcelona, Spain 2013)

Talks

1. Di Clemente, R., Belmonte, A., & Buldyrev, S. V. On the spatial distribution of the Italian primary school-size ***SigmaPhi2014*** *International Conference on Statistical Physics* (Rhodes, Greece 2014).

2. Di Clemente, R., Belmonte, A., & Buldyrev, S. V. On the spatial distribution of the Italian primary school-size ***ECCS'*** 14 *European Conference on Complex Systems* (Lucca, Italy 2014).

Invited Talk

1. 06/12/2013 **UCL**, London, Statistical Agent Based Modelization of the Phenomenon of Drug Abuse.

Abstract

The multidisciplinary approach to problem solving involves drawing appropriately from different viewpoints to redefine problems outside of normal boundaries and reach solutions based on a new understanding of complex situations. Social and economics science have always borrowed and embraced tools and instruments from mathematics and physics to develop their theories. Historically a real multidisciplinary methodology to economic and social issues has been neglected by the academic researchers due to a widespread gap in their formal approach. Recently a new interdisciplinary framework has been developed connecting together social and economics theories with the complex systems analysis; this approach reveals new conceptual prospectives and methodologies thanks to its multiple-level viewpoint, which are able to disclose novel challenges and problems. This thesis collects three different multidisciplinary approaches to social and economic behaviors formalized with the complex systems tools.

Chapter 2

We introduce a statistical agent based model to describe the phenomenon of drug abuse and its dynamical evolution at the individual and global level. The agents are heterogeneous with respect to their intrinsic inclination to drugs, to their budget attitude and social environment. The various levels of drug use were inspired by the professional description of the phenomenon and this permits a direct comparison with all available data. We show that certain elements have a great importance to start the use of drugs, for example the rare

events in the personal experiences which permit to overcome the barrier of drug use occasionally. The analysis of how the system reacts to perturbations is very important to understand its key elements and it provides strategies for effective policy making. The present model represents the first step of a realistic description of this phenomenon and can be easily generalized in various directions.

Chapter 3

We characterize the statistical law according to which Italian primary school-size distributes. We find that the school-size can be approximated by a log-normal distribution, with a fat lower tail that collects a large number of very small schools. The upper tail of the school-size distribution decreases exponentially and the growth rates are distributed with a Laplace PDF. These distributions are similar to those observed for firms and are consistent with a Bose-Einstein preferential attachment process. The body of the distribution features a bimodal shape suggesting some source of heterogeneity in the school organization that we uncover by an in-depth analysis of the relation between schools-size and city-size. We propose a novel cluster methodology and a new spatial interaction approach among schools which outline the variety of policies implemented in Italy. Different regional policies are also discussed shedding lights on the relation between policy and geographical features.

Chapter 4

By analyzing the distribution of revenues across the production sectors of quoted firms we suggest a novel dimension that drives the firms diversification process at country level.

Data show a non trivial macro regional clustering of the diversification process, which underlines the relevance of geopolitical environments in determining the microscopic dynamics of economic entities. These findings demonstrate the possibility of singling out in complex ecosystems those microfeatures that emerge at macro-levels, which could be of particular relevance for decision-makers in selecting the appropriate parameters to be acted upon in order to achieve desirable results. The understanding of this micro-macro information exchange is further deepened through the introduction of a simplified dynamic model.

Chapter 1

Introduction

1.1 Complex Systems apply to Economic & Social Science

Over the past twenty years the multidisciplinary approach of complex systems in the physical and social-economic sciences has been developed. It has led to new conceptual perspectives and methodologies that are of value not only to researchers but also to professionals and policy-makers [1]. Complex systems approaches enable researchers to study aspects of the real world in which events and actions have multiple causes and consequences, and where order and structure coexist at many different scales of time and space. The specificity of complex systems, generally under investigated or simply not addressed by traditional science, resides in the emergence of non-trivial superstructures that often dominate the system's behavior and cannot be easily traced back to the properties of the constituent entities [1]. The interdisciplinary approach of Complex Systems can raise universal questions that can be expressed across a broad spectrum of disciplines from biology to computer networks to human societies [2]. The methods to disentangle these questions also belong to different disciplines which spread between computer science, mathematics and physics. This research approach can overcome

the standard methods in specialized domains which rarely takes in consideration the multiple-level viewpoint. The complexity portrays economy and social systems not as deterministic, predictable, and mechanistic, but as process dependent, organic, and always evolving mechanism [3]. Statistical Physics in the Complexity framework has then evolved towards the application to specific problems, many of them in interdisciplinary areas. The explosion of the complex network analysis has shown that many phenomena, even those not closely related to the original physical area, can be represented with these tools. The importance of this and other applications now has to be evaluated with respect to the impact these ideas and methods have in other fields where there are applied. This situation poses new challenges and problems that have been discussed in various articles and editorials [4, 5, 6]. This thesis is the collection of three different complex systems approaches to the socio-economic systems introduced in the following paragraphs.

1.2 Introduction to chapter 2

This chapter (reproduction of [7]) consists in the development of an Agent Based Model (ABM) for the phenomenon of drug abuse. It is an interdisciplinary piece of work in which the concepts and methods of statistical physics are applied to investigate the socially relevant phenomenon of drug abuse.

The model we developed is part of the class of the Agent Based Models (ABMs), which have recently been used in many socio-economic areas [8, 9]. The relation of ABM models with statistical physics and social behaviors is clear because one attempts to describe the competition between interaction and noise, the heterogeneity of the agents, the origin of large fluctuations and the spontaneous development of critical situations. The available literature presents only highly simplified and preliminary approaches on the drug abuse subject [10, 11] and we believe our model provides a basis for a scientific approach to this problem, its understating and possibly its optimal control. Our ABM has a much

higher level of complexity and realism and it provides information on all elements of the phenomenon both at the individual and global level. To this end, it is important to note that our main conclusions could not have been achieved from these previous models.

The model proposed is directly related to the concepts and the parameters used by professionals in the field at the international level and it makes optimal use of all the available information. This allows for a direct comparison to the present and future data and can be easily generalized to explain more complex and realistic environment. At the present level, the model permits to identify the crucial, most important parameters, and as well as those that play a minor role. In this respect one, of the critical parameters to overcoming the natural barrier to start drug consumption, for the first time, is the tail fluctuations of personal experience (environment and interactions). On the contrary the economic barrier plays a relatively minor role with an appreciable importance only at the beginning of the consumption. Our model permits to track the individual history of an agent and in turn, this information could be directly compared with medical and other data. We stress our model to analyze the response of the consumption to external changes of some parameters (consumers behaviors, environment and price). All this informations can help to optimize the control, to define a suitable policy, and to lead to a basic understanding of the complexity of the phenomenon.

1.3 Introduction to chapter 3

This chapter (reproduction of [12]) investigates the statistical distribution of the sizes of Italian primary schools. It is an interdisciplinary work in which the concepts and methods of complex systems are applied to investigate the Italian school distribution according to geographical features and population density. Many scholars have shown surprisingly complex features governing cities, firms, or social groups, uncovering power law distributions, fat tails and other properties incomprehensible in the scope of orthodox approaches. This framework proposes new challenges in this respect and problems in studying socio-economic sys-

tems that have been discussed in various articles in the most important journals, such as Nature, Science, and American Economic Review [13, 14, 15].

We introduce a definition of school-size as the number of students currently enrolled in each school and we show that the PDF can be well approximated by a lognormal distribution. School system shares certain elements with firms and cities [16, 17, 18] such as too many small schools in the left tail of the size distribution comparatively to its left tail which exponentially decays. But it also has specific and important puzzles: the body of the PDF distribution manifests a bimodality characteristic. The evidence of the bimodality underlies the interplay between different processes that define thresholds and boundaries that are very peculiar for the Italian primary school-size distribution.

Motivated by the absence of any territorial constraint in school choice, and despite the fact that the fat left tail of the school-size distribution has been particularly targeted in the past years by political interest and legislative attempts [19, 20], there is no evidence in the rising of the size of the schools in the lower quantiles. To disentangle the bimodality source we introduce a measure of the average spatial interaction intensity between a school and the surrounding ones in different Italian regions. We find that interactions are very weak, on average, for small schools, especially for countryside-based regions. This pattern involves small villages, with only one school, whose closeness coincides with the proximity of schools.

Our conclusions indicate that the bimodality of the Italian primary school-size distribution is very likely to be due to a mixture prevalently driven by the population density and, in turn, by the geographical features of the territory. This empirical work is directly aimed to address policy schooling decision, providing a first formal complex system approach to develop a better education system. Our analysis can be easily generalized to other countries and many other complex systems such as hospitals and other facilities that are strongly related to the city-size.

1.4 Introduction to chapter 4

There is a growing literature on benefit of diversification at firm level in emerging as well as developed economies [21, 22, 23]. The effects of institutions, policies and economic environments under which diversification has an impact on firm performances have been extensively studied [23, 24, 25]. In some cases diversification in less developed economies has been claimed to be related to the difficulty of stipulating effective contracts at firms level. On the other hand, in economies with well established capital markets, diversification may have limited value due to the fact that the institutional context enables smaller, specialized firms to raise capital. Mainstream literature tends to relate firms diversification to such economic statements. It has also been suggested [26] that diversification can originate from group affiliation strategies used to create internal market.

In the present work we propose a different perspective according to which the market maturity plays a secondary role with respect to entrepreneurial culture. In particular we show empirical evidences of a clear difference between western (Anglo-Saxon) markets and the rest of the world, independently of the degree of development of the market itself.

In the framework of economic Complexity first [27] and then [28] introduce a new metrics for comparing the competitiveness of countries through a fitness model based on countries' products complexity. This analysis take in consideration the binary export matrix M built thanks to the Revealed Comparative Advantage (RCA) [29].

The rearranged M matrix, ordering rows and columns with respect to the fitness and complexity metrics shows a almost-triangularity shape. It appears that the higher specialization is not a suitable strategy for countries. Analyzing the distribution of the M, we can see that the poorest countries are those which are specialized in the export of few products. On the contrary, the more developed countries are the ones with the more diversification of exports.

Within the Economic Complexity framework we analyzed the firm

differentiation process. In fact Countries and firms are fundamental actors sharing complex economic and social ecosystems, but while firms are specialized entities, countries display diversified features which have been demonstrated to be crucial in understanding their competitiveness. This raises a question on the mechanisms driving specialized entities to organize themselves into diversified super-structures which are a quite ubiquitous question in natural and social sciences. Coherently with the evidence of a triangular structure of country-product matrix in [27, 28, 30, 31], in the present analysis the same triangular feature is also found in the country-sector matrix obtained by aggregating firms on the basis of its legal address. In the particular case of firm diversification we identify in this chapter a novel country-dependent micro signature, the revenue diversification barrier, which we indicate to be the factor governing the diversification dynamics.

In this chapter (reproduction of [32]) we analyze the Bloomberg data of the distribution of revenues across production sectors of quoted firm aggregated by country. This distribution manifests a peculiar (country dependent) triangular shape, in which a clear lower boundary appears, indicating that firm diversification is positively correlated with revenues: to increment its diversification a firm must increase its incomes by a minimum critical amount which is country-dependent. We design a simple mathematical model to shed new light on these dynamics. The model suggests that the diversification process develops over time and depends on the competitive environment in which the firms are embedded.

The country dependence of the firm diversification barrier shows a non trivial geographic clustering, possibly reflecting different cultural and entrepreneurial patterns. Our findings show that within the framework of economic complexity it is possible to identify the mechanism responsible for micro-macro information exchange.

Chapter 2

Statistical Agent Based Modelization of the Phenomenon of Drug Abuse

2.1 Introduction

The ambition to develop a quantitative description of people's behavior and introduce novel ideas and methods in socio-economic disciplines [4, 5, 6] (see also the future ict project www.futurict.eu) is one of the main challenges of statistical physics and complexity theory. Agent based models (ABMs) [8, 9] represent a broad framework to address these questions. They can describe some of the most important properties that get inspiration from physical phenomena. These are for example the importance of heterogeneity, large (critical) fluctuations, self-organized criticality and the lack of cause-effect relation. The the key features of ABM is the suitable choice of the nature of the agents, their interactions and their dynamical evolution.

In this chapter, we present an Agent Based Model which aims to

describe, at a reasonably microscopic level, the phenomenon of drug abuse, its evolution and control. The scheme of the model was inspired by the participation the authors to the national initiative PREVO.LAB [33, 34, 35] whose objective is to analyze and control the phenomenon of drug abuse in Italy. Therefore, the model is closely aligned with the professional analysis in this field and permits a direct comparison of concepts and parameters with those actually observed and analyzed in reality. In this respect, the model is rather realistic and suitable for direct applications, including control, forecast of the phenomenon and for policymaking. The information in this field is very scattered and ranges from highly accurate information about people who become hospitalized to the little known detail about those at the beginning of the process. The model aim is to provide a complete framework to describe the phenomenon, whose parameters are fixed by the best known facts. In this way, the model is able to extrapolate the knowledge of the less known submerged elements. This Agent Based Model, with its parameters fixed by real observations, exhibits the following features:

- The tail fluctuations of personal experience (environment and interactions) are critical components for overcoming the natural barrier to start drug consumption, for the first time. This also implies that the heterogeneous nature of the social-network is very important and cannot be represented by an average situation.
- The economic barrier plays a relatively minor role with an appreciable importance only at the beginning.
- The model permits to track the individual history of an agent and in turn, this information could be directly compared with medical and other data.
- One can analyze the response of the system to external changes of some parameters. This can help optimizing the control and defining a suitable policy.
- The model is flexible and can be easily improved by introducing specific elements that may be inspired by new observations. For

example, it could be implemented in a specific social complex network [36] and one can also introduce different drugs with different market socioeconomic characteristics, related for example to the age, the economic background or the gender.

In summary, we introduce an ABM with variables and parameters defined in a way that makes efficient use of all data and information available in the field. This permits us to achieve a complete description of the phenomenon both at the individual and global level, and to identify its crucial elements and its responsiveness to changes in any of its parameters.

The study of drug consumption and trend predictions began in the '70s with the analysis of trend historical data of New York City [37]. In the '90s Everingham and Rydell [38] proposed a Markovian process to analyze and predict cocaine consumption in the USA. The first ABM model introduced was *Drugmart* by Agar and Wilson [10, 39], revised by the UK Office of Science and Technology report [40], which is a simple network model. Another more complex model is the Australian *SimDrug* [11] which simulates the consumption of heroin in Melbourne with accurate people's dynamic reconstruction.

With respect to these preliminary models, our ABM has a much higher level of complexity and realism and it provides information on all elements of the phenomenon both at the individual and global level. To this end, it is important to note that our main conclusions could not have been derived from these previous models.

2.2 Results

2.2.1 The Model

In our model, an agent i, with $i \in [1; N]$, is characterized by a number of parameters and interacts with the environment and with the other agents at each time $t \in [1; T]$. In the dynamic the unit time step will correspond to one week. We now briefly introduce the key elements of the model, which are described in more detail in the section 2.4.

Age $y_i(t) = [10 : 100]$ years. The starting distribution of the agents' age at t_0 is taken from the Italian data of ISTAT. In a first stage, we consider the property of a static situation $y_i(t) = y_i(t_0)$. Later, we also introduce the dynamic evolution of the system by increasing the agents' age with time.

Consumers type $S_i(t) = 0; 1; 2; 3; 4$ defines the consumption stage and the level of dependence reached by a consumer: non users ($S_i = 0$); mini user with age $y_i(t) < 26$ ($S_i = 1$); occasional user ($S_i = 2$); frequent user not pathological ($S_i = 3$); heavy user pathological ($S_i = 4$). These categories are inspired by the professional classification of drug abuse from "Diagnostic and Statistical Manual of Mental Disorders" [41]. The mini user with $y_i < 26$ are agents with low income that can be in contact with the normal dose and with a lower quantity of drug, the mini dose. This mini dose, as we will see later, was introduced by drug cartel with the purpose of stimulating in the young consumers a possible future dependency on drug [33, 34].

Personal budget and saving behavior. Behavioral analysis [34] has shown that the money a person would spend on initial drug consumption is rather related to personal saving attitude for leisure opportunities, than to the total personal budget. Therefore, the only variation we make in the budget $m_i(y_i(t))$ refers to the difference between the adult working population with age $y_i(t) \geq 26$ with $m_i(y_i(t)) = 1$, and young unemployed or student population with age $y_i(t) < 26$ with $m_i(y_i(t)) = 0.2$. The threshold of age 26 represents the average age of the change of spending and work habits [34]. The crucial point will be the parameter $\gamma_i \in [0; \infty]$ (saving behavior), which describes the tendency of an agent to save or spend money. Inspired by various indicators of social behavior, we assign to it a lognormal distribution [42] with variance Γ_σ and average Γ_μ. A value of $\gamma_i \approx 0$ means that the agent is inclined to spend his leisure budget while $\gamma_i \geq 1$ corresponds to a strong willingness to save money and do not use it for buying drugs.

Personal exposure e_i describes every possible contact between the agent and the drug world, in particular the input the agent receives about drug consumption. In the present model we adopt a simplified model

of interaction among agents but the generalization to a complex social network would be a natural extension in the model [36, 43, 44].

Inclination to drug β_i controls the intrinsic (genetic, educational, cultural, etc.) agent's barrier towards drugs. Also in this case, inspired by the social studies [45], we adopt lognormal distribution with $\sigma = B_\sigma, \mu = B_\mu$. For every agent we define the barrier to be overcome in order to start using drugs: $b_i^*(t) = \beta_i/\alpha(y_i(t))$; where $\alpha(y_i(t))$ is a function dependent on the agent's age and it has the characteristic to increase the barrier for older agents. This function is described by a simple analytical form which reproduces the data of Prevo.Lab [33].

Drug addiction: describes how an agent becomes addicted to drugs and how this modifies her behavior. This property is defined by the function $d_i(\pi_i(t), S_i(t))$ which can increase or decrease $b_i^*(t)$, depending on the persistence of the agent in a given stage, where $\pi_i(t)$ is an exponential function of consecutive drug assumption. The inclusion of this effect redefines the drug dependence function: $b_i(t) = b_i^*(t) - d_i(\pi_i(t), S_i(t))$.

In the drug market there are two different dose levels: mini-dose and dose. The first has a lower quantity and a cheaper price. It is used for attracting young people to drug consumption while the second represents the normal dose for adults and for young usual users. The mini-dose is very important for the analysis of drug consumption because it is often the way in which the young generations start the drug use. Moreover it represents the cheapest and easiest way of transmission of drug among young people [33]. In our model the mini-dose could symbolize a light drug that can lead to the intake of the normal one. The mini-dose price $p_m(t)$ is supposed to be substantially smaller than a dose. We have chosen an lower magnitude order as reasonable estimation of the mini dose price which is about one tenth of the dose [34]. The price of the dose $p_d(t)$ is based on a simplified relation between supply and demand. In our case we assume that the total supply is fixed so the price is simply due to the total number of users.

$$p_d(t) = \frac{\sum_{k=1}^{N} \delta(S_k(t)=1,2,3,4)}{N} \qquad p_m(t) = \frac{p_d(t)}{10} \tag{2.1}$$

This relation could be easily made more realistic with a variable level of

total drug amount.

The agents interact and evolve through the comparison of the two major barriers: the *Economic Barrier* "EB". and the *Socio-Emotional Barrier* "SEB". The first barrier compares the price of the drug with the buying capacity and it is defined by:

$$\begin{cases} \frac{p_m(t)}{m_i(y_i(t))} < 1 - \gamma_i & \text{if} \quad y_i(t) < 26 \\ \frac{p_d(t)}{m_i(y_i(t))} < 1 - \gamma_i & \text{if} \quad y_i(t) \geq 26 \end{cases} \tag{2.2}$$

The second barrier has the role of describing the social opportunities and the emotional process involved in drug consumption. The social opportunities are defined in terms of the events by which the agents are subjected during a time unit. We describe this *"daily noise"* by $r_i(t)$, random number drawn each time for each agent from a Gaussian distribution, $R_\mu = 0$, R_σ. Our choice to use the Gaussian distribution is due in large part to the lack of quantitative information that is available. It was also chosen because of the symmetric shape that could lead to a daily event that could generate a positive or negative effect in the drug intake. When more information becomes available, it will be easy to modify accordingly. The condition to overcame the SEB is defined by: $b_i(t) + r_i(t) < e_i(t)$.

We have now all the elements that define the dynamics and the interaction between the agents. In summary the most important parameters of the model are γ_i, $r_i(t)$ and β_i which are respectively money saving behavior, daily noise and intrinsic inclination to drug.

We can define the transition probability of each agent from the stage $S_i(t)$ to $S_i(t+1)$ as indicated in Figure 2.1a. The corresponding rates are described in detail in the section 2.4.2.

2.2.2 Tuning the parameters

The first point we consider is to tune the model parameters to a realistic description of the observed data. A crucial parameter in this respect is the observed fraction of people using drugs. In order to address this question we have made several simulations with $N = 1000$ and $T = 1000$

(the time unit t is one week). The strategy is to fix two distributions and examine the change of drug using population during the variation of the third one. Then we repeat the operation for all three distributions. In Figure 2.1b-c-d we see the change depending on the three distributions. We choose to have in our population about 10% of drug users (a realistic value considering all different drugs [33, 46, 47], which could be adjusted when better data is available). When studying the real drug consumption, in addition to these total values, it is important to split the agents in four age ranges [33] labeled as: young $(15-25)$, young-adult $(26-35)$, adult $(36-45)$ and senior $(45-100)$. Also, the individual percentage of users within the four age ranges is used to fix our parameters. These procedures define the three distributions, which we name *"Usual Distributions"*, UD:

$$\begin{cases} \beta_i & \dashrightarrow \text{ lognormal}(B_\mu = 1.2, B_\sigma = 2.7) \\ \gamma_i & \dashrightarrow \text{ lognormal}(\Gamma_\mu = 0.6, \Gamma_\sigma = 0.9) \\ r_i(t) & \dashrightarrow \text{ gaussian}(R_\mu = 0, R_\sigma = 0.6) \end{cases} \tag{2.3}$$

Therefore, these optimized distributions provide a complete and realistic parametrization of our model which can now be used to investigate a variety of problems. We start by considering the individual personal history per agent.

2.2.3 Individual agent history

It is interesting to note that the individual history can be vastly different in reality and we are going to observe that our ABM is able to reproduce a dynamic variety of agent's history. The Table 2.1 shows $\Delta t = 68$ consecutive iterations of eight users, and it outlines some of the different users typology that can be produced by the model:

- $\#64$ is a senior agent that never tested drugs;
- $\#141$ is a young agent that, as an occasional consumer, tries the normal dose;
- $\#1$(young-adult) and $\#343$(young) are two agents that are occasional consumers;

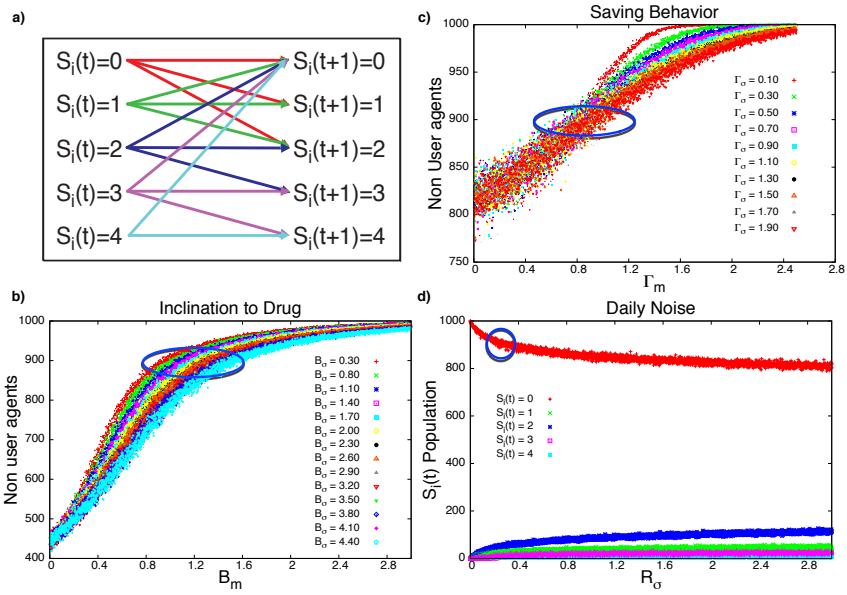

Figure 2.1: Stage transitions and parameter settings. a. All possible stage transitions for an agent at stage $S_i(t)$. **b. c. d.** Simulations intended to select the main parameter of the model in relation to the realistic stage of the phenomenon (see text for more details). In particular in **b.** we show the effect of the parameter β_i (inclination to drug) distribution $(B_m; B_\sigma)$, in **c.** effect of a parameter γ_i (saving behavior) distribution $(\Gamma_m; \Gamma_\sigma)$ and in **d.** r_i (daily noise) distribution $(R_m = 0; R_\sigma)$. The blue circles identify the realistic areas to fix the values of the three distributions so that they correspond to a total of 10% of users. This parameter optimization refers also to the age ranges discussed in the text.

- $\#15$(young-adult) and $\#130$(young) are both heavy consumers that evolve into addicted;
- $\#45$(adult) and $\#494$(young) are two agents that try drugs on extremely rare occasions.

The fact that our model reproduces a realistic heterogeneity of agents' history is a very important element. This realistic variability is in fact crucial to have an accurate description of the phenomenon that could

#Agent	Age	Personal History
1	26	00020000000020000000000000000000000000020002202000020202000000000020002
15	29	22202002000000020022333333333333333333333333333333344444444444444444400
45	36	00000000000000002000000000000000000000000000000000000000000000000000000
64	68	00000000000000000000000000000000000000000000000000000000000000000000000
130	16	33333333012333333330101233333333333333333333333334444444444444444444
141	20	11100101010110111100001001111100011101000101010012000110111111110101
343	23	01101110010000100010101001110100011010101001101101001000000110101111
494	25	00000000000000000000001000000000000000000000100000000000000000010000

Table 2.1: Individual agent histories. a. The table shows the histories of eight agents for $\Delta t = 68$ consecutive times (One time unit corresponds to one week). One can see that our model is able to reproduce a variety of a realistic heterogeneity in agents' history. This history ranges: from the agent $\#64$, who never tried drugs, to agents characterized as heavy users $\#15; \#130$.

be compared to a considerable amount of real data. A possible upgrade would be analyzing and contrasting the real drug user clinical histories. By means of the above we could provide a probabilistic prediction of detoxification options for each clinical case. In order to gain more insight in the individual evolution, we can observe the trend of the barrier for inclination to drugs use, the Social Emotional Barrier (SEB) and the stage reached by an individual user i at anytime t. The Figure 2.2a shows the history of a young user $y_i = 19$ with a low intrinsic inclination barrier β_i, for $\Delta t = 300$ consecutive iterations. The user stage $S_i(t)$ is indicated in red dots, the left term of SEB inequality barrier in green, and in blue the starting value of inclination to drug $b_i^*(t)$. We can see that this agent develops all the stages with important fluctuations and finally ends up being an addict. At this level the green barrier line declines strongly and finally we can also observe that the agent detoxifies. In the case above the effect of aging is not yet included $b_i^*(t) = b_i^*(t_0)$.

Now we consider a complete dynamic evolution in which aging is explicitly included. We set the year length $\tau = 50t$, in which one iteration corresponds to a real week time spend. We elaborate a set of rules to define the death of an agent due to the aging and the birth of the new one in order to have a constant number of agents N (sec. 2.4.3). We also

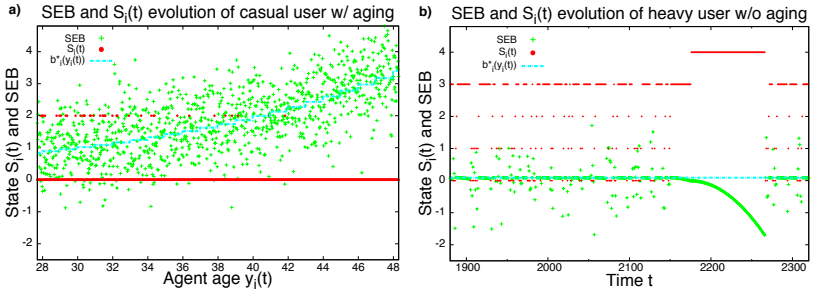

Figure 2.2: Personal agent history. a. Personal agent history without aging: the agent becomes a heavy user because after various fluctuations she reaches the stage $S_i(t) = 4$. The green line describes the left term of Social Emotional Barrier and one can see that this barrier is drastically lower when the stage of dependence is reached. **b.** Personal agent history with aging. In this case the agent is a casual user, meaning that agent has frequently used drugs at a younger age. Then the barrier of inclination to drug use (blue line) $b_i^*(y_i(t))$ increases with the age. We can see that, in this way, the use of drugs $S_i(t) = 2$ becomes more and more rare. Finally when the agent gets older the use of drug is completely eliminated. This example shows that the model provides a detailed description of individual histories, which could be directly compared to real data.

define a simple process, from the data [11, 35, 46], which describes the possibility that an agent can die of overdose and be replaced by a new one. In Fig. 2.2b we reproduce the personal history of an agent in the aging process. The figure shows the increase of the barrier to $b_i^*(t)$ due to the agent's aging. The agent starts with a high value of β_i (low tendency to use drug) but the daily noise induces to try several drugs $(S_i(t) = 2)$. However, when the agent grows up, $b_i^*(t)$ rises, so the daily noise becomes less effective and finally the agent no longer uses drugs. This example shows that the possibility to monitor a single agent's history can be an important tool to compare with real medical data. Once again the future analysis could study the probability of stage change and compare it with real data from detoxification centers.

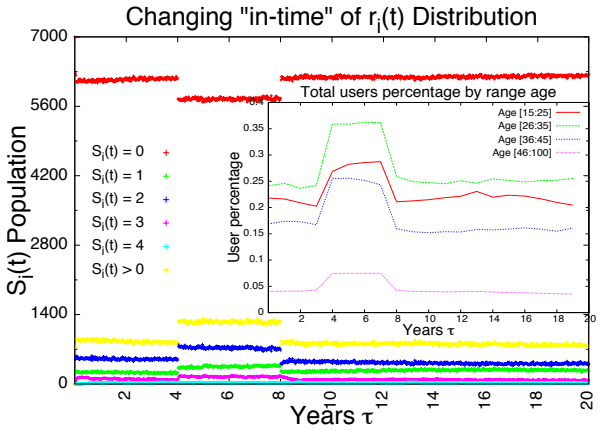

Figure 2.3: Fluctuation of the life style. Here we test how the system reacts to increased fluctuations for the rare events (tail of the $r_i(t)$ distribution) that induce the starting of abuse. During the period $4\tau < t < 8\tau$ the distribution of daily noise increased ($R_m = 0.1; R_\sigma = 1.3$), the various ranges of population are affected by this change. In the insert we show how the total number of drug users is distributed among the various age ranges. From these studies, one can conclude that the system is strongly sensitive to this tail fluctuations, which represents a very important point for the control of this phenomenon.

2.2.4 Global property and effects of perturbations

Our model also allows us to examine how the users population reacts to various sources stress, ranging from a different opportunity to be involved in drug use (r_i), a new agent's intrinsic education (β_i) or a change in saving behavior (γ_i). With the purpose of analyzing the agent's dynamic under these social modifications, we fixed two of the UD and analyzed the effect of changing *in-time* of the third one. We will activate the *in-time* modification between the time steps, $t^+ < t < 2t^+$.

In Figure 2.3 we change the daily noise distribution from the UD to $R_\mu = 0.3, R_\sigma = 1.6$, with $t^+ = 4\tau$. The Figure shows how an increase of noise leads to an increase in the number of agents in the drug stage. As it is shown from the insert of Figure 2.3 the $r_i(t)$ distribution has different

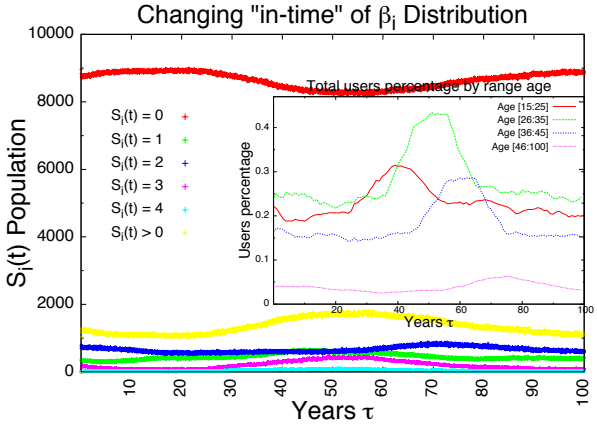

Figure 2.4: Cultural fluctuations. Here we show the effect of the changes in the genetic cultural barrier of inclination to drugs. We change the distribution of inclination to drug β_i for the new agent during the period $20\tau < t < 40\tau$ with $(B_m = 1.2; B_\sigma = 2.7)$. The lower intrinsic barrier leads to a creation of a new users generation more inclined to drugs abuse. We can see that the effect is not instantaneous and we can follow how it develops over the years. Also in this case, as we show in the insert, one can see the large difference of age effect between the various age ranges. The pattern revealed can be directly compared to real data and is important for monitoring and controlling the phenomenon.

impact for each age range, given the dependency of $b_i^*(t)$ on age.

Figure 2.4 exhibits the change of the population stages $S_i(t)$ due to the modification of the intrinsic inclination toward drug use distribution $(B_\mu = 1.2, B_\sigma = 2.7)$ for the new born agent between $20\tau < t < 40\tau$. The agents who are born during the perturbation periods are more inclined to use drugs than the others. The insert of Figure 4 shows the trends between the age ranges and how the aging of the new agents increases temporally the number of drug users.

In Figure 2.5 we change the tendency of agent spending behavior. We replace the distribution of γ_i from the UD with $(\Gamma_m = 0.1; \Gamma_\sigma = 0.3)$ for the new born between $20\tau < t < 40\tau$. During this perturbation period

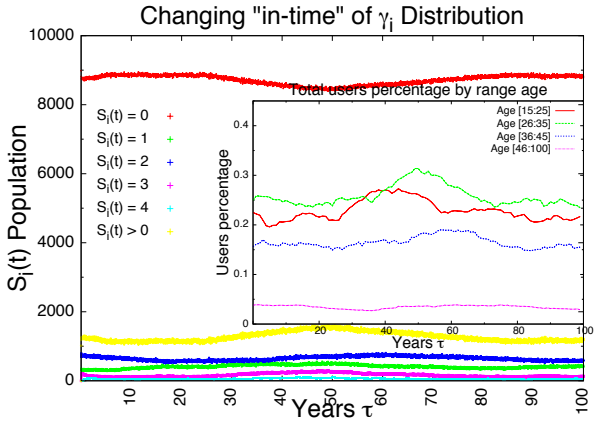

Figure 2.5: Fluctuations of saving money behavior. Here we examine the effect of the change of the saving behavior. We change the distribution of γ_i for the new agent during the period $20\tau < t < 40\tau$ with $(\Gamma_m = 0.1; \Gamma_\sigma = 0.3)$. A lower value of γ_i means that an agent tends to spend money easily. In particular a value near to zero for an agent means that she has no problems for drug purchase. We can see that the effect is similar to the cultural fluctuation but it is of minor impact. The effect of spending capacity decreases with the age and in particular it does not affect the senior agents.

the new born agents are more inclined to spend money for leisure, in this case drug. One can see from the insert of the picture the different trend among age ranges. According to the analysis, the saving behavior does not appear to have an important role in explaining drug assumption. The decrease in-time the γ_i distribution produces a limited impact in the number of users compared to the effect β_i.

Finally we focus on the role that mini users have on the introduction of drug among the new generations. This is an important point because it is hard to monitor and as an appreciable effect only after a long time. In Figure 2.6, we compare two situations of a population evolution. For the first 15τ years the simulation does not consider the mini-dose. Consecutively we included the mini user. We can see that the existence of mini user stage leads after many years to a large increase in the total number

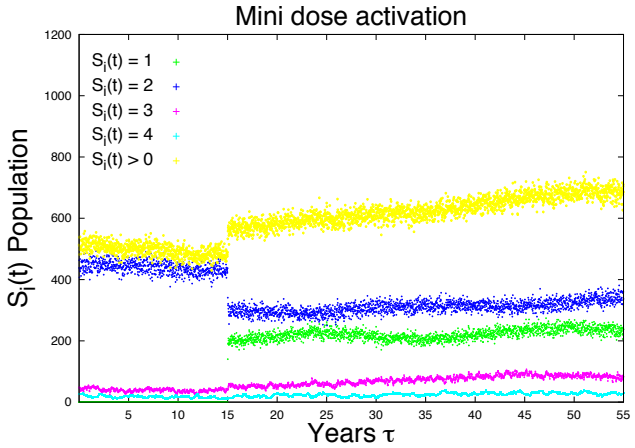

Figure 2.6: Mini dose activation. We consider a population of $N = 7000$ agents. At the beginning the market is without the mini dose, at a time $t = 15\tau$ the mini dose is introduced in the system for young people. We only show the drug user stages and one can see an immediate change of the global users followed by a further increase of addict population. The results show that the effect of mini dose for young generation represents and extremely important element for the dynamics of the system.

of drug users.

2.3 Discussion

In conclusion, we have developed an agent based model for the phenomena of drug use, which is relatively simple, but at the same time rather realistic. Its concept and parameters have been established in close relation with the professional studies of the real phenomenon. The model permits the study of the problem at the individual and global level. The individual histories of the agents can be extremely diverse and appear to provide a realistic description of the real situation which could be easily compared to medical data directly. At the global level we can identify the crucial most important parameters, as well as those that play a minor role. This leads to a basic understanding of the complexity of the phe-

nomenon and to the possibility of analyzing how the situation changes if the parameters are modified. This point is essential in defining the strategies for controlling and reducing the phenomenon.

One of the main conclusions of our studies is the fact that drug use is usually trigged by a rare event in the personal experience. This shows the importance of the tail of these heterogeneous experiences, which allows us to predict that the introduction of additional heterogeneity in the structure of the society will lead to important effects. In particular, our model's interactions among agents are assumed to be similar on average. Therefore we expect that the introduction of a complex network distribution of this iteration will be relevant. In this perspective we can already predict that the hubs of this distribution will be very important for policymaking.

The model is meant to represent a realistic basis of analysis and discussion, and can be easily generalized with a complex network of social interactions and with multiple drugs with different social characteristics. We believe that these studies can lead to a higher level of understanding of these phenomenona and can be useful for a more effective policy making.

2.4 Methods

2.4.1 Model parameters specification

Following we explain in detail the parameters that characterize the agents.

Personal exposure

The personal exposure e_i, describes every possible contact between the agent and the drug world and in particular describes how agents feel about drugs in general and their consumption in particular. It is defined by:

$$e_i(t) = \frac{\left[\sum_{k=1}^{N} \delta_{(S_k(t)=1,2,3,4)}\right] - \delta_{(S_i(t)=1,2,3,4)}}{N-1} \tag{2.4}$$

This expression simply counts the fraction of drug users within the entire population, but does not include the single agent under observation. The latter represents a type of mean field interaction, which could be easily generalized to complex network situations [36, 43, 44].

Age function

The function $\alpha(y_i(t))$ is dependent on the agent's age and its purpose is to increase the barrier for older agents:

$$\alpha(y_i(t)) = \frac{\exp\left(\frac{-(-c+\lg(y_i(t)))^2}{2d^2}\right)}{ad} \tag{2.5}$$

The values of $a = 1; d = 0.5; c = 3$ are chosen in order to outline the importance of age in the consumption of drugs [34]. This expression is a simple mathematical representation of how age influences the probability of drug assumption Fig. 2.7a. The main characteristic of this function is that the peack of this function is located between sixteen and twenty-four years and then it slowly decreases, as it has been shown by the studies of Prevo.Lab. As we can see from the Figure 2.7b the $\alpha(y_i(t))$ increases the average value of the b_i^* due to age behavior.

Drug addiction

The drug addiction describes how an agent becomes addicted to drugs. This property is defined by the function dependence $d_i(\pi_i(t), S_i(t))$ which has the purpose to decrease the barrier $b_i^*(t)$. The function $\pi_i(t)$ *"stage permanence"* is an experience counter:

$$\begin{cases} \pi_i(t) = 1 & \text{if} \quad S_i(t) \neq S_i(t-1) \\ \pi_i(t) = \pi_i(t-1) + 1 & \text{if} \quad S_i(t) = S_i(t-1) \end{cases} \tag{2.6}$$

and the function dependence is:

$$d_i(\pi_i(t), S_i(t)) = l(t)(\exp(\pi_i(t)S_i(t) * \chi(z)) - 1) + d_i(\pi_i(t-1), S_i(t-1)) \tag{2.7}$$

This expression represents an exponential growth of the addiction parameter with the persistence of the certain drug user stage. The term

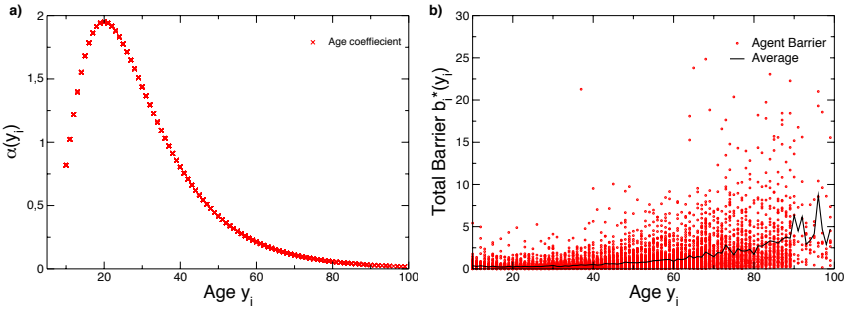

Figure 2.7: Studies of barrier age dependence. a. The function age coefficient $(\alpha(y_i))$ is described by a simple analytical form which is a reproduction of Prevo.Lab's data. **b.** This is an example of a distribution of the intrinsic agent barrier with age dependence $b_i^*(y_i)$ for a simulation of 10000 agents.

$\chi(z)$ represents the *drug coefficent*, which describes how addictive the drug of type z is. In this regard, it would be easy to introduce drugs of different types. For the present study we consider a single case in which $\chi(z) = .0001$. This parameter could be changed for increasing or decreasing the effect of the drug addiction. The parameter $l_i(t)$ is the *first experience coefficient*, which describes the quality of the agent's first experience in using the given drug:

$$\begin{cases} l_i(t) = -2 & \text{if} \quad k_i(t) > 0,75 \;\&\; S_i(t) < 3 \\ l_i(t) = 1 & \text{Otherwise} \end{cases} \tag{2.8}$$

where $k_i(t)$ is a random number with uniform distribution between $[0 : 1]$. The two possible values of $l_i(t)$ are chosen to give a mathematical representation of the social hypothesis, that a negative experience has more psychological impact than an equivalent positive one [48]. We can redefine the barrier of inclination to drugs with:

$$b_i(t) = b_i^*(t) - d_i(\pi_i(t), S_i(t)) \tag{2.9}$$

in order to outline the role of dependence and first experience in the drug consumption and addiction.

2.4.2 Rules for the stage transitions

We define the rules that enable agents to have stage transitions from stage $S_i(t)$ to $S_i(t+1)$. The Figure 2.1a in the chapter shows all the possible transition stages. Each transition has some different condition due to the initial stage $S_i(t)$.

From $S_i(t) = 0$

From the stage $S_i(t) = 0$ we can have stage transitions into:

- $S_i(t+1) = 1$ if the EB and SEB barrier are verified for an agent with age $y_i(t) < 26$;
- $S_i(t+1) = 2$ if the EB and SEB barrier are verified for an agent with age $y_i(t) \geq 26$;
- otherwise the agent remains in the stage of non user $S_i(t) = 0$.

As we can see from Eq.2.10 the difference between the final stage reached by the agent is dependent only on age:

$$\begin{cases} S_i(t+1) = 1 & \text{if} \quad [y_i < 26 \;\&\; \frac{p_m(t)}{m_i(y_i)} < 1 - \gamma_i \;\&\; b_i(t) + r_i(t) < e_i(t)] \\ S_i(t+1) = 2 & \text{if} \quad [y_i \geq 26 \;\&\; \frac{p_d(t)}{m_i(y_i)} < 1 - \gamma_i \;\&\; b_i(t) + r_i(t) < e_i(t)] \\ S_i(t+1) = 0 & \qquad\qquad\qquad\qquad \text{Otherwise} \end{cases} \tag{2.10}$$

From $S_i(t) = 1$

Only an agent with $y_i(t) < 26$ can reach this stage. The possible transitions as we show in Eq.2.11 are:

- Becomes non-user, if EB or SEB are not verified;
- Persists in the stage of mini-user if the agent gets the sufficient daily noise (opportunity) and has a budget;
- Change in $S_i(t+1) = 2$ if the agent has funds to purchase the full dose and the SEB is verified without need of the daily noise.

The possibility for an agent to get in contact with the full dose is dependent on personal behavior associated with dependence level rather than the daily occasions.

$$\begin{cases} S_i(t+1) = 2 & \text{if} \quad [p_d(t)m_i(y_i) < 1-\gamma_i \quad \& \quad b_i(t) < e_i(t)] \\ S_i(t+1) = 0 & \text{if} \quad [p_m(t)m_i(y_i) > 1-\gamma_i \quad \text{or} \quad b_i(t) + r_i(t) > e_i(t)] \\ S_i(t+1) = 1 & \text{Otherwise} \end{cases} \tag{2.11}$$

An agent becomes $S_i(t+1) = 2$ when the function dependence and the first experience coefficient decreases the SEB.

From $S_i(t) = 2$

In this stage the agent is an occasional user of normal dose, the agent can:

- Return to non-user stage, if EB or SEB are not verified;
- Stay in $S_i(t+1) = 2$ if both conditions are verified;
- Become a frequent user if the agent's inclination barrier to drug is reduced by the dependence (SEB without daily noise).

The following equation shows the transition possibilities:

$$\begin{cases} S_i(t+1) = 0 & \text{if} \quad [\frac{p_d(t)}{m_i(y_i)} > 1-\gamma_i \quad \text{or} \quad b_i(t) + r_i(t) > e_i(t)] \\ S_i(t+1) = 3 & \text{if} \quad [\frac{p_d(t)}{m_i(y_i)} < 1-\gamma_i \quad \& \quad b_i(t) < e_i(t)] \\ S_i(t+1) = 2 & \text{Otherwise} \end{cases} \tag{2.12}$$

As we can see, the transition rules are quite similar to the previous stage. We have considered that the stage $S_i(t) = 2$ is reachable by all agents without being contingent on age.

From $S_i(t) = 3$

The agent is a frequent user. She does not care about the EB because the addiction raises the necessity of a dose and the money becomes a negligible constraint. She can become:

- Heavy user, pathological, if her inclination barrier to drug decreases by a factor 5 or more;
- She can decide to stop using drugs if the social environment changes drastically, or with the 10% probability, which represents the possibility to stop using drugs due to exogenous environmental factors.
- Otherwise the agent i continues to take drug as a frequent user.

Here we show the transition rules:

$$\begin{cases} S_i(t+1) = 4 & \text{if} \quad 5b_i(t) < e_i(t) \\ S_i(t+1) = 0 & \text{if} \quad [b_i(t) > e_i(t) \text{ or } c_i(t) > 0.9] \\ S_i(t+1) = 3 & \quad \text{Otherwise} \end{cases} \tag{2.13}$$

Where $c_i(t)$ is a random number from uniform distribution $[0:1]$. There is no more the daily noise $r_i(t)$ and the transitions depend on the variation of $b_i(t)$ due to the dependence function and the stage persistence. The factor 5 of the first condition is based on our decision [33, 34, 35] to discriminate in a significant way the difference between the two stages $S_i(t) = 3$ and $S_i(t+1) = 4$. The $b_i(t) > e_i(t)$ therefore represents the decreasing possibility of the exposure factor due to a sudden change in the social interactions.

From $S_i(t) = 4$

When an agent becomes a heavy user, she has only two possible choices according to the transition table showed in the Fig. 2.1a. The possibilities are:

- The agent can leave drug with the probability of 5%;
- The agent can persist in the stage $S_i(t+1) = 4$.

At this stage, the situation of the agent is independent of the condition EB and SEB due to every addiction. Therefore we assume only a simple probabilistic transition:

$$\begin{cases} S_i(t+1) = 0 & \text{if} \quad l_i(t) > 0.95 \\ S_i(t+1) = 4 & \quad \text{Otherwise} \end{cases} \tag{2.14}$$

Where $l_i(t)$ is a random number from uniform distribution $[0 : 1]$. The probability of the 5% is an estimation suggested by the Prevo.lab studies and is related to the possibility of recovering from drug addiction.

In this way we have defined all the rules that allow an agent to evolve and reach all possible stages of consumption. So far, the agent cannot die and the age $y_i(t)$ is extracted at the beginning of the simulation and remains constant for every t. Later we define the rules that describe the death and birth of the agent and its aging dynamics.

2.4.3 Aging

Here we show the rules that define the process of an agent's vital cycle (death and birth), and the dynamic that leads an agent to overdose.

Death and birth condition

Based on the 2007 ISTAT data we consider the probability of death in a given year for a person of a given age. We label this probability $P_{y_i(t)}$ and every $\tau = 50t$ consecutive iteration we check this rule for each of N agents:

$$\begin{cases} u_i(t) < P_{y_i(t)} & y_i(t+1) = y_i(t) + 1 \\ \text{Otherwise} & i \text{ dies } y_i(t+1) = 10 \end{cases} \tag{2.15}$$

Where $u_i(t)$ is a random number from uniform distribution $[0 : 1]$. When the agent i dies, she will be replaced by another agent with age $y_i(t+1) = 10$ and she will be characterized by the new extraction of the parameters β_i, γ_i from the current value of the distributions at time t. Instead, if she lives, she will be one year older.

Overdose rule

By adding the possibility of death, it is possible to improve also some rules of the $S_i(t) = 4$ stage transition. An agent in $S_i(t) = 4$ can become a non-user by either detoxifying in a treatment center or by dying of overdose. When the agent dies of an overdose, the agent will be replaced by a new agent with the new extraction of the parameters β_i, γ_i from the current value of the distributions at time t. The following equation

expresses the overdose death condition:

$$\begin{cases} S_i(t+1) \to 0 & \text{if } g_t(i) > 0.1 \quad \text{lives} \\ S_i(t+1) \to 0 & \text{Otherwise} \quad \text{dies} \end{cases} \tag{2.16}$$

Where $g_i(t)$ is a random number from uniform distribution $[0:1]$. The agent with the probability of 10% dies. Otherwise she lives. The probability to survive after an overdose is taken from the data of SimDrug and Sert [11, 35]. With our dynamics we can replicate the same overdoses trend of SimDrug model too (see the following section 2.5.1).

2.5 Other analysis

2.5.1 Overdoses Comparison

One of the main achievement proposed by the SimDrug model [11] is the study of the amount of fatal and non-fatal overdoses among the simulated population. Thanks to the SimCity like framework SimDrug model is able to reproduce the trend of overdoses within Melbourne city. Fig. 2.8a shows the results of SimDrug for fatal and non-fatal overdoses. This simulation was carried out with 3000 agents for $1400t$ time step iterations. Each time step t represents a single day and one simulated years is defined by $350t$.

We set up our parameters to have the same environment of SimDrug model: $N = 3000$ agents for 4τ iterations with $\tau = 350t$. For our simulations we keep the main three distribution tune to the UD except for the average value of the intrinsic inclination to drug β_i that we set up at: $B_m = 1$. Fig. 2.8b shows that our model is able to reproduce the same trend of a more complicate and time-consuming simulation defined by more complex rules.

2.5.2 Probability Studies

Detoxify probability

Drug addiction studies and report [33, 34, 35] always try to quantify the correct amount time, needed for a person, to stop the drug abuse depending of its current stage of addiction. In our model is possible to analyze

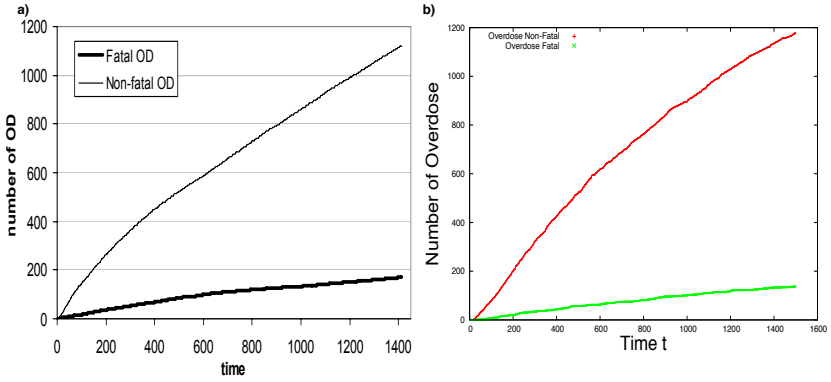

Figure 2.8: Overdoses Comparison. a. SimDrug [11] simulation with 3000 agents for $1400t$ time step. The SimDrug model is able to reproduce the same trend of the overdoses within Melbourne city. **b.** Overdose study of our model. Here we show the our model is able to reproduce the same result of more complicate and time-consuming simulation defined by more complex rules. The simulation was made with $N = 3000$ agent for 4τ, with the UD except for the average value of the intrinsic inclination to drug β_i set up at: $B_m = 1$.

the probability for an agent i in a particular stage $S_i(t_0) \neq 0$ at time t_0 to come back for the first time to the stage $S_i(t_0 + t) = 0$ after t model iteration. Considering the agent's path though the addiction stages as a Markov process, we can define this probability as:

$$P(S_i(t_0 + t) = 0 | S_i(t_0) \neq 0) = P(S_i(t_0 + t) = 0 | S_i(t_0 + t - 1) \neq 0) \times \\ \times \prod_{k=1}^{t-1} P(S_i(t_0 + n) \neq 0 | S_i(t_0 + n - 1) \neq 0) \tag{2.17}$$

It is very difficult to calculate rigorously the entire transition matrix of the Markov chain except for the values of $S(t_0) = 4$. Straightforward we have already defined from Eq. 2.14 the value of the $P(S_i(t_0 + t) = 0 | S_i(t_0) = 4) = 0.05$ for $t = 1$.

Fig. 2.9 shows the empirical prediction of this probability. To have a good statistics we carried out 10 simulations of $N = 50000$ agents for

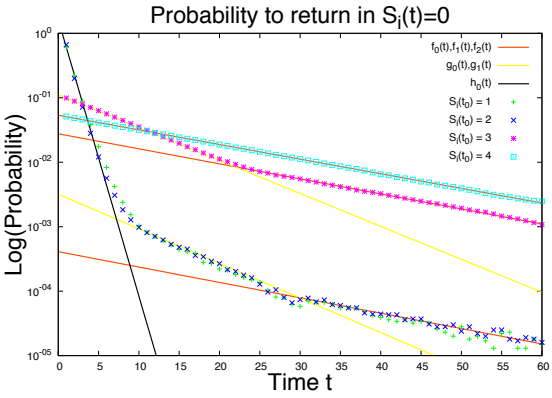

Figure 2.9: Detoxify probability. Here we show the probability for an agent i in a particular stage $S_i(t_0) \neq 0$ at time t_0 to come back for the first time to the stage $S_i(t_0+t) = 0$ after t model iteration. One can see the three different pattern outline by the color fit depending of the current addict stage of the agent in consideration. The black fit represents the $S_i(t_0 + t) = 1, 2$, the yellow $S_i(t_0 + t) = 3$ and the red $S_i(t_0 + t) = 4$. The red line fitted is in agreement with the Markov model define by 2.14

20τ, with $\tau = 50t$ time iterations. One can see that the transition probability of return in $S_i(t_0 + t) = 0$ for the first time with $S(t_0) = 4$ is in agreement with the expected data (red fit line). Furthermore the figure highlights three distinctive pattern depending on the agent stages in t. The changes of the fitting coefficient in the probability outline the changing in the stage of the agent depending of t. The black line fit is the peculiar coefficient for agents at the stage $S_i(t_0 + t) = 1, 2$, the yellow fits line represent the behavior of the agents in the stage $S_i(t_0 + t) = 3$. As we can see the agent in $S_i(t_0) = 1, 2$ after approximately $30t$ will reach the stage $S_i(t_0 + t) = 4$, if they dod not detoxify before, as outlined by the red fit line.

This analysis gives us a snapshot of the agents addiction timing of our model and it could be easily compared directly to medical data. The model is now tuned thanks to the national macro data provide by Prevo.lab. The fine microscopic set up,thought this analysis, could improve considerably the model performance and reliability.

Chapter 3

The Italian primary school-size distribution and the city-size: a complex nexus

3.1 Introduction

There is a growing literature that nowadays sheds light on complexity features of social systems. Notable examples are firms and cities [16, 17, 18, 49], but many others have been proposed [50, 51]. These systems are perpetually out of balance, where anything can happen within well-defined statistical laws [52, 53]. Italian schools system seems to not escape from the same characterization and destiny. Despite several attempts of the Italian Ministry of education to reduce the class-size to comply with requirements stated by law [47, 54, 55], no improvements have been made and still heterogeneity naturally keeps featuring the size distribution of the Italian primary schools.

In this chapter we characterize the statistical law according to which the size of the Italian primary schools distributes. Using a database pro-

vided by the Italian Ministry of education in 2010 we show that the Italian primary school-size approximately distributes (in terms of students) as a log-normal distribution, with a fat lower tail that collects a large number of very small schools. Similarly to the firm-size [56, 57], we also find the upper tail to decrease exponentially. Moreover, the distribution of the school growth rates are distributed with a Laplassian PDF. These distributions are consistent with the Bose-Einstein preferential attachment process. These results are found both at a provincial level and aggregate up to a national level, i.e. they are universal and do not depend on the geographic area.

The body of the distribution features a bimodal shape suggesting some source of heterogeneity in the school organization. The evidence of the bimodality underlies the interplay between different processes that define thresholds and boundaries that are very peculiar for the Italian primary school-size distribution. The question that we attempt to address in this chapter is whether such regularity might depend on the complex geographic features of the Country that in turn determines the way population (and in particular young people) distributes. We address these questions by analyzing in depth the spatial distribution of the schools, with particular regard to the areas where commuting is more effortful. We then proceed by investigating the complex link between schools and comuni, the smallest administrative centers in Italy, addressed by the introduction of a new binning methodology and a new spatial interaction analysis. Our conclusions indicate that the bimodality of the Italian primary school-size distribution is very likely to be due to a mixture of two laws governing small schools in the countryside and bigger ones in the cities, respectively.

Several examples of different regional schooling organizations are analyzed and discussed. We use GPS code positions for schools in two very different Italian Regions: Abruzzo and Tuscany. We introduce a measure of the average spatial interaction intensity between a school and the surrounding ones. We show that in regions like Abruzzo, that are mainly countryside, a policy favoring small schools uniformly distributed across small comuni has been implemented. Abruzzo small schools are generally located in low density populated zones, in correspondence of very small comuni. They are also very likely to have another small school as

closest and the median distance between them is 8 km that is also the distance between small comuni. In Tuscany, a flatter region with a very densely populated zone along the metropolitan area composed by Florence, Pisa and Livorno, we conversely find:

- higher school density;
- stronger interaction between small and big schools;
- greater average proximity among schools.

We address these stylized facts by arguing that the Italian primary school organization is basically the result of a random process in the school choice made by the parents. Primary education is not felt so much determinant to drive housing choice, like in US, because of the absence of any territorial constraint in school choice. Even if there is a certain mobility *within* a comune toward the most appealing schools, primary students generally do not move *across* comuni to attend a school. As a result, school density and school-size are prevalently driven by the population density and then by the geographical features of the territory. This generates a mixture in the schooling organization that turns into a bimodal shape distribution.

3.2 Results

Empirical evidence

We analyze a database on the primary school-size distribution in Italy that provides information on public and private schools, locations, and the number of classes and students enrolled. Data are collected, at the beginning of every academic year, by the Italian Ministry of education to be used for official notices. Our dataset covers $N = 17187$ primary schools in 2010 of which 91.31% were public. Almost seven thousands are located in mountain territories, (which represent the 40%) and 4101 are spread among administrative centers (provincial head-towns).

In Italy primary education is compulsory for children aged from six to ten. However, the parents are allowed to choose any school which they prefer, not necessarily the school closest to their home, [55]. We define x_i

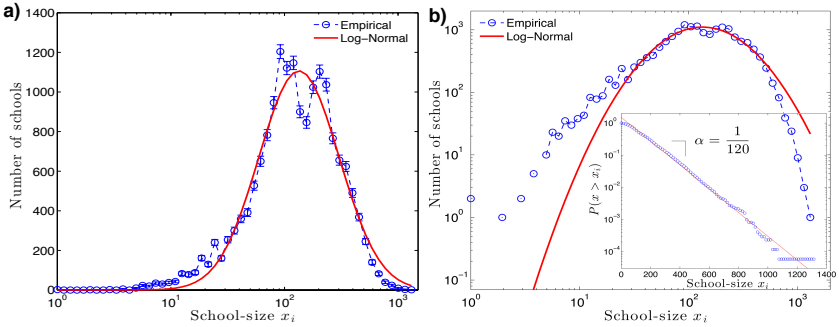

Figure 3.1: School-size distribution. **a.** Italian primary school-size distribution according to the number x_i of student per school $i \in [1, \ldots, N]$ for the year 2010. The empirical distribution is drawn in blue (each circle is a bin); the red line stands for the Gaussian fit with mean $\hat{\mu} = 4.77$ ($\hat{\mu}/\ln(10) = 2.07$) and standard deviation $\hat{\sigma} = 0.85$ ($\hat{\sigma}/\ln(10) = 0.37$). On a non-logarithmic scale, $\exp(\hat{\mu}) = 118$ and $\exp(\hat{\sigma}) = 2.34$. $N = 17187$. Statistical errors (SE) are drawn in correspondence of each bin, according to $\sqrt{N_{bin}}$. SE are bigger in the body of the distribution and tinier in the tails. Nevertheless, central bins space from the two peaks, $m_1 = 1.7$ and $m_2 = 2.3$, at least 6 times the SE, equals on average to $\sqrt{10^3} = 32$. In this case the probability to have a non bimodal shape under our distribution is 4×10^{-15}. **b.** Italian primary school-size distribution in log-log scale. As expected, the theoretical distribution has drawn as a perfect parabola (the red curve), $y = ax^2 + bx + c$, such that $\hat{\mu} = -b/2a$ and $\hat{\sigma} = -1/2a$. Conversely, the empirical distribution does not plot as a parabola, at least for what regards to the tails which deviate from the log-normal. The inset figure shows a functional form of the right tail of the empirical distribution. We plot the cumulative distribution, $P(X > x_i) = \exp(-\alpha x_i)$, of school sizes in semi-logarithmic scale with characteristics size $\alpha = 0.0084$. This in turn means that there are approximately 120 students per school.

the size of the school $i \in [1, \ldots, N]$ as the number of students enrolled in each school. Fig. 3.1(a) shows the histogram of the logarithm of the size of all primary schools in Italy. The red solid curve is the log-normal fit to the data

$$P(\ln x) = \exp\left(-\frac{(\ln x - \hat{\mu})^2}{2\hat{\sigma}^2}\right)\frac{1}{\sqrt{2\pi}\hat{\sigma}} \tag{3.1}$$

using the estimated parameters $\hat{\mu} = 4.77$ ($\hat{\mu}/\ln(10) = 2.07$), the mean of the $\ln x$ of the number of students per school, and its standard deviation, $\hat{\sigma} = 0.85$ ($\hat{\sigma}/\ln(10) = 0.37$). On a non-logarithmic scale, $\exp(\hat{\mu}) = 118$ and $\exp(\hat{\sigma}) = 2.34$ are called the location parameter and the scale parameter, respectively [58]. The histogram in Fig. 3.1a suggests that log-normal fits data quite well. However, even a quick glance reveals that there are too many schools with a small dimension and much less mass in the upper tail with respect to the fit, suggesting that the number of students of the largest schools is smaller than would be the case for a true log-normal. In other words, similarly with firms-size distribution [59], tails seem to distribute differently from the log-normal distribution. Also Fig. 3.1a reveals a bimodal shape of the school-size distribution that we will extensively investigate below.

These findings can be detected in a more powerful way by plotting the histogram in a double logarithmic scale, comparing the tails of the log-normal distribution with those of the empirical one. We do this in Fig. 3.1b where y-axes represents the logarithm of the number of schools in the bins whereas in the x-axes the logarithm of the number of students stands. The empirical distribution differs significantly from the theoretical distribution which is a perfect parabola (the red curve), both in the tails and in the central bimodal part. A functional form of the right tail of the empirical distribution is revealed in the inset of Fig. 3.1b where we plot the cumulative distribution $P(X > x)$ of school sizes in semi-logarithmic scale. The straight line fit suggests that the right tail decreases exponentially $P(X > x) = \exp(-x\alpha)$ with a characteristics size $\alpha = \frac{1}{120}$. This in turn means that there are approximately 120 students per school and also that the distribution of large schools declines exponentially. The exponential decay of the right tail of size distribution is consistent with Bose-Einstein preferential attachment process and is observed in the distribution of sizes of universities and firms.

Next we investigate the growth rates of elementary schools. Since temporal data are not currently available, we look at the single academic

year, the 2010, and define the growth rate g_i as follows:

$$g_i \equiv \frac{x_i^1 - x_i^5}{\sum_{j=1}^5 x_i^j} = \lambda_i - \mu_i, \tag{3.2}$$

where x_i^j stands for the number of students attending the j-*th* grade in school i, with $j \in [1, 5]$; $\lambda_i \equiv x_i^1 / \sum_{j=1}^5 x_i^j$ is the fraction of students that have been enrolled in the first grade at six years old in school i, whereas $\mu_i \equiv x_i^5 / \sum_{j=1}^5 x_i^j$ is the fraction of students that exit the school after the 5-*th* grade. Fig. 3.2a shows the relation between growth rate g_i and school-size x_i. The numbers of grades j provided by each school i, named J_i, is defined by the color gradient bar on the right side of the Fig. 3.2a. Blue circles identify schools with $J_i = 1$. Such a group collects schools just established only providing the 1-*st* grade, i.e. with $\lambda_i = 1$ *and* $\mu_i = 0$, or that are going to close providing only the 5-*th* grade, i.e. with $\mu_i = 1$ *and* $\lambda_i = 0$. As soon as more grades are provided (colors switching to the warm side of the bar) schools tend to cluster around a null growth rate.

In Fig. 3.2b we investigate the growth/size relationship in depth. We demonstrate the applicability of the Gibrat law that states that the average growth rate is independent on the size [60, 61]. We define the average of the school size in each bin c as $\langle x_i \rangle_c$. The number of school in each bin n_c is represented by the size of the circle and the average number of grades $\langle J_i \rangle_c$ is depicted according to the color gradient on the right side (the same of Fig. 3.2a). Independently from the size and the number of grades provided, schools do not grow on average. Nevertheless, we find more variability in smaller schools, apart from schools with $x_i < 10$, namely hospital-based schools mostly similar to one another, and the standard deviation of the growth rate $\sigma_{g(\langle x_i \rangle_c)}$ is found to be decreasing as $\langle x_i \rangle_c^{-\beta}$ with school-size by a rate of $\beta \approx .60$ (subFig. 3.2b inset). This is consistent with what has been found for other complex systems like firms or cities [13, 14, 56, 62, 63, 64].

In Fig. 3.3a we study the growth rate distribution, where the probability density function $P(g = g_i)$ of growth rate has been plotted. The blue line represents the full sample (all the schools) distribution. Black and red colors identify the full capacity schools ($J_i = 5$) and the schools with $J_i < 5$, respectively. Regardless of the number of grades provided,

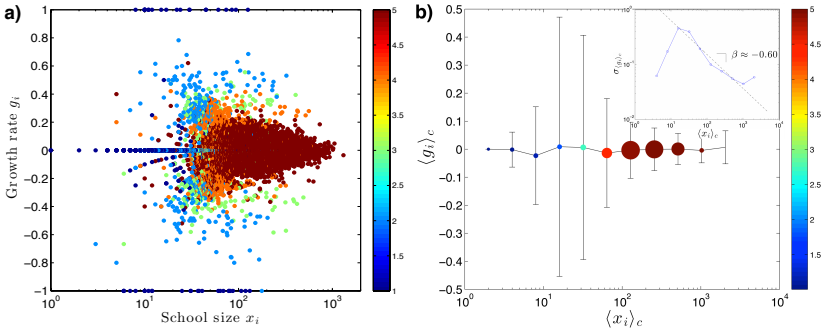

Figure 3.2: The growth rate and school-size relationship. The growth rate g_i is defined according to Eq. (3.2). **a.** Colors, according to the vertical bar on the right-hand side of the graph, are the number of grades J_i provided by the school i. Smaller schools (in blue) with $J_i = 1$ are both the newest one (just created, with $\lambda = 1$) and schools that are going to close (with $\mu = 1$). They can also be schools that do not grow yet providing just one grade (i.e. $j = 3$). **b.** The mean growth rate clusters around zero across different subsets c that are differently populated by n_c schools according to the size of the circles. The color of the circles stand for the average number of grades J_i (the same gradient color bar of Fig. 2(a) is used here). The variability within each cluster c is shown in the inset figure. Apart from schools with $x_i < 10$, namely hospital-based schools mostly similar to one another, the standard deviation is found to be decreasing with school-size by a rate of $\beta \approx .60$.

the growth distribution underlines a Laplace PDF in the central part of the sample [65]. The not-fully covered schools show a three peak behavior, where the left peak represents schools which are going to close, the central peak gathers schools that provide several grades but still in equilibrium phase, and the right peak is made up by the growing schools. Fig. 3.3b reports empirical tests for the tails of the PDF of the growth rate of the full sample (the upper one in blue, and the lower one in black). The asymptotic behavior of g can be well approximated by power laws with exponents $\zeta \approx 4$ (the magenta dashed line), bringing support to the hypothesis of a stable dynamics of the process [62]. All these findings are consistent with the Bose-Einstein process according to which the size dis-

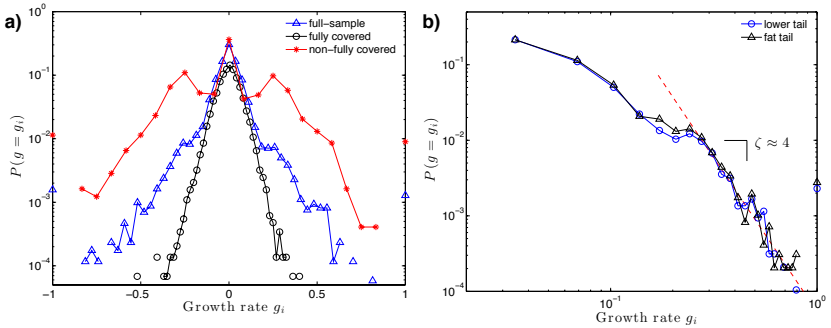

Figure 3.3: The growth rate distribution of the Italian primary schools in 2010. a. The probability density function $P(g = g_i)$ of growth rate has been plotted underlying a Laplace PDF in the body around $P(g) = 1$ and $P(g) \approx 10^{-1.5}$. Blue triangles ($\triangle$) stand for the full sample distribution, black circles ($\circ$) indicate mature schools with $J_i = 5$, and red stars ($*$) schools with $J_i = 1$. **b.** The plot reports empirical tests for the tails parts of the PDF of growth rate, the upper one in blue ($\circ$), and the lower one in black ($\square$). The asymptotic behavior of g can be well approximated by power laws with exponents $\zeta \approx 4$ (the magenta dashed line).

tribution has an exponential right tail, a tent-shaped distributed growth rate g_i, with a Laplace cap and power law tails, the average growth rate is independent of the size, and the size-variance relationship is governed by the power law behavior with exponent $\beta \approx 0.5$ [66].

3.2.1 City size and school size

Fig. 3.1a features the coexistence of two peaks, the first peak corresponding to $\log_{10} x_i \equiv m_1 = 1.7$ and the second one to $\log_{10} x_i \equiv m_2 = 2.3$, divided by a splitting point in correspondence of $\log_{10} x_i \equiv \bar{m} \approx 2.1$. The school sizes corresponding to these features are $\mu_1 = 10^{m_1} = 50$, $\mu_2 = 10^{m_2} = 200$, and $\bar{\mu} = 10^{\bar{m}} = 128$, with $\bar{\mu}$ approximately equal to the average school size. 39% of the Italian primary schools distribute on the right of $\bar{\mu}$, and more than 60% distribute on the left side. We test the alternative hypothesis of unimodality by looking at the probability that the numbers of schools in the two central bins n_1, n_2 are not smaller and

the numbers of schools in the next three bins n_3, n_4, n_5 are not larger than a certain number n^* provided that the standard deviation of the number of schools in these bins due to small statistics is $\sqrt{n^*}$. This probability is equal to $p(n^*) = \prod_i \mathrm{erfc}(|n_i - n^*|/\sqrt{2n^*})/2$ and it reaches maximum $p_{\max} \approx 4 \times 10^{-15}$ at $n^* = 980$. Accordingly, we establish the bimodality with a very high confidence. This is also consistent with the bimodality index that we find to be equal to $\delta = (\mu_1 - \mu_2)/\sigma = .45$, [67].

In this section we investigate the source of this heterogeneity that we find to be related to geographical and political features of the country and remarkably to the size of the comuni, the smallest administrative centers in Italy (information on comuni are provided by the Italian statistical institute, ISTAT), also here referred interchangeably as cities regardless of the size, p_k. A particular treatment is devoted to the nexus between the school-type (private versus public) and the geographical features of the comuni in the supplementary information, where we show that private schools are much less variable in size than public schools and have a narrow unimodal distribution peaked at approximately 100 students which contributes to the left peak of the entire school size distribution (Figure 3.14).

We denote a comune with letter $k = [1, \ldots, K]$. In 2010, $K = 8,092$ comuni have been counted in Italy, the 40% of which located in the mountains. We define $\mathcal{M}$ the set of mountains comuni and, accordingly, we call school i a mountain school iff it resides in a comune $k \in \mathcal{M}$ (in the sec. 3.5 we explain the mechanism according to which Italian comuni are classified as mountains). Each city k has $n_k \geq 0$ schools (more than 15% of the cities have no schools) and population p_k, which distributes approximately as a log-normal PDF (see Fig. 3.4a), except for the right tail that is distributed according to a Zipf law, i.e. $p_k \sim r(p_k)^{-\xi}$ with slope $\xi \approx 1$ [15, 16, 18, 68, 69]. In Fig. 3.4b we find $\xi \approx .80$, in Italy, that is exactly the slope of the power law $p_k \sim r(n_k)^{-\zeta}$ which links the population p_k with the rank of this city in terms of number of schools n_k (blue circles in Fig. 3.4b), i.e. $\zeta = \xi \approx .80$. This means that the first city, Rome, has almost the double number of schools than Milan, and triple of Naples, while Rome has almost the double of inhabitants of Milan, and the triple of Naples. This amounts to say that n_k is a good proxy for the city-size.

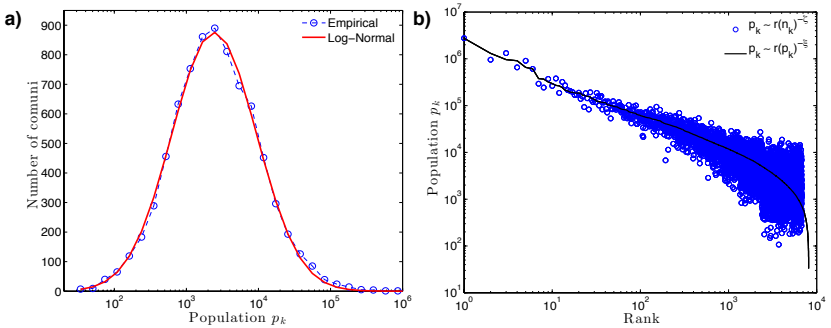

Figure 3.4: Cities features. a. The Italian city-size distribution for $K = 8092$ observations. Blue circles stand for each city-bin whereas the red solid line draws the log-normal fit of the data. Conversely to the school-size distribution depicted in Fig. 3.1a, the city-size PDF features single-peakedness, but similarly it has a power-law decay in the upper tail. **b.** Zipf plot for Italian cities according to the size p_k and the number of schools n_k. The black line draws the classical Zipf plot $p_k \sim r(p_k)^{-\xi}$, with cities ranked according to population p_k. Blue circles instead depict the Zipf plot $p_k \sim r(n_k)^{-\zeta}$, with cities ranked according to the number of schools n_k. Consequently, the sample reduces to $M = 6726$ over $N = 8092$ since more of the 15% of the cities have no schools.

We use the number of schools to assign comuni to different clusters $h \in [1, \dots, H]$, according to

$$h = \{\forall\, k \in [1, \dots, K] : 2^{h-1} \leq n_k < 2^h\}. \tag{3.3}$$

Accordingly, the first bin $h = 1$ gathers all the comuni with only one school; the second one collects all the comuni with $n_k = [2, 3]$, and so on. Though we find the average population $\langle p \rangle_h$ to increase across different city-clusters h, less comuni K_h lie in more populated clusters (the magenta and black lines in Fig. 3.5a). Interestingly, we find the interaction term $K_h \langle p \rangle_h$, the green line in Fig. 3.5a, to distribute uniformly across different comuni-clusters, meaning that in small comuni with $n_k = 1$ live the same population than in bigger ones with much more schools.

Nevertheless, population is differently composed across city-clusters and a smaller fraction of young people is found in smaller comuni. To

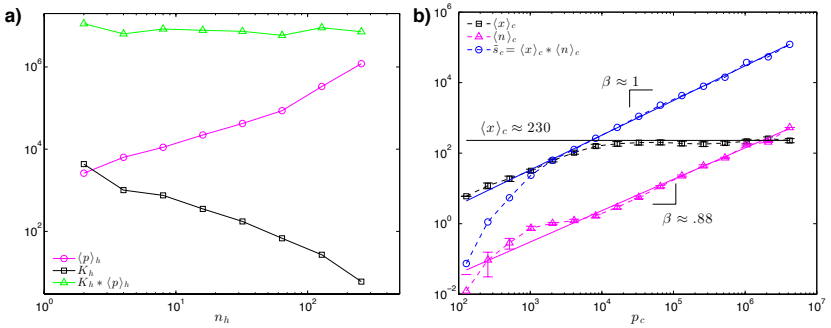

Figure 3.5: Population features. a. Each comune is assigned to 8 clusters, according to Eq. 3.3, and scattered against population, the magenta line (○) and the number of cities K_h, the black line (□). The interaction term, $K_h * \langle p\rangle_h$, the green line (△), represents the total population living in each city-cluster h. **b.** According to Eq. 3.4 K cities are assigned to $C = 16$ clusters. In the x-axis the number of inhabitants in cluster $c = \{7, 22\}$ is scattered against the average number of schools (magenta line (△)) and the average school-size $\langle x\rangle_c$ (the black line (□)). The interaction term (○), representing the typical number of schooling-aged population in cluster c, $\tilde{s_c} = \langle x\rangle_c * \langle n\rangle_c$ distributes as a power law with coefficient $\beta \approx 1$ for cities bigger than 10^3 inhabitants, and it is drawn in green. For smaller comuni, instead, the line drops meaning that a smaller fraction of young people features them.

see that we also introduce a clusterization of comuni according to population. Each comune is assigned to a cluster $c \in [1, \ldots, C]$ composed by all the comuni k with population p_k ranging from ψ^{c-1} to ψ^c, i.e.

$$c = \{\forall\, k \in [1, \ldots, K] : \psi^{c-1} < p_k \leq \psi^c\}. \tag{3.4}$$

Setting the parameter[1] $\psi = 2$ yields $C = 23$ clusters. Although the first seven sets are empty because no comuni in Italy has less than 128 inhabitants, the first (non-empty) cluster, $c = 8$, collects very small comuni with $p_k \in (128, 256]$. The last one, $c = 23$, conversely, is composed by the biggest cities with $p_k \in (2^{22}, 2^{23}]$. In Fig. 3.5b we plot the average number

[1]It is possible to change the value of ψ without having any effect on the shape of the distributions

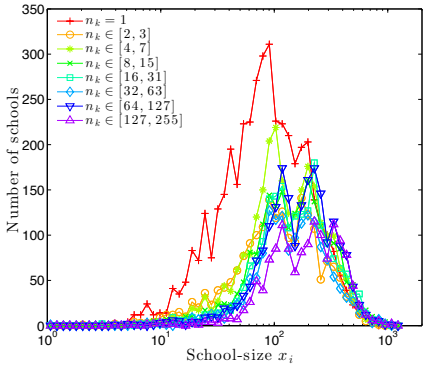

Figure 3.6: City vs. number schools. School-size distribution for different city-samples clustered according to the number of schools, i.e. to Eq. 3.3. Only comuni with $n_k = 1$ show a single peak school-size distribution, clustered around m_1 (the $+$-red line on the top). They have an average population of 2000 inhabitants and the 81% are located in mountain territories.

of schools $\langle n \rangle_c$ (magenta line) and the average school-size $\langle x \rangle_c$ (the blue line) against the comuni size p_c for each non-empty cluster c. We find that the average number of schools increases as a power law with coefficient $\beta = 0.88$. This is consistent with the literature [15, 16, 18, 69] that has stressed the emergence of scale-invariant laws that characterize the city-size distribution. The average school-size increases with the population of the city reaching an asymptotic value at $\langle x \rangle_c \simeq 230$ students per school in the large cities. As expected, the interaction term, representing the average number of school-aged population in comuni belonging to cluster c, $\tilde{s}_c = \langle x \rangle_c * \langle n \rangle_c$, behaves linearly with the comuni size except for small comuni with $p_c < 10^3$, for which the school-aged population constitutes a smaller fraction of the total population than in large cities.

In Fig. 3.6-3.7 we investigate the school-size distribution according to the comuni features. To this end, Fig. 3.6 draws the distributions of $\log_{10} x_i$ conditionally on the number of schools, n_k, in the comune k. It yields 8 curves, one for each cluster h defined in Eq. 3.3. The first cluster is drawn in red ($+$) distributing all the schools located in comuni where only one school is provided. The orange line ($\circ$) distributes all the schools provided in comuni with two or three schools (i.e. $h = 2$);

and so on. The interesting point of Fig. 3.6 is that only the school-size distribution of the smallest comuni (with $n_k = 1$) features a unimodal shape. The reason for that relies on the fact that comuni with only one school are geographically similar: they are the 57% of the total, with little more than 2000 inhabitants, the 81% of which are located in mountain territories.

The relationship between school-size and altitude is investigated in Fig. 3.7a. Instead of conditioning on $\mathcal{M}$, here we propose a more consistent exercise according to which comuni are assigned to different bins on the basis of the altitude. In such a way, we can analyze comuni with 1,000 meters above the sea differently to those with 600 meters of altitude that would be gathered in the same cluster $\mathcal{M}$ throwing away informative heterogeneity. It yields 5 bins: the first bin (drawn as a blue $+$ line) gathers all the comuni whose altitude is lower than 125 meters above the see level (labeled 125 in Fig. 3.7a). Comuni with an altitude between 125 and 250 meters above the sea level composed the second bin (the green $\circ$ line). These two distributions cluster around the second mode m_2 and in the Supplementary Information we additionally demonstrate that the hypothesis of bimodality can be rejected for the latter distribution. However, the greater the altitude of the comuni the greater is the shift of the correspnding school-size distribution toward the small schools and the greater is the contribution of these comuni to the first mode m_1. Such a shift becomes evident for comuni with an altitude between $250m$ and $500m$ (red $\triangle$ line). Comuni located between $500m$ and $1000m$ (cyan $\times$ line) and above $1000m$ (purple $\square$ line) clusterize around m_1.

Even the largest cities are very different from each other in terms of their school size distribution. This heterogeneity is very likely to be driven by geographical features. where we restrict our interest on the largest Italian cities belonging to cluster $h = 8$ (and to the first two bins in terms of altitude in Fig. 3.7a). These cities provide a number of schools n_k between 127 and 255, whose overall size distribution shows a three-peak shape with a third peak around 300 students absent in smaller cities (the bottom violet $\triangle$ line in Fig. 3.6). The presence of the three peaks around 100, 200 and 300 students might suggest the presence of architectural standards of school buildings supporting these particular sizes. However, by plotting the distribution by city, Fig. 3.7b, we show that

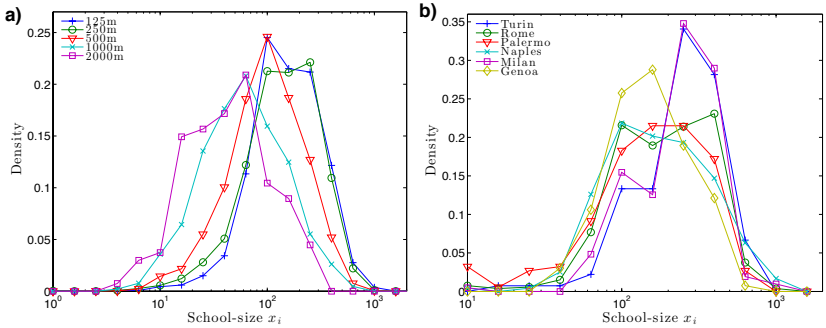

Figure 3.7: School-size distribution conditional on comuni features. a. School-size distribution for different city-samples clustered according to the altitude. The altitude of the comune shift the school-size distribution (shift location effect) as higher comuni are generally smaller schools. **b.** School-size distribution in the six biggest Italian cities. With the exception of Rome, the hypothesis of unimodality may not be rejected in none of the biggest cities. In particular, flatter cities, such as Milano and Torino, mostly contribute to second mode m_2, whereas in Genova, Italian city built upon mountains that steeply ended on the sea, all the school-size distribution stands on the left side.

all the traces of trimodality disappear. In particular flatter cities, such as Milano and Torino, mostly contribute to second mode m_2, whereas in Genova, an Italian city built upon mountains that steeply slope towards the sea, the school-size distribution is unimodal contributing mostly to the first mode m_1.

Another way to look at the effect of geography on the comunal school-size is to compute the fraction of large schools on the total within each comune k:

$$P_k(x_i > \bar{\mu} | \forall\, i \in k) \equiv \frac{n_k(x_i > \bar{\mu})}{n_k} \quad \forall\, i \in k, \tag{3.5}$$

where $n_k(x_i > \bar{\mu})$ stands for the number of schools that, in each comune k, are larger than the minimum $\bar{\mu}$ of the school-size distribution shown in Fig. 3.1a. It can also be interpreted as the contribution rate of a comune k to the second mode m_2. The upper panel of Fig. 3.8 diagrammatically explains how $P_k(\cdot)$ is computed.

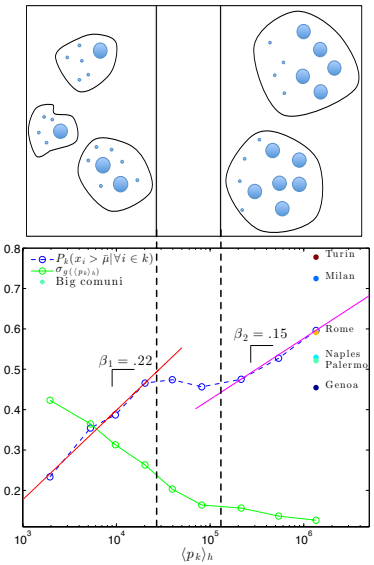

Figure 3.8: Fraction of large schools in comune k. The panel above shows the process according to which each comune, with population p_k defined by the size of the the black circles, is assigned to either patterns on the basis of the size of the schools provided in there (the small blue circles). The panel below shows that more populated clusters of cities are, on average, more likely to have schools sized around m_2. The relationship, depicted in blue, is however non monotonic. In correspondence of each bin h, the standard deviations has been computed, underlining the outstanding variability in very small cities (the green line).

We firstly study the relationship between $P_k(\cdot)$ and population, then looking at the spatial distribution across the Italy. In Fig. 3.8, we clusterize comuni according to Eq. 3.3, and for each bin h we compute the average $\langle P_k(x_i > \bar{\mu} | \forall\, i \in k) \rangle_h$ and population $\langle p_k \rangle_h$. Interestingly, the plot shows that $P_k(\cdot)$ does not increase monotonically with population, demonstrating the existence of two city-patterns. More precisely, cities with less than 10^4 inhabitants follow a pattern according to which the fraction of big schools, with $x_i > \bar{\mu}$, increases, on average, with population at a rate of $\beta_1 \approx .22$; in cities with more than 10^5 we find the effect

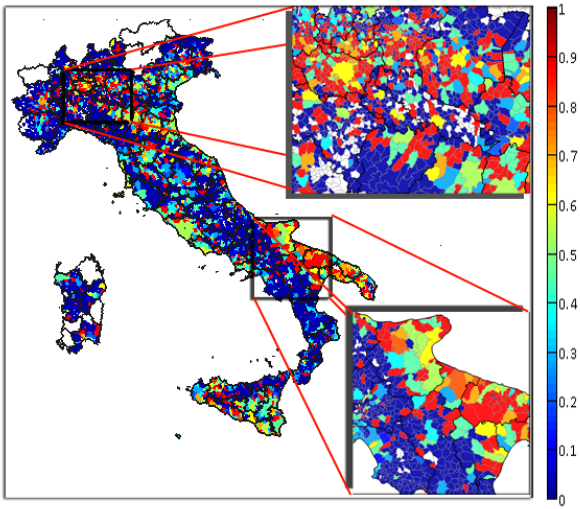

Figure 3.9: Spatial distribution of cities according to $P_k(x_i > \bar{\mu} | \forall\, i \in k)$. Warmer territories stand for cities more likely of having schools distributed around m_2. The two figure inset underline the region around Milan (in the North), on the top, and the regions of Basilicata (mostly mountain, at the left side) and of Apulia (mostly flat, at the right side), on the bottom. Maps generated with Matlab.

of population to be smaller, corresponding to $\beta_2 \approx .15$. For the cities with population between 10^4 and 10^5, the fraction of large schools does not increase with size suggesting that exogenous shocks such as altitude, rugged terrain and age might shift a city in this transition zone to either mode m_1 or mode m_2.

Overall, the distribution of $P_k(x_i > \bar{\mu} | \forall\, i \in k)$ is strongly correlated with the geographical features of the comuni territory. The map in Fig. 3.9 clarifies this point; all the mountain territories, Apennines that represent the spine of the peninsula and the Alps on the northern side, turns to be comuni with small schools, since the share of small schools in mountain comuni is equal to $P(x_i \leq \bar{\mu} | k \in \mathcal{M}) = 0.72$. As soon as the probability to contribute to m_2 increases the colors get warmer; but this is very unlikely to be in mountain territories, because less than 30% of moun-

tain comuni contribute to the antimode. Some regional patterns are also shown in the insets. The first upper panel depicts the area around Milan, which is surrounded by warm colors that mostly dye the Pianura Padana around. On the south side, Apennines approach and colors get blue with a lot of comuni with no schools (depicted in white). This pattern is more evident in the lower panel, which maps the region of Apulia, flat and mostly red, and the Basilicata on the left side, mountainous and mostly blue colored.

3.2.2 Countryside versus dense regions

In this last section, we bring more evidence on the effect of geography and comuni organization on the school-size by restricting our attention at two Italian regions: Abruzzo and Tuscany. But same results stand by looking at regions with the same geographical features. The two regions have very peculiar and representative geographical and administrative characteristics. Abruzzo is a mostly mountain region with a little flat seaside; it has four main head towns divided from each other by mountains. Conversely, Tuscany has many flat zones in the center and the mountain areas shape the region boundaries. Remarkably, it has a very high densely populated zone along the metropolitan area composed by Florence, Pisa and Livorno.

These two regions also differ in terms of administrative organizations, Abruzzo favoring the establishment of comuni with a smaller size due to the presence of mountains. Fig. 3.10 shows the comuni population distributions in Tuscany and Abruzzo. We clusterize comuni using the algorithm in Eq. 3.4. As Fig. 3.10a makes clear, comuni distribute approximately as a log-normal pdf in both regions, i.e. as a parabola in a log-log scale (the green-$\circ$ line stands for Abruzzo pdf, the magenta-$\triangle$ for Tuscany). Nevertheless, Tuscany has bigger cities. Figure 3.10a also shows the average number of schools as function of the population. The fact that $\langle n_K \rangle$ is less than one for small comuni reflects the fact that many of these comuni do not provide schools. Abruzzo has a larger number of small comuni that do not provide schools. The first bin collects comuni with a bit more than 100 inhabitants. They are 7 in Abruzzo (none in Tuscany), none of them providing any school services. The second bin accu-

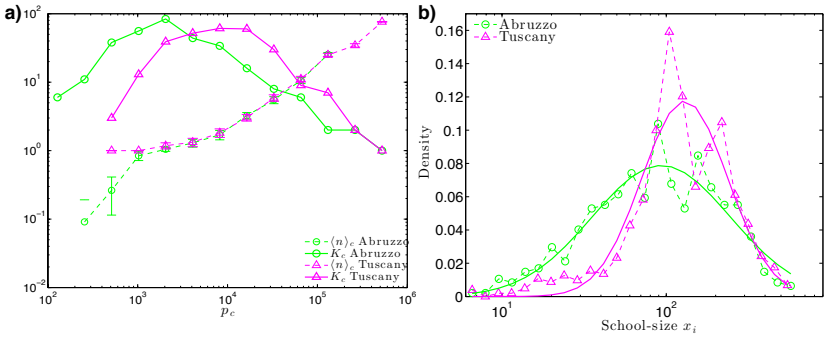

Figure 3.10: Regional analysis I. a. The figure distributes the city-size in Abruzzo (○-green) and Tuscany (△-magenta) by plotting the number of comuni, K_c, against the number of inhabitants, p_c. Also shown is the average number of schools in a comune in Abruzzo and Tuscany, belonging to a bin c defined by Eq. 3.4, by the circled- and triangled-connected lines respectively. **b.** School-size distribution in Abruzzo (○-green) and Tuscany (△-magenta). Both pdf are approximately lognormal and bimodal with splitting point equal to 128 and 151 students per school respectively.

mulates 10 comuni in Abruzzo with 300 inhabitants (none in Tuscany), of which only one has a school. Comuni with about 600 inhabitants are 40 in Abruzzo and only 7 in Tuscany. Only 30% of them have one school in Abruzzo while 80% of them have at least one school in Tuscany. Overall, there are 53 comuni in Abruzzo without schools; only 3 in Tuscany.

Such a differences reflects on the school-size distribution, depicted in Fig. 3.10b. Although primary schools distribute in both regions in terms of size with two peaks, both Abruzzo m_1 and m_2 are shifted on the left w.r.t. the Tuscany ones. The average school-size is smaller in Abruzzo ($\hat{\mu}_{ABR} = 4.56$ ($\hat{\mu}_{ABR}/\ln(10) = 1.98$) versus $\hat{\mu}_{TOS} = 4.91$ ($\hat{\mu}_{TOS}/\ln(10) = 2.13$)), and, remarkably, the lower tail is fatter in the former region. The cutoff for splitting the mixed distributions amounts to 128 in Abruzzo and 151 in Tuscany, and 31% of the schools are clustered in the second peak in the former region; $P(x_i > \bar{\mu}_{TOS} | \forall\, i \in TOS) = 0.38$ in the latter.

In Fig. 3.11a we show, following the same clustering technique used in Fig. 3.8, that in both regions the fraction of big schools within comune

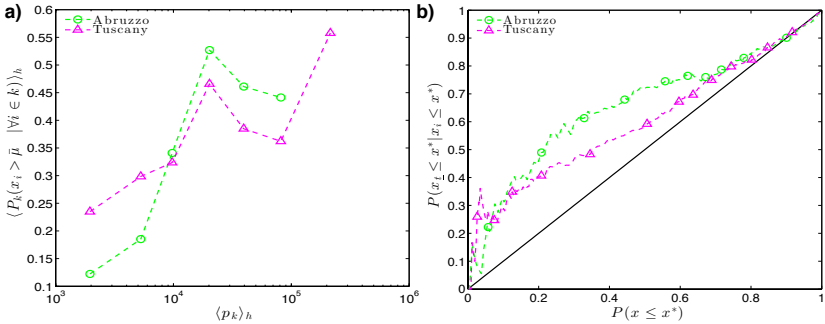

Figure 3.11: Regional analysis II. a. Average fraction of big schools in each comuni bin, defined by Eq. 3.3, in Abruzzo (○-green) and Tuscany △-magenta). The plot shows that more populated comuni are, on average, more likely to have schools sized around m_2, in both regions. Yet, in mountain regions, such as Abruzzo, smaller comuni have also smaller schools on average. **b.** The conditional probability is plotted in the y-axis, for an arbitrary school size x^*, as function of x^* against the cumulative probability $P(x_i \leq x^*)$. The conditional probability is equal to the cumulative in correspondence of the black line. Along these points, there is no attraction between schools of the same size. This is not the case in both the two regions.

k, $P_k(x_i > \bar{\mu}|\forall\, i \in k)$, increases monotonically with respect to the number of inhabitants for $\langle p_k\rangle_h < 20000$.

In this interval, a comparison with figures for entire Italy, plotted in Fig. 3.8, reveals that both regions follow the same national pattern. Yet, mountain regions, such as Abruzzo, have a significantly smaller concentration of big schools. In particular in Abruzzo, only about $1/10$ of comuni with just one school, with an average population of roughly 2000, have a school with more than 125 students. In Tuscany, they are the 25%, about the same as national ratio. In larger comuni, with an average population of 5000 and two schools provided (the second bin), the probability of having big schools raises to 0.2 in Abruzzo, still smaller than Tuscany where $\langle P_k(x_i > \bar{\mu}|\forall\, i \in k)\rangle_{h=2} = 0.3$.

Small schools are mainly located in the countryside, and for that reason they cluster together, i.e. it is more likely to find a small school near

a small one. In Abruzzo this clustering effect is stronger than in Tuscany. We investigate this point in Fig. 3.11b, where we compute, and plot on the x-axis, the cumulative probability $P(x_i \leq x^*)$, as function of x^*, and the correspondent conditional probability $P(x_{\underline{t}} \leq x^* | x_i \leq x^*)$, on the y-axis, which is the fraction of schools with the size smaller than x^* among the schools closest to a school of size x^*. This quantity is equal to 74% and 65% for $x^* \equiv \bar{\mu}_{reg}$ in Abruzzo and Tuscany respectively, meaning that there is a greater probability that a small school matches with another of the same kind in the former region. If the conditional probability were equal to the cumulative, as indicated by the black line in Fig. 3.11b, the sizes of neighboring schools would be independent. This is not the case in either the two regions. The probability that a small school has a smaller nearest neighbor is larger than the probability that any school is smaller than a given one. Indeed, the two curves (green for Abruzzo and magenta for Tuscany) are significantly above the 45 degree line for $P(x_i < x^*) < 0.6$ in Tuscany and for $P(x_i < x^*) < 0.7$ in Abruzzo. These probability values roughly correspond to the probabilities $P(x_i < \bar{\mu})$ in respectively Tuscany and Abruzzo, indicating that in both regions small schools are likely to belong to the small mountainous comuni, whose nearest neighbors are of the same class.

We further study the attraction intensity among small schools by disentangling the effect between the countryside and dense zones. To this end, we analyze the GPS location of the schools in the two regions and, for each school i, we compute the number of schools n_m^i belonging within a circle of radius r_m centered at each school j. We exclude from n_m^i all the schools which do not belong to Tuscany or Abruzzo, respectively. To eliminate the effect of region's boundaries, we also compute areas D_m^j as the areas of the intersections of these circles with a given region (Abruzzo or Tuscany). Thus $D_m^i \leq \pi (r_m^i)^2$, because these areas do not include the seaside and administrative territories of other regions. The difference between two subsequent circles yields the area of the annulus $A_m^i = D_m^i - D_{m-1}^i$. The density of schools in the area A_m^i is then defined as:

$$\rho_m^i = \frac{n_m^i - n_{m-1}^i}{A_m^i}, \tag{3.6}$$

and the average density of schools as function of a distance to a randomly

selected school is

$$\langle \rho_m \rangle_i = \frac{\sum_N n_i^i - \sum_N n_{m-1}^i}{\sum_N A_m^i}. \tag{3.7}$$

In Fig. 3.12a red lines represent the average school-density around *all* the schools in Tuscany and Abruzzo, which are 472 in the former and 1037 in the latter region. Green lines describe the average school density around a small school with $x_i \leq \bar{\mu}$, named S_1, whereas the blue lines describe the density around large schools, S_2. 64% of the schools in Abruzzo belong to the S_1 group, 53% in Tuscany. Fig. 3.12a collects evidence about the fact that small schools S_1 are located in low school density zones and, accordingly, have a smaller probability to be surrounded by competitor schools than large schools (S_2) located in densely populated areas. In both regions, in fact, the green line goes under the blue one, for at least first $50km$. In particular, within this distance, in Abruzzo the density stays almost constant at approximately 0.053 meaning that 1 school is provided every $20km^2$. In Tuscany, this figure goes up to 0.07, because of a generally higher population density, but yet small.

The correlation coefficients between the school size and the distance to it nearest neighbor are negative in both regions, but the magnitude is quite different, equal to 0.34 in Abruzzo, that is 1.7 times greater than in Tuscany (0.20). To reduce the noise, we proceed by clusterizing schools according to their size. Fig. 3.12b confirms this pattern by showing that small schools have on average more distant nearest schools. We look at the size of each school in both regions, and we define the geodesic euclidean distance between the school i and its nearest neighboring school (which we denote by subscript $\underline{t}$) as $d(x_i, x_{\underline{t}})$. The binning algorithm used is to base 2:

$$l = \{\forall\, i \in [1, \ldots, N] :\, 2^{l-1} \leq x_i < 2^l\}. \tag{3.8}$$

This clusterization yields 8 bins, with different average sizes plotted on the x-axis of Fig. 3.12b. On the y-axis, we plot the average distance between the school i, that belongs to the bin l, and its nearest, i.e. $\langle d(x_i, x_{\underline{t}}) \rangle_l$. Each school-bin l is depicted by green circles for Abruzzo and magenta triangles for Tuscany. The average distance between the closest schools decreases with respect to the average school size $\langle x_i \rangle_l$ for $\langle x_i \rangle_l > 32$ in both regions meaning that, in general, small schools are sparser than

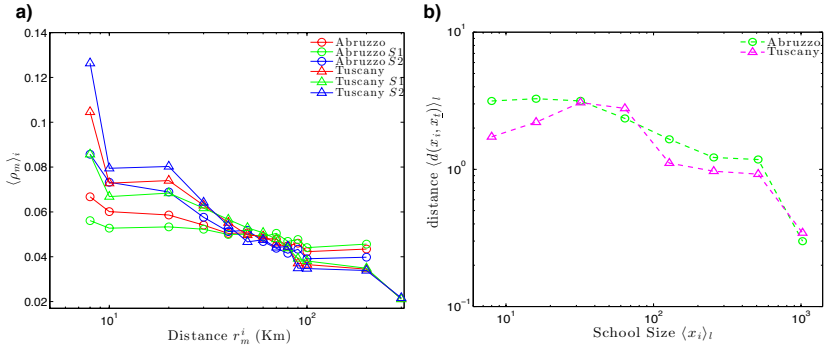

Figure 3.12: Regional spatial analysis. a. $\langle \rho_m \rangle_i$ has been plotted, based on Eq. 3.6, and 3.7, for the region of Abruzzo (□) and Tuscany (△). The red line draws the trajectory averaging among *all* the schools in Italy. Green and blue lines stand for small schools, i.e. $x_i \leq \bar{\mu}$, called S_1, and big schools, i.e. $x_i > \bar{\mu}$, called S_2, respectively. **b.** The average distance, in km, between the closest schools, $\langle d(x_i, x_{\underline{i}}) \rangle_l$, is plotted in Abruzzo (○-green) and Tuscany (△-magenta) with respect to the average size, $\langle x_i \rangle_l$. Each cluster l has been obtained by aggregating schools with near size according to Eq. 3.8. In Tuscany, the schools provided in small islands, at least $20km$ far from the coast, have been removed in order to eliminate any artificial bias from the spatial analysis, whereas the 18% of the schools, with no address provided in the MIUR dataset, have been geocoded in Tuscany according to the GPS localization of the city hall of the comune in which they stand. The average distance between the closest schools decreases in both regions with respect to the average size meaning that, in general, small schools are sparser than large schools that are more likely to be located in very dense zones, like cities.

large schools that are more likely to be located in very dense zones, like cities. The non-monotonic behavior of this quantity for $\langle x_i \rangle_l < 32$ in Tuscany can be explained by the fact that such small schools in Tuscany are usually hospital schools which are located in densely populated areas. Whereas the schools provided in small islands, at least $20km$ far from the coast, have been removed in order to eliminate any artificial bias from the spatial analysis. The three first magenta bins are all below the green ones, confirming, in accordance with the geographical features

of the two regions, that in Abruzzo small schools are sparser and more likely to be located in the countryside where the school density is low (see Fig. 3.12a). Moreover, small schools on average have a distance to the nearest neighbor of $4 - 5km$ which is the average distance between a small comune and a more school-dense one (see the Methods section).

The two regions then outline very different patterns of the school system in the countryside. In Abruzzo small schools are uniformly distributed across small comuni, as a result of a policy favoring the disaggregation of the comuni and school organization, due to a tight geographical constraint. In Tuscany, instead, a different system has been implemented, according to geographic features and a higher population density, where small comuni are larger and do not necessarily have small schools, especially if they stand in very populated zones.

3.3 Discussion

We have studied the main features of the size distribution of the Italian primary schools, including the sources of the bimodality, and we have investigated its relation to the characteristics of the Italian cities. The fat left tail of the distribution is the consequences of political decisions to provide small schools in small (mostly countryside) comuni, instead of increasing the efficiency of public transportations. This is most probably caused by the topographical features of the hilly terrain making transportation of students dangerous and costly. The evidence of this conclusions is that hilly cities like Palermo, Napoli, an, above all, Genoa, with steep mountains that end up into the sea, have higher fraction of small schools than mainly flat cities like Torino and Milano.

The analysis of schools growth rates highlights that the schools dynamics follows the Gibrat law, and both the growth rate distribution and the size distribution are consistent with a Bose-Einstein process. Alternatively, the exponential decay of the upper tail can be explained by a constraint by the size of the building or a traveling distance and transportation cost.

Despite our results are conducted using data on Italian primary schools, they predict that schooling organization would be different in

another country with different geographical features. Flat territory would lead to open schools in the main villages allowing the children residing in the smallest ones to travel daily. This result is additionally supported by the fact that no territorial constraint has been imposed to the schooling choice. Despite parents can enroll children in the most preferred school, primary students generally do not move *across* comuni to attend a school. Accordingly, we find that school density and school-size are prevalently driven by the population density and then by the geographical features of the territory, as a result of a random process in the school choice made by the parents. This goes in the opposite direction with what has been found in other countries such as USA where school choices influence residential preferences of parents and drive the real estate prices in townships depending on the quality of their schools [70].

The availability of new longitudinal school data will be relevant to a more in-depth analysis and further discussions. Moreover, the availability of data for other similar countries would favor comparison and would be useful to assert our theory. We believe that this study, and future research, can lead to a higher level of understanding of these phenomena and can be useful for a more effective policy making.

3.4 Methods

In this section we propose a novel algorithm for the analysis of spatial distribution of primary schools in entire Italy. This algorithm is needed if the exact coordinates of individual schools are not available, but instead, the centers and the territories of all the comuni are known. For each comune k, we define a gravity center g_k of its territory corresponding to the GPS location of its city hall, and t_k as the area of the comune administration. In Italy the city hall is located in the center of the densely populated part of the administrative division, in order to be easily reachable by the majority of inhabitants. We develop a novel spatial-geographical approach consisting of a sequence of geographic regions bounded by two concentric circles, that we exemplified in Fig. 3.13a for a comune in Abruzzo. First we define a set Z_m^k of comuni whose city halls are within

a circle of radius r_m^k and the center at the city hall of comune k. Formally,

$$Z_m^k = \{\forall\, j \in [1, .., K] : d(g_k, g_i) \leq r_m^k\}. \tag{3.9}$$

Next we compute the number of schools provided by the comuni which are members of set Z_m^k that is defined by

$$n_m^k = \sum_{j \in Z_m^k} n_j \tag{3.10}$$

and their area

$$D_m^k = \sum_{j \in Z_m^k} t_j, \tag{3.11}$$

where t_j is the area of comuni j. Next we compute the area associated with all the comuni in the m-th concentric annulus surrounding comune k as the difference between the area associated with the larger circle m of radius r_m^k and the area associated with the smaller circle $m-1$ of radius r_{m-1}^k, i.e. $A_m^k = D_m^k - D_{m-1}^k$. In Fig. 3.13a, each comune territory is colored with different colors according to the annulus in which they belong.

The density of schools in the area A_m^k is then defined as:

$$\rho_m^k = \frac{n_m^k - n_{m-1}^k}{A_m^k} \tag{3.12}$$

Then we compute the average density of schools around any school in Italy as:

$$\langle \rho_m \rangle_k = \frac{\sum_K n_m^k - \sum_K n_{m-1}^k}{\sum_K A_m^k} \tag{3.13}$$

In Fig. 3.13b, we plot $\langle \rho_m \rangle_k$ averaged over all the $K = 8092$ Italian comuni as a function of the radius r_m that goes up to 10^3 Km across the entire Italy. The red line represents the average school-density among *all* the cities in Italy. On average, Italian comuni stand within very dense zones providing almost 1 school per $10km^2$. The dense zones generally last for $10km$ and, after that, a smoothed depletion zone is experienced. However, the average distance between a comune k and a very large city with many schools is about $100km$, accordingly we see a second peak in the average school density at distance $100km$.

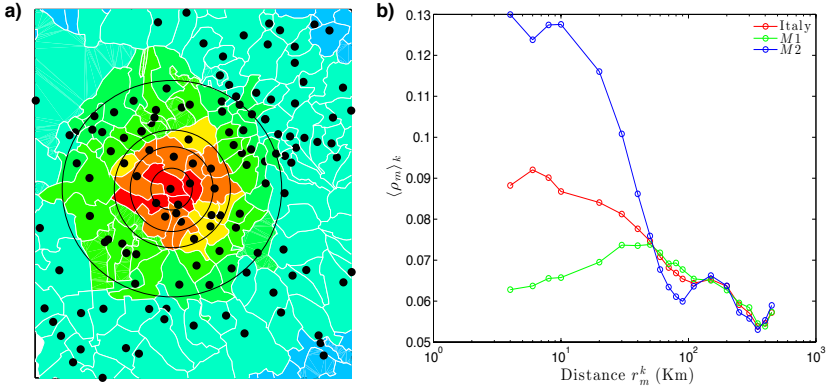

Figure 3.13: Spatial analysis. a. Graphical example for a small comune in Abruzzo of the algorithm used in Fig. 3.13b, based on the Eq. 3.11, 3.12, and 3.13. Different comuni are colored according to the annulus in which they belong. **b.** $\langle \rho_m \rangle_k$ has been plotted for a radius r_m^k of length 10^3 across Italy. The red line draws the trajectory averaging among *all* the cities in Italy. Green and blue lines stand for cities with probability $P_k(x_i > \bar{\mu} | \forall i \in k) \leq 1/2$, labeled $M1$, and $P_k(x_i > \bar{\mu} | \forall i \in k) > 1/2$, labeled $M2$, respectively. Maps generated with Matlab.

The full sample analysis basically averages heterogeneous characteristics that feature different types of comuni. The interaction among schools can be better understood by splitting the sample according to $P_k(x_i > \bar{\mu} | \forall i \in k)$. In Fig. 3.13b, comuni with $P_k(x_i > \bar{\mu} | \forall i \in k) \leq 1/2$, i.e. with predominantly small schools, are named $M1$. The others, with predominantly big schools, are called $M2$.

- $M2$-comuni, the blue line, are (on average) more likely to be surrounded by school-dense cities. They are cities located in densely populated areas (depicted in red in Fig. 3.9) where the school density is large (1.3 schools stand on average within $10km^2$). As far as the distance increases mountainous areas (and hence $M1$-comuni) are encountered and, as a result, the density of schools is found to dramatically decrease.

- The green line describes instead cities labeled $M1$ where a smaller

school density is found. Within $10km$, in fact, almost 1 school every $20km^2$ are encountered on average, about the half of what we find for the $M2$-comuni. This is because $M1$-comuni mainly stand along the countryside (those depicted in blue in Fig. 3.9) where school density slowly increases with distance and reach a maximum at approximately $40km$, which can be interpreted as a typical distance to a densely populated area in a neighboring mountain valley. After this distance the density of schools around $M1$ and $M2$ comuni behave approximately in the same way.

3.5 Supplementary Information

3.5.1 Italian private primary schools versus public primary schools: a comparison.

In this chapter we addressed the source of the bimodality by considering all the Italian primary schools. Here we focus on the potential effect of school type on the school-size distribution. Our dataset collects $N = 17,187$ primary schools in Italy. The fraction of private schools was always low during the past century. In Italy only the 9% of the total of primary school are private.

The main source of primary school privatization within the country is religion. Most of the private schools are venues where education is strictly connected with the Catholic confession. Among the private schools more than 73% are of Catholic inspiration. Straightforward historical roots are expected to explain the location of the Italian Catholic private schools and only marginal are the geographical reasons: private schools are in fact only the 6.54% of the mountain schools.

We define $\mathcal{M}$ the set of comuni k that are in mountains that, according to the Law n. 991/1952, are those that have at least the 80% of their territories above the 600 meters above the sea and an altitude gap between the higher and the lower point not least than 600 meters. Each comune k has n_k schools and a fraction of private schools in this comune defined as $P(i \in \mathcal{P}|\forall i \in k) \equiv \eta_k$, where i is the school ID. We also define the school-size of a private school i that resides in a mountain comune as $x_{i\in\mathcal{P},\mathcal{M}}$. Analogously, $x_{i\in\bar{\mathcal{P}},\bar{\mathcal{M}}}$ stands for the size of a public school residing in a

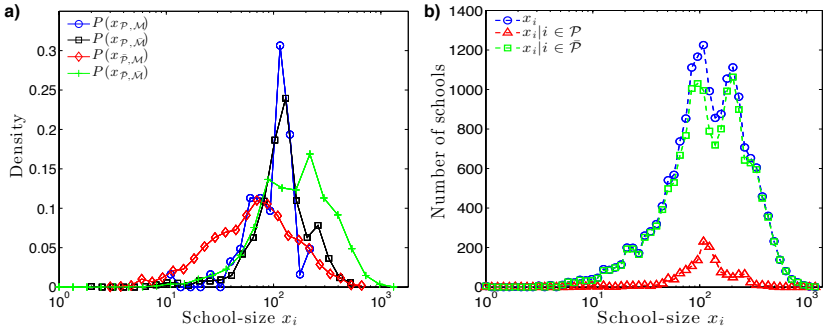

Figure 3.14: **a.** Italian primary school-size distribution disentangled by school type (private, $\mathcal{P}$, versus public, $\bar{\mathcal{P}}$) and geography (mountain, $\mathcal{M}$, versus non-mountain, $\bar{\mathcal{M}}$). **b.** Italian primary school-size distribution by school-type. The blue pattern replicates Fig. 3.1a.

non-mountain comune.

Figure 3.14a shows that neither private mountain schools $(\mathcal{P}, \mathcal{M})$ nor private schools that reside in flat territories $(\mathcal{P}, \bar{\mathcal{M}})$ seem to contribute significantly to the left tail of the school-size distribution. Both the $(\circ)$ blue and the $(\square)$ black lines, respectively, depict two relatively narrow school-size distributions around 100 students per school, the $(+)$ green $(\bar{\mathcal{P}}, \bar{\mathcal{M}})$ and the $(\diamond)$ red lines $(\bar{\mathcal{P}}, \mathcal{M})$. In accordance with the results shown in the main text, mountain public schools mostly contribute to the left tail of the distribution. Finally, the distributions of private schools both for mountain and flat regions are almost identical even though there are only 449 mountain private schools and one might expect large statistical uncertainty.

Figure 3.14b draws the school-size distribution without considering geography but only distinguishing with respect to the school-type. Frequencies are then shown for private (red $\triangle$) and public (green $\square$) schools and compared with the distribution of all the Italian primary schools (in blu $\circ$) that replicates Figure 1a in the main text. It confirms that private schools play only a slight role in generating the left peak which is still present in the the size distribution of public schools.

Figure 3.15a plots the fraction $\eta_c = n_{\mathcal{P},c}/n_c$ of private schools among

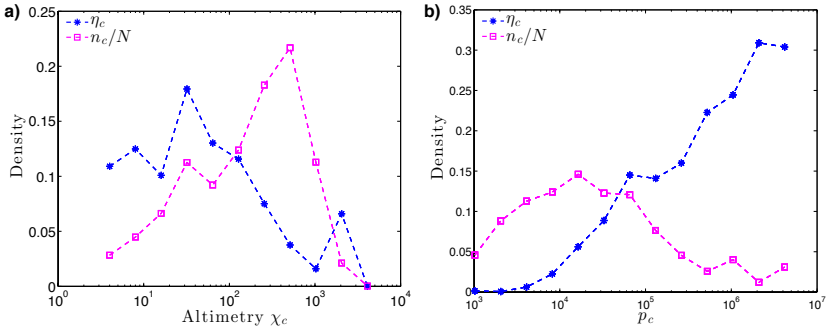

Figure 3.15: Private schools analysis II. a. The fraction of private schools, $\eta_c = n_{\mathcal{P},c}/n_c$ among all schools in bin c with a given altitude above the sea level, χ_c, with respect to the total number of schools n_c in that bin ($*$, blue lines). The distribution of all the schools among the altitude bins n_c/N ($\square$, magenta lines). **b**. The fraction of private schools $\eta_c = n_{\mathcal{P},c}/n_c$ among all schools in bin c, formed by comuni with a given number of inhabitants, p_c. As a robust check we also plot in ($\square$) magenta the distribution of schools (both private and public) in each bin c, n_c/N, versus the population.

all schools (public and private) in each bin c of comuni with given altitude, against their altitude above the sea level, χ_c. In order to reduce the noise, we binned comuni according to the altimetry:

$$c = \{\forall\, k \in [1, \ldots, K] :\, 2^{c-1} < \chi_k \leq 2^c\}. \tag{3.14}$$

It yields 11 bins, $c \in [1, \ldots, 11]$, each of them collecting comuni according to the meters above the sea level. The figure also shows the distribution n_c/N of all the schools among the elevation bins. The figure provides evidence that the majority of private schools (in $*$ blue) are located in the comuni with low altitude $\chi_c < 128m$. In contrast the distribution of number schools with given altitude (both private and public, in $\square$ magenta) reaches the maximum for comuni with altitude $\chi_c = 512m$. The hill-shape of this distribution can be explained by the unequal territory covered by different bins. We conclude that there is a greater fraction of private schools in the planes than in the mountains.

Finally, using the same binning algorithm in Eq. 3.4, Fig. 3.15b shows strong positive correlation between the fraction of private schools $\eta_c =$

$n_{\mathcal{P},c}/n_c$ among all schools in bin c, η_c, of comuni with given number of inhabitants, p_c (in $*$ blue), confirming that the location of the Italian Catholic private schools mainly roots in the more populated comuni.f As a robust check we also plot in ($\square$) magenta the distribution, n_c/N, of all schools (both private and public) among population bins c, which, consistently with the analysis of Fig. 3.4-3.5, yields the Italian city-population distribution which has a slightly skewed shape. In very small comuni ($p_c < 10^4$), where a greater quantity of schools is provided, we count only a small fraction of private ones. Conversely, the fraction of private schools in the large comuni is very large (e.g. private schools constitute 30% of all schools located in Rome, in comparison to the 9% nationwide).

Large flat comuni are then very likely to be the places where most of Italian private primary schools are located. We conclude that privatization has been driven across the years for religious confessional purposes rather than following the unmatched education demand in the countryside due to the lack of the public system.

3.5.2 Testing unimodality in the school-size distributions of flat comuni.

In this section we address concerns on bimodality on the school-size distribution of flat comuni. In the section 3.2.2 we have demonstrated that geography is the main source of bimodality in the school-size pdf showing that mountain schools clusterize around m_1. Yet there might be other confounding factors that might keep a second peak, i.e. m_1, in the school-size pdf of the schools that reside in flat comuni.

In Fig. 3.6b we distribute schools according to the number of students, x_i, conditional on the altimetry of comuni. As we discuss in the section 3.2.1 this analysis gives five distributions which correspond to different elevation bins. The PDFs of mountain schools stand on the left and on the right we have flat schools. The ($\circ$) green line shows the school-size distribution for $N_{250m} = 3,033$ schools that reside in comuni with around 250 meters from the sea level. Despite this PDF does not show a sharp peak corresponding to m_2, and thus potentially might be bimodal, here we demonstrate that *statistically* the hypothesis in favor of

unimodality can not be rejected.

To see that we use the complementary error function to estimate the probability that the number of schools in the central bin n_1 is not significantly smaller than and the numbers of schools in the neighboring two bins n_2, n_3 are not significantly larger than a certain number n^* provided that the standard deviation of the number of schools in these bins due to small statistics is $\sqrt{n^*}$:

$$p(n^*) = \frac{1}{2}\Pi_i \text{erfc}\left(\frac{|n_i - n^*|}{\sqrt{2n^*}}\right) \tag{3.15}$$

This is equivalent to test the hypothesis that the distribution is unimodal. In the school-size distribution for schools that reside in comuni with around 250 meters above the sea, the central bin collects $n_1 = 639$ schools. On either sides there are two other bins that collect $n_2 = 670$ and $n_3 = 646$ respectively. The probability that the distribution is not bimodal is maximum for $n^* = 646$ where it is equal to $p_{max}(n^* = 646) = 0.15$. Fixing a level of confidence of 0.10 we, therefore, cannot reject the hypothesis of unimodality.

Chapter 4

Diversification versus specialization in complex ecosystems

4.1 Introduction

Countries and firms are fundamental actors sharing complex economic and social ecosystems. Their evolutive paths lead to structurally different scenarios: firms are specialized entities while countries, as recently shown, are diversified [27, 28]. This raises a question on the mechanisms driving specialized entities to organize themselves into diversified super-structures. Is diversification a matter of size, of time horizon, or both? Are there other hidden dimensions governing the diversification process?

A similar scenario holds in biological ecosystems [71]: species (firms) tend to be substantially specialized, while groups of species competing on the same ecosystem (countries), appear to be diversified. Inspired by this argument in this paper we investigate the key mechanisms this picture is grounded on. It has been recently shown that this kind of analogy between economic and biological systems could gives rise to fruitful insights on elementary mechanisms [72].

Identifying the diversification drivers at the various scales is a challenging task in all disciplines since diversification processes are ubiquitous in nature [73] and economic systems [74, 75]. In our view economic ecosystems represent an ideal (paradigmatic) playground for an empirical investigation.

We therefore analyze the distribution of revenues across production sectors of quoted firms aggregated by country (Bloomberg database [76]). Not surprisingly the analysis confirms that country competitiveness is mainly driven by diversification of productive systems, while firms' competitiveness is mainly a matter of specialization. The macroscopic signature of these macro-micro level discrepancies is reflected by the nested triangular structure of the country-sector binary matrix contrasting the essential randomness of the firm-sector binary matrix.

We argue that this is a specific observation of a general feature of complex systems: the shift from the macro to the micro level generally entails the loss of those features characterizing the former level. As in biology [77], the emerging diversification at macro level cannot be properly addressed at the level of individual species/firms. However, the environment in which the micro level is embedded preserves a sort of a macro level *memory* which enables to identify those micro level features that could emerge at larger scales [78].

Guided by this idea we show that, in the specific case of economic ecosystems, the microscopic feature emerging ad the macro scale is the firm's diversification barrier α (see fig. 4.1). Moreover the α's of different countries aggregate on macro-regional (multi-country) scale. This *zoom-in zoom-out* framework thus enables the identification of the proper micro-variable selecting the emerging (aggregated) macro-properties. This is of particular relevance in socio-economic systems, since it may help decision-makers to select the correct variable to be acted upon at the (micro) specialized level, in order to achieve desirable results at the (macro) diversified level.

We are confident that this type of macro/micro level of exchange of information is a general property of competitive environments, not confined to economic ecosystems. However, in economic ecosystem the empirical identification of diversification drivers is simplerfor the following reasons:

1. Micro and macro level for social ecosystems are the result of an artificial evolution induced by mankind unlike other ecosystems driven by Nature's laws, making it easier to identify the *right* micro/macro information exchange.

2. Economic datasets, as a result of the previous consideration, are very large and highly standardized.

3. From the lowest to the highest level of aggregation (individuals, firms, industrial districts, regions, countries, macro-regions) there exists a unique underlying metrics: the economic value conventionally measured in terms of capital.

In this respect the traditional economic literature has extensively studied the effect of institutions, policies and economic environments under which diversification has an impact on firm performance [23, 24, 25]. However, the general picture which emerges from the standard approach is usually non conclusive as to whether diversification patterns affect firm performances. Instead as mentioned, in the present work, we find that firm performances are correlated to diversification, but the signature of this correlation appears in a highly non-trivial way as a selection rule which prevents firms from occupying a part of the diversification-revenues plane. We argue that the subtleness of this dependence - highly diversified firms are necessarily also highly performing while highly performing, firms can be both diversified or not - is at the basis of the strongly debated economic literature about this field.

We also propose a simple mathematical model mimicking the firms diversification dynamics in which firms evolve via a random walk in a random potential. Firm's survival rate depends on the values of the potential in the state reached by a particular firm (firm performance) through a parameter. This parameter models the toughness of the economic environment in which firms compete. Surviving firms tend to diversify in time with a given probability. Such a minimal model is able to reproduce the main features observed in the data analysis.

4.2 Results

The dataset we use consists of annual revenues of quoted firms disaggregated into Bloomberg's sector code and downloaded in May 2013. The database contains about 38000 firms and about 2000 sectors.

We proceed similarly to the work of [28] where an archival export dataset is considered to measure intangible assets determining the competitiveness of countries. It is worth noticing that in both analyses the datasets were not collected with the purpose of the analyses in which they were subsequently used.

As previously mentioned, the identification of the diversification drivers at the various scales is a challenging task in all disciplines. In Economics, in particular, it is unclear, but crucial, how the dynamics at micro-level determines the one at the macro-level and *vice versa*. This paper aims to shed some light on this very relevant question which affects how the economy should support the concrete implementation of economic policy decisions with a more scientific grounding.

The analysis confirms the recent finding [28] that, contrary to classical predictions [79], country competitiveness is mainly driven by diversification of productive systems.

Coherently with the evidence of a triangular structure of country-product matrix in [27, 28, 30, 31], in the present analysis the same triangular feature is also found in the country-sector matrix obtained by aggregating firms on the basis of its legal address (see sec. 4.5). The same matrix constructed at the firm level looses its nestedness and is similar to a random matrix with the same density (for further discussion see sec. 4.6), reflecting firm specialization. This raises a rather fundamental question: what is the mechanism that organizes the information present into an almost random matrix, at the firm's level, in a nested matrix, at the country level?

To address this issue - within the general specialization trend for companies - we investigate whether there exist non trivial and country dependent patterns of diversification. We identify in the *revenue diversification barrier* (hereafter α) the micro signature of these country dependent patterns. It is interesting to note that this barrier α organizes itself at even higher level: this barrier tends to reflect geographical vicinity and

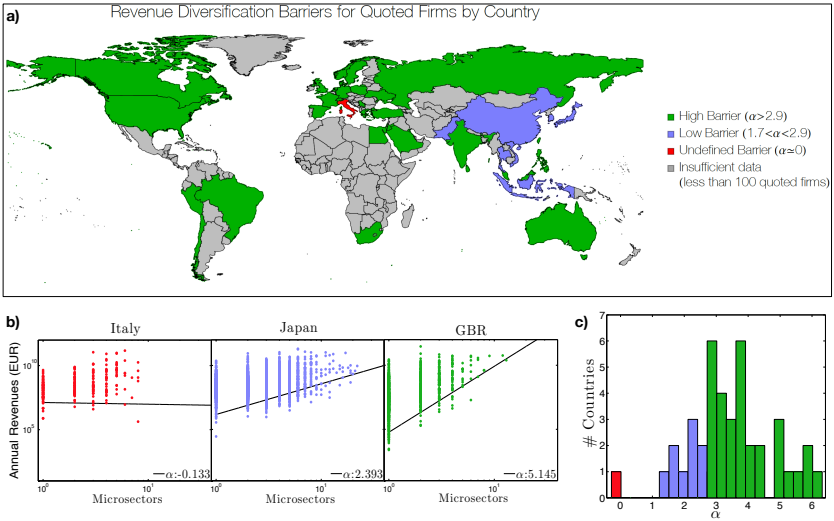

Figure 4.1: **a.** The worldwide distribution of the *revenue diversification barrier* α. The α tends to reflect the geographical vicinity and to cluster at macro regional level. **b.** The scatter plot of firm revenues against firm diversification for thee paradigmatic countries. Except for Italy the data draw a peculiar shape with an evident lower boundary. The angular coefficient of this linear boundary is what we define as the *revenue diversification barrier* α. **c.** The histogram of α. Colors are consistent with those used in panels a. and b.

to cluster at macro regional level. This can be observed in Fig. 4.1 panel a where we report worldwide distribution of α.

In panel b we report the scatter plot of firms' revenues (measured in EUR) against the firm diversification for three paradigmatic countries. With the exception of Italy, for all countries for which data are significant we observe a peculiar shape in which a clear lower boundary appears in the scatter plot. This means that while firms with high revenues can be either diversified or not, revenues of diversified firms are necessarily higher than non-diversified one. This suggests the existence of a *revenue diversification barrier* necessary to successfully diversify in a competitive market. In the double logarithmic space, the stiffness of this lower en-

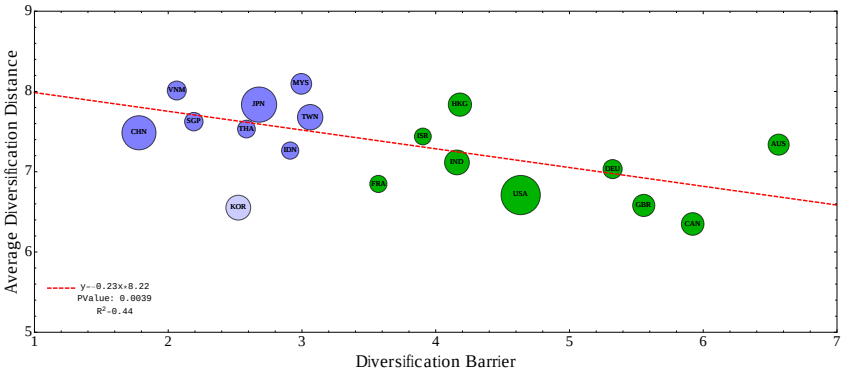

Figure 4.2: Diversification distance against revenue diversification barrier. The plot shows a clear negative correlation between these two variables. Blue and Green markets are clearly separated by both variables suggesting that firms in diversification-prone markets tend to diversify more and more coherently (i.e. with a smaller diversification distance). South Korea (lighter blue) appears to be an outlier and removing it from the regression improves the quality of the fit (PValue decreases and R^2 increases)

velope naturally defines the barrier (for further details on the definition and robustness of the measure of α see sec. 4.7).

In panel c we show the evidence for the nontrivial geographical clustering of the values of α. All the countries with low diversification barriers (blue) appear to belong to the Asian macro area with the notable exception of India, Hong-Kong and the Philippines. We speculate that these blue colored markets share a higher tolerance to diversification. In fact the diversification success of a firm is the result of the evolution in a competitive environment. The nature of this competition determines the stiffness of the barrier. On the other hand, the firms competing in green-colored markets are embedded in an environment which is operating a stronger selection of firms and consequently are characterized by a lower survival rate with respect to their diversification opportunities. Despite the fact that India, Hong-Kong and the Philippines are Asian countries, it is not surprising to find them among stiff markets because their value of α may reflect the strong anglo-saxon imprinting of the eco-

nomic organization of these countries. Italy features an economy with different diversification dynamics. The substantially 0 value of α characterizing this market may mean that firm diversification is not driven by market selectiveness but rather by exogenous (with respect to this scheme) mechanisms.

To further characterize blue and green markets and consequently firm diversification patterns, we analyze the relation between α and the average diversification coherence of firms. The average diversification coherence is related to the typical distance among occupied sectors by a firm: the greater this typical distance, the lower the coherence (mathematical details of the definition of this measure are provided in sec. 4.8). These two variables prove to be anti-correlated as shown in Fig. 4.2, indicating that the difference between blue and green markets is not only a matter of diversification barrier but also of diversification structure: firms operating in green markets tend to have revenues in sectors which are closer than those of firms living in blue markets. In terms of diversification, green markets are characterized by more coherent firms supporting the argument that selection rules are stricter in these economic systems.

4.3 Model

We propose an extremely simplified model that embodies in our view the minimal traits necessary to shed light on the meaning of the revenue diversification barrier α. Firms are mimicked as random walkers moving in a potential, seeking local minima. The height of such minima is representative of a firm's performance: the lower the value of the potential, the better the performance. Markets (countries) differ in their tolerance (τ) with respect to poor performances, i.e. in the probability of a firm to fail given its level of performance. Surviving firms, i.e. those with good performances, have the chance (P_{div}) to diversify, while failed firms are replaced with new ones with the lowest possible level of diversification. Mathematical details on the implementation of the model are provided in the sec. 4.9. By making an analogy between the performance as defined in the present model and the revenues of a firm, we can observe in Fig. 4.3 how the model produces patterns very similar to those observed

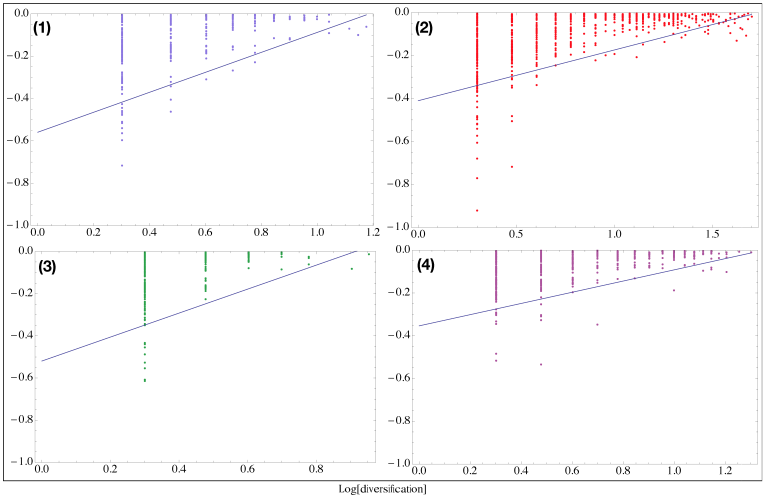

Figure 4.3: Performance versus diversification in the model. It is possible to observe a lower boundary extremely similar to those observed in the real data, even in its functional form. The numbered labels indicate respectively the phase zone in Fig. 4.4

in the real dataset. Interestingly there is still a linear lower bound in the doubly logarithmic diversification vs. performance scatter plot. Within this model the diversification is clearly proportional to the life span of a given firm. The similarity between real data scatter plot and the model produced data can thus be interpreted in view of the question raised in the introduction: diversification is a dynamic process that develops over time and the boundary in the diversification-performance relation is set by the competitiveness of the environment in which the economic entities are immersed. In other words what we observe in real data is compatible with diversification being a dynamic process that goes on as long as a firm is able to survive. How long it will survive given its profits depends on the tolerance of the ecosystem. The differences in tolerance generates the differences in the diversification boundaries that we observe across countries.

The values of α have a clear dependence on τ and P_{div} as shown in the

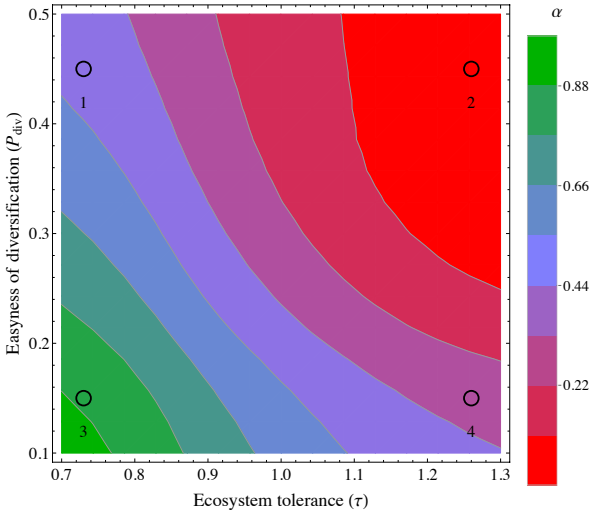

Figure 4.4: The phase diagram of the model obtained numerically. The diversification barrier α decreases in tolerant ecosystems and with increasing easiness of diversification P_{div}. The numbers indicate the phase diagram zones explored by the model "countries" whose scatter plot of performance versus diversification are reported in figure 3. As green zone are populated by high diversification barrier "countries", while purple zone by the lower barrier "countries".

phase diagram in Fig. 4.4. In particular α decreases when the ecosystem tolerance increases. P_{div} acts as a simple multiplier of the life span of a firm in determining its diversification.

4.4 Conclusions

The analysis of the distribution of firm revenues across production sectors aggregated by country manifests a peculiar triangular shape. This enables us to define a country dependent revenue diversification barrier "α", which represents a novel macroscopic dimension driving the microscopic diversification process.

We have shown that this new macro feature shows a non trivial geographical clustering, which points out the importance and implication of the geo-political environment in the diversification patterns. α can be interpreted as the microscopic signature responsible for micro-macro information exchange showing that though the economic complexity methods it is possible to single out the microscopic variables governing the macroscopic dynamic.

Within our finding the microscopic firms' differentiation dynamics can be interpreted as a *"Darwinan"* competitive process in which the firms survival to diversification depends on the characteristics of the macroscopical (country like) environment. To further confirm this picture, a time dependent analysis on similar data is called for. Moreover, to better understand the meaning of this newly introduced dimension α, a comparison with other economic country indicators could also be implemented in the future.

4.5 Data definition

4.5.1 BICS hierarchical Classification system

The Bloomberg Industry Classification Systems (BICS) is a proprietary hierarchical classification system, which classifies firms' general business activities.

BICS for stock companies contains 10 *macro* sectors, which represent the broadest classification of general business activities. Each sector is further broken down into a hierarchical system of sectors (up to 8 levels of detail), which are classified into more narrowly defined business activities. The whole classification system counts up to 2294 unique sectors. Each *macro* sector is defined by a code composed by two digits. Sectors (or subsectors) are hierarchically defined by attaching further couples of digits to a parent element code. The deepest sector is defined by a 16-digits code. The figure 4.5 shows the first two levels of the Bloomberg BICS stocks hierarchical classification system for stock companies. The figure 4.6 shows the hierarchical tree system of BICS.

Level 1		Level 2	
CODE	**Macro Sector**	**CODE**	**First Level Microsector**
10	Communications	1010	Media Content
		1011	Telecom
11	Consumer Discretionary	1110	Apparel & Textile Products
		1111	Automotive
		1112	Consumer Discretionary Srvcs
		1113	Distributors
		1114	Home & Office Products
		1115	Leisure Products
		1116	Recreation Facilities & Srvcs
		1117	Retail Discretionary
		1118	Travel, Lodging & Dining
12	Consumer Staples	1210	Consumer Products
		1211	Dist/Whsl - Consumer Staples
		1212	Retail Staples
13	Energy	1310	Oil, Gas & Coal
		1311	Renewable Energy
14	Financials	1410	Asset Management
		1411	Banking
		1412	Institutional Financial Srvc
		1413	Insurance
		1414	Real Estate Oper & Srvcs
		1415	REIT
		1416	Specialty Finance
15	Health Care	1510	Biotech & Pharma
		1511	Health Care Facilities/Srvcs
		1512	Medical Equipment/Devices
16	Industrials	1610	Aerospace & Defense
		1611	Electrical Equipment
		1612	Engineering & Const Srvcs
		1613	Industrial Distribution
		1614	Machinery
		1615	Manufactured Goods
		1616	Transportation & Logistics
		1617	Transportation Equipment
		1618	Waste&Envrnmt Srvc Equip&Fac
17	Materials	1710	Chemicals
		1711	Construction Materials
		1712	Containers & Packaging
		1713	Forest & Paper Products
		1714	Iron & Steel
		1715	Metals & Mining
18	Technology	1810	Design, Mfg & Distribution
		1811	Hardware
		1812	Semiconductors
		1813	Software
		1814	Technology Services
19	Utilities	1910	Utilities

Figure 4.5: first two levels of the Bloomberg BICS stocks hierarchical Classification system for stock companies.

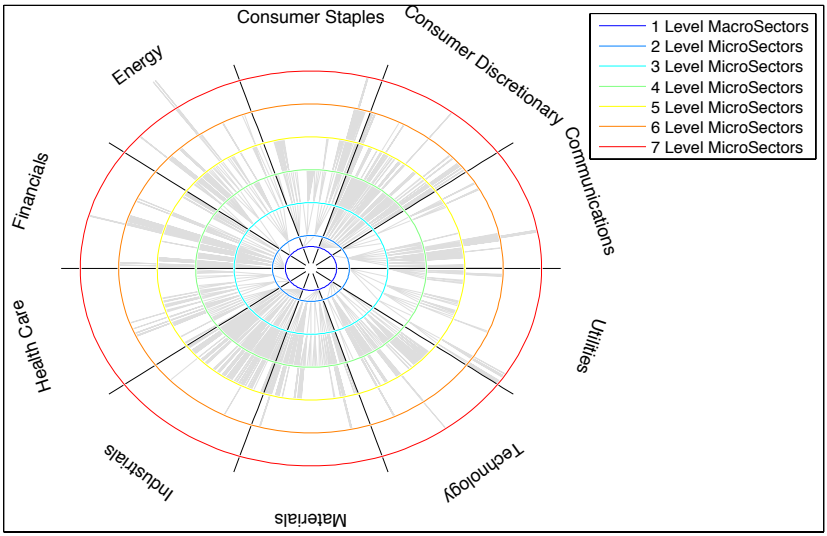

Figure 4.6: The Figure shows the reconstruction of BICS hierarchical tree system with its 2294 hierarchical Sectors. The colored lines show the classification levels.

4.5.2 Dataset

The Bloomberg platform collects data about firms' revenues categorized using BICS. The data used in this work have been downloaded from Bloomberg platform on May 2013. The data contains information on annual revenues of 38274 traded firms broken down in 2294 sectors. The dataset covers 99 stock markets. Each firm is associated to its country of domicile, as reported by Bloomberg. the country of domicile is the country where the firm senior management is established legally. All the analyses have been performed on countries where at least 100 traded companies are domiciled. Those countries are listed in Table 4.1.

4.5.3 Data sanitation

We organize our data in a firm-sector rectangular matrix. Most of the elements of this matrix are 0s. We perform various data sanitation pro-

Country	# Companies	Country	#Companies
USA	4415	SWE	272
JPN	3366	POL	250
CHN	3112	ITA	245
TWB	1433	CHE	231
KOR	1341	GRC	213
IND	1338	LKA	210
HKG	1115	ZAF	198
CAN	1035	SRB	176
GBR	970	PHL	170
AUS	834	EGY	154
MYS	794	CHL	154
DEU	620	KWT	153
VNM	605	UKR	152
RUS	567	DNK	142
SGP	554	JOR	140
THA	483	NOR	135
IDN	439	ESP	122
FRA	437	NDL	118
ISR	406	PAK	115
TUR	297	PER	113
BRA	276	SAU	109

Table 4.1: Countries where at least 100 traded companies are domiciled.

cedures in order to extract relevant information from such matrix.

From a total of 216888 non-0 entries we eliminate 7333 negative revenues, since their meaning appears unclear. In particular we notice an abnormal number of repeated identical values, which clusterize on a national basis but span on different firms. The dataset provides for each subsector the aggregate of the firms' revenues on that particular subsector plus the sum of the revenues of the full hierarchy of subsectors below it. We perform the necessary subtractions to ensure that only the pertaining revenue is assigned to each subsector.

4.6 Triangularity vs. randomness

The firm diversification level is the number of sectors developed by the firm. The real binary firm-sector matrix has a density close to 0.05. We generate a random matrix with same size and density of the real one. In

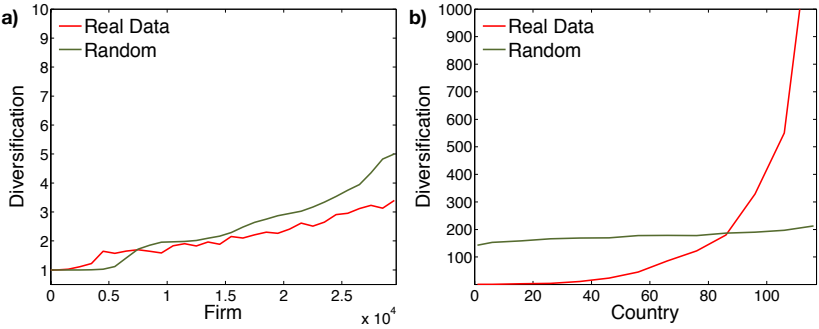

Figure 4.7: Comparison between the real data (red) and a random realization with same density (green). **a.** The firm-sector matrix exhibits a pattern similar to a random case emphasizing the firm' specialization. **b.** On the contrary aggregating the data on country level a non-random pattern emerges, corresponding to the presence of a nested structure.

figure 4.7a we show a comparison of the firm diversification, sorted by fitness [28], between the real data (depicted in red) and the random case (green). The two diversification trends show a similar pattern. This outlines the firms' high specialization and the absence of triangular structure in the matrix. Instead, in Fig. 4.7b, the real country-sector matrix, generated aggregating firms at country level on the basis of the legal address, exhibits a clearly nested (triangular) structure such as the country-product matrix [28].

4.7 Definition of the revenue diversification barrier and its robustness

The diversification barrier α is measured as the slope of the lower boundary of the scatter plot of diversification vs. revenues in logarithmic space. The lower boundary is defined as the lower 5th percentile of the distribution of revenues for a given diversification level.

We check the sensitivity of α with respect to a variation of the percentile used to define the lower bound.

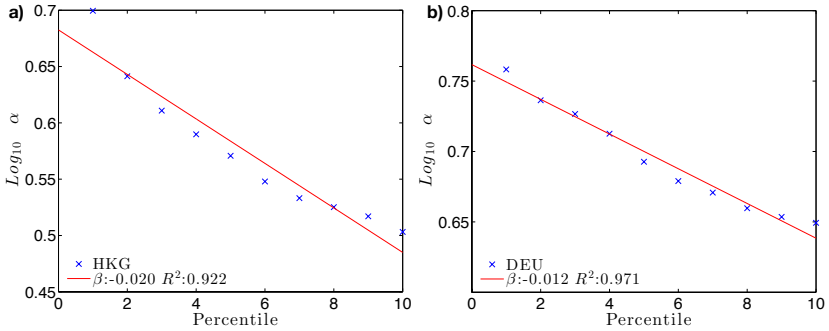

Figure 4.8: Dependency of α on different percentile cut-offs for two sample countries. The decay is well fitted by an exponential law $y = Ae^{\beta x}$ for all the countries examined. Values of β from regressions are shown in fig. 4.9

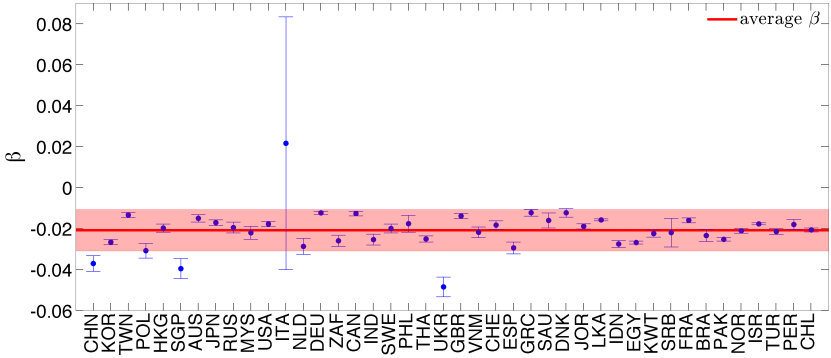

Figure 4.9: Each blue dot represents the coefficient β and its standard error for a specific country. The solid red line is the average value of β on all countries with more than 100 quoted firms. The shaded area in the plot marks one standard deviation. Most of the countries display a consistent decay of α with the percentile used thus making the particular choice of a percentile not relevant.

In fig. 4.8 different values of α for different percentiles are shown, for each country with at least 100 quoted firms. The plot clearly shows a decay trend which is common to (almost) all the countries. We then study

in detail this decay of α. In figure 4.9 we show the angular coefficient (β) of a linear regression between the logarithm of α and the percentile, together with the respective standard error, for each country. For the majority of the countries β lies within one standard deviation from the average (red solid line). This shows that the consistency of our analysis is not affected by a particular choice of the percentile. Italy shows an anomalous sensitivity dependence with respect to other countries. The χ^2 test over the β regressions in the fifth percentile accept the linear hypothesis at 95% for all the countries.

4.8 Diversification coherence

As mentioned, the BICS classification itself defines a topological distance between the codes, more precisely a tree. Each node in the tree corresponds to a more fine specification of the parent element.

Relying on this information we want to develop a measure of how coherently a firm is diversified. In particular we want to be able to weight diversification by a distance among the BICS categories in which diversification occurs: a company diversified in many very close subsectors might be considered less diversified than a company which has revenues only in two very distant sectors.

To this purpose we must take into account the fact that having revenues in a given sector and in one of its subsectors, at any level, does not add to the diversification. For this reason we cannot use the simple topological distance defined by the hierarchical tree implied by the BICS codes. Our approach is to define a new directed network, which is derived from the relations present in the BICS categorization, but with appropriate distances (or link weights). On such a network we use the total weight of the minimal (directed) spanning tree between all the nodes in which a company has revenues as a measure of its coherency.

To this end we need to define a distance (or link weights) that needs to have the following properties:

1. The distance between a sector and one of his subsectors must be 0 (producing pens and red pens does not add to diversification)

2. The distance between two subsectors of the same sector is propor-

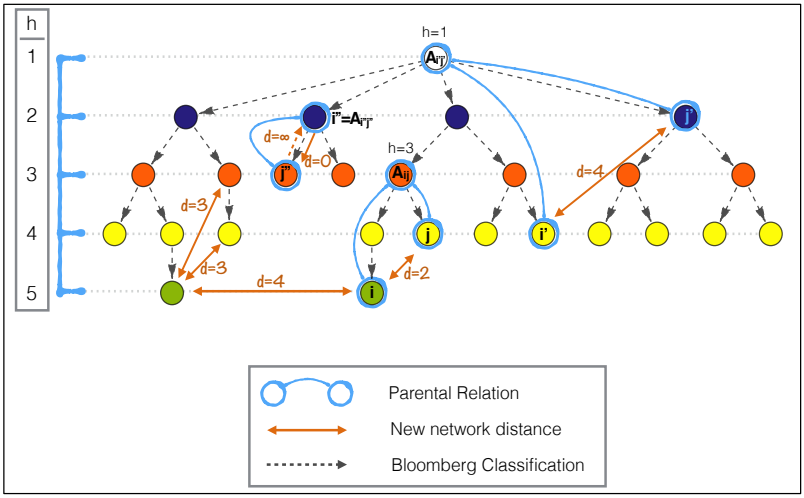

Figure 4.10: Network distance. Examples of network distance as defined in Eq. 4.1

tional to the depth of the two subsectors (red pens and blue pens are more far apart than red pens with wooden body and red pens with plastic body)

3. As a consequence of the first property the distance between two sectors (A and B) and two of their respective subsectors (Aa and Bb) must be the same (pens are as distant from rulers as red pens are from metal rulers)

4. The distance between a subsector and its parent element sector must be infinite (to avoid 0 cost spanning trees between subsectors).

As depicted in Fig. 4.10 this translates in the fact that the distance between two nodes must be a function of depth of the nearest common parent element, except when one of the two nodes is a subsector of the other one, in which case the distance is asymmetric (0 or ∞). In formulae the distance is written as follows:

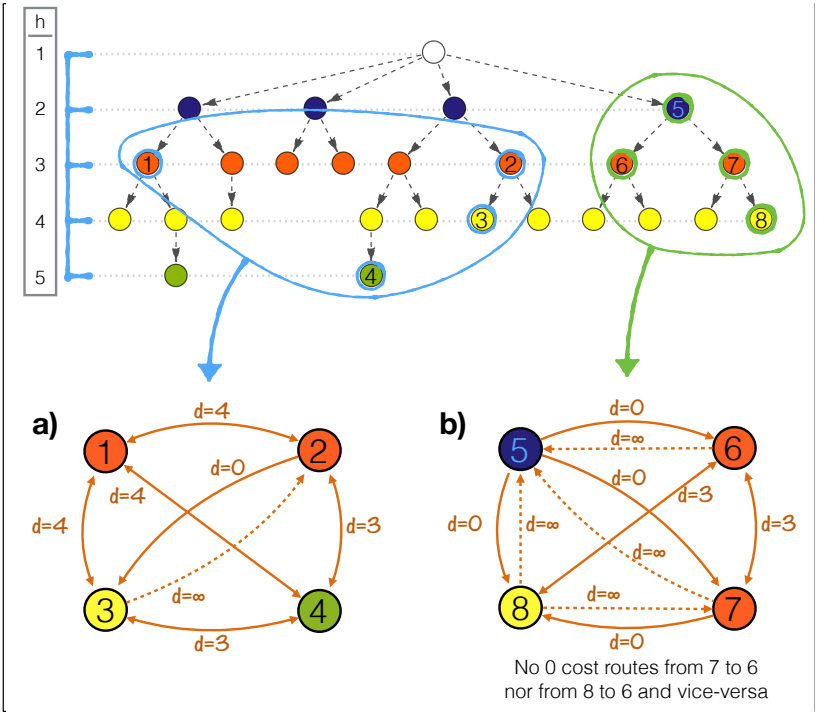

Figure 4.11: Network examples. The resulting networks with link weights equal to $d_{i,j}$ for two hypothetical situations are shown in panels a and b. On these networks minimal spanning trees are determined via the Chu-Liu/Edmond's algorithm

$$d_{i,j} = \begin{cases} H - h(A_{i,j}) & \text{if } A_{i,j} \neq i \wedge A_{i,j} \neq j \\ 0 & \text{if } A_{i,j} = i \\ \infty & \text{if } A_{i,j} = j \end{cases} \tag{4.1}$$

where $A_{i,j}$ is the nearest common janitor to the nodes i and j, $h(A_{i,j})$ is its depth in the tree and H is the total depth of the tree plus 1. The application of this definition is illustrated in Fig. 4.11 where the resulting networks, with link weights equal to $d_{i,j}$, for two hypothetical situations

are shown in panels a and b. On these networks minimal spanning trees are determined via the Chu-Liu/Edmond's algorithm [80, 81, 82].

4.9 Mathematical model for the diversification barrier

The model described in the main text provides a very simple mechanism that can explain the emergence of the diversification barrier α together with its functional form. A firm is depicted as a random walker seeking local minima in a random potential. The random potential is a realization of a simple gaussian discrete random walk, with 0 mean and unit variance. We generate 100 equally spaced discrete points of the potential. The potential $V(x)$ is then made periodic via a reflection, and is made continuous via a linear interpolation, the period being 200. Thus $V(x) = V(x + k * 200)$ holds for any real x and for any integer k. Finally $V(x)$ is scaled to have maximum equal to 1 and minimum equal to 0.

The firms' dynamics is implemented as follows. Each firm starts at a random x_0 coordinate and is made to evolve as a brownian particle in the potential defined by $V(x)$. It seeks for local minima by evolving with the Metropolis-Hastings algorithm.

At each time step a proposal $x_t^{(p)}$ for a new value of x_t is drawn from a gaussian distribution $N(x_t, \sigma)$. The parameter σ needs to be chosen such that the typical jump distance for a firm will be inside a typical local minima. This typical width is of order 1, by construction, thus we have chosen $\sigma = 0.1$. The proposal is then accepted with probability $P = e^{(V(x_{t-1}) - V(x_t^{(p)}))/T}$. If the proposal is accepted we set $x_t = x_t^{(p)}$ else $x_t = x_{t-1}$.

We define the performance of a firm as $P(t) = 1 - V(x_t)$. Every 100 time-steps we compute the average performance $\overline{P}$ in such time window: the firm either survives with probability $1 - \overline{P}^{\tau}$ or fails. If the firm survives it has the chance to increase its diversification of 1, with probability P_{div}.

References

[1] Jacobson, M. J. & Wilensky, U. Complex systems in education: Scientific and educational importance and implications for the learning sciences. *The Journal of the learning sciences* **15**, 11–34 (2006). 1

[2] Bourgine, P., Johnson, J. & Chavalarias, D. The css roadmap for copmlex systems science and its application 2012-202 (2012). 1

[3] Arthur, W. B. Complexity and the economy. *science* **284**, 107–109 (1999). 2

[4] Pietronero, L. Physicists get social. *Nature Physics* **6**, 641–640 (2010). 2, 7

[5] Barabási, A.-L. The network takeover. *Nature Physics* **8**, 14 (2011). 2, 7

[6] Vespignani, A. Modelling dynamical processes in complex socio-technical systems. *Nature Physics* **8**, 32–39 (2011). 2, 7

[7] Di Clemente, R. & Pietronero, L. Statistical agent based modelization of the phenomenon of drug abuse. *Scientific Reports* **2**, 1–8 (2012). 2

[8] Cristelli, M., Pietronero, L. & Zaccaria, A. Critical overview of agent-based models for economics. *arXiv preprint arXiv:1101.1847* (2011). 2, 7

[9] Chakraborti, A., Toke, I. M., Patriarca, M. & Abergel, F. Econophysics: Empirical facts and agent-based models. *arXiv preprint arXiv:0909.1974* (2009). 2, 7

[10] Agar, M. H. & Wilson, D. Drugmart: Heroin epidemics as complex adaptive systems. *Complexity* **7**, 44–52 (2002). 2, 9

[11] Perez, P. *et al.* Monograph 11 simdrug exploring the complexity of heroin use in melbourne (2005). 2, 9, 16, 28, 29

[12] Belmonte, A., Di Clemente, R. & Buldyrev, S. V. The italian primary school-size distribution and the city-size a complex nexus. *Scientific Reports* **4**, 1–11 (2014). 3

[13] Axtell, R. L. Zipf distribution of us firm sizes. *Science* **293**, 1818–1820 (2001). 4, 36

[14] Stanley, M. H. *et al.* Scaling behaviour in the growth of companies. *Nature* **379**, 804–806 (1996). 4, 36

[15] Eeckhout, J. Gibrat's law for (all) cities. *The American Economic Review* **94**, 1429–1451 (2004). 4, 39, 42

[16] Gabaix, X. Zipf's law for cities: an explanation. *The Quarterly journal of economics* **114**, 739–767 (1999). 4, 31, 39, 42

[17] Gabaix, X. Power laws in economics and finance. Tech. Rep., National Bureau of Economic Research (2008). 4, 31

[18] Allen, P. M. *Cities and regions as self-organizing systems: models of complexity* (Psychology Press, 1997). 4, 31, 39, 42

[19] *Decreto del Presidente della Repubblica del 20 marzo 2009 n. 81,* (Gazzetta Ufficiale, 2 luglio 2009). 4

[20] *Disposizioni concernenti la riorganizzazione della rete scolastica, la formazione delle classi e la determinazione degli organici del personale della scuola.* (Decreto Ministeriale 331, 24 luglio 1998). 4

[21] Chakrabarti, A., Singh, K. & Mahmood, I. Diversification and performance: evidence from east asian firms. *Strategic Management Journal* **28**, 101–120 (2007). 5

[22] Guillen, M. F. Business groups in emerging economies: A resource-based view. *academy of Management Journal* **43**, 362–380 (2000). 5

[23] Fauver, L., Houston, J. & Naranjo, A. Capital market development, international integration, legal systems, and the value of corporate diversification: A cross-country analysis. *Journal of Financial and Quantitative Analysis* **38**, 135–158 (2003). 5, 65

[24] Campa, J. M. & Kedia, S. Explaining the diversification discount. *The Journal of Finance* **57**, 1731–1762 (2002). 5, 65

[25] Khanna, T. & Palepu, K. Is group affiliation profitable in emerging markets? an analysis of diversified indian business groups. *The Journal of Finance* **55**, 867–891 (2000). 5, 65

[26] Claessens, S., Djankov, S., Fan, J. P. & Lang, L. H. *Corporate diversification in East Asia: the role of ultimate ownership and group affiliation*, vol. 2089 (World Bank, Financial Operations Vice Presidency, Financial Economics Unit, 1999). 5

[27] Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences* **106**, 10570–10575 (2009). 5, 6, 63, 66

[28] Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & Pietronero, L. A new metrics for countries' fitness and products' complexity. *Scientific reports* **2** (2012). 5, 6, 63, 66, 76

[29] Balassa, B. 'revealed'comparative advantage revisited: An analysis of relative export shares of the industrial countries, 1953–1971. *The Manchester School* **45**, 327–344 (1977). 5

[30] Cristelli, M., Gabrielli, A., Tacchella, A., Caldarelli, G. & Pietronero, L. Measuring the intangibles: A metrics for the economic complexity of countries and products. *PloS one* **8**, e70726 (2013). 6, 66

[31] Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & Pietronero, L. Economic complexity: conceptual grounding of a new metrics for global competitiveness. *Journal of Economic Dynamics and Control* **37**, 1683–1691 (2013). 6, 66

[32] Di Clemente, R., Chiarotti, G., Cristelli, M., Tacchella, A. & Pietronero, L. Diversification versus specialization in complex ecosystems. *Submitted* (2014). 6

[33] Dipendenze, P. D. Ricerca - studio "previsione dell'evoluzione dei fenomeni di abuso". bollettino previsionale previsione 2012. Tech. Rep., Prevo.Lab, MI (2009). 8, 10, 11, 13, 26, 28

[34] Dipendenze, P. D. Ricerca - studio "previsione dell'evoluzione dei fenomeni di abuso". area ricerche sul mercato. Tech. Rep., Prevo.Lab, MI (2009). 8, 10, 11, 22, 26, 28

[35] Dipendenze, P. D. Ricerca - studio "previsione dell'evoluzione dei fenomeni

di abuso". area interviste testimoni privileggiati strumento prevo.meter. Tech. Rep., Prevo.Lab, MI (2009). 8, 16, 26, 28

[36] Palla, G., Barabási, A.-L. & Vicsek, T. Quantifying social group evolution. *Nature* **446**, 664–667 (2007). 9, 11, 22

[37] Levin, G., Roberts, E. B. & Hirsch, G. B. *The persistent poppy: A computer-aided search for heroin policy* (Ballinger Publishing Company Cambridge, Massachusetts, 1975). 9

[38] Everingham, S. S., Rydell, C. P. & Center, A. *Modeling the demand for cocaine* (Rand, 1994). 9

[39] Agar, M. H. Agents in living color: Towards emic agent-based models. *Journal of Artificial Societies and Social Simulation* **8** (2005). 9

[40] Jones, R., Morris, K. & Nutt, D. Drugs futures 2025? foresight: Brain science, addiction and drugs state of science review (2005). 9

[41] Association, A. P. & on DSM-IV., A. P. A. T. F. *Diagnostic and statistical manual of mental disorders: DSM-IV.* (Amer Psychiatric Pub Inc, 1994). 10

[42] Limpert, E., Stahel, W. A. & Abbt, M. Log-normal distributions across the sciences: keys and clues. *BioScience* **51**, 341–352 (2001). 10

[43] Caldarelli, G. Scale-free networks: complex webs in nature and technology. *OUP Catalogue* (2007). 11, 22

[44] Pastor-Satorras, R. & Vespignani, A. Epidemic spreading in scale-free networks. *Physical review letters* **86**, 3200 (2001). 11, 22

[45] Smart, R. G., Whitehead, P. C. & Laforest, L. The prevention of drug abuse by young people: an argument based on the distribution of drug use. *Bulletin on Narcotics* **23**, 11–15 (1971). 11

[46] EMCDDA. 2009 national report (2008 data) to the emcdda, new development, trends and in-depth information on selected issues. Tech. Rep., Reitox Italian Focal Point & Presidenza Del Consiglio Dei Ministri Dipartimento delle politiche antidroga (2009). 13, 16

[47] Dei Ministri, P. D. C. Relazione annuale al parlamento sullo stato delle tossicodipendenze in italia. *Presidenza del Consiglio dei Ministri, Roma* (2007). 13, 31

[48] Kahneman, D., Slovic, P. & Tversky, A. *Judgment under uncertainty: Heuristics and biases* (Cambridge University Press, 1982). 23

[49] Amaral, L. A. N., Buldyrev, S. V., Havlin, S., Salinger, M. A. & Stanley, H. E. Power law scaling for a system of interacting units with complex internal structure. *Physical Review Letters* **80**, 1385 (1998). 31

[50] Byrne, D. *Complexity theory and the social sciences: an introduction* (Routledge, 2002). 31

[51] Caves, R. E. Industrial organization and new findings on the turnover and mobility of firms. *Journal of economic literature* 1947–1982 (1998). 31

[52] Bak, P. *How nature works* (Oxford University Press Oxford, 1997). 31

[53] Kauffman, S. *At home in the universe: The search for the laws of self-organization and complexity* (Oxford University Press, 1995). 31

[54] Belmonte, A. & Pennisi, A. Education reforms and teachers needs: a longterm territorial analysis. *IJRS* **12**, 87–114 (2013). 31

[55] *Decreto del Presidente della Repubblica dell'8 marzo 1999 n. 275* (Gazzetta Ufficiale, 10 Agosto 1999). 31, 33

[56] Fu, D. *et al.* The growth of business firms: Theoretical framework and empirical evidence. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 18801–18806 (2005). 32, 36

[57] De Wit, G. Firm size distributions: An overview of steady-state distributions resulting from firm dynamics models. *International Journal of Industrial Organization* **23**, 423–450 (2005). 32

[58] Pitman, E. The estimation of the location and scale parameters of a continuous population of any given form. *Biometrika* **30**, 391–421 (1939). 35

[59] Stanley, M. H. *et al.* Zipf plots and the size distribution of firms. *Economics letters* **49**, 453–457 (1995). 35

[60] Sutton, J. Gibrat's legacy. *Journal of economic Literature* 40–59 (1997). 36

[61] Gibrat, R. *Les inégalités économiques* (Recueil Sirey, 1931). 36

[62] Pammolli, F. *et al.* A generalized preferential attachment model for business firms growth rates. *The European Physical Journal B* **57**, 127–130 (2007). 36, 37

[63] Buldyrev, S. V. *et al.* A generalized preferential attachment model for business firms growth rates. *The European Physical Journal B* **57**, 131–138 (2007). 36

[64] Growiec, J., Pammolli, F., Riccaboni, M. & Stanley, H. E. On the size distribution of business firms. *Economics Letters* **98**, 207–212 (2008). 36

[65] Ayebo, A. & Kozubowski, T. J. An asymmetric generalization of gaussian and laplace laws. *Journal of Probability and Statistical Science* **1**, 187–210 (2003). 37

[66] Buldyrev, S. V., Pammolli, F., Riccaboni, M. & Stanley, H. E. *The Rise and Fall of Business Firms* (To be published, 2014). 38

[67] Wang, J., Wen, S., Symmans, W. F., Pusztai, L. & Coombes, K. R. The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer informatics* **7**, 199 (2009). 39

[68] Clauset, A., Shalizi, C. R. & Newman, M. E. Power-law distributions in empirical data. *SIAM review* **51**, 661–703 (2009). 39

[69] Reed, W. J. The pareto, zipf and other power laws. *Economics Letters* **74**, 15–19 (2001). 39, 42

[70] Black, S. E. Do better schools matter? parental valuation of elementary education. *The Quarterly Journal of Economics* **114**, 577–599 (1999). 54

[71] Garlaschelli, D., Caldarelli, G. & Pietronero, L. Universal scaling relations in food webs. *Nature* **423**, 165–168 (2003). 63

[72] Haldane, A. G. & May, R. M. Systemic risk in banking ecosystems. *Nature* **469**, 351–355 (2011). 63

[73] Knoll, A. H. & Carroll, S. B. Early animal evolution: emerging views from comparative biology and geology. *Science* **284**, 2129–2137 (1999). 64

[74] Gomez-Mejia, L. R. Structure and process of diversification, compensation strategy, and firm performance. *Strategic management journal* **13**, 381–397 (1992). 64

[75] Ansoff, H. I. Strategies for diversification. *Harvard business review* **35**, 113–124 (1957). 64

[76] Bloomberg. Index methodology global fixed income family (2013). 64

[77] Cracraft, J. Biological diversification and its causes. *Annals of the Missouri Botanical Garden* 794–822 (1985). 64

[78] Walker, I. Biological memory. *Acta biotheoretica* **21**, 203–235 (1972). 64

[79] David, R. Principles of political economy and taxation. *publicado en* (1817). 66

[80] Chu, Y.-J. & Liu, T.-H. On shortest arborescence of a directed graph. *Scientia Sinica* **14**, 1396 (1965). 81

[81] Edmonds, J. Optimum branchings. *Journal of reserach of the national bureau of standards section B- Mathematical Science* **4**, 233 (1967). 81

[82] *https://github.com/mlbright/edmonds* (April, 24th. 2014). 81