An analysis of a large dataset on immigrant integration in Spain. The Statistical Mechanics perspective on Social Action.

Adriano Barra,[1] Pierluigi Contucci*,[2] Rickard Sandell,[3] and Cecilia Vernia[4]

[1] *Dipartimento di Fisica, Sapienza Università di Roma*
[2] *Dipartimento di Matematica, Università di Bologna*
[3] *Departamento de Ciencias Sociales, Universidad de Carlos III de Madrid*
[4] *Dipartimento di Scienze Fisiche Informatiche e Matematiche, Università di Modena e Reggio Emilia*

Supplementary Material

This appendix is used to explain the technical specifications and computational details for both the data analysis and elaboration (first section) and the theoretical mathematical part on statistical mechanics (second section).

Data Analysis and Elaboration

To investigate the functional dependence of the quantifiers on γ, we first tested the time series of the parameters involved. Starting from the raw data, we consider the average values of each quantifier Q and of the migrant's density γ in the two databases as a function of the quarters t. The result is shown in Supplementary Fig. S1a. The newborns with mixed parents display a very regular linear increase over time. The mixed marriages behave similarly but have an added seasonal periodicity as expected from cultural behavior. The two labor quantifiers exhibit a more complex behavior over time. The two lower panels of Supplementary Fig. S1a show how the density of immigration increases over time in the two databases. Using those functions $\gamma(t)$ and inverting them in $t(\gamma)$, we can plot each quantifier $Q(t)$ in terms of γ, thereby obtaining $Q(t(\gamma))$ represented in the four panels of Supplementary Fig. S1b.

As we can see, apart from a vague functional dependence on the newborns, all of the other quantifiers display erratic behavior and escape a functional law. It is evident that the time fluctuations through which these graphs are obtained contain spurious external effects. To explain better this point a parallel with thermodynamics is very enlightening: the derivation of the law that relates magnetization and temperature from time series would be extremely difficult in a condition in which energy is pumped into a ferromagnet from a random source. In fact we should first derive the laws of magnetization with time and temperature with time. Then we should eliminate the time variable from those data. Large fluctuation would make the data useless. Moreover when those fluctuations seem to be absent (the newborns case), the two processes of marginalization over time and inversion yield a very poor output. The bottom line is that the time series approach is not the suitable method for obtaining the functional dependence we are looking for since it loses relevant information and propagates spurious external effects.

To fully use the rich information of the two databases and extract from them the functional dependence of the quantifiers in terms of γ, we first proceed by identifying the empirical probability distribution ensemble for each dataset. We do so by merging into a unique catalogue the data entries in each database, regardless of their coordinates in space and time, and ordering them by increasing values of γ. The observed time windows cover a time scale much larger than the typical time scales involved in the dynamics of the jobs market or marriages/newborns that we focused on. Analysis of their density versus γ (Supplementary Fig. S2) shows that, for both datasets, only about one percent of the data are found for $\gamma \geq 0.4$. To efficiently model the macroscopic behavior of the integration quantifiers with robust statistics, we limit our study below that threshold. We also notice that data density decreases for small γ. The reason for this is that our observation window started when the migration phenomena was already running and the density of migrants in Spain was larger than zero. The immigrant densities appear to have tails with power law distribution. For the marriages and newborn dataset, as highlighted in the inset, we find for $\gamma \geq 0.2$ the law $\mu(\gamma) \sim \gamma^{\delta}$ with $\delta = -3.241$.

Supplementary Figure S3 shows the raw data *clouds* for each quantifier. An apparent anomaly is the presence in the lower left panel of horizontal lines where the data agglomerate. A further analysis shows that their values are due to the fractions with small denominators, that is, municipalities where the total number of marriages within the observed quarter does not exceed the few units. The explanation of this anomaly is found in the strong cyclical behavior due to seasonal preferences about the appropriate time for marriage in Spain (see also Supplementary Fig. S1). People in Spain prefer to marry in the summer rather than in the winter, which is unsurprising. Aggregating the data from quarters to years would wash away the anomaly. Nevertheless, as we explain below, it turns out that roughening the data in this way is not necessary, and it is possible to keep the dataset as it is, and hence preserve its richness.

Since we are interested in the quantifier's averages as functions of γ and since all quantifiers are ratios, there are two possible ways of computing the averages. For a given bin of γ one can compute 1) the statistical average of the ratios, or 2) the ratio between the statistical average of numerators and the statistical average of the denominators. The first is the usual mean of the ratios and the second is their global *mediant*. As will be explained later the difference in the results obtained from the two distinct procedures is in the range of 0.1 to 0.2 percent (see Supplementary Fig. S4) and they can consequently be considered as effectively equivalent.

We then proceed by grouping the data into bins over γ in which the averages can be evaluated. We found that 15 to 20 bins optimizes the job market dataset, while 35 to 40 optimizes the dataset on marriages and newborns. For the binning criteria, we tested the method of constant information and that of constant bin width. The advantage of the first criterion is a constant robustness quality across all bins. Clearly, with this approach, the width of the bin will vary over γ (their width increases at high values of γ due to the data density decrease reported in Supplementary Fig. S2). This can be avoided if we instead use constant bin width. However, with this approach the tradeoff is that as γ increases, the amount of data inside each bin may diminish (in particular for $\gamma \geq 0.4$). As we confined our analysis to $\gamma_{max} = 0.4$ for robustness requirements, the two criteria produce essentially the same results as one can see in Supplementary Fig. S4.

Figure 3 (upper panels) in the main text shows the outcome of the average criteria and coarse graining procedure for J_p and J_t. The dot's plots are made of 17 bins. On each bin the dot represents the average value of 200 data. On each figure the black curve is the *free fit*, that is, the curve of type $c_F\gamma(1-\gamma)$ best fitting the experimental points. Their goodness of fit, reported in the relative captions, is estimated as $R^2_{J_p(\gamma)} \sim 0.985$ and $R^2_{J_t(\gamma)} \sim 0.963$.

Figure 3 (lower panels) in the main text shows the results for M_m and B_m. In this case, the plots of the dots are made of 38 bins and each of them comes from 700 points. In this case, the free fit has a much lower goodness of fit: $R^2_{M_m(\gamma)} \sim 0.855$ and $R^2_{B_m(\gamma)} \sim 0.789$. In particular, the data show an anomalously high growth rate for small γ and a low one for large γ. For this reason we tested another family of curves, whose genesis we aim to explain through statistical mechanics in the next sections, and ultimately account for interactions among persons. Remarkably, all of these curves scale as $c_I\sqrt{\gamma(1-\gamma)}\theta(\gamma-\gamma_c)$ where θ is the step function. The agreement of the fit with the experimental data is clearly shown by the values $R^2_{M_m(\gamma)} \sim 0.992$ and $R^2_{B_m(\gamma)} \sim 0.984$. In the above formula, one may note the classical exponent *one half* typical of theories that account for imitative interaction. The presence (or lack) of a critical value γ_c is determined by the underlying social network. It is known that in ferromagnetic theories network dilution eventually decreases γ_c to zero, without affecting the critical exponent for a large class of social topologies [33][35][36][37][38][39][40][41][42]. Accordingly, we found empirical values $\gamma_c \sim 10^{-3}$ as reported in the caption.

For each quantifier Q of Figure 3 the single dot, representing the average within the bin, can be obtained using two averaging process (mean and mediant) and two kind of binning criteria for γ (bins with constant information and with constant width). In Supplementary Fig. S4 we test the robustness of our findings against these possible choices. Therefore, for each quantifier Q in the γ-bin we compute Q_A, the mean value using binning with constant information, Q_B the mean value using binning with constant width, Q_C the mediant with constant information binning and Q_D the mediant with constant width binning. In Supplementary Fig. S4 we represent the relative errors ($|Q_i - Q_j|/Q_j$, $i \neq j$ $i,j \in$ {A,B,C,D}) and absolute errors ($|Q_i - Q_j|$, $i \neq j$ $i,j \in$ {A,B,C,D}) as a function of γ made by comparing in couples all this possible choices for mixed marriages and newborns from mixed couples. Note that the relative errors are expected to increase at high values of γ, while in that region they are reduced with respect to those for smaller γ by at least one order of magnitude in all observables. The apparent increase of the relative error sizes at small γ is due instead to a ratio between small numbers and is a simple and harmless consequence of numerical noise as confirmed by the behaviors of their absolute values.

For completeness we also analyse the quantifier fluctuations from the averages, whose results are shown in Supplementary Fig. S5. To obtain those distributions, we proceed as follows. Let us focus on a concrete example, e.g. the permanent jobs reported in the main text in Fig. 1 (green plot) and the corresponding quantifier's fluctuations shown in Supplementary Fig. S5 (green plot). We take each dot in the first picture (representing the average of J_p within its corresponding $\Delta\Gamma$ bin) to be the center of the (green) distribution reported in Supplementary Fig. S5. Of course, within the chosen $\Delta\Gamma$ bin of Fig. 1, not all the data values contributing to obtain the average will be exactly the value of the mean (represented by the green spot), but some one will be smaller (and fall on the right in the corresponding panel of Supplementary Fig. S5) and some other will be bigger (so to fall on the left). Summing on all the deviations from the corresponding means of each spot in the same (green) plot of Fig. 1, when all the points have been investigated, the green histogram of Supplementary Fig. S5 is then finished and we move to another quantifier. Still in Supplementary Fig. S5, beyond the reported bare results, the latter are also fitted with standard distribution laws as Gaussian, Logistic, Gumbel, and Cauchy (the goodness of fits are reported in the captions). The rescaled distribution for $M_m(\gamma)$ is slightly asymmetric to the right: the Gumbel distribution (compared with Logistic and Gauss distribution) gives the best fit. The rescaled distribution for $J_p(\gamma)$, $J_t(\gamma)$ and $B_m(\gamma)$ is more symmetric: in these cases the Logistic distribution (compared with Cauchy and Gauss distribution) yields the best fit. In the insets the normal probability plots for the rescaled quantifiers show that the tails of the sample distributions are non-Gaussian.

Statistical Mechanics Methodology

This section is devoted to Methods, in particular to fill the gap between the empirical laws and the theoretical apparatus of statistical mechanics, especially devoted to the readers who are closer to the hard sciences. Our aim is the introduction of a mathematical model based on individual choices, with an internal structure (microscopic theory) whose emergent social behavior (macroscopic theory) reproduces all the observed quantifiers. The model reduces to the standard discrete choice theory [24] when peer-to-peer interaction is negligible and to its interacting version when imitation is instead the dominating factor.

To introduce the statistical mechanics model, we assign to each person their own tendency to marry versus remaining single. In addition, each couple (i, j) has their own likelihood to marry or not. Similarly, each person has an individual tendency to have children, and each couple too. All of these phenomena are then described by individual random variables and couple random variables.

The two observables about marriages and children are of course different: for marriages, the monogamy law only allows each individual to belong to a single couple. Newborn couplings instead may not only have multiplicity but individual may have children with different partners regardless of being married or not.

All of these rules, from the mathematical point of view, turn into topological constraints in the configurational space like the *hard-core* interaction of monogamy, or probabilistic constraints such as the concentration of children per couple around small integers. The rule structure can be described as follows: given a set of points $1, ..., N$, a configuration of marriages M is a set of links among the N points with the property that no points belong to more than one link. We indicate the unpaired individuals (singles) by S_M and the paired ones (married couples) by C_M. In principle, one should account for possible different numbers between the two genders; however, such refinements are small corrections of the main theory due to the heterogeneous origin of migrants, which balances polarized incomes from countries with eradicated cultural constraints. We call the set of marriage configurations $\mathcal{M}$. We want to describe a system in which we assign to each configuration a statistical weight and a partition function built on individual random variables as well as on couple random variables. Calling s_i the weight of the person i in the single state and $c_{i,j}$ that of the couple (i, j) in the married state (both the c's and the s's are positive real numbers), the partition function (grancanonical) of the system is given by

$$Z^{(\mathcal{M})} \;=\; \sum_{M\in\mathcal{M}} \prod_{(i,j)\in C_M} \epsilon_{i,j} c_{i,j} \prod_{i\in S_M} s_i \;, \tag{1}$$

where the numbers $\epsilon_{i,j} \in \{0, 1\}$ are the acquaintance matrix elements of the population, i.e. they specify the reciprocal knowledge among two individuals. It may be worth stressing that since no topological insights were available within our data we lack specific information on $\epsilon_{i,j}$. For this reason we carried our analysis with the diluted mean-field approach. While a fully connected mean-field approach would set a critical value $\Gamma_c > 0$, the assumption of an (over-percolated) underlying topology (e.g. a random graph a' la Erdos-Renyi [35] or a small world a' la Strogats-Watts [33]), implies a rescaling of the critical value with the size N [39][42][48], hence $\Gamma_c \to 0$, in full agreement with our experimental findings.

Similarly a configuration of filiations F is a set of links among the N points with the property that for a given couple (i, j) (not necessarily married) the number of children (links) is distributed according to (say) a Poisson distribution ρ of given average λ. The choice of the Poisson distribution is the most reasonable, but our conclusions are robust, pathological cases apart. We indicate individuals without children by U_F (undescended) and the couples with children by P_F (parents). We call the set of filiations $\mathcal{F}$. Calling u_i the weight of the person i in the undescended state and $p_{i,j}$ that of the child (i, j) in the parental state (both the u's and the p's are positive real numbers) the partition function of the system is then given by

$$Z^{(\mathcal{F})} \;=\; \sum_{F\in\mathcal{F}} \rho(F) \prod_{(i,j)\in P_F} \epsilon_{i,j} p_{i,j} \prod_{i\in U_F} u_i \;. \tag{2}$$

The random variables (c, s, p, u) are taken to be constant on mean field models like the one we treat explicitly in this work. Calling K_M the total number of links in the configuration M and defining the frequency as $\nu_M = K_M/(N/2)$, the expected value of the marriage frequency can be computed as

$$P_{\mathcal{M}} \;=\; \mathrm{Av} \frac{\sum_{M\in\mathcal{M}} \nu_M \prod_{(i,j)\in C_M} \epsilon_{i,j} c_{i,j} \prod_{i\in S_M} s_i}{\sum_{M\in\mathcal{M}} \prod_{(i,j)\in C_M} \epsilon_{i,j} c_{i,j} \prod_{i\in S_M} s_i} \;, \tag{3}$$

where the average operation Av is computed on the acquaintance matrix ensemble. Similarly for the newborn problem, calling K_F the total number of links in the configuration F and defining the fraction $\nu_F = K_F/(N\lambda/2)$, its expected

value is

$$P_{\mathcal{F}} \;=\; \mathrm{Av}\frac{\sum_{F\in\mathcal{F}}\nu_F\rho(F)\prod_{(i,j)\in P_F}\epsilon_{i,j}p_{i,j}\prod_{i\in U_F}u_i}{\sum_{F\in\mathcal{F}}\rho(F)\prod_{(i,j)\in P_F}\epsilon_{i,j}p_{i,j}\prod_{i\in U_F}u_i}\;. \tag{4}$$

For each population, the previous probability measure provides an average value of the two observables.
We now turn to the theory of bi-populated systems where s_i and u_i take two values each, depending only on the individual being on Imm or Nat and the couple variables ($c_{i,j}$ and $p_{i,j}$) take only three values for the three cases (Imm, Imm), (Nat, Imm) and (Nat, Nat). We may include an imitative ($J \geq 0$) interaction between the two populations with the introduction of a suitable mean-field Hamiltonian [35][36][30][31]

$$H(M) = -J_M \sum_{i\in Nat, j\in Imm} \epsilon_{i,j}\sigma_i\sigma_j \;, \tag{5}$$

where

$$\sigma_i = \begin{cases} +1 & \text{if } i \text{ belongs to a mixed marriage} \\ -1 & \text{otherwise} \;, \end{cases} \tag{6}$$

and a similar definition for $H(F)$:

$$H(F) = -J_F \sum_{i\in Nat, j\in Imm} \epsilon_{i,j}\tau_i\tau_j \;, \tag{7}$$

where

$$\tau_i = \begin{cases} +1 & \text{if } i \text{ has a child within a mixed couple} \\ -1 & \text{otherwise} \;. \end{cases} \tag{8}$$

Note that σ's and τ's configurations are uniquely determined by monomer-dimer configurations. We point out that the case where imitation and anti-imitation coexist [6] leads to a different scenario (see [19] for a case study). The two complete partition functions are then:

$$Z^{(\mathcal{M})} \;=\; \sum_{M\in\mathcal{M}} e^{-H(M)} \prod_{(i,j)\in C_M} \epsilon_{i,j}c_{i,j} \prod_{i\in S_M} s_i \;, \tag{9}$$

$$Z^{(\mathcal{F})} \;=\; \sum_{F\in\mathcal{F}} e^{-H(F)}\rho(F) \prod_{(i,j)\in P_F} \epsilon_{i,j}p_{i,j} \prod_{i\in U_F} u_i \;. \tag{10}$$

We point out that the introduction of an exponential Hamiltonian deformation of the monomer-dimer model is a working hypothesis to be tested against experimental data and it has the same significance of the logit distribution assumption in the original McFadden discrete choice theory. For a discussion about its paramount justification in terms of Entropy variational principles see [11] and references therein.

Calling M_M the number of mixed 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

$$P_{\mathcal{M}}^{(Nat,Imm)} = \mathrm{Av}\frac{\sum_{M\in\mathcal{M}} f_M e^{-H(M)} \prod_{(i,j)\in C_M}\epsilon_{i,j}c_{i,j}\prod_{i\in S_M}s_i}{\sum_{M\in\mathcal{M}} e^{-H(M)} \prod_{(i,j)\in C_M}\epsilon_{i,j}c_{i,j}\prod_{i\in S_M}s_i} \tag{11}$$

and analogously the probability of mixed children

$$P_{\mathcal{F}}^{(Nat,Imm)} = \mathrm{Av}\frac{\sum_{F\in\mathcal{F}} f_F e^{-H(F)}\rho(F) \prod_{(i,j)\in P_F}\epsilon_{i,j}p_{i,j}\prod_{i\in U_F}u_i}{\sum_{F\in\mathcal{F}} e^{-H(F)}\rho(F) \prod_{(i,j)\in P_F}\epsilon_{i,j}p_{i,j}\prod_{i\in U_F}u_i} \tag{12}$$

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 in the configuration F. Although an exact solution for the general model introduced in this section is not yet available, one can still obtain results for a wide variety of cases that include the mono-populated and bi-populated mean field limits [19][30][31]. The latter shows two regimes according to the ratio of J (the coupling J_M and J_F tuning the strength of the imitative behavior encoded in the Hamiltonians (14) and (16), hereafter called J for

simplicity) and the monomer-dimer pressure $p = \ln Z/N$. Given the lack of phase transitions in the Monomer-Dimer model (see [32, 44] for a rigorous proof of the non-random mean-field case and [45] for the random case), we can focus on the extreme regimes: the imitative regime $J \gg p$ in which the interaction J dominates on the Monomer-Dimer interaction (hard core or Poisson), and the free regime $J \ll p$ where the Monomer-Dimer interaction dominates on the imitative one. The structural difference between them is the presence of some divergence of the derivative of the P's in formulas (11) and (12)

$$\frac{\partial}{\partial \gamma} P(\gamma) \ . \tag{13}$$

In the free regime one finds for the P's a γ dependence of the type

$$P(\gamma) \ = \ c_F \gamma(1-\gamma) \ , \tag{14}$$

where the constant c_F depends only on the a priori probabilities of the Monomer-Dimer interaction. The expression can also be obtained from purely probabilistic (or combinatorial) reasoning and always displays a finite derivative in the origin.

In the imitation regime, on the other hand, the interactions among agents encoded in the Hamiltonians favor the imitative behavior, while the Monomer-Dimer interaction term, accounted for by the adjacency matrix of the reciprocal relation, plays the role of a phase-selecting perturbation (the + phase of the Hamiltonians) similar to a small magnetic field in spin models. The diluted mean-field Hamiltonian we introduced has a size proportional to $\gamma(1-\gamma)$, and for various adjacency matrices $\epsilon_{i,j}$ defining diluted topologies (i.e., random graphs, small worlds, etc. [38][35][42][46]), the model predicts a zero (or very small) critical γ_c and a functional behavior of the type:

$$P \ = \ c_I [\gamma(1-\gamma)]^{\frac{1}{2}} \ , \tag{15}$$

where the constant c_I, as much as the c_F in the free case, depends only on the a priori probabilities of the Monomer-Dimer interaction. The mechanism underlying such behavior is depicted, as far as the (social) network is over-percolated [47], and the interactions among agents are only imitative, by the mean-field ferromagnet with the critical exponent *one half*. Its relevance in social sciences has been clearly advocated by Durlauf [13]. Mathematically the critical value of γ turns out to be very close to zero as a consequence of dilution [35], [42], [46], [48]. Due to the large ensemble of data we analyzed, we can invoke the law of large numbers [20], which allows us to compare experimental frequencies reported in the first part of the paper, with probabilities obtained through the statistical mechanics method. The theoretical models illustrated reach the precise functional behavior of equation (15) in the limit of infinitely many particles (thermodynamic limit).

[1] The Global Approach to Migration and Mobility, EU Report, Commission 743, Sec 1353, (2011), http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0743:FIN:EN:PDF [Accessed 13 April 2013].

[2] *International Dialogue on Migration N. 17 - Migration and Social Change.* (International Organization for Migration, Geneva, 2011).

[3] Penninx, R., Spencer, D. & Van Hear, N., *Migration and Integration in Europe: The State of Research*, (Economic and social research council, Oxford, 2008).

[4] EU actions to make integration work: Common Basic Principles. http://ec.europa.eu/ewsi/en/EU_actions_integration.cfm, [Accessed 13 April 2013].

[5] 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. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2005:0389:FIN:EN:PDF, [Accessed 13 April 2013].

[6] Mézard, M., Parisi, G. & Virasoro, M.A., *Spin glass theory and beyond*, (World Scientific, Singapore, 1987).

[7] Mantegna, R.N. & Stanley, H.E., *Introduction to econophysics: correlations and complexity in finance*, (Cambridge University Press, 1999).

[8] Stanley, H.E., Econophysics and the Current Economic Turmoil, *APS News*, **11**, 17 (2008).

[9] Bouchaud, J.P. & Potters, M., *Theory of financial risks: from statistical physics to risk managemenent*, (Oxford University Press, 2004).

[10] Ballerini, M, & al., Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study, *Proc. Natl. Aca. Sc.* **105**, 1232-1237 (2008).

[11] Bialek, W., & al., Statistical mechanics for natural flocks of birds, *Proc. Natl. Acad. Sc.* **109**, 4786-4791 (2012).

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

[13] Durlauf, S.N., How can statistical mechanics contribute to social science?, *Proc. Natl. Acad. Sc.* **96**, 10582-1058 (1999).
[14] Durlauf, S.N., Statistical mechanics approaches to socioeconomic beavior, *NBER Technical WP 203*, Cambridge (1996).
[15] Brock, W. & Durlauf, S.N., Discrete choices with social interactions, *Rev. Econ. St.* **68**, 235-260 (2001).
[16] Montanari, A. & Saberi, A., The spread of innovations in social networks, *Proc. Natl. Acad. Sc.* **107**, 20196-20202 (2010).
[17] Castellano, C., Fortunato, S., & Loreto, V., Statistical physics of social dynamics, *Rev. Mod. Phys.* **81**, 591-646 (2009).
[18] Contucci, P. & Giardina', C., Mathematics and Social Sciences: A Statistical Mechanics Approach to Immigration, *ERCIM News* **73**, 8241 (2008).
[19] Barra, A. & Contucci, P., Toward a quantitative approach to migrants social integration, *Europhys. Lett.* **89**, 68001-68007 (2010).
[20] Feller, W., *An Introduction to Probability Theory and Its Applications*, (Wiley Series in Prob. and Math. Stat., 1968).
[21] Curie, P., Propriété ferromagnétique des corps a diverse temperatures, *Ann. de Chim. et de Phys.*, Series 7, **5** 289-405 (1895).
[22] Weiss, P., L' hypothése du champ moléculaire e la propriété ferromagnétique, *J. de Phys.*, **6** , 661-689 (1907).
[23] Weber, M., *Economy and Society: An outline of interpretative sociology*, [p.23], (California University Press, 1978).
[24] McFadden, D., Economic choices, *The Amer. Econ. Rev.* **91**, 351-378 (2001).
[25] Gallo, I., Barra, A. & Contucci, P., Parameter Evaluation of a Simple Mean-Field Model of Social Interaction, *Math. Mod. and Meth. in Appl. Sci.* **19**, 1427-1439 (2009).
[26] Bouchaud, J.P., Crises and collective socio-economic phenomena: simple models and challenges, *J. Stat. Phys*, **151**, 567-606 (2013).
[27] Byrne, D., *The attraction paradigm*, (Academic Press, New York, 1971).
[28] Hedstrom, P., Sandell, R., Stern, C., Meso-level Networks and the Diffusion of Social Movements, *Amer. J. Soc.* **106**, 145-172 (2001).
[29] Thompson, C.J., *Mathematical Statistical Mechanics*, Series of book in applied mathematics, (Princeton University Press, 1972).
[30] Contucci, P., Gallo, I. & Menconi, G., Phase transitions in social sciences: two-populations mean field theory, *Int. Jou. Mod. Phys. B* **22**, 14, 1-14 (2008).
[31] Contucci, P. & Gallo, I., Bipartite Mean Field Spin Systems. Existence and Solution, *Math. Phys. Elec. Jou.* **14**, 1, 1-22 (2008).
[32] Heilmann, O.J. & Lieb, E.H., Monomers and dimers, *Phys. Rev. Lett.* **24**, 25, 1412-1416 (1970).
[33] Watts, D.J. & Strogatz, S.H, Collective dynamics of small world networks, *Nature* **393**, 6684-6686 (1998).
[34] Sandell, R., Social Influences and Aggregated Immigration Dynamics: The Case of Spain 1999- 2009, *Int. Migrat. Rev.* **46**, 971-1004 (2012).
[35] Agliari, E. & Barra, A., A Hebbian approach to complex network generation, *Europhys. Lett.* **94**, 10002-10008 (2011).
[36] Barra, A. & Agliari, E., A statistical mechanics approach to Granovetter theory, *Physica A*, **391**, 10, 3017 (2012).
[37] Dodds, P.S., Muhamad, R. & Watts, D.J., An Experimental Study of Search in Global Social Networks, *Science* **301**, 827-829 (2003).
[38] Barra, A. & Agliari, E., Equilibrium statistical mechanics on correlated random graphs, *J. Stat. Mech.* **2011.02**, 02027 (2011).
[39] Barrat, A., Barthelemy, M, Pastor-Satorras, R. & Vespignani, A., The architecture of complex weighted networks, *Proc. Natl. Acad. Sc.* **101**, 11, 3747-3752 (2004).
[40] Barabasi, A.L. & Albert, R., Emergence of Scaling in Random Networks, *Science* **286**, 509-512 (1999).
[41] Liljeros, F., Edling, C.R., Amaral, L., Stanley, H.E. & Aberg, Y., The web of human sexual contacts, *Nature* **411**, 907-908 (2001).
[42] Bianconi, G., Mean field solution of the Ising model on a Barabasi-Albert network, *Phys. Lett. A* **303**, 166-168 (2002).
[43] Horst, U., Scheinkman, J.A., Equilibria in systems of social interactions, *J. Econ. Theory*, **130**, 44-47 (2006)
[44] Heilmann, O.J. & Lieb, E.H., Theory of monomer-dimer systems, *Comm. Math. Phys.* **25**, 190 (1972).
[45] Alberici, D. & Contucci, P., Solution of the monomer-dimer model on locally tree-like graphs. Rigorous results., *http://arxiv.org/abs/1305.0838* (2013).
[46] Leone, M., Vasquez, A., Vespignani, A. & Zecchina, R., Ferromagnetic ordering in graphs with arbitrary degree distribution, *Europ. Phys. Journ. B* **28**, 191-197 (2002).
[47] Bollobas, B., *Random Graphs*, (Cambridge University Press, 2000).
[48] Dembo, A. & Montanari, A., Ising models on locally tree-like graphs, *Ann. Appl. Probab.* **20**, 565-592 (2010).

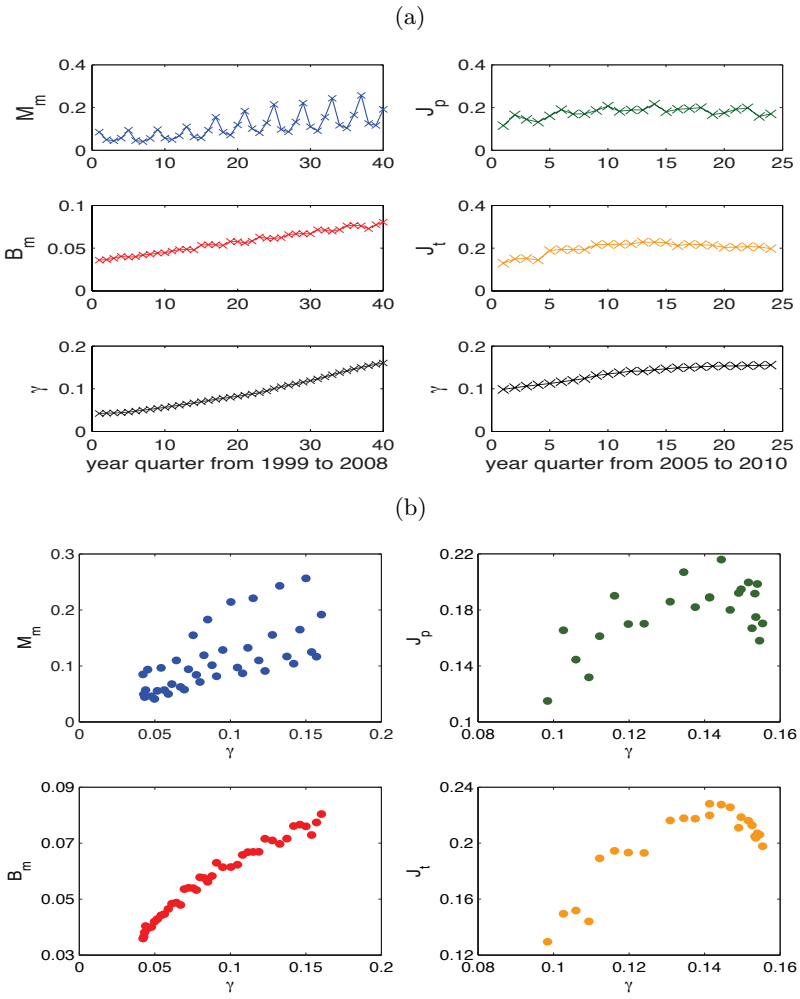

FIG. 1: Panel (a): time series representing the quantifiers **Q** (mixed marriages M_m, newborns from mixed parents B_m, permanent jobs J_p, temporary jobs J_t) and migrant's densities γ versus year quarters **t** in the two databases. Each point in the plot is the average value of Q in the quarter t. Panel (b): quantifiers **Q** versus γ obtained from time series of panel (a), i. e. $Q(t(\gamma))$, where $t(\gamma)$ is the inverse of $\gamma(t)$ (two lower panels of (a))

.

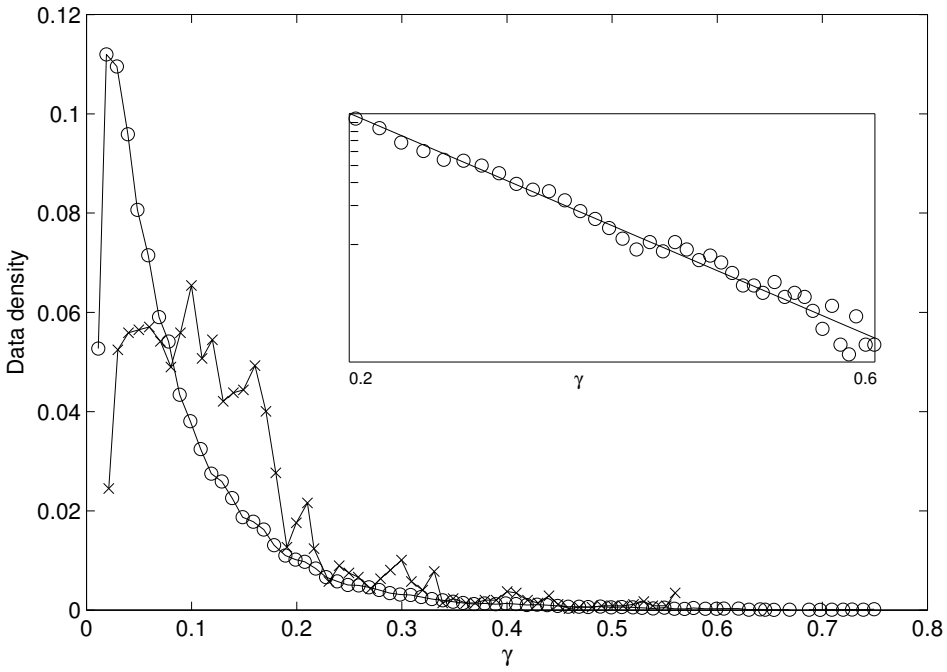

FIG. 2: Density of the marriage and newborn dataset (circles) and of the job market dataset (crosses) as a function of γ. In the inset the marriage and newborn data density is fitted, for $\gamma \geq 0.2$, with the power-law behavior (in log-log scale) where $\mu(\gamma) \propto \gamma^{\delta}$, $\delta = -3.241 \pm 0.024$.

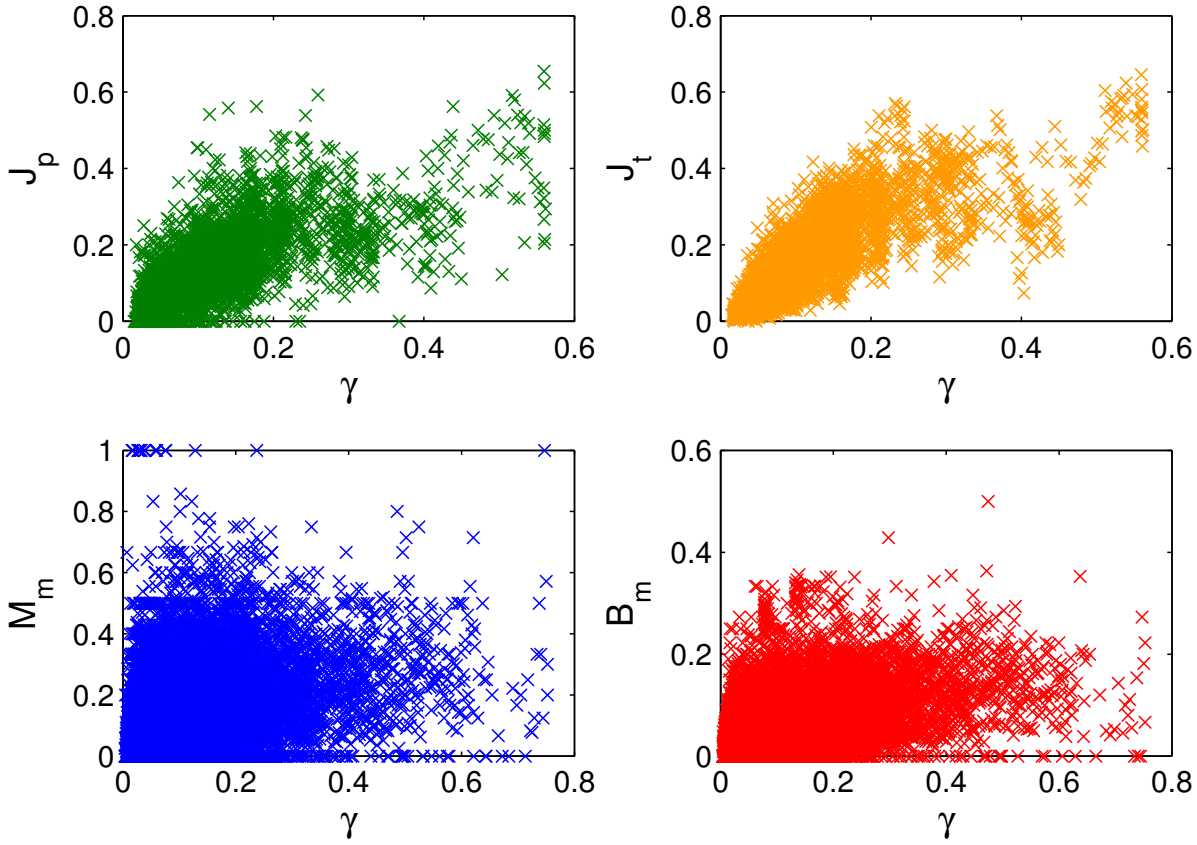

FIG. 3: Raw data versus γ. Green points represent the fraction of permanent job positions held by migrants in a municipality where a percentage γ of migrants is present (apart from restrictions outlined in the introduction, the whole of Spain is sampled over the entire analyzed timeframe); similarly orange points account for temporary jobs. Further, blue points represent the fraction of mixed marriages, while red ones mirror the newborns from mixed parents. 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 seasonal preferences in weddings. See the discussion within the paper.

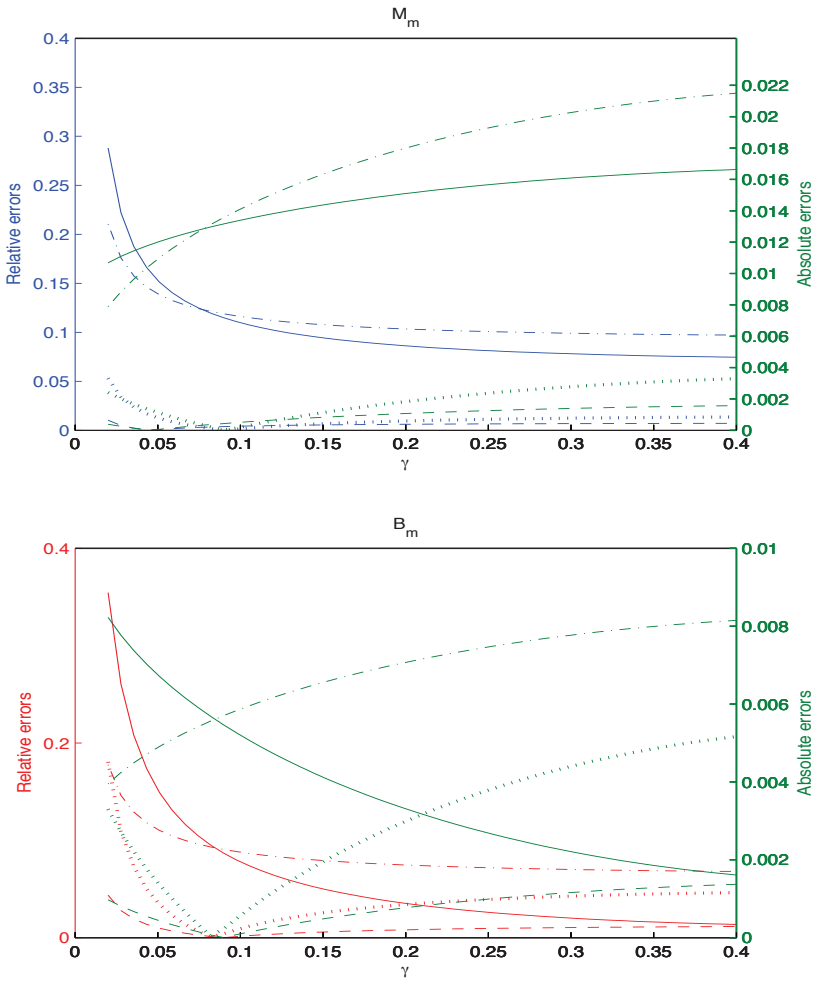

FIG. 4: Upper panel: Relative errors $|Q_i - Q_j|/Q_j$, $i \neq j$ $i,j \in$ {A,B,C,D} (blue lines) and absolute errors $|Q_i - Q_j|$, $i \neq j$ $i,j \in$ {A,B,C,D} (green lines) as a function of γ for $Q = M_m(\gamma)$. Lower panel: Relative errors $|Q_i - Q_j|/Q_j$, $i \neq j$ $i,j \in$ {A,B,C,D} (red lines) and absolute errors $|Q_i - Q_j|$, $i \neq j$ $i,j \in$ {A,B,C,D} (green lines) as a function of γ for $Q = B_m(\gamma)$. Q_A denotes the mean value using binning with constant information, Q_B the mean value using binning with constant width, Q_C the mediant with constant information binning and Q_D the mediant with constant width binning. For both panels the continuous lines refer to $i = B, j = D$ and the dash-dotted lines to $i = A, j = C$ and represent the errors made using the mean approach with respect to the mediant one with constant step binning and with constant information binning, respectively. The dotted lines ($i = D, j = C$) and the dashed lines ($i = B, j = A$) represent the errors made using the constant step binning with respect to the constant information, with the mean approach and with the mediant approach, respectively.

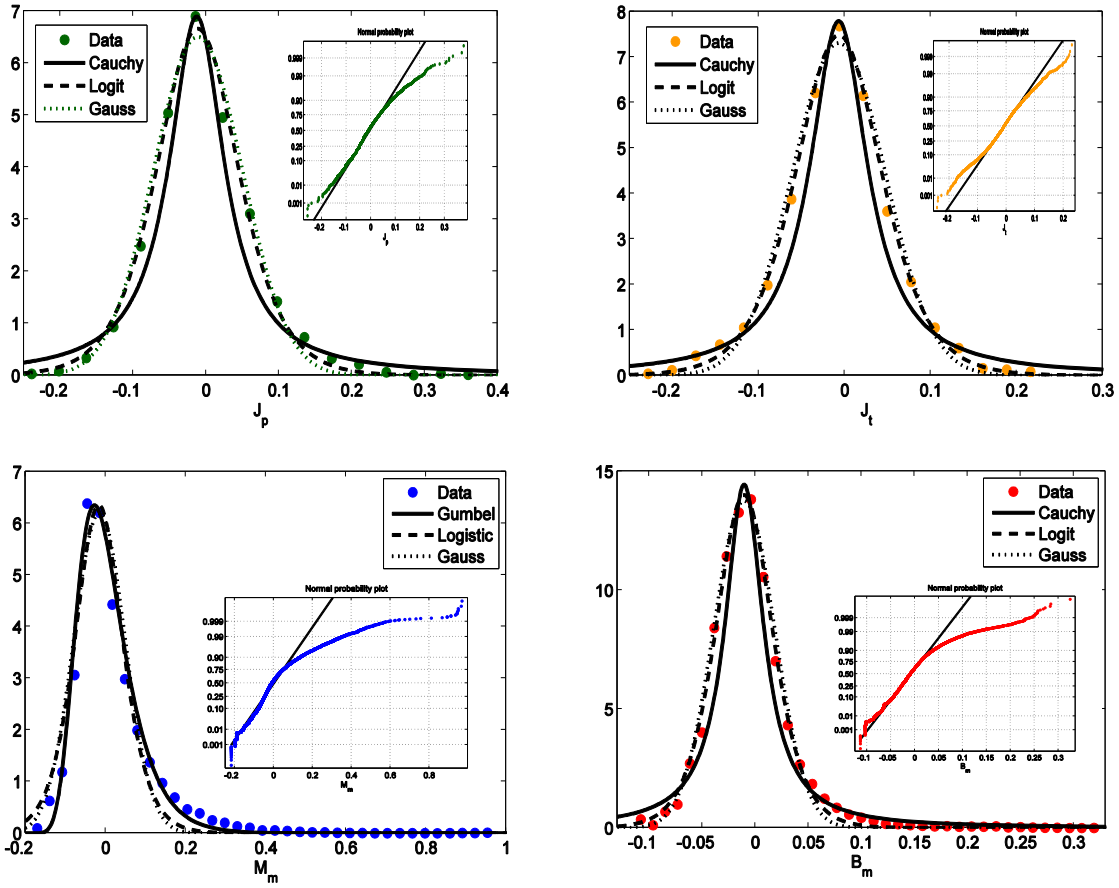

FIG. 5: Left upper panel: rescaled distribution of $J_p(\gamma)$. Fit with standard distribution laws are reported: Cauchy distribution $R^2 = 0.951$, Logistic distribution $R^2 = 0.994$ and Gaussian distribution $R^2 = 0.986$. Right upper panel: rescaled distribution of $J_t(\gamma)$. Fit with standard distribution laws are reported: Cauchy distribution $R^2 = 0.956$, Logistic distribution $R^2 = 0.996$ and Gaussian distribution $R^2 = 0.984$. Left lower panel: rescaled distribution of $M_m(\gamma)$. Fit with standard distribution laws are reported: Gumbel distribution $R^2 = 0.979$, Logistic distribution $R^2 = 0.956$ and Gaussian distribution $R^2 = 0.941$. Right lower panel: rescaled distribution of $B_m(\gamma)$. Fit with standard distribution laws are reported: Cauchy distribution $R^2 = 0.958$, Logistic distribution $R^2 = 0.994$ and Gaussian distribution $R^2 = 0.988$. For all panels the inset represents the normal probability plot.