UNIVERSITA' DEGLI STUDI DI MODENA E REGGIO EMILIA

M2CSC School of Graduate Studies in Multiscale Modelling, Computational Simulations and Characterization in Material and Life Sciences

A quantitative approach to the modelling of interacting systems from empirical data: the statistical mechanics perspective and a case study from social sciences

doctoral thesis

presentata da

Dr. Francesco De Pretis

Direttore	Relatore
Prof.ssa Ledi Menabue	Prof. Claudio Giberti

XXVI CICLO (2011 – 2013)

Abstract

In this work, we analyze complex systems characterized by embodying many particles with a specific focus on those ones arising in social sciences. The starting point of our analysis is a statistical mechanics approach: in particular, we are concerned with models where individuals can either interact or non-interact. Within such theoretical framework, one of the main aspects to be investigated is distinguishing whether the social phenomena that one may consider through empirical data are mainly due to imitative interaction among individuals rather than individual choices.

In specific, we base our analysis on the exploitation of two large datasets collected by the Italian Institute of Statistics (ISTAT) and the Emilia Romagna Region, the former one containing information recorded in all Italian municipalities regarding marriages and births occurred during an eleven years span, from 2001 to 2011; the latter one containing information recorded in all Emilia Romagna Region municipalities regarding marriages occurred during a sixteen years span, from 1995 to 2010.

Focusing on classical quantifiers of integration such as the fraction of marriages with spouses of mixed origin (native and immigrant) and the fraction of newborns with parents of mixed origin, several analyses performed over the datasets reveal that average measurements of these quantifiers exhibit diverse patterns. In particular, if data belonging to the ISTAT dataset are partitioned into two subsets independent from time, separating

events coming from large-sized cities (defined by embodying more than ten thousand inhabitants) by those ones occurring in small-sized cities, two different patterns clearly emerge: average measurements of data fit very well a linear and then a square root function. Such patterns visibly emerge even if we observe integration phenomena from the very first years: a predictive analysis – in this case considering time as a variable – has shown that such kind of patterns are valid even if we perform the same analyses, cumulating data year by year until we fall again into the whole datasets.

According to the statistical mechanics models we have initially considered, the theoretical interpretation we offer is that immigrants living in large-sized cities mainly make independent choices related to integration phenomena while the contrary holds true in small-sized cities, where imitative behaviors seem to play a more dominant role. The result emerged with Italian data unveils even more complex patterns of integration with respect to previous seminal work carried over similar data coming from Spain.

At last, this work shows that statistical mechanics models can be successfully applied in quantitatively describing social integration phenomena, catching subtle patterns already perceived by classical sociology authors more than a century ago.

Sintesi

In questo lavoro, analizziamo sistemi complessi caratterizzati dall'incorporare molte particelle con un interesse specifico verso quelli che si manifestano nelle scienze sociali. Il punto di partenza della nostra analisi è un approccio meccanico statistico: in particolare, concentriamo il nostro interesse verso modelli dove individui possano interagire o non interagire. All'interno di questa cornice teorica, uno dei principali aspetti da investigare è distinguere se i fenomeni sociali che si possono considerare tramite dati empirici sono principalmente dovuti ad interazione imitativa fra gli individui piuttosto che a causa di scelte individuali.

Nello specifico, basiamo la nostra analisi sullo studio di due grandi insiemi di dati raccolti dall'Istituto Italiano di Statistica (ISTAT) e dalla Regione Emilia Romagna, il primo contenente informazioni registrate in tutti i comuni italiani per ciò che riguarda i matrimoni e le nascite avvenute nell'arco di undici anni, dal 2001 al 2011; il secondo contenente informazioni registrate in tutti i comuni della Regione Emilia Romagna per ciò che riguarda i matrimoni celebratisi in un arco di sedici anni, dal 1995 al 2010.

Concentrandoci su quantificatori classici di integrazione, come la frazione di matrimoni con coniugi di origine mista (nativi e immigrati) e la frazione di neonati con genitori di origine mista, le analisi effettuate sugli insiemi di dati rivelano che le misure medie di questi quantificatori mostrano pattern diversi. In particolare, se si considera l'insieme di dati

dell'ISTAT e si fa una partizione dello stesso, suddividendolo in due sottoinsiemi indipendenti dal tempo, separando eventi provenienti dalla città di grandi dimensioni (definite dall'incorporare più di 10 000 abitanti) da quelli che si verificano in città di piccole dimensioni, due pattern differenti chiaramente emergono: le misure medie di dati fittano in modo evidente nel primo caso una funzione lineare e nel secondo caso una funzione a radice quadrata. Tali pattern emergono visibilmente anche se osserviamo i fenomeni di integrazione fin dai primi anni : un'analisi predittiva – in questa caso considerante il tempo come una variabile in gioco - ha dimostrato che questo tipo di pattern sono validi pure se si eseguono le stesse analisi, cumulando i dati anno per anno fino a quando non si ricada di nuovo nell'intero insieme di dati.

In accordo con i modelli meccanico statistici precedentemente menzionati, l'interpretazione teorica che offriamo è che gli immigrati che vivono in città di grandi dimensioni compiono principalmente scelte indipendenti legate a fenomeni di integrazione, mentre il contrario è vero in città di piccole dimensioni, dove i comportamenti imitativi sembrano svolgere un ruolo più dominante. Il risultato emerso con i dati italiani svela pattern ancora più complessi di integrazione rispetto a quanto emerso in precedente lavoro compiuto su dati simili provenienti dalla Spagna .

Questo lavoro dimostra infine che i modelli della meccanica statistica possono essere applicati con successo nel descrivere quantitativamente i fenomeni di integrazione sociale, cogliendo sottili pattern già percepiti da autori classici della sociologia più di un secolo fa.

Index

Chapter 1

Introduction

Le savant doit ordonner ; on fait la science avec des faits comme une maison avec des pierres ; mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison.

The scientist must set in order; science is built up with facts, as a house is with stones; but a collection of facts is no more a science than a heap of stones is a house.

Jules Henri Poincaré, La Science et l'Hypothèse (1902)

The work presented here is the result of a three year experience at M2CSC, the School of Graduate Studies in Multiscale Modelling, Computational Simulations and Characterization in Material and Life Sciences at the University of Modena and Reggio Emilia, Italy.

M2CSC is a Ph.D. school characterized by a very multidisciplinary approach, enrolling students coming from several and different fields (like mathematics, chemistry and engineering) and organizing didactical courses and seminars which reflect this cultural background, looking at science and its methods as a unifying bridge among diverse disciplines. Such framework has definitely influenced this work, which has been conducted both investigating theoretical problems and applying mathematical models to real-world phenomena.

At a first glance, this work has been centered on Statistical Mechanics models, applied either in describing ferromagnetism (**chapter 2**) or in modelling social phenomena (**chapter 3**). In the latter case, the Statistical Mechanics approach has been used to interpreter the output of several data analyses performed over a very large database released by ISTAT (the Italian National Statistical Institute), containing information on immigration phenomena occurred in Italy from 2001 to 2011 (**chapter 4**).

Before deepening into the details and the results which has come out from this work, we give a more profound explanation of the reasons which has led to discuss such problems, the motives which are beneath the methodologies applied and the yearned outcome of such work, thinking how this can hopefully contribute to the current scientific discussion.

The inability of giving credible forecasts over phenomena that we observe every day is something which has been perceived even from the most conservative establishment: as reported by the British newspaper The Daily Telegraph (Pierce, 2008), in occasion of the opening of the academic year at London School of Economics on November 2008, Her Majesty Elizabeth II abruptly asked the economists "why did nobody notice it ?", referring to the ongoing financial crisis.

The question whether none could really predict the so-called "credit-crunch" started in 2007 with the default of Lehman Brothers (at that time, the fourth-largest investment bank in the US) and – to mention another example – the reason why it has been so difficult to handle immigration flows in Europe (especially in Italy) since the beginning of the '90s (Willekens, 1994; Bijak & Wiśniowski, 2009) put in doubt the actual modelling of economic and social phenomena. Juxtaposing immigration and economic phenomena could sound unusual but, on the other hand, it turns out that they have more things in common than one might initially expect: such phenomena are processes characterized by occurring among a very large number of people, more resembling the particles of gas rather than a spare group of players; for this reason, their analogy with thermodynamic systems has been advocated not only by physicists (Georgescu-Roegen, 1971).

However, despite these properties, such phenomena have been for a long time modeled and studied by social scientists through a classical physics approach (Jevons, 1871), designing social processes as an idealized system of perfectly rational, optimizing institutions and individuals, who, by trading in markets, bring the economy and society to a balanced, efficient equilibrium; a "postcard-scenery" which has been completely blasted by the turmoil of the ongoing financial crisis.

The disappointment related to the low predictability of such models has found endorsers even in the most unexpected people, such as real economic agents, seen as "insiders" in the world of finance. This is the case – for example – of George Soros, who claims to have been for a long time aware of the low consistency of the models used to give explanation to social phenomena; in 2009 such beliefs pushed Soros to contribute with a 50 million US dollars pledge to the foundation of the Institute for New Economic Thinking (INET), a New York

City-based nonprofit think tank whose purpose is to support academic research and teaching in economics "outside the dominant paradigms of efficient markets and rational expectations" (Weber T., 2011).

The ideas beneath INET and the new vision of social science the institute embodies are well exemplified by the initial part of a speech Soros gave at the INET Conference at King's College on April 2010: "Economic theory has modeled itself on theoretical physics. It has sought to establish timelessly valid laws that govern economic behavior and can be used reversibly both to explain and to predict events. But instead of seeking laws capable of being falsified through testing, economics has increasingly turned itself into an axiomatic discipline consisting of assumptions and mathematical deductions – similar to Euclidean geometry. Rational expectations theory and the efficient market hypothesis are products of this approach. Unfortunately they proved to be unsound. To be useful, the axioms must resemble reality. Euclid's axioms meet that condition; rational expectations theory does not. It postulates that there is a correct view of the future to which the views of all the participants tend to converge. But the correct view is correct only if it is universally adopted by all the participants – an unlikely prospect. Indeed, if it is unrealistic to expect all participants to subscribe to the theory of rational expectations, it is irrational for any participant to adopt it. Anyhow, rational expectations theory was pretty conclusively falsified by the crash of 2008 which caught most participants and most regulators unawares. The crash of 2008 also falsified the Efficient Market Hypothesis because it was generated by internal developments within the financial markets, not by external shocks, as the hypothesis postulates." (Soros, 2010)

As Soros correctly outlines, economics and social sciences need new paradigms and new models, able to embrace the complex processes that take place into the real world between human beings. Such complexity cannot be captured any more by a classical physics approach: the point of view underlying this work is that this new way, this novel approach to the study of such phenomena necessarily starts with statistical mechanics, a field between mathematics and physics used for "statistical probabilistic predictions about systems which either contain elements which are too small to see or too numerous to keep track of; or usually both" (Susskind, 2013).

Which reasons could drive us to endorse such shift, this passage from the classical approach to the statistical one? In order to reply to such question, it is worth to recall in this place

which are the main differences between classical and statistical mechanics, observing the methodologies such fields implement to give reason to the phenomena under their own lens.

As summarized in (Susskind, 2013), classical mechanics mainly concerns with the concept of perfect predictability: the rough idea is that while witnessing the phenomena in a "closed system", (i.e. separated and not interacting with anything else outside such system), one can make predictions with a maximum precision or at a given level of precision (or predictability), just knowing two things, i.e. the initial conditions and the laws of evolution of the system – a vision that dates backs already the beginning of the nineteenth century, as popularized by Pier Simon Laplace (Laplace, 1814).

The point is that complete predictability can turn out to be totally useless. This assumption has not to be viewed as a criticism since the basic laws of classical mechanics have shown to be very powerful in their predictability: the point – again – is that in many cases such laws turns to be very useless to actually analyzing what it is really going on. For example, having a list of the positions and velocities of every particle in a certain room would not be very useful to us, since the list would be too long and would not give explicit account of other macroscopic properties of major interest, such temperature, pressure and so on. To shift such example to the set of economic and social phenomena, it could be not such worth (or useful) to cast predictions over the amount of each single bank account of a given population, to know the exact price with which each barrel of crude oil is exchanged world-wide at a given time or with whom and when each mixed marriage among people coming from two populations takes place. Surely instead, it would be of foremost interest to know with a good predictability the gross domestic product of a nation, the international price of crude-oil or the way mixed marriages and other integration phenomena occur in a given nation, in a given time. In other words, it could sound more interesting to know not what the single does but what comes out when many singles interact together.

This is exactly the case of statistical mechanics, a theory which provides tools that are applicable when the initial conditions of a system are not known with infinite precision (the so-called "starting point"), the laws of evolution of the system are neither completely known or when the system is not closed (for example, if it is interacting with other elements outside). Therefore, to say that in (Susskind, 2013) words: "when ideal predictability is not possible or it is not practicable way, you resort to probability and statistical mechanics".

However, the renounce to classical mechanics total predictability does not mean that statistical mechanics is not itself a sound field which can offer precise and viable answers. It is known that when the number of elements in a systems tend to be very large, also probability tend to be a very precise predictor (when – of course – the law of large numbers is applicable), so that statistical mechanics itself can be highly predictable but not for everything: even for the more complicated properties of a given system, like fluctuations and large deviations, statistical mechanics is always a feasible tool not to forecast predictions over the exact time when such unusual events might happen but to give precise predictions of the probability associated to such unusual events. From another point of view, statistical mechanics cannot predict information on every element of a given system but can afford precise and reliable information about collective properties arising from the interaction of a large number of elements, providing a framework for relating the microscopic properties of individual elements to the macroscopic bulk properties of a given system or population.

The importance of the statistical mechanics approach in investigating economic and social phenomena has been strengthen by Pierluigi Contucci and Jean-Philippe Bouchaud, whose words – expressed at the "Disorder in Probability and Statistical Mechanics" conference held in Modena on June 2012 – impressively match with those ones previously reported by Soros (De Pretis, 2012).

"The current crisis puts classical economics thinking under huge pressure. In theory, deregulated markets should be efficient, thanks to perfectly rational agents that correct ("arbitrage'') any mispricing or forecasting error. These equilibrated markets should be stable: crises can only be triggered by acute exogenous disturbances, but certainly not precipitated by the dynamics of the market itself." – stated Bouchaud at the 2012 Modena conference – "Recent data, instead, allow to cast doubts on several pillars of the classical dogma. In particular, market fluctuations seem to result from the endogenous dynamics of a complex system, that spontaneously exhibits jumps and shocks."

Therefore, assuming as scientific perspective that one belonging to the statistical mechanics, this work has started from one its models, that is the so-called Curie-Weiss model (**chapter 2**), which was first introduced at the beginning of the twentieth century by Pierre Weiss (Weiss, 1907) to give explanations of several experimental observations carried out in previous years by Pierre Curie (Curie, 1895).

The Curie-Weiss model can be pictured as a statistical mechanics model (Bovier, 2006; Bolthausen & Bovier, 2007) in which the focus is posed on the behavior of a larger number of magnets (whose inner magnetic spin can be tracked by a dichotomous variable assuming for example the value $\pm$ 1) interacting each other under the effect of a possible external magnetic field. In modelling such two different types of interaction (an internal interaction among the spins and an external interaction between the spins and an external magnetic field), two parameters have been introduced in the Hamiltonian function defining the Curie-Weiss model to express and eventually tune the interactions occurring in the system.

Besides the mathematical aspects, the Curie-Weiss model has risen interests into the scientific community for the multiple applications it can exhibits, notably in the field of opinion formation (Montanari, 2007), discrete choice theory and social interaction (Contucci, Gallo, & Menconi, 2008; Gallo, Barra & Contucci, 2009; Gallo, 2008), just to mention the closest applications to economic and social phenomena modelling.

According to such perspective, in (Burioni, Contucci, Fedele, Vernia & Vezzani, 2014) the Curie-Weiss model and one of its generalizations have been employed to provide forecasts for health screening campaigns, given an empirical dataset containing anonymized information regarding the participation of a large group of women to cancer screening tests. The study has concerned an inverse problem related to the Curie-Weiss model: authors wanted to determine the parameters of the Hamiltonian function, relying on the empirical data they gathered. In its strictest version, this inverse problem has been investigated always in **chapter 2**, where, after having restated it as combinatorial optimization problem over a very numerous set, we provide some numerical attempts for its solution, based on Computational Statistics methods.

Besides that, another approach in modelling social phenomena through Statistical Mechanics has been drawn in **chapter 3**: here, we present an interacting model (Barra, Contucci, Sandell, & Vernia, 2013) concerning this time integration phenomena occurring within a mixed population containing natives and immigrants. Authors of this study claim that such phenomena carry inside the "fingerprint of the mean-field imitation theory of statistical mechanics" and have experimentally proved this result, tackling a very large database containing information over marriages, births and work contracts which occurred in Spain from 1999 to 2010 in municipalities whose population exceeded ten thousand inhabitants.

Similarly, in **chapter 4** the work has first focused on the analysis of a smaller database containing information about marriages in all the municipalities of Emilia Romagna (an Italian region) from 1996 until 2010 – whose gathering from public accessible source on internet has required automation of web contents wrapping techniques (**appendix 1**) – and then it has considered an even larger database – if compared with the Spanish one – containing over one million information about marriages and births in Italy in a time period lasting from 2001 to 2011 with data coming from all the Italian municipalities, officially released by ISTAT.

As it will be shown in **chapter 4**, such data have undergone several analyses, leading to notable results that confirm in one way those ones already attained in (Barra, Contucci, Sandell, & Vernia, 2013) and in another way provide a different and even more sophisticated perspective over the sociological processes that occur in case of integration of immigrants in a foreign country. The analyses performed over the Italian data reveal that in "small-sized municipalities" (defined as municipalities whose population is under ten thousand inhabitants) imitation processes are the main drive of integration, whereas independent choices primarily occur when we look at people belonging to "large-sized municipalities" (defined as complementary to the "small-sized municipalities"). In other words, the statistical mechanics model underling such analysis seems to be able to give a rough but meaningful division between two very different patterns in integration processes: mixed marriages and births from mixed couples seem to occur in "small-sized municipalities" mainly because of imitation processes, whereas it is true the contrary – according to the data – if we observe the behavior of mixed couples in "large-sized municipalities". More explicitly, integration processes are not scale-invariant: the scale, instead, matters and defines in some way the behavior of immigrants and natives. A similar result was already perceived more than a century ago by sociologist Emile Durkheim, when he noticed different patterns in suicidal events occurring in "small villages" and "large industrial cities" (Durkheim, 1897).

As the reader will see, multiscale modelling and data-analysis have been widely used in accordance with the spirit of this Ph.D. school: furthermore, the hope is that the statistical mechanics approach presented here could find more and more recognition in describing and explaining economic and social phenomena.

Chapter 2

The Curie-Weiss model

As outlined in **chapter 1**, statistical mechanics will be our starting point to analyze social phenomena: the model we present here is the so-called Curie-Weiss model, which has recently risen an increasing interests into the scientific community for the multiple applications it can exhibits, notably in the field of opinion formation (Montanari, 2007), discrete choice theory and social interaction (Contucci, Gallo & Menconi, 2008; Gallo, Barra & Contucci, 2009; Gallo, 2008) and eventually integration phenomena (Barra, Contucci, Sandell, & Vernia, 2013), just to mention the closest applications to economic and social phenomena modelling.

However, the Curie-Weiss model did not originate from social sciences but arose from physical experiments conducted in the nineteenth century, mainly concerning the status of magnetization of a series of materials, according to their internal temperature. As the name of the model would state, the Curie-Weiss model could be thought as a mutual work between two French scientists, Pierre Curie and Pierre-Ernest Weiss, even though – as it will be shown later – German physicist Ernst Ising played a very important role in defining the ordinary statistical mechanics framework which today defines the model.

Back to 1895, Nobel prize winner Pierre Curie conducted a series of experiments at the École municipale de physique et de chimie industrielles in Paris (Curie, 1895) aimed to

investigate the magnetization of a set of materials such as iron, nickel and magnetite, showing that such property was dependent to the internal temperature of such materials.

As already witnessed by Michael Faraday (Curie, 1895), iron and nickel – today known as ferromagnetic materials – are characterized by the experimental fact of retaining magnetization and becoming permanent magnets, after having exposed to a certain magnetic field. Nonetheless, such property – as it was proved by Curie – is abruptly lost if the internal temperature of such material reaches the so-called Curie temperature (also known as Curie point), a breaking threshold beyond which permanent magnetization changes to induced magnetization: in other words, trespassing the Curie temperature makes such materials lost its property of spontaneous magnetism. Curie also noticed that each material had a specific temperature.

In 1907, Pierre-Ernest Weiss produced a model which could give reasons to the experimental measurements carried on by Curie almost fifteen years earlier: Weiss was the first to theorize that atoms in ferromagnetic materials experience a molecular field, a sum of all the fields produced by all other internal particles in the material, aside a possible external field. This assumption led him to write a balance equation where such kind of internal interaction is explicitly taken into account: the theoretical data then derived from such equation looked to be in perfect accordance with those ones collected by Curie and by the experiments carried by Weiss himself at his laboratory at the Eidgenössische Technische Hochschule (ETH) in Zurich (Weiss, 1907). The model later became known as Curie-Weiss model.

However, the current statistical mechanics framework that today is used for Curie-Weiss model was introduced only more than fifteen years later by Ernst Ising.
In his Ph.D. thesis defended at the University of Hamburg (Germany) in 1924 (Ising, 1925), Ising tackled the special case of a linear chain of magnetic moments, which are only able to take two positions, "up" and "down," and which are coupled by interactions between nearest neighbors. The Curie Weiss model can be seen as an Ising model where the nearest neighbors interaction is then replaced by another extreme choice, that is each magnetic moment interacts with each other at any site with the same strength. Nevertheless, differently from Weiss, Ising was the first to introduce the Hamiltonian notation to define the interactions which take place within and externally to the system, even though the ideas of interaction within spins and with another external field were inspired by Weiss' previous work.

Surprisingly, the one-dimensional model defined by Ising fails to provide any spontaneous magnetization for positive temperatures (Huang, 1987), while such physical phenomenon is suitably replicated by the Curie-Weiss model (Ellis, 1985): this first-order phase transition is instead correctly simulated in the two-dimensional Ising model, whose complex mathematical treatment made it solvable only later in 1944 by Lars Osanger (Onsager, 1944). Furthermore, the two-dimensional Ising model correctly predicts some other physically measurable quantities for ferromagnetic materials, known as critical exponents, because its nearest neighbors interaction looks like to be more similar to the expected short range interaction that occurs in quantum mechanics phenomena, whereas the Curie-Weiss model "total interaction" of the internal spins produces incorrect results for such exponents (Huang, 1987; Baxter, 1982).

Although the one and two-dimensional Ising models are exactly solvable models, up to now at higher dimensions analytical solutions have not been found yet. For this reason, a very important tool to explore the properties of such models is the mean field theory (known also as self-consistent field theory), an approach that analyzes the behavior of large and complex stochastic models by studying a simpler model. According to this perspective, already the Curie-Weiss model can be seen as a mean-field version of the Ising model (Bovier, 2006), as we have seen that in this case the nearest neighbors interaction is replaced by a more general interaction among spins. Moreover, both the Curie-Weiss model and the Ising model can be treated through the so-called mean field approximation (Nishimori, 2001; Evans, 2006), a mathematical artifice that reduces the internal interaction of the spins (in any way this interaction has been defined) to an external mean field, eventually merged to an already present external field. Such approximation makes the models more easily treatable from a mathematical point of view and paves the way to the proof of the so-called consistency equation which – as it will be shown later – is embodied in both models. The first scientists who attempted in some way to introduce such artifice were British physicists William Lawrence Bragg and Edward John Williams who introduced the so-called Bragg-Williams approximation in 1934 (Bragg & Williams, 1934).

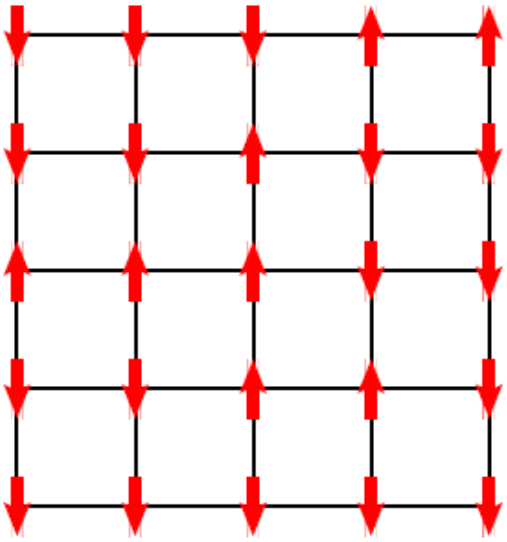
In the following paragraphs, we present the Curie-Weiss model and the one-dimensional Ising model. Moreover, we derive through a mean field approach the so-called consistency equation, which arises in both models. Eventually, we outline an inverse problem related to the Curie-Weiss model and we provide the reader with some theoretical and computational attempts which have been pursued for its investigation. It is worth to notice that in our

treatment - differently from other authors - the inverse temperature (usually indicated with β) will be included in the parameters of the model and not explicitly specified.

More extensive descriptions of the Curie-Weiss model and its relation to the Ising model can be found in (Thompson, 1979; Baxter, 1982; Ellis, 1985; Chandler, 1987; Huang, 1987; Parisi, 1988; Simon, 1993; Nishimori, 2001; Bovier, 2006; Bolthausen & Bovier, 2007; Mézard & Montanari, 2009; Talagrand, 2010)

2.1 Mathematical definition of the model

We consider a subset $\Omega_N \subset \mathbb{Z}^d$ of a d-dimensional lattice, containing N sites. In each i-site (with $i = 1 \dots N$), we assign a variable – named spin σ_i – which can assume the value + 1 or – 1: we call Λ_N the whole *configuration space*, defined as $\Lambda_N = \{+1, -1\}^N$. A *spin configuration* $\boldsymbol{\sigma}$ is an element of Λ_N and represents one possible set of N spins states (physically speaking, this is what is observable in a given time), so that $\boldsymbol{\sigma} = (\sigma_1 \dots \sigma_N)$. Moreover, a pair of sites (ij) takes the name of *bond*, whereas a given *set of bonds* will be denoted by B; as it will be shown later, the choice made over the set of bonds will be crucial to define the mechanical statistics model under our lens. Anyway, given a whatever set of bonds, we assign an *interaction energy* $-J\sigma_i\sigma_j$ to each bond in the set B, where J is a parameter known as *coupling constant*. This interaction energy is then $-J$ if the states of the two bonded spins are the same (i.e. if $\sigma_i = \sigma_j$) and J otherwise: in case of same sign spins, the interaction energy is therefore lower if $J > 0$. Aside from mathematical considerations, physical evidence regarding magnetism phenomena confirms that same sign spins bonds are more stable, requiring less energy; in case of interaction, spins thus tend to be oriented in the same direction. This positive interaction can lead to macroscopic magnetism (ferromagnetism) and into this framework J is called ferromagnetic interaction constant: if taken negative, it is on the contrary called anti-ferromagnetic interaction constant.



Aside from spins interaction, each i-site can be thought to have already have its own energy of the type $-h\sigma_i$ where h is an *external field* (physically, the so-called Zeeman energy in magnetism). Again, if σ_i is aligned to the field h, it gives a lower energy contribution.

Example of subset $\Omega_{25} \subset \mathbb{Z}^2$, taken from a 2-dimensional lattice.
Pointing up and down arrows represent spins whose value can be +1 or -1

Putting together the energy due to bond interactions and the energy due to an external field, the total energy of the system can be therefore expressed through an *Hamiltonian function*, depending from a certain spin configuration $\boldsymbol{\sigma}$:

$$H_N(\boldsymbol{\sigma}) = -J \sum_{(ij)\in B} \sigma_i \sigma_j - h \sum_{i=1}^{N} \sigma_i \tag{2.01}$$

The latter function is also known as *configurational energy*. In the following passages, we assume the parameter J to be positive. The particular case $J = 0$, which does not consider interaction, is known as *ideal lattice gas of non-interacting particles*.

The physical evidence invoked above for energy minimization has also fostered a usual statistical mechanics approach where we assume as probability distribution for a certain spin configuration $\boldsymbol{\sigma}$ – given an Hamiltonian $H_N(\boldsymbol{\sigma})$ expressing the energy of a system – a particular type of probability distribution, known as *Boltzmann-Gibbs distribution*:

$$P_N(\boldsymbol{\sigma}) = \frac{e^{-H_N(\boldsymbol{\sigma})}}{Z_N} \tag{2.02}$$

where Z_N is a normalization factor known as *partition function* and defined as:

$$Z_N = \sum_{\boldsymbol{\sigma}\in\Lambda_N} e^{-H_N(\boldsymbol{\sigma})} \tag{2.03}$$

If we now consider a subset $\Lambda_N \subset \mathbb{Z}$ of a one-dimensional lattice and we define B as a set of bonds containing only pairs of nearest neighbors sites, we can rewrite the Hamiltonian function as:

$$H_N^{(NN)}(\boldsymbol{\sigma}) = -J \sum_{i=1}^{N} \sigma_i \sigma_{i+1} - h \sum_{i=1}^{N} \sigma_i \tag{2.04}$$

The Hamiltonian $H_N^{(NN)}(\boldsymbol{\sigma})$ defines the *one-dimensional Ising model* (Ising, 1925) and – as stated above – explicitly embodies a nearest neighbors interaction: we can note this by observing the first summation term containing at each time only σ_i and σ_{i+1} . However, spins at sites $i = 1$ and $i = N$ lack of neighbors: to avoid such issue, the simplest way is by considering our model on a circle, that is imposing $\sigma_{N+1} = \sigma_1$ – a periodic boundary condition (Bovier, 2006).

A mean-field *version* (Bovier, 2006) of the one-dimensional Ising model is obtained modifying the set B and opening it to all possible pairs of sites. Doing that, the nearest neighbors interaction is replaced by a total interaction among spins and the Hamiltonian defining such systems now becomes:

$$H_N^{(CW)}(\boldsymbol{\sigma}) = -\frac{J}{N}\sum_{i,j=1}^{N}\sigma_i\sigma_j - h\sum_{i=1}^{N}\sigma_i \tag{2.05}$$

The Hamiltonian $H_N^{(CW)}(\boldsymbol{\sigma})$ defines this time the *Curie-Weiss model* and the choice of using a parameter J scaled by N is done to avoid the configurational energy taking values higher then $O(N)$, for extensiveness bounds (Bovier, 2006).

2.2 Empirical magnetization and site magnetization

A macroscopic variable associated to both one-dimensional Ising and Curie-Weiss model is $m_N(\boldsymbol{\sigma})$, the *empirical magnetization* of a given spin configuration $\boldsymbol{\sigma}$, defined as:

$$m_N(\boldsymbol{\sigma}) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i \tag{2.06}$$

With the term *site magnetization*, expressed as $\langle\sigma_i\rangle$, it is instead indicated a particular type of average, named *thermal average*, which is computed for a generic spin σ_i through the Boltzmann-Gibbs distribution introduced in **equation (2.02)**, as:

$$\langle\sigma_i\rangle = \sum_{\boldsymbol{\sigma}\in\Lambda_N}\sigma_i P_N(\boldsymbol{\sigma}) \tag{2.07}$$

In the next paragraphs, we will always refer to average operators based on such kind of probability distribution only through angle brackets $\langle * \rangle$.

Since in a given spin configuration $\boldsymbol{\sigma}$ all sites are equivalent and the system is translationally invariant (being it previously defined on a circle, for periodic boundary conditions), we have $\langle\sigma_1\rangle = \langle\sigma_2\rangle = \cdots = \langle\sigma_N\rangle$. This invariance can be exploited to prove that the site magnetization is equivalent to the thermal average of the empirical magnetization.

In fact we can state that:

$$\langle \sigma_i \rangle = \frac{1}{N} \sum_{k=1}^{N} \langle \sigma_k \rangle \tag{2.08}$$

The second term of the **equation (2.08)** can be thus rewritten as:

$$\frac{1}{N} \sum_{\boldsymbol{\sigma} \in \Lambda_N} \left(\sum_{k=1}^{N} \sigma_k \right) P_N(\boldsymbol{\sigma}) \tag{2.09}$$

which leads to the following expression:

$$\sum_{\boldsymbol{\sigma} \in \Lambda_N} m_N(\boldsymbol{\sigma})\, P_N(\boldsymbol{\sigma}) = \langle m_N(\boldsymbol{\sigma}) \rangle \tag{2.10}$$

Combining **equation (2.08)**, **(2.09)** and **(2.10)**, the relation

$$\langle m_N(\boldsymbol{\sigma}) \rangle = \langle \sigma_i \rangle \tag{2.11}$$

trivially holds for each $i = 1, \dots, N$.

2.3 The mean-field approximation and the consistency equation

As mentioned above, a way to attack the Curie-Weiss model is the so-called mean-field approximation: we adapt here passages given in (Nishimori, 2001; Evans, 2006; Contucci & Giardinà, 2012) to prove the consistency equation for the Curie-Weiss model.

Instead of looking at the entire Hamiltonian function, we now consider all the contributions given only by a certain spin σ_j

$$H_j^{(CW)}(\boldsymbol{\sigma}) = -\frac{J}{N} \sigma_j \sum_{k=1}^{N} \sigma_k - h\sigma_j \tag{2.12}$$

The mean-field approximation applied to this contribution works by considering the empirical magnetization – introduced above in the **equation (2.06)** – and renaming it equal to a constant M (an average field which is caused by the other spins σ_k's), so that:

$$\frac{1}{N}\sum_{k=1}^{N}\sigma_k = M \tag{2.13}$$

Substituting M into **equation (2.12)**, we have a new Hamiltonian function $H_j^{(MF)}(\sigma_j)$, this time depending only on σ_j:

$$H_j^{(MF)}(\sigma_j) = -J\sigma_j M - h\sigma_j = -\sigma_j(JM + h) \tag{2.14}$$

The terms into brackets can be shortened in this form:

$$(JM + h) = h^{(MF)} \tag{2.15}$$

as if the spin σ_j would be under the effects of only one external "mean" field $h^{(MF)}$, so that the **equation (2.14)** becomes:

$$H_j^{(MF)}(\sigma_j) = -h^{(MF)}\sigma_j \tag{2.16}$$

If we look now at the entire Hamiltonian function, we therefore have:

$$H_N^{(MF)}(\boldsymbol{\sigma}) = \sum_{j=1}^{N} H_j^{(MF)}(\sigma_j) \tag{2.17}$$

Furthermore, the quantity M (**equation (2.13)**) has to satisfy the following auto-consistency condition:

$$M = \langle\frac{1}{N}\sum_{k=1}^{N}\sigma_k\rangle_{MF} = \langle m_N(\boldsymbol{\sigma})\rangle_{MF} \tag{2.18}$$

Here, the thermal average has been computed according to mean-field Hamiltonian.

At this point, we start working on the Boltzmann-Gibbs distribution as defined in the **equation (2.02)**. For the previous Hamiltonian function $H_N^{(MF)}(\boldsymbol{\sigma})$, it holds:

$$P_N^{(MF)}(\boldsymbol{\sigma}) = \frac{e^{-H_N^{(MF)}(\boldsymbol{\sigma})}}{\sum_{\boldsymbol{\sigma}\in\Lambda_N} e^{-H_N^{(MF)}(\boldsymbol{\sigma})}} \tag{2.19}$$

This Boltzmann-Gibbs distribution can be equivalently written explicitly exploiting the **equation (2.16)** and then factorizing the summations terms:

$$P_N^{(MF)}(\boldsymbol{\sigma}) = \frac{e^{-\sum_{i=1}^{N} H_i^{(MF)}(\sigma_i)}}{\sum_{\boldsymbol{\sigma}\in\Lambda_N} e^{-\sum_{j=1}^{N} H_j^{(MF)}(\sigma_j)}} = \frac{\prod_{i=1}^{N} e^{-H_i^{(MF)}(\sigma_i)}}{\sum_{\boldsymbol{\sigma}\in\Lambda_N} \prod_{j=1}^{N} e^{-H_j^{(MF)}(\sigma_j)}} \tag{2.20}$$

If we operate over the denominator of $P_N^{(MF)}(\boldsymbol{\sigma})$, exchanging the summation with the product, this leads to:

$$P_N^{(MF)}(\boldsymbol{\sigma}) = \frac{\prod_{i=1}^{N} e^{-H_i^{(MF)}(\sigma_i)}}{\prod_{i=1}^{N} \sum_{\sigma_j=\pm 1} e^{-H_i^{(MF)}(\sigma_i)}} = \prod_{i=1}^{N} \check{P}_i^{(MF)}(\sigma_i) \tag{2.21}$$

where

$$\check{P}_i^{(MF)}(\sigma_i) = \frac{e^{-H_i^{(MF)}(\sigma_i)}}{\sum_{\sigma_i=\pm 1} e^{-H_i^{(MF)}(\sigma_i)}} = \frac{e^{h^{(MF)}\sigma_i}}{e^{h^{(MF)}\sigma_i} + e^{-h^{(MF)}\sigma_i}} \tag{2.22}$$

meaning that the Boltzmann-Gibbs probability distribution $P_N^{(MF)}(\boldsymbol{\sigma})$ factorize. We will exploit such results straightaway, applying it to the site magnetization $\langle\sigma_j\rangle_{MF}$.

In fact, if in this mean field model we focus again only on a spin σ_j, its site magnetization $\langle\sigma_j\rangle_{MF}$ will defined according to **equation (2.07)** and **(2.21)**:

$$\langle\sigma_j\rangle_{MF} = \sum_{\boldsymbol{\sigma}\in\Lambda_N} \sigma_j P_N^{(MF)}(\boldsymbol{\sigma}) = \sum_{\sigma_1=\pm 1 \,\dots\, \sigma_N=\pm 1} \sigma_j \prod_{i=1}^{N} \check{P}_i^{(MF)}(\sigma_i) \tag{2.23}$$

In **equation (2.23)**, if we highlight the term $\check{P}_j^{(MF)}(\sigma_j)$, we can rewrite the summation in the following manner:

$$\langle\sigma_j\rangle_{MF} = \sum_{\sigma_1=\pm 1 \,\dots\, \sigma_N=\pm 1} \left[\prod_{\substack{i=1 \\ i\neq j}}^{N} \check{P}_i^{(MF)}(\sigma_i)\right] \left[\sigma_j \check{P}_j^{(MF)}(\sigma_j)\right] \tag{2.24}$$

Now, splitting **equation (2.24)** into two factors and exchanging the summation with the product, we obtain:

$$\langle\sigma_j\rangle_{MF} = \left[\prod_{\substack{i=1 \\ i\neq j}}^{N} \sum_{\sigma_i=\pm 1} \check{P}_i^{(MF)}(\sigma_i)\right] \left[\sum_{\sigma_j=\pm 1} \sigma_j \check{P}_j^{(MF)}(\sigma_j)\right] \tag{2.25}$$

Considering that the first factor of **equation (2.25)** is equal to one (in fact, $\sum_{\sigma_i=\pm1} \check{P}_i^{(MF)}(\sigma_i) = 1, \forall i = 1 \dots N$), the site magnetization therefore becomes:

$$\langle \sigma_j \rangle_{MF} = \sum_{\sigma_j=\pm1} \sigma_j \check{P}_j^{(MF)}(\sigma_j) \tag{2.26}$$

Combining **equation (2.22)** and **(2.26)**, we have:

$$\langle \sigma_j \rangle_{MF} = \sum_{\sigma_j=\pm1} \sigma_j \frac{e^{h^{(MF)}\sigma_j}}{e^{h^{(MF)}\sigma_j} + e^{-h^{(MF)}\sigma_j}} = \frac{e^{h^{(MF)}} - e^{-h^{(MF)}}}{e^{h_N^{(MF)}} + e^{-h^{(MF)}}} \tag{2.27}$$

The latter expression of exponential functions can be shortened as hyperbolic tangent function, so that **equation (2.27)** becomes:

$$\langle \sigma_j \rangle_{MF} = tanh\left(h^{(MF)}\right) \tag{2.28}$$

Recalling **equations (2.15)**, **(2.18)** and the relation given in **equation (2.11)**, we eventually obtain:

$$\langle m_N(\boldsymbol{\sigma}) \rangle_{MF} = tanh(J\langle m_N(\boldsymbol{\sigma}) \rangle_{MF} + h) \tag{2.29}$$

We now consider **equation (2.28)** in the *thermodynamic* limit, that is for $N \to \infty$, applying a limit to both its members:

$$\lim_{N\to\infty} \langle m_N(\boldsymbol{\sigma}) \rangle_{MF} = \lim_{N\to\infty} tanh(J\langle m_N(\boldsymbol{\sigma}) \rangle_{CW} + h) \tag{2.30}$$

This *thermodynamic* limit exists (Bovier, 2006; Gallo, 2008) and leads to the result of this paragraph, the so-called *consistency equation* for the mean field model:

$$m = tanh(Jm + h) \tag{2.31}$$

where m takes name of *average magnetization*[1].

[1] The *average magnetization* takes also name of *spontaneous magnetization* with a vanishing external field (i.e. for $h \to 0$) (Mézard & Montanari, 2009).

As final consideration we draw over the **equation (2.31)**, we stress that, in analogy to what it has been outlined above, it is possible to prove the same consistency equation also for the Curie-Weiss model.

2.4 An inverse problem arising in social phenomena modelling

As outlined at the beginning of this chapter, the Curie-Weiss model has risen interest outside its strict employment in describing physical phenomena, finding applications also in the framework of economic and social phenomena modelling: in such context, it may be used to model a unique population of individuals which interact each other in an uniform way. According to such perspective, in (Burioni, Contucci, Fedele, Vernia & Vezzani, 2014) the Curie-Weiss model has been employed to provide forecasts for health screening campaigns, given an empirical dataset containing anonymized information regarding the participation willingness of N women ($N \sim 10^4$) to cancer screening tests organized by the Local Health Authority of the municipality of Parma, Italy. The study results to be of particular interest in order to introduce an inverse problem related to the Curie-Weiss model: for this reason, we now offer a brief account of the some features beneath (Burioni, Contucci, Fedele, Vernia & Vezzani, 2014).

As internationally recognized by the medical community, screening activities are a very important tool to prevent the spread of cancerous diseases, since the high rate of success in curing early-stage patients and the lower social costs connected in avoiding treatments of diagnosed patients. For these two reasons, screening activities have to be encouraged in many ways, making the public opinion aware of the importance of prevention in the health care field. Thus, how to foster such activities? Which is the best way to spread screening prevention in a population of women? Does it turn out to be more important to have a testimonial, maybe a celebrity who could publicly endorse a campaign aimed to support screening activities? Or in reality does it occur that most of the women who decide to take part to screening activities are influenced by other women closer to them?

As it can be seen, it is easy to find an analogy between the behavior of a group of women in front of the decision of taking part to screening activities and that one of magnets whose magnetic spins can be ether influenced by the interactions with other magnets or be mainly oriented by an external field. Translated in statistical mechanics terms, if the dynamics occurring within this group of women was pictured through a Curie-Weiss model, it would

be interesting to investigate which mechanism is prevailing, i.e. if it is higher the role of J (internal interaction) compared with that one of h (external interaction).

In the bid of replying to such question, it is possible to project the empirical dataset we mentioned above as an ensemble of spins, whose signs ± 1 depend on the positive or negative reply to the participation of the screening activities: basing our analysis on this framework, we observe a unique configuration of spins coming from a Curie-Weiss model and the *inverse problem* to attack becomes then extracting from such configuration information about J and h.

The following paragraphs will describe the theoretical and computational attempts moved in solving such *inverse problem*, starting from the consideration of a more general case when one may observe several configurations of spins: however, in this context the experimental bound of a unique configuration of spins remains well-posed since the high difficulty in collecting such kind of data.

2.5 The multi configurations problem

In (Fedele, Vernia & Contucci, 2013), authors treat the problem of reconstructing the values of the parameters J (coupling constant) and h (external field) embodied into a Curie-Weiss Hamiltonian – introduced in **equation (2.05)** – given a certain number of samples made of M independent spins configurations $\{\boldsymbol{\sigma}^{(1)}, \dots, \boldsymbol{\sigma}^{(M)}\}$ all distributed according to the Boltzmann-Gibbs measure (**equation (2.02)**), where a generic spin configuration $\boldsymbol{\sigma}^{(i)} = (\sigma_1^{(i)} \dots \sigma_N^{(i)})$, $i = 1, \dots, M$ is an element of $\Lambda_N = \{+1, -1\}^N$.

In order to tackle the problem, authors' starting point is the *average magnetization* as defined by:

$$\lim_{N\to\infty} \langle m_N(\boldsymbol{\sigma}) \rangle_{CW} = m \tag{2.32}$$

We recall here that $m_N(\boldsymbol{\sigma})$ has been introduced in **equation (2.06)**. Moreover, the average $\langle * \rangle$ (**equation (2.07)**) has been computed according to a Curie-Weiss Hamiltonian. For the average magnetization, the consistency equation holds, as already presented in **equation (2.30)**.

If we derive equation (2.32), assuming that the passage of the limit under the derivative sign is valid, we have:

$$\lim_{N\to\infty} \frac{\partial}{\partial h} \langle m_N(\boldsymbol{\sigma}) \rangle_{CW} = \chi \tag{2.33}$$

where the quantity χ (known as *magnetic susceptibility)* is obtained deriving **equation (2.30)**:

$$\chi = \frac{\partial m}{\partial h} = \frac{1 - m^2}{1 - J(1 - m^2)} \tag{2.34}$$

and

$$\frac{\partial}{\partial h} \langle m_N(\boldsymbol{\sigma}) \rangle_{CW} = N(\langle m_N^2(\boldsymbol{\sigma}) \rangle_{CW} - \langle m_N(\boldsymbol{\sigma}) \rangle_{CW}^2) \tag{2.35}$$

Combining **equations (2.32)**, **(2.33)**, **(2.34)** and **(2.35)**, we can write the following relation for J:

$$J = \frac{1}{1 - \lim_{N\to\infty} \langle m_N(\boldsymbol{\sigma}) \rangle_{CW}^2} - \frac{1}{\lim_{N\to\infty} N(\langle m_N^2(\boldsymbol{\sigma}) \rangle_{CW} - \langle m_N(\boldsymbol{\sigma}) \rangle_{CW}^2)} \tag{2.36}$$

whereas the external field h is obtained by inverting the consistency equation:

$$h = \tanh^{-1}\left(\lim_{N\to\infty} \langle m_N(\boldsymbol{\sigma}) \rangle_{CW}\right) - J \lim_{N\to\infty} \langle m_N(\boldsymbol{\sigma}) \rangle_{CW} \tag{2.37}$$

Given a sample made of M independent spins configurations $\{\boldsymbol{\sigma}^{(1)}, \dots, \boldsymbol{\sigma}^{(M)}\}$, authors derive two estimators for average magnetization and magnetic susceptibility at finite size, following a maximum likelihood approach. They find:

$$\widehat{m}_{(M)} = \frac{1}{M} \sum_{i=1}^{M} m_N(\boldsymbol{\sigma}^{(i)}) \tag{2.38}$$

and

$$\hat{\chi}_{(M)} = N\left(\frac{1}{M} \sum_{i=1}^{M} m_N^2(\boldsymbol{\sigma}^{(i)}) - \widehat{m}_{(M)}^2\right) \tag{2.39}$$

Combing **equations (2.36)**, **(2.37)**, **(2.38)** and **(2.39)**, we eventually write the estimators for the parameters J and h, that is:

$$\hat{J} = \frac{1}{1 - \hat{m}_{(M)}^2} - \frac{1}{\hat{\chi}_{(M)}} \tag{2.40}$$

and

$$\hat{h} = \tanh^{-1}\left(\hat{m}_{(M)}\right) - \hat{J}\hat{m}_{(M)} \tag{2.41}$$

As final remark, (Fedele, Vernia & Contucci, 2013) shows that numerical tests considering 20 samples, each one containing a number $M = 20\,000$ of spins configurations (with $N = 10\,000$), lead to the calculation of a set of estimators $\left(\hat{J}^{(s)}, \hat{h}^{(s)}\right)\ s = 1 \dots 20$, whose arithmetical average show minimal difference with the parameters J and h.

2.6 The mono configuration problem

The problem we want to attack is similar to that one outlined in **paragraph 2.5**, though here we have a stricter constraint, due to the experimental reasons outlined in **paragraph 2.4**. In fact, we are bounded on the analysis of a unique sample (that is $M = 1$), only containing unique spin configuration $\boldsymbol{\sigma} = (\sigma_1 \dots \sigma_N)$ with $N = 10\,000$, a spin cardinality in agreement with (Burioni, Contucci, Fedele, Vernia & Vezzani, 2014).

The necessity of artificially recreating a higher cardinality of spins configurations in order to exploit the inference methods shown in (Fedele, Vernia & Contucci, 2013), has led to the exploration of bootstrap or jackknife techniques (Efron, 1979), which are usually implemented in suck kind of context: the basic idea is to generate new spins configurations, randomly sampling with replacement the original spin configuration $\boldsymbol{\sigma}$. However, numerical attempts based on such techniques have not led to any concrete result.

Another path which has been explored is a subsampling of the unique spin configuration $\boldsymbol{\sigma}$: in analogy to the structure of a sample made of M spins configurations $\left\{\boldsymbol{\sigma}^{(1)}, \dots, \boldsymbol{\sigma}^{(M)}\right\}$, each one containing N spins, we imagine here to handle a sample made of R subsamples $\left\{\tilde{\sigma}^{(1)}, \dots, \tilde{\sigma}^{(R)}\right\}$ of the unique spin configuration $\boldsymbol{\sigma}$, where a generic subsample $\tilde{\sigma}^{(t)} = \left(\sigma_1^{(t)}, \dots, \sigma_L^{(t)}\right)\ t = 1, \dots, R$ contains L spins. Moreover, we impose that the relation $N = R \cdot L$

holds and that $\{\tilde{\sigma}^{(1)}, \dots, \tilde{\sigma}^{(R)}\}$ is constructed to be a partition of the unique spin configuration.

Likewise above, we can write two estimators for average magnetization and magnetic susceptibility at finite size, based on the sample $\{\tilde{\sigma}^{(1)}, \dots, \tilde{\sigma}^{(R)}\}$:

$$\check{m}_{(R)} = \frac{1}{R}\sum_{i=1}^{R} m_L\left(\tilde{\sigma}^{(i)}\right) \tag{2.42}$$

and

$$\check{\chi}_{(R)} = N\left(\frac{1}{R}\sum_{i=1}^{R} m_L^2\left(\tilde{\sigma}^{(i)}\right) - \check{m}_{(R)}^2\right) \tag{2.43}$$

Similarly, **equations (2.42)** and **(2.43)** lead to an estimator for the parameter J:

$$\check{J} = \frac{1}{1-\check{m}_{(R)}^2} - \frac{1}{\check{\chi}_{(R)}} \tag{2.44}$$

whereas a similar **equation** to **(2.41)** can be drawn for $\check{h}$.

However, also this approach has proved to be unfruitful: inferring the parameters J and h over a certain number of random samples made of R subsamples of the unique spin configuration $\boldsymbol{\sigma}$ has not led to any reliable result, even exploring different values of R in agreement with the previous bound we set.

The difficulties arisen with the previous attempts – despite the manifest analogy with the methods implemented in (Fedele, Vernia & Contucci, 2013) – have directed to the definition of an optimal procedures of subsampling.

More explicitly, an optimal subsampling procedure can be stated as a minimization of the absolute error committed in predicting J through $\check{J}$ over a set called Φ_N, holding all the possible partitions of the spin configuration $\boldsymbol{\sigma} = (\sigma_1 \dots \sigma_N)$, given the above shown constraint $N = R \cdot L$:

$$\min_{\Phi_N} \left|J - \check{J}\right| \tag{2.45}$$

The ultimate goal of such approach is to study the optimal solutions one may find for the **equation (2.45)**, that is those samples which are partitioned in a way that enables a correct reconstruction of the parameters J and h.

2.7 The Harmony Search algorithm

Since the high[2] cardinality of the set Φ_N, the implementation of computational statistics methods seemed the natural framework to attack a problem definitely based on combinatorial optimization. In specific, a particular kind of meta-heuristic model – Harmony Search algorithm (HSA) – has been applied to explore optimal solutions to the stated problem. This decision has been taken for its versatility of approach in exploring very different problems requiring optimization over extremely numerous sets (Geem, Kim & Loganathan, 2001; Geem, 2009; Geem, 2010).

Now, we give a description of the algorithmic design implemented to solve the problem stated in **equation (2.45)**. HSA starts with the creation of the *Harmonic Matrix*, a $\mathsf{S} \times (R+1)$ matrix containing S elements of Φ_N, that is S samples made of R subsamples of the unique spin configuration $\boldsymbol{\sigma} = (\sigma_1 \dots \sigma_N)$. In the $(R+1)$th column the *harmony values* are stored, i.e. the images of the S elements of Φ_N through a proper fitness function, which has to be defined *ab initio*[3]. Worth to note, both the S elements of Φ_N and their related R subsamples are created through a random process simulating uniform sampling from the unique spin configuration $\boldsymbol{\sigma}$.

After its creation, the Harmonic Matrix is sorted according to the column of the harmony values; then a new sample made of R subsamples of the unique spin configuration $\boldsymbol{\sigma} = (\sigma_1 \dots \sigma_N)$ is created in two possible ways which are described below.

With a probability equal to p_{HS}, this new sample is randomly created in the same way as the previous S elements of Φ_N embodied into the Harmonic Matrix.

With a probability equal to $(1 - p_{HS})$, this new sample is randomly picked up from a set containing only the first $B_{HS}\%$ of the S samples – ordered according to their own harmony

[2] Due to the constraint $N = R \cdot L$, $|\Phi_N|$ depends on the factorization of N. If N is not prime, $|\Phi_N| \gg N$: for instance $|\Phi_N| > 10^3$ already if we consider $N = 10$ and $|\Phi_N| > 10^9$ for $N = 20$.
[3] We will explicitly define the fitness function in **paragraph 2.8**.

value – of the Harmonic Matrix. In this case, the new sample is then slightly modified through a process of shifting: the V_{HS}% of the elements of each subsample (belonging to the new sample) randomly migrates from one subsample to another, in a way to differentiate the new sample from its initial structure.

Through this two different ways of creating new samples, HSA guarantees both a wide range of variability in the search of the optimal solution, both the exploration of locality for already existing samples embodied into the initial Harmonic Matrix.

In fact, after the creation of a new sample, HSA evaluates its harmony value and if such harmony is higher with respect to that one belonging to the worst (always in terms of harmony value) sample already embodied into the Harmonic Matrix, a process of substitution between the two samples occurs. After that, the Harmonic Matrix is sorted again according to the column of the harmony values.

This process of creation of new samples and their evaluation and eventual insertion into the Harmonic Matrix is iterated for a number of times, called cycles of HSA. At the end of the iteration process, the optimal solution is considered to be the sample showing the best harmony value in the Harmonic Matrix.

2.8 Numerical experiments

In order to test the algorithm, the fitness function has defined as exactly **equation (2.45)**, that is we have considered a benchmark problem, embodying in the fitness function itself the value of J. In this case, HSA has proved to be very affordable algorithm: several numerical tests (based on different values of the parameters J and h) have shown that HSA was always able to find a partition of the spin configuration $\boldsymbol{\sigma} = (\sigma_1 \dots \sigma_N)$ with $L = 1\,000$, whose $\check{J}$ differed on average from J for less than 0.1%. Moreover, the search routine was not time-consuming at all: considering that all simulations have been performed on a eMachine E525 with a Celeron processor Dual-Core CPU (T3100 @ 1.90 GHz, 1.90 GHz) and processed through Matlab language (version 7.9.0 R2009b at 32bit), the process of search has required on average less than 20 seconds.

Worth to note, the number S of samples embodied into the Harmonic Matrix, the probability p_{HS} of choosing between the two possible ways of creating a new sample, the

parameters tuning the modification of already existing samples (B_{HS} and C_{HS}) and eventually the number of cycles to be performed in the iteration process may be set in different ways, affecting the speed of the algorithm and its convergence to optimal solutions. Numerical simulations – in agreement with (Geem, 2009) – have suggested that the following setting $(S, p_{HS}, B_{HS}, V_{HS}) = (150, 0.3, 20, 5)$ seemed the best ensemble of parameters to produce valuable results in terms of the optimality defined by the previous fitness function.

The good achievements reached in solving the aforementioned benchmark problem have pushed towards the definition of an optimal subsampling procedure based on a fitness function not embodying this time the parameter J, in order to solve the problem stated in **equation (2.45)**.

Investigations made over the optimal solutions obtained by HSA for the benchmark problem has anyway shown that **equation (2.45)** appears to be a very difficult problem to attack: sometimes minimal differences between samples determine huge changes in $\check{J}$, the estimator of the parameter J. From another perspective, this phenomenon gives reason to the fact that occasionally minimal changes (in the order of 10^{-5}) in the value of $\check{\chi}_{(R)}$, the estimator of the magnetic susceptibility, produce a class of very different $\check{J}$.

In order to define a new fitness function for HSA, it has seemed reasonable that an optimal sample may be built up of R subsamples which were the most possible intrinsically independent. Such supposition has directly led to the idea of defining a function of covariance for the R subsamples, defined by a $L \times L$ covariance matrix C.

Given the possibility of computing such a covariance matrix C for each element of Φ_N, we have explored several possibilities of introducing a metric able to determine distances between two or more elements of Φ_N.

Among many possible definitions, the so-called *Förstner-Moonen* measure (Förstner & Moonen, 1999) has seemed a suitable choice:

$$d(C_A, C_B) = \sqrt{\sum_{i=1}^{n} ln^2 \lambda_i(C_A, C_B)} \tag{2.46}$$

where C_A and C_B represent two covariance matrices and the i-eigenvalue $\lambda_i(C_A, C_B)$ is coming from the equation $|\lambda C_A - C_B| = 0$.

However such kind of measure has not been implemented since dire problems raising from the eigenvalues calculation. In fact, the covariance matrices associated with the elements of Φ_N contain on average a considerable number of elements very close to zero and show no clear structure, so that a correct eigenvalues calculation has resulted almost impossible to carry out, even applying different algorithms of eigenvalues calculation (i.e. Cholesky factorization, Arnoldi iteration) (Arnoldi, 1951; Horn & Johnson, 1985).

Another alternative function of covariance has been introduced, this time relaying on a scalar number defined as:

$$d = \sum_{i>j} |C_{ij}| \tag{2.47}$$

where $C_{(i,j)}$ represent the (i,j) element of the covariance matrix.

Numerical simulations designed to predict the parameter J have shown that no clear pattern is exhibited if the HSA fitness functions is defined as in **equation (2.47)** and the results continue to be unstable and very sensible to initial conditions for different values of the parameter J.

Nevertheless, a small modification to **equation (2.47)**, that is defining it without the absolute value:

$$d = \sum_{i>j} C_{ij} \tag{2.48}$$

leads to the appearance (**figure 2.01**) of a clear decreasing pattern for $\check{J}$, estimator of the parameter J.

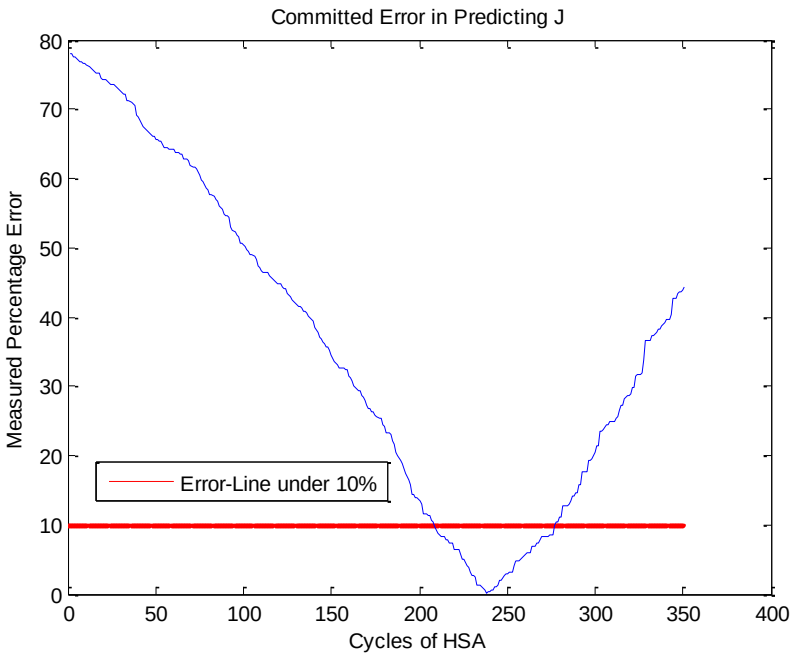


Figure 2.01 *Prediction of the parameter* J. On the horizontal axis, we find the cycles of HSA, whereas on the vertical axis the prediction error measure $\left|J - \check{J}\right|$ in percentage terms is represented. The graph illustrates the results of a numerical experiment where the parameters $J = 1.2$ and $h = 0.1$. The sample associated with the best prediction of the parameter J was found at cycle 238 and showed a percentage error equal to 0.2%.

Several numerical experiments as that one shown in **figure 2.01** have proved similar decreasing pattern for a wide class of values associated to the parameters J and h. Also in this case, computational time were quite satisfying (on average lasting less than one minute) taking into account that such experiments were launched on a Sony Vaio with an Intel Pentium processor Dual-Core (B950 @ 2.10 GHz, 2.10 GHz) and processed through Matlab language (version 7.9.0 R2009b at 32bit).

However, aside from the hint given by the pattern found above, the problem still remains open. In the current framework, the quest for a "stop criterion" (i.e. a procedure able to identify a suitable number of cycles of HSA for a given spin configuration $\boldsymbol{\sigma}$) may be implemented as it would represent a key to solve the problem stated in **equation (2.45)**.

Chapter 3

A statistical mechanics model concerning immigrant integration

After having given an outline of the Curie-Weiss model in the previous chapter, here we focus on a related model presented in (Barra, Contucci, Sandell & Vernia, 2013), always grounded on statistical mechanics methods but this time centered on describing social rather than ferromagnetic phenomena. To be more explicit, the model explores social phenomena through the description of interacting systems behaviors occurring in a given topology, that in a certain way model the concept of integration which may arise within a certain type of mixed population. For the sake of clarity, hereunder we provide the reader with some information which we deem essential to offer a first theoretical picture of the model, before deepening into more technical details in the following paragraphs.

First of all, as mixed population we imagine a set of N individuals which – in a given time – can be partitioned according to two opposite characteristics: individuals being immigrant and individuals being native, where we imagine natives to be a major subset of the population consisting of individuals sharing the same (dominant) citizenship. For social phenomena of integration, instead, we refer to social facts as mixed marriages (i.e. marriages involving an immigrant and a native) or mixed newborns originated by a mixed couple, that is a pair of individuals always formed by an immigrant and a native but not necessarily married.

Given these social phenomena, the model we consider is aimed at describing a scenario where mixed marriages or mixed newborns can be the possible results of two different modalities in which an individual may choose and act. In particular, an individual (whatever being native or not) may take a choice over the idea of marrying a native or an immigrant in an independent way (i.e. choosing by himself, without being influenced by other people) or he may take the same decision because of the influence that is exercised over him by other individuals of the population. The model is therefore intended to embody these two fundamental characteristics: it has to give reason of individual choices when these ones are assumed in total independence, along with classical discrete choice theory (McFadden, 2001), or when such choices are assumed through peer-to-peer interaction, when imitation among individuals is instead the dominating factor. The latter case is known in sociology as *social action* and its paramount importance in affecting social phenomena was highlighted by a classical sociologist like Max Weber[4], already almost one century ago (Weber M., 1922).

It is manifest that such a model resounds of the theoretical structure the Curie-Weiss model entails (**chapter 2**), in a way that is possible to draw a partial analogy between individuals and magnets, social phenomena and ferromagnetic ones. Within such perspective, the model is similarly based on an internal structure (microscopic theory) whose emergent social behavior (macroscopic theory) aims to reproduce through probability measures the aggregated choices of individuals over the aforementioned social facts (i.e. marriages and newborns).

The chapter will be organized following the path marked in (Barra, Contucci, Sandell & Vernia, 2013). In the first paragraph, we outline a first version of the model (i.e. the mono-populated one) where the population of N individuals includes only natives: in describing such model we do not consider the possibility that individual choices regarding marriages or newborns can be taken because of interaction among individuals. In the second paragraph, instead, we consider a population of N individuals which can be partitioned into two subsets, this time including into the population a certain amount of immigrants which is in any case lower with respect to the remaining number of natives. This bi-populated model embodies also the possibility of interaction among individuals, realized through the

[4] In his last work titled "Wirtschaft und Gesellschaft ", Weber noticed: "*Social action is not identical with the similar actions of many persons […]thus, if at the beginning of a shower a number of people on the street put up their umbrellas at the same time, this would not ordinarily be a case of action mutually oriented to that of each other, but rather of all reacting in the same way to the like need of protection from the rain*".

implementation of a Hamiltonian function, likewise to those ones introduced in **chapter 2**. Alongside the statistical mechanics formalization, both in the second and third paragraph we exploit monomer-dimer graphs (Heilmann & Lieb, 1970), a discrete mathematics representation of the models' elements through a set of vertices and edges. Furthermore, both for the mono-populated and bi-populated models we derive probability measures reproducing the frequencies first of marriages and newborns and then of mixed marriages and mixed newborns. Eventually, in the third paragraph we consider the mean field limit of the bi-populated model, highlighting how the probability measure changes in functional form with respect to the level of interactions among individuals.

3.1 Mono-populated model

The first type of model we describe is the mono-populated one, where no immigrants are present.
Given a population, that is a set of N individuals named $I = \{1, \dots, N\}$, we define a marriage configuration M as the union of two sets S_M and C_M, that is $M = S_M \cup C_M$.

The subset $C_M \subset I^2$ represents all the paired individuals (i.e. possible[5] married couples) of the M configuration. Each element $(i,j) \in C_M$ is denoted by having the following properties:

- $i \neq j$ (no self-loops)
- $(i,j) = (j,i)$ (symmetric relation)
- $if\ (i,j) \in C_M \rightarrow (i,k) \notin C_M\ \forall k \neq j$ (monogamy constraint)
- (i,j) is connected *only* by one link, symbolizing a possible marriage union

The subset $S_M \subset I$ represents instead the unpaired individuals (i.e. singles) of the M configuration. The following relation therefore holds true:

$$|S_M| = N - 2|C_M| \tag{3.01}$$

[5] The complete information concerning real married couples will be given accounting also the $\varepsilon_{i,j}$ elements of the acquaintance matrix, which will be outlined below.

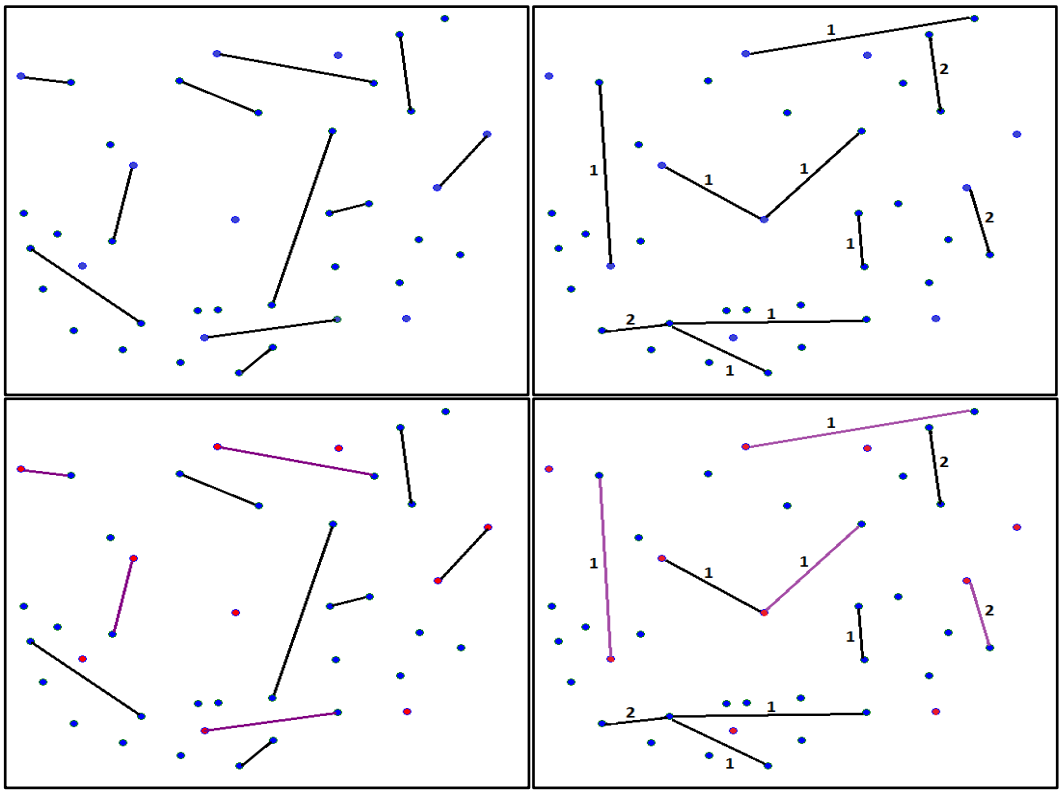


Figure 3.1 Upper panels: *a possible marriage configuration* (**left**) and *a possible filiation configuration* (**right**) in the mono-populated model. Blue spots represent a population of 40 individuals. Black lines are in turn possible marriages or newborns, whose multiplicity per couple is labelled through a number. **Lower panels:** *a possible marriage configuration* (**left**) and *a possible filiation configuration* (**right**) in the bi-populated model. Blue spots represent 30 natives, whereas red spots signal 10 immigrants. Pink lines are in turn possible marriages or newborns within mixed couples, whereas black lines mirror only those ones within individuals sharing the same citizenship.

Following a discrete mathematics approach, a marriage configuration $\boldsymbol{M}$ can be also visually represented through a particular type of graph named monomer-dimer. In such graph, each individual is represented as a point which can be isolated (monomer) or undirectedly linked to another one (dimer). **Figure 3.1** (left upper panel) shows a certain marriage configuration $\boldsymbol{M}$: we note 11 dimers (symbolizing 11 possible marriages) and 18 monomers (singles). Looking at the panel, it is manifest that this set $\boldsymbol{M}$ is only one representation of all possible ways of constructing such a monomer-dimer graph. Thus, we call $\boldsymbol{\mathcal{M}}$ the set of *all* marriage configurations $\boldsymbol{M}$.

Since each marriage configuration M is univocally determined by its subset C_M, a combinatorial reasoning like in (Alberici, Contucci & Mingione, 2013) proves that $\mathcal{M}$ cardinality can be computed as:

$$|\mathcal{M}| = \sum_{d=0}^{[N/2]} \frac{N!}{d!\,(N-2d)!} 2^{-d} \tag{3.02}$$

Aside from visual representation, if we focus on the elements of the two subset S_M and C_M of a given marriage configuration M, we could imagine that each element of the set S_M shows a certain inclination to marry versus remaining singles and each element of the set C_M has its own likelihood to marry. We call therefore s_i this tendency for the i-element of the set S_M and $c_{i,j}$ that one for the couple (i,j) of the set C_M. Both s_i and $c_{i,j}$ can be thought as weights for the i-single and for the (i,j) couple, imposing that these parameters have to be positive real numbers.

Eventually, the model is fully defined only together with E – the acquaintance matrix of the N individuals of the population – whose elements $\varepsilon_{i,j} \in \{0,1\}$ signal whether the i-individual and the j-individual of the (i,j) couple are acquainted ($\varepsilon_{i,j} = 1$) or not. Worth to note (Bianconi, 2002; Agliari & Barra, 2011; Barra & Agliari, 2012), the $\varepsilon_{i,j}$ elements account for the topology of the graph, like d-dimensional lattices, or the complete graph of N-points or more refined structures like Erdos-Renyi and small world graphs.

Given the weights s_i and $c_{i,j}$ and the $\varepsilon_{i,j}$ acquaintance matrix elements, we can write the partition function (Alberici, Contucci & Mingione, 2013) of this system as:

$$Z^{(\mathcal{M})} = \sum_{M \in \mathcal{M}} \prod_{(i,j) \in C_M} \varepsilon_{i,j} c_{i,j} \prod_{i \in S_M} s_i \tag{3.03}$$

Similarly to what implemented for the marriages, we define a configuration of filiation F as the union of two sets U_F and P_F, that is $F = U_F \cup P_F$.

The subset $U_F \subset I$ represents the unpaired individuals (i.e. undescendents) of filiation configuration F, whereas the subset $P_F \subset I^2$ mirrors all the paired individuals (i.e. only those couples – this time not necessarily married – who are characterized by possibly[6]

[6] Again, the complete information concerning real couples with children will be given accounting also the $\varepsilon_{i,j}$ elements of the acquaintance matrix.

having children) of the same configuration. For each element $(i,j) \in P_F$, we define a function $l: P_F \to \mathbb{N}^+$, where $l(i,j)$ counts the number of links, i.e. children belonging to the (i,j) couple. Each element $(i,j) \in P_F$ is denoted by having the similar but not equal properties to those ones previously described for the elements of the C_M set, namely:

- $i \neq j$ (no self-loops)
- $(i,j) = (j,i)$ (symmetric relation)
- (i,j) is connected *at least* by one link ($l(i,j) \geq 1$), symbolizing in this case a newborn originated by the couple.

The number $l(i,j)$ for each couple (i,j) may be modeled through a certain probability distribution, likely to be a Poisson distribution[7] ρ of a given average λ. More explicitly, we claim that:

$$\rho_\lambda\big(l(i,j)\big) = \exp(-\lambda)\frac{\lambda^{l(i,j)}}{[l(i,j)]!} \tag{3.04}$$

Therefore, we may associate a Poisson probability measure to the entire filiation configuration F, namely:

$$\rho(F) = \prod_{(i,j)\in P_F} \exp(-\lambda)\frac{\lambda^{l(i,j)}}{[l(i,j)]!} = \exp(-|P_F|\lambda)\frac{\exp\left[\left(\sum_{(i,j)\in P_F} l(i,j)\right)\ln\lambda\right]}{\prod_{(i,j)\in P_F}[l(i,j)]!} \tag{3.05}$$

Figure 3.1 (right upper panel) shows a certain filiations configuration F: we note 13 dimers (symbolizing possible children born from 10 different couples) and 23 monomers (undescendents). With respect to C_M elements, we can stress that the two main differences in this monomer-dimer graph are the lack of the monogamy-constraint property (i.e. individuals may have children not only with another unique partner, meaning that an individual may belong to more than one couple) and that there can be more than one link connecting two individuals (that is, a couple may have more than one child). Contrarily from what shown for the subset C_M and S_M in **equation (3.1)**, the fact that individuals may belong to more than one couple (that is, having children with different partners) precludes the possibility of writing a cardinality relation between $|U_F|$ and $|P_F|$.

[7] Although the choice of such a distribution can sound reasonable for this context, the following results do not strictly depend on it.

For the same reasoning made above for the marriages, we call $\mathcal{F}$ the set of all filiation configurations. Moreover, if we focus on the elements of the two subset U_F and P_F of a given filiation configuration F, we could imagine that – in analogy to what previously drawn for S_M and C_M – each element of the set U_F shows an individual tendency to have children and each element of the set P_F too. We call therefore u_i this tendency for the i-element of the set U_F and $p_{i,j}$ that one for the couple (i,j) of the set P_F. Again, both u_i and $p_{i,j}$ can be thought as weights for the i-undescendent and for the (i,j) parents, imposing that these parameters have to be positive real numbers.

Likewise to **equation (3.3)**, considering the $\varepsilon_{i,j}$ acquaintance matrix elements, the partition function of this system is given by:

$$Z^{(\mathcal{F})} = \sum_{F \in \mathcal{F}} \rho(F) \prod_{(i,j) \in P_F} \varepsilon_{i,j} p_{i,j} \prod_{i \in U_F} u_i \tag{3.06}$$

Looking at the partition functions in **equations (3.3)** and **(3.6)**, at a first stage one may consider both $(s_i, c_{i,j})$ and $(u_i, p_{i,j})$ parameters as random variables so that the Z's would define disordered models. However, in order to pave the way to a more treatable approach of this model, it would sound reasonable to set these couples of parameters to be constant. This will be explicitly done in the mean field model that we treat in **paragraph 3.3**.

After having defined the partition functions for the model, we now focus on probability measure.
Calling K_M the total number of links[8] in the configuration M and defining the frequency as $\nu_M = 2K_M/N$, the expected value of the marriage frequency can be computed as

$$P_{\mathcal{M}} = \mathbf{Av} \frac{\sum_{M \in \mathcal{M}} \nu_M \prod_{(i,j) \in C_M} \varepsilon_{i,j} c_{i,j} \prod_{i \in S_M} s_i}{Z^{(\mathcal{M})}} \tag{3.07}$$

where the average operation $\mathbf{Av}$ is computed on the acquaintance matrix ensemble.

[8] It holds that $K_M = |C_M|$

Analogously, calling K_F the total number of links[9] in the configuration F and defining the frequency as $\nu_F = 2K_F/N$, the expected value of the newborn frequency is

$$P_{\mathcal{F}} = \mathbf{Av}\frac{\sum_{F\in\mathcal{F}} \nu_F \rho(F) \prod_{(i,j)\in P_F} \varepsilon_{i,j} p_{i,j} \prod_{i\in U_F} u_i}{Z^{(\mathcal{F})}} \tag{3.08}$$

3.2 Bi-populated model

The second type of model we describe is the bi-populated one, considering now that the population of N individuals can be partitioned into two subsets, representing the natives (whose cardinality is N_{nat}) and the immigrants (whose cardinality is N_{imm}). For this reason, we set both parameters s_i and u_i to take two value each, depending only on the individual being immigrant (Imm) or native (Nat) and the same holds true also for the couples parameters ($c_{i,j}$ and $p_{i,j}$) which take only three values for three cases (Imm, Imm), (Nat, Imm) and (Nat, Nat). Furthermore, in this bi-populated model we include also an imitative ($J \geq 0$) interaction between the two populations with the introduction of a suitable mean-field Hamiltonian function, consistently with (Gallo & Contucci, 2008; Contucci, Gallo & Menconi, 2008; Agliari & Barra, 2011; Barra & Agliari, 2012):

$$H(M) = -J_M \sum_{i\in Nat, j\in Imm} \varepsilon_{i,j}\sigma_i\sigma_j \tag{3.09}$$

where

$$\sigma_i = \begin{cases} +1 & \text{if } i \text{ belongs to a mixed marriage} \\ -1 & \text{otherwise} \end{cases} \tag{3.10}$$

and J_M is the interaction parameter for the marriage configuration M.

[9] Differently from the previous note, here the inequality $|P_F| \leq K_F = \sum_{(i,j)\in P_F} l(i,j)$ holds.

A similar Hamiltonian can be introduced to model the interactions within a given filiation configuration F, so that:

$$H(F) = -J_F \sum_{i \in Nat, j \in Imm} \varepsilon_{i,j} \tau_i \tau_j \tag{3.11}$$

where

$$\tau_i = \begin{cases} +1, & \text{if i has a child within a mixed couple} \\ -1, & \text{otherwise} \end{cases} \tag{3.12}$$

and J_F is the interaction parameter for the filiation configuration F.

Again the $\varepsilon_{i,j}$ term we find in the Hamiltonian functions is a generic element of the acquaintance matrix. Both Hamiltonians depend on the configurations M and F since σ's and τ's do and since the Hamiltonian functions embody the product $\sigma_i \sigma_j$ and $\tau_i \tau_j$ (where i-element is a generic native and the j-element is a generic immigrant), they result to be proportional to the quantity $\gamma(1-\gamma)$ where

$$\gamma = \frac{N_{imm}}{N} \tag{3.13}$$

is defined as the density of immigrants for the population of N individuals.

In **figure 3.1** (lower panels), a monomer-dimer representation for certain configurations M and F is again pictured, this time according to the bi-populated model: here, blue spots represent natives, whereas red ones immigrants. Both panels show a population of 40 individuals, of which 10 immigrants. In the **left lower panel**, possible marriages are represented by colored lines connecting two points according to whether a particular marriage is among the same colors (black line) or among different ones (pink lines). In the panel, 4 couples out of 11 are characterized by a mixed marriage. 18 individuals are singles. In the **right panel:** colored lines connect two points according to whether a newborn is originated among the same colors (black line) or among different ones (pink lines). In the panel, 4 couples out of 10 are mixed couples and 5 children out of 13 belong to mixed couples. 23 individuals are undescendent.

As eventually done for the mono-populated model, we now describe the partition functions for the bi-populated model and we compute the expected values of the frequency of the two observables, following the same path shown above.

Given the Hamiltonians introduced in **equation (3.09)** and **(3.11)**, the bi-populated partition functions are:

$$Z_H^{(\mathcal{M})} = \sum_{M \in \mathcal{M}} e^{-H(M)} \prod_{(i,j) \in C_M} \varepsilon_{i,j} c_{i,j} \prod_{i \in S_M} s_i \tag{3.14}$$

$$Z_H^{(\mathcal{F})} = \sum_{F \in \mathcal{F}} e^{-H(F)} \rho(F) \prod_{(i,j) \in P_F} \varepsilon_{i,j} p_{i,j} \prod_{i \in U_F} u_i \tag{3.15}$$

Furthermore, calling M_M the number of mixed marriages and K_M the total number of marriages in the configuration M and defining the frequency of mixed marriages $f_M = M_M/K_M$ we have that its expected value, that is, the probability of mixed marriages is:

$$P_{\mathcal{M}}^{(Nat,Imm)} = \mathbf{Av} \frac{\sum_{M \in \mathcal{M}} f_M e^{-H(M)} \prod_{(i,j) \in C_M} \varepsilon_{i,j} c_{i,j} \prod_{i \in S_M} s_i}{Z_H^{(\mathcal{M})}} \tag{3.16}$$

And analogously, the probability of mixed children

$$P_{\mathcal{F}}^{(Nat,Imm)} = \mathbf{Av} \frac{\sum_{F \in \mathcal{F}} f_F e^{-H(F)} \rho(F) \prod_{(i,j) \in P_F} \varepsilon_{i,j} p_{i,j} \prod_{i \in U_F} u_i}{Z_H^{(\mathcal{F})}} \tag{3.17}$$

is given in terms of frequency of children from mixed couples $f_F = M_F/K_F$ where M_F is the number of children from mixed couples and K_F the total number of children in the configuration F.

3.3 Mean-fields limits

Albeit an exact solution of the general model introduced above is not yet available, we can still obtain results for a wide variety of cases that include the mono-populated and bi-populated mean fields limits, as proved in (Gallo & Contucci, 2008; Contucci, Gallo & Menconi, 2008; Barra & Contucci, 2010). In particular, the bi-populated model shows two regimes: the *imitative regime* in which the interaction parameter J introduced in the Hamiltonians (**equations (3.09)** and **(3.11)**) dominates the dynamics of the system and *the free regime* where the contrary holds true.

In particular, if we express **equations (3.16)** and **(3.17)** as a function of the quantity $\gamma(1-\gamma)$, where γ is the density of immigrants introduced in **equation (3.13)**, and we consider the free regime (the equivalent of an ideal gas of independent particles), one find for the probability measures a dependence of the type:

$$P(\gamma) \propto \gamma(1-\gamma) \tag{3.18}$$

In the imitative regime, on the other hand, the interactions among individuals encoded in the Hamiltonians favor another type of behavior with respect to what shown above. As highlighted in **paragraph 3.2**, the mean field Hamiltonians have a size proportional to $\gamma(1-\gamma)$ and, for various adjacency matrices $\varepsilon_{i,j}$ defining diluted topologies like small words or random graphs (Bianconi, 2002; Leone, Vasquez, Vespignani & Zecchina, 2002; Agliari & Barra, 2011; Barra & Agliari, 2011), the model predicts a behavior of the type:

$$P(\gamma) \propto [\gamma(1-\gamma)]^{\frac{1}{2}} \tag{3.19}$$

The mechanism underlying such behavior is – as far as the network is over-percolated (Bollobas, 2000) and the interactions among agents are the only dominant drive – the mean-field ferromagnet with the critical exponent one half, that is the Curie-Weiss model which has been outlined in **chapter 2**. In this context, the critical value of γ (that is the threshold over which **equation (3.19)** holds true) turns out to be very close to zero as shown in (Bianconi, 2002; Leone, Vasquez, Vespignani & Zecchina, 2002; Dembo & Montanari, 2010; Agliari & Barra, 2011).

Comparing **equations (3.18)** and **(3.19)**, we can therefore underline that the two regimes show a structural difference, due to some divergence of the derivative of the previous P's formulas. In fact, the regimes are determined by the behavior of the probability measures in

a neighborhood of $\gamma = 0$, where in one case the derivative is singular at $\gamma = 0$, whereas in the other case it is not.

In social sciences, the relevance of such models have been clearly recognized by Durlauf (Durlauf, 1999): such perspective, along with that one presented in (Barra, Contucci, Sandell & Vernia, 2013), will be the theoretical guideline for the following chapter.

Chapter 4

Empirical evidences of a statistical mechanics model concerning immigrant integration

In this chapter, we want to cast light over the way immigrant integration phenomena occur, assuming a quantitative point of view based on the exploitation of large empirical datasets. Starting from the seminal work on Spanish immigration data (Barra, Contucci, Sandell & Vernia, 2013) – whose mathematical model has been presented in **chapter 3** – we share the simple observation that very little is known about the mechanisms that bring about integration. For example, elementary questions concerning how integration responds to an increase in immigration density or to what extent the intensity of interaction[10] modifies the level of integration still beg coherent empirical and theoretical answers. It is therefore manifest that the missing of knowledge over the mechanisms internally governing such phenomena undermines the effectiveness of formulating social policies concretely able to promote integration. This work thus proposes new perspectives according the theoretical modelling based on statistical mechanics methods, already outlined in the previous chapter and whose growing importance in social sciences (Galam & Moscovici, 1991; Durlauf, 1996; Durlauf, 1999; Brock & Durlauf, 2001; Castellano, Fortunato & Loreto, 2009; Montanari & Saberi, 2010) has been extensively underlined in **chapter 1**.

[10] In this context, we refer to *social* interaction, that is a sequence of social actions between individuals (or groups) who modify their actions and reactions due to the actions by their interaction partners. The underlying notion of social action has been introduced in **chapter 3**.

According to this perspective, the main result of this chapter is the identification and theoretical interpretation of two empirical laws (a linear one and a square root one) which connect two quantifiers of integration (i.e. the percentage of mixed marriages and the percentage of mixed newborns) to the density of immigrants related to the geographical areas where the previous quantifiers have been measured. This result has been obtained analyzing three large datasets containing over 10^6 information regarding marriages and births registered in Italy during an eleven years span and only marriages occurred in the Emilia Romagna region (Italy), during a sixteen years span. It is important to stress that whereas data coming from the latter dataset present only one of these empirical laws (the square root law), both Italian datasets have been characterized by showing the two distinct patterns only when partitioning data according to the size of their municipalities' population.

This work has required many analytical processes which will be outlined throughout this chapter. After having presented in the first paragraph the social indicators of integration we consider, in the second paragraph we give a review of data collection and classification issues, describing the process and methodologies implemented in storing data. In the third paragraph, we provide a graphical representation of data through scatter-plots, analyzing their cardinality, whereas in the fourth paragraph we focus on data density issues and on the way immigrants and natives distribute themselves with respect to small-sized and large-sized municipalities, given a division threshold between these two types of municipalities at 10 000 inhabitants. In the fifth and sixth paragraphs we discuss binning procedures – since our interest is directed towards the quantifiers' averages as a function of the density of immigrants – and we present the process of curve fitting which has led to the emergence of the two aforementioned empirical laws. Whereas the previous paragraphs were considering analyses conducted on data independently from the time period they were referring to, in the seventh paragraph we draw a time-dependent analysis; its outcome leads to a confirmation of our previous results, proving that the empirical laws we found emerge also from the very first years of data observation, thus making them also predictive. The eighth paragraph finally offers a theoretical interpretation of these two empirical laws, linking them to the statistical mechanics model we outlined in **chapter 3**. In specific, Italian data suggest that in small-sized municipalities imitative phenomena mainly take place (as for all Emilia Romagna data) while independent choices seem to be the most common patterns in large-sized municipalities. This result turns out to be even more impressive if compared with subtle considerations made by classical sociology authors, who understood a crisp

difference in the emergence of social actions occurring in small-sized or large-sized municipalities already more than one hundred years ago. The final paragraph provides an outlook over the next analyses that could be drawn over the three datasets. Centered on this work, an article – co-authored by Prof. Pierluigi Contucci (University of Bologna) and Prof. Cecilia Vernia (University of Modena and Reggio Emilia) – can be found in **Appendix 3**.

4.1 Social indicators

We consider a population of N individuals which can be partitioned into two subsets: N_{nat}, the number of natives and N_{imm}, the number of immigrants. We recall here the definition of the density of immigrants as given in **equation (3.13)**:

$$\gamma = \frac{N_{imm}}{N_{imm} + N_{nat}} \in [0,1] \tag{4.01}$$

Along with the social phenomena of integration introduced in **chapter 3**, we therefore consider the set of marriages M, occurring within the population N, and we partition it into three subsets: M_{nat} representing the number of marriages occurring between natives couples, M_{imm} representing the number of marriages occurring between immigrant couples and M_{mix} representing the number of marriages occurring between mixed couples.

The fraction of mixed marriages M_m is therefore defined as:

$$M_m = \frac{M_{mix}}{M_{mix} + M_{nat} + M_{imm}} \in [0,1] \tag{4.02}$$

In the same way, we consider as well the set of newborns B and we partition it into three subsets: B_{nat} representing the number of newborns originating from native couples, B_{imm} representing the number of newborns originating from immigrant couples and B_{mix} representing the number of newborns originating from mixed couples.

The fraction of mixed newborns B_m is therefore defined as:

$$B_m = \frac{B_{mix}}{B_{mix} + B_{nat} + B_{imm}} \in [0,1] \tag{4.03}$$

While γ is an observable that measures the presence of immigrants in a given population, the quantifiers M_m and B_m are observables that weight the level of social integration between the two subsets of natives and immigrants. M_m and B_m can be therefore seen as social indicators that numerically describe the magnitude of social phenomena occurring in a given population containing both natives and immigrants in a proportion described by γ.

As written in the above introduction, our goal is a statistical mechanics predictive theory by which the magnitude of these integration quantifiers can be expressed as a function of the density of immigrants. More specifically, we are interested in describing the different quantifiers as a function of the quantity $\Gamma = \gamma(1-\gamma) \in \left[0, \frac{1}{4}\right]$, since this variable tunes the total number of available cross-links couplings among two populations (native and immigrant), according to the relation $N_{imm}N_{nat} = \Gamma N^2$.

4.2 Data collection and classification

As in many other countries, in Italy marriages and newborns are registered by civil registration officers, whose registry is embodied into the *Anagrafe* office[11]. In this sense, data collection is first made directly by civil registration officers that transcript data which are – by force of law – personally communicated by citizens. Moreover, through SISTAN (Sistema Statistico Nazionale), data registered in each Italian municipality are exchanged with other federated institutions containing statistics offices (like regions, ministers, prefectures or chambers of commerce) and totally grouped by ISTAT, the Italian Institute of Statistics. We can therefore state that there are two levels of data collection: the first level is played by every singular civil registry office, whereas the second level is played by all other state institutions belonging to SISTAN, which have the exclusive right of accessing to and eventually grouping such information.

Since the particular structure of data collection, there is another important issue to highlight: data coming from civil registry offices are example of a census. A census refers to data collection about everyone or everything in a group or statistical population and has advantages such as accuracy and detail, and disadvantages such as cost and time. Census is opposed to sampling techniques, which extract valuable information only from a part of the statistical population at cost of being less accurate and reliable.

[11] A mandatory structure present in each municipality.

As reported above, the work has been centered on two large datasets released by ISTAT and the Emilia Romagna Region, the former one containing information recorded in all Italian municipalities regarding marriages and births occurred during an eleven years span, from 2001 to 2011; the latter one containing information recorded in all Emilia-Romagna Region municipalities regarding marriages occurred during a sixteen years span, from 1995 to 2010. A deeper description of each dataset is presented below. However, it is already important to highlight that in each dataset, a "native individual" points to an individual bearing Italian citizenship (or Sammarinese one, as explained below), whereas an "immigrant" is a foreigner not bearing the Italian citizenship.

Emilia-Romagna

The dataset from Emilia-Romagna Region was the smallest one to be put under analysis. Before deepening into its description, we offer here some brief information regarding Emilia Romagna Region and its population.

Emilia Romagna is an administrative Region of the Italian Republic, situated in northern part of the Italy, comprising an area of 22 446 km^2 and about 4.4 million inhabitants (4 429 766 inhabitants resulted at the 2011 ISTAT census). In November 2013, the Emilia-Romagna Region results to be divided into nine provinces, containing 348 municipalities. These municipalities significantly vary in population's size: according to 2010 ISTAT population data, in Emilia Romagna region the most populated municipality resulted Bologna with 380 181 inhabitants, whereas the least populated one was Zerba, a municipality belonging to the Province of Piacenza, whose population accounts only for 94 people.

The regional statistics office website[12] offers many valuable information over social phenomena occurring in the region: focusing on integration quantifiers, the website releases sensible data regarding marriages occurred in all Emilia Romagna municipalities. In specific, data are accessible per every municipality (identified by its own name and its related ISTAT code, a numerical code univocally determining each Italian municipality) at a given year and consist in the number of registered marriages, separated per citizenship of spouses according to two categories: Italians and Foreigners. Since the particular way of accessing to data (per municipality at a given year), collecting the entire dataset is not a

[12] http://statistica.regione.emilia-romagna.it/servizi-online/statistica-self-service [Accessed 2013, December 15]

trivial problem. Automated techniques of web-contents wrapping have been employed to solve this task, as described in the **Appendix 1** presented at the end of this thesis[13].

After the data downloading, information regarding marriages registered in a given municipality at a given year are thus stored in a 2x2 matrix (as in **table 4.1**), whose rows are referring to the nationality of the male spouse, whereas the columns are referring to the nationality of the woman spouse[14].

Spouse Citizenship	Italy, San Marino Rep.	Foreign Countries
Italy, San Marino Rep.	1.371	77
Foreign countries	41	14

Table 4.1 Year 1995. Registered marriages in the municipality of Bologna. Numbers in counterdiagonal represent mixed marriages.

All considered data are censual and are available from 1990 until 2010. Nonetheless, information about citizenship of spouses is available only from 1995: for this reason only a 16 years span from 1995 to 2010 was considered.

Together with the information on marriages, immigrant density γ is fundamental information required for the model presented in **chapter 3**. In this case, information regarding immigrant population and total population is available online from 2001 from

[13] As technical note, all the experimental computations presented from this point forward have been conducted on a Sony Vaio with an Intel Pentium processor Dual-Core (B950 @ 2.10 GHz, 2.10 GHz) and processed through Matlab language (version 7.9.0 R2009b at 32bit). Since the particular nature of the task, web-data wrapping has been instead processed through Python language (version 2.7.3 at 32bit).

[14] To be noted, spouses bearing the Sammarinese citizenship were considered as Italian spouses, even if they legally belong to another independent state. This exception has to be stressed since San Marino is located within the Emilia Romagna region: anyway, the impact of this choice has to be considered low, since San Marino population accounts today for slightly more 32 000 inhabitants and even a smaller amount in the past twenty years (source: 2013-1993 population data, Ufficio Informatica, Tecnologia, Dati e Statistica, Repubblica di San Marino).

ISTAT website[15]. For previous years, the regional statistics office has released such information for the years from 1995 until 1998, whereas in 1999 and 2000 the survey on immigrant population was suspended by ISTAT. For these two specific years, the immigrant density γ was linearly estimated, according to border values in 1998 and in 2001.

Data collected were at last displaced in a 5 463 x 9 matrix containing therefore 49 167 entries.

The number of rows of this matrix is 5 463, resulting from data collected in 341 municipalities for a 15 years span (1995-2009), plus data coming from 348 municipalities collected in 2010. The imbalance in the number of municipalities is due to the fact that 7 new municipalities have been aggregated to the Emilia Romagna region in 2010.

Moreover, the 9 columns of this matrix contain:

- the name of the municipality
- the ISTAT numerical code
- the year data were referring to
- the immigrant population N_{imm}
- the total population N
- the number of marriages occurring only between Italian spouses M_{nat}
- the number of marriages occurring only between immigrant spouses M_{imm}
- the number of marriages occurring between Italian male spouses and immigrant female spouses
- the number of marriages occurring between immigrant male spouses and Italian female spouses.

The sum of the two latter entries is M_{mix}, the number of marriages occurring between mixed couples.

It is important to notice that embodying the ISTAT numerical code is a redundant choice motivated by the fact that there are many municipalities in Emilia Romagna Region sharing similar names.

[15] http://demo.istat.it [Accessed 2013, December 15]

Italy

The Italian Republic is a country comprising an area of 301 338 km^2 and about 60 million inhabitants (59 433 744 inhabitants resulted at the 2011 ISTAT census). In November 2013, the Italian Republic results to be divided into one hundred ten provinces: the provinces themselves contain 8 093 municipalities. The number of municipalities is not forcedly fixed but has varied year per year, due to administrative reforms which have in turn settled new municipalities or erased old ones.

The actual 8 093 municipalities of the Italian Republic significantly vary in population's size: according to 2011 ISTAT census, in Italy the most populated municipality resulted Rome with 2 614 263 inhabitants, whereas the least populated one was Pedesina, a municipality belonging to the Province of Sondrio (Lombardia Region), whose population accounts only for 31 people.

Differently from the Emilia Romagna region dataset – which required a complex task of collection through downloading – Italian data were directly released by ISTAT, in spreadsheet format. Data contained information regarding marriages and newborns recorded in all Italian municipalities during a 11 years span, from 2001 to 2011. Instead, information regarding immigrant population and total population was again available online from 2001 from ISTAT website. These data were crucial to compute immigrant density $\boldsymbol{\gamma}$, which was subsequently matched to the previous marriages and newborns data.

Data collected were at last displaced in a 89 093 x 13 matrix containing therefore 1 158 209 entries.

The number of rows of this matrix is 89 093, resulting from data collected in a varying number of municipalities for an 11 years span (2001-2011), as presented in **figure 4.1**.

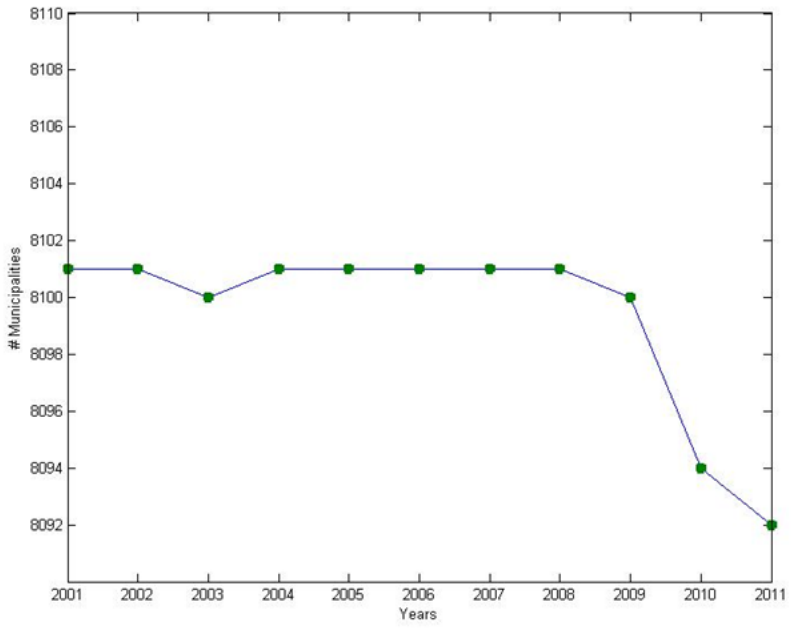


Figure 4.1 Number of Italian municipalities from 2001 to 2011

Moreover, the 13 columns of this matrix contain data classified as in the previous 9-column matrix (that one referring to Emilia Romagna region dataset), plus other 4 columns which contain:

- the number of newborns originating only from Italian couples B_{nat}
- the newborns originating only from immigrant couples B_{imm}
- the number of newborns originating from a couple including an Italian male and an immigrant female
- the number of newborns originating from a couple including an immigrant male and an Italian female.

The sum of the two latter entries is B_{mix}, the number of newborns originating only from mixed couples.

In this case, embodying the ISTAT numerical code is not only redundant choice but a mandatory one since there are exactly 16 municipalities in Italy bearing the same name, according to ISTAT data.

4.3 Graphical representation of data and cardinality analysis

The aforementioned matrices permit the calculation of immigrant density γ and of mixed marriages M_m and mixed newborn B_m quantifiers.

The first step of our analysis is to collect all the observed data for each dataset in a Cartesian plane, representing in the horizontal axis the immigrant density and in the vertical axis the value of the integration quantifier. For a given quantifier, the graphical output is therefore a scatter plot, i.e. a set of points, each one representing information coming from a given municipality in a given year. This representation of data has been conducted according to the prescriptions of a time-independent analysis, i.e. data have been plotted together independently from the year they were referring to. As represented in **figure 4.2**, **figure 4.3** and **figure 4.4**, the sketched sets quite resemble clouds, whose cardinality and shape are not the same but vary according to minor events occurring in tiny municipalities. These differences will be investigated below, putting the focus on data cardinality issues.

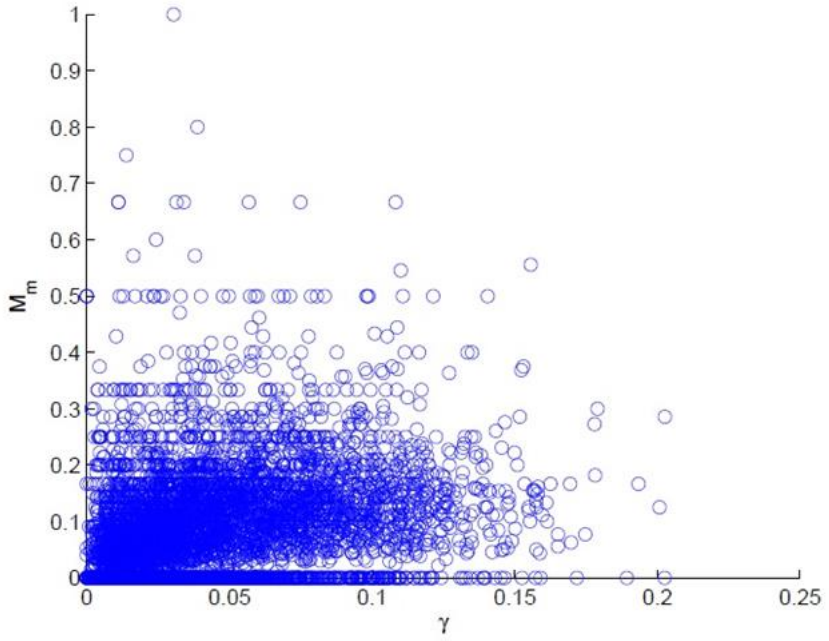


Figure 4.2 *Emilia Romagna dataset.* Raw data versus γ. Blue points represent the fraction of mixed marriages occurred in a municipality where a percentage γ of immigrants is present.

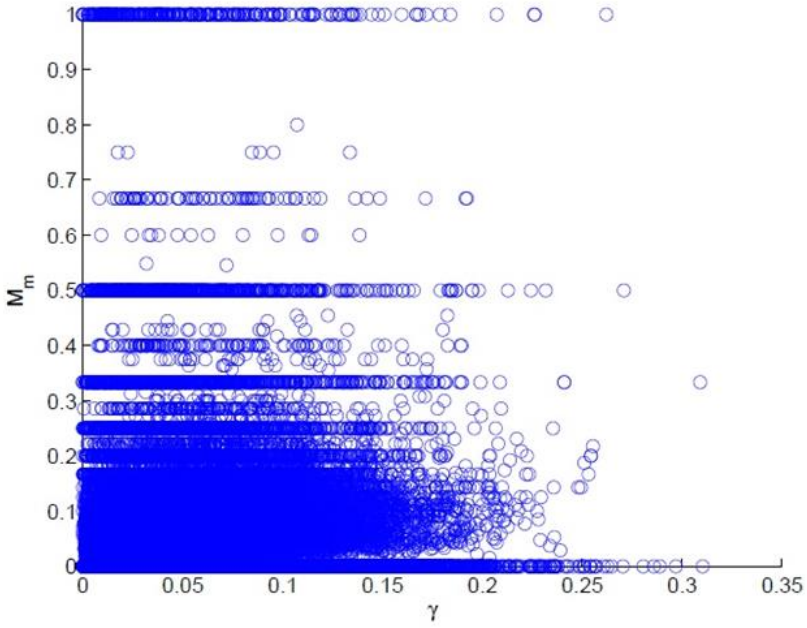


Figure 4.3 *Italian dataset*. Raw data versus γ. Blue points represent the fraction of mixed marriages occurred in a municipality where a percentage γ of immigrants is present.

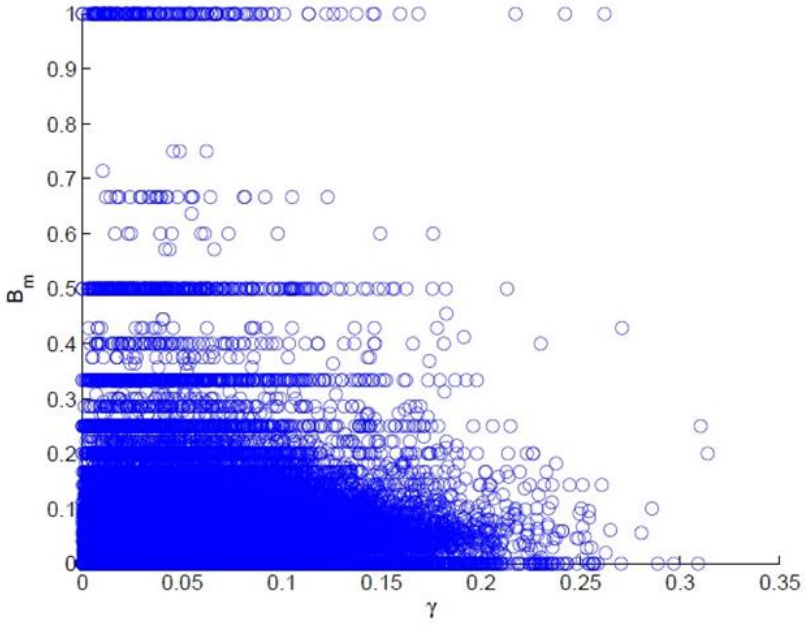


Figure 4.4 *Italian dataset*. Raw data versus γ. Blue points represent the fraction of mixed newborns recorded in a municipality where a percentage γ of immigrants is present.

Regarding data cardinality of the clouds, there are two particular issues to be discussed.

First, since each point in the ensemble plane represents information coming from a given municipality in a given year, one may expect that each municipality shall appear T times, where T is the number of years considered in a given dataset. For example, if we look at **figure 4.4**, information coming from a given municipality should appear 11 times, since the period considered goes from 2001 to 2011 and data were yearly recorded. In effect, this seems to be the case for many municipalities but not for all municipalities: as shown in **figure 4.1** the number of municipalities is not forcedly fixed but varies year per year. Moreover, there is another reason for which information coming from a given municipality in a given year shall not appear T times. In fact, since in some years there are some tiny municipalities in which no marriages at all have taken place or no newborns at all have been registered, the clouds contain fewer points. In specific, **figure 4.2** is formed only by 5 406 points because in the remaining approximately 1% of the total dataset no marriages at all were recorded. Same considerations can be drawn for **figure 4.3**, where in about 7.7% of the total dataset, no marriages at all were recorded, and for **figure 4.4**, where in about 3.6% of the total dataset, no newborns at all were recorded.

Secondly, another cardinality issue to check is the number of points lying on the horizontal axis and vertical axis of a given cloud.

Points lying on the horizontal axis are characterized by the fact of having a quantifier equal to zero associated with a not equal to zero immigrant density γ, signaling municipalities in which in a given year no mixed marriages or mixed newborns have been recorded, even if a percentage γ of immigrant was present. Aerograms in **figure 4.5**, **figure 4.6** and **figure 4.7**. show that these points account for a substantial part of all points contained in each dataset. In particular, it is interesting to highlight a huge difference in terms of red area (quantifier $M_m = 0$) between Emilia Romagna region and Italy (**figure 4.5** and **figure 4.6**), though data are not completely comparable since they are referring to partially different periods.
A particular point to be put under focus is the origin, characterized by having both the immigrant density γ and the quantifier equal to zero. **Figure 4.2** has about 0.2% points with $(\gamma, Q) = (0,0)$ where Q is the given quantifier, whereas these points account for about 1% of the entire clouds referring to the Italian datasets. Eventually, from a theoretical point of view there should not be any point lying on the vertical axis, since it would mean that for an immigrant density γ equal to zero, we would observe a quantifier different from zero,

i.e. we would witness mixed marriages or mixed newborns without having any immigrant recorded in those municipalities at that time. However, in **figure 4.2**, **figure 4.3** and **figure 4.4** it is respectively represented a not null but very small fraction of points lying on the vertical axis, accounting for only about 0.09% (Emilia Romagna dataset) and 0.03% (Italian datasets) of the entire clouds. Albeit these points can be ignored as statistically meaningless, they should not be considered *a priori* errors in transcription of civil registries. For instance, according to Italian law the celebration of a marriage does not require both spouses to be registered *ab initio* in a given municipality: the registration of the non-resident spouse can be completed by 180 days after the ceremony. This event is likely to be highly emphasized in very tiny municipalities with absent immigrant population, like exactly those ones such points belong to.

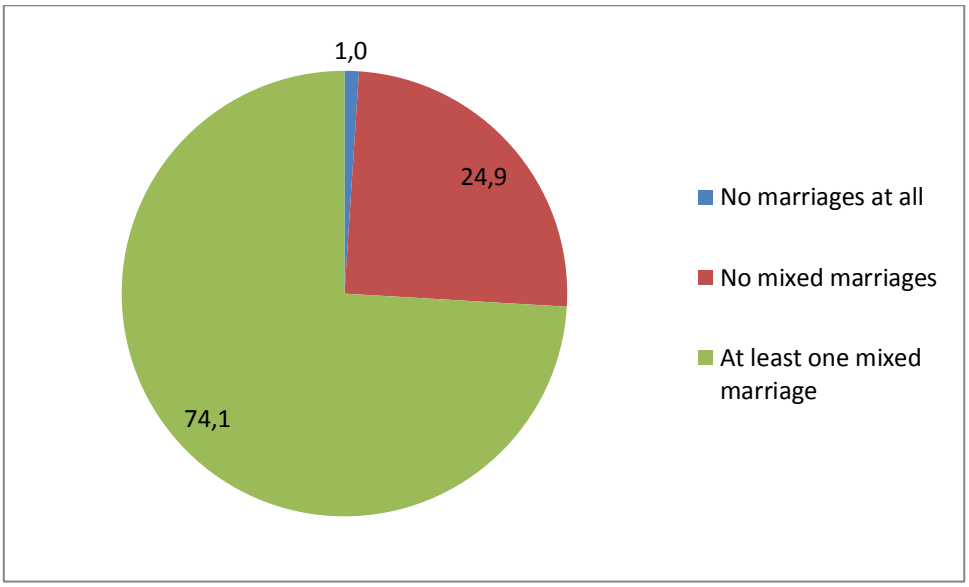


Figure 4.5 *Emilia Romagna dataset*. Percentage distribution of data according to the quantifier M_m cardinality

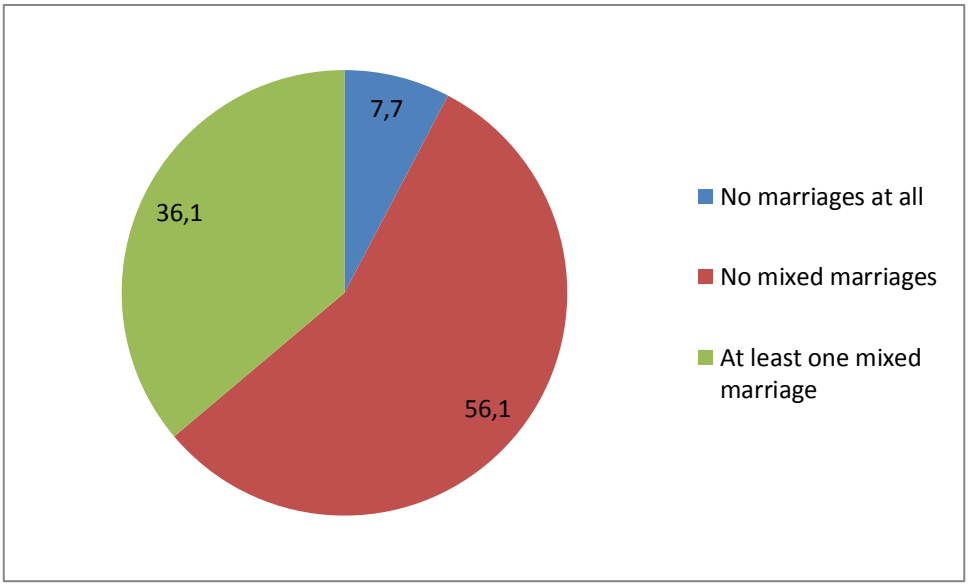


Figure 4.6 *Italian dataset.* Percentage distribution of data according to the quantifier M_m cardinality

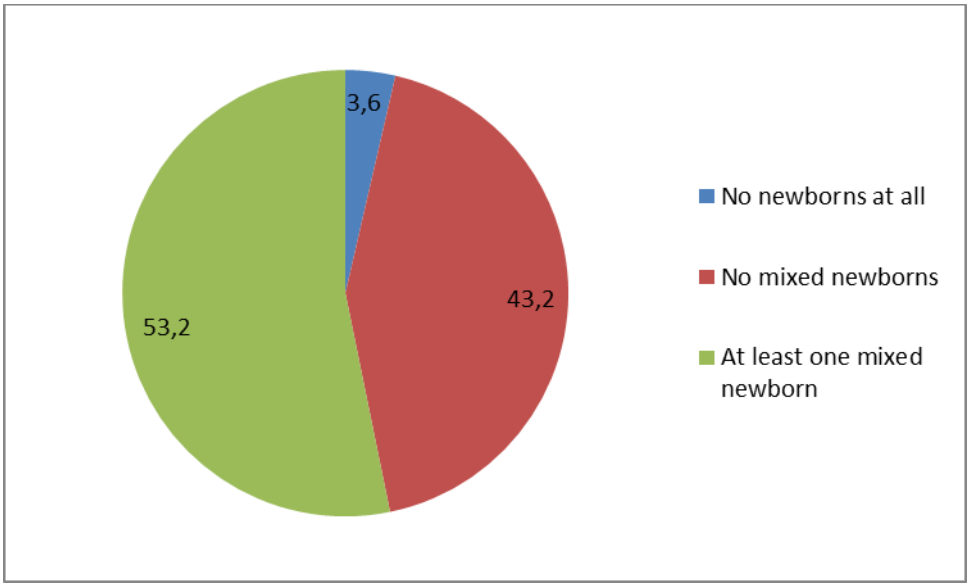


Figure 4.7 *Italian dataset.* Percentage distribution of data according to the quantifier B_m cardinality

Eventually, there is a particular pattern that seems to appear in all the clouds, independently from the considered quantifier. If we take for instance **figure 4.3**, one may note that a certain group of data seem to lie along horizontal lines displaced according to m/n, with $m \leq n$ $(m,n) \in \mathbb{N}$. Lines are clearly visible for $n = 2, 3, 4 \ldots$

The reason why data agglomerate along such horizontal lines is due to what one may call a small-sized municipalities effect. If we make a partition of the Emilia Romagna region dataset and we decide to separate small-sized municipalities from other ones, choosing a threshold at 10 000 inhabitants (a dimensional choice that could sound reasonable in a logarithmic scale since Italian municipalities may generally contain population spanning from 10^2 to 10^6 inhabitants), we obtain **figure 4.8** and **figure 4.9** as new clouds emerging from scatter plotting of data. Undoubtedly, **figure 4.8** – which represents data coming only from small-sized municipalities – is characterized by still having horizontal lines like those ones in **figure 4.3**, whereas such pattern disappears in **figure 4.9**. We claim therefore that this effect is due to events normally occurring in very tiny municipalities, where the total number of marriages or newborns is very small. In fact, being the quantifier a ratio m/n defined in an interval from 0 to 1, it can assume at maximum $n + 1$ values. Thus, if n (i.e. the total number of marriages or newborns) is small, the quantifier may forcedly assume only a small number of values, aggregating therefore data around few horizontal lines. Moreover, being the total number of marriages or newborns very small, it is worth to note that in very tiny municipalities the presence of mixed couples or mixed newborns is greatly emphasized, raising up the quantifiers value (e.g. in a small-sized municipality where only two marriages have been celebrated in one year, one involving a mixed couple, the quantifier M_m is one half; for example this happened in 2010 in Saludecio, a municipality belonging to the province of Rimini).

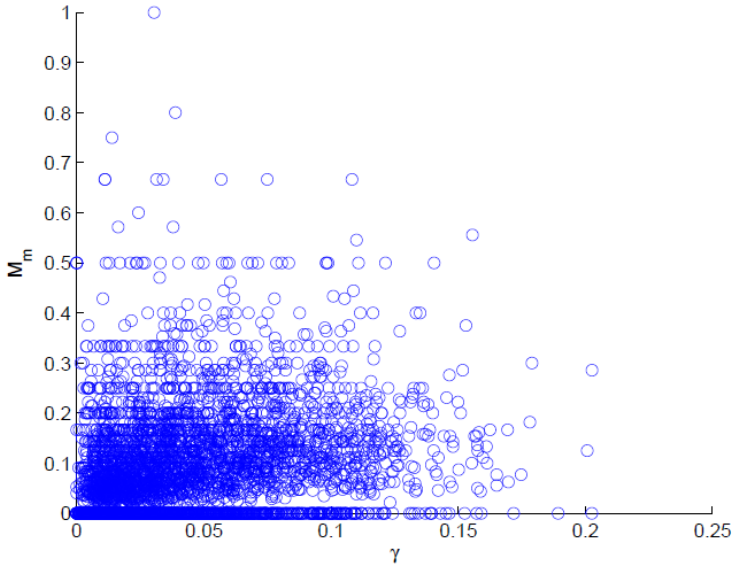


Figure 4.8 *Emilia Romagna dataset.* Raw data versus γ. Blue points represent the fraction of mixed marriages occurred in municipalities with less than 10 000 inhabitants where a percentage γ of immigrants is present.

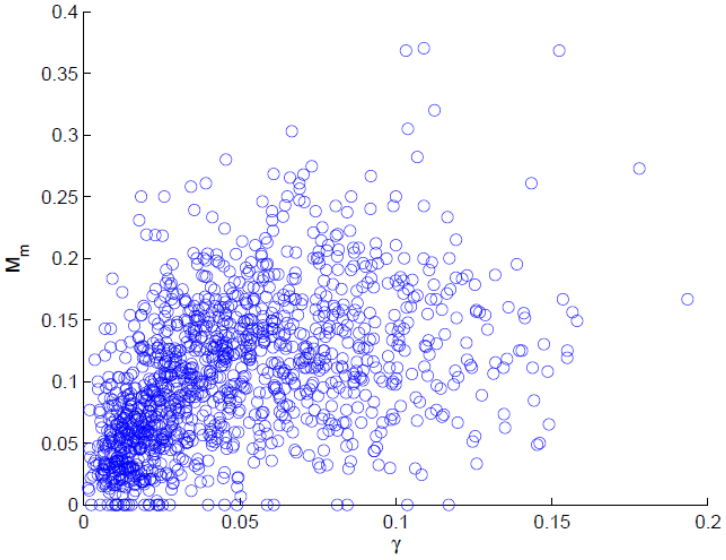


Figure 4.9 *Emilia Romagna dataset.* Raw data versus γ. Blue points represent the fraction of mixed marriages occurred in municipalities with more than 10 000 inhabitants where a percentage γ of immigrants is present.

Same results are found when partitioning the Italian dataset according to the 10 000 inhabitants threshold. **Figure 4.10** shows the four clouds obtained through splitting **figure 4.3** and **figure 4.4**: note that in right panels horizontal lines fade since there data coming from large-sized municipalities are represented.

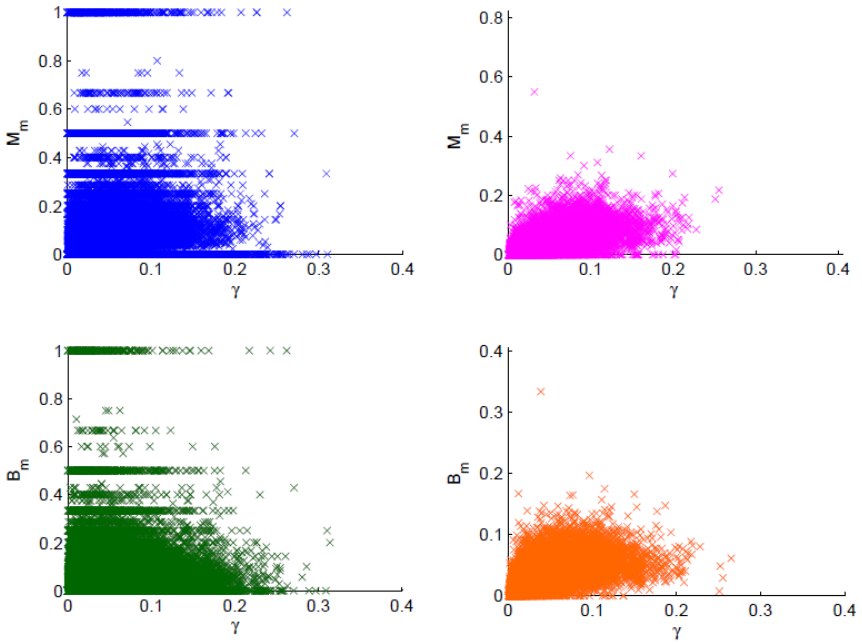


Figure 4.10 *Italian dataset.* Raw data versus γ. Blue points represent the fraction of mixed marriages occurred in municipalities with less than 10 000 inhabitants where a percentage γ of migrants is present. Similarly green points account for newborns from mixed couples registered in small-sized municipalities. Further, pink points represent the fraction of mixed marriages occurred in municipalities with more than 10 000 inhabitants, while orange ones mirror the newborns from mixed couples registered in large-sized municipalities.

4.4 Data density and distribution of immigrants and natives

After the representation of data through scatter-plots, one of the first analyses to be taken into account is a data density check for the explanatory variable γ: this drives us to the definition of an histogram plot of the immigrant density γ. As shown in **figure 4.11** for the Emilia Romagna dataset, on the horizontal axis we represent the immigrant density γ and on the vertical axis its relative frequency. **Figure 4.12** shows the same kind of data density plot for the Italian datasets, partitioned exactly in the same way as in **figure 4.10**.

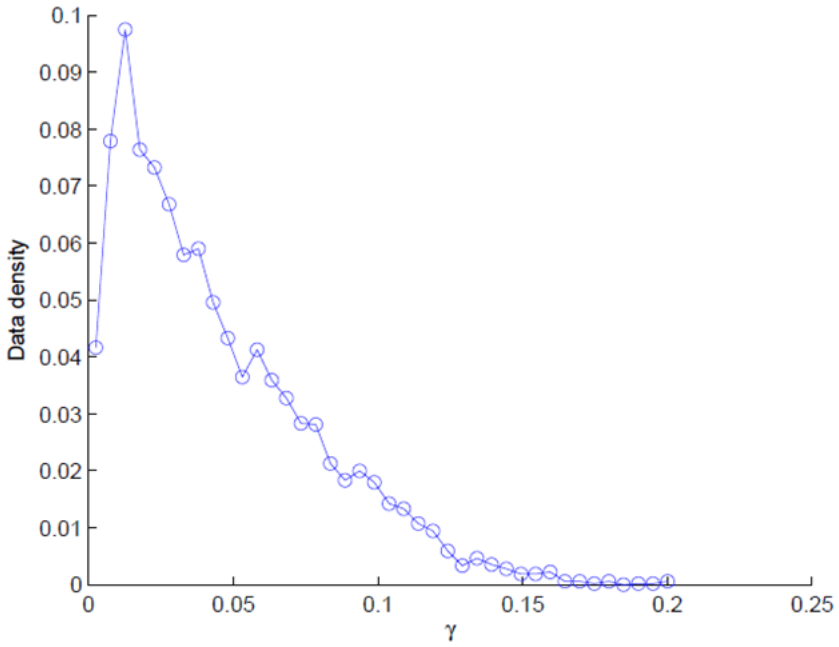


Figure 4.11 *Emilia Romagna dataset.* Relative frequencies of data as a function of γ. Data have been split into 40 bins.

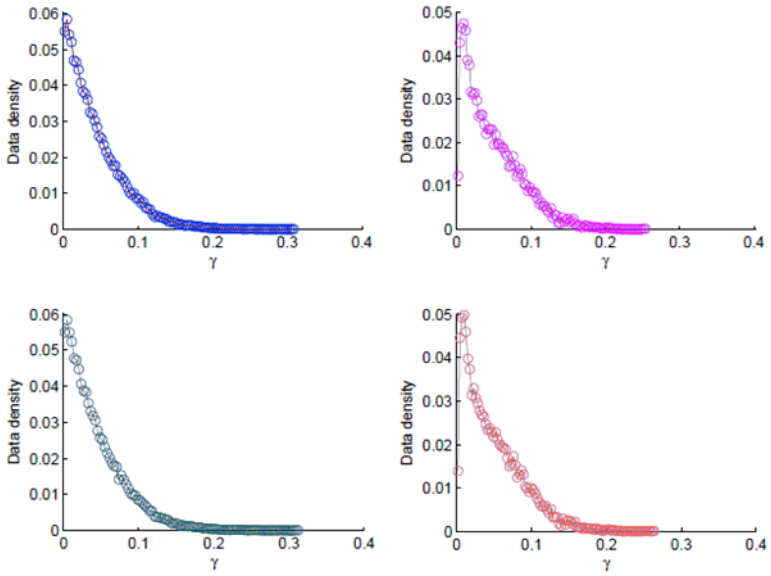


Figure 4.12 *Italian datasets*. Relative frequencies of data as a function of γ. Upper panels: datasets of marriages occurred in small-sized (left) and large-sized (right) municipalities. Lower panels: datasets of newborns occurred in small-sized (left) and large-sized (right) municipalities. In each panel, data have been split into 100 bins.

Some observations can be stated in analyzing both these histograms. First of all, in each figure data density begins to decrease for a small but not null value of γ. The reason for which this occurs traces back to the fact that our observation window began when the migration phenomena were already ongoing (1995 for the Emilia Romagna dataset, 2001 for the Italian datasets) and the immigrant density γ in the whole country or region was already larger than zero.

Secondly, such histograms provide useful information regarding the distribution of data, especially for the tails as pointed by **table 4.2**. This information will be exploited in the next paragraphs.

	0.1	0.05	0.03	0.01
ER	0.0948	0.1108	0.1206	0.1453
M_m^s	0.0952	0.1165	0.1329	0.1657
M_m^b	0.1017	0.1208	0.1347	0.1604
B_m^s	0.0957	0.1173	0.134	0.1671
B_m^b	0.1018	0.121	0.135	0.1608

γ >0.1	γ >0.15
0.0825	0.0075
0.086	0.0168
0.1062	0.0179
0.088	0.0175
0.1062	0.0181

Table 4.2 Complementary cumulative frequencies of data as a function of γ. Rows indicate the referring quantifier (*ER* means Emilia Romagna dataset, followed by Italian partitioned datasets – apexes point to small-sized or large-sized municipalities). In the left table, values of gamma γ for given complementary cumulative frequencies are represented. In the right table, the contrary is shown.

Furthermore, the histograms tails appear to be power law distributed for all datasets. This result has emerged through the evaluation of several model functions, tested using the coefficient of determination R^2 to check the goodness of fit.

More explicitly, given a dataset containing $y_1 \dots y_n$ observed values, each of which has an associated modelled value f_i (computed through a model function), we define the mean of the observed data $f_i = f(y_i)$ as:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4.04}$$

With the mean of observed data, we can define the total sum of squares as:

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{4.05}$$

whereas, the sum of squares of residuals is:

$$SS_{res} = \sum_{i=1}^{n} (y_i - f_i)^2 \tag{4.06}$$

The coefficient of determination R^2 is given by combining the above quantities as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{4.07}$$

A coefficient of determination R^2 close to one signifies that the model function well explains the observed data, whereas when R^2 appears to be far from one, the model function is unfit to represent the observed data.

Within this framework, **table 4.3** shows exponents δ, related R^2 and margins of error E for the law $\mu(\gamma) \propto \gamma^{\delta}$, computed for $\gamma \geq 0.1$ for histograms presented in **figure 4.11** and **4.12**.

	δ	R^2	E
ER	-5.288	0.9638	$\pm$ 0.718
M_m^s	-4.1	0.9877	$\pm$ 0.168
M_m^b	-4.469	0.967	$\pm$ 0.312
B_m^s	-4.108	0.9901	$\pm$ 0.152
B_m^b	-4.542	0.9722	$\pm$ 0.291

Table 4.3 Histograms tails ($\gamma \geq 0.1$). As previously, rows indicate the referring quantifier. In the columns, we find the estimation of power law exponents, related R^2 and margins of error for 95% confidence intervals.

Figure 4.12 poses another question to investigate which links the immigrant density γ and the partition of the Italian datasets, first introduced in **paragraph 4.3**. As already indicated above, this partition was defined to give explanation to the horizontal lines initially appearing in the global scatter-plots. We want to check if the immigrant density γ in any given municipality can be considered almost the same, independently from the size of the municipalities it refers to, or instead it significantly changes. **Figure 4.13** and **4.14** answer such question. The first figure is an histogram of the size S of the municipalities (in terms of total population), plotted in logarithmic scale for the Italian dataset concerning marriages[16]. One may note that the aforementioned threshold at 10 000 inhabitants (i.e. $x = 4$) still belongs to the central part of the distribution of data. Moreover, the size the size S of the municipalities seems to be – at least for the tails –log-normally distributed in agreement with (Berry & Okulicz-Kozaryn, 2012). The second figure instead pictures the immigrant density γ as a functions of the size S of municipalities (always expressed through

[16] As pointed out in **paragraph 4.3**, the two Italian datasets referring to marriages and newborns contain a different number of records and therefore do not carry the same information regarding immigrant density γ. Hereunder we present analyses implemented over the Italian dataset concerning marriages but similar results can be retrieved for the other dataset concerning newborns.

a logarithmic scale): the plot, computed through different binning processes according to a mean measure (for technical details, refer to **paragraph 4.5**), clearly highlight two facts: first, on average the immigrant density γ increases as the Italian municipalities size rise (note the peak in the right part of the graph, corresponding to the municipality of Milan); second and more important, the average values for γ referring to small-sized municipalities ($\gamma = 0.0432$) and large-sized municipalities ($\gamma = 0.0480$) can be considered almost equal.

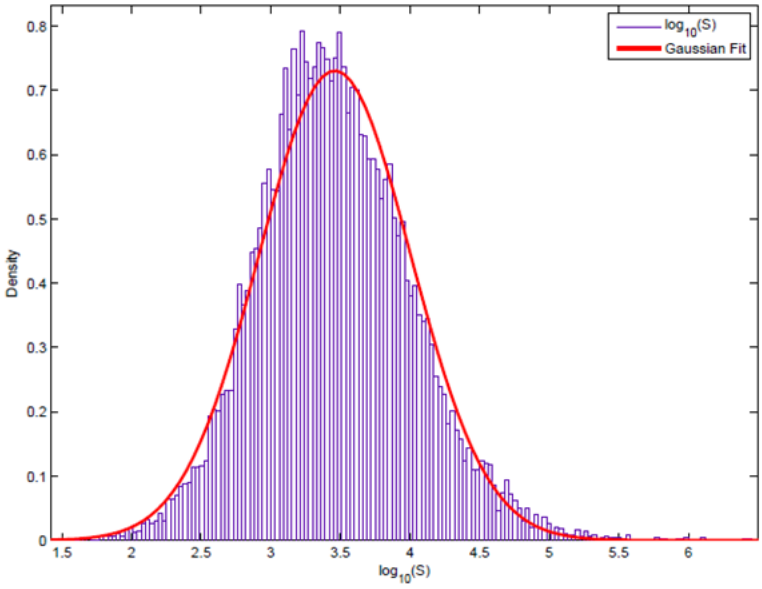


Figure 4.13 *Italian dataset (marriages).* Histogram for S, size (i.e. total population) of municipalities, plotted in logarithmic scale. In red, a Gaussian fit for the distribution.

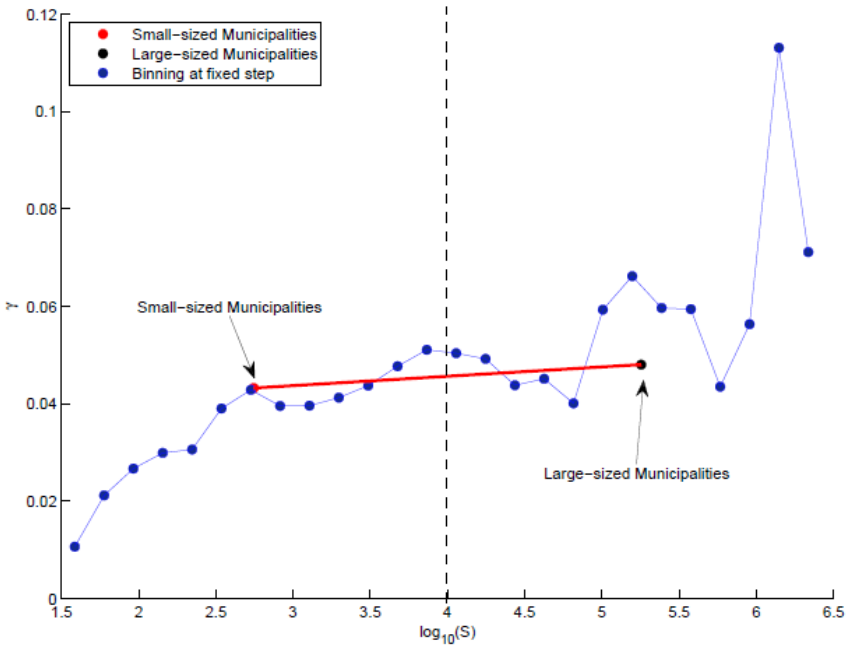


Figure 4.14 *Italian dataset (marriages).* γ as a function of S, size (i.e. total population) of municipalities, plotted in logarithmic scale. Binning through mean measure. First binning (blue dots): 26 bins at constant bin width (step 0.19). Second binning (red and black dots): data are split only into 2 bins, one for small-sized municipalities (red dot) and another (black dot) for large-sized municipalities.

Figure 4.15 confirms the considerations offered above: in this graph, always for the Italian dataset concerning marriages, we find on the horizontal axis the eleven years the dataset refers to and on the vertical axis the immigrant density γ. The black line shows how immigrant density γ changes over time in the large-sized municipalities and the red line illustrates the same variation over time in the small-sized municipalities. At first glance, it is manifest that the values of γ are more or less the same in absolute terms: the maximum difference is 0.0121 between the two curves for the year 2010. Moreover, the two curves highly show linear correlation.

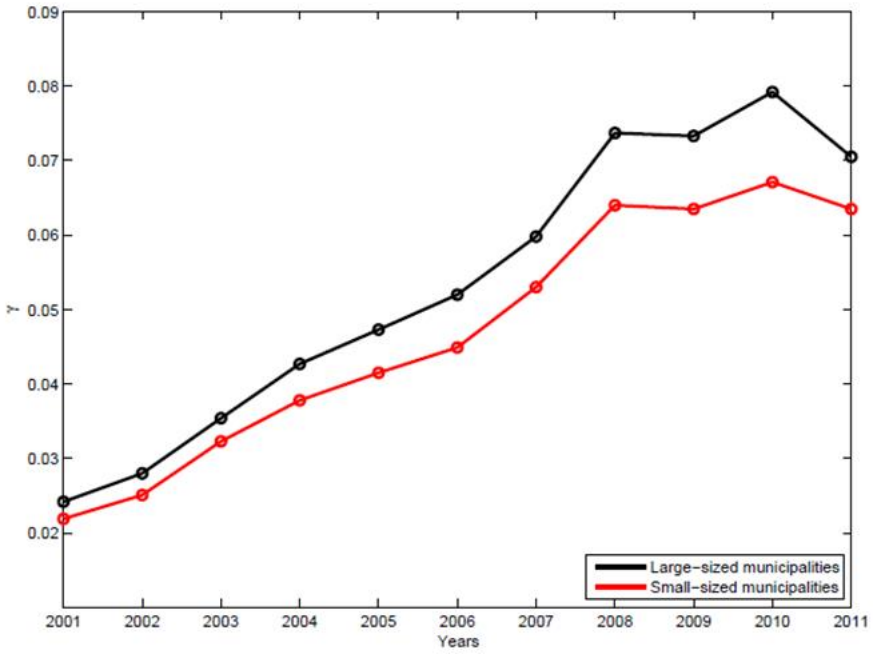


Figure 4.15 *Italian dataset (marriages).* Black line represents the immigrant density γ in large-sized municipalities from 2001 to 2011. Red line mirrors the same quantity in the same period referred to small-sized municipalities.

In fact, given two time-series x and y, we can introduce the correlation coefficient ρ as:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{4.08}$$

Where σ_{xy} is the covariance and σ_x and σ_y are the standard deviations computed over the aforementioned time-series.

Assuming for instance x to be the time-series of immigrant density γ in large-sized municipalities and y to be the time-series of immigrant density γ in small-sized municipalities, we find $\rho = 0.9983$, a result very close to one, signaling that the two time-series show a great level of linear dependence. This result has to be noted because it evidences that immigrant fluxes in Italy during the period 2001-2011 have operated almost in the same way in populating small-sized and large-sized municipalities.

Additionally, the comparison between immigrant density γ in small-sized and large-sized municipalities has led also to the control of the proportion of immigrants π_{imm} in small-sized municipalities against proportion of natives π_{nat} in small-sized municipalities.

We define the proportion of immigrants in small-sized municipalities as:

$$\pi_{imm} = \frac{N_{imm}^{s}}{N_{imm}} \tag{4.09}$$

where N_{imm} is the total number of immigrants and N_{imm}^{s} are only those ones living in small-sized municipalities.

Similarly, we define the proportion of natives in small-sized municipalities as:

$$\pi_{nat} = \frac{N_{nat}^{s}}{N_{nat}} \tag{4.10}$$

where N_{nat} is the total number of natives and N_{nat}^{s} are only those ones living in small-sized municipalities.

In **figure 4.16**, focusing again on the Italian dataset concerning marriages – this time only in small-sized municipalities[17] – we find on the horizontal axis the eleven years the dataset refers to and on the vertical axis the proportion π. The blue line shows how the proportion of natives π_{nat} in small-sized municipalities changes over time and the green line illustrates the same for the proportion of immigrants π_{imm} in small-sized municipalities. At first glance, we can state that both proportions π_{nat} and π_{imm} are close to the value $\pi = 0.3$, meaning that there are no substantial differences in the way immigrants and natives distribute themselves between small-sized and large-sized municipalities. Moreover, the two curves show a high level of linear dependence: the variation of proportion π_{nat} against the variation of proportion π_{imm} presents a coefficient of correlation $\rho = 0.9190$, which becomes even higher ($\rho = 0.9713$) if we exclude the last year 2011 in the time-series.

[17] Again, similar results can be retrieved for the Italian dataset concerning newborns in small municipalities.

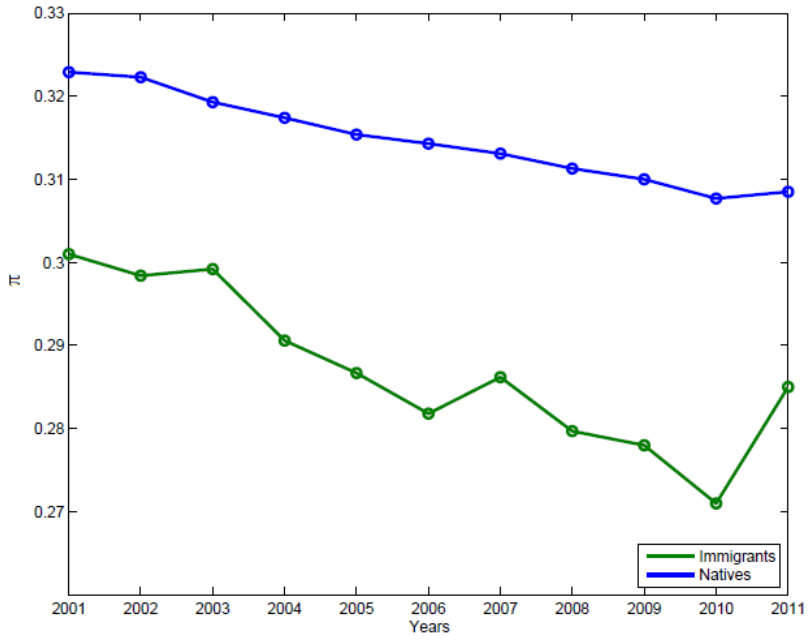


Figure 4.16 *Italian dataset (marriages in small-sized municipalities).* Blue line represents the proportion π_{nat} of natives living in small-sized municipalities from 2001 to 2011. Green line shows the proportion π_{imm} of immigrants living in small-sized municipalities in the same period.

In other words, this datum is not fit to discriminate between natives and immigrants and between small-sized and large-sized municipalities.

4.5 Binning

Besides the representation of data through scatter plots, our main interest has been the analysis of the quantifiers' averages as functions of Γ. Since all quantifiers are ratios, there are two possible ways to compute the averages: in fact, for a given bin of Γ (i.e. an interval of Γ containing a certain amount of points), one may compute the usual mean of the ratios or the global mediant[18] of the ratios.

[18] The name mediant is due to the fact that it strictly lies between the smallest and largest fraction of which it is the mediant.

More explicitly, we consider a set $C = \{(\Gamma_1, Q_1), \dots, (\Gamma_n, Q_n)\}$ containing all n points of a given cloud. Each i-point (Γ_i, Q_i) has coordinates belonging to the non-negative rational numbers set $\mathbb{Q}_0^+$, so that we can write $(\Gamma_i, Q_i) = \left(\frac{s_i}{t_i}, \frac{u_i}{v_i}\right)$ with s_i, t_i, u_i and v_i belonging to $\mathbb{N}$ (t_i and $v_i \neq 0$).

Let $\mathcal{B}$ be a partition of C so that a generic subset $B_j = \left\{\left(\Gamma_1^j, Q_1^j\right), \dots, \left(\Gamma_{k_j}^j, Q_{k_j}^j\right)\right\}$, containing k_j points (with $k_j \leq n$), is one of $\mathcal{B}$'s elements.

We define now two different types of binning processes made according a mean measure or according to a mediant measure.

A binning process computed according to a mean measure reduces all k_j points of each subset B_j to one singular point $\overline{b_j}$ so that

$$\overline{b_J} = \left(\frac{1}{k_j}\sum_{i=1}^{k_j}\Gamma_i^j, \frac{1}{k_j}\sum_{i=1}^{k_j}Q_i^j\right) \tag{4.11}$$

Instead, exploiting the fact that both Γ_i^j and Q_i^j are non-negative rational numbers, a binning process computed according to a mediant measure reduces all k_j points of each subset B_j to one singular point b_j^* where

$$b_j^* = \left(\frac{\sum_{i=1}^{k_j}\mathrm{s}_i^j}{\sum_{i=1}^{k_j}\mathrm{t}_i^j}, \frac{\sum_{i=1}^{k_j}\mathrm{u}_i^j}{\sum_{i=1}^{k_j}\mathrm{v}_i^j}\right) \tag{4.12}$$

Furthermore, defined mean and mediant as possible tools to compute the averages, it is crucial to identify a criterion to split data into a certain number of bins of Γ. If k_j is constant $\forall B_j \in \mathcal{B}$, we have a *constant information binning* process, whereas if $\left|\Gamma_{k_j}^j - \Gamma_1^j\right|$ is constant $\forall B_j \in \mathcal{B}$, we have a *constant bin width binning* process. Assuming a non-uniform distribution of data, it is manifest that the two binning process are not equivalent, since the first criterion makes vary the width of the bins, whereas the second criterion makes vary the number of points belonging to each bin of Γ.

Theoretically, given two options both for the way to compute the averages (mean or mediant), both for the way of splitting data into bins of Γ (constant information or constant bin width), there are thus four possibilities of conducting a binning and average process. The choice which has resulted to the be the optimal one has been a constant

information binning process with averages computed through a mediant measure. The reasons beneath this choice will be given in the next paragraph.

4.6 Evaluation of patterns for quantifiers' averages

After having compacted the initial information of raw data into bins, a curve fitting (based on linear least squares methods) has taken place. In particular, following (Barra, Contucci, Sandell & Vernia, 2013) whose model has been outlined in **chapter 3**, we have evaluated linear and square root functions with two degrees of freedom (i.e. $f(\Gamma) = a\Gamma + b$ and $f(\Gamma) = a\sqrt{\Gamma} + b$) according to the coefficient of determination R^2, as introduced in **paragraph 4.4**.

As previously, the first dataset to be put under analysis has been that one containing information from the Emilia Romagna region. All data coming from this dataset have fitted a square root function (**figure 4.17**) with a coefficient of determination R^2 being over 95%, highly reproducing the result obtained in (Barra, Contucci, Sandell & Vernia, 2013). Among the possibilities in binning processing, the best one appeared to be constant information binning matched with a mediant measure, since the other choices resulted to be noisy, i.e. not stable according to the binning procedure. More explicitly, we found that 18 bins (each one containing about 300 points) were optimizing the Emilia Romagna dataset, that is this number of bins was associated with the highest R^2 against $f(\Gamma) = a\sqrt{\Gamma} + b$, whereas choosing a different number of bins (always in a neighborhood of 18 bins) was associated by lower but comparable values of R^2.

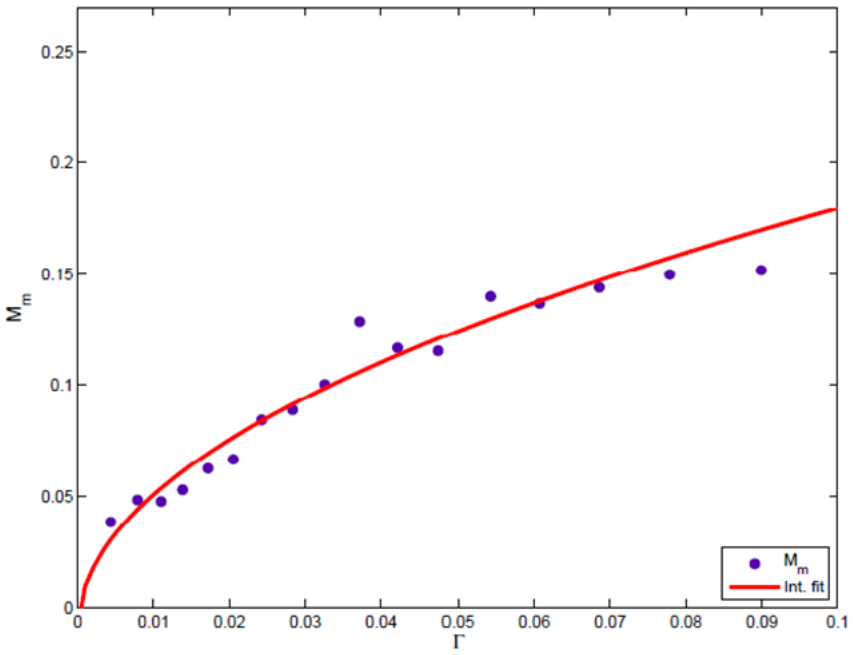


Figure 4.17 *Emilia Romagna dataset*. Dots are average quantities versus Γ. Quantifier M_m (blue dots), fraction of mixed marriages occurred from 1995 to 2010 in all the municipalities located in Emilia Romagna region, with the best square root fit (red curve) $a\sqrt{\Gamma} + b$ (a = 0.5943 ± 0.0757, b= -0.008631 ± 0.014019, goodness of fit $R^2 = 0.9529$ computed for $\Gamma < 0.078$).

Looking at **figure 4.17**, it is worth to note that the evaluation of parameter b is compatible with the hypothesis that it can be null, as prescribed by the statistical mechanics model presented in **chapter 3** that we will use for results interpretation in a next paragraph. Other information regarding the analyses conducted over the Emilia Romagna region dataset can be found in (De Pretis & Vernia, 2013), which is reported at the end of this work as **Appendix 2**.

Secondly, data coming from the ISTAT Italian datasets – if taken as whole non-partitioned datasets – have always reported to be noisy, not resulting stable according to any binning procedure. This instability – probably due to the presence of data coming from very small-sized municipalities, sometimes embodying less than one hundred inhabitants – has led to the idea of analyzing the datasets according to the partition based on the municipality population, as previously presented in **paragraph 4.3**. The choice of 10 000 inhabitants as threshold to divide small-sized municipalities from large-sized municipalities has been made

also along with (Barra, Contucci, Sandell & Vernia, 2013) - where analyzed data were concerning municipalities whose population only exceeded 10 000 inhabitants – and proved to be right for the Italian datasets, since the employed binning procedures resulted to be sufficiently stable if applied in agreement with such threshold. In particular, the best results have been again obtained for constant information binning matched with a mediant measure. Afterwards, according to this partition of the datasets, highly recognizable patterns have emerged as shown in **Figure 4.18**.

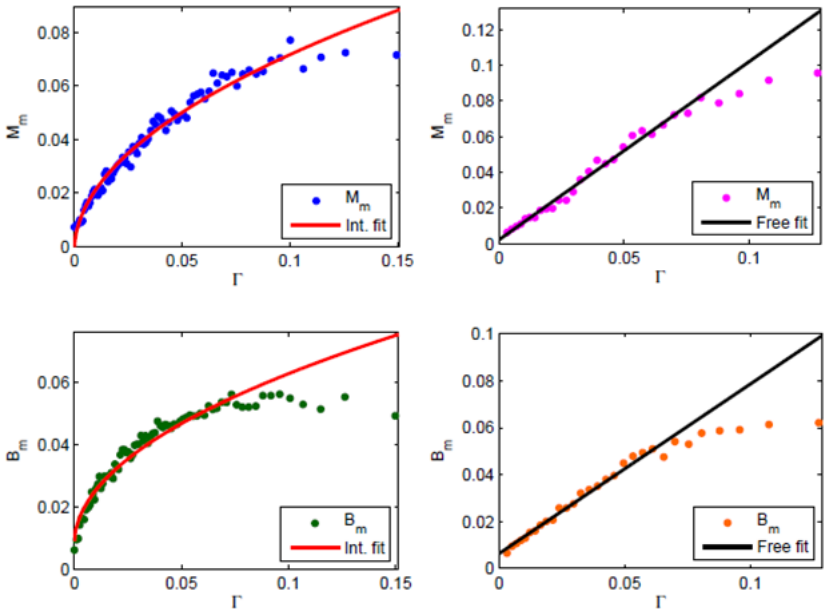


Figure 4.18 *Italian datasets*. Dots are average quantities versus Γ. Left upper panel: quantifier M_m^s (blue dots), fraction of mixed marriages occurred in municipalities with less than 10 000 inhabitants, fitted by $a\sqrt{\Gamma} + b$ (red curve). Right upper panel: quantifier M_m^b (pink dots), fraction of mixed marriages occurred in municipalities with more than 10 000 inhabitants, fitted by $a\Gamma + b$ (black curve). Similar analyses conducted in lower panels for quantifier B_m both for small-sized and large-sized municipalities.

In **table 4.4**, we report an array containing the estimation of the parameters a and b (and relative margins of error E_a and E_b for 95% confidence intervals), together with the coefficient of determination R^2, used for the fitting of average data against $f(\Gamma) = a\sqrt{\Gamma} + b$ (M_m^s and B_m^s, i.e. quantifiers referred to small-sized municipalities) and against

$f(\Gamma) = a\Gamma + b$ (M_m^b and B_m^b, i.e. quantifiers referred to large-sized municipalities). The last column of the table highlights the Γ interval related to the fitting procedure.

	a	E_a	b	E_b	R^2	Bounds
M_m^s	0.2329	± 0.0086	-0.002043	± 0.00168815	0.9746	Γ<0.126
M_m^b	1.002	± 0.046	0.001994	± 0.002	0.9879	Γ<0.081
B_m^s	0.1735	± 0.00815	0.007938	± 0.0015215	0.9614	Γ<0.096
B_m^b	0.7764	± 0.04485	0.005253	± 0.001424	0.9908	Γ<0.07

Table 4.4 *Italian datasets*. Estimation of parameters used for fitting of average data

Aside from direct visual effect seen in **figure 4.18**, the quantitative measures described in **table 4.4**, showing values of coefficient of determination R^2 always over 96%, prove that in small-sized municipalities a clear square root pattern emerges whereas a linear one appears for large-sized municipalities. These empirical laws have been verified against different types of fitting models with various degrees of freedom and still hold true. Once more, as for the Emilia Romagna dataset, the coefficient b can be always considered almost null.

Furthermore, it is important to highlight that the emergence of these two different empirical laws – according to different municipality's population size – appears only in the Italian datasets because, even partitioning the Emilia Romagna dataset in relation to the same modalities (i.e. putting a division threshold at 10 000 inhabitants[19]) we always find the same square root patterns as when fitting the whole dataset, instead of – like for the Italian data – a linear pattern for the subset of data belonging to large-sized municipalities. Such difference with respect to the Emilia Romagna dataset has led to a cross-check of the results obtained for the Italian datasets, that is evaluating a square root fit for large-sized municipalities and evaluating a linear fit for small-sized municipalities. The analysis has confirmed that the previously identified patterns (as visualized in **figure 4.18**) always display a higher coefficient of determination R^2 with respect to the other checked patterns. The question whether these identified patterns are valid even on a time-scale has been investigated in the following paragraph.

[19] With such threshold, we find that about 76 % of data of the Emilia Romagna dataset belong to municipalities whose population is under 10 000 inhabitants, whereas this number rises to about 85% for the Italian datasets.

4.7 Time-dependent analysis

Hitherto, we have conducted analyses over Italian data which have made emerge the two empirical laws described in the previous paragraph, only considering data as a unique time-independent ensemble.

However, one may conjecture whether these empirical laws emerge also considering the immigrant integration phenomena on a time-scale, questioning also when this may eventually occur. For this reason, the same type of analysis processed in the previous paragraphs (i.e. binning of data and fitting against curves $f(\Gamma) = a\Gamma + b$ and $f(\Gamma) = a\sqrt{\Gamma} + b$) has been applied to eleven different data ensembles for each partitioned Italian dataset. These data ensembles have been built in a cumulative manner: the first data ensemble contained only data recorded in 2001 (the first year of observation for the Italian datasets), the second data ensemble contained only data recorded in 2001 *and* 2002, and so on until falling again into the whole datasets previously analyzed in **paragraph 4.6**. The results of this analysis are shown below in **figure 4.19**. In each panel, the rightmost dot on the highest curve mirrors the R^2 value presented in **table 4.4**.

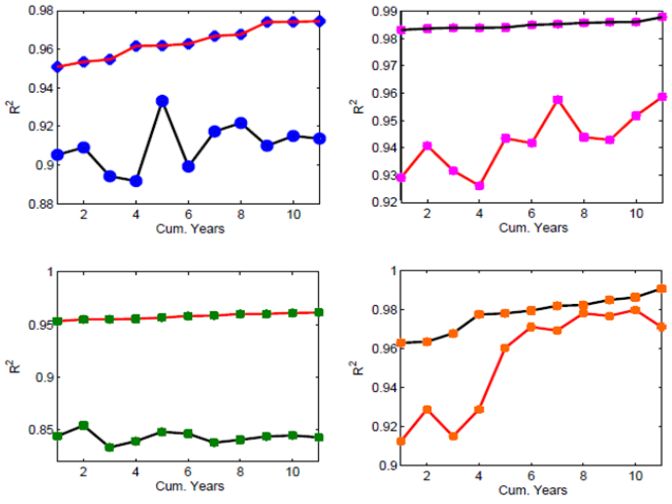


Figure 4.19 *Italian datasets.* R^2 measurements for goodness of fit against an increasing years span. R^2 pictured through a red curve (when referring to a *square root fit*) and through a black curve (*linear fit*). Left upper panel: R^2 (blue dots) related to mixed marriages occurred in municipalities with less than 10 000 inhabitants. Right upper panel: R^2 (pink dots) related to mixed marriages occurred in municipalities with more than 10 000 inhabitants. Left lower panel: R^2 (green dots) related to newborns from mixed couples occurred in municipalities with less than 10 000 inhabitants. Right lower panel: R^2 (orange dots) related to newborns from mixed couples occurred in municipalities with more than 10 000 inhabitants.

The main result of this time-dependent analysis gives evidence that the two previous identified empirical laws have also a predictive content, since they own a goodness of fit (always measured by R^2) in fitting data since the very first year of phenomena observation. For instance, focusing on **figure 4.19** left lower panel, we see that already considering only mixed newborns registered in 2001 in small-sized municipalities, the square root fit (red curve) is characterized by having a coefficient of determination higher than 95%, whereas the same data fit against a linear curve (black curve) with a quite lower R^2 under 85%. Moreover, the separation of the two curves continues almost unchanged while adding data from other years with a slightly increasing value of R^2 for the square root fit and a not steady pattern for the linear fit. This pattern is also confirmed if we analyze mixed marriages in small-sized municipalities and – with opposite results – if we look at right panels, that is the phenomena occurring in large-sized municipalities. The only panel which may cause doubts is the right lower one, i.e. that one comprising mixed newborns data in large-sized municipalities. Here, although the linear pattern always dominates the square root one, the distance between the curves is at the beginning significant but then becomes negligible as cumulating data year by year until falling again in the whole dataset. In this case, we have evaluated that the linear fit provided anyway the best approximation of data with respect to the square root one: in fact, the linear fit was undoubtedly better capturing the data located nearby the origin (i.e. characterized by having a small value of Γ), whose importance in terms of critical value of γ has been highlighted at the end of **chapter 3**.

4.8 Theoretical interpretation of emerged patterns

The patterns emerged through the analyses processed in **paragraph 4.6 and 4.7** can be now interpreted according to the statistical mechanics model previously introduced in **chapter 3**.

For the Emilia Romagna dataset, the emergence of a square root pattern signals that imitative phenomena may occur as predominant behaviors within the population of immigrants and natives, meaning that social actions based on peer-to-peer interactions play a fundamental role. More explicitly, our work therefore claims that – crudely talking on average terms – in Emilia Romagna municipalities the establishment of new mixed marriages seems to be severely influenced by the imitation or learning of experiences belonging to other people included in the population. Seen in network terms, we could observe a population of connected individuals which communicate and share information

with each other. This is exactly, what other authors have detected, analyzing similar data regarding mixed marriages in Spanish municipalities in (Barra, Contucci, Sandell & Vernia, 2013).

The Italian datasets analysis opens instead a different and more complex perspective over the way immigrant integration phenomena occur in the whole country. Always taking an average point of view over the integration actions we considered, that is mixed marriage and newborns, it is manifest that the same kind of patterns and therefore behaviors we identified in Emilia Romagna municipalities are found only in small-sized Italian municipalities, where the population embodies less than 10 000 inhabitants. There, we can state that still imitative behaviors centered on a deep interaction of individuals are embodied in social actions for integration. On the other side, looking at large-sized Italian municipalities, the patterns detected are no more proportional to a square root function but exhibit a clear linear shape. According to the same theoretical model we have just exploited for the previous interpretation of data, a linear fit of data may signal that individual choices, rather than imitative behaviors, seem to be the propellant that drives the institution of new mixed marriages or the occurrence of the new mixed newborns. In other words, even if in large-sized municipalities the total population is greater, the immigrant density γ is almost the same with respect to small-sized municipalities (see **figure 4.15**) and the proportion of immigrants living there is comparable with that one of natives (see **figure 4.16**), our data analysis suggests that here we are observing social phenomena that seem to arise from a different network of individuals, where connections and interactions look like to be not so prevailing and grounded as in the small-sized municipalities case, paving the way to the emergence of individuality rather than social acting.

Considering both cases, we could therefore claim that the social phenomena we are witnessing look like to be not scale-invariant, seeming to be more influenced by the number N of individuals we consider, when aggregating raw data according to their geographical origin. This result turns out to be even more impressive if compared with subtle considerations made by classical sociology authors, who understood a crisp difference in the emergence of social actions occurring in small-sized or large-sized municipalities already more than one hundred years ago. Among others, French sociologist Émile Durkheim was the first to realize in his milestone work over suicides (Durkheim, 1897) that certain types of social actions occur in a not uniform manner but heavily depend by social laws and size of the municipality where they are recorded. In particular, through an accurate classical statistical analysis on registered suicides in France during the second industrial revolution

period, Durkheim suggested that a high rate of suicides could be due to anomie, "a condition in which society provides little moral guidance to individuals" (Gerber & Macionis, 2010). In Durkheim's thought, anomie can be regarded as a breakdown of social bonds between an individual and the community: this condition was sensibly higher in large industrial cities and was at that time severely affecting individuals migrating from small villages in search of labor.

Although our work stems from a hard-science background, providing quantitative evidences to immigration phenomena, we could not hide that our results resound in Durkheim's insights. Our hope is that this study may be source of debate and discussion not only for the mathematical physics community but also for modern sociologists involved in immigration phenomena research.

4.9 Future perspectives

In the next future, due to the richness of the ISTAT Italian dataset, several studies will be conducted to deepen the phenomena entailed in such amass of data. The possibility of accessing to data belonging to singular municipalities per year assures the reconstruction of several geographical macro-structures like provinces or regions, allowing the implementation of new multi-scale comparisons that could go further the analysis presented so far. In specific, one of the first step could be the control of a geographical dependence of the patterns which have been tracked by this work, in order to observe if Italian municipalities belonging to different geographical areas behave in a different way respect to what already nationwide highlighted in terms of immigrant integration. Eventually, a more theoretical work shall address the questions concerning the emergence of different patterns in comparable-sized municipalities clusters.

Appendix 1

Automation of Web contents wrapping techniques through a Python-based algorithm

As outlined in **chapter 4**, Self-Service Statistics web-resources[20] of the Emilia Romagna Region Statistics Office represent a valuable source for many information concerning social phenomena occurring in the region: focusing on integration quantifiers, the website releases sensible data regarding marriages occurred in all Emilia Romagna municipalities in a period lasting from 1990 to 2010. In specific, data are accessible per every municipality (identified by its own name and its related ISTAT code) at a given year and consist in the number of registered marriages, separated per citizenship of spouses (only since 1995) according to two categories: Italians (including Sammarinese citizens) and Foreigners. Since the particular way of accessing to data (per municipality at a given year), collecting the entire dataset is not a trivial problem. For this reason, automated techniques of web-contents wrapping have been employed to solve this task. For instance, taking as task the collection of marriages data for the municipality of Bologna in the year 2010, we show the technical difficulties linked to collecting the entire dataset.

Figure A1.01 represents the first screenshot the user interfaces: here, two queries have to be answered to access the second screenshot, namely the selection of the year data will refer to and the type of elaboration required (table, histogram or aerogram). Since we are

[20] http://statistica.regione.emilia-romagna.it/servizi-online/statistica-self-service [Accessed 2013, December 15]

interested in a numerical collection of data, table (i.e. a matrix) is our default option.

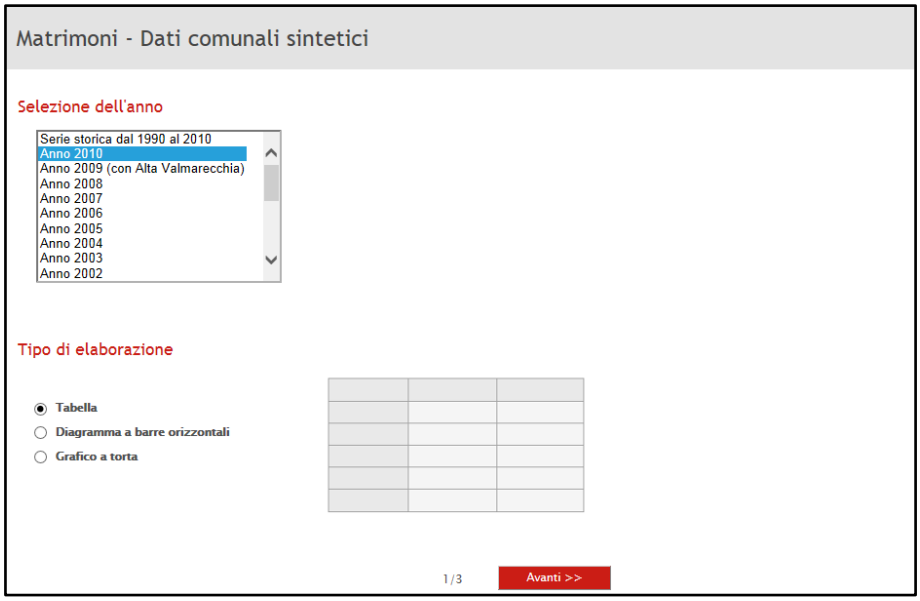


Figure A1.01. Data collection. Year and elaboration-type queries

Figure A1.02 represents the second screenshot the user interfaces: here again, two queries have to be answered to access the third and final screenshot, namely the selection of the geographical area data will refer to (the entire region, one or more provinces, one or more municipalities) and the two type of parameters defining the table (those ones defining the rows and columns entries of the matrix, which can be chosen among citizenship of male and female spouse, type of celebration (civil or religious), month of celebrations, birthplace of male and female spouse, etc. etc.). Here, we select the municipality of Bologna as geographical area of interest and citizenship of male spouse for rows entries and citizenship of female spouse for columns entries. An important option (which will play a fundamental role in the storing process of data) is the total sums for rows and columns entries which is selected by default.

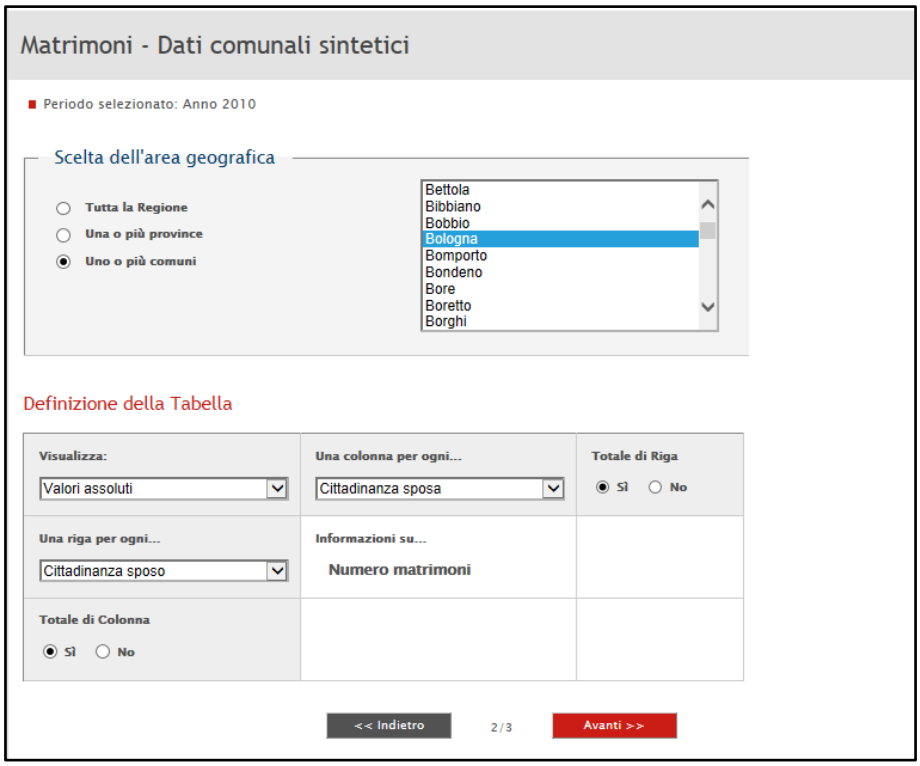


Figure A1.02. Data collection. Year and elaboration-type queries

Eventually, **figure A1.03** represents the final screenshot the user observes: it contains the data required by previous queries, possibly exportable as PDF or spreadsheet file.

Figure A1.03. Data collection. Final screenshot containing marriages data

Theoretically, randomly posing a query referring to data belonging to a certain municipality in a given year, the user can observe five different types of final screenshots, which can be mathematically defined as an empty matrix, a 2x2 matrix, a 3x2 matrix, a 2x3 matrix and a 3x3 matrix (as shown in **figure A1.03**). These matrices correspond to all the possible combinatorial matches, given male and female individuals belonging to two different sets, i.e. Italians and Foreigners.

All the operations previously shown are easily reproducible to download data for any municipality at any given year: however, the main problem which arises is connected with the time spent in each single operation. A hand-operated download of data presented in the previous example, simulated on a Sony Vaio laptop powered by an Intel Pentium processor Dual-Core (B950 @ 2.10 GHz, 2.10 GHz) through Internet Explorer (version 11.0.2) via a 7 Mbps DSL connection costs around 65 seconds (35 seconds to access data, plus other 30 seconds to process and save data into a spreadsheet). This means that the operation of downloading the entire dataset is highly time-consuming, requiring at least 100 hours to collect information from 5 568 final screenshots (348 municipalities for 16 years). Such time dimension already proves the necessity of automatizing the previous assignments also in order to avoid hand-made errors, which could be easily committed in achieving a similar task.

In this context, automatizing data downloading generates two different kind of problems to solve: the first one is the automation of browser actions (i.e. surfing web-pages, selections on dropdown-list and dropdown-menu, plus clicks on internal navigation buttons), while the second one is the web-contents wrapping of required information. All these actions require an algorithm, containing the necessary information to solve the tasks for one single operation (i.e. downloading data for one municipality in a given year), to be subsequently repeated for all the 5 568 final screenshots. Due to its usage and versatility in scripting and Internet-based computing, the programming language chosen for the algorithm has been Python (version 2.7.3 at 32bit) with the following modules for browser automation (Selenium version 2.0) and web-contents wrapping (Beautiful Soup version 3.2.0).

As it will be presented in the code below, the algorithm design has been shaped on an unique reiterated connection instead of a series of multiple connections. The decision has been taken to avoid the banning (for a given quantity of time) of the IP address instructing the connection: as many other servers, also the server hosting the Emilia Romagna region web-site prevents DoS (Denial of Service) attacks, disabling connections with IP addresses

accessing an anomalous number of times to hosted web-pages. Computer experiments have shown instead that an unique reiterated connection is not feasible any ban from the server, whereas a multiple connections design could result in a ban lasting hours.

We present here the two scripts coded in Python which build up the algorithm outlined above.
The first script is **downloading.py** which initially opens a connection with the Emilia Romagna web-site and then starts an automatic surfing within the three screenshots described above, downloading all marriages data for every municipality at a given initial year (in this example, the year 2010). Each final screenshot has its HTML code first stored in a file named **final_screenshot.txt**, which is subsequently put under web-wrapping process to extract the matrix associated to it. Year, name of the municipality and numbers belonging to such matrix are then stored in another file named **er_data_encrypted.txt**. Important to note, all final screenshots are produced by dynamic HTML web-pages, meaning that their contents have been generated by a web-application, in this case under the particular requests of a user (i.e. the data specifics). In this case, inner HTML (the code of the web-page) and outer HTML (the code of the screenshot) do not match. This discrepancy has required the implementation of a simple Java script, to read data directly from outer HTML. Eventually, the script has been programmed to await a certain quantity of seconds, each time it loads new web-pages (the awaiting time has been optimized according to previous computer experiments) and to automatically reboot in case of crash or denial of connection.

downloading.py

```
***
import time
import math
import codecs
import urllib
import operator
from selenium import webdriver
from selenium.common.exceptions import TimeoutException
from BeautifulSoup import BeautifulSoup, SoupStrainer

start_time = time.time()

year="Anno 2010"
```

```
er_cities=['Agazzano', 'Albareto',…, 'Zibello', 'Zocca', 'Zola Predosa']
j=0

for k in range(1,10000):
   try:
          driver = webdriver.Ie()
          driver.get("http://sasweb.regione.emiliaromagna.it/statistica
          /SceltaElaborazione.do?analisi=matrimonisint ")
          inputElement = driver.find_element_by_id("annoSS")
          inputElement.send_keys(year)
          driver.find_element_by_id("avanti").send_keys("\n")
          time.sleep(20)
          for n in range(1,10000):
              database=[]
              city=er_cities[j]
              database.append(year)
              database.append(city)
              driver.find_element_by_id("postFilter1").send_keys("root.Descrizione
              Comune")
              driver.find_element_by_name("root.Descrizione
              Comune").send_keys(city)
              if n==1:
                  driver.find_element_by_name("dimension1").send_keys("Cittadinanza
                  sposo")
                  driver.find_element_by_name("dimension2").send_keys("Cittadinanza
                  sposa")
              driver.find_element_by_id("avanti").send_keys("\n")
              time.sleep(15)
              data=driver.execute_script("data_java=document.getElementsByTagName
              ('html')[0].outerHTML; return data_java;");
              f=codecs.open('C:\Python27\final_screenshot.txt', 'w', 'utf-8')
              f.write(data)
              f.close()
              response = urllib.urlopen("file:///C:/Python27/final_screenshot.txt")
              for tag in BeautifulSoup(response,parseOnlyThese=SoupStrainer('td')):
```

[21] The variable **er_cities** contains the names of all 348 municipalities belonging to Emilia Romagna region. For spatial convenience, we omit to report this huge list.

```
                try:
                    control=unicode((tag.text))
                    database.append(control)

                except:
                    continue
            data2write=''
            for i in range(0,len(database)):
                data2write=data2write+str(database[i])+' '
            data2write=data2write[0:-1]+'\n'
            f=open('C:\Python27\er_data_encrypted.txt', 'r')
            old_data=f.read()
            f.close()
            f=open('C:\Python27\er_data_encrypted.txt', 'w+')
            f.writelines(old_data)
            f.writelines(data2write)
            f.close()
            print time.time() - start_time, "seconds"
            print data2write
            j=j+1
            driver.find_element_by_id("indietro").send_keys("\n")
            time.sleep(15)
    except:
        driver.quit()
        time.sleep(120)
```

downloading.py script alone is not able to complete all the algorithm tasks. In effect, related to the initial example of downloading the number of marriages occurred in the municipality of Bologna in the year 2010, the first script writes the following line in the **er_data_encrypted.txt** file:

```
Anno 2010 Bologna 810 97 907 40 44 84 850 141 991
```

whereas, the output we would like to expect is:

```
Anno 2010 Bologna 810 97 40 44
```

where the last four numbers read as the number of marriages occurred only between Italians, the number of marriages where the male spouse was Italian and the female spouse was Foreigner, the number of marriages where the male spouse was Foreigner and the female spouse was Italian, the number of marriages only between Foreigners.

However, comparing this line with the previous one, it is manifest that the task of decryption is quite trivial, given the total sums inserted in the sequence of numbers. For the decryption, a second script named **deciphering.py** has been therefore set: here, data loaded from **er_data_encrypted.txt** file are processed and then saved in a new file named **er_data_deciphered.txt**, which can be then easily imported in any spreadsheet program since all sensible data are space-separated. The choice of not merging this script with the previous one is motivated by the fact that such separation avoids problems linked to the integration in a unique algorithm of both online and offline operations (i.e. **deciphering.py** does not require any Internet connection to be executed, whereas **downloading.py** does).

deciphering.py

```
***
f=open('C:\Python27\er_data_encrypted.txt','r')
data=f.readlines()
f.close()

k=10
s=0

missing_cities=[]

for i in range(0,len(data)):
    while data[i][k].isdigit()==False:
        k=k+1
    city=data[i][10:k-1]
    numbers=data[i][k:-1].split(' ')
    print city
    if len(numbers)==9:
       new_numbers=[numbers[0],numbers[1],numbers[3],numbers[4]]
    if len(numbers)==4:
       new_numbers=[numbers[0],0,0,0]
    if len(numbers)==6:
       if numbers[0]==numbers[1] and numbers[2]==numbers[3] and numbers[4]==numbers[5]:
          new_numbers=[numbers[0],0,numbers[2],0]
       else:
          new_numbers=[numbers[0],numbers[1],0,0]

    k=10
    f=open('C:\Python27\er_data_deciphered.txt','r+')
    old_data=f.read()
```

```
if i==0:
   old_data=''
f.close()
f=open('C:\Python27\er_data_deciphered.txt','w+')
line2write=str(old_data+'1995,'+str(city)+','+str(new_numbers[0])+',
'+str(new_numbers[1])+','+str(new_numbers[2])+','+str(new_numbers[3])+'\n')
f.write(line2write)
f.close()
```

Appendix 2

A statistical mechanics approach to immigrant integration in Emilia Romagna (Italy)

A statistical mechanics approach to immigrant integration in Emilia Romagna (Italy)

Francesco De Pretis[1] and Cecilia Vernia[2]

[1] M2CSC School of Graduate Studies, Università degli Studi di Modena e Reggio Emilia, Italy
francesco.depretis@unimore.it
[2] Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università degli Studi di Modena e Reggio Emilia, Italy
cecilia.vernia@unimore.it

Abstract. Integration phenomena are social processes among human beings that take place every day when an autochthone population is experiencing the arrival of new immigrants. Although being a rising phenomenon (involving now over one billion people according to United Nations) which questions societies and policy-makers all over the world, numerical measurements capable to give robust insights over the way immigrant integration occurs are still far from what is usually considered an affordable standard in mathematical and physical sciences. Basing our analysis on previous seminal works, we follow here a statistical physics approach to the analysis of immigrant integration. In specific, we consider a large dataset collected by the Emilia Romagna region office of statistics (Italy), containing information over all marriages occurred amid the regional population during a sixteen years span, from 1995 to 2010. We define as quantifier of integration the percentage of marriages with spouses of mixed origin and we perform several analyses over the dataset, including binning and data fitting. The final outcome consists in an emerging pattern: quantifier's average measurements align around a square root fit when considered with respect to a suitable function of the immigrant density. The theoretical interpretation we offer is that such result agrees with a suitable version of the Curie-Weiss model used in statistical mechanics to describe ferromagnetisms. More explicitly, immigrants living in Emilia Romagna municipalities seem to present mainly imitative behavior's phenomena in making social actions for integration. The result emerged with Emilia Romagna data complies with previous works concerning similar data coming from Spain.

1 Introduction

Integration of immigrants is a political priority in many countries: there are over one billion migrants all over the world, one quarter of which are international migrants [1]. Even though it is not clear how sensitive integration is to an increase of immigrant density and to what extent social interaction goes into higher integration, it is easy to guess that social interaction between immigrants and autochthonous popula-

tion is a necessary condition for immigrant integration. Curie-Weiss models have been used in the last years in the quest to model social interactions and processes of decision taken by individual human beings [2,3,4,5,6]. In this paper we follow a statistical physics approach to the study of immigrant integration using methods and models already explored in a previous seminal work concerning a large collection of Spanish data [7]. Given a large dataset described in the subsequent section, we focus on a classical quantifier of integration such as the fraction of marriages with spouses of mixed origin (native – i.e. bearing Italian citizenship – and immigrant)

$$M_m = \frac{number\ of\ mixed\ marriages}{number\ of\ marriages}.$$

Within this framework, our goal is a statistical mechanics theory by which the magnitude of the above-mentioned quantifier can be expressed as a function of the density of immigrants, i.e. the ratio between the number of immigrants N_{imm} and the total population $N = N_{imm} + N_{nat}$ where N_{nat} is the number of natives:

$$\gamma = \frac{N_{imm}}{N_{imm} + N_{nat}} \in [0,1]$$

For a better representation of the integration quantifier – based on combinatorial reasoning [7] – we are interested in studying its dependence on the quantity $\Gamma = \gamma(1-\gamma)$. Afterwards, we seek an empirical function from real data, able to entail the observed collective behavior. As it will be reported in the results section, this work confirms that it is possible to discriminate, using quantitative methods, whether the value of the integration quantifier follows from people acting according to some individual preferences independently of other people (*independent choices*), or whether it follows as a result of social interaction with other ones (*imitative behaviors*). These two opposite cases are described in statistical mechanics theory either as perfect gas of independent particles (in this case, average measurements of the quantifier against Γ follow a linear growth) or interacting theory with possible phase transitions (in this case, average measurements of the quantifier against Γ align around a square root curve).

2 Data Description and Methods

As stated above, the perspective of this work belongs to the statistical mechanics methods used to explain social integration phenomena, starting from the analysis of real data. More precisely, the work has been centered on a large dataset collected by the Emilia Romagna region office of statistics (Italy), containing information recorded in all Emilia Romagna region municipalities (348 cities) regarding marriages occurred amid the population during a sixteen years span, from 1995 to 2010. In particular, for each municipality, the database provides the reference year, the number of marriages between Italians and foreigners, the number of marriages only between foreigners, the number of marriages only between Italians and the total amount of marriages.

Regarding the sources of data, the Emilia Romagna region dataset was somehow freely downloadable from the regional office of statistics website (data were accessible per municipality at given year, so that techniques of automatic web-contents wrapping have been employed to collect the entire dataset). All data were real (i.e. not estimated) and were subsequently matched with the density of immigrants for each municipality at each given year (information freely retrievable from ISTAT – Italian National Institute of Statistics – sources): the density of immigrants was estimated only for two specific years (1999 and 2000) since the recording of immigrant population was suspended during that period. Given the dimension of the considered dataset, the work can be somewhat inscribed in a *big data exploitation*: to give a rough idea of the computational efforts pursued, around 50.000 data have been processed in order to compute the above described quantifier (y-axis) matched with the density of immigrants (x-axis), producing the scatter plot reported below in Figure 1. It is worth to note that according to prescriptions of a time-independent analysis, data have been plotted together independently from the year they were referring to.

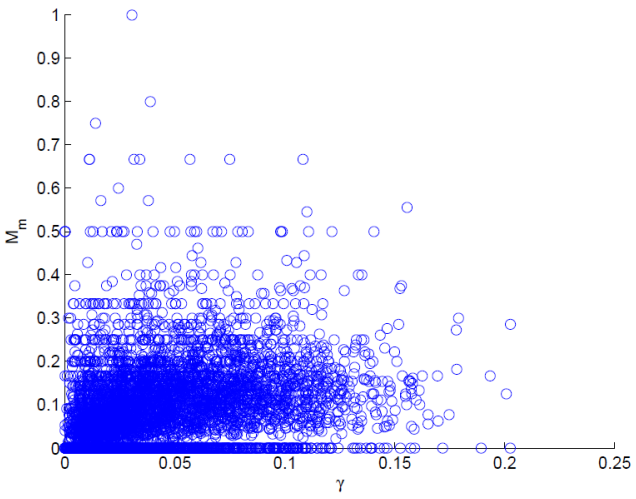


Fig. 1. – Raw data versus γ. Blue points represent the fraction of mixed marriages occurred from 1995 to 2010 in all municipalities located in Emilia Romagna region where a percentage γ of migrants is present.

Besides the representation through a scatter plot, we have been interested in the quantifier's average measurements as functions of Γ: since the quantifier M_m is a ratio (i.e. the total number of mixed marriages over the total number of marriages) and we were concerned in looking at patterns in the global scale of the dataset, a natural way to compute averages has been a *mediant* measure. This means that for a given bin of γ, we have computed the ratio between the statistical average of numerators and the statistical average of the denominators. This processing has been performed according to a constant information binning, i.e. each bin contained a fixed number of points. After having compacted the initial information of raw data into bins, a classical pro-

cedure of curve fitting has taken place. In particular, following the example of previous work conducted in [7], we have evaluated linear and power functions according to the R^2 coefficient of determination.

3 Results

After having performed the procedures described in the methods section, the following result has been obtained: quantifier's average measurements computed for all data coming from the Emilia Romagna region dataset have fitted around a visible square root pattern with a R^2 coefficient of determination being over 95%, highly reproducing results obtained with similar Spanish data in [7]. Since the analysis of the data density versus Γ shows that only about the 7% of the data are found for Γ greater than 10%, we limit our study below this threshold. It is worth of note that in Emilia Romagna region in 2010 the percentage of immigrants over the total population is about 11%, the highest density of immigrants with respect to any other Italian region.

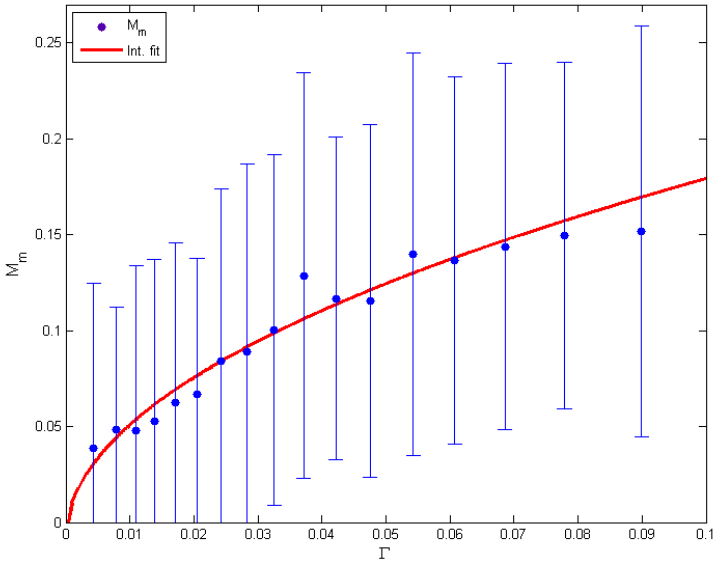


Fig. 2. – Emilia Romagna dataset. Dots are average quantities versus Γ, whereas lines denote error bars. Quantifier M_m (blue dots), fraction of mixed marriages occurred from 1995 to 2010 in all the municipalities located in Emilia Romagna region, with the best square root fit (red curve) $a\sqrt{\Gamma} + b$ (a = 0.5943 ± 0.0757, b= -0.008631 ± 0.014019, goodness of fit R^2 = 0.9529 computed for $\Gamma < 0.078$). Parameter b evaluation is compatible with the hypothesis that it can be null, as prescribed by the statistical mechanics model we use for results interpretation.

The result has been verified according to various types of binning (i.e. changing the number of bins) and various families of functions (for instance, linear functions). In

the end, a square root fit emerged as the best estimation for the quantifier's average measurements, since with linear and other fittings, the outcomes reported lower R^2 coefficient of determination associated with noisy fits highly depending on the nature of binning.
The mathematical model that supports these results is a generalization of the monomer-dimer model [8] with the addition of an imitative interacting social network component of small world-type [9]. The model, proposed and described in [7], reduces to the classical discrete choice theory [10] (or perfect gas of independent particles) with linear growth of the quantifier as a function of Γ, when imitation is negligible, and to the square root behavior when imitation is dominant. The social network structure explains why the integration starts very close to $\Gamma = 0$ when the choice is dependent on other agent behavior.

Therefore, translated in statistical mechanics terms according to the theoretical interpretation shown in [7], the result of an empirical square root function for the quantifier's average measurements offers an interesting picture of immigrant integration issues in Emilia Romagna region. In specific, even though we do not deal with the possible origins of such cooperative influence, we simply conclude that data suggest that in Emilia Romagna municipalities imitative phenomena mainly take place against the possibility of independent choices carried on by the same immigrants.

Acknowledgement. Authors express their gratitude to Pierluigi Contucci and Claudio Giberti for the inspiring insights and comments that helped writing this work.

References

1. The Global Approach to Migration and Mobility, EU report, Commission 743, Sec 1353, (2011)
2. S.N. Durlauf, Statistical mechanics approaches to socioeconomic behavior, Technical Working Paper, 203, Natl. Bur. Econ. Res. (1996).
3. A. Barra, P. Contucci, Toward a quantitative approach to migrants social integration, Europhys. Lett. 89, 68001, 68007, (2010)
4. I. Gallo, An equilibrium approach to modelling social interaction, PhD Thesis, Università di Bologna (2009)
5. P. Contucci, I. Gallo, G. Menconi, Phase transitions in social sciences: two-populations mean field theory, Int. J. Mod. Phys. B, Vol. 22, N. 14, 1-14 (2008)
6. A. Barra, P. Contucci, I. Gallo, Parameter Evaluation of a Simple Mean-Field Model of Social Interaction, Mathematical Models and Methods in Applied Science, Vol. 19, 1427-1439, (2009)
7. A. Barra, P. Contucci, R. Sandell, C. Vernia, Integration indicators in immigration phenomena. A statistical mechanics perspective, http://arxiv.org/abs/1304.4392 (2013)
8. O.J. Heilmann, E.H. Lieb, Monomers and dimers, Phys. Rev. Lett. 24, 25, 1412, (1970).
9. D.J. Watts, S.H. Strogatz, Collective dynamics of small world networks, Nature 393, 6684, (1998)
10. D. McFadden, Economic choices, The Amer. Econ. Rev. 91, (2001).

Appendix 3

The Statistical Physics approach on immigrant integration in Italy

The Statistical Physics approach on immigrant integration in Italy.

Pierluigi Contucci *, **Cecilia Vernia** †, **Francesco De Pretis** ‡

*Dipartimento di Matematica, Università di Bologna, Italy,†Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio Emilia, Italy, and ‡M2CSC School of Graduate Studies, Università di Modena e Reggio Emilia, Italy

How does integration of immigrants change when their percentage goes from 1% to 2% and up? We analysed data on classical integration quantifiers in Italy about the frequency of mixed marriages and newborn children from mixed couples. We discovered that, up to the average immigrant density of 7%, there are two different growth laws according to the size of the municipality. For large-sized municipalities the integration quantifiers grow linearly on the immigrant density. This means that if the quantifier has grown of a given amount Q when the immigrant density is 1% it will grow to 2Q at 2%, 3Q at 3% and so on. In small-sized municipalities instead in order to reach an integration of 2Q the immigrant density has to grow to 4%, to reach 3Q it has to grow to 9% and so on. While the first law has a constant growth rate the second, that follows the square root function, has an anomalous high growth rate when the phenomenon starts, which quickly decreases at increasing immigrant densities. We argue that the difference depends on the different level of social interaction in large-sized and small-sized municipalities. The statistical mechanics approach that we follow provides a natural interpretation for the two growth laws: the linear one is typical for those systems where individuals make their choices independently, while the square root growth signals the presence of imitative interaction among individuals on diluted networks. Our conclusions agree with classical theories of social alienation on large-sized municipalities and propose an explicit quantitative measure for it. Our work provides both an interpretative and predictive tool ready to be used by policy makers.

Despite the recent economic crisis happened in many countries immigration phenomena haven't shown any relevant signal of decrease [1]. Enormous socio-economic disparities among different places in the world are constantly fueling immigration fluxes. The necessity to build a functional coexistence of immigrants and natives with different socio-cultural backgrounds has become a top priority of political agendas [2].

In order to understand what are the suitable policies to ease and promote this process it is of paramount importance to know what are the natural mechanisms governing the integration phenomena. In particular it is necessary to have a quantitative description of them including, possibly, reliable forecasts on how the phenomena would change in case some of the parameters happen to change or are pushed to [3].

Immigrant density is among the important quantities and the way integration observables depend on it is the main topic of this letter. In a recent work [4] a new approach to study some integration quantifiers, applied to the case study of Spain, was proposed and based on methods, models and ideas whose origins can be traced to statistical physics. We follow here a similar approach for the immigration phenomena in Italy and we investigate a new set of extensive statistical data on mixed marriages and newborns to mixed couples, annually recorded by ISTAT (Italian Institute of Statistics) for all Italian municipalities in the time period from 2001 to 2011. The dataset contains over 1 100 000 data, yearly describing - per around 8100 municipalities - the total population, the number of immigrants, the number of marriages and newborns originating from different types of couples (either mixed or only foreigner or Italian ones). The first step of our analysis is to collect all the observed data for the municipalities in the *ensemble plane* representing in the horizontal axis the immigrant density γ and in the vertical axis the value of the integration quantifier. More precisely

$$\gamma = \frac{N_{imm}}{N_{imm} + N_{nat}}\,, \qquad [1]$$

where N_{imm} is the number of immigrants and N_{nat} is the number of the natives.

Alongside immigrant density, we focus on two classical integration quantifiers: the fraction of marriages with spouses of mixed origin (native and immigrant) M_m and the fraction of newborns with parents of mixed origin B_m. For a given quantifier, the graphical output is therefore a set of points, each one representing a municipality in a given year. The sketched sets - one for the marriages, the other for the newborns - quite resemble *clouds*, whose cardinality and shape are not the same but vary according to minor events occurring in tiny municipalities.

At first stage, this time-independent representation of data is not easily readable: data grouped in a unique ensemble, indifferently from the time they were referring to, do not show any clearly recognizable pattern (functional form). Anyway, we aim at studying average behaviors, i.e. estimating the average percentage of a mixed marriage or a mixed newborn, for a given immigrant density. This is our second step of analysis: beyond the randomness of all events recorded in the clouds, we are interested in the quantifier's averages as functions of γ. Since all quantifiers are ratios, a natural way in mathematics to tackle the averages is through a *mediant* measure. This means that for a given interval of γ values (a γ bin), we compute the ratio between the statistical average of numerators and the statistical average of the denominators. The output of this analysis can be seen as a binning process: a spare set of points concentrate all information previously coming from the entire clouds.

However, an initial binning performed over the two clouds results unsatisfactory: the average quantities turns out to be noisy and highly depending on the binning's parameters setting. This situation led to the idea of partitioning each dataset, separating small-sized municipalities from other ones. The implemented threshold (i.e. 10 000 inhabitants) was initially attempted according to previous work [4], where the considered municipalities had a population only exceeding 10 000 inhabitants. Ex-post testing has proved the threshold to be a sufficiently robust choice among others.

After the partition, the ensemble planes are presented as in **figure 1**. Left panels show data belonging to small-sized municipalities, whereas large-sized municipalities data are found in the right ones. Furthermore, upper panels picture data coming from the marriages dataset, while the births data are diplayed in the lower panels. This graphical layout applies also for the following figures.

Data density is shown in **figure 2**. Both for small-sized and large-sized municipalities, indipendently from the considered quantifier, data are statistically robust for γ<0.16, since about one percent of the data are always found over this immigrant density. Moreover, we notice here that the density decreases for γ very near to zero in each panel. The reason for this is that our observation window begins in 2001 when the migration phenomena was already ongoing and the entire density of migrants in Italy was larger than zero.

In **figure 3** we see the same binning process described above applied this time to the partitioned datasets. Since the product $N_{imm}N_{nat} \propto \Gamma$ explicitly counts the number of possible cross-group links, it is more convenient to map horizontally the quantity

$$\Gamma = \gamma(1-\gamma)\,. \qquad [2]$$

The quadratic function $\gamma \to \Gamma$ has the only effect to map the interval $[0, 1/2]$ monotonically onto $[0, 1/4]$. Moreover our whole analysis is a study of the phenomena in a suitable interval of Γ that turns out to be at least within the 7%. Eventually, bins have been fitted according to a square-root fit (left panels) and to a linear fit (right panels). Data excellently fit such curves with a R^2 coefficient of determination spanning from 96 % to 99 %.

These results lead to ask why data follow two different empirical laws if we observe the immigrant integration phenomena at different scales. The interpretation we offer is that such results agree with a statistical physics model already presented in [4]. More openly, we state that it can be possible to introduce an analogy between the behaviors of a large group of people in terms of social actions and the behaviors of a large group of magnets in terms of attractions [5] [6] [7] [8]. In both cases, the occurring interactions can be due either to internal phenomena or to external phenomena to the group. In specific, our analysis suggests that in small-sized municipalities imitative phenomena [9] mainly take place among immigrants, that in some way seem to be influenced by the behaviors of other people, whereas independent choices [10] look like to be the most common patterns in large-sized municipalities. These results turn out to be even more impressive if compared with subtle considerations made by classical sociology authors, who understood a crisp difference in social actions [11] occurring in small-sized or large-sized municipalities already more than one hundred years ago [12].

Eventually, these patterns are clearly distinguishable one from another: in **figure 4**, taking as example what is shown in the left panels, fitting small-sized municipalities data against a linear fit (last dot of the black curve) always produces worse results in terms of R^2 measures with respect to the square root fit (last dot of the red curve). Moreover, this figure highlights that these empirical laws continue to visibly emerge also if we observe integration phenomena from the very first years: a predictive analysis - considering time as another variable - shows that such patterns are manifest even if we perform the same analyses first only on 2001 data, then on 2001-2002 data and so on, cumulating data year by year until we fall again into the whole datasets.

In conclusion, this predictability indeed enhances the theoretical interpretation we offer, making this work a viable apparatus for decision-making in immigration policies.

1. The Global Approach to Migration and Mobility, EU Report, Commission 743, Sec 1353, (2011)
2. http://ec.europa.eu/ewsi/en/EU_actions_integration.cfm, and also, European Union,Commission of the European Communities. (2005). A Common Agenda for Integration Framework for the Integration of Third-Country Nationals in the European Union. COM(2005) 389 final.
3. Rinus Penninx, Dimitrina Spencer and Nicholas Van Hear, Migration and Integration in Europe: The State of Research, University of Oxford (2008)
4. Barra, A., Contucci, P., Sandell, R., Vernia, C., The Statistical Mechanics of Social Actions in Immigrant Integration. Preprint. Accepted to be published on Scientific Reports, Nature (2014)
5. S.N. Durlauf, How can statistical mechanics contribute to social science?, Proc. Natl. Acad. Sc. USA 96, (1999).
6. C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics, Rev. Mod. Phys. 81, 591-646, (2009).
7. P. Contucci, C. Giardina, Mathematics and Social Sciences: A Statistical Mechanics Approach to Immigration, ERCIM News, Vol. 73, (2008).
8. A. Barra, P. Contucci, Toward a quantitative approach to migrants social integration, Europhys. Lett. 89, 68001, 68007, (2010).
9. W. Brock, S. Durlauf, Discrete choices with social interactions, Rev. Econ. St. 68, (2001).
10. D. McFadden, Economic choices, The Amer. Econ. Rev. 91, (2001).
11. Max Weber, Economy and Society, [1921] 1978, p. 23.
12. Durkheim, E. (1897). Le Suicide: Etude de sociologie. Paris: Flix Alcan.

SUPPLEMENTARY INFORMATION concerning data description is presented below, beyond this paper.

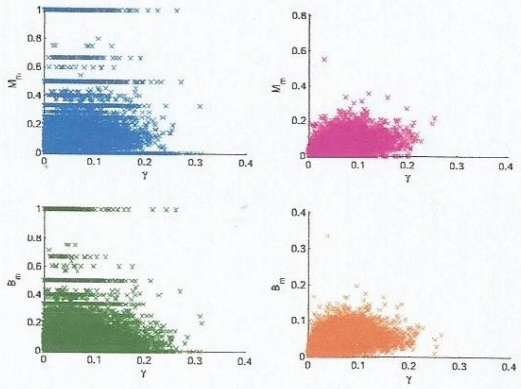


Fig. 1. Raw data versus γ. Blue points represent the fraction of mixed marriages occurred in municipalities with less than 10.000 inhabitants where a percentage γ of migrants is present; similarly green points account for newborns from mixed couples. Further, pink points represent the fraction of mixed marriages occurred in municipalities with more than 10.000 inhabitants, while orange ones mirror the newborns from mixed couples. One may note that data in the left panel seem to lie along horizontal lines displaced according to $1/n$, with $n \in \mathbb{N}$ due to effects seen in small-sized municipalities. See the discussion within the supplementary information.

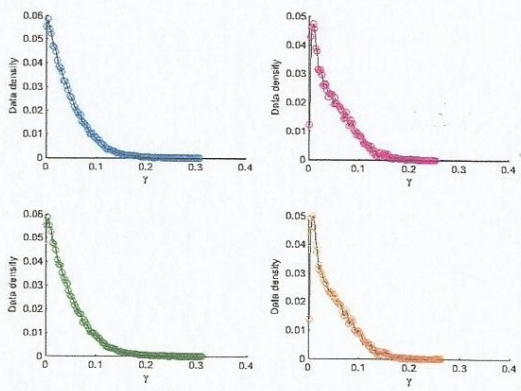


Fig. 2. Density of the mixed marriages datasets and of the newborns belonging to mixed couples datasets as a function of γ. Left upper panel: density of the mixed marriages (blue circles) dataset for municipalities with less than 10.000 inhabitants. Right upper panel: density of the mixed marriages (pink circles) dataset for municipalities with more than 10.000 inhabitants. Left lower panel: density of the newborns belonging to mixed couples (green circles) dataset for municipalities with less than 10.000 inhabitants. Right lower panel: density of the newborns belonging to mixed couples (orange circles) dataset for municipalities with more than 10.000 inhabitants.

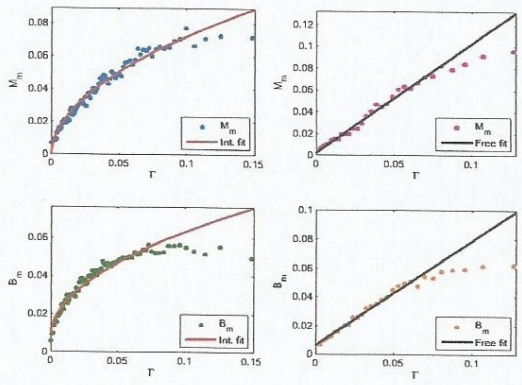


Fig. 3. Dots are average quantities versus Γ. Left upper panel: quantifier M_m (blue dots), fraction of mixed marriages occurred in municipalities with less than 10.000 inhabitants, with the best square root fit (red curve) $a\sqrt{\Gamma} + b$ (a = 0.2329 $\pm$0.0086, b = -0.002043$\pm$0.00168815, with a goodness of fit R^2 = 0.9746 computed for Γ<0.126). Right upper panel: quantifier M_m (pink dots), fraction of mixed marriages occurred in municipalities with more than 10.000 inhabitants, with the best linear fit (black curve) $a\Gamma + b$ (a = 1.002 $\pm$0.046, b = 0.001994$\pm$0.002, with a goodness of fit R^2 = 0.9879 computed for Γ<0.081). Left lower panel: quantifier B_m (green dots), fraction of newborns with mixed parents, born in municipalities with less than 10.000 inhabitants, with the best square root fit (red curve) $a\sqrt{\Gamma} + b$ (a = 0.1735 $\perp$0.00815, b = 0.007938$\perp$0.0015215, with a goodness of fit R^2 = 0.9614 computed for Γ<0.096). Right lower panel: quantifier B_m (orange dots), fraction of newborns with mixed parents, born in municipalities with more than 10.000 inhabitants, with the best linear root fit (black curve) $a\Gamma + b$ (a = 0.7764 $\perp$0.04485, b = 0.005253$\perp$0.001424, with a goodness of fit R^2 = 0.9908 computed for Γ<0.07)

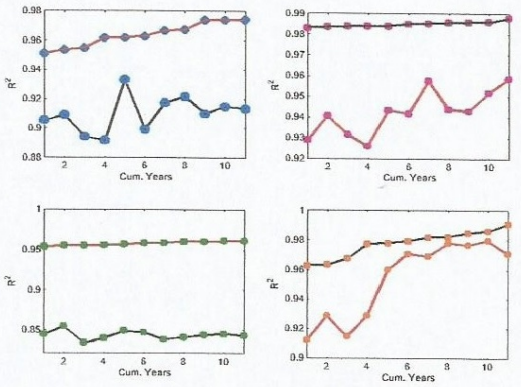


Fig. 4. Cumulative time behaviour: R^2 measurements for goodness of fit against an increasing years span. R^2 pictured through a red curve when related to a square root fit and through a black curve when related to a linear fit. Left upper panel: R^2 (blue dots) related to mixed marriages occurred in municipalities with less than 10.000 inhabitants. Right upper panel: R^2 (pink dots) related to mixed marriages occurred in municipalities with more than 10.000 inhabitants. Left lower panel: R^2 (green dots) related to newborns from mixed couples occurred in municipalities with less than 10.000 inhabitants. Right lower panel: R^2 (orange dots) related to newborns from mixed couples occurred in municipalities with more than 10.000 inhabitants.

SUPPLEMENTARY INFORMATION

Data description

In this section, we give a deeper description of the data we analysed, putting the focus on cardinality and density issues and on the distribution of immigrants and natives with respect to small-sized and large-sized municipalities.

The records about marriages come from about 8100 Italian municipalities - number which has slightly changed over years, as shown in **figure 5** - and span over a period of time of eleven years, from 2001 to 2011; measurements have been registered yearly and divided into four categories: marriages between people exclusively bearing Italian citizenship, marriages between people exclusively bearing non-Italian citizenship, marriages where only the husband is bearing the Italian citizenship and marriages where only the wife is bearing the Italian citizenship. The two latter ones have been defined as mixed marriages.

The entire marriages dataset contains 89093 records; however, combining together the two scatter-plots shown in the upper panels of **figure 1**, we count only 82208 points. This discrepancy is due to the fact that in all other records no marriages at all were registered: this is the null contribution given by very tiny municipalities in some given years, accounting for about 7.73% of the whole dataset. Moreover, in about 56.13% of all records no mixed marriages have occurred. Eventually, considering the partition of the dataset into two subsets (as shown in the article), we observe that records belonging to municipalities whose population is under 10000 inhabitants count up to 84.43%.

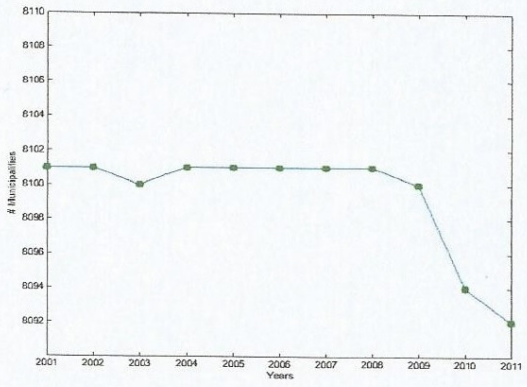


Fig. 5. Marriages dataset. Number of Italian municipalities from 2001 to 2011. This number has not been constant because of administrative reforms which have merged or created new municipalities in Italy.

Looking at data density plots (**figure 2**, upper panels), we find that marriages data are statistically robust already for a value of γ around 0.16, being the last percentile of data found over this threshold. Moreover, both data density tails appear to be power law distributed. For the restriction to small-sized municipalities, we find for $\gamma \geq 0.1$ the law $\mu(\gamma) \propto \gamma^{\delta}$ with $\delta = -4.1 \pm 0.168$ with a goodness of fit measured by the coefficient of determination R^2=0.9877. Instead, for the restriction to large-sized municipalities, we find for $\gamma \geq 0.1$ the law $\mu(\gamma) \propto \gamma^{\delta}$ with $\delta = -4.469 \pm 0.312$ with a goodness of fit measured by the coefficient of determination R^2=0.967.

A similar description to the marriages dataset can be drawn for the births dataset. The time period data refer to is the same and a similar classification of births has been made: births originating from couples whose partners are exclusively bearing Italian citizenship, births originating from couples whose partners are exclusively bearing non-Italian citizenship, births originating from couples where only the father is bearing the Italian citizenship and births originating from couples where only the mother is bearing the Italian citizenship. The two latter ones have been defined as mixed births.

Considering cardinality issues, we observe that the null contribution of births given by very tiny municipalities in some given years accounts this time for about 3.61% of the whole dataset, whereas in about 43.18% of all records no mixed newborns have occurred. Furthermore, same percentage weigths of data are found when partitioning the birth dataset into two subsets, fixing the population threshold at 10000 inhabitants.

As for the marriages data, the analysis of data density plots (**figure 2**, lower panels) signals that births data are again statistically robust already for a value of γ around 0.16. Alike, both births data density tails appear to be power law distributed. For the restriction to small-sized municipalities, we find for $\gamma \geq 0.1$ the law $\mu(\gamma) \propto \gamma^{\delta}$ with $\delta = -4.108 \pm 0.152$ with a goodness of fit measured by the coefficient of determination R^2=0.9901. Instead, for the restriction to large-sized municipalities, we find for $\gamma \geq 0.1$ the law $\mu(\gamma) \propto \gamma^{\delta}$ with $\delta = -4.542 \pm 0.291$ with a goodness of fit measured by the coefficient of determination R^2=0.9722.

The partition of the marriages and births datasets which has led to the main results of this article has been investigated under various perspectives. A first question is wheater immigrant density γ varies according to such partition, i.e. if this indicator changes according to the size of the municipality where it has been measured. **Figure 6** casts light over such question: in this graph, for the marriages dataset (similar results can be proved for the births dataset) we find on the horizontal axis the eleven years the dataset refers to and on the vertical axis the immigrant density γ.

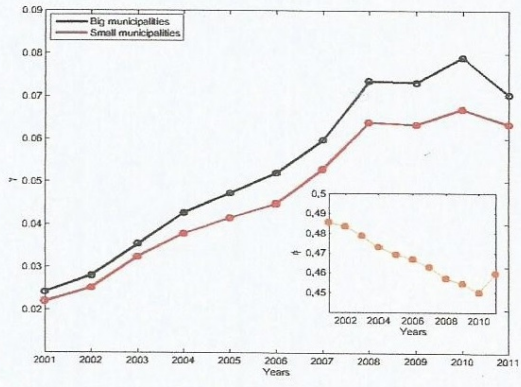


Fig. 6. Marriages dataset. Black line represents the growth of immigrant density γ^b in municipalities with more than 10.000 inhabitants from 2001 to 2011. Red line mirrors the same growth referred to immigrants living in municipalities with less than 10.000 inhabitants. In the inset it is shown for the same period the variation of ϕ, i.e. population living in small-sized municipalities over population living in large-sized municipalities

The black line shows how immigrant density γ^b changes over time in large-sized municipalities and the red line illustrates the same variation over time in small-sized municipalities. At first glance, it is manifest that the values of γ^b and γ^s are comparable in absolute terms: the maximum difference is 0.0121 between the two curves for the year 2010. Moreover, the two curves show a high level of linear correlation with the correlation coefficient $\rho = 0.9983$. We can therefore infer that immigrants seem to be distributed in the same manner, independently from the size of the municipality they belong to. Eventually, in the **figure 6** inset we find the variation of ϕ, a ratio defined as

$$\phi = \frac{N^s}{N^b} \quad [3]$$

measuring the proportion between total population living in small-sized (N^s) and large-sized municipalities (N^b).

Always related to the partition of datasets, a second question to be posed is a comparison between the proportion of immigrants π^s_{imm} in small-sized municipalities and the proportion of natives π^s_{nat} belonging to the same geographical areas. Being N^s_{imm} the immigrants living in small-sized municipalities, we define the proportion of immigrants in small-sized municipalities as:

$$\pi^s_{imm} = \frac{N^s_{imm}}{N_{imm}} \quad [4]$$

Similarly, being N^s_{nat} the natives living in small-sized municipalities, we define the proportion of natives in small-sized municipalities as:

$$\pi^s_{nat} = \frac{N^s_{nat}}{N_{nat}} \quad [5]$$

Worth to note, both π^s_{imm} and π^s_{nat} can be obtained as function of γ and ϕ. In fact, we have

$$\pi^s_{imm} = \frac{\phi\gamma^s}{\phi\gamma^s + \gamma^b} \quad [6]$$

whereas to compute π^s_{nat}, it is sufficient to substitute the values of γ with their complementary $1 - \gamma$.
Focusing again on the marriages dataset - this time considering data coming only from small-sized municipalities - in **figure 7** we find on the horizontal axis the eleven years the dataset refers to and on the vertical axis the proportion π. The blue line shows how the proportion of natives π^s_{nat} in small-sized municipalities changes over time and the green line illustrates the same variation over time for the proportion of immigrants π^s_{imm} in small-sized municipalities. At first glance, we can state that both proportions π^s_{nat} and π^s_{imm} can be approximated at 30%, meaning that there are no substantial differences in the way immigrants and natives distribute themselves in small-sized municipalities (and conversely also in large-sized municipalities). Furthermore, the two curves can be estimated to be linearly correlated: the variation of proportion π^s_{nat} against the variation of proportion π^s_{imm} presents a coefficient of correlation $\rho = 0.9190$, which becomes even higher ($\rho = 0.9713$) if we exclude the last year 2011 in the time-series. Same results are found also for the births dataset, enhancing the conclusions we previously drew.

Eventually, we claim that the partition of data highlights features that are present only in small-sized municipalities, independently from the considered quantifier: looking at **figure 1**, one may note that data in left panels seem to lie along horizontal lines displaced according to $1/n$, with $n \in \mathbb{N}$. This optical effect is due to events occurring in very tiny municipalities, where the total number of marriages or newborns is so small that the presence of mixed couples or mixed newborns is greatly emphasized, raising up the quantifiers value (e.g. in a small village where only two marriages have been celebrated in one year, one involving a mixed couple, the quantifier Q is one half).

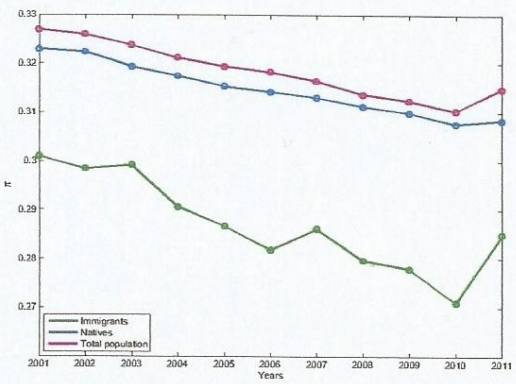


Fig. 7. Marriages dataset (restriction to small-sized municipalities). Blue line represents the proportion π^s_{nat} of natives living in municipalities with less than 10.000 inhabitants from 2001 to 2011. Green line shows the proportion π^s_{imm} of immigrants living in the same municipalities during the same period. Pink line mirrors the same for the proportion π^s of total population (i.e. immigrants *and* natives) living in small-sized municipalities

Acknowledgements

I would like to express my gratitude to my advisor Prof. Claudio Giberti for his important help.

Special thanks go also to Prof. Pierluigi Contucci and Prof. Cecilia Vernia for the interesting discussions which influenced this work.

References

Agliari, E. & Barra, A. (2011). A Hebbian approach to complex network generation, Europhys. Lett. 94, 10002.

Alberici, D., Contucci, P. & Mingione, E. (2013). A mean-field monomer-dimer model with attractive interaction. The exact solution. ArXiv, 1-32.

Arnoldi, W. E. (1951). The principle of minimized iterations in the solution of the matrix eigenvalue problem, Quarterly of Applied Mathematics, volume 9, 17–29.

Barra, A. & Agliari, E. (2011). Equilibrium statistical mechanics on correlated random graphs, J. Stat. Mech. 02, 02027.

Barra, A. & Agliari, E. (2012). A statistical mechanics approach to Granovetter theory, Physica A, 391, 10, 3017.

Barra, A. & Contucci, P. (2010). Toward a quantitative approach to migrants social integration, Europhys. Lett. 89, 68001, 68007.

Barra, A., Contucci, P., Sandell, R. & Vernia, C. (2013). Integration indicators in immigration phenomena. A statistical mechanics perspective. ArXiv, 1-11.

Baxter, R. J. (1982). Exactly Solved Models in Statistical Mechanics. London: Academic Press.

Berry B.J.L. & Okulicz-Kozaryn A. (2012). The city size distribution debate: Resolution for US urban regions and megalopolitan areas, Cities, Vol. 29, Sup. 1, S17–S23

Bianconi, G. (2002). Mean field solution of the Ising model on a Barabasi-Albert network, Phys. Lett. A 303, 166-168.

Bijak, J. & Wiśniowski, A. (2009). Forecasting of immigration flows until 2025 for selected European countries using expert information. IDEA Working Papers n° 7.

Bollobas, B. (2000). Random Graphs, Cambr. Stud. Adv. Math. Cambridge, UK: Cambridge University Press.

Bolthausen E. & Bovier, A. (2007). Spin Glasses. Berlin: Springer-Verlag.

Bovier, A. (2006). Statistical Mechanics of Disordered Systems: a Mathematical Perspective. Cambridge, UK: Cambridge University Press.

Bragg, W. L. & Williams, E. J. (1934). The effect of thermal agitation on atomic arrangement in alloys. Proceedings of the Royal Society of London, 145A, 699-730.

Brock W. & Durlauf S.N. (2001). Discrete choices with social interactions, Rev. Econ. St. 68.

Burioni R., Contucci P., Fedele M., Vernia C. & Vezzani A. (2014) Statistical mechanics forecasts for health screening campaigns, Preprint.

Castellano C., Fortunato S. & Loreto V. (2009). Statistical physics of social dynamics, Rev. Mod. Phys. 81, 591-646.

Chandler, D. (1987). Introduction to modern statistical mechanics. Oxford, UK: Oxford University Press.

Contucci, P., Gallo, I. & Menconi, G. (2008). Phase transitions in social sciences: two-populations mean field theory. Intenational Journal of Modern Physics B, 22(14), 1-14.

Contucci, P. & Giardinà, C. (2012). Perspectives on Spin Glasses. Cambridge, UK: Cambridge University Press.

Curie, P. (1895). Propriétés magnétiques des corps à diverses températures. Annales de chimie et de physique, 5(7), 289-405.

Dembo, A. & Montanari, A. (2010). Ising models on locally tree-like graphs, Ann. Appl. Probab. 20, 565-592.

De Pretis, F. (2012, June 27). Nella tempesta globale quali numeri ci salveranno? La Stampa.

De Pretis, F. & Vernia C. (2013). A statistical mechanics approach to immigrant integration in Emilia Romagna (Italy). Preprint. Accepted to be published in Complex Networks, Studies in Computational Intelligence 424, Springer-Verlag.

Durkheim, É. (1897). Le Suicide : Étude de sociologie. Paris: Félix Alcan.

Durlauf S.N. (1996). Statistical mechanics approaches to socioeconomic behavior, Technical Working Paper, 203, Natl. Bur. Econ. Res.

Durlauf S.N. (1999). How can statistical mechanics contribute to social science?, Proc. Natl. Acad. Sc. USA 96.

Efron, B. (1979). Bootstrap methods: Another look at jackknife. The Annals of Statistics, 7(1), 1-26.

Ellis, R. S. (1985). Entropy, Large Deviations, and Statistical Mechanics. New York: Springer-Verlag.

Evans, M. (2006). Phase Transitions and the Ising Model. Edinburgh: The University of Edinburgh.

Fedele, M., Vernia, C. & Contucci, P. (2013). Inverse problem robustness for multi-species mean field spin models. Journal of Physics A: Math. Theor. Vol 46, 065001.

Förstner, W. & Moonen, B. (1999). A metric for covariance matrices, Technical report, Dept. of Geodesy and Geo-informatics, Stuttgart University

Galam S. & Moscovici S. (1991). Towards a theory of collective phenomena: Consensus and attitude changes in groups, Europ. J. Social. Psycho. 21, 1, 49-74.

Gallo, I. (2008). An equilibrium approach to modelling social interaction. Bologna: Bologna University.

Gallo, I., Barra, A. & Contucci, P. (2009). Parameter Evaluation of a Simple Mean-Field Model of Social Interaction. Mathematical Models and Methods in Applied Science, 19, 1427-1439.

Gallo, I. & Contucci, P. (2008). Bipartite Mean Field Spin Systems. Existence and Solution, Math. Phys. Elec. Jou. 14, 1, 1-22.

Geem, Z. W., Kim, J. H. & Loganathan, G. V. (2001). A New Heuristic Optimization Algorithm: Harmony Search. Simulation, 76(2), 60-68.

Geem, Z. W. (2009). Music-Inspired Harmony Search Algorithm. Berlin: Springer-Verlag.

Geem, Z. W. (2010). Recent Advances in Harmony Search Algorithm. Berlin: Springer-Verlag.

Georgescu-Roegen, N. (1971). The Entropy Law and the Economic Process. Cambridge, Massachusetts: Harvard University Press.

Gerber, J.J. & Macionis, L. M. (2010). Sociology. Toronto: Pearson Canada. p. 97.

Heilmann, O.J. & Lieb, E.H. (1970). Monomers and dimers, Phys. Rev. Lett. 24, 25, 1412.

Horn, R. A. & Johnson, C. R. (1985). Matrix Analysis. Cambridge, UK: Cambridge University Press.

Huang, K. (1987). Statistical Mechanics. Cambridge: Massachusetts: Wiley.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. Zeitschrift für Physik, 31, 253-258.

Jevons, W. S. (1871). The Theory of Political Economy. London: Macmillan & Co.

Laplace, P. S. (1814). Essai philosophique sur les probabilités. Paris: Courcier.

Leone, M., Vasquez, A., Vespignani, A. & Zecchina, R. (2002). Ferromagnetic ordering in graphs with arbitrary degree distribution, Europ. Phys. Journ. B 28, 191.

McFadden, D. (2001). Economic choices, The American Economic Review, 91.

Mézard, M. & Montanari, A. (2009). Information Physics and Computation. Oxford: Oxford University Press.

Montanari, A. (2007). The Curie-Weiss model. Stanford, California: Stanford University.

Montanari A. & Saberi A. (2010). The spread of innovations in social networks, Proc. Natl. Acad. Sc. USA, N.107, 20196.

Nishimori, H. (2001). Statistical physics of spin glasses and information processing. Oxford: Oxford University Press.

Onsager, L. (1944). Crystal Statistic I. A Two-Dimensional Model with an Order-Disorder Transition. Physical Review, 65, 117-149.

Parisi, G. (1988). Statistical Field Theory. New York, NY: Addison-Wesley.

Pierce, A. (2008, November 5). The Queen asks why no one saw the credit crunch coming. The Daily Telegraph.

Simon, B. (1993). The Statistical Mechanics of Lattice Gases. Princeton, NJ: Princeton University Press.

Soros, G. (2010). Anatomy of Crisis – The Living History of the Last 30 years: Economic Theory, Politics and Policy. INET Inaugural 2010 Conference (p. 1-10). King's College Cambridge, England: INET.

Susskind, L. (2013). The Theoretical Minimum Lectures. Stanford, USA: Stanford University Web-Resources.

Talagrand, M. (2010). Mean Field Models for Spin Glasses. Paris: Springer-Verlag.

Thompson, C. J. (1979). Mathematical Statistical Mechanics. Princeton: Princeton University Press.

Weber, M. (1922). Wirtschaft und Gesellschaft. Tübingen: Mohr Verlag.

Weber, T. (2011, January 26). Davos 2011: Soros warns UK cuts could cause recession. BBC News website.

Weiss, P. (1907). L'hypothèse du champ moléculaire et la propriété ferromagnétique. Journal de Physique Théorique et Appliquée, 6(4), 661-690.

Willekens, F. (1994). Monitoring international migration flows in Europe. European Journal of Population / Revue européenne de Démographie, 10(1), 1-42.